İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN EDEBİYAT FAKÜLTESİ

ORACLE DATA MINER İLE ÖĞRENCİ KAYITLARI ÜZERİNE BİR VERİ MADENCİLİĞİ UYGULAMASI

BİTİRME ÖDEVİ

Ceyda DURMAZ
Murat KOCAMIŞ

Anabilim Dalı: MÜHENDİSLİK BİLİMLERİ

Tez Danışmanı: Öğr. Gör. Dr. Ahmet KIRIŞ

MAYIS 2008

Ceyda DURMAZ Murat KOCAMIŞ	**ORACLE DATA MINER İLE ÖĞRENCİ KAYITLARI ÜZERİNE BİR VERİ MADENCİLİĞİ UYGULAMASI**	**2008**
Ceyda DURMAZ Murat KOCAMIŞ	**ORACLE DATA MINER İLE ÖĞRENCİ KAYITLARI ÜZERİNE BİR VERİ MADENCİLİĞİ UYGULAMASI**	**2008**
Ceyda DURMAZ Murat KOCAMIŞ	**ORACLE DATA MINER İLE ÖĞRENCİ KAYITLARI ÜZERİNE BİR VERİ MADENCİLİĞİ UYGULAMASI**	**2008**
Ceyda DURMAZ Murat KOCAMIŞ	**ORACLE DATA MINER İLE ÖĞRENCİ KAYITLARI ÜZERİNE BİR VERİ MADENCİLİĞİ UYGULAMASI**	**2008**
Ceyda DURMAZ Murat KOCAMIŞ	**ORACLE DATA MINER İLE ÖĞRENCİ KAYITLARI ÜZERİNE BİR VERİ MADENCİLİĞİ UYGULAMASI**	**2008**

ÖNSÖZ

Bu çalışmayı hazırlarken, bizden yardımını ve desteğini hiçbir zaman esirgemeyen Sayın Hocamız Öğr. Gör. Dr. Ahmet KIRIŞ' a, başta Kaan YAREN, Mehmet AYDOĞAN, M. Zahid ARIKAN, Pınar ÖZTÜRK ve Serkan KARAKAVAK olmak üzere tüm değerli arkadaşlarımıza, hayatımız boyunca bize sevgi, güven ve her türlü desteği veren ailelerimize en içten teşekkürlerimizi sunarız.

Mayıs, 2008

Ceyda DURMAZ

Murat KOCAMIŞ

İÇİNDEKİLER

ÖZET

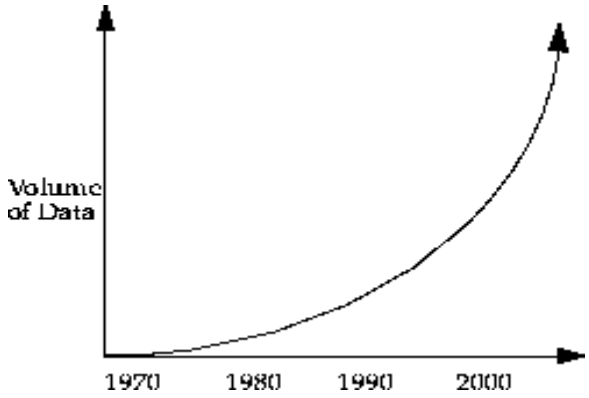
Bu çalışmada öğrenci kayıtları üzerine bir veri madenciliği uygulaması yapılmıştır. Bu uygulama ile İstanbul Teknik Üniversitesi' nde verilmekte olan matematik havuz derslerinin bölümlere ve öğretim üyelerine göre başarı oranlarının hesaplanarak bir veri madenciliği modeli oluşturulması amaçlanmaktadır. Ancak burada amaç bu modelin kullanılarak tüm yeni öğrencilerin başarı notlarının tahmin edilmesi değildir ve nitelikli bir öğretim sisteminde verilen bilgilerle bunun tahmin edilemeyeceği de açıktır. Erişilmek istenen amaç yüksek güvenilirlikte böyle bir tahmin yapılabilecek sıra dışı durumların varlığını araştırmaktır.

Bu amaçla öğrenci kayıtları alınmış, veriler ayıklanarak derslere göre tablolara ayrılmıştır. Not sisteminin yapısından dolayı veri madenciliği probleminin sınıflandırma modeline uygun olduğu belirlenmiştir. Bu problemi çözmek için de sınıflandırma modelinin algoritmalarından biri olan Naive Bayes seçilmiştir.

Veri madenciliği uygulamasını gerçekleştirmek için ilk olarak sunucu üzerine Oracle veritabanının 11g sürümü, istemci üzerine ise Oracle Data Miner paket programı yüklenmiştir. Daha sonra oluşturulan tablolar veritabanına aktarılmış ve gerekli modeller oluşturulmuştur. Bu modeller yardımıyla da istenilen tahminler elde edilerek, bulunması hedeflenen bazı sıra dışı durumların varlığı gözlenmiştir.

1. GİRİŞ

Son 20 yıl içerisinde elektronik ortamda saklanan veri miktarında büyük bir artış meydana gelmiştir. Yeryüzündeki bilgi miktarı her 20 ayda bir iki katına çıkmakta, buna bağlı olarak yeni veritabanı sayısı da hızla artmaktadır (Şekil.1.1). Süpermarket alışverişi, banka kartları kullanımı, telefon aramaları gibi günlük hayatta kullanılan birçok faaliyetin yanında birçok farklı bilim disiplininden elde edilen veriler, hava tahmini simülasyonu, sanayi faaliyet testleri büyük veritabanlarında kayıt altına alınmaktadır.



Şekil.1.1: Veri miktarındaki artış [1]

Yüksek kapasiteli işlem yapabilme gücünün ucuzlaması ile birlikte, veri elde etme ve saklama ucuzlamaya başlamıştır. Bilginin iş dünyasında çok değerli olduğu ve iş dünyasının kalbini oluşturduğu, karar alıcıların bu bilgi ışığında şirket stratejileri geliştirdiği aşikardır. Veritabanı yönetim sistemleri, geleneksel sorgulama yöntemlerini kullanarak eldeki verilerden sınırlı çıkarım yapabilmektedir. Geleneksel çevrimiçi işlem sistemleri (on-line transaction processing systems), OLTP, veritabanındaki bilgiye hızlı, güvenli ve verimli erişim imkanı sağlamakta, fakat veriden analizler yapılarak anlamlı çıkarımlar elde edilmesini sağlayamamaktadır. Büyük veri yığınları içerisinden anlamlı çıkarımlar elde etme ihtiyacı, mevcut yöntemlerin bu ihtiyaca yeterli çözümü sunamaması uzmanları yeni arayışlara yöneltmiştir. Geleneksel yöntemlerle elde edilemeyen bu bilgiye erişebilmek için,

bilgi keşfi (Knowledge Discovery in Databases-KDD) adı altında çalışmalar yürütülmüş ve bunun bir sonucu olarak veri madenciliği (Data Mining) kavramı ortaya çıkmıştır. Veri madenciliğinin temel amacı, çok büyük veri tabanlarındaki ya da veri ambarlarındaki veriler arasında bulunan ilişkiler, örüntüler, değişiklikler, sapma ve eğilimler, belirli yapılar gibi bilgilerin matematiksel teoriler ve bilgisayar algoritmaları kombinasyonları ile ortaya çıkartılması ve bunların yorumlanarak değerli bilgilerin elde edilmesidir.

1.1. Tanım

Günümüzde sadece veriye ulaşmak değil, eldeki veriden maksimum bilgiyi üretmek de önemli bir olgu halini almıştır. Veri madenciliği bu konuda diğer teknolojilerden bir adım öndedir. Veri madenciliği kavramı ile ilgili bazı tanımlar aşağıda verilmiştir:

1. "Veri madenciliği, veritabanında bilgi keşfi (KDD) eldeki verilerden, önceden bilinmeyen fakat potansiyel olarak yararlı olabilecek bilgileri çıkarmaktır. Bu kümeleme, veri özetlemesi, öğrenme sınıflama kuralları, değişikliklerin analizi ve sapmaların tespiti gibi birçok farklı teknik bakış açısını içine alır [2].

2. "Veri madenciliği anlamlı kuralların ve örüntülerin bulunması için geniş veri yığınları üzerine yapılan keşif ve analiz işlemleridir." [3].

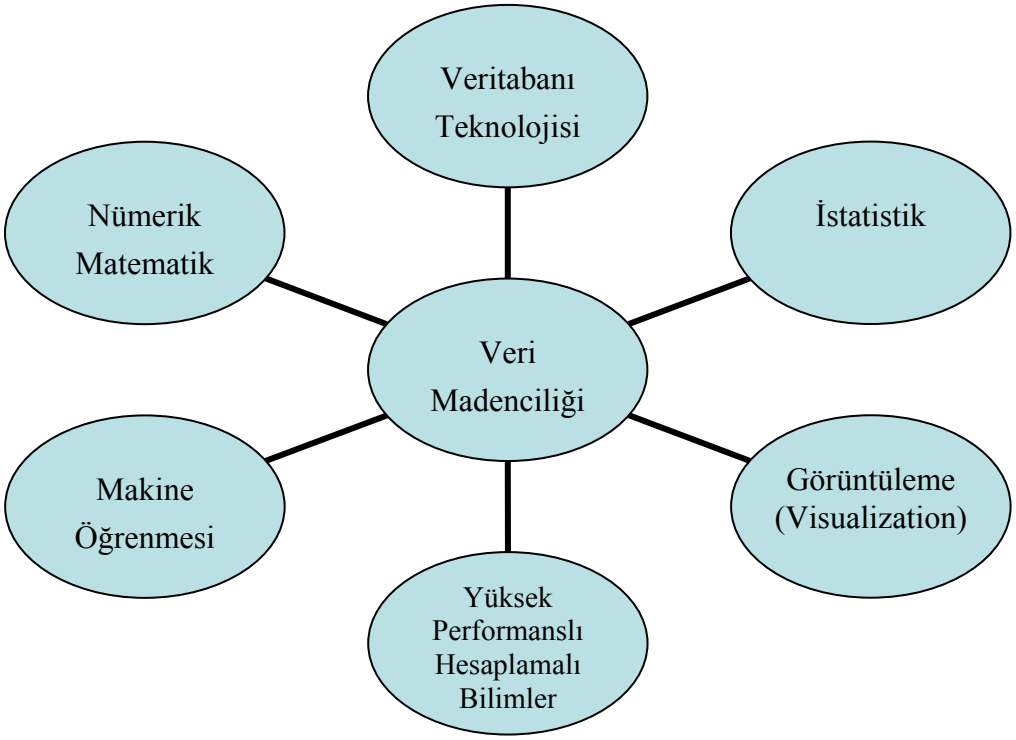
3. "Veri madenciliği çok büyük veritabanları içindeki veriler arasındaki bağlantılar ve örüntüleri araştırarak, gizli kalmış yararlı olabilecek verilerden değerli bilginin çıkarılması sürecidir." [4].

4. "Veri madenciliği veri ambarlarında tutulan çok çeşitli verilere dayanarak daha önceden keşfedilmemiş bilgileri ortaya çıkarmak, bunlara karar vermek ve eylem planını gerçekleştirmek için kullanma sürecidir." [5].

Bu tanımlardan hareket ederek veri madenciliği kavramına şu şekilde bir tanım yapılabilir: Veri madenciliği, bilgi keşfi büyük veri yığınları içerisinde veriler arasındaki ilişkiler üzerinde matematiksel yöntemler aracılığıyla analiz yapma; farkında olunmayan, gizli, ancak işe yarayabilecek verilerin ortaya çıkarılması ve bu

veriler içerisindeki bağıntıların, örüntülerin modellenerek işe yarar bilgi çıkarılması sürecidir.

Veri madenciliğinin temel olarak dört ana başlıktan oluştuğu kabul edilmiştir. Bunlar sınıflama, kategori etme, tahmin etme ve görüntülemedir. Veri madenciliği istatistik, makine bilgisi, veritabanları ve yüksek performanslı işlem gibi temelleri de içerir. Bu yapı aşağıdaki şekilde gösterilmiştir.



Şekil 1.2. Veri madenciliği disiplinler arasıdır [6]

1.2. Tarihsel Gelişim

Veri madenciliği kavramının ortaya çıkışı 1960lı yıllara kadar dayanmaktadır. İlk olarak veri taraması (data dredging), veri yakalanması (data fishing) gibi adlandırmalar kullanılmış ve bilgisayar yardımıyla gerekli sorgulama yapıldığında istenilen bilgiye ulaşılacağı sonucu kabul edilmiştir. 1990lara gelindiğinde veri madenciliği ismi ortaya atılmış ve bu yeni kavrama neden olan olgunun da geleneksel istatistiksel yöntemler yerine veri analizinin algoritmik bilgisayar modülleri tarafından değerlendirilmesi olduğu kabul edilmiştir. Veri madenciliğinin tarihsel evrimi Tablo.1.1 de gösterilmiştir:

Tablo.1.1 Veri madenciliğinin tarihsel gelişimi

Tarih	Basamaklar	Sorular	Kullanılabilir Teknolojiler	İlgili Yazılımlar
1960lar	Veri toplama, Veritabanı Yönetim Sistemleri	Benim son 5 yıldaki toplam kârım nedir?	Bilgisayar, Disk, Düz dosyalar	Fortran
1980ler	Veriye ulaşım, Veri sorgulama	Geçen Mart İstanbul'daki birim satış miktarı nedir?	Daha hızlı ve ucuz bilgisayarlar, daha fazla depolama alanı, ilişkisel veritabanları	Oracle,IBM DB, SQL
1990lar	Veri ambarları, Karar destek sistemleri	Geçen Mart İstanbul'daki birim satış miktarı nedir? Ankara ile karşılaştırmalı olarak görmek istiyorum.	Daha hızlı ve ucuz bilgisayarlar, Daha fazla depolama alanı, İlişkisel veritabanları, OLAP, Çok boyutlu veritabanları, Veri ambarları	SQL Standart, Veri Ambarları, OLAP, Darwin, IBM Intelligent Miner, SPSS Crisp DM, SAS Miner, Angoss Knowledge Seker
1990ların sonu 2000ler	Veri madenciliği Web madenciliği	Ankara'da gelecek ayki birim satışlarım ne durumda olacak ? Neden?	Daha hızlı ve ucuz bilgisayarlar, Daha fazla depolama alanı, İlişkisel veritabanları, Gelişmiş bilgisayar algoritmaları	Oracle Data Miner, IBM DB2 UDB Mining, SPSS Clementine, SAS Enterprise Miner

1.3. Kullanım Alanları

Gelişen teknoloji ile birlikte veriler çok hızlı bir şekilde toplanabilmekte, depolanabilmekte, işlenebilmekte ve bilgi olarak kurumların hizmetine sunulabilmektedir. Günümüzde bilgiye hızlı erişim özellikle ani ve maksimum kazancı sağlayacak karar vermeyi gerektiren iş dünyasında çok büyük önem arz etmekte ve bunun için birçok araştırma yapılmaktadır. Araştırmacılar, büyük hacimli ve dağınık veri setleri üzerinde çalışmalar yapmış, hızlı ve güvenli bilgi ihtiyacını karşılayabilmek için veri madenciliği üzerine yoğunlaşmışlardır. Aşağıda veri madenciliğinin kullanım alanları belli bir hiyerarşi içerisinde verilmiştir [1].

1.3.1. Perakende/ Pazarlama

- Tüketicilerin tüketim eğilimlerinin belirlenerek, alım alışkanlıklarının belirlenmesi
- Tüketicilerin demografik karakteristikleri arasındaki ilişkilerin belirlenmesi
- E-posta kampanyalarına tepkinin tahmin edilmesi
- Pazar analizi yapılarak piyasaya sürülecek bir ürüne verilecek tepkilerin tahmin edilmesi

1.3.2 Bankacılık

- Kredi kartı kullanımında oluşabilecek dolandırıcılık durumlarının tespiti
- Bankaya sadık müşterilerin tespiti
- Kart kullanımı profili oluşturmak, müşteri kart kullanımındaki değişiklikleri tahmin edilmesi
- Kullanıcı gruplarının kredi kartı harcamalarını saptanması
- Farklı finansal göstergeler arasındaki gizli korelasyonun bulunması
- Tarihsel pazar verileri kullanılarak belirli kuralların oluşturulması

1.3.3. Sağlık Hizmetleri ve Sigortacılık

- Sigorta poliçesi üzerinden ödenecek para analizi yapılması
- Hangi müşterilerin yeni bir sigorta poliçesi alacağının tahmin edilmesi
- Riskli müşterilerin davranış kalıplarının tespit edilmesi
- Dolandırıcılık davranışlarının tanımlanması

1.3.4 Tıp

- Hasta davranışlarının tahmin edilerek karakterize edilmesi
- Farklı hastalıklar üzerinde yapılan başarılı tıbbi terapilerin tanımlanması
- Demografik ve tarihi veriler ışığında bölgelerin incelenerek potansiyel hastalık tehlikelerinin tahmin edilmesi

1.3.5. Ulaştırma

- Dağıtım listelerine karar verilmesi, araçların dağıtım kanallarının belirlenmesi
- Yük modeli analizinin yapılması ve bunun sonucunda yükleme durumunun saptanması

1.3.6. Eğitim

- Eğitim modelleri ve öğrenci başarı durumları incelenerek, eğitimde başarıyı artırıcı durumların saptanması
- Öğretmen-ortam-öğrenci ilişkisi içerisinde verimlilik artışı için tahminler üretilmesi

1.3.7. Ekonomi

- Eldeki verilerin incelenerek saptamalarda bulunulması, ekonomik eğilim ve düzensizliklerin tespiti
- Eldeki verilerin analiz edilerek ülke ekonomisi için ekonomik politikalar oluşturulması, senaryo testi tahmini yapılması

1.3.8. Güvenlik

- Uydulardan gelen uzaysal datanın ve görüntülerin değerlendirilerek düşman kuvvetlerin nerelerde hangi araç ve teçhizatlarda yoğunlaştığının ve konuşlanmaya uygun arazi yapılarının belirlenmesi
- Internet sayfalarının anahtar kelimelerle taranarak lehte ve aleyhte propaganda yapan sayfaların belirlenmesi
- Haberleşme araçları takip edilerek terörist faaliyetlerin belirlenmesi

1.4. Veri Madenciliği Modelleri

IBM, veri işleme operasyonları için iki çeşit model tanımlamıştır [1].

1.4.1. Doğrulama Modeli

Doğrulama modeli kullanıcıdan bir hipotez alarak, testler yaparak bu hipotezin geçerliliğini araştırır.

1.4.2. Keşif Modeli

Keşfetme modelinde sistem önemli bilgileri otomatik olarak gizli veriden elde eder. Veri yaygın olarak kullanılan modeller, genelleştirmeler ile ayıklanır ve başka bir aracıya gereksinim duyulmaz.

1.5. Veri Madenciliği Uygulamaları İçin Temel Adımlar

Veri madenciliği uygulamalarında izlenmesi gereken temel aşamalar aşağıda sistematik bir şekilde verilmiştir.

1.5.1. Uygulama Alanının Ortaya Konulması

Bu adımda veri madenciliğinin hangi alanda ve hangi amaç için yapılacağı belirlenir.

1.5.2. Hedef Veri Grubu Seçimi

Belirlenen amaç doğrultusunda belirli kriterler çerçevesinde aynı veya farklı veritabanlarından veriler toplanarak hedef veri grubu seçilir.

1.5.3. Model Seçimi

Veri madenciliği problem veya problemlerinin seçimi yapılır. (Sınıflandırma, Kümelendirme, Birliktelik Kuralları, Şablonların ve İlişkilerin Yorumlanması v.b.)

1.5.4. Ön İşleme

Bu aşamada seçilen veriler temizlenir, gereksiz veriler ayıklanarak silinir, kayıp veri alanları ile ilgili stratejiler belirlenir, veriler yeniden düzenlenerek tutarlı bir hale getirilir. Kısacası bu aşamada uyumlandırma işlemi yapılır. (Data temizleme ve data birleştirme)

1.5.5. Dönüştürme / İndirgeme

Verinin kullanılacak modele göre içeriğinin korunarak şeklinin dönüştürülmesi işlemidir.

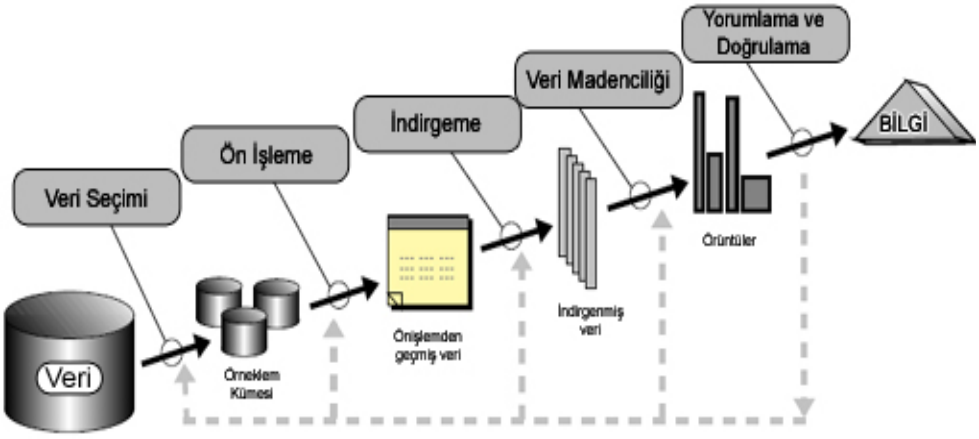
1.5.6. Algoritmanın Belirlenmesi

Bu aşamada indirgenmiş veriye ve kullanılacak modele hangi algoritmanın uygulanacağına karar verilir, seçime uygun veri madenciliği yazılımı seçilir veya algoritmaya uygun program geliştirilir.

1.5.7. Yorumlama ve Doğrulama

Uygulanan yöntemler sonucunda elde edilen veriler üzerine yorumlarda bulunulur, bu yorumlar test edilerek doğrulanır ve elde edilen çıkarımlar sonucunda işe yarayacak bilgiye ulaşılır.

Veri madenciliği aracılığıyla bilginin bulunması süreci Şekil.1.3 te görsel olarak verilmiştir.



Şekil.1.3. Bilgi keşfi sürecinde veri madenciliğinin yeri [7]

1.6. Temel Veri Madenciliği Problemleri ve Çözüm Yöntemleri

Bilgi keşfi sürecinde karşılaşılan farklı tipteki problemler, farklı veri madenciliği algoritmaları ile çözülmektedir. Genellikle, veri madenciliği görevleri iki başlık altında toplanır.

- Tanımlayıcı veri madenciliği görevleri eldeki verinin genel özelliklerinin belirlenmesidir.

- Kestirimci/Tahmin edici veri madenciliği görevleri ulaşılabilir veri üzerinde tahminler aracılığıyla çıkarımlar elde etmek olarak tanımlanmıştır.

Veri madenciliği algoritmaları aşağıda açıklanmış ve Şekil.1.4' te gösterilmiştir [8-9].

1.6.1. Karakterize Etme (Characterization)

Veri karakterizasyonu hedef sınıfındaki verilerin seçilmesi, bu verilerin genel özelliklerinin belirlenmesi ve bu özellikler sayesinde karakteristik kuralların oluşturulması olayıdır.

Örnek: Perakende sektöründe faaliyet gösteren, uluslararası XYZ şirketinin binlerce kurumsal müşterisi olsun. XYZ şirketinin pazarlama biriminde, büyük kurumsal müşterilere yönelik kampanyalar için her yıl düzenli olarak bu şirketten 10 milyon YTL ve üstü alım yapan kurumsal müşteriler hedeflenmektedir. Veritabanından hedef grup belirlenerek genelleme yapılır ve genel kurallar oluşturulur.

1.6.2. Ayrımlaştırma (Discrimination)

Hedef sınıf ve karşıt sınıf elemanlarının özellikleri arasında karşılaştırma yapılmasını sağlar. Karakterize etme adımından farkı mukayese yöntemini kullanmasıdır.

Örnek: XYZ kurumsal müşterilerinden her yıl 10 milyon YTL ve üstü alışveriş yapan fakat geri ödeme konusunda riskli olan müşteri grubunun belirlenmesi

1.6.3. Sınıflandırma (Classification)

Eldeki sorgulanmış veriler sınıflandırma algoritması kullanılarak sınıflandırma kuralları oluşturulur. Sonra oluşturulan bu sınıflandırma kuralları kullanılarak veriler sınıflandırılır. Böylece daha sonradan girilecek veriler önceden tanımlanmış bu sınıflardan karakteristik özellik bakımından uygun olan sınıfa eleman olarak atanır. Sınıflandırma problemleri için "Oracle Data Miner" (ODM) ın desteklediği çözüm yöntemleri Naive Bayes (NB), Karar Destek Vektörleri (SVM), Karar Ağaçları ve Lojistik Regresyon (GLM) dir.

Örnek : XYZ şirketi müşterilerinin alım durumlarını göz önünde bulundurarak, alım gücüne göre "Yüksek", "Orta", "Düşük" şeklinde sınıflandırır. Müşterilerinin risk durumlarını sınıflandırmak için de "Risksiz", "Riskli", "Çok Riskli" şeklinde etiketlerle sınıflandırılabilir.

1.6.4. Tahmin Etme (Prediction)

Geçmiş kayıtların analizi sonucu elde edilmiş bazı bilgiler kullanılarak gelecekte oluşacak durumların tahmininin ve trend analizinin yapılmasıdır. Örneğin XYZ şirketi geçen yılın satışlarını bölge bazlı sınıflandırmış ve bu sene için bir trend analizi yaparak her bölgede oluşacak talebi tahmin etmiştir. Bu tür problemler için ODM nin kullandığı regresyon analizi yöntemi SVM dir.

1.6.5. Birliktelik Kuralları (Association Rules)

Birliktelik kuralları gerek birbirini izleyen gerekse eş zamanlı durumlarda araştırma yaparak, bu durumlar arasındaki ilişkilerin tanımlanmasında kullanılır. Bu model yaygın olarak Market Sepet Analizi uygulamalarında kullanılmaktadır. Örneğin bir süpermarkette X ürününden alan müşterilerin büyük bir kısmı Y ürününden de almıştır. Birliktelik kuralı ile bu durum ortaya çıkarılarak, süpermarketin X ve Y ürününü aynı veya yakın raflara koyması sağlanır. ODM bu problem sınıfı için de Birliktelik Kuralları modelini kullanmaktadır.

1.6.6. Kümeleme (Clustering)

Yapı olarak sınıflandırmaya benzeyen kümeleme metodunda birbirine benzeyen nesnelerin aynı grupta toplanarak kümelenmesi sağlanır. Sınıflandırma metodunda oluşturulan sınıfların kuralları, sınırları ve çerçevesi bellidir ve veriler bu kriterlere göre sınıflara atanır. Kümeleme metodunda ise sınıflar arası bir yapı mevcut olup, benzer özellikte olan verilerle yeni gruplar oluşturmak temel esastır. ODM "K-means" ve "O-Cluster" kümeleme yöntemlerini desteklemektedir.

1.6.7. Aykırı Değer Analizi (Outlier Analysis)

İstisnalar veya sürprizler olarak da isimlendirilen aykırı değerler, bir sınıf veya kümelemeye tabii tutulamayan/atanamayan, belirli bir gruba dahil olmayan veri tipleridir. Aykırı değerler bazı problemlerde gürültü, atılması gereken değerler olarak ele alınırken bazı problemlerde ise aykırı değerler çok önemli bilgiler olarak değerlendirilebilmektedir.

Örneğin bir markette müşterilerin hep aynı ürünü iade etmesi bu metodun araştırma konusu içine girer. ODM temizleme, eksik değer, aykırı değer analizi gibi birçok yöntemi veri hazırlama aşaması içine almakta ve desteklemektedir.

1.6.8. Zaman Serileri (Time Series)

Birçok uygulamada kullanılan veriler statik değildir ve zamana bağlı olarak değişiklikler göstermektedir. Bu analiz tipi belirli bir veya daha fazla özelliğin belli bir zaman aralığı içerisinde değişimi, zamana bağlı eğilimdeki sapma gibi problemleri inceler. Bir zaman aralığında ölçülebilir değerler ve tahmin edilen/ beklenen değerleri karşılaştırmalı olarak inceleyerek, sapmaları belirler.

Örneğin XYZ şirketinin Ocak-Haziran 2008 dönemi için önceki yılın satış miktarları göz önünde tutularak bir hedef ortaya konulmuştur. 2008 ve 2007 değerleri karşılaştırmalı olarak incelenerek sapma miktarı belirlenir. ODM her ne kadar çeşitli histogramlarla kullanıcıya görsel destek sağlasa da tam anlamıyla bu tür problemleri desteklememektedir.

1.6.9. Veri Görüntüleme (Visualization)

Çok boyutlu veriler içerisindeki karmaşık bağlantıların/bağıntıların görsel olarak yorumlanabilmesi problemini inceler. Grafik araçları veri ilişkilerini görsel/grafiksel olarak sunar. ODM zaman serilerinde olduğu gibi histogramlarla bu problem grubunu kısmen desteklemektedir.

1.6.10. Yapay Sinir Ağları (Artificial Neural Networks)

Yapay sinir ağları insan beyninden esinlenerek geliştirilmiş, ağırlıklı bağlantılar aracılığıyla birbirine bağlanan işlem elemanlarından oluşan paralel ve dağıtılmış bilgi işleme yapılarıdır. Yapay sinir ağları öğrenme yoluyla yeni bilgiler türetebilme ve keşfedebilme gibi yetenekleri hiçbir yardım almadan otomatik olarak gerçekleştirebilmek için geliştirilmişlerdir. Yapay sinir ağlarının temel işlevleri arasında veri birleştirme, karakterize etme, sınıflandırma, kümeleme ve tahmin etme gibi veri madenciliğinde de kullanılan metotlar mevcuttur. Yüz ve plaka tanıma sistemleri gibi teknolojiler yapay sinir ağları kullanılarak geliştirilen teknolojilerdendir.

1.6.11. Genetik Algoritmalar (Genetic Algorithms)

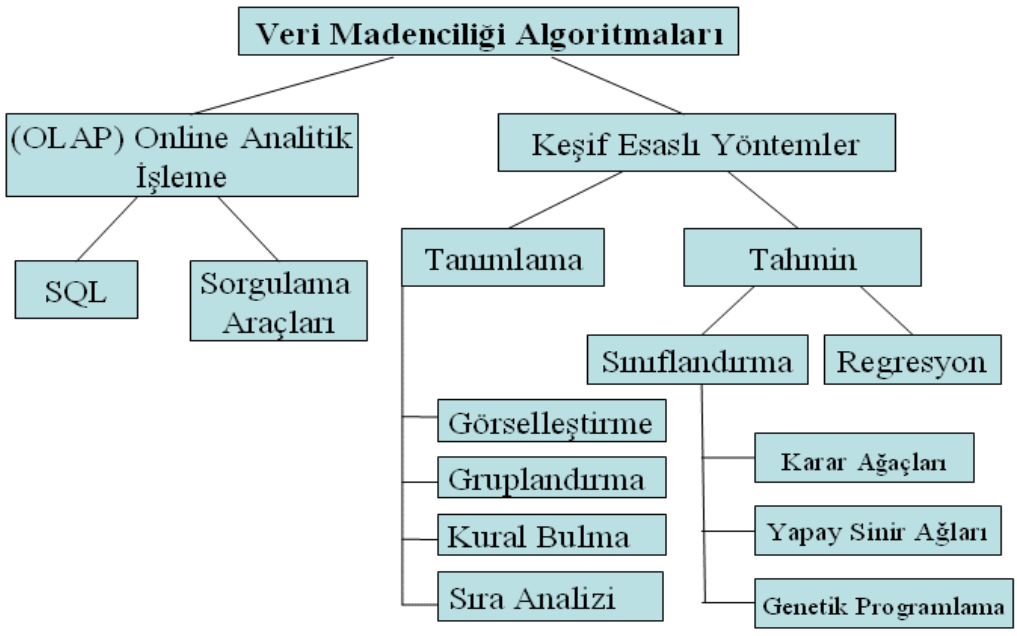
Genetik algoritmalar doğada gözlemlenen evrimsel sürece benzeyen, genetik kombinasyon, mutasyon ve doğal seçim ilkelerine dayanan bir arama ve optimizasyon yöntemidir. Genetik algoritmalar parametre ve sistem tanılama, kontrol sistemleri, robot uygulamaları, görüntü ve ses tanıma, mühendislik tasarımları, yapay zeka uygulamaları, fonksiyonel ve kombinasyonel eniyileme problemleri, ağ tasarım problemleri, yol bulma problemleri, sosyal ve ekonomik planlama problemleri için diğer eniyileme yöntemlerine kıyasla daha başarılı sonuçlar vermektedir.

1.6.12. Karar Ağaçları (Decision Trees)

Ağaç yapıları esas itibarıyla kural çıkarma algoritmaları olup, veri kümelerinin sınıflanması için "if- then" tipinde kullanıcının rahatlıkla anlayabileceği kurallar inşa edilmesinde kullanılırlar. Karar ağaçlarında veri kümesini sınıflamak için "Classification and Regression Trees (CART)" ve "Chi Square Automatic Interaction Detection (CHAID) " şeklinde iki yöntem kullanılmaktadır.

1.6.13. Kural çıkarma (Rules Induction)

Veri temelinde istatistiksel öneme sahip yararlı "if-else" kurallarının ortaya çıkarılması problemlerini inceler.



Şekil.1.4. Veri Madenciliği Algoritmaları [8]

1.7. Uygulama Ana Hatları, Kapsam ve Amaç

Bu bitirme çalışmasında yukarıda anlatılanlar çerçevesinde bir veri madenciliği uygulaması yapılacaktır. Bu uygulama ile İstanbul Teknik Üniversitesi' nde verilmekte olan matematik havuz derslerinin bölümlere ve öğretim üyelerine göre başarı oranlarının hesaplanarak bir veri madenciliği modeli oluşturulması amaçlanmaktadır. Ancak burada amaç bu modelin kullanılarak tüm yeni öğrencilerin başarı notlarının tahmin edilmesi değildir ve nitelikli bir öğretim sisteminde verilen bilgilerle bunun yapılamayacağı da açıktır. Erişilmek istenen amaç yüksek güvenilirlikte böyle bir tahmin yapılabilecek sıra dışı durumların varlığını araştırmaktır.

Bu doğrultuda İTÜ Öğrenci İşleri Daire Başkanlığı' ndan MAT101(E), MAT102, MAT103(E), MAT104(E), MAT201(E), MAT202 ve MAT261 derslerinin 1999-2008 yılları arasındaki yaklaşık 80000 kaydı (Numara, Bölüm, Öğretim Görevlisi, Yıl, Dönem, Not) formatında alınmıştır. Alınan bu kayıtlardan sözü edilen dersler için gerekli tablolar oluşturularak Oracle veritabanına (11g) aktarılmıştır. Daha sonra her ders için ODM yardımıyla Naive Bayes algoritması kullanılarak bir sınıflandırma modeli oluşturulmuştur. Elde edilen bu modeller tablolara uygulanarak hedeflenen sıra dışı durumların varlığı gözlenmiştir.

Bu bitirme ödevinde 2. bölümde genel olarak ORACLE veritabanı ve temel kavramlarından bahsedilmiştir. Uygulamada kullanılan "Oracle Data Miner" 3. bölümde, sınıflandırma algoritmalarından biri olan Naive Bayes yöntemi matematiksel altyapısıyla birlikte 4. bölümde anlatılmıştır. Veri madenciliği uygulamasının ayrıntıları ve elde edilen sonuçlar ise 5. bölümde tartışılmıştır.

2. ORACLE VERİTABANI YÖNETİM SİSTEMİ

Bu bölümde Oracle veritabanı yönetim sistemi ile ilgili temel kavramlar ve veritabanına erişim, kullanıcıların oluşturulması, gerekli hakların verilmesi gibi temel işlemlere kısaca değinilerek, sistem hakkında genel bir bilgi edinilmesi amaçlanmıştır.

2.1. Tanım

Oracle gelişmiş bir ilişkisel veritabanı yönetim sistemidir ve diğer tüm ilişkisel veritabanı sistemleri gibi büyük miktarda veriyi çok kullanıcılı ortamda depolama ve güvenli bir şekilde erişim işlemlerini yönetmektedir.

Oracle veritabanı yazılımları özellikle kurumsal alanda kullanılan yaygın bir veritabanı sistemidir. Oracle çok sayıda araçtan oluşur ve uygulama geliştiricilerinin kolay ve esnek uygulamalar geliştirmesini sağlar. Diğer veritabanı yönetim sistemlerinde olduğu gibi, saklı yordam (stored procedure), paketler, tetikleyici (trigger) gibi bileşenler yer alır.

2.2. Tarihçe

Larry Ellison ve arkadaşları 1979 yılında “Software Development Laboratories" şirketini kurmuşlardır. Aynı yıl içinde şirketin adını "Relational Software Inc." (RSI) olarak değiştirmiş ve Oracle 2 sürümü adı altında ilk ilişkisel veritabanı modellerini piyasaya sürmüşlerdir. Bu versiyon Oracle veritabanının atası olmuştur. Bu sürümde basit sorgu işlemleri ve eklenebilirlik özellikleri bulunmaktaydı. 1982 yılında RSI adı değiştirilerek "Oracle Corporation" olarak isimlendirilmiştir. 1983'te Oracle sürüm 3, 1984'te Oracle sürüm 4 piyasaya sürülmüştür. 1985'te istemci-sunucu modeli ve dağıtık uygulamalar eklenmiştir. 1988'de Oracle 6, 1989' da ERP (Kurumsal Kaynak Planlaması) paketi piyasaya sunulmuştur. 1997 yılına kadar Oracle yeni sürümleri çıkarılarak nesne yönelimli bir yapıya kavuşturulmuş ve 1999 yılında Oracle 8i sürümüne Java Virtual Machine paketi eklenmiştir. 2005 yılına

gelene kadar Oracle 11i, 10g sürümleri çıkarılmış ve son olarak da 2007 yılında Oracle 11g sürümü piyasaya sunulmuştur [10].

2.3. Temel Terimler

Oracle veritabanı ile ilgili sık kullanılan bazı terimler aşağıda listelenmiştir [11]:

Anlık sorgu (Ad hoc Query): Basit (bir kereliğine yazılan) sorgulara verilen isimdir. Bu tür sorgulara örnek olarak belirli verilerin listelenmesi komutlarını verebiliriz:

SELECT * FROM MUSTERI (müşteri tablosundaki bütün kayıtları listele)

Blok : Oracle veritabanlarının depolanmasında kullanılan en küçük birime blok denir. Bir blok 2 KB-16KB boyutları arasında büyüklüğe sahiptir.

Ara Bellek (Buffer): Verileri depolamak için kullanılan bellek miktarıdır. Bir ara bellek kullanılmış (anlık) veriyi içerir. Birçok durumda, ara bellekler disk üzerindeki verilerin bellekteki kopyasıdır.

Ön Bellek (Cache): Verilere hızlı erişim için kullanılan ara bellek alanlarıdır. Mantık olarak son erişilen bilgilerin durduğu ve aynı bilginin bir kere daha istendiğinde ana belleğe gitmeden ön bellekten alıdığı bir erişim mekanizmasıdır.

Kontrol noktası (Checkpoint): Bellekteki verilerin disk dosyalarına yazılması işlemidir.

Veritabanı: İlişkili verilerin toplandığı veri kümesidir. Ana veri düzenleme sisteminde veritabanı temeldir.

Veri Sözlüğü: Tabloların oluşturduğu bir veri sözlüğüdür. Veritabanı hakkında bilgi bu sözlükte yer alır.

Veritabanı Yöneticisi (DBA): Veritabanı yönetiminden sorumlu olan kişidir. Sistem yöneticisi ya da veritabanı yöneticisidir.

Dinamik Performans Tabloları (Dynamic Performance Tables): Başlatılan kopyanın performansını saklamak için kullanılan dinamik tablolardır.

Fonksiyon: Belli bir işlemi yerine getirmek için kullanılan komut kümeleridir. Veritabanı programlamasında sunucu tarafında yazılan kodlar fonksiyon ve yordam olarak yazılır.

Yordam (Procedure): Belli bir işlemi yerine getirmek için kullanılan komut kümeleridir. Veritabanı programlamasında sunucu tarafında yazılan kodlar fonksiyon ve yordam olarak yazılır.

Sorgu: Bir veritabanı üzerinde çalıştırılan komut kümesidir. Örneğin SELECT deyimiyle başlayan komutlardır.

Şema: Veritabanı nesnelerinin şemasıdır.

İşlem Bilgisi/Hareketi (Transaction): Bir ya da birden çok SQL deyimi bir işlem bilgisi olarak tanımlanır. "Transaction" 'lar özel bir alanda depolanır ve verilerin bütünlüğünün sağlanması için kullanılır. Bir "transaction" içindeki işlemlerin tamamı (birkaç güncelleme komutu) yerine getirilir ve işlem onaylanır (commit). Aksi takdirde işlem geri çevrilir (roll back).

Tetikleyici (Trigger): Yordam ve fonksiyonların otomatik olarak başlatılmasını sağlayan mekanizma ya da yordamın otomatik olarak çalıştırılmasıdır. Tetikleyiciler tipik olarak tablo üzerinde INSERT, UPDATE ya da DELETE deyimi işletildiğinde başlar.

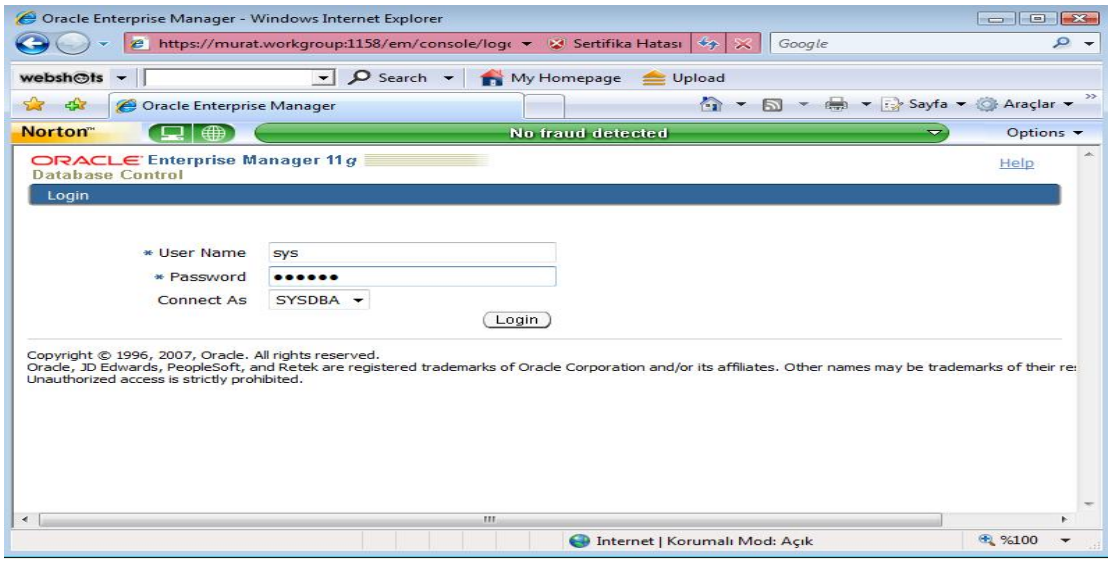
2.4. Veritabanı Yönetim Konsolu

Veritabanı yönetim konsolu kullanıcılar ve veritabanı yöneticilerinin veritabanına erişimlerini ve verilen haklar çerçevesinde istedikleri özellikleri kontrol etmelerini sağlayan bir arayüzdür. Aşağıda Oracle veritabanı yönetim konsolu ile ilgili bazı ekran görüntüleri ve açıklamaları verilmektedir.

2.4.1 Giriş Ekranı

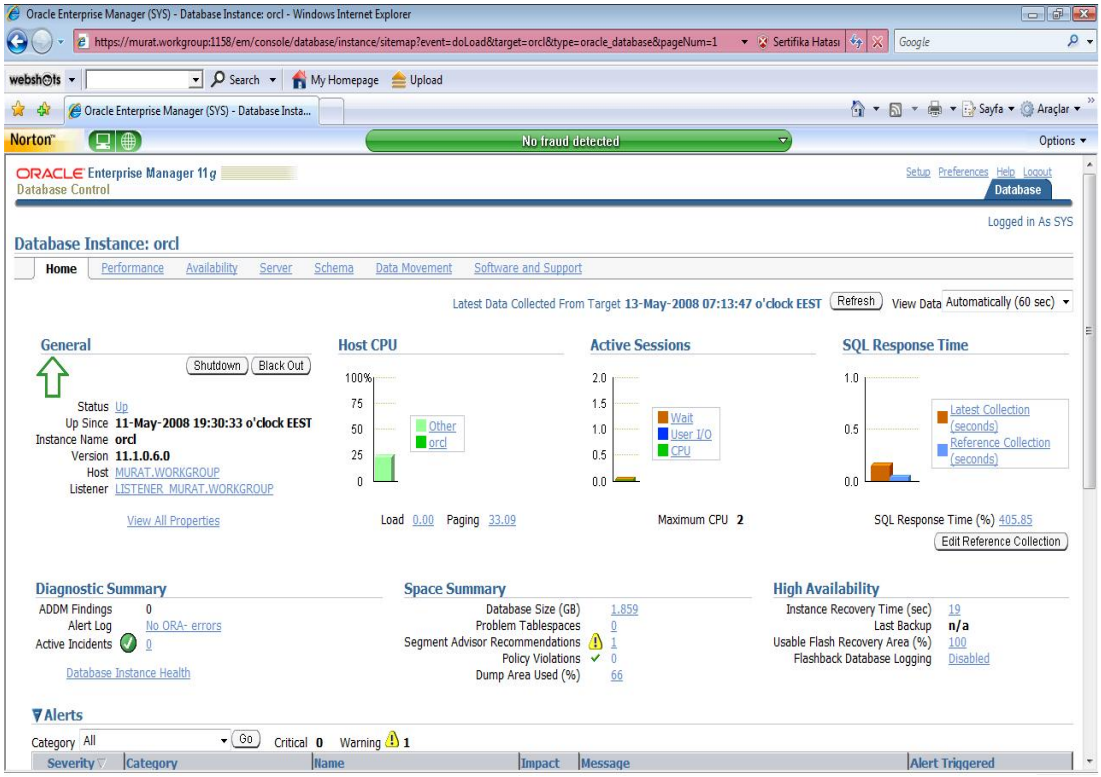
Aşağıda verilen ekrana ulaşmak için Oracle ile ilgili servisler otomatik başlat olarak ayarlanmamışlarsa öncelikle bu servisler başlatılmalı, sonrasında da internet tarayıcısına https://<localhost>:1158/em yazılarak çalıştırılmalıdır.

Oracle Database 11g programı yüklendikten sonra "SYS" kullanıcısının kilidi (varsayılan) açılmıştır ve bu kullanıcı "Süper Yönetici" olduğundan bütün işlemleri yapmaya yetkilidir. "SYS" kullanıcısı ile sisteme SYSDBA olarak girilir.



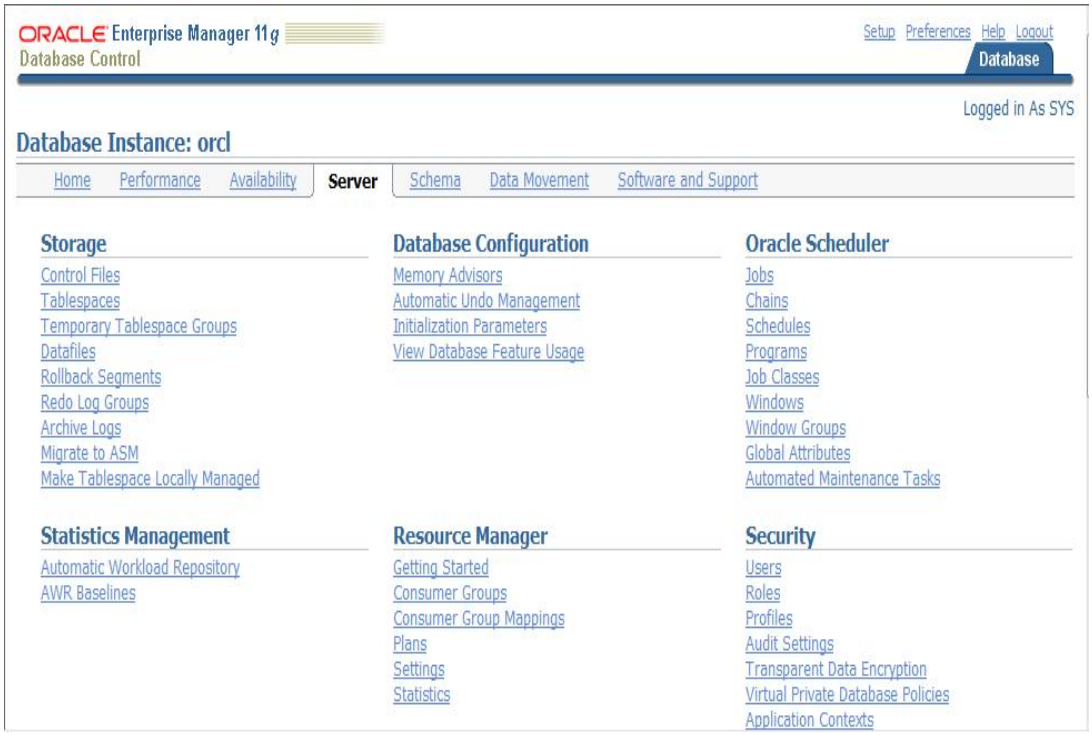
2.4.2 Veritabanı Ana Yönetim Ekranı

Sisteme "SYS" kullanıcısı olarak girildikten sonra aşağıdaki ekran karşımıza çıkar. Bu ekran aracılığıyla veritabanındaki bütün özellikler; Performans, Ulaşılabilirlik, Sunucu, Şema, Veri Hareketleri, Yazılım ve Destek gibi bileşenler kontrol edilebilir.

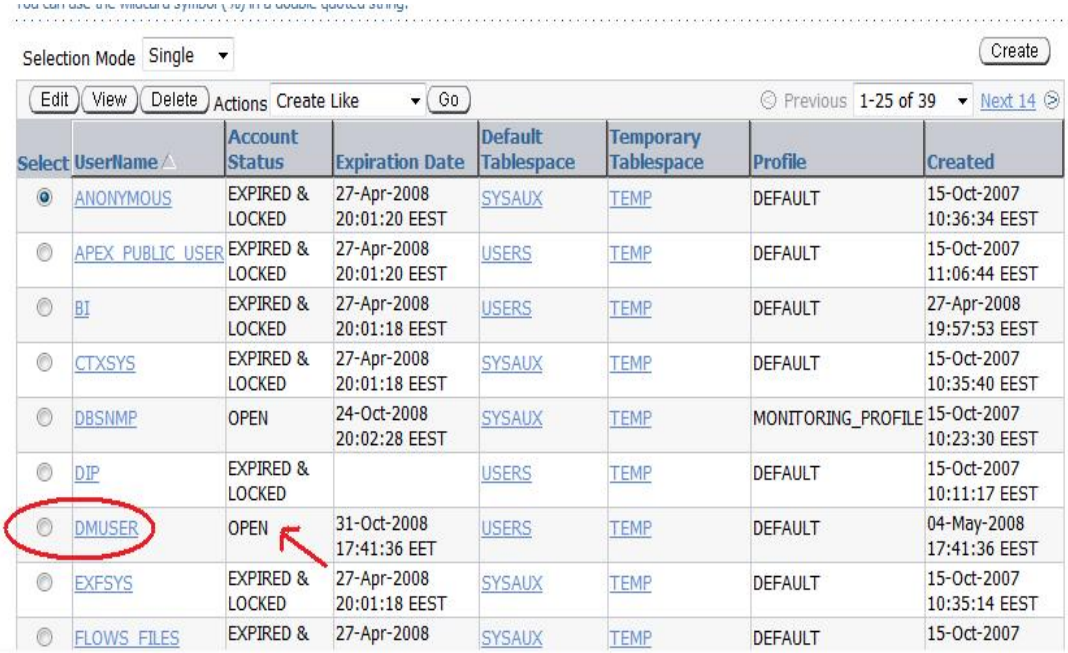


2.4.3. Sunucu Ekranı

Veritabanı ile ilgili tüm bilgilere erişip, değişiklikler yapabileceğimiz ekrandır.

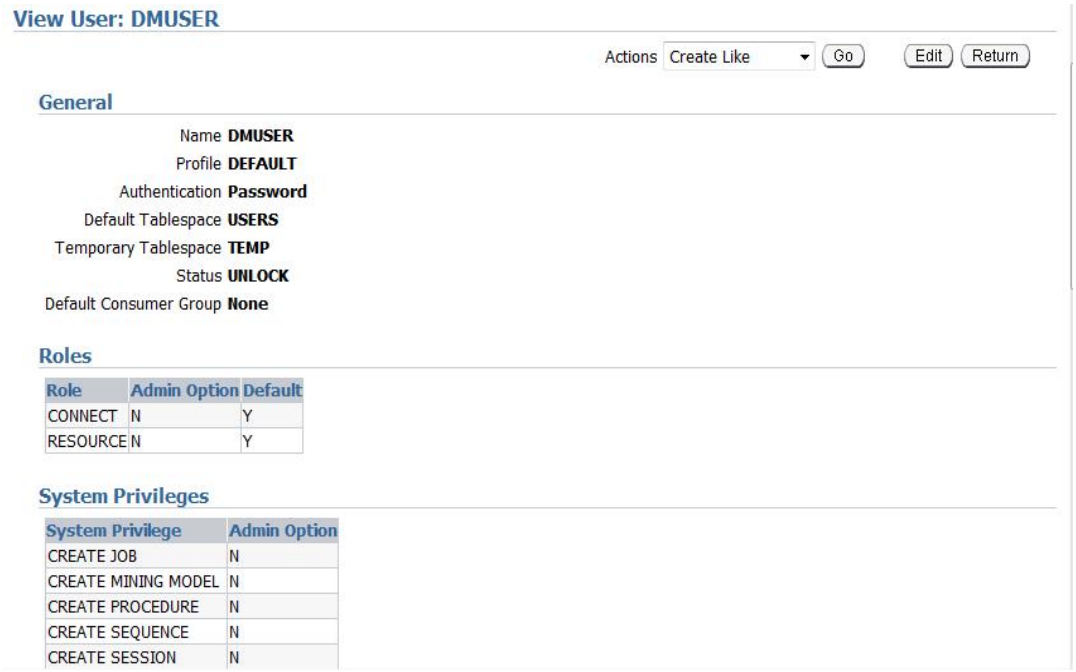


Buradan "Security->Users" yolundan hareket ederek veritabanında tanımlı tüm kullanıcılar aşağıdaki gibi görülebilir. Veri madenciliği uygulaması için yükleme sırasında DMUSER isimli yeni bir kullanıcı oluşturulmuş, kilidi kaldırılmış ve gerekli haklar verilmiştir.



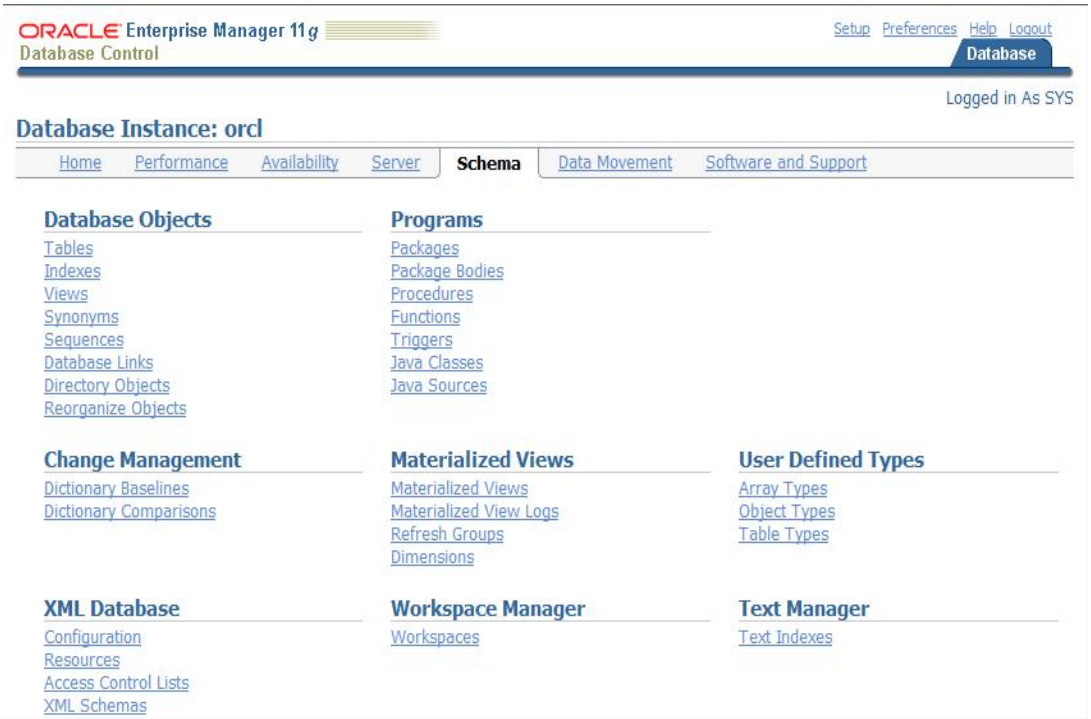
Select	UserName	Account Status	Expiration Date	Default Tablespace	Temporary Tablespace	Profile	Created
◉	ANONYMOUS	EXPIRED & LOCKED	27-Apr-2008 20:01:20 EEST	SYSAUX	TEMP	DEFAULT	15-Oct-2007 10:36:34 EEST
○	APEX_PUBLIC_USER	EXPIRED & LOCKED	27-Apr-2008 20:01:20 EEST	USERS	TEMP	DEFAULT	15-Oct-2007 11:06:44 EEST
○	BI	EXPIRED & LOCKED	27-Apr-2008 20:01:18 EEST	USERS	TEMP	DEFAULT	27-Apr-2008 19:57:53 EEST
○	CTXSYS	EXPIRED & LOCKED	27-Apr-2008 20:01:18 EEST	SYSAUX	TEMP	DEFAULT	15-Oct-2007 10:35:40 EEST
○	DBSNMP	OPEN	24-Oct-2008 20:02:28 EEST	SYSAUX	TEMP	MONITORING_PROFILE	15-Oct-2007 10:23:30 EEST
○	DIP	EXPIRED & LOCKED		USERS	TEMP	DEFAULT	15-Oct-2007 10:11:17 EEST
○	DMUSER	OPEN	31-Oct-2008 17:41:36 EET	USERS	TEMP	DEFAULT	04-May-2008 17:41:36 EEST
○	EXFSYS	EXPIRED & LOCKED	27-Apr-2008 20:01:18 EEST	SYSAUX	TEMP	DEFAULT	15-Oct-2007 10:35:14 EEST
○	FLOWS_FILES	EXPIRED &	27-Apr-2008	SYSAUX	TEMP	DEFAULT	15-Oct-2007

DMUSER kullanıcısı seçilirse kullanıcı özelliklerini belirten bir sayfa açılır.



Burada kullanıcıya ait genel özellikler, roller, kullanıcı sistem hakları, nesneler üzerindeki hakları, kontenjan bilgisi, Proxy kullanıcısı, tüketici grup hakları gibi kısımlar bulunur.

2.4.4. Şema Ekranı



Bu kısımda veritabanı nesneleri (tablolar, indeksler, görünümler... gibi) , programlar, kullanıcı tanımlı tipler, XML veritabanı, Text Manager bulunmaktadır. “Database Objects->Tables” yolunda şema kısmına DMUSER yazılarak bu kullanıcıya ait tablolar görüntülenebilir.

	DMUSER	DR$TEST_TEXT_IDX$N	USERS	NO		
	DMUSER	DR$TEST_TEXT_IDX$R	USERS	NO		
	DMUSER	EUL_GW_COLS	USERS	NO	0	05-May-2008 22:01:05 EEST
	DMUSER	EUL_GW_FILTERS	USERS	NO	0	05-May-2008 22:01:06 EEST
	DMUSER	EUL_GW_FKS	USERS	NO	0	05-May-2008 22:01:06 EEST
	DMUSER	EUL_GW_FK_COLS	USERS	NO	0	05-May-2008 22:01:06 EEST
	DMUSER	EUL_GW_OBJS	USERS	NO	0	05-May-2008 22:01:04 EEST
	DMUSER	EUL_GW_OBJ_FK_USGS	USERS	NO	0	05-May-2008 22:01:06 EEST
	DMUSER	EUL_GW_SCHEMAS	USERS	NO	1	05-May-2008 22:01:04 EEST
	DMUSER	EUL_GW_UKS	USERS	NO	0	05-May-2008 22:01:06 EEST
	DMUSER	EUL_GW_UK_COLS	USERS	NO	0	05-May-2008 22:01:06 EEST
	DMUSER	MAT101	USERS	NO	12344	05-May-2008 22:01:19 EEST
	DMUSER	MAT101B610749152	USERS	NO	6138	12-May-2008 22:00:48 EEST
	DMUSER	MAT101T327288048	USERS	NO	6206	12-May-2008 22:00:49 EEST
	DMUSER	MAT102E	USERS	NO	6288	05-May-2008 22:01:11 EEST
	DMUSER	MAT103	USERS	NO	13692	05-May-2008 22:01:21 EEST
	DMUSER	MAT104	USERS	NO	9490	05-May-2008 22:01:17 EEST
	DMUSER	MAT104103	USERS	NO	6988	05-May-2008 22:01:35 EEST
	DMUSER	MAT104103TEKLI	USERS	NO	6573	09-May-2008 22:07:32 EEST
	DMUSER	MAT201	USERS	NO	17094	12-May-2008 22:00:50 EEST
	DMUSER	MAT202	USERS	NO	5367	05-May-2008 22:01:09 EEST
	DMUSER	MAT261	USERS	NO	13602	05-May-2008 22:01:22 EEST

Edit View Delete With Options Actions Create Like Go Previous 25 76-100 of 123 Next 23

3. ORACLE DATA MINER

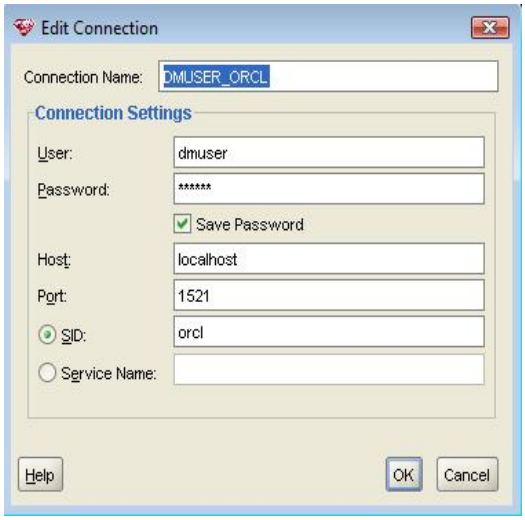
Bu bölümde Oracle Data Miner paketi tanıtılacaktır. Bu nedenle öncelikle "Oracle Data Mining" ve "Oracle Data Miner" (ODM) kavramları açıklanmalıdır.

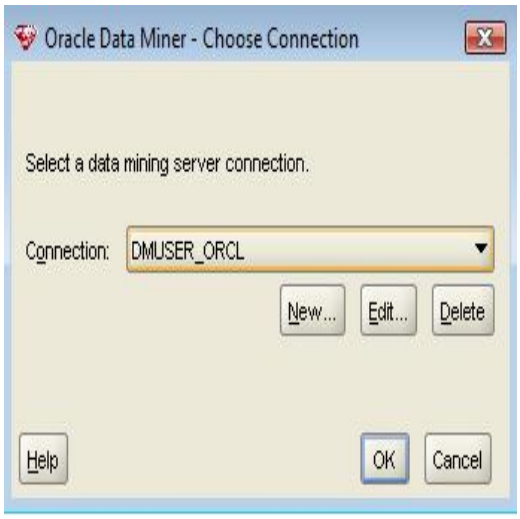
"Oracle Data Mining" kavramı veriler arasından bazı metodolojiler kullanılarak yararlı bilgi elde etmeye yarayan, Oracle veritabanına gömülü olan bir süreci ifade eder. ODM ise Oracle veritabanı ile birlikte çalışan ve "Oracle Data Mining" işlemini kontrol etmek için kullanılan bir araçtır. Başka bir ifadeyle, ODM veritabanı içinde gömülü olan bu işleme bağlanmak için istemciye kurulması gereken bir pakettir. ODM paket programı Oracle veritabanı ile bütünleşik olmadığından ayrı olarak kurulması gerekir.

Oracle veritabanının 10g öncesi sürümleri Darwin Veri madenciliği arayüzünü desteklemektedir. "Oracle Data Mining" işlemi ise Oracle10g ve Oracle 11g sürümleri tarafından desteklenmektedir. Bu nedenle sunucu üzerine Oracle11g veritabanı, istemciye ise ODM paketi yüklenerek veritabanına erişim sağlanmıştır.

3.1. Erişim

ODM paketinin veritabanına bağlantı ekranları aşağıda verilmiştir. Ancak ODM ilk kez çalıştırıldığında bazı erişim bilgilerinin girilmesi gerekmektedir.

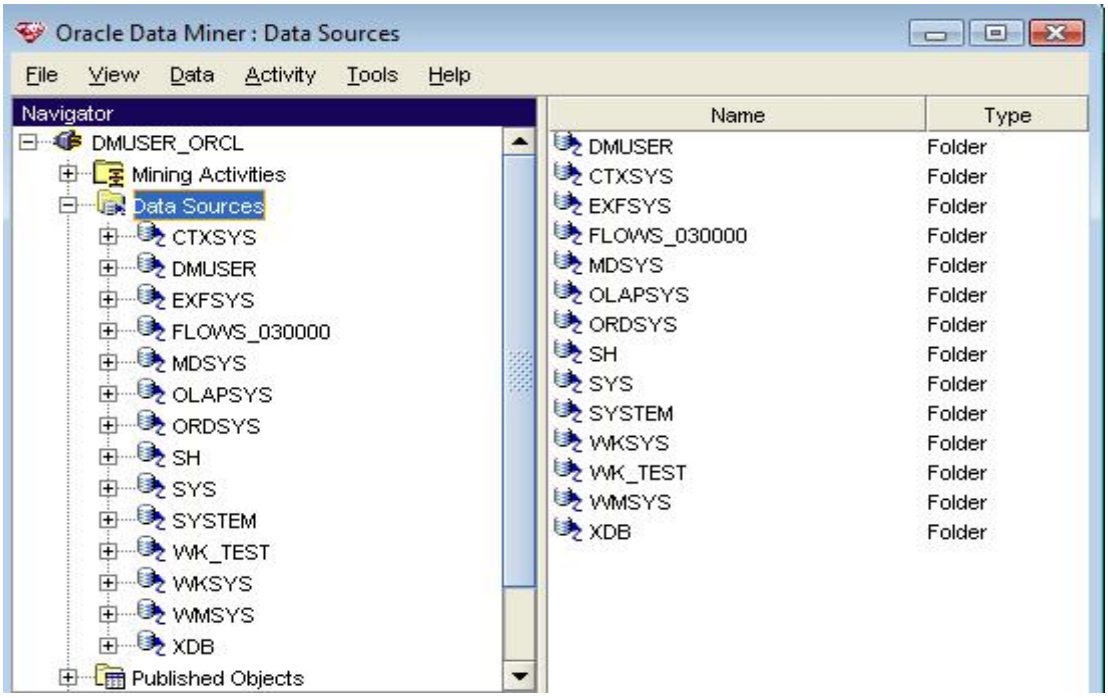


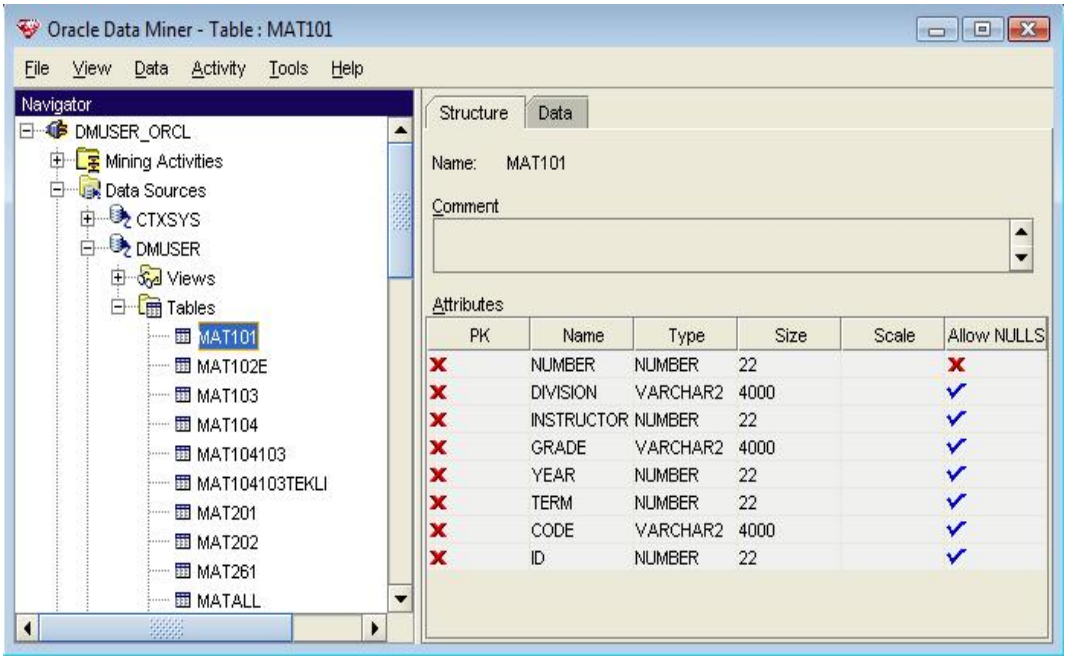


3.2. ODM Ana Ekranı

Bağlantı sağlandığında ODM programının genel görünüşü aşağıdaki gibidir. Sol tarafta yer alan "Navigator" kısmı ağaç yapısında olup, modeller, tablolar, testler, sonuçlar ve görevlere kolay bir şekilde erişimi sağlar.

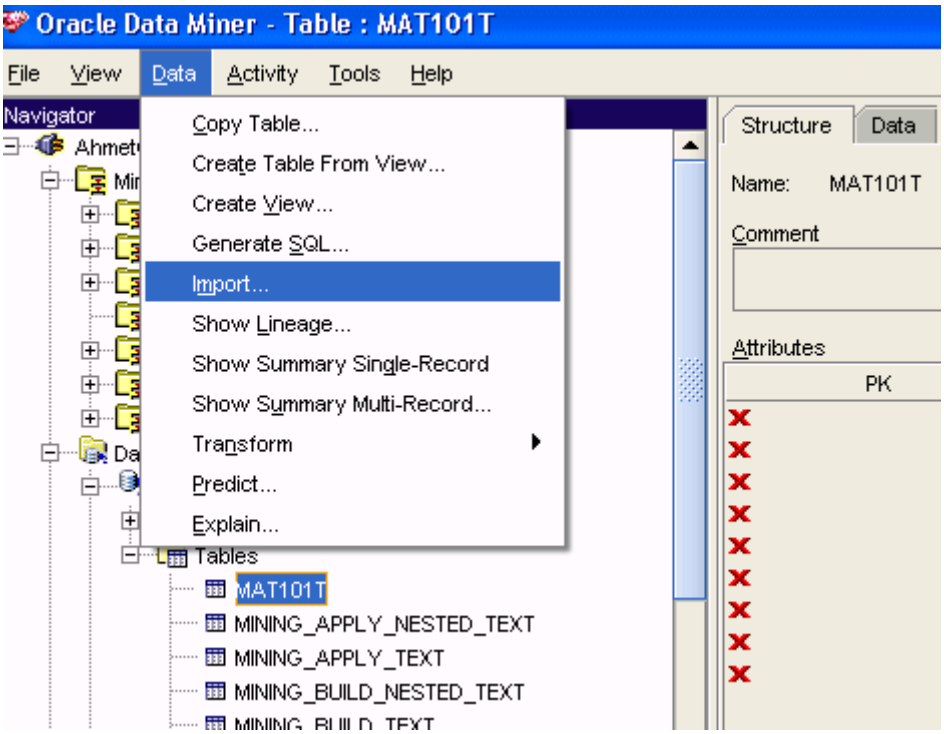
"Data Sources" katmanı, veritabanında bulunan kullanıcılar ve bu kullanıcılara ait tablo ve görünümleri içermektedir. Burada önemli bir kısımdan bahsetmek gerekmektedir. "Oracle Data Mining" işlemine erişim hakkı sadece kurulum sırasında tanımlanan "DMUSER" ve "SH" kullanıcılarına verilmiştir. Dolayısıyla veri madenciliği yapılacak tablolar bu kullanıcılardan biri ile bağlanılarak veritabanına aktarılmalıdır.

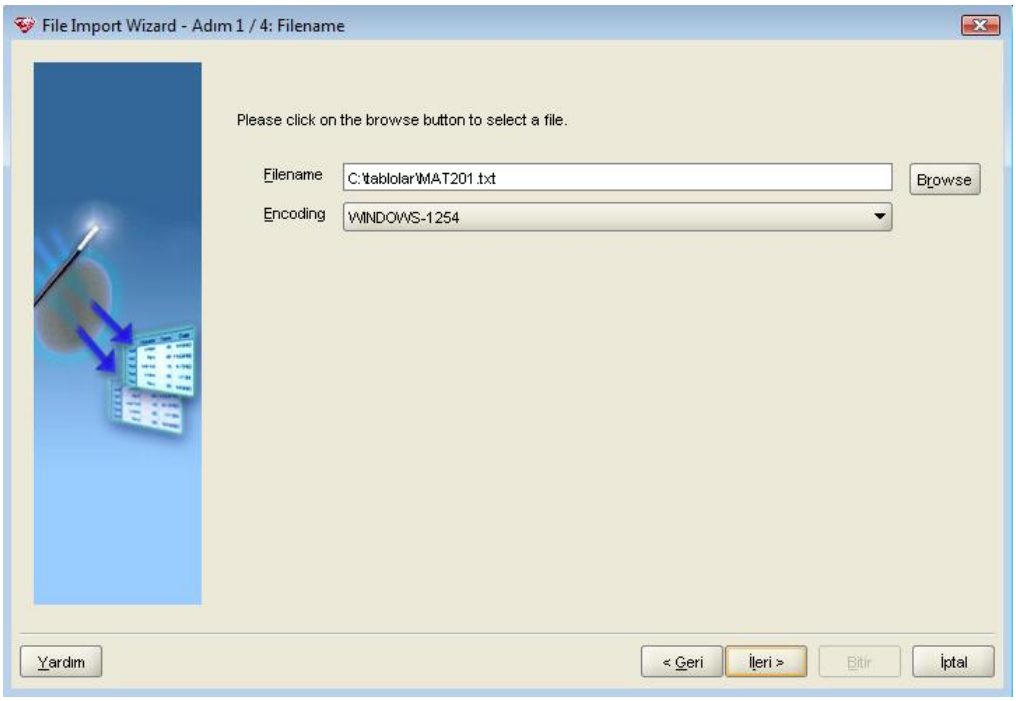


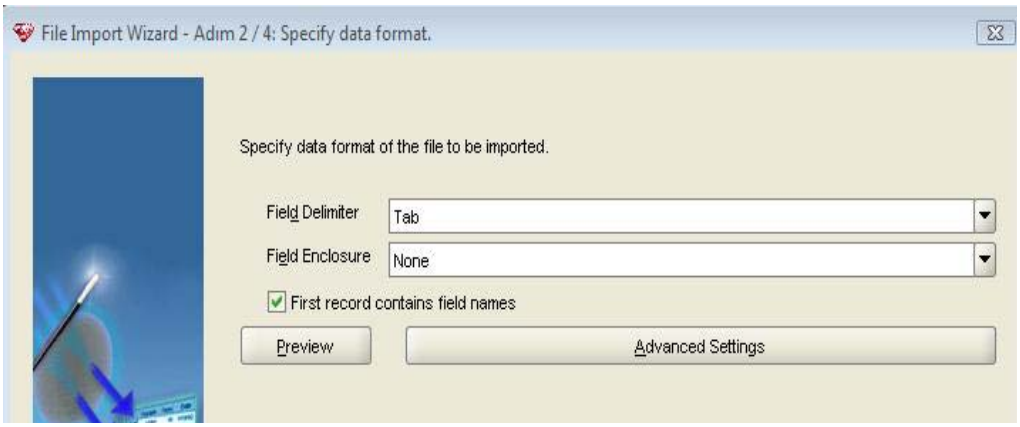


3.3. Veri Aktarımı

ODM nin "Data->Import" yolu ile veritabanına dışarıdan veri aktarımı yapılabilir.

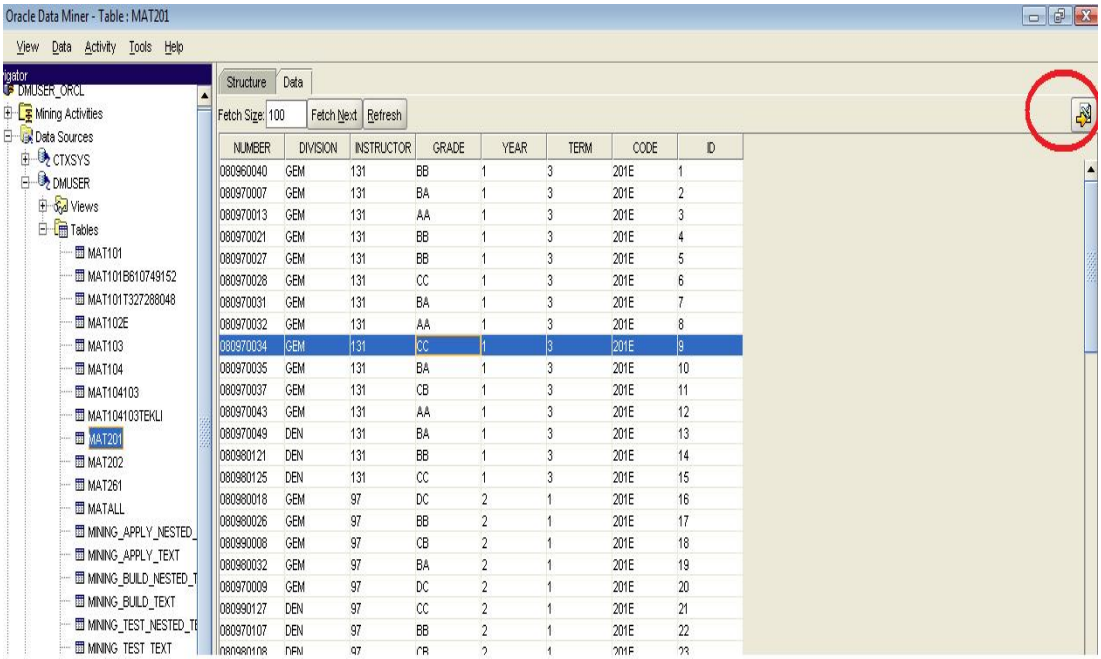



File Import Wizard - Adım 1 / 4: Filename
Please click on the browse button to select a file.
Filename
C:\tablolar\MAT201.txt
Browse
Encoding
WINDOWS-1254
Yardım
< Geri
İleri >
Bitir
İptal


File Import Wizard - Adım 2 / 4: Specify data format.
Specify data format of the file to be imported.
Field Delimiter
Tab
Field Enclosure
None
First record contains field names
Preview
Advanced Settings

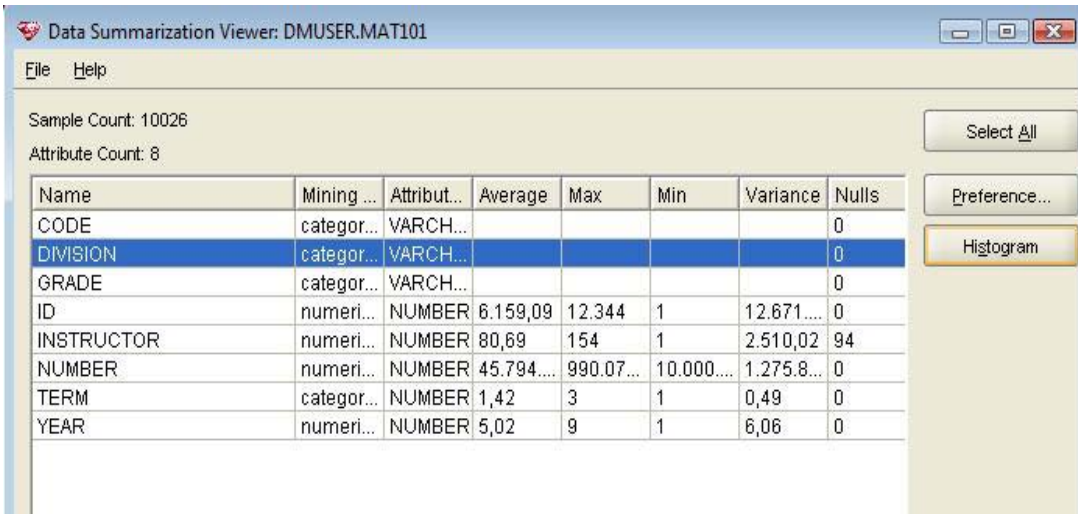

File Import Wizard - Adım 3 / 4: Target field definitions.
Specify field definitions corresponding to fields in imported record. Please press enter key after you have modified any cell to ensure the new value has been recorded.
Column Name
Data Type
Data Size/Fo...
Null If
"NUMBER"
VARCHAR2(SIZE)
4000
"DIVISION"
VARCHAR2(SIZE)
4000
"INSTRUCTOR"
NUMBER
N/A
"GRADE"
VARCHAR2(SIZE)
4000
"YEAR"
NUMBER
N/A
"TERM"
NUMBER
N/A
"CODE"
VARCHAR2(SIZE)
4000
"ID"
NUMBER
N/A
VARCHAR
NUMBER
Preview
Clean Up Names

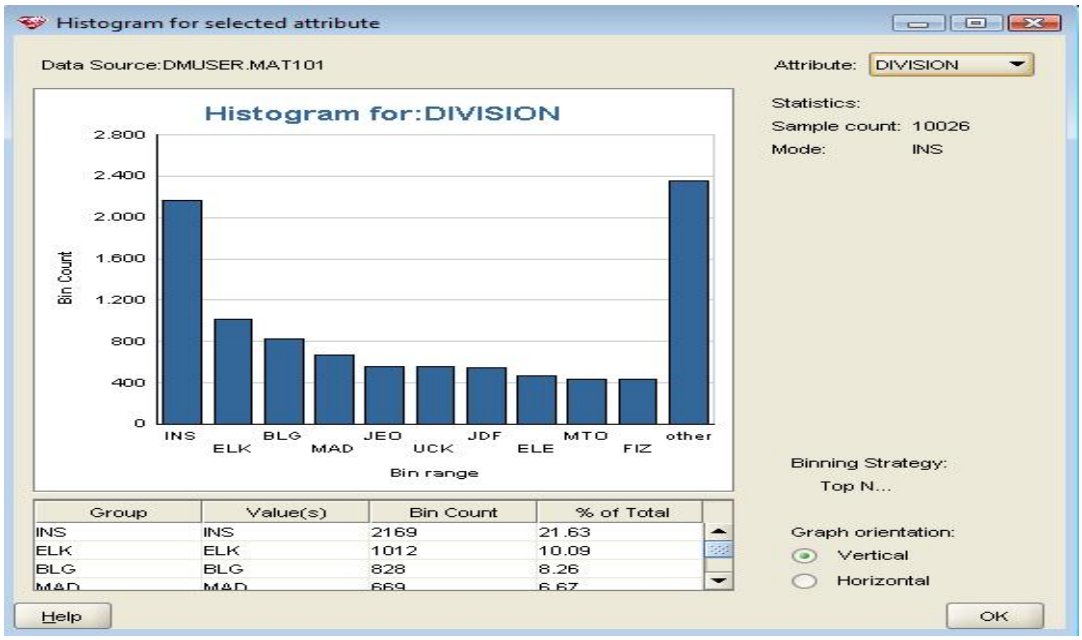
Veritabanında herhangi bir tablo görüntülenip, sağ üst köşede bulunan “export” tuşuna basılarak ilgili tablo dışarı aktarılabilir.



3.4. İstatistik İşlemleri

Herhangi bir tablo ile ilgili istatistiksel verileri görmek için tablonun format özelliklerinden birini seçmek gerekir. Böylece herhangi bir sütunun maksimum, minimum değerleri, ortalaması, standart sapması, varyansı ve dağıtılmış değerlerin histogramla gösterilmesi sağlanabilir. Aşağıda “Data->Show Summary Single Record” için bir örnek verilmiştir.



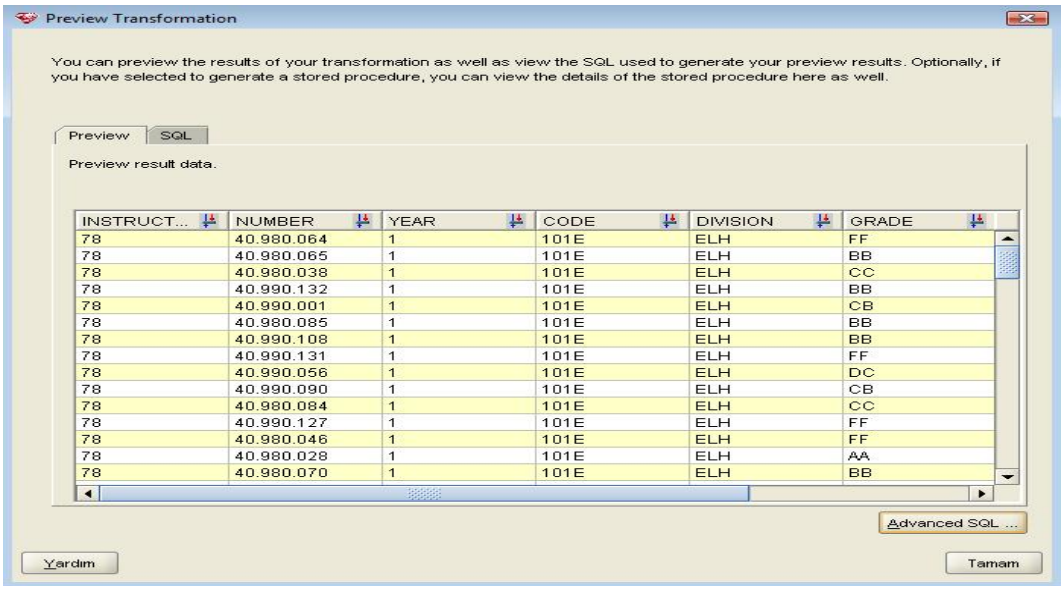


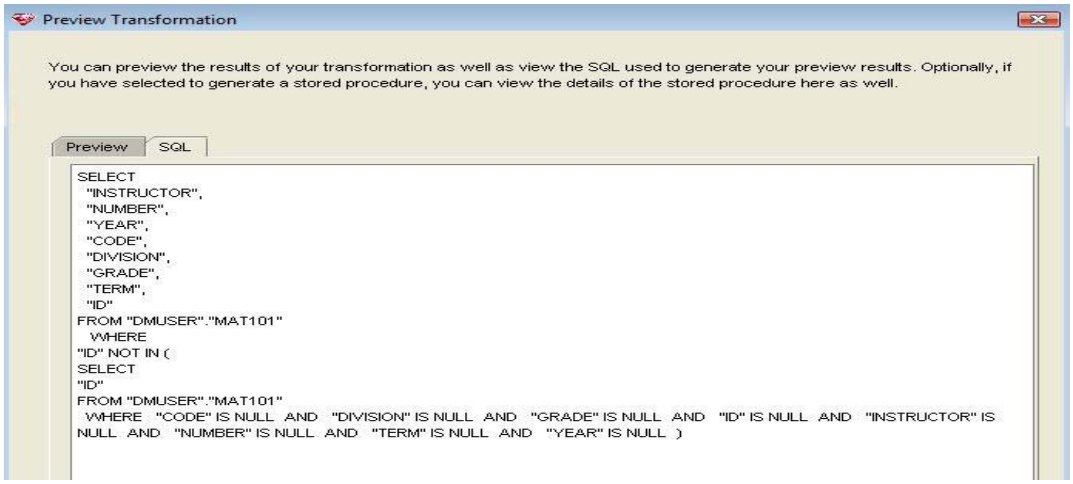
3.5. Veri Temizleme ve Hazırlama İşlemleri

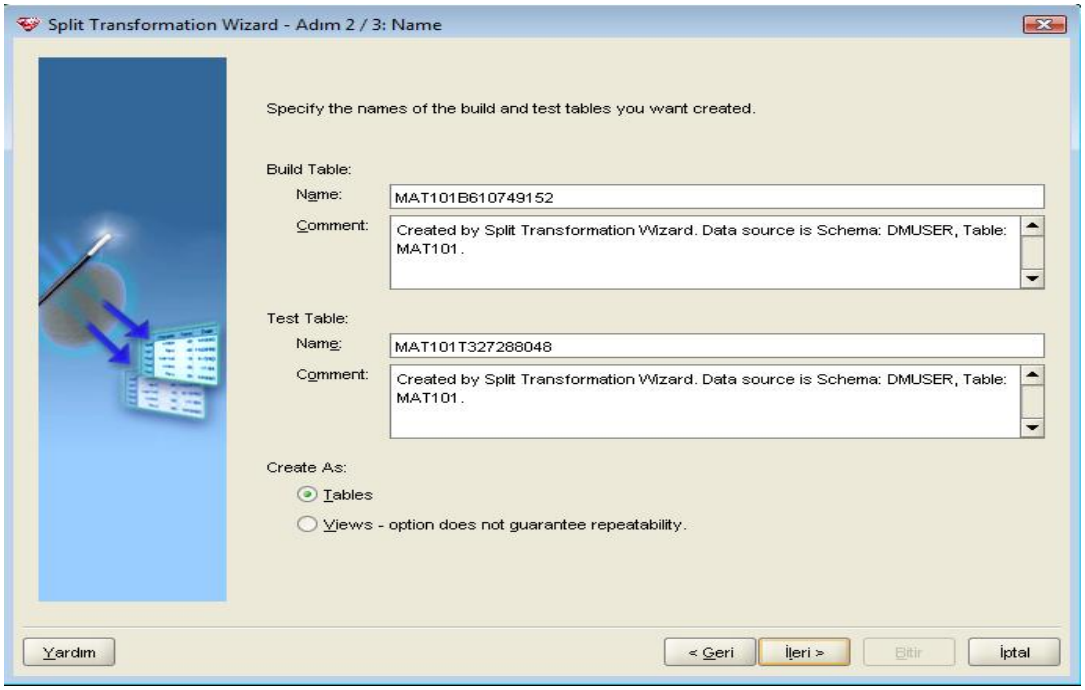
Bu bölümde hatalı, çelişkili ve ekstrem kayıtlar veriden çıkarılır ve eksik veri alanları tespit edilir. Bu tür kayıtlar doğrudan silinerek veri temizlenebileceği gibi, bunlar yerine ortalama değerler gibi daha tutarlı kayıtlarında konulması da tercih edilebilir. Ayrıca bir özelliğin, çalışanların maaşları gibi çok sayıda farklı kayıt içermesi durumunda bu bilgiden daha iyi yararlanmak amacıyla yüksek, orta, düşük gibi gruplara ayrılarak yeni değerler verilebilir.

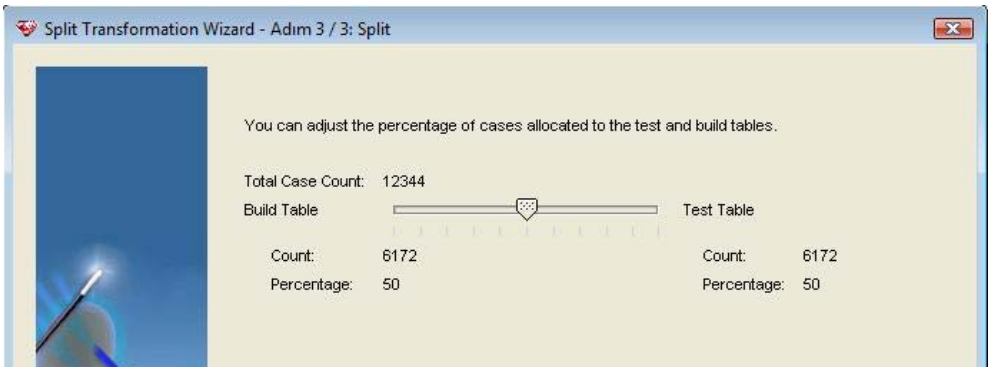
Bazı veri madenciliği algoritmaları özellikler nümerik olduğunda bu özelliklerin çok farklı değerler olmasına karşı duyarlıdır. Örneğin "yaş" genelde 100' den küçük iken, "gelir" milyon mertebesinde değerler alabilmektedir. Bu durumda belirli hesaplamalar için "gelir" sütununun "yaş" sütununa göre değer ölçüsünden dolayı çok önemli olduğu gibi bir yanılgı ortaya çıkmaktadır. Oysa hem "gelir" hem de "yaş" sütunundaki değerler normalizasyon ile 0-100 aralığına çekilerek yaş ile karşılaştırılabilir. ODM bir sütunun nümerik değerlerini belirli bir aralığa göreceli ağırlıkları ile orantılı olarak taşımak için bir normalizasyon fonksiyonu sağlar.

Ayrıca verinin model oluşturma ve oluşturulan modelin test edilmesi için bölünmesi gerekebilir. Buna benzer bazı veri hazırlama aşamalarının ekran çıktıları aşağıda verilmiştir.


Preview Transformation
You can preview the results of your transformation as well as view the SQL used to generate your preview results. Optionally, if you have selected to generate a stored procedure, you can view the details of the stored procedure here as well.
Preview
SQL
Preview result data.
INSTRUCT... NUMBER YEAR CODE DIVISION GRADE
78 40.980.064 1 101E ELH FF
78 40.980.065 1 101E ELH BB
78 40.980.038 1 101E ELH CC
78 40.990.132 1 101E ELH BB
78 40.990.001 1 101E ELH CB
78 40.980.085 1 101E ELH BB
78 40.990.108 1 101E ELH BB
78 40.990.131 1 101E ELH FF
78 40.990.056 1 101E ELH DC
78 40.990.090 1 101E ELH CB
78 40.980.084 1 101E ELH CC
78 40.990.127 1 101E ELH FF
78 40.980.046 1 101E ELH FF
78 40.980.028 1 101E ELH AA
78 40.980.070 1 101E ELH BB
Advanced SQL ...
Yardım
Tamam


Preview Transformation
You can preview the results of your transformation as well as view the SQL used to generate your preview results. Optionally, if you have selected to generate a stored procedure, you can view the details of the stored procedure here as well.
Preview
SQL
SELECT
"INSTRUCTOR",
"NUMBER",
"YEAR",
"CODE",
"DIVISION",
"GRADE",
"TERM",
"ID"
FROM "DMUSER"."MAT101"
WHERE
"ID" NOT IN (
SELECT
"ID"
FROM "DMUSER"."MAT101"
WHERE "CODE" IS NULL AND "DIVISION" IS NULL AND "GRADE" IS NULL AND "ID" IS NULL AND "INSTRUCTOR" IS NULL AND "NUMBER" IS NULL AND "TERM" IS NULL AND "YEAR" IS NULL)


Split Transformation Wizard - Adım 2 / 3: Name
Specify the names of the build and test tables you want created.
Build Table:
Name: MAT101B610749152
Comment: Created by Split Transformation Wizard. Data source is Schema: DMUSER, Table: MAT101.
Test Table:
Name: MAT101T327288048
Comment: Created by Split Transformation Wizard. Data source is Schema: DMUSER, Table: MAT101.
Create As:
Tables
Views - option does not guarantee repeatability.
Yardım
< Geri
İleri >
Bitir
İptal

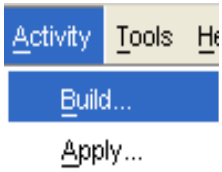


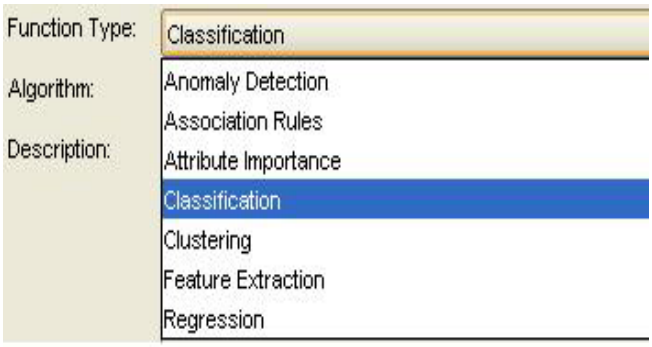
3.6. Veri Madenciliği Etkinlikleri

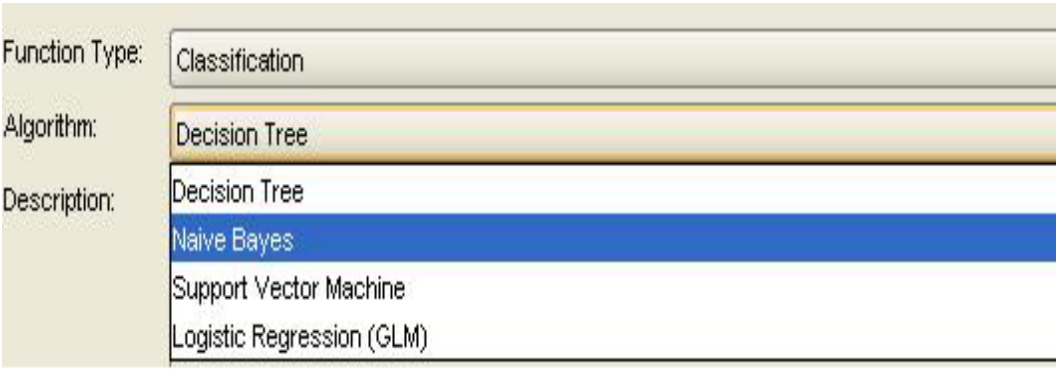
Bu bölümde model oluşturma, oluşturulan modeli test etme ve modelin yeni veriye uygulanarak sonuçların elde edilmesi aşamaları anlatılmaktadır.

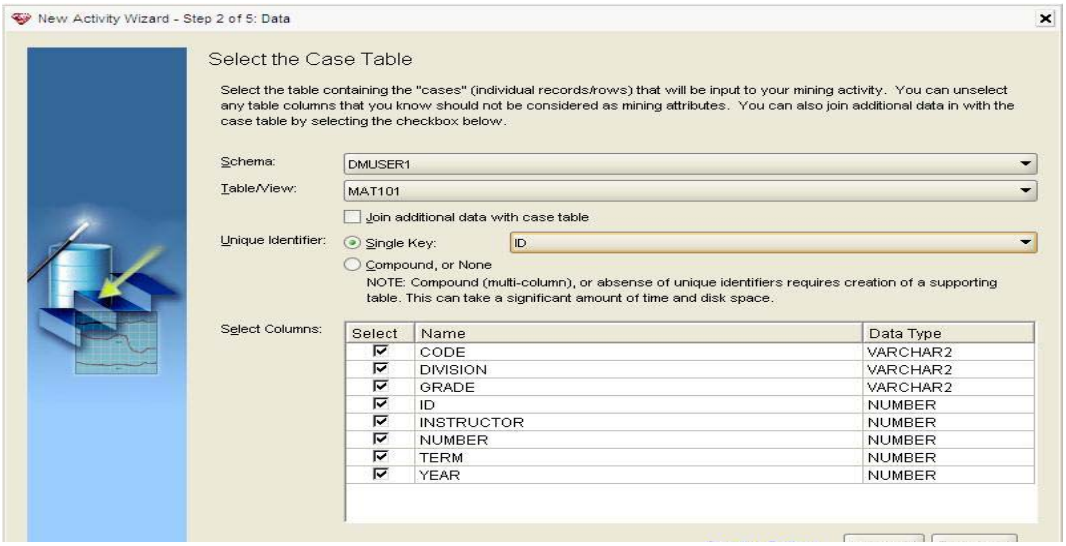
3.6.1. Model Oluşturma

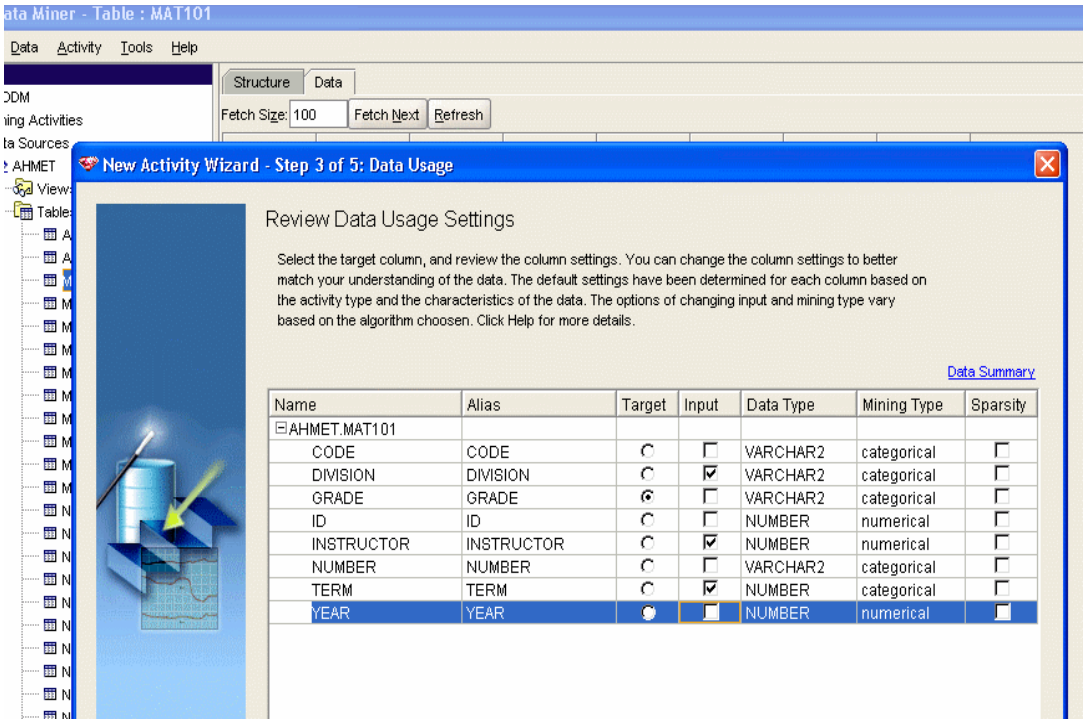
Burada uygulanacak veri madenciliği problemi, bunun çözümü için geçerli algoritmalar, uygulanacak veri kümesi ve ilgili diğer parametreler seçilerek model oluşturulur.



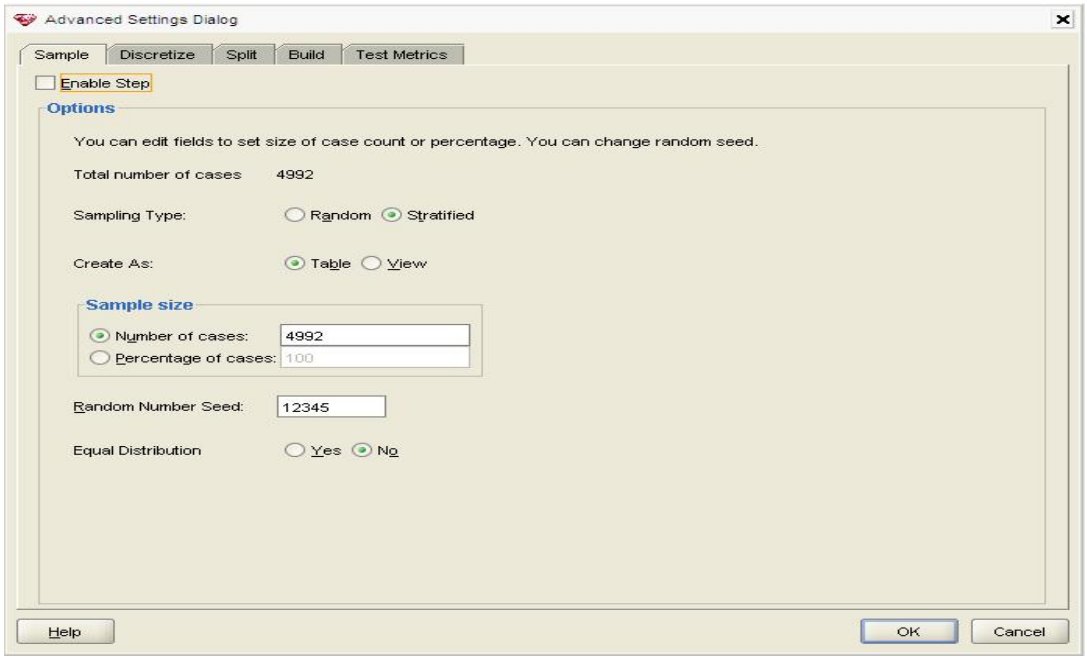


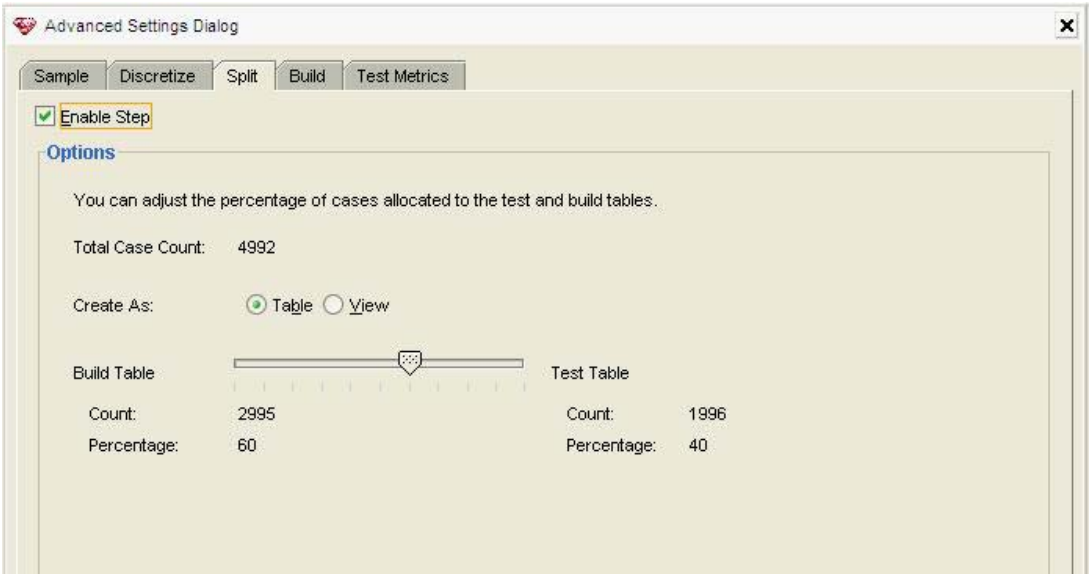


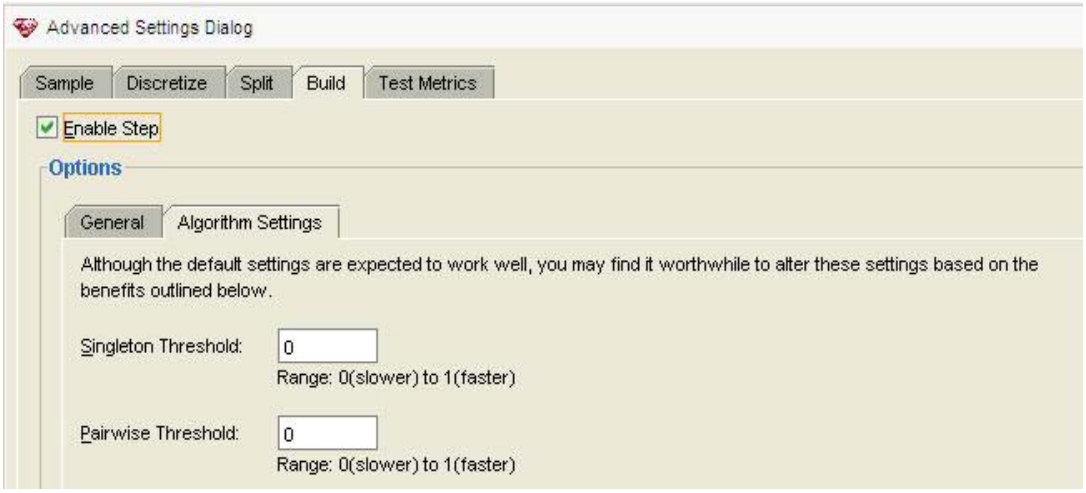


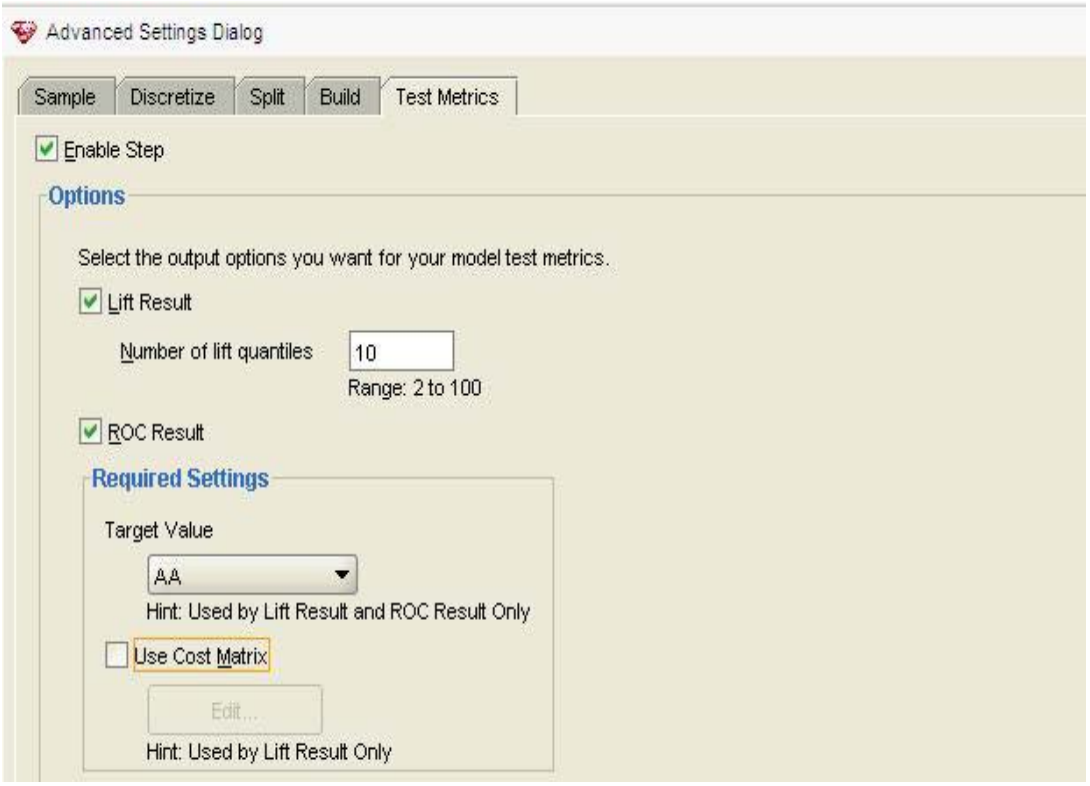


Yukarıda veirlen ekran model oluşturmanın en önemli adımıdır. Burada hangi sütunun tahmin edilmek istendiği ve bu tahminin diğer hangi sütunlarla yapılacağı seçilmektedir. Dolayısıyla burada tahmini etkilemeyecek olan örneğin "öğrenci numarası" gibi sütunlar "girdi" olarak seçilmemeli, sadece tahmini etkileyebilecek sütunların seçilmesine özen gösterilmelidir.

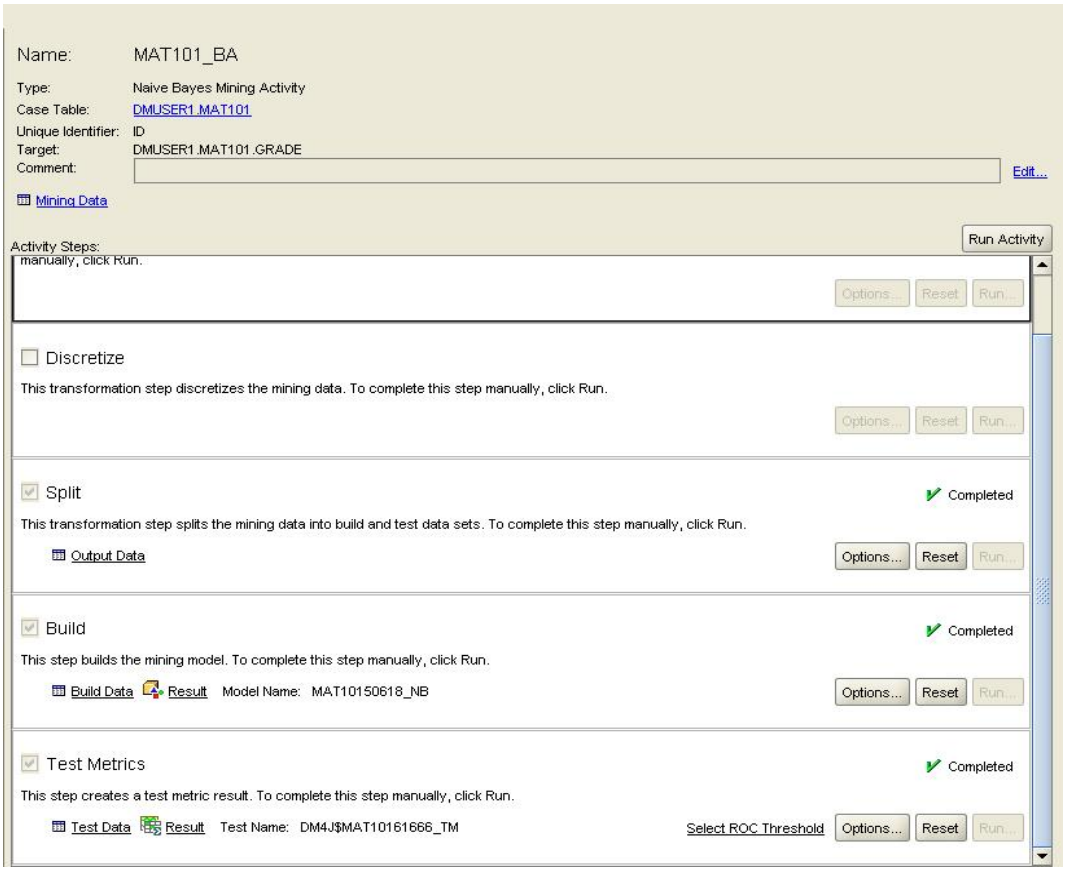

Advanced Settings Dialog
Sample
Discretize
Split
Build
Test Metrics
Enable Step
Options
You can edit fields to set size of case count or percentage. You can change random seed.
Total number of cases 4992
Sampling Type: Random Stratified
Create As: Table View
Sample size
Number of cases: 4992
Percentage of cases: 100
Random Number Seed: 12345
Equal Distribution Yes No
Help
OK
Cancel


Advanced Settings Dialog
Sample
Discretize
Split
Build
Test Metrics
Enable Step
Options
You can adjust the percentage of cases allocated to the test and build tables.
Total Case Count: 4992
Create As: Table View
Build Table
Test Table
Count: 2995
Percentage: 60
Count: 1996
Percentage: 40


Advanced Settings Dialog
Sample
Discretize
Split
Build
Test Metrics
Enable Step
Options
General
Algorithm Settings
Although the default settings are expected to work well, you may find it worthwhile to alter these settings based on the benefits outlined below.
Singleton Threshold: 0
Range: 0(slower) to 1(faster)
Pairwise Threshold: 0
Range: 0(slower) to 1(faster)

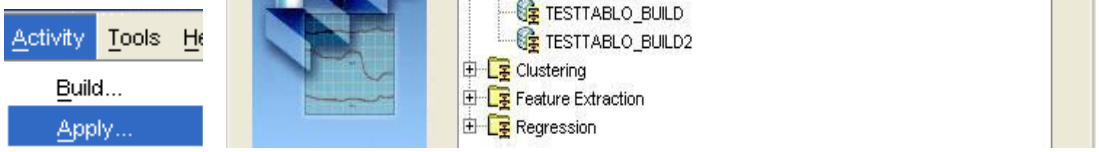


Yukarıdaki ekran çıktıları ile model oluşturma işlemi tamamlanır. Model oluşturmanın sonuç ekran çıktısı aşağıda verilmiştir.

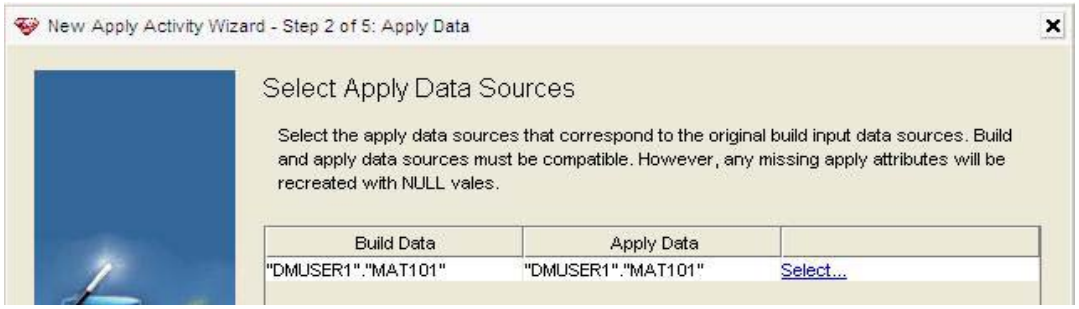


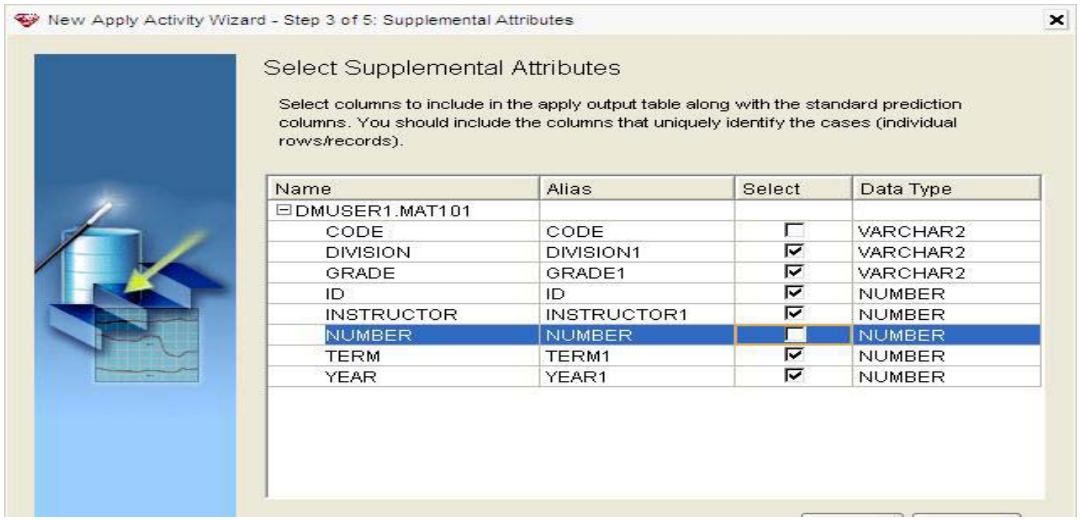
3.6.2. Modelin Uygulanması

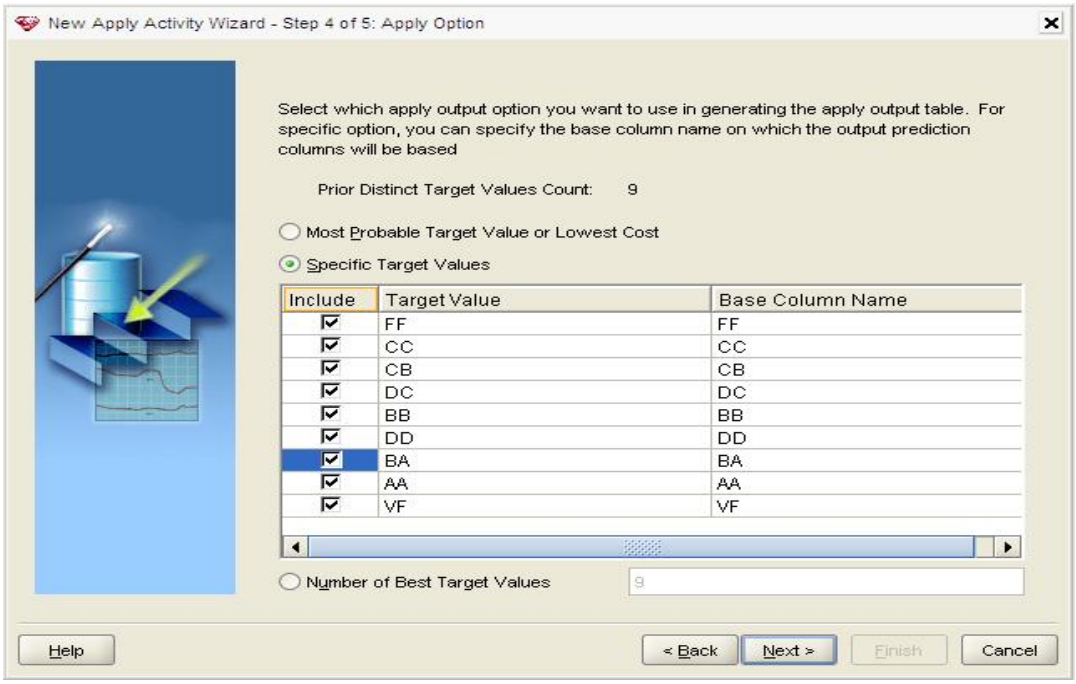
Bu bölümde veri madenciliği probleminin son adımı olarak bir önceki adımda elde edilen modelin yeni veri kümesine uygulanması adımları verilecektir.

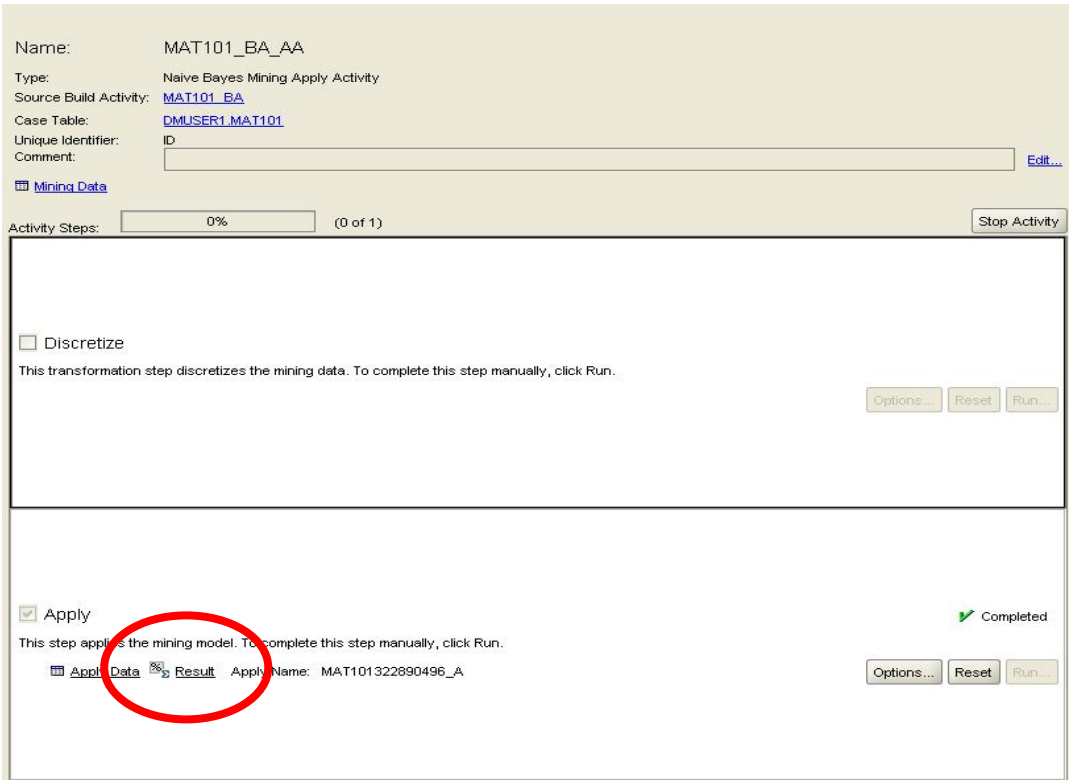


Yukarıdaki ekranda daha önce oluşturulan hangi modelin uygulamasının yapılacağı, aşağıdaki ekranda ise hangi veriye uygulanacağı seçilmektedir.

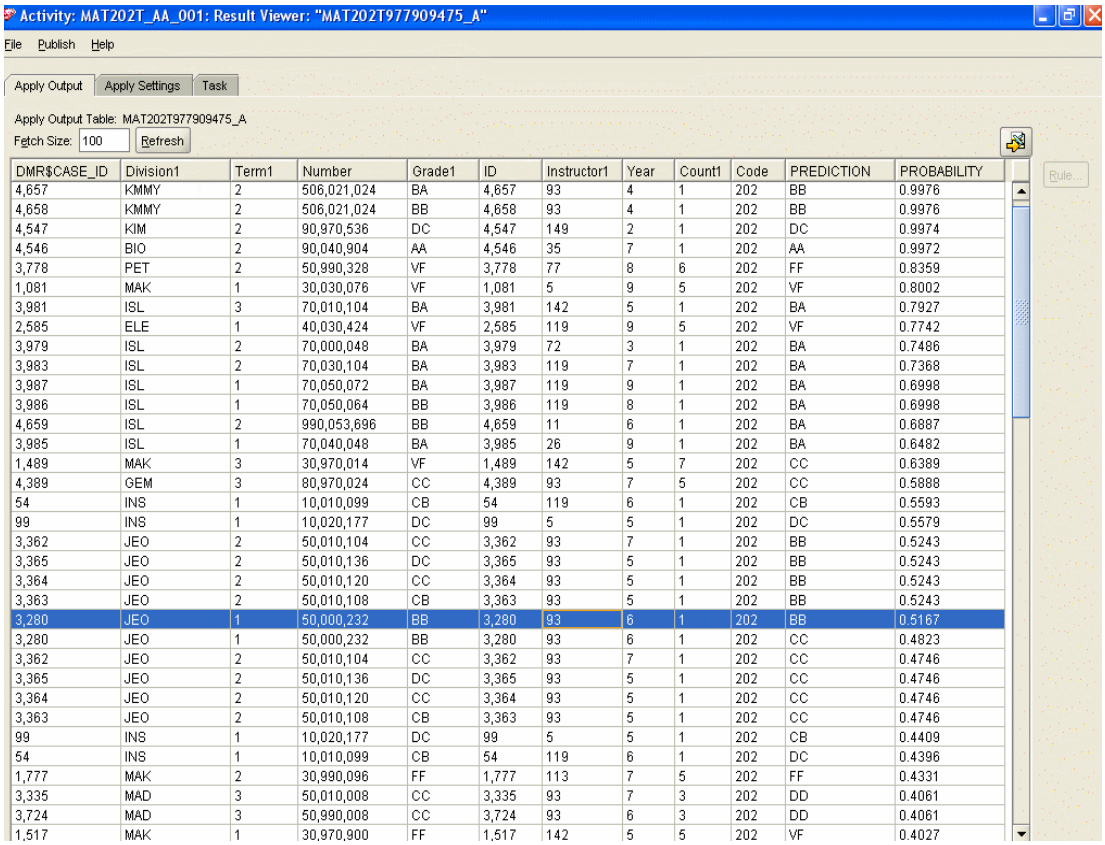








Uygulama ekranının son adımı yukarıda verilmiştir. Burada sonuç ekranına basılarak elde edilen sonuçlar görülebilir.



DMR$CASE_ID	Division1	Term1	Number	Grade1	ID	Instructor1	Year	Count1	Code	PREDICTION	PROBABILITY
4,657	KMMY	2	506,021,024	BA	4,657	93	4	1	202	BB	0.9976
4,658	KMMY	2	506,021,024	BB	4,658	93	4	1	202	BB	0.9976
4,547	KIM	2	90,970,536	DC	4,547	149	2	1	202	DC	0.9974
4,546	BIO	2	90,040,904	AA	4,546	35	7	1	202	AA	0.9972
3,778	PET	2	50,990,328	VF	3,778	77	8	6	202	FF	0.8359
1,081	MAK	1	30,030,076	VF	1,081	5	9	5	202	VF	0.8002
3,981	ISL	3	70,010,104	BA	3,981	142	5	1	202	BA	0.7927
2,585	ELE	1	40,030,424	VF	2,585	119	9	5	202	VF	0.7742
3,979	ISL	2	70,000,048	BA	3,979	72	3	1	202	BA	0.7486
3,983	ISL	2	70,030,104	BA	3,983	119	7	1	202	BA	0.7368
3,987	ISL	1	70,050,072	BA	3,987	119	9	1	202	BA	0.6998
3,986	ISL	1	70,050,064	BB	3,986	119	8	1	202	BA	0.6998
4,659	ISL	2	990,053,696	BB	4,659	11	6	1	202	BA	0.6887
3,985	ISL	1	70,040,048	BA	3,985	26	9	1	202	BA	0.6482
1,489	MAK	3	30,970,014	VF	1,489	142	5	7	202	CC	0.6389
4,389	GEM	3	80,970,024	CC	4,389	93	7	5	202	CC	0.5888
54	INS	1	10,010,099	CB	54	119	6	1	202	CB	0.5593
99	INS	1	10,020,177	DC	99	5	5	1	202	DC	0.5579
3,362	JEO	2	50,010,104	CC	3,362	93	7	1	202	BB	0.5243
3,365	JEO	2	50,010,136	DC	3,365	93	5	1	202	BB	0.5243
3,364	JEO	2	50,010,120	CC	3,364	93	5	1	202	BB	0.5243
3,363	JEO	2	50,010,108	CB	3,363	93	5	1	202	BB	0.5243
3,280	JEO	1	50,000,232	BB	3,280	93	6	1	202	BB	0.5167
3,280	JEO	1	50,000,232	BB	3,280	93	6	1	202	CC	0.4823
3,362	JEO	2	50,010,104	CC	3,362	93	7	1	202	CC	0.4746
3,365	JEO	2	50,010,136	DC	3,365	93	5	1	202	CC	0.4746
3,364	JEO	2	50,010,120	CC	3,364	93	5	1	202	CC	0.4746
3,363	JEO	2	50,010,108	CB	3,363	93	5	1	202	CC	0.4746
99	INS	1	10,020,177	DC	99	5	5	1	202	CB	0.4409
54	INS	1	10,010,099	CB	54	119	6	1	202	DC	0.4396
1,777	MAK	2	30,990,096	FF	1,777	113	7	5	202	FF	0.4331
3,335	MAD	3	50,010,008	CC	3,335	93	7	3	202	DD	0.4061
3,724	MAD	3	50,990,008	CC	3,724	93	6	3	202	DD	0.4061
1,517	MAK	1	30,970,900	FF	1,517	142	5	5	202	VF	0.4027

Sonuç ekranında görülmesi istenenerek seçilen sütunların yanısıra, tahmin (prediction) ve bu tahminin gerçekleşme olasılığı olmak özere iki ek sütun yer almaktadır. Burada daha önceden sonuçları varolan bir veri kümesi üzerinde uygulama gerçekleştirildiği için, hem tahmin hemde gerçekte öğrencinin hangi notu aldığı (Grade) görülebilmektedir. Bu bölümde amaç sadece ODM yi tanıtmak olduğu için, sonuçlar üzerinde durulmamış ve yeni sonuçların elde edilmesi ve bunların yorumlanması işlemleri 5. bölüme bırakılmıştır.

4. TEMEL KAVRAMLAR VE MATEMATİKSEL ALTYAPI

Bu bölümde uygulama olarak seçilen sınıflandırma modeli ve bu modeli çözmek için kullanılacak olan Naive Bayes yönteminin matematiksel altyapısından bahsedilmiş ve öncesinde gerekli temel kavramlar kısaca özetlenmiştir.

4.1. Temel Kavramlar

Kayıt *:* Bir tabloda her yeni girdi, her yeni satır bir kayıt olarak isimlendirilir.

Özellik *:* Bir tabloda her sütun özellik olarak isimlendirilir ve i. kaydın j. özelliği i.satır j. sütundaki verinin değeridir.

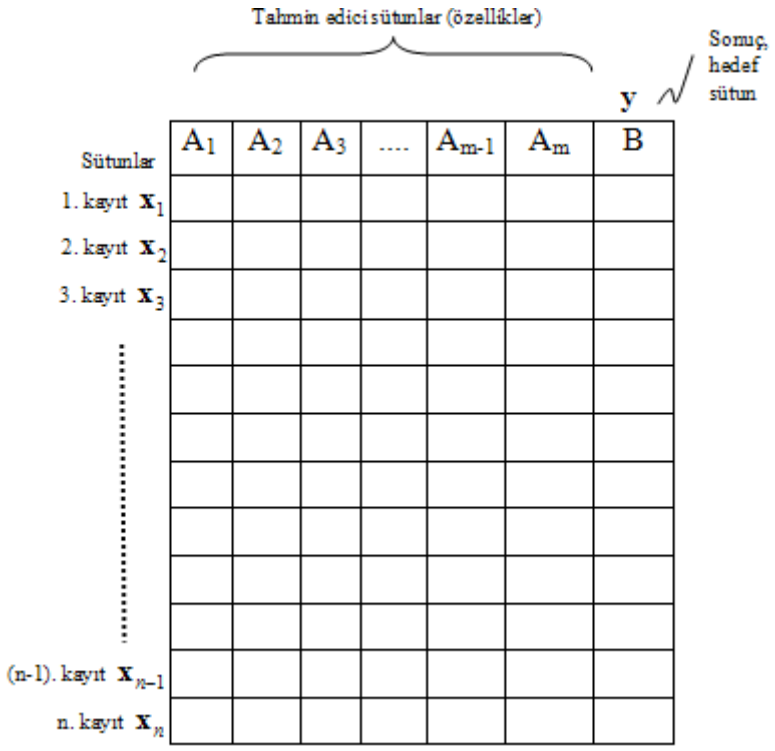
Olasılık : Bir olayın olabilirliğinin ölçüsüdür, [0-1] arasında değer alabilir, $P(A)$ ile gösterilir ve

$P(A)=1$, A olayının mutlaka gerçekleşeceğini

$P(A)=0$, A olayının gerçekleşmesinin mümkün olmadığını ifade eder.

Vektör *:* Burada kullanacağımız anlamıyla bir vektör $\mathbf{x}=\{x_1, x_2, x_3, ..., x_{m-1}, x_m\}$ şeklinde *m* elemanı ile belirlenen ve i. elemanı x_i ile verilen bir büyüklüktür.

Veri madenciliği uygulanacak olan veri kümesi üzerinde, aşağıdaki şekilde gösterilen tanımlamalar yapılacaktır; Tabloda her satır bir vektör ($\mathbf{x}_i$) olarak düşünülür, $\mathbf{x}_i$ vektörünün j. elemanı i. kaydın A_j sütunundaki değerine karşı gelir. Son sütun (*B*) yani $\mathbf{y}$ vektörü veri madenciliği ile tahmin edilmek istenen hedef özelliktir. Dolayısıyla *n* kayıt ve (*m+1*) sütundan oluşan bir tabloda her biri *m* boyutlu *n* tane belirleyici $\mathbf{x}_i$ vektörü ve bir tane hedef sütun (*B*) yani $\mathbf{y}$ vektörü vardır.



Şekil.4.1. Bir veri kaydı örneği

4.2. Naive Bayes Yöntemi

Sınıflandırma problemlerinin çözümünde kullanılan Naive Bayes yöntemi temel olarak olasılık teorisini kullanmaktadır. Bu bölümde önce yöntemin teorisi, sonrasında da yöntemle ilgili küçük örnekler verilmiştir.

4.2.1. Teori

Naive Bayes yöntemi ile sınıflandırma koşullu olasılık hesabına dayanmaktadır. Fig.4.1' de görüldüğü üzere tüm değerleri belirli geçmiş bir veri kümesinde, B yani sonuç sütunu, diğer $A_i, (i = 1,...,m)$ sütunlarına bağlı kabul edilerek, $P(B = b_j \mid A_i = a_{ik},...(i = 1,...,m))$, olasılıkları hesaplanır, burada $j = 1,...,s$ ve $k = 1,...,m_i$ dir. Bu ifade ile, her biri m_i tane farklı gruptan oluşan A_i sütunları a_{ik} değerlerini aldıklarında, bu A_i sütunlarına bağlı olarak, B sütununda bulunan s tane farklı grubun b_j değerlerinden her birini alma olasılıkları hesaplanmaktadır. Geçmiş veri kümesi yardımıyla hesaplanan bu olasılıklar, yeni gelecek verinin hangi gruba dahil edileceğinin yani B sütununun tahmininde kullanılacaktır. Ancak konuyu

anlaşılır kılmak için, tahmin edici sütun önce bir tane, A_1, sonra iki tane, A_1, A_2 alınarak, B sütununun bunlara bağlı olasılıkları hesaplanarak problem basitleştirilmiş daha sonra ise *m* sütun alınarak problem genelleştirilmiştir. Tüm bu işlemleri gerçekleştirmek için öncelikli olarak koşullu olasılık kavramının açıklanması gerekmektedir.

A ve B iki olay olmak üzere, bu olayların olma olasılıkları $P(A)$ ve $P(B)$ ile verilir. Eğer A ve B olaylarının gerçekleşmesi birbirine bağlı değilse, bu iki olayın birlikte olma olasılığı

$$P(A,B) = P(A) \times P(B) \tag{4.1}$$

ile verilir. Örneğin A olayı, o gün havanın yağmurlu olması ve B olayı ise atılan bir madeni paranın yazı gelme olasılığı ise, bu iki olay birbirinden bağımsızdır ve bu iki olayın birlikte olma olasılıkları her bir olayın olma olasılıklarının çarpımına eşittir.

Eğer A ve B olayları birbirine bağlı ise, bu iki olayın birlikte olma olasılıkları; A' nın olma olasılığı ile A' dan sonra B' nin olma olasılığının çarpımı ile yani

$$P(A,B) = P(A)\,P(B \mid A) \tag{4.2}$$

veya B' nin olma olasılığı ile B' den sonra A' nın olma olasılığının çarpımı ile yani

$$P(A,B) = P(B)\,P(A \mid B) \tag{4.3}$$

ile verilir. Dolayısıyla buradan (4.2) ve (4.3) denklemleri birbirine eşitlenerek, A olayından sonra B olayının olma olasılığı

$$P(B \mid A) = \frac{P(B)P(A \mid B)}{P(A)} \tag{4.4}$$

ile verilir. Örneğin A olayı havanın yağmurlu olması, B olayı ise Ali*' nin* balığa çıkma olayı ise, B olayının A olayına bağlı olduğu açıktır ve A olayından sonra B olayının olma olasılığı yani hava yağmurlu iken Ali' nin balığa çıkma olayı (4.4) ifadesiyle hesaplanır. Bir olayın olması ve olmaması olasılıkları toplamı $P(B) + P(B^{\perp}) = 1$ dir. Burada "$^{\perp}$" üst indisi B olayının değilini göstermektedir.

Dolayısıyla Ali hava yağmurlu değilken balığa çıktığı gibi, yağmur yağmazken de balığa çıkabilir, yani bir *B* olayına bağlı olarak *A* olayının olma olasılığı

$$P(A) = P(A,B) + P(A,B^{\perp}) = P(B)P(A \mid B) + P(B^{\perp})P(A \mid B^{\perp}) \tag{4.5}$$

şeklinde verilir. Bu ifade, (4.4)' te kullanılırsa,

$$P(B \mid A) = \frac{P(B)P(A \mid B)}{P(B)P(A \mid B) + P(B^{\perp})P(A \mid B^{\perp})} \tag{4.6}$$

elde edilir. Eğer *A* ve *B* olayları farklı değerler alabiliyorsa, örneğin Ali' nin balığa çıkması (b_1), işe gitmesi (b_2), spor yapması (b_3) gibi üç farklı *B* olayı varsa bu durumda $P(B = b_1) + P(B = b_2) + P(B = b_3) = 1$ dir. (4.5) ifadesine benzer bir şekilde bu kez *A* olayı r tane ayrık a_k ve *B* olayı s tane ayrık b_j değeri alıyorsa;

$$P\left(A = a_k\right) = \sum_{j=1}^{s} P\left((A = a_k), (B = b_j)\right) = \sum_{j=1}^{s} P\left(B = b_j\right) P\left((A = a_k) \mid (B = b_j)\right) \tag{4.7}$$

elde edilir. (4.7) ifadesi (4.4)' te yerine yazıldığında ise,

$$P\left((B = b_j) \mid (A = a_k)\right) = \frac{P\left(B = b_j\right) P\left((A = a_k) \mid (B = b_j)\right)}{\sum_{k=1}^{s} P\left(B = b_k\right) P\left(A \mid (B = b_k)\right)} \tag{4.8}$$

şeklinde yazılabilir. (4.8) ifadesinin *A* ve *B* olaylarının ikiden fazla değer alabildikleri durum için (4.6) ifadesinin genelleştirilmiş hali olduğu açıktır. Bu ifade Şekil.4.1' de verilen tabloda *B* sonuç sütununu tahmin edici tek bir A_1 sütunu olması halinde *B* sütununun alabileceği değerlerin olasılıklarının hesaplanmasında kullanılır. Ancak gerçek hayatta sadece biri tahmin edici, diğeri hedef sütun olmak üzere iki sütun olması değil, hedef sütunu tahmin edici bir çok sütun bulunması beklenir.

Bu nedenle (4.8) ifadesinde *A* gibi sadece bir tahmin edici sütun yerine *m* tane A_i sütunu olduğunu ve bunların her birinin r_i tane bağımsız değer alabildiği yani örneğin A_1 sütunu $r_1 = 5$, A_2 sütunu $r_2 = 3$ farklı değer alabildiğini varsayalım. Bu durumda (4.8) ifadesinde *A* yerine $A_1, A_2, ..., A_m$ gibi m tane olay alınırsa

$$P\left(B=b_j \mid A_1=a_{1j_1}, A_2=a_{2j_2},..., A_m=a_{mj_m}\right)=$$
$$\frac{P\left(B=b_j\right)P\left(A_1=a_{1j_1}, A_2=a_{2j_2},..., A_m=a_{mj_m} \mid B=b_j\right)}{\sum_{k=1}^{s} P\left(B=b_k\right)P\left(A_1=a_{1j_1}, A_2=a_{2j_2},..., A_m=a_{mj_m} \mid B=b_k\right)} \quad (4.9)$$

ifadesi elde edilir. Tahmin edici her sütunun yani her A_i olayının birbirinden bağımsız olduğu kabulü yapılırsa, sonuç olarak

$$P(B=b_k \mid A_1=a_{1j_1}, A_2=a_{2j_2},..., A_m=a_{mj_m})=\frac{P(B=b_k)\times\prod_{i=1}^{m} P(A_i=a_{ij_i} \mid B=b_k)}{\sum_{\forall r \mid b_r \in B}\left(P(B=b_r)\times\prod_{i=1}^{m} P(A_i=a_{ij_i} \mid B=b_r)\right)} \quad (4.10)$$

ifadesi elde edilir. Burada $j_i=1,...,m_i$ ve $k=1,...,s$ için bu olasılık değerleri hesaplanmalıdır, ayrıca $\forall r \mid b_r \in B$ terimi hedef sütunun alabileceği tüm farklı değerler üzerinde toplam alınacağını ifade etmektedir.

4.2.2. Örnekler

Örnek.1. Baş gösteren bir salgın sonucu bir bölgede yaşayanların % 30' unun hasta olduğu tahmin ediliyor ve hastalık taşıyan kişileri belirlemek için ön bir sağlık testi yapılıyor. Testin geçmiş uygulamalarından eğer gerçekten hasta olan bir kişiye uygulanmış ise %95 doğru sonuç, gerçekte hasta olmayan bir kişiye uygulanmış ise %10 yanlış sonuç verdiği biliniyor. Bu durumda;

a. Testin pozitif sonuç (kişinin hasta olması sonucunu) verme
b. Pozitif sonuç verilen bir kişinin gerçekte hasta olma
c. Negatif sonuç verilen bir kişinin gerçekte sağlam olma
d. Testin uygulandığı kişinin yanlış sınıflandırılma

olasılıklarını hesaplayınız.

Çözüm;

"*T*" : "Testin pozitif (hasta) sonuç verme"
"*H*" : "Kişinin gerçekten hasta olma"
"*Y*" : "Kişinin yanlış sınıflandırılma"

olayları olsun. Bu durumda verilenlerden

$P(T \mid H) = 0.95$ (Kişi gerçekte hasta iken testin de pozitif (hasta) sonuç verme olasılığı),

$P(T \mid H^{\perp}) = 0.1$ (Kişi gerçekte hasta değilken testin pozitif (hasta) sonucunu verme olasılığı),

$P(H) = 0.3$ (Topluluktan rastgele seçilen birinin hasta olma olasılığı)

yazılabilir.

a. Testin pozitif sonuç vermesi iki şekilde mümkün olabilir; Kişi hastadır ve test pozitif sonuç vermiştir veya kişi sağlamdır ama test yine de pozitif (hasta) sonuç vermiştir. O halde, (3.5) denkleminden;
$P(T) = P(H)P(T \mid H) + P(H^{\perp})P(T \mid H^{\perp}) = 0.3\ 0.95 + 0.7\ 0.1 = 0.355$
ihtimalle test pozitif sonuç verir.

b. Pozitif sonuç verilen bir kişinin gerçekten hasta olma olasılığı $P(H \mid T)$, (3.6) denkleminden;
$$P(H \mid T) = \frac{P(H)\ P(T \mid H)}{P(H)P(T \mid H) + P(H^{\perp})P(T \mid H^{\perp})} = \frac{0.3\ 0.95}{0.3\ 0.95 + 0.7\ 0.1} = 0.802817$$
olarak elde edilir.

c. Negatif (hasta değil) sonuç verilen bir kişinin gerçekten sağlam olma olasılığı $P(H^{\perp} \mid T^{\perp})$, (3.6) denkleminden
$$P(H^{\perp} \mid T^{\perp}) = \frac{P(H^{\perp})\ P(T^{\perp} \mid H^{\perp})}{P(H)P(T^{\perp} \mid H) + P(H^{\perp})P(T^{\perp} \mid H^{\perp})} = \frac{0.7\ 0.9}{0.3\ 0.05 + 0.7\ 0.9} = 0.976744$$
olarak elde edilir.

d. Kişinin yanlış sınıflandırılma olasılığı, $P(Y)$,
$P(Y) = P(H^{\perp})P(T \mid H^{\perp}) + P(H)P(T^{\perp} \mid H) = 0.7\ 0.1 + 0.3\ 0.05 = 0.085$
olarak hesaplanır.

Örnek.2. *Tek boyut için:* Yapılan bir anket sonucunda 1000 deneğin gelir durumları "düşük", "orta", "iyi" ve "yüksek" olarak gruplanmış ve "Ev sahibi" olup olmadıkları ise ikinci bir sütunda Tablo.4.1.a' da ki gibi belirtilmiş olsun.

Her ne kadar, ODM bu olasılık hesaplarını arka planda otomatik olarak işleyip kullanıcıya sadece sonucu bildirse de, burada amaç doğrultusunda arka planda neler döndüğünü açıklanmıştır. Burada kısaltma amacıyla Gelir=G, Evet=E, Hayır=H

şeklinde sembolize edilecektir. Tablo.4.1.a verisinden bir Sql cümleciği aracılığıyla elde edilen her farklı gruptaki kişi sayısı Tablo.4.1.b ile gösterilmiştir.;

Gelir	Ev
Düşük	Evet
Düşük	Evet
Düşük	Hayır
Orta	Hayır
Yüksek	Hayır
Düşük	Evet
İyi	Hayır
⋮	⋮

Tablo.4.1.a. Gelir-Mülk ilişkisi

Gelir	Ev=E	Ev=H
Düşük	250	150
Orta	50	200
İyi	150	80
Yüksek	100	20

Tablo.4.1.b. Her gruptaki kişi sayısı

Tablo.4.1.b yardımıyla sözü edilen olasılıklar (4.8) ifadesi yardımıyla

$$P(Ev=E\,|\,Gelir=D)=\frac{P(Ev=E)\,P(G=D\,|\,Ev=E)}{P(Ev=E)\,P(G=D\,|\,Ev=E)+P(Ev=H)\,P(G=D\,|\,Ev=H)}$$

$$=\frac{\frac{550}{1000}\frac{250}{550}}{\frac{550}{1000}\frac{250}{550}+\frac{450}{1000}\frac{150}{450}}=0.625,$$

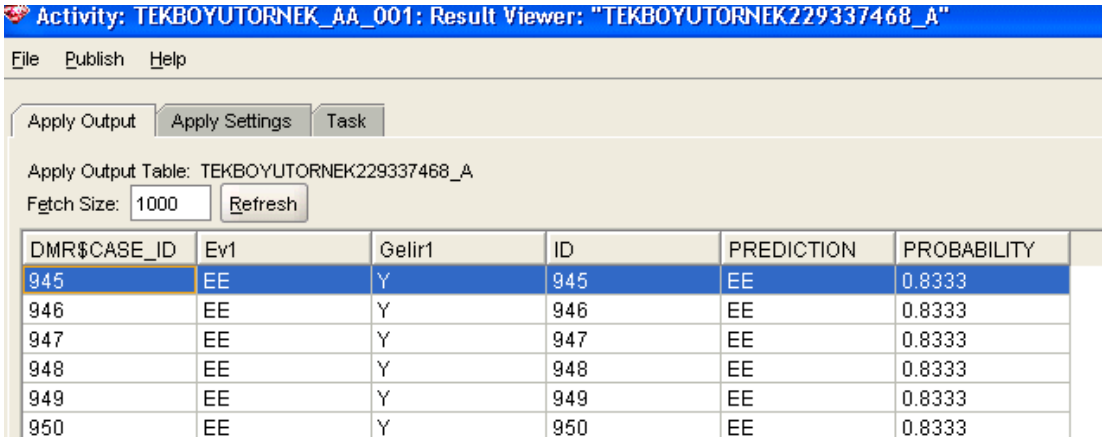
$$P(Ev=E\,|\,Gelir=D)=\frac{250}{400}=0.625$$

$$P(Ev=H\,|\,Gelir=D)=1-0.625=0.375$$

olarak hesaplanabilir. Burada bu sonuçlar çok daha kolay bir şekilde Tablo.4.1.b den de görülmektedir. Ama hem hedef özelliğin ikiden fazla hem de kestirimci özellik sayısının birden fazla olduğu durumlarda tablodan okuma zorlaşacak ve yukarıdaki formülün daha kolay uygulanabileceği açıktır. Benzer şekilde diğer olasılıklarda hesaplanarak;

$$P(Ev=E\,|\,Gelir=O)=0.2,\qquad P(Ev=H\,|\,Gelir=O)=1-0.2=0.8,$$
$$P(Ev=E\,|\,Gelir=\dot{I})=0.652174,\qquad P(Ev=H\,|\,Gelir=\dot{I})=1-0.652174=0.347826,$$
$$P(Ev=E\,|\,Gelir=Y)=0.833333,\qquad P(Ev=H\,|\,Gelir=Y)=1-0.833333=0.166667$$

yazılabilir. Yukarıdaki hesaplamaların ODM nin elde ettiği sonuçlarla aynı olduğunu göstermek için aşağıda ODM nin bu örneğe uygulanması sonucu elde edilen ekran çıktısı verilmiştir.



DMR$CASE_ID	Ev1	Gelir1	ID	PREDICTION	PROBABILITY
945	EE	Y	945	EE	0.8333
946	EE	Y	946	EE	0.8333
947	EE	Y	947	EE	0.8333
948	EE	Y	948	EE	0.8333
949	EE	Y	949	EE	0.8333
950	EE	Y	950	EE	0.8333
943	EE	Y	943	EE	0.8333
944	EE	Y	944	EE	0.8333
593	H	O	593	H	0.8
594	H	O	594	H	0.8
595	H	O	595	H	0.8
463	H	O	463	H	0.8
464	H	O	464	H	0.8
651	EE	II	651	EE	0.6522
652	EE	II	652	EE	0.6522
653	EE	II	653	EE	0.6522
879	H	II	879	EE	0.6522
880	H	II	880	EE	0.6522
385	H	DD	385	EE	0.625
386	H	DD	386	EE	0.625
387	H	DD	387	EE	0.625
63	EE	DD	63	EE	0.625
64	EE	DD	64	EE	0.625
385	H	DD	385	H	0.375
386	H	DD	386	H	0.375
387	H	DD	387	H	0.375
63	EE	DD	63	H	0.375
64	EE	DD	64	H	0.375
843	H	II	843	H	0.3478
844	H	II	844	H	0.3478
845	H	II	845	H	0.3478
713	EE	II	713	H	0.3478
714	EE	II	714	H	0.3478
593	H	O	593	EE	0.2
594	H	O	594	EE	0.2
595	H	O	595	EE	0.2
463	H	O	463	EE	0.2
464	H	O	464	EE	0.2
945	EE	Y	945	H	0.1667
946	EE	Y	946	H	0.1667
947	EE	Y	947	H	0.1667

ODM, Naive Bayes yöntemiyle sınıflandırmaya mümkün olan her durum için olasılıkları hesaplayarak bir model oluşturup, bu modeli yukarıdaki gibi aynı tablo üzerinde veya yeni kayıtların durumunu tespit için kullanmaktadır. Modelin doğruluğunun test edilmesi amacıyla formüllerle yapılacak işlemlerde aynı veri ve hesaplanan olasılıklar kullanılarak yeni tahmin tablosu oluşturulabilir. Örneğin geliri orta olan kişinin ev sahibi olma olasılığı 0.2 olduğu için modelin tahmini hayır ve sonucun güvenilirliği 0.8 olacaktır. Geliri düşük olanın ev sahibi olma olasılığı ise 0.625 olarak hesaplandığı için tahmin evet ve sonucun güvenilirliği 0.625 tir. Bu şekilde işleme devam edilerek tüm tablo yeniden oluşturulur. Modelin tüm güvenilirliği ise gerçek değerler ile tahmini değerlerin karşılaştırılması sonucu elde edilen aşağıdaki güvenilirlik matrisi ile verilebilir.

Tablo.4.2. Güvenilirlik matrisi

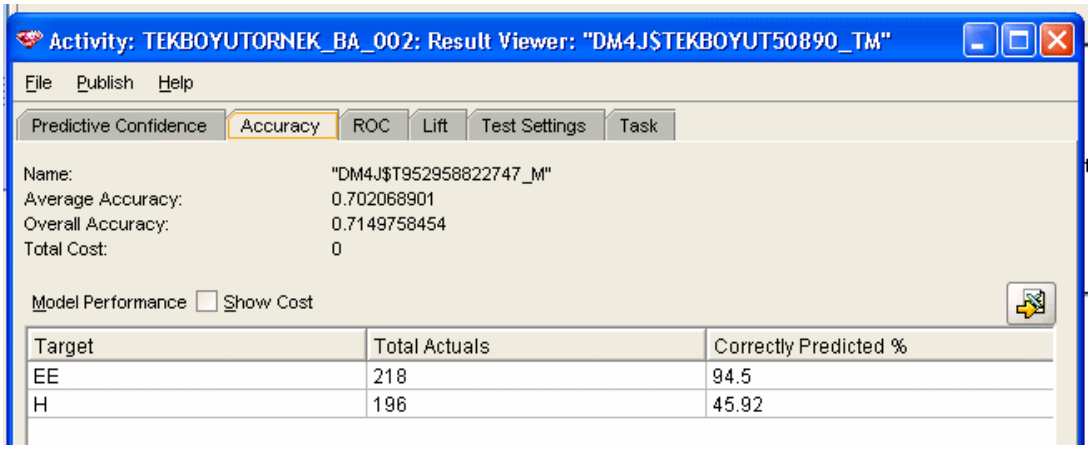
	E	**H**
E	517	85
H	115	283

Doğruluk = 0.8379

Tablo.4.2 de görülen güvenilirlik matrisinde satırlar gerçek değerleri, sütunlar ise tahmin sonuçlarını göstermektedir. Örneğin gerçekte evi varken, modelin de evet yani "evi var" olarak tahmin ettiği kayıt sayısı 517 (doğru), gerçekte evi varken modelin hayır olarak tahmin ettiği kayıt sayısı (yani yanlış) 85 tir. Dolayısıyla matrisin köşegeni doğru kayıt sayısını, köşegen dışı ise yanlış kayıt sayısını göstermektedir. Buradan modelin doğruluğu

$$\frac{517+283}{(517+283)+(115+85)} = 0.8$$

olarak elde edilir. Modelin güvenilirliği ODM kullanılarak da hesaplanabilir. Ancak ODM ile model oluştururken verinin bir kısmını model, bir kısmını test için ayırma zorunluluğundan dolayı yukarıdaki veri %60 oranında model, %40 oranında test için ayrılarak ODM den elde edilen güvenilirlik sonucu aşağıdaki ekran çıktısında verilmiştir. Model, formüllerle hesaplanan duruma göre daha az veri kullandığı için güvenilirliğin biraz daha kötü çıkması doğaldır.



Örnek.3. *İki Boyut için:* Yapılan bir anket sonucunda 100 deneğin gelir durumları "Düşük", "Orta" ve "Yüksek"; ev sahibi olmaları ise "Var=1" ve "Yok=0" olarak belirlenmiş ve bu özelliklerdeki deneklerin araba sahibi olmaları ve varsa hangi model olduğu ise 3. bir sütunda verilmiş olsun (Tablo.4.3.a). Kayıtların dağılımı Tablo.4.3.b' de verilmiştir.

Tablo.4.3.a Model verisi

Gelir	Ev	Araba
Orta	0	BMW
Düşük	1	Yok
Yüksek	1	Yok
Orta	0	BMW
Orta	1	Megane
Yüksek	1	Yok
Düşük	0	Yok
Yüksek	1	Yok
Düşük	1	Megane
.	.	.
.	.	.
.	.	.

Tablo.4.3.b Kayıt dağılımları

Gelir	Ev	Araba	Kayıt Sayısı
Düşük	0	Yok	16
Düşük	1	Yok	9
Orta	0	Yok	0
Orta	1	Yok	6
Yüksek	0	Yok	0
Yüksek	1	Yok	22
Düşük	0	Megane	6
Düşük	1	Megane	1
Orta	0	Megane	2
Orta	1	Megane	3
Yüksek	0	Megane	0
Yüksek	1	Megane	2
Düşük	0	BMW	2
Düşük	1	BMW	0
Orta	0	BMW	23
Orta	1	BMW	1
Yüksek	0	BMW	4
Yüksek	1	BMW	3

Tablo.4.3.b yardımıyla gelir grupları ve ev sahiplerinin hangi arabaya sahip olduklarını (4.10) ifadesiyle belirlenmeye çalışılmıştır. Burada iki belirleyici özellik olduğundan (4.10) ifadesi bu tabloya uygun formda yazılmalıdır. Burada A=Araba, G=Gelir, E=Ev, D=Düşük, Y=Yüksek, O=Orta, MG=Megane' ı ifade etmektedir.

(4.10) ifadesi örneğe uygulandığında,

$$P(A=Yok \mid G=D, E=0) = \frac{P(A=Yok)xP(G=D \mid A=Yok)xP(E=0 \mid A=Yok)}{\left(\begin{array}{l} P(A=Yok)xP(G=D \mid A=Yok)xP(E=0 \mid A=Yok) \\ +P(A=MG)xP(G=D \mid A=MG)xP(E=0 \mid A=MG) \\ +P(A=BMW)xP(G=D \mid A=BMW)xP(E=0 \mid A=BMW) \end{array}\right)}$$

$$= \frac{\frac{53}{100}x\frac{25}{53}x\frac{16}{53}}{\left(\frac{53}{100}x\frac{25}{53}x\frac{16}{53}\right)+\left(\frac{14}{100}x\frac{7}{14}x\frac{8}{14}\right)+\left(\frac{33}{100}x\frac{2}{33}x\frac{29}{33}\right)} = 0.567254$$

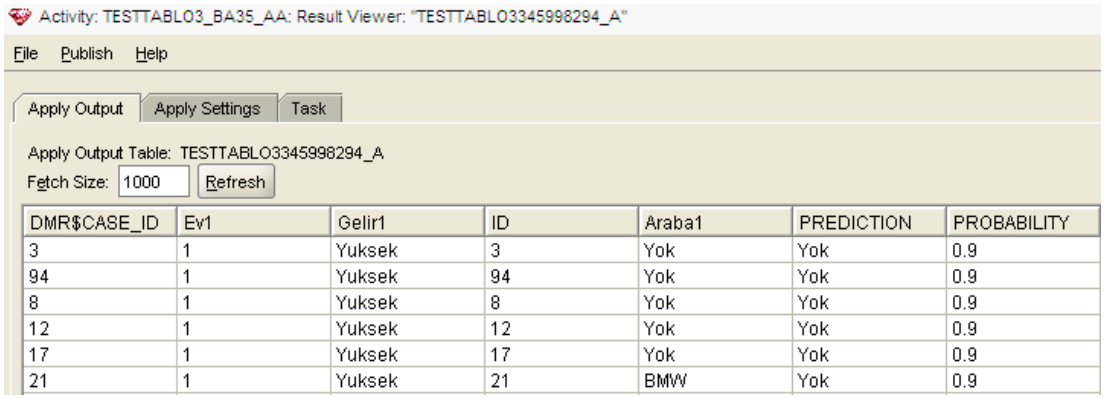
sonucu elde edilir. Bu sonuca göre geliri düşük ve evi olmayan kişilerin %56.72 olasılıkla arabası olmayacaktır. Benzer şekilde diğer bazı olasılıklar da hesaplanırsa;

$P(A=Yok \mid G=Y, E=1) = 0.9001$ $\quad$ $P(A=Yok \mid G=Y, E=0) = 0.4766$

$P(A=Yok \mid G=D, E=1) = 0.8433$ $\quad$ $P(A=MG \mid G=D, E=0) = 0.3006$

$P(A=BMW \mid G=O, E=0) = 0.8188$ $\quad$ $P(A=Yok \mid G=O, E=0) = 0.0703$

Yukarıdaki hesaplamalarla ODM nin elde ettiği sonuçları karşılaştırmak için aşağıda ODM nin bu örneğe uygulanması sonucu elde edilen ekran çıktısı verilmiştir ve yukarıdaki sonuçlarla tutarlı olduğu görülmektedir.



DMR$CASE_ID	Ev1	Gelir1	ID	Araba1	PREDICTION	PROBABILITY
3	1	Yuksek	3	Yok	Yok	0.9
94	1	Yuksek	94	Yok	Yok	0.9
8	1	Yuksek	8	Yok	Yok	0.9
12	1	Yuksek	12	Yok	Yok	0.9
17	1	Yuksek	17	Yok	Yok	0.9
21	1	Yuksek	21	BMW	Yok	0.9
6	1	Yuksek	6	Yok	Yok	0.9
2	1	Dusuk	2	Yok	Yok	0.8433
9	1	Dusuk	9	Megane	Yok	0.8433
20	1	Dusuk	20	Yok	Yok	0.8433
24	1	Dusuk	24	Yok	Yok	0.8433
98	1	Dusuk	98	Yok	Yok	0.8433
44	1	Dusuk	44	Yok	Yok	0.8433
65	1	Dusuk	65	Yok	Yok	0.8433
81	1	Dusuk	81	Yok	Yok	0.8433
96	1	Dusuk	96	Yok	Yok	0.8433

35	1	Dusuk	35	Yok	Yok	0.8433
1	0	Orta	1	BMW	BMW	0.8188
99	0	Orta	99	BMW	BMW	0.8188
13	0	Orta	13	BMW	BMW	0.8188
15	0	Orta	15	BMW	BMW	0.8188
19	0	Orta	19	BMW	BMW	0.8188
22	0	Orta	22	BMW	BMW	0.8188
25	0	Orta	25	BMW	BMW	0.8188

4	0	Orta	4	BMW	BMW	0.8188
7	0	Dusuk	7	Yok	Yok	0.5673
100	0	Dusuk	100	Yok	Yok	0.5673
16	0	Dusuk	16	Yok	Yok	0.5673
18	0	Dusuk	18	Megane	Yok	0.5673
23	0	Dusuk	23	BMW	Yok	0.5673
29	0	Dusuk	29	Megane	Yok	0.5673
33	0	Dusuk	33	Yok	Yok	0.5673

14	0	Dusuk	14	Yok	Yok	0.5673
11	0	Yuksek	11	BMW	Yok	0.4766
95	0	Yuksek	95	BMW	Yok	0.4766
64	0	Yuksek	64	BMW	Yok	0.4766
27	0	Yuksek	27	BMW	Yok	0.4766
5	1	Orta	5	Megane	Yok	0.4533

7	0	Dusuk	7	Yok	Megane	0.3006
100	0	Dusuk	100	Yok	Megane	0.3006
16	0	Dusuk	16	Yok	Megane	0.3006
18	0	Dusuk	18	Megane	Megane	0.3006
23	0	Dusuk	23	BMW	Megane	0.3006
29	0	Dusuk	29	Megane	Megane	0.3006

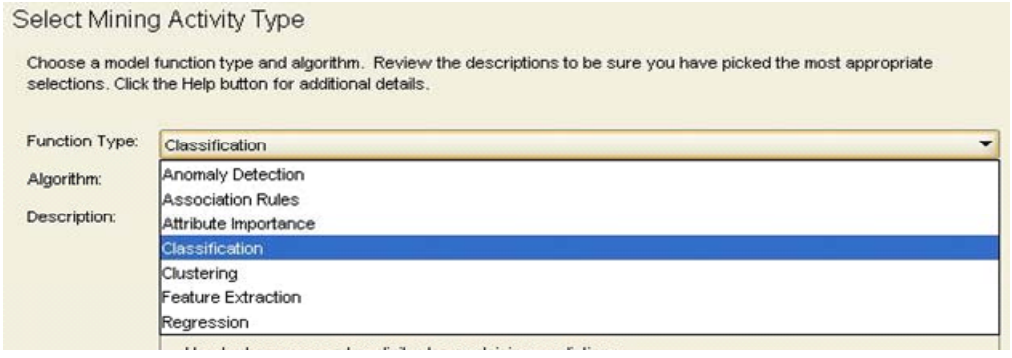
27	0	Yuksek	27	BMW	Megane	0.082
1	0	Orta	1	BMW	Yok	0.0703
99	0	Orta	99	BMW	Yok	0.0703
13	0	Orta	13	BMW	Yok	0.0703
15	0	Orta	15	BMW	Yok	0.0703
19	0	Orta	19	BMW	Yok	0.0703
22	0	Orta	22	BMW	Yok	0.0703
25	0	Orta	25	BMW	Yok	0.0703
30	0	Orta	30	BMW	Yok	0.0703

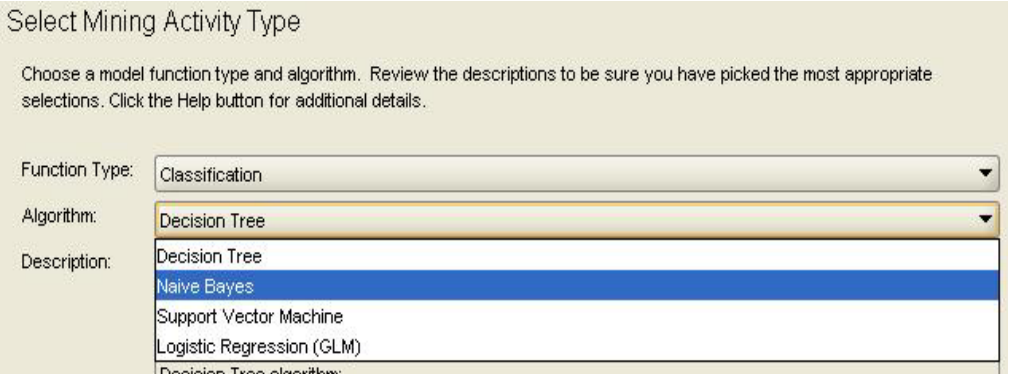
5. UYGULAMA VE SONUÇLAR

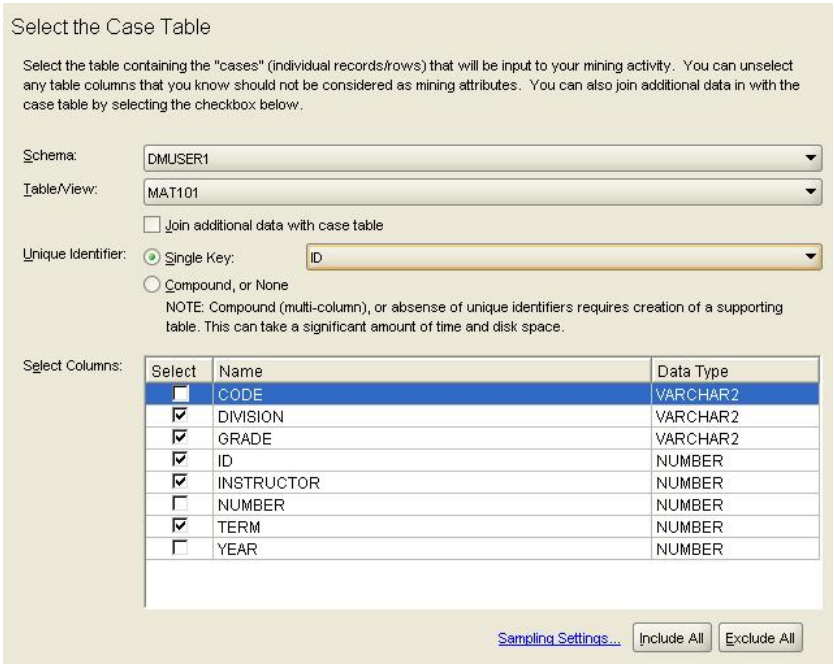
Bu bölümde İstanbul Teknik Üniversitesi' nde verilmekte olan matematik havuz derslerinin bölüm ve öğretim üyelerine göre başarı oranlarının hesaplandığı bir veri madenciliği modeli oluşturulmuş ve elde edilen sonuçlar detaylı olarak incelenerek yorumlanmıştır. İlk olarak İTÜ Öğrenci İşleri Daire Başkanlığı' ndan MAT101(E), MAT102, MAT103(E), MAT104(E), MAT201(E), MAT202 ve MAT261 derslerinin 1999-2008 yılları arasındaki yaklaşık 80000 kaydı (Numara, Bölüm, Öğretim Görevlisi, Yıl, Dönem, Not) formatında alınmıştır. Veriler ayıklanmış ve her ders için bir tablo oluşturulmuştur. Bu tablolar ODM içerisine aktarılarak her tablo için model oluşturulmuştur. Oluşturulan bu modeller ilgili tablolara uygulanarak tahminler elde edilmiştir. Burada her ne kadar 3. bölümde ODM tanıtılırken model oluşturma ve uygulama aşamalarına değinildiyse de, konu bütünlüğü ve uygulamanın öğrenci kayıtları üzerine olması nedeniyle en azından sadece MAT101 dersi için tüm aşamalar daha geniş açıklamalı ve ekran çıktıları ile birlikte verilerek, elde edilen tüm sonuçlar yorumlanmıştır.

5.1. Model Oluşturma

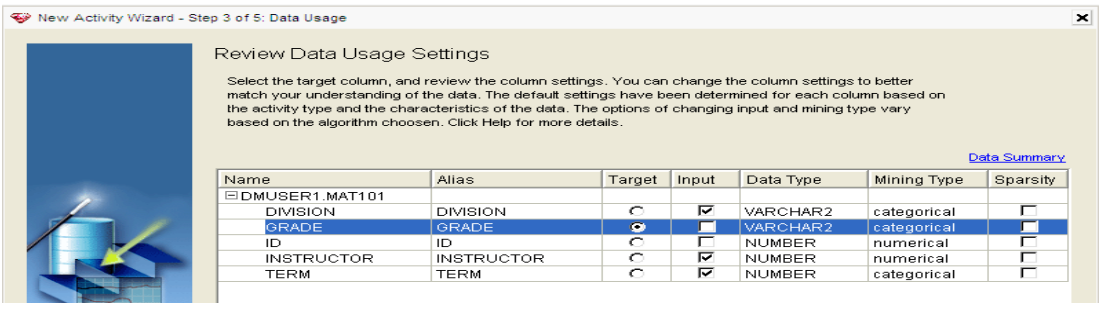
Not sistemi üzerinde yapılan analiz çalışması sonrasında problemin sınıflandırma modeline uygun yapıda olduğu belirlenmiş ve bu bilgiler fonksiyon tipi olarak "Classification", çözüm algoritması olarak da "Naive Bayes" seçilerek MAT101 dersi için model oluşturma aşamaları ekran çıktıları ile birlikte aşağıda verilmiştir.



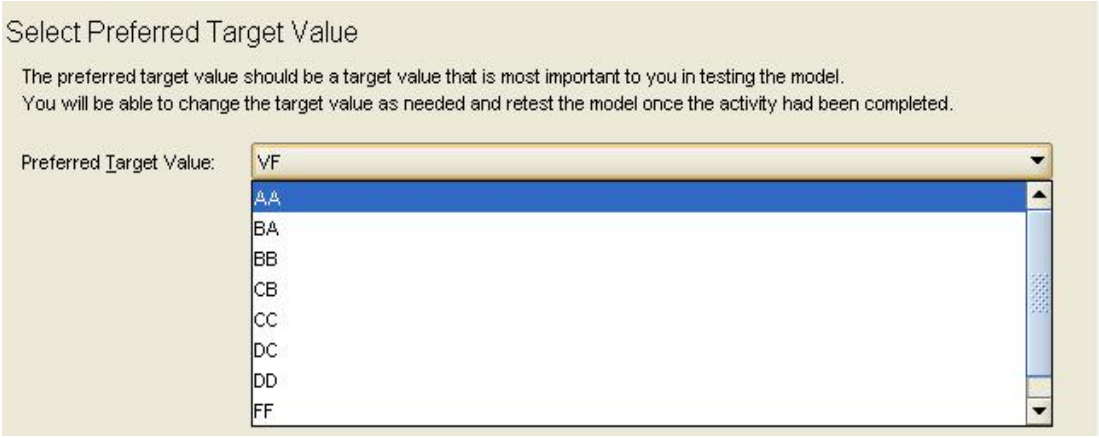




MAT101 dersinin modeli oluşturulacağından bu aşamada tablo olarak MAT101 isimli tablo seçilmiştir. “Unique Identifier” kısmı zorunlu olup, tabloda benzersiz değerlere sahip olan sütun “single key” olarak seçilir. Bu tabloda benzersiz değer ID sütunudur. Bu modelde kullanılacak sütunlar “DIVISION”, “GRADE”, “ID”, “INSTRUCTOR” ve “TERM” olarak belirlenmiştir. Burada seçilmeyen “NUMBER” ve “YEAR” sütunları yeni gelecek öğrenciler için mevcut olmadığından modele dahil edilmemiştir.

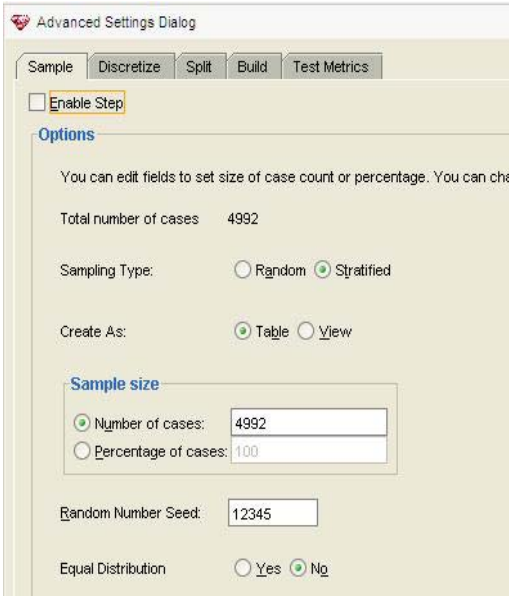


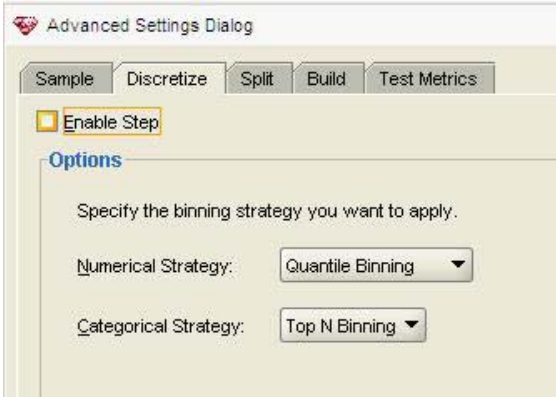
Hedef sütun olarak başarı notu yani “GRADE” seçilmiş ve tahmin edici özellik olarak seçilmemesi gereken “ID” ve “GRADE” özellikleri girdi olarak alınmamıştır.



Tercih edilen hedef değeri “AA” olarak seçilerek bir sonraki adımda model ismi verilmiştir.

Son adımda yer alan “Advanced Settings” seçilerek model oluşturma aşamasında yürütülecek adımlar üzerinde gerekli ayarlar yapılmıştır.

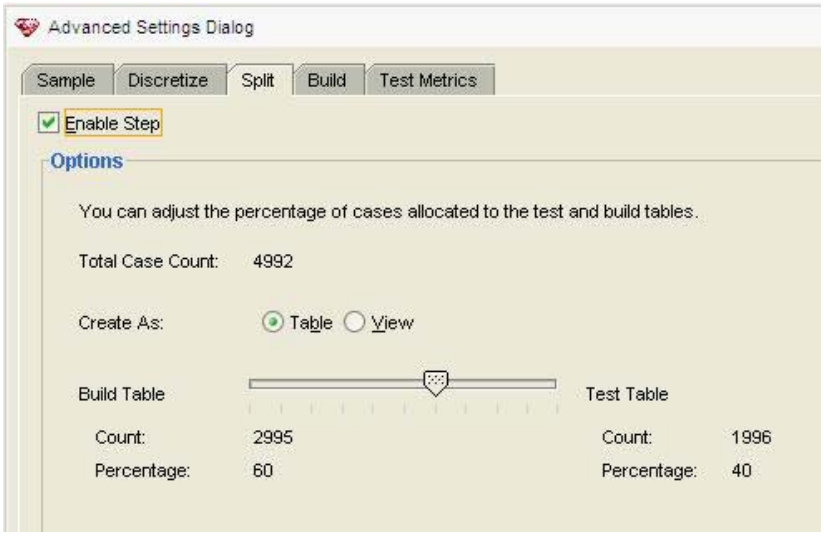




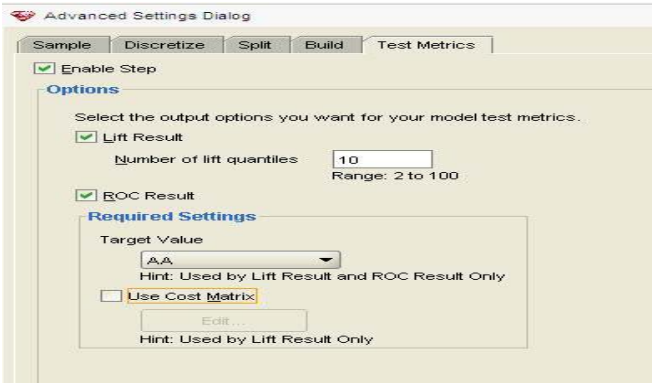
"Sample" seçeneği verinin çok fazla kayıt içermesi durumunda, bu veri ile oluşturulacak model için çok fazla zaman gerekeceğinden, verinin tamamı ile değil de, içinden örnekleme yapılarak daha az veri ile model oluşturmak amacıyla kullanılmaktadır.

"Discretize" seçeneği ise içinde çok farklı ve fazla kayıt bulunduran sayısal verilerden gruplar oluşturma amacıyla kullanılmaktadır.

Yukarıdaki tanımlar çerçevesinde, bu uygulama için "Sample" ve "Discretize" seçeneklerinin kullanılmaması gerektiği açıktır ve bu nedenle modelde bunların seçili olmamasına dikkat edilmelidir.

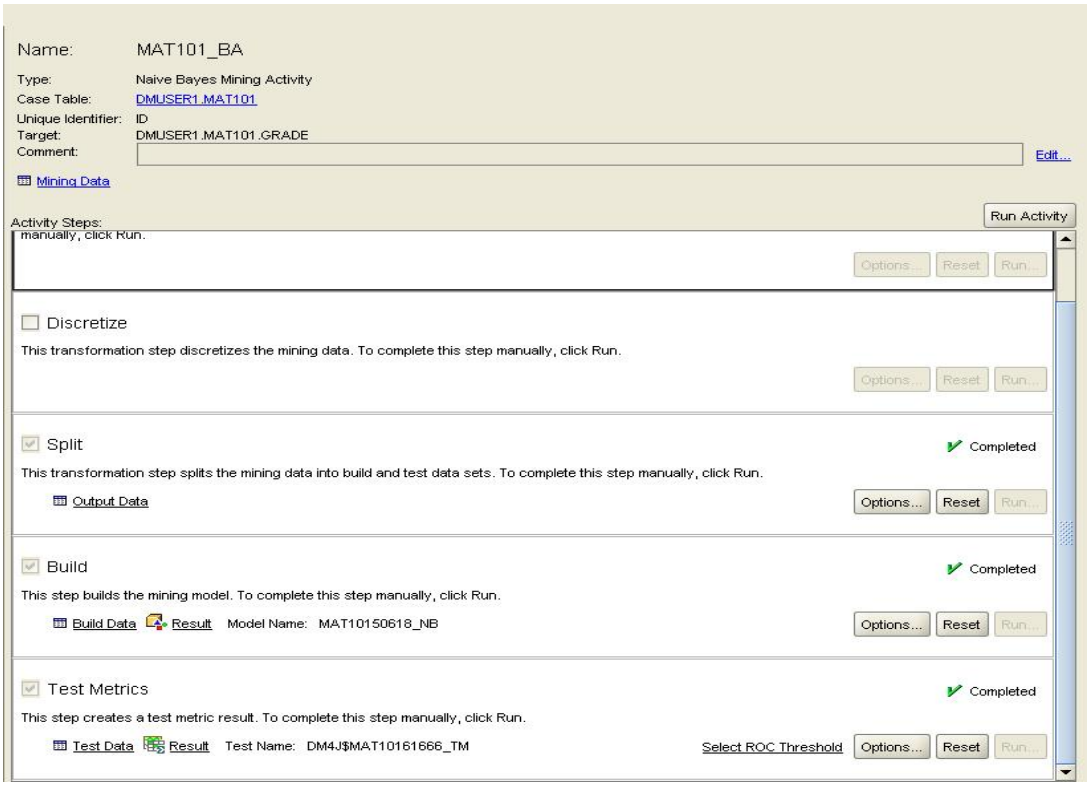


"Split" sekmesinde, test için ayrılacak veri yüzdesi ayarlanmıştır. Model oluşturma adımı "Test" adımını da kapsadığı için, ayrıca test yapılmasına gerek görülmemiştir. Burada ayrılmış olan test verisine, oluşturulan model uygulanmakta ve tahmin edilen değerler gerçek değerlerle karşılaştırılarak, kurulan model için bir güvenilirlik değeri hesaplanmaktadır.



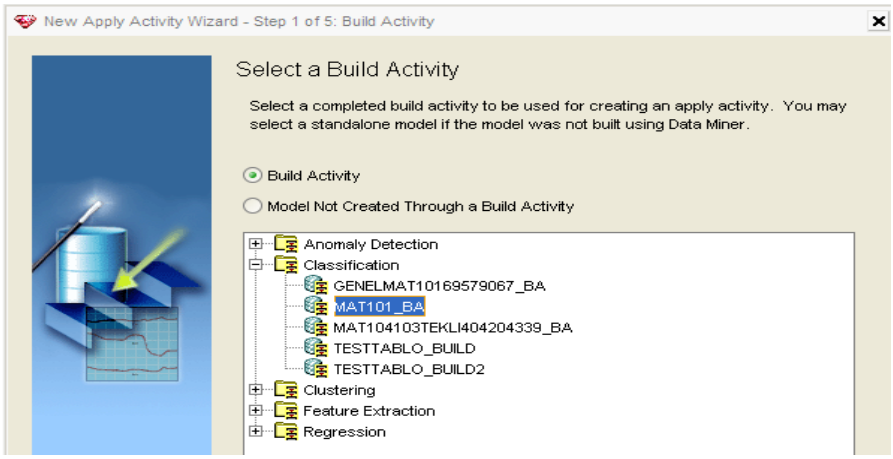
"Test Metrics" sekmesinde bulunan "Cost Matrix" seçeneği de kaldırılmalıdır, çünkü burada tahmin edilecek olan başarı notu sütununda diğerlerine göre daha iyi tahmin edilmesi istenen bir değer yoktur.

Gelişmiş ayarlar yapıldıktan sonra model oluşturma süreci başlatılır. Bu sürecin tamamlanmış hali aşağıdaki gibidir.

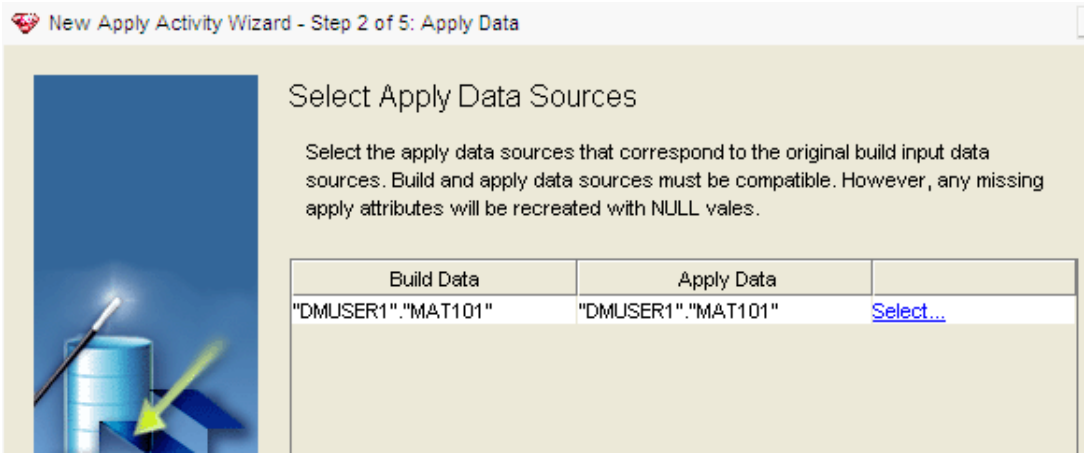


5.2. Modelin Uygulanması

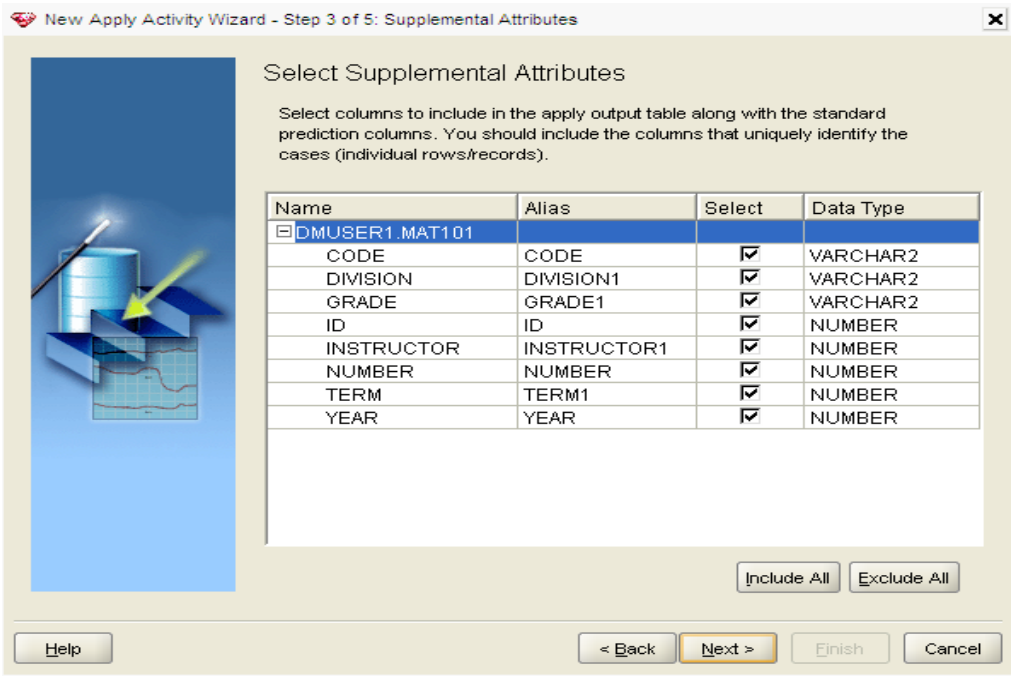
Bu aşamada önceki adımda oluşturulan modelin ders kayıtlarına uygulanması ayrıntılı olarak anlatılmaktadır.



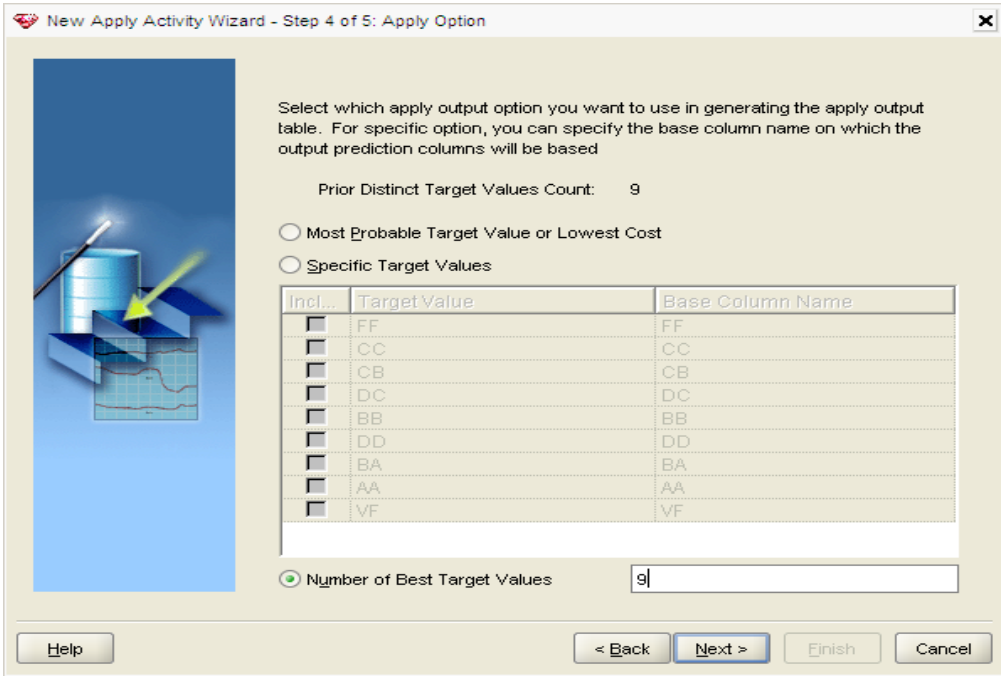
Uygulama için listelenen modellerden, bir önceki aşamada oluşturulan MAT101_BA modeli seçilmiştir ve bir sonraki adımda modelin uygulanacağı tablo aşağıdaki gibi görüntülenmiştir.



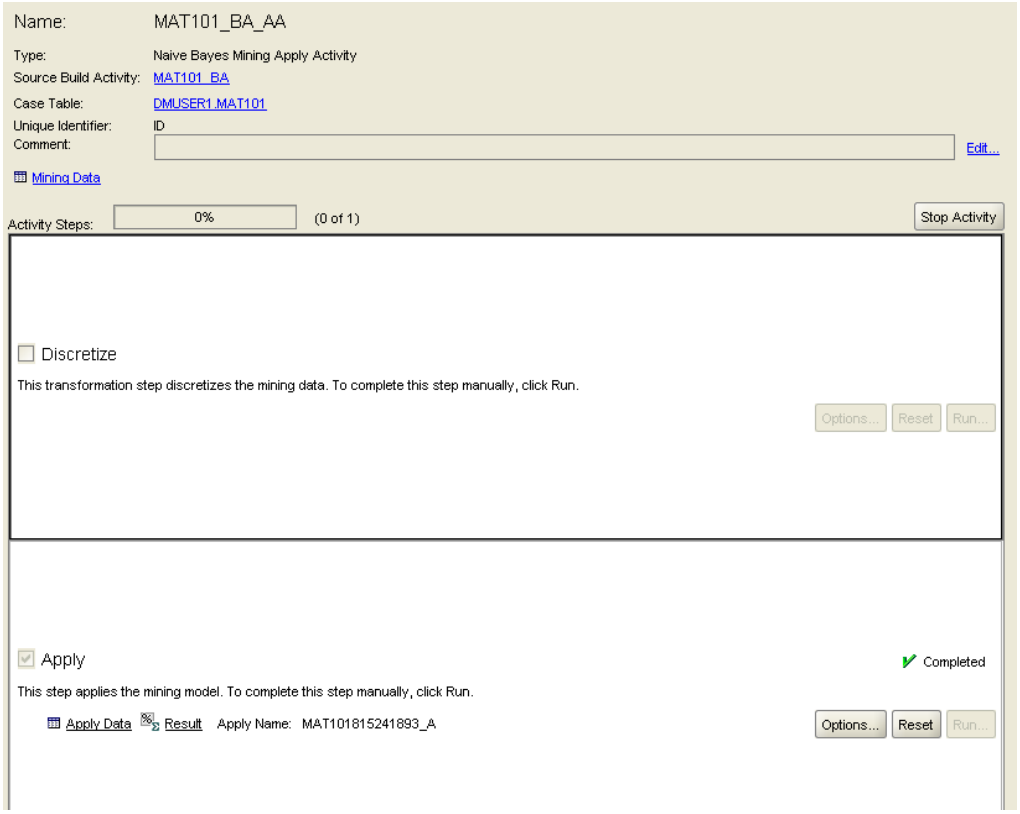
Uygulama sonucunda görüntülenmek istenen sütunlar seçilerek bir sonraki adıma geçilmiştir.



Her bir kayıt için en yüksek olasılığa sahip hedef değerler görüntülenmek istendiğinden en iyi hedef değerleri sayısı, farklı not değerleri 9 adet olduğundan 9 olarak atanmıştır.



Son adımda uygulamaya isim verilerek uygulama süreci başlatılmıştır.



Süreç tamamlandıktan sonra "Result" yolundan istenen olasılık ve tahminler görüntülenebilmektedir.

5.3 Sonuçlar ve Yorumlar

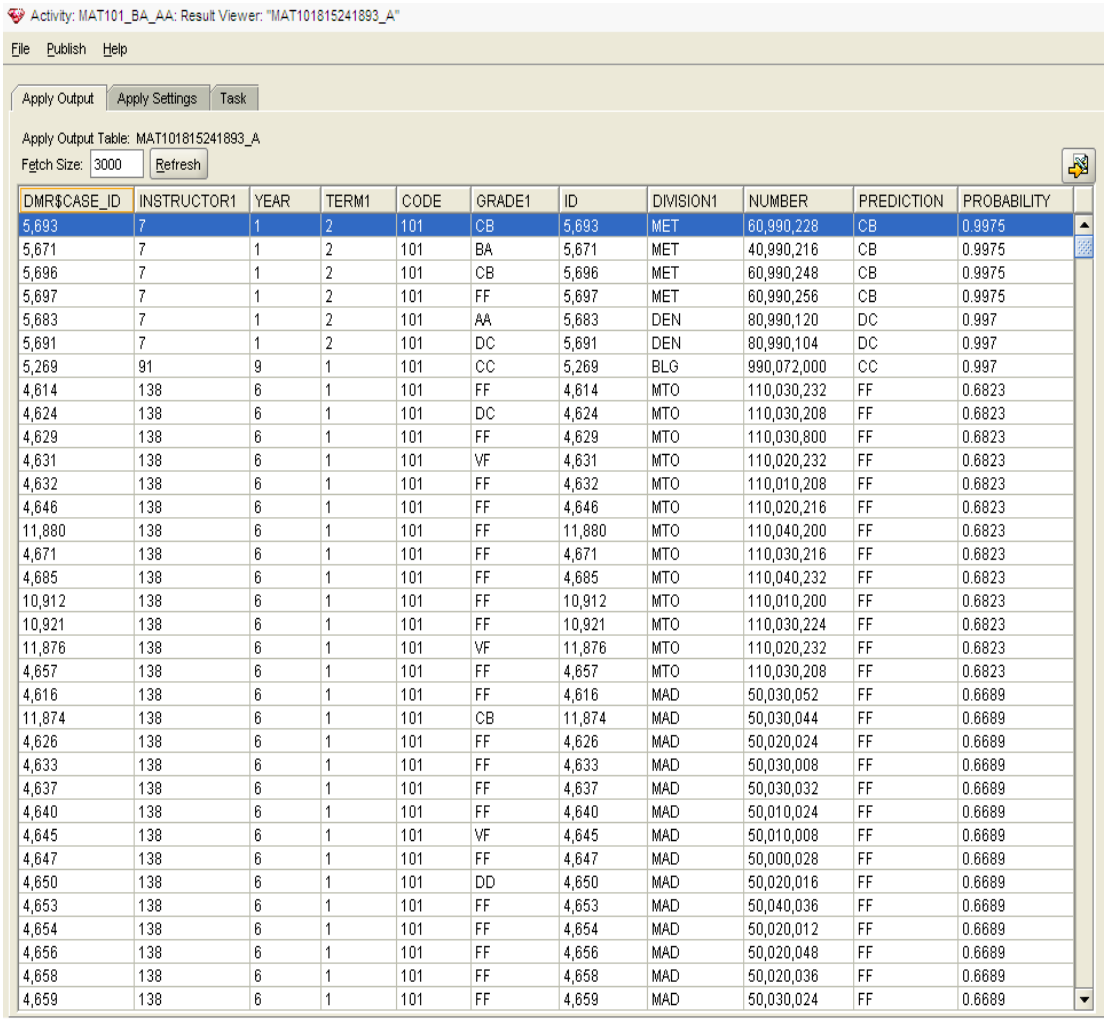
Ders tablolarından elde edilen sonuçlara geçmeden önce, önemi nedeniyle daha önce de belirtilen amacı bir kez daha vurgulamak faydalı olacaktır. Bu veri madenciliği uygulamasında yapılmak istenen tüm öğrencilerin başarı notlarının tahmin edilmesi değildir ve verilen bilgilerle bunun tahmin edilememesi gerektiği de açıktır. Erişilmek istenen amaç, "Acaba veriler içinde öyle sıra dışı durumlar var mı ki böyle bir tahmin yapılabilsin ve yapılan bu tahmin yüksek güvenilirlikte olsun" sorusunun cevabını bulmaktır. Dolayısıyla oluşturulan modelde sadece yüksek olasılıklı tahminlerin göz önüne alınması, diğer düşük olasılık değerine sahip tahminlerle ilgilenilmemesi gerektiği açıktır. Hatta bu uygulama amacının dışına çıkılarak düşük olasılıklı tahminlerin de göz önüne alındığı genel bir başarı notu tahmin aracı olarak düşünülmemelidir. Bu durumda beklendiği üzere düşük olasılıklı tahminlerin çok sayıda olması nedeniyle modelin tüm doğruluğunun da düşük çıkacağı açıktır.

Bu bitirme ödevinde temel amaç, bu verilerden yüksek doğrulukta sıra dışı durumların varlığının tespitidir.

Aşağıda sonuç olarak verilen tüm ekran çıktılarında, bölüm ve gizlilik nedeniyle kodlanan öğretim görevlilerine göre not tahminleri "Prediction" sütununda, tahminlerin olma olasılığı ise "Probability" sütununda görülmektedir. Diğer sütunlar ise girdi verilerinde de bulunan ve karşılaştırma için buraya da dahil edilen sütunlardır.

Örneğin aşağıdaki tabloda görüldüğü üzere, MTO kodlu Meteoroloji Mühendisliği bölümünde kayıtlı bir öğrencinin 138 kodlu öğretim görevlisinden MAT101 dersini alması halinde %68.23 olasılıkla FF notunu alacağı gibi sıra dışı bir durum tespit edilmiştir.

Ayrıca yine bu tabloda daha yüksek olasılıklı MET kodlu Metalürji ve Malzeme Mühendisliği bölümünde kayıtlı bir öğrencinin 7 kodlu öğretim görevlisinden MAT101 dersini alması halinde %99.75 olasılıkla CB notunu alacağı gibi bir durum daha ortaya çıkmaktadır.



DMR$CASE_ID	INSTRUCTOR1	YEAR	TERM1	CODE	GRADE1	ID	DIVISION1	NUMBER	PREDICTION	PROBABILITY
5,693	7	1	2	101	CB	5,693	MET	60,990,228	CB	0.9975
5,671	7	1	2	101	BA	5,671	MET	40,990,216	CB	0.9975
5,696	7	1	2	101	CB	5,696	MET	60,990,248	CB	0.9975
5,697	7	1	2	101	FF	5,697	MET	60,990,256	CB	0.9975
5,683	7	1	2	101	AA	5,683	DEN	80,990,120	DC	0.997
5,691	7	1	2	101	DC	5,691	DEN	80,990,104	DC	0.997
5,269	91	9	1	101	CC	5,269	BLG	990,072,000	CC	0.997
4,614	138	6	1	101	FF	4,614	MTO	110,030,232	FF	0.6823
4,624	138	6	1	101	DC	4,624	MTO	110,030,208	FF	0.6823
4,629	138	6	1	101	FF	4,629	MTO	110,030,800	FF	0.6823
4,631	138	6	1	101	VF	4,631	MTO	110,020,232	FF	0.6823
4,632	138	6	1	101	FF	4,632	MTO	110,010,208	FF	0.6823
4,646	138	6	1	101	FF	4,646	MTO	110,020,216	FF	0.6823
11,880	138	6	1	101	FF	11,880	MTO	110,040,200	FF	0.6823
4,671	138	6	1	101	FF	4,671	MTO	110,030,216	FF	0.6823
4,685	138	6	1	101	FF	4,685	MTO	110,040,232	FF	0.6823
10,912	138	6	1	101	FF	10,912	MTO	110,010,200	FF	0.6823
10,921	138	6	1	101	FF	10,921	MTO	110,030,224	FF	0.6823
11,876	138	6	1	101	VF	11,876	MTO	110,020,232	FF	0.6823
4,657	138	6	1	101	FF	4,657	MTO	110,030,208	FF	0.6823
4,616	138	6	1	101	FF	4,616	MAD	50,030,052	FF	0.6689
11,874	138	6	1	101	CB	11,874	MAD	50,030,044	FF	0.6689
4,626	138	6	1	101	FF	4,626	MAD	50,020,024	FF	0.6689
4,633	138	6	1	101	FF	4,633	MAD	50,030,008	FF	0.6689
4,637	138	6	1	101	FF	4,637	MAD	50,030,032	FF	0.6689
4,640	138	6	1	101	FF	4,640	MAD	50,010,024	FF	0.6689
4,645	138	6	1	101	VF	4,645	MAD	50,010,008	FF	0.6689
4,647	138	6	1	101	FF	4,647	MAD	50,000,028	FF	0.6689
4,650	138	6	1	101	DD	4,650	MAD	50,020,016	FF	0.6689
4,653	138	6	1	101	FF	4,653	MAD	50,040,036	FF	0.6689
4,654	138	6	1	101	FF	4,654	MAD	50,020,012	FF	0.6689
4,656	138	6	1	101	FF	4,656	MAD	50,020,048	FF	0.6689
4,658	138	6	1	101	FF	4,658	MAD	50,020,036	FF	0.6689
4,659	138	6	1	101	FF	4,659	MAD	50,030,024	FF	0.6689

MAT101 için yukarıda anlatılan tüm adımlar her bir tablo için tekrarlanmış ve çeşitli sonuçlar elde edilmiştir. Bunlar içerisinde yukarıdaki MAT101 örneğinde olduğu gibi sıra dışı durumların gözlendiği bazı sonuçlar aşağıda verilmiştir;

MAT104103E tablosu, MAT103E ve MAT104E kodlu derslerin kayıtları birleştirilerek oluşturulmuştur. Bu tabloya uygulanan model ve uygulama sonuçlarına bakılarak benzer sıra dışı durumların varlığı gözlenmiştir.

Örneğin, 44 kodlu öğretim görevlisinden MAT103E dersini alıp, notu AA olan END kodlu Endüstri Mühendisliği öğrencisinin, aynı öğretim görevlisinden MAT104E kodlu dersi alması durumunda %70 olasılıkla AA alacağı tahmin edilmiştir.

Term1	2LecCo...	2Grad...	LecCod...	2ID1	2Te...	ID	2Instru...	Divisio...	2Year1	Number	Grade1	Instructor1	Year	PREDICTION	PROBABILI
1	103	FF	104	4,592	1	8,827	62	GID	6	60,030,800	FF	91	6	FF	0.9242
1	103	DD	104	5,541	3	8,789	133	GID	8	60,050,208	FF	130	9	FF	0.8154
3	103	VF	104	12,277	2	8,474	138	MET	8	60,020,156	FF	7	6	FF	0.7689
1	103	DD	104	12,157	3	7,523	133	IML	8	30,060,208	FF	130	9	FF	0.7529
1	103E	AA	104E	761	2	8,885	19	END	2	70,980,352	AA	46	3	AA	0.7363
2	103E	AA	104E	785	2	612	19	END	2	70,000,392	AA	38	3	AA	0.7341
2	103E	AA	104E	8,234	2	708	48	END	5	70,000,312	DC	20	3	AA	0.7341
2	103E	AA	104E	772	2	5,321	19	END	2	70,000,336	AA	44	3	AA	0.7341
2	103E	AA	104E	343	2	5,015	154	END	1	70,990,304	AA	44	2	AA	0.7089
2	103E	AA	104E	11,632	2	357	154	END	1	70,990,360	DC	38	2	AA	0.7089
2	103E	AA	104E	857	1	5,331	44	END	3	70,000,312	AA	44	3	AA	0.7032
2	103E	AA	104E	12,835	1	750	44	END	3	70,000,320	CB	19	3	AA	0.7032
2	103E	AA	104E	7,166	1	5,079	44	END	2	70,000,328	AA	44	2	AA	0.7032
2	103E	AA	104E	7,594	1	5,332	44	END	3	70,000,328	AA	44	3	AA	0.7032
2	103E	AA	104E	11,686	1	5,354	44	END	3	70,000,344	AA	44	3	AA	0.7032
2	103E	AA	104E	13,373	1	759	44	END	3	70,000,344	BA	19	3	AA	0.7032
2	103E	AA	104E	7,143	1	5,062	44	END	2	70,000,344	AA	44	2	AA	0.7032
2	103E	AA	104E	486	1	5,021	44	END	2	70,000,344	AA	44	2	AA	0.7032
2	103E	AA	104E	7,095	1	5,035	44	END	2	70,000,352	CB	44	2	AA	0.7032
2	103E	AA	104E	7,559	1	8,587	44	END	3	70,000,360	AA	19	3	AA	0.7032
2	103E	AA	104E	7,577	1	5,316	44	END	3	70,000,360	AA	44	3	AA	0.7032
2	103E	AA	104E	7,556	1	329	44	END	3	70,000,368	BB	38	2	AA	0.7032
2	103E	AA	104E	7,157	1	5,013	44	END	2	70,000,368	AA	44	2	AA	0.7032
2	103E	AA	104E	870	1	5,324	44	END	3	70,000,368	AA	44	3	AA	0.7032
2	103E	AA	104E	865	1	580	44	END	3	70,000,376	BA	44	3	AA	0.7032
2	103E	AA	104E	989	1	617	38	END	3	70,000,376	CB	38	3	AA	0.7032
2	103E	AA	104E	854	1	5,345	44	END	3	70,000,384	AA	44	3	AA	0.7032
2	103E	AA	104E	13,376	1	5,308	44	END	3	70,010,200	BA	44	3	AA	0.7032
2	103E	AA	104E	7,634	1	547	44	END	3	70,010,200	AA	44	3	AA	0.7032
2	103E	AA	104E	7,612	1	5,349	44	END	3	70,010,208	AA	44	3	AA	0.7032
2	103E	AA	104E	7,626	1	5,342	44	END	3	70,010,216	AA	44	3	AA	0.7032
2	103E	AA	104E	7,611	1	587	44	END	3	70,010,224	AA	44	3	AA	0.7032
2	103E	AA	104E	7,628	1	744	44	END	3	70,010,224	BA	19	3	AA	0.7032
2	103E	AA	104E	7,625	1	584	44	END	3	70,010,224	AA	44	3	AA	0.7032
2	103E	AA	104E	7,608	1	752	44	END	3	70,010,248	AA	19	3	AA	0.7032
2	103E	AA	104E	7,610	1	557	44	END	3	70,010,256	AA	44	3	AA	0.7032
2	103E	AA	104E	7,638	1	733	44	END	3	70,010,264	CB	19	3	AA	0.7032
2	103E	AA	104E	7,639	1	568	44	END	3	70,010,264	AA	44	3	AA	0.7032
2	103E	AA	104E	12,371	1	5,339	44	END	3	70,010,272	AA	44	3	AA	0.7032
2	103E	AA	104E	7,637	1	606	44	END	3	70,010,288	AA	44	3	AA	0.7032
2	103E	AA	104E	7,640	1	573	44	END	3	70,010,904	AA	44	3	AA	0.7032
2	103E	AA	104E	6,847	1	8,853	44	END	1	70,980,304	AA	44	1	AA	0.7032
2	103E	AA	104E	7,103	1	5,039	44	END	2	70,990,320	AA	44	2	AA	0.7032
2	103E	AA	104E	6,865	1	144	44	END	1	70,990,328	BB	44	1	AA	0.7032
2	103E	AA	104E	7,140	1	5,017	44	END	2	70,990,336	AA	44	2	AA	0.7032
2	103E	AA	104E	7,124	1	8,540	44	END	2	70,990,344	AA	44	2	AA	0.7032
2	103E	AA	104E	7,105	1	5,095	44	END	2	70,990,352	AA	44	2	AA	0.7032
2	103E	AA	104E	460	1	5,060	44	END	2	70,990,360	BA	44	2	AA	0.7032
2	103E	AA	104E	6,893	1	168	44	END	1	70,990,368	AA	44	1	AA	0.7032

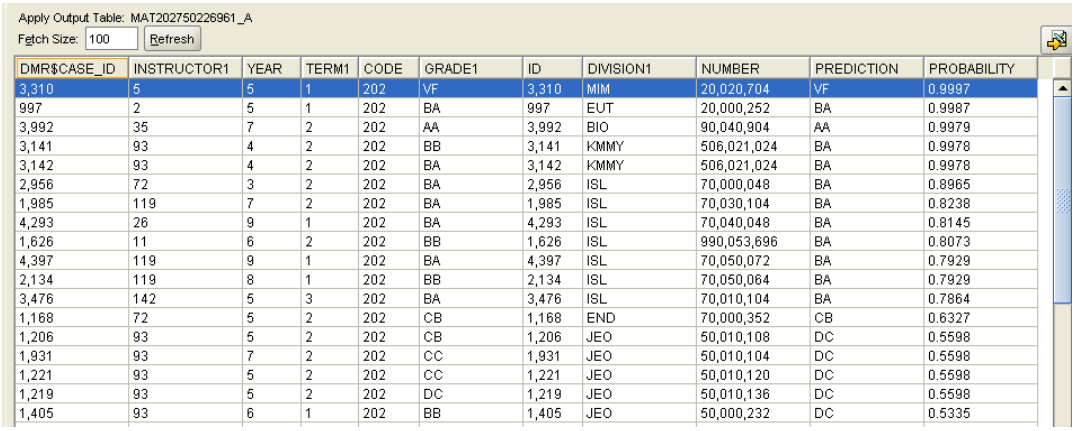
Örneğin, 39 kodlu öğretim görevlisinin verdiği MAT103E kodlu dersten FF notunu alan DEN kodlu Deniz Teknolojisi Mühendisliği öğrencisinin, aynı öğretim görevlisinden MAT104 kodlu dersi alması durumunda %100 olasılıkla VF notunu alacağı tahmin edilmiştir. Aynı şekilde 148 kodlu öğretim görevlisinin verdiği MAT103E kodlu dersten CC notunu alan DUI kodlu Deniz Ulaştırma İşletme Mühendisliği öğrencisinin, 30 kodlu öğretim görevlisinden MAT104 kodlu dersi alması durumunda %100 olasılıkla DD notunu alacağı tahmin edilmiştir.

Apply Output Table: MAT104103E251829084_A

Fetch Size: 100 Refresh

DM...	2Lec...	Term1	2Grade11	2ID1	2Ter...	ID	Lec...	2Instructor11	Division1	2Year1	N...	Grade1	Instructor1	Year	PREDICTION	PROBABIL
80,00	103E	2	FF	8,824	1	2,8	104	39	DEN	8	80,0	VF	39	7	VF	1
130,0	103E	2	CC	2,235	2	6,9	104	148	DUI	6	130,0	DD	30	7	DD	1
60,99	103E	2	CB	7,308	1	7,6	104	41	GID	2	60,9	CB	122	4	CB	0.9998
90,03	103E	3	BB	8,682	1	7,4	104	140	KIM	7	90,0	BB	55	8	BB	0.9998
30,99	103E	2	CC	7,272	1	3,9	104	41	MAK	2	30,9	DC	133	5	CB	0.9995
130,0	103E	2	DD	8,505	2	4,5	104	138	DUI	6	130,0	CB	126	7	CB	0.9992
30,98	103E	1	FF	602	1	7,5	104	139	MAK	2	30,9	VF	130	9	VF	0.9733
140,0	103E	2	FF	8,952	3	3,0	104	102	TEK	8	140,0	FF	39	8	FF	0.9645
10,04	103E	3	DD	13,482	3	2,8	104	102	CEV	8	10,0	BB	80	7	FF	0.9068
10,03	103E	2	FF	3,341	1	6,4	104	20	CEV	9	10,0	FF	122	6	FF	0.8998
30,03	103E	2	DD	8,537	3	3,1	104	128	MAK	6	30,0	FF	39	8	FF	0.8981

Benzer şekilde, MIM kodlu Mimarlık bölümünden bir öğrenci 5 kodlu öğretim görevlisinden MAT202 dersini alırsa %99.97 olasılıkla VF notunu alacağı tahmin edilmiştir. EUT kodlu Endüstri Ürünleri Tasarımı bölümünden bir öğrenci 2 kodlu öğretim görevlisinden MAT202 dersini alırsa %99.87 olasılıkla BA notunu alacağı tahmin edilmiştir

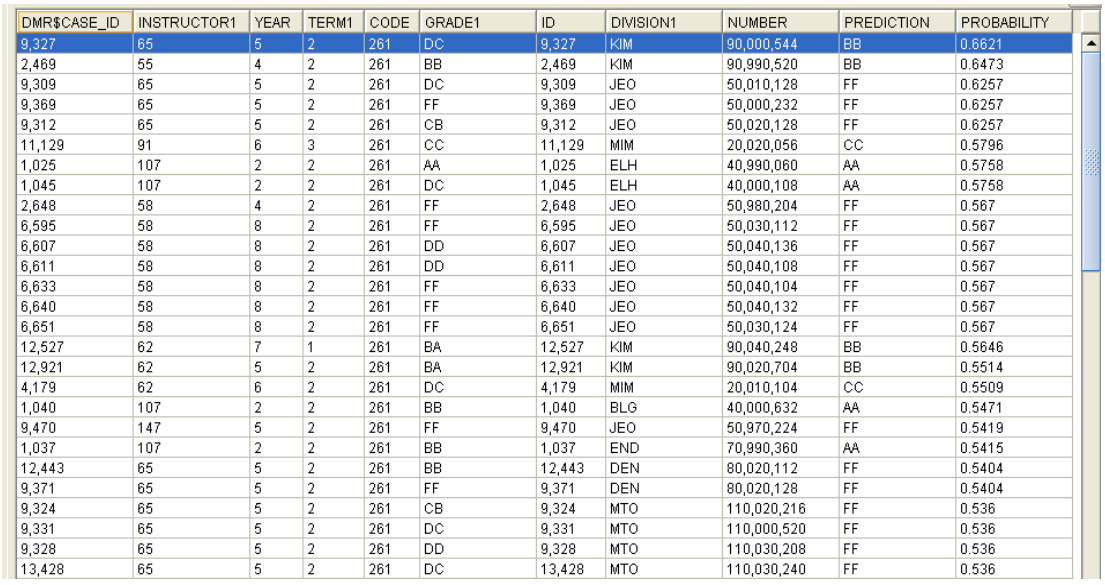


Apply Output Table: MAT202750226961_A

Fetch Size: 100 Refresh

DMR$CASE_ID	INSTRUCTOR1	YEAR	TERM1	CODE	GRADE1	ID	DIVISION1	NUMBER	PREDICTION	PROBABILITY
3,310	5	5	1	202	VF	3,310	MIM	20,020,704	VF	0.9997
997	2	5	1	202	BA	997	EUT	20,000,252	BA	0.9987
3,992	35	7	2	202	AA	3,992	BIO	90,040,904	AA	0.9979
3,141	93	4	2	202	BB	3,141	KMMY	506,021,024	BA	0.9978
3,142	93	4	2	202	BA	3,142	KMMY	506,021,024	BA	0.9978
2,956	72	3	2	202	BA	2,956	ISL	70,000,048	BA	0.8965
1,985	119	7	2	202	BA	1,985	ISL	70,030,104	BA	0.8238
4,293	26	9	1	202	BA	4,293	ISL	70,040,048	BA	0.8145
1,626	11	6	2	202	BB	1,626	ISL	990,053,696	BA	0.8073
4,397	119	9	1	202	BA	4,397	ISL	70,050,072	BA	0.7929
2,134	119	8	1	202	BB	2,134	ISL	70,050,064	BA	0.7929
3,476	142	5	3	202	BA	3,476	ISL	70,010,104	BA	0.7864
1,168	72	5	2	202	CB	1,168	END	70,000,352	CB	0.6327
1,206	93	5	2	202	CB	1,206	JEO	50,010,108	DC	0.5598
1,931	93	7	2	202	CC	1,931	JEO	50,010,104	DC	0.5598
1,221	93	5	2	202	CC	1,221	JEO	50,010,120	DC	0.5598
1,219	93	5	2	202	DC	1,219	JEO	50,010,136	DC	0.5598
1,405	93	6	1	202	BB	1,405	JEO	50,000,232	DC	0.5335

Nitelikli bir öğretim sisteminde yukarıda varlığı gözlenen sıra dışı durumların mümkün olduğunca az ve bu girdilerle tahmin olasılıklarının küçük olması beklenmelidir. Ancak öğrencinin devam durumu, ödev ve kısa sınav notları gibi yeni girdilerin eklenmesiyle tahminlerin doğruluğu yükselebilir ve bu durumda yeni öğrencilerin başarı notlarının tahmin edilmesi sağlanabilir. Yukarıdaki girdilerle genel beklentiye uygun olarak, örneğin MAT261 dersinde hiçbir sıra dışı durum gözlenmemiştir.



DMR$CASE_ID	INSTRUCTOR1	YEAR	TERM1	CODE	GRADE1	ID	DIVISION1	NUMBER	PREDICTION	PROBABILITY
9,327	65	5	2	261	DC	9,327	KIM	90,000,544	BB	0.6621
2,469	55	4	2	261	BB	2,469	KIM	90,990,520	BB	0.6473
9,309	65	5	2	261	DC	9,309	JEO	50,010,128	FF	0.6257
9,369	65	5	2	261	FF	9,369	JEO	50,000,232	FF	0.6257
9,312	65	5	2	261	CB	9,312	JEO	50,020,128	FF	0.6257
11,129	91	6	3	261	CC	11,129	MIM	20,020,056	CC	0.5796
1,025	107	2	2	261	AA	1,025	ELH	40,990,060	AA	0.5758
1,045	107	2	2	261	DC	1,045	ELH	40,000,108	AA	0.5758
2,648	58	4	2	261	FF	2,648	JEO	50,980,204	FF	0.567
6,595	58	8	2	261	FF	6,595	JEO	50,030,112	FF	0.567
6,607	58	8	2	261	DD	6,607	JEO	50,040,136	FF	0.567
6,611	58	8	2	261	DD	6,611	JEO	50,040,108	FF	0.567
6,633	58	8	2	261	FF	6,633	JEO	50,040,104	FF	0.567
6,640	58	8	2	261	FF	6,640	JEO	50,040,132	FF	0.567
6,651	58	8	2	261	FF	6,651	JEO	50,030,124	FF	0.567
12,527	62	7	1	261	BA	12,527	KIM	90,040,248	BB	0.5646
12,921	62	5	2	261	BA	12,921	KIM	90,020,704	BB	0.5514
4,179	62	6	2	261	DC	4,179	MIM	20,010,104	CC	0.5509
1,040	107	2	2	261	BB	1,040	BLG	40,000,632	AA	0.5471
9,470	147	5	2	261	FF	9,470	JEO	50,970,224	FF	0.5419
1,037	107	2	2	261	BB	1,037	END	70,990,360	AA	0.5415
12,443	65	5	2	261	BB	12,443	DEN	80,020,112	FF	0.5404
9,371	65	5	2	261	FF	9,371	DEN	80,020,128	FF	0.5404
9,324	65	5	2	261	CB	9,324	MTO	110,020,216	FF	0.536
9,331	65	5	2	261	DC	9,331	MTO	110,000,520	FF	0.536
9,328	65	5	2	261	DD	9,328	MTO	110,030,208	FF	0.536
13,428	65	5	2	261	DC	13,428	MTO	110,030,240	FF	0.536

KAYNAKLAR

[1] **Dilly R.,** 1995. *Data Mining, An introduction Student Notes*, http://www.pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_1.html.

[2] **Frawley W. J., Shapiro G. P., Matheus C. J.**, 1992. *Discovery in Databases: An Overview*, AI Magazine, **13**-3, 57-70.

[3] **Berry J. A., Linoff G.**, 1997. *Data mining techniques for marketing, sales and customer support*, John Wiley & Sons Inc., New York.

[4] **Holshemier M., Siebes A.**, 1994. *Data mining* http://www.pcc.qub.ac.uk.

[5] **Swift R.**, 2001. *Accelerating customer relationship*, Prentice Hall PTR, NJ.

[6] **Christen, P.,** 2005. *A very short introduction to data mining, Lecture Notes*, http://datamining.anu.edu.au.

[7] **Simoudis E., Fayyad U., Han J.**, 1996, *Proceeding of the second international conference on knowledge discovery and data mining*, AAAI Press.

[8] **Baykasoğlu A.,** 2005. *Veri madenciliği ve çimento sektörüne bir uygulama, Akademik Bilişim Konferansı*, pp.82-83, Gaziantep Üniversitesi, Gaziantep.

[9] **Sander J.,** 2002. *Principles of Knowledge Discovery in Databases, Computing Science 690 lecture notes*, http://www.cs.ualberta.ca /~joerg/courses/cmput690/

[10] http://en.wikipedia.org/wiki/Oracle_Corporation

[11] **Çubukçu F.,** *İlişkisel veritabanları ve Oracle' ın temelleri*, http://www.farukcubukcu.com.