İSTANBUL TEKNİK ÜNİVERSİTESİ ★FEN EDEBİYAT FAKÜLTESİ
MATEMATİK MÜHENDİSLİĞİ PROGRAMI

ORACLE DATA MINER İLE BORSA ALANINDA TEKNİK ANALİZDE KULLANILAN GÖSTERGELER (INDICATORS) ÜZERİNE BİR VERİ MADENCİLİĞİ UYGULAMASI

BİTİRME ÖDEVİ

Tolga KAPAN 090070077

Birgül AYAZ 090080023

Tez Danışmanı: Yar. Doç. Dr. Ahmet KIRIŞ

MAYIS 2012

ÖNSÖZ

Bu çalışmayı hazırlamamızda bize yol gösteren, bizden yardımını ve desteğini hiçbir zaman esirgemeyen Sayın Hocamız Yrd. Doç. Dr. Ahmet KIRIŞ' a, geçmiş yılların hisse senedi fiyat verilerine ulaşmamızda bize yardımcı olan FOREX şirketi çalışanlarına, arkadaşlarımıza ve hayatımız boyunca bize sevgi, güven ve her türlü desteği veren ailelerimize en içten teşekkürlerimizi sunarız.

Mayıs, 2012

Tolga KAPAN
Birgül AYAZ

MAYIS 2012

İÇİNDEKİLER

Sayfa

Error! No table of contents entries found.

ÖZET

Borsada alım-satım kararı verilirken dikkate alınan teknik analiz uygulamalarından göstergelerin (indicators), kar elde etmek için alım ya da satım durumu ile ilişkisi üzerine bir veri madenciliği uygulaması yapılmıştır. Bu uygulamada göstergeler yardımı ile alım ya da satım kararlarının önceden tahmin edilmesi ve daha kullanışlı bir gösterge oluşturulması amaçlanmaktadır. Alım satım kararları önceden tahmin edilerek veya daha kullanışlı bir gösterge oluşturularak kar elde edilebilir veya zarar en aza indirgenebilir.

Bu amaçla FOREX şirketinden geçmiş yıllara ait hisse senedi verileri alınmış, bu veriler düzenlenerek Excel dosyası oluşturulmuştur. Bu Excel dosyasındaki verilerle göstergelerin formülleri oluşturularak gösterge değerleri elde edilmiştir. Gösterge değerleri ile hisse senetleri fiyatlarının günlük kapanış değerleri kullanılarak tablolar oluşturulmuş ve hisse senetleri alınmalı, satılmalı ya da işlem yapılmamalı olarak sınıflandırılıp problem veri madenciliği sınıflandırma modeline uygun hale getirilmiştir. Bu problemi çözmek için sınıflandırma modelinin algoritmalarından olan Naivé Bayes ve Karar Ağaçları (Decision Tree) yöntemleri kullanılmıştır.

Veri madenciliği uygulamasını gerçekleştirmek için Oracle veritabanının 11g sürümü ve üzerine Oracle Data Miner paket programı yüklenmiştir. Daha sonra oluşturulan tablolar veritabanına aktarılmış ve gerekli modeller oluşturulmuştur. Bu modeller yardımıyla tahminler elde edilmiş ve sonuçlar yorumlanmıştır.

1. GİRİŞ

Gelişen teknoloji ve tüm Dünya ülkelerinde bilgisayarın yaygın kullanılması elektronik ortamda saklanan veri miktarında büyük bir artış meydana getirmiştir. Bilgi miktarının her 20 ayda bir iki katına çıkması veritabanı sayısında hızlı bir artışa neden olmaktadır. Birçok farklı bilim dallarından toplanan veriler, hava tahmini simülasyonu, sanayi faaliyet testleri, süpermarket alışverişi, banka kartları kullanımı, telefon aramaları gibi veriler, daha büyük veritabanlarında kayıt altına alınmaktadır. Veri tabanında kayıt altına alınan bu verilerin tahmin edilemeyecek boyutta faydaları vardır. Bu verilerden elde edilecek bilgiler doğrultusunda finansal alanda gelecek için stratejiler belirlenebilir.

Veri madenciliği uygulaması finansal alanda kullanılarak, hisse senedi alım satımında göz önünde bulundurulan göstergelerin geçmiş yılların hisse senedi fiyatlarına göre verdiği kararlar ve vermesi gereken kararlar karşılaştırılmış ve model oluşturulmuştur. Bu model ile hisse senedi alım satım kararları test verileri üzerinden tahmin edilmeye çalışılmıştır. Bitirme projesi kapsamında da hisse senetlerinin alım satım kararları değerlendirilirken kullanılan göstergeler ile stratejiler geliştirilmesi ve yeni göstergeler elde edilmesi amaçlanmıştır. Alım-satım kararı verilirken göstergelerin nasıl elde edildiği ve veri madenciliği uygulamasına neden gereksinim duyulduğu bir süreç dâhilinde aşağıda ifade edilmektedir.

2. GÖSTERGELER (INDICATORS)

Teknik analiz yapılırken göstergelerin verdiği alım-satım sinyalleri göz önünde bulundurulmaktadır. Karar verme aşamasında kararların bu durumlarla uyum içinde olması önemlidir. Oracle Data Miner ile göstergelerin geçmişte verdiği alım-satım kararlarının gerçek zamanlı sonuçlarla uyumluluğu değerlendirilip gelecekte verilecek kararların tahmin edilmesi amaçlanmıştır.

Hisse senedi alım-satım kararı vermek için önemli uygulamalardan birisi teknik analizdir. Teknik analiz yaparken, trendleri belirlemek ve oluşumlarını çözümlemek zordur. Ayrıca bu tespitleri yaparken, objektif olmak ve trend dönüşümlerinin süresini ve büyüklüğünü tespit etmek güçtür. Bu güçlüğü aşmak, trend değişimlerini oluştukları sırada belirlemek amacıyla fiyat ve işlem hacmi verileri esas alınarak çeşitli hesaplamalar yapılmaktadır. Grafik örneklere yardımcı olmak amacıyla fiyat ve işlem hacmi verilerinden türetilen bu kurallara "göstergeler" adı verilmektedir.

Bitirme projesi kapsamında alım-satım kararlarında kullanılan göstergeler ve açıklamaları aşağıdaki gibidir.

2.1 Toplama- Dağıtım Endeksi (Accumulation- Distribution Index, ADI)

Toplama- Dağıtım Endeksi işlem hacmi ve fiyatlardaki hareketleri birleştirerek fiyatların trendinin sürüp sürmeyeceği konusunda fikir veren bir göstergedir. Her günkü işlem hacminin belli bir oranı bir önceki günün kümülatif toplamına eklenir veya çıkarılır. Formülü aşağıdaki gibidir:

$$ADI_i = \sum_{i=1}^{\infty} \left[\left(\frac{(C_i - L_i) - (H_i - C_i)}{(H_i - L_i)} \right) \times V_i \right] \tag{2.1}$$

burada,

C= Kapanış fiyatı, (Close)

H= Günün en yüksek fiyatı, (High)

L= Günün en düşük fiyatı, (Low)

V= İşlem hacmi, (Volume)

ADI= Accumulation- Distribution Index

olarak verilmektedir. Buradaki kısaltmalar, (C,H,L,V) çalışmanın geri kalan bölümlerinde de aynı ifadeleri tanımlamaktadır.

Eğer kapanış günün en yüksek fiyatına yakınsa günlük işlem hacminin hesaplanan oranı kadar bir miktar bir önceki günün ADI değerine eklenir. Kapanış günün en düşük fiyatına daha yakınsa işlem hacminin hesaplanan oranı kadar bir miktar bir önceki günün ADI değerinden çıkartılır. Gösterge bir dip yapıp yukarı dönerek yükselişe başlarsa alım, tepe yaptıktan sonra düşmeye başlarsa satım kararı verilir.

2.2 Mal Kanal Endeksi (Commodity Channel Index, CCI)

Mal Kanal Endeksi(CCI)senet fiyatının kendi istatistiksel ortalamasından aşağı veya yukarı ne kadar saptığını gösterir. "+100" ile "–100" limit değerlerdir. "+100" ün üzerine çıkması senedin aşırı alım seviyesinde olduğu,-100'ünaltında olması ise aşırı satım bölgesinde olduğunu gösterir. Mal Kanal Endeksi, Donald R.Lambert tarafından geliştirilen bir fiyat momentum göstergesidir. Piyasa trendinin başlangıç ve bitişini bulmak için dizayn edilmiştir. Formülü aşağıdaki gibidir:

$$A_i = \frac{(H_i - L_i + C_i)}{3} \quad B_i = \frac{\left(\sum_{i=1}^{14} \frac{(H_i + L_i + C_i)}{3}\right)}{14}$$
$$D_i = \frac{\left(\sum_{i=1}^{14} | A_i - B_i |\right)}{14}, \quad CCI_i = \frac{(A_i - B_i)}{(D_i - 0.015)} \tag{2.2}$$

Burada CCI= Mal Kanal Endeksidir. Bu endeks "–100" seviyesinin altından üstüne çıkıldığında alım, "+100" seviyesinin üstünden altına düşüldüğünde satım yapılmalıdır.

2.3 Hareketli Ortalamaların Birleşmesi-Ayrılması Göstergesi(Moving Average Convergence Divergence, MACD)

MACD (kesiksiz çizgi) göstergesi Tetik Çizgisini(kesikli çizgi) aşağıdan yukarı doğru kestiğinde alım, yukarıdan aşağı doğru kestiğinde satım sayesinde karar vermeyi kolaylaştırır. 12 günlük üssel hareketli ortalamadan 26 günlük üssel hareketli ortalama çıkartılınca MACD bulunur. Bu MACD formülünün ayrıca 9 günlük üssel hareketli ortalaması kesikli çizgilerle çizdirilir ve bu kesikli çizgiye tetik çizgisi(trigger) adı verilir. MACD göstergesi tetik çizgisini aşağıdan yukarı

Kestiği zaman alım, yukarıdan aşağı kestiği zamanda satım kararı verilir.

$$
\begin{aligned}
&A_i = (\alpha \times C_i) + (1-\alpha) \times A_{i-1}, \quad n = 12 \\
&B_i = (\alpha \times C_i) + (1-\alpha) \times B_{i-1}, \quad n = 26 \\
&MACD_i = A_i - B_i \\
&T_i = (\alpha \times MACD_i) + (1-\alpha) \times T_{i-1}, \quad n = 9 \\
&\alpha = \frac{2}{(n+1)}
\end{aligned}
\tag{2.3}
$$

Burada

n = Üssel hareketli ortalama için seçilen gün sayısı

MACD= Hareketli Ortalamaların Birleşmesi-Ayrılması Göstergesi

T= Tetik Çizgisi (trigger)

MACD (kesiksiz çizgi) göstergesi olarak alınmıştır.

Tetik Çizgisini(kesikli çizgi) aşağıdan yukarı doğru kestiğinde alım, yukarıdan aşağı doğru kestiğinde satım yapılır.

2.4 Aroon Osilatörü (Aroon Oscillator)

Aroon Osilatörü, Tushar Chande tarafından geliştirilmiştir. Aroon, şafağın ilk ışıkları anlamına gelen Sanskritçe bir kelimedir. Trend yapan piyasadan yatay piyasaya geçişte, senet fiyatındaki değişimleri önceden anlamaya olanak verir.

Gösterge iki göstergeden oluşmaktadır. İlki, belirlenen periyottaki en yakın en yüksek fiyatın görülmesinden bu yana geçen periyot sayısını gösteren Yukarı Aroon (Aroon Up) göstergesi; ikincisi ise, belirlenen periyottaki en yakın en düşük fiyatın görülmesinden bu yana geçen periyot sayısını gösteren Aşağı Aroon (Aroon Down) göstergesidir. Aroon Oscillator' ü ise Aroon Up ve Aroon Down göstergelerinin

arasında fark olarak tanımlanabilecek tek bir çizgi şeklindedir. Aroon Up ve Aroon Down göstergeleri "0" ile "+100" arasında dalgalanırken, Aroon Oscillator' ü "-100" ile "+100" arasında dalgalanmaktadır.

$$AroonUp_i = \left[\frac{\left(n - \left(\max\left(H_i\right)\right)\right)}{n} \right] \times 100$$

$$AroonDown_i = \left[\frac{\left(n - \left(\min\left(L_i\right)\right)\right)}{n} \right] \times 100 \qquad (2.4)$$

$$AroonOsc_i = AroonUp_i - AroonDown_i$$

Burada,

n=Hesaplama için seçilen gün sayısı

$\max\left(H_i\right)$ = En yüksek güniçi yüksek değerin gerçekleştiği günden bugüne süre

$\min\left(L_i\right)$=En düşük güniçi düşük değerin gerçekleştiği günden bugüne geçen süre olarak alınmıştır.

Aroon Up ve Aroon Down çizgileri, üst üste beraber hareket ediyorlarsa, yatay piyasaya işarettir. Aroon Up çizgisi, Aroon Down çizgisini aşağıdan yukarı aşarak kestiğinde alım ve yukardan aşağı kestiğinde satım sinyali vermektedir.

2.5 Fiyat Osilatörü (Price Oscillator, PO)

Fiyat Osilatörü (Price Oscillator), hisse senedi fiyatının iki hareketli ortalaması arasındaki farkı gösterir. Hareketli ortalamalar arasındaki fark rakamsal veya yüzdesel olarak belirtilebilir. Fiyat osilatörü hesaplanırken çeşitli hareketli ortalamalar kullanılabilir. Formülü aşağıdaki gibidir:

$$KHVO_I = \frac{\sum_{i=1}^{n} C_i}{n} \qquad UVHO_i = \frac{\sum_{i=1}^{n} C_I}{n}$$

$$POSC_I = \left(\frac{KHVO_I - UVHO_i}{UVHO_i} \right) \times 100 \qquad (2.5)$$

Burada POSC=Fiyat Osilatörü'dür.

Fiyat Osilatörü göstergesinin "0" seviyesinin üzerine çıkması alım, altına düşmesi satım sinyali olarak kullanılabilir.

2.6 Bollinger Bantları (Bollinger Bands)

John Bollinger tarafından geliştirilmiştir. Fiyatlar bu bandın içinde hareket etme eğiliminde olup bandın daralması gelecekte muhtemel bir fiyat değişimine işaret etmektedir. Fiyatlardaki aşağı veya yukarı hareketlilik arttıkça bandın genişliği artarken, yatay ve durgun fiyat hareketi dönemlerinde band daralacaktır. Formülü aşağıdaki gibidir:

$$\sigma_i = \sqrt{\frac{\sum_{i=1}^{n}\left(C_I - Y_I\right)^2}{n}} \qquad Y_I = \frac{\sum_{i=1}^{n} C_i}{n}$$
$$Üstband_i = Y_i + \left(d \times \sigma_i\right) \tag{2.6}$$
$$Ortaband_i = Y_i$$
$$Altband_i = Y_i - \left(d \times \sigma_i\right)$$

Burada,

σ =Standart Sapma

Y=Basit Hareketli Ortalama

d= Standart Sapma Çarpanı olarak alınmıştır.

Fiyatlar üst banda değdiğinde kısa vadeli satım, alt banda değdiğinde kısa vadeli alım yapılabilir. Orta bandı aşarak yükselişini sürdüren hisse senedini tutmaya devam etmek doğru bir yoldur. Orta bandı kırarak düşüşünü sürdüren hisse senedinden alım yapmak için bir süre daha beklenmelidir.

2.7 Chaikin Para Akım Göstergesi (Chaikin Money Flow, CMF)

Chaikin Para Akım Göstergesi, Chaikin Toplama- Dağıtım Osilatör' ü temel alır. Toplama- Dağıtım' ı (Accumulation- Distribution) açıklarsak, sırasıyla, eğer hisse senedinin o günkü kapanış fiyatı o günün orta noktasının [(en yüksek + en düşük) /2] üzerinde olursa, o gün toplanma günüdür. Tam tersi, o günkü kapanış fiyatı o günün orta noktasının altında olursa, o gün dağılım günüdür. Chaikin Para Akım Göstergesi (CMF), 21 günlük A/D değerlerinin toplamının 21 günlük işlem miktarı toplamına bölünmesiyle elde edilir. Formülü aşağıdaki gibidir:

$$A_i = \sum_{i=1}^{21}\left[\left(\frac{\left(C_i - L_i\right)-\left(H_i - C_i\right)}{\left(H_i - L_i\right)}\right)\times V_i\right]$$
$$B_i = \sum_{i=1}^{21} V_i \qquad CMF_i = \frac{A_i}{B_i} \tag{2.7}$$

CMF göstergesi "0" çizgisinin üzerine çıktığında alım, altına düştüğünde satım yapılır.

2.8 Chande Momentum Osilatörü (Chande Momentum Oscillator, CMO)

Chande Momentum Osilatörü (CMO), Tushar Chande tarafından geliştirilmiştir. "saf momentum" u elde etmeyi amaçlamaktadır.

"+50" seviyesinin üzeri aşırı alım, "-50" seviyesinin aşağısı aşırı satım bölgesidir. Formülü aşağıdaki gibidir.

$$CU_i = \begin{cases} C_i - C_{i-1} & eger \quad C_i - C_{i-1} > 0 \\ 0 & eger \quad C_i - C_{i-1} < 0 \end{cases} \qquad CMU_i = \sum_{i=1}^{14} CU_i$$
$$CD_i = \begin{cases} 0 & eger \quad C_i - C_{i-1} > 0 \\ C_i - C_{I-1} & eger \quad C_i - C_{i-1} < 0 \end{cases} \qquad CMD_i = \sum_{i=1}^{14} CU_i \tag{2.8}$$
$$CMO_i = \left[\frac{\left(CMU_i - CMD_i\right)}{\left(CMU_I + CMD_I\right)}\right]\times 100$$

Burada,

CU=Yükselen Günlerin Değeri; CD=Düşen Günlerin Değeri;

CMU=Yükselen Günlerin Kümülatif Toplamı;

CMD=Düşen Günlerin Kümülatif Toplamı

CMO=Chande Momentum Osilatörü olarak alınmıştır.

"-50" seviyesinden alım, "+50" seviyesinden satım yapılır. Gösterge "0" çizgisinin üzerine çıktığında kesin alım, altına düştüğünde de kesin satım yapılmalıdır.

2.9 Doğrusal Regresyon Göstergesi (Linear Regression Indicator)

Doğrusal Regresyon Göstergesi, belirlenen süre boyunca hisse senedi fiyat trendini temel alır. En küçük kareler yöntemini kullanarak doğrusal (lineer) regresyon trend çizgisi hesaplanır. Örneğin 14 günlük çizilen Doğrusal Regresyon trend çizgisinin son gün aldığı değer, 14 günlük Doğrusal Regresyon Göstergesi (Linear Regression Indicator)' nin değerine eşittir.

Fiyatlar göstergenin üzerine çıktığında alım, altına düştüğünde satım yapılır.

2.10 Momentum Göstergesi

Momentum göstergesi, belirli bir süre içinde fiyatların yüzdesel değişimini gösterir. Referans çizgisi "100" dür. Momentum belirli bir zaman aralığı için sürekli fiyat değişimlerinin aldığı değerle ölçülür. Formülü aşağıdaki gibidir:

$$MOM_I = \left(\frac{C_i}{C_{i-n}} \right) \times 100 \quad n = 12 \tag{2.9}$$

Burada MOM= Momentum Göstergesi'dir. "100" referans çizgisinin üzerine çıktığında alım, altına düştüğünde satım yapılmalıdır.

2.11 Fiyat Değişim Oranı (Price Rate of Change, P-ROC)

Fiyat Değişim Oranı (Price Rate of Change, P-ROC) göstergesi fiyatların belirli bir süreye göre yüzdesel olarak ne kadar arttığını veya azaldığını gösterir. Momentum indikatörüne benzemektedir, şu farkla, P-ROC yüzdesel değişimi gösterir. Formülü aşağıdaki gibidir:

$$PROC_I = \left(\frac{C_i - C_{i-n}}{C_{i-n}} \right) \times 100 \tag{2.10}$$

Burada PROC=Fiyat Değişim Oranı'dır. "0" seviyesinin üzerine çıktığı zaman alım, altına düştüğü zaman satım yapılır.

2.12 Göreceli Momentum Endeksi (Relative momentum Index, RMI)

Göreceli Momentum Endeksi, Roger Altman tarafından geliştirilmiştir. Göreceli Güç Endeksi (Relative Strength Index, RSI)'ın bir varyasyonudur. RSI gibi kapanıştan kapanışa artan ve azalan günleri hesaplayacağına, RMI kapanıştan x gün

önceki kapanışa göre artan ve azalan günleri hesaplar. İndikatörün iki parametresi bulunmaktadır. Süre (Time Periods) ve Momentum. Süresi 14 ve momentumu 1 olan RMI, 14 günlük RSI'a eşittir. Formülü aşağıdaki gibidir:

$$A_i = \begin{cases} C_i - C_{i-x} & C_i > C_{i-1} \\ 0 & C_i < C_{i-1} \end{cases} \qquad B_i = \begin{cases} 0 & C_i > C_{i-1} \\ |C_i - C_{i-x} & C_i > C_{i-1} \end{cases}$$

$$D1_i = \frac{\sum_{i=1}^{m} A_i}{m} \qquad D2_i = \frac{\sum_{i=1}^{m} B_i}{m} \qquad (2.11)$$

$$RM_i = \begin{cases} 0 & ,D2_i = 0 \\ \dfrac{D1_i}{D2_i} & ,D2_i \neq 0 \end{cases} \qquad RMI_i = \left(\frac{RM_i}{1 + RM_i} \right) \times 100$$

Burada RMI= Göreceli Momentum Endeksi'dir.

2.13 Göreceli Güç Endeksi (Relative Strength Index, RSI)

Göreceli Güç Endeksi (Relative Strength Index, RSI) göstergesi, iki farklı hisse senedini karşılaştırmaz, tek bir hisse senedinin içsel gücünü ölçmektedir. RSI, "0" ile "100" aralığında değerler alan ve fiyat takip eden bir göstergedir. Welles Wilder tarafından geliştirilmiştir. Wilder, 14 günlük RSI hesaplamasını önermektedir. Ayrıca 9 ve 25 günlük RSI'larda kullanılmaktadır. Hesaplamada kullanılan periyot azaldıkça, indikatördeki değişkenlik artmaktadır. Genellikle "0" ile "30" aralığında dip, "70" ile "100" aralığında tepe yapmaktadır.

Genel olarak RSI grafik oluşumlarını tanımlamakta yararlı olması yanı sıra hisse senedi fiyatının kısa dönem için pahalı ya da ucuz olduğu seviyeyi belirlemekte kullanılır. Ayrıca klasik grafik metot ve şekilleri ile görülmesi zor olan destek ve direnç seviyelerini çok defa daha önce ve daha belirgin olarak gösterir. En popüler yatay piyasa (anti-trend) göstergelerinden olup aşırı alım-aşırı satım bölgelerinin tanımlanması, tepe ve diplerin tahmin edilmesinde kullanılır. Formülü aşağıdaki gibidir:

$$PD_i = \begin{cases} C_i - C_{i-1}, & C_i > C_{i-1} \\ 0 \quad , & C_i \le C_{i-1} \end{cases} \qquad ND_i = \begin{cases} 0 \quad , & C_i > C_{i-1} \\ \left|C_i - C_{i-1}\right| \quad , & C_i \le C_{i-1} \end{cases}$$

$$OPD_i = \frac{\sum_{i=1}^{n} PD_i}{n} \qquad OND_i = \frac{\sum_{i=1}^{n} ND_i}{n} \qquad RSI_i = 100 - \left[\frac{100}{1 + \left(\frac{OPD_i}{OND_i} \right)} \right] \tag{2.12}$$

Burada RSI=Göreceli Güç Endeksi'dir.

2.14 Stokastik Momentum Endeksi (Stochastic Momentum Index, SMI)

Stokastik Momentum Endeksi (Stochastic Momentum Index, SMI) William Blau tarafından geliştirilmiştir. Stokastik Osilatör, hisse senedi kapanış fiyatıyla belirlenen süre içindeki fiyat aralığını karşılaştırırken; SMI, hisse senedi kapanış fiyatıyla belirlenen süre içindeki fiyat aralığının orta noktasını karşılaştırmaktadır. SMI, "-100" ile "+100" aralığında hareket etmektedir. Kapanış, fiyat aralığının orta noktasından büyükse, SMI pozitiftir. En geniş artı değer kapanış değeri yükseklerin en yükseğine eşit olduğunda gerçekleşir. Tersi durumda, SMI negatiftir. En geniş düşük değer kapanış değeri düşüklerin en düşüğüne eşit olduğu zaman gerçekleşir. Formülü aşağıdaki gibidir:

$$\begin{aligned} HHV_i &= \max\{H_i, H_{i-1}, H_{i-2}, ...H_{i-n}\} \\ LLV_i &= \min\{L_i, L_{i-1}, L_{i-2}, ...L_{i-n}\} \\ UZ_i &= C_i - MP_i \\ A_i &= (\alpha \times UZ_i) + \left[(1-\alpha) \times A_{i-1}\right] \\ B_i &= (\alpha \times A_i) + \left[(1-\alpha) \times B_{i-1}\right] \\ D_i &= HHV_i - LLV_i \\ E_i &= (\alpha \times D_i) + \left[(1-\alpha) \times E_{i-1}\right] \\ F_i &= (\alpha \times E_i) + \left[(1-\alpha) \times F_{i-1}\right] \\ G_i &= \frac{F_i}{2} \qquad SMI_i = \left(\frac{B_i}{G_i}\right) \times 100 \qquad \alpha = \frac{2}{(n+1)} \end{aligned} \tag{2.13}$$

Burada n=Üssel hareketli ortalama için seçilen gün sayısıdır.

2.15 Stokastik Osilatörü (Stochastic Oscillator, SO)

Stokastik Osilatörü, hisse senedinin kapanış fiyatını belirlenen süre içindeki fiyat

aralığı ile karşılaştırmaktadır. Stokastik Osilatörü iki eğri ile gösterilir. Kesiksiz bir çizgi olarak gösterilen ana eğri %K olarak adlandırılırken, noktalı çizgilerle gösterilen %D eğrisi, %K'nın hareketli ortalamasıdır. Alım satım kararları iki şekilde verilebilir:

-Osilatör, "20" seviyesinin altına düşüp sonra üzerine çıktığı zaman alım, "80" seviyesinin üzerine çıkıp sonra altına düştüğü zaman satım kararı verilir.

-Kimi zaman gösterge "20" ve "80" seviyelerine ulaşmadan ters yönlere hareket edebilmektedir. Bu durumlarda %K eğrisi, hareketli ortalaması olan %D eğrisini aşağıdan yukarı kestiği zaman alım, yukarıdan aşağı kestiği zaman da satım kararı verilir.

İndikatörün sinyal seviyesini yukarı veya aşağı kesmesine göre alım veya satım yapılabileceği gibi, "80" nin üzerinden aşağı döndüğünde satım ve "20"'nin altından yukarı döndüğünde de alım yapılmalıdır. Formülü aşağıdaki gibidir:

$$HHV_i = \max\{H_i, H_{i-1}, H_{i-2}, ...H_{i-n}\}$$
$$LLV_i = \min\{L_i, L_{i-1}, L_{i-2}, ...L_{i-n}\}$$
$$A_i = C_i - LLV_i \qquad B_i = HHV_i - LLV_i$$
$$SUM1_i = \sum_{i=1}^{m} A_i \qquad SUM2_i = \sum_{i=1}^{m} B_i \qquad (2.14)$$
$$STO_i = \left(\frac{SUM1_i}{SUM2_i}\right) \times 100 \qquad \%D_i = \frac{\sum_{i=1}^{k} STO_i}{k}$$

Burada STO Stokastik Osilatör'dür.

2.16 Wilder'ın Düzeltme Göstergesi (Wilder's Smoothing, WS)

Wilder'ın Düzeltme göstergesi, adından da anlaşılacağı üzere, Welles Wilder tarafından geliştirilmiştir. Wilder, bu düzeltme indikatörünü diğer çalışmalarının bir parçası gibi kullanmaktadır. İndikatör, temel olarak üssel (exponential) metoda benzeyen bir hareketli ortalama tipidir çünkü serideki bütün tarihi verinin gittikçe azalan daha küçük bir yüzdesini muhafaza etmektedir.

Kapanış fiyatları indikatörün üstüne çıktığı yerlerde alım sinyali ve altına düştüğü yerlerde satım sinyali üretmektedir.

2.17 Williams'ın %R Göstergesi (Williams' %R)

Williams'ın %R göstergesi aşırı alım ve aşırı satım seviyelerini ölçen bir momentum göstergesidir. Larry Williams tarafından geliştirilmiştir. Stokastik Osilatör'e benzemekle birlikte gösterge "0" ile "–100" arasında dalgalanır. "0" ile "–20" aralığında aşırı alım, "–80" ile "–100" aralığında ise aşırı satım seviyelerine ulaşılmış demektir. Formülü aşağıdaki gibidir:

$$HHV_i = \max\{H_i, H_{i-1}, H_{i-2}, ...H_{i-n}\}$$
$$LLV_i = \min\{L_i, L_{i-1}, L_{i-2}, ...L_{i-n}\} \qquad (2.15)$$
$$WIL\%R_i = \left[\frac{(HHV_i - C_i)}{(HHV_i - LLV)}\right] \times (-100)$$

Burada WIL%R=Williams'ın %R Göstergesidir. WIL%R göstergesi "-80" seviyesinin altından yukarı doğru döndüğünde alım, "-20" seviyesinin üzerinden aşağı doğru döndüğünde de satım yapılmalıdır.

2.18 Williams'ın Toplama Dağıtım Göstergesi(Williams' Accumulation/ Distribution)

Williams'ın Toplama Dağıtım göstergesi bir fiyat göstergesidir. Williams'a göre eğer hisse senedi yeni bir tepe yaptığında gösterge yeni bir tepe yapıp düşerse, satım yapılmalıdır. Eğer senet yeni bir dip yaptığında gösterge yeni bir dip yapıp yükselirse, alım yapılmalıdır. Formülü aşağıdaki gibidir:

$$TRH_i = \max\{C_{i-1}, H_i\} \qquad TRL_i = \min\{C_{i-1}, L_i\}$$
$$AD_i = \begin{cases} C_i - TRL_i & , C_i > C_{i-1} \\ 0 & , C_i = C_{i-1} \\ C_i - TRH_i & , C_i < C_{i-1} \end{cases} \qquad (2.16)$$
$$WILADI_i = \sum_{i=1}^{\infty} AD_i$$

Burada WILADI= Williams'ın Toplama Dağıtım Göstergesi' dir.

2.19 Piyasa Kolaylık Endeksi (Market Facilitation index, MFI)

Piyasa Kolaylık Endeksi (Market Facilitation Index), Dr. Bill Williams tarafından geliştirilmiştir. Fiyat hareketi ve işlem hacmini birleştiren bir yöntemdir. Günlük işlem aralığının işlem hacmine bölünmesiyle elde edilir. İşlem hacmi başına fiyat

hareketi ölçerek fiyat hareketinin etkinliğini gösterir. Formülü aşağıdaki gibidir:

$$MF_i = \left(\frac{H_i - L_i}{V_i} \right) \qquad MF_i = \begin{cases} 1 & eger \quad MF_i > MF_{i-1} \\ 0 & eger \quad MF_i = MF_{i-1} \\ -1 & eger \quad MF_i < MF_{i-1} \end{cases} \tag{2.17}$$

Burada MFI= Piyasa Kolaylık Endeksi' dir.

2.20 İzdüşüm Osilatörü (Projection Oscillator)

İzdüşüm Osilatörü (Projection Oscillator), yine Mel Widner tarafından geliştirilmiştir. Osilatör, " 20" seviyesinin altına düşüp sonra bu seviyenin üstüne çıktığı zaman alım, "80" seviyesinin üstüne çıkıp sonra bu seviyenin altına düştüğü zaman satım yapılmalıdır. Yüksek değerler ("80"in üstü) aşırı iyimserliği, düşük değerler ("20"nin altı) aşırı kötümserliği gösterir. Eğer piyasa trend yapmıyorsa, bu indikatör aşırı alım ve aşırı satım göstergesi olarak iyi sonuçlar verecektir. Eğer piyasa trend yapıyorsa, osilatör trendin yönünde işlemler yapmak için kullanılabilir.

2.21 Güniçi Momentum Endeksi (Intraday Momentum Index, IMI)

Güniçi Momentum Endeksi (Intraday Momentum Index, IMI), Tushar Chande tarafından geliştirilmiştir. Göreceli Güç Endeksi (Relative Strength Index, RSI) ile mum grafikleri analizinin karışımıdır.

IMI'nın hesaplaması, RSI'a benzemekle birlikte, burada günlük açılış fiyatlarıyla kapanış fiyatları arasındaki ilişki kullanılarak günün yükseliş mi yoksa düşüş mü yaptığı tanımlanır. Eğer kapanış, açılışın üstündeyse yükseliş; altındaysa düşüş söz konusudur. Formülü aşağıdaki gibidir:

$$D1_i = \begin{cases} C_i - O_i & eger \; C_i - O_i > 0 \\ 0 & eger \; C_i - O_i < 0 \end{cases} \qquad SUM1_i = \sum_{i=1}^{n} D1_i$$

$$D2_i = \begin{cases} 0 & eger \; C_i - O_i > 0 \\ |C_i - O_i| & eger \; C_i - O_i > 0 \end{cases} \qquad SUM2_i = \sum_{i=1}^{n} D2_i \tag{2.18}$$

$$IMI_i = \left[\frac{SUM1_i}{SUM1_i + SUM2_i} \right] \times 100$$

2.22 Q-STICK Göstergesi (Qstick Indicator)

Q-Stick Göstergesi, Tushar Chande tarafından geliştirilmiştir. Q-Stick Göstergesi, açılış ve kapanış fiyatları arasındaki farkın hareketli ortalamasıdır. Fiyatlar ortalamanın üzerine çıktığında alım, altına düştüğünde satım yapılmalıdır.

Formülü aşağıdaki gibidir:

$$QSTICK_i = \frac{\left(\sum_{i=1}^{n}(C_i - O_i)\right)}{n} \tag{2.19}$$

Burada QSTICK Q-Stick Göstergesidir. Göstergenin sinyal seviyesini yukarı veya aşağı kesmesine göre alım veya satım yapılabileceği gibi, "+50"nin üzerinden aşağı döndüğünde satım ve "-50"nin altından yukarı döndüğünde de alım yapılabilir.

3. VERİ MADENCİLİĞİ

Günümüzde kullanılan veri tabanı yönetim sistemleri eldeki verilerden sınırlı çıkarımlar yaparken geleneksel çevrimiçi işlem sistemleri (on-line transaction processing systems) de bilgiye hızlı, güvenli erişimi sağlamaktadır. Fakat ikisi de eldeki verilerden analizler yapıp anlamlı bilgiler elde etme imkanını sağlamakta yetersiz kalmışlardır. Verilerin yığınla artması ve anlamlı çıkarımlar elde etme ihtiyacı arttıkça uzmanlar Knowledge Discovery in Databases (KDD) adı altında çalışmalarına hız kazandırmışlardır. Bu çalışmalar sonucunda da veri madenciliği (Data Mining) kavramı doğmuştur. 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.

3.1 Tanım

İlişkisel veri tabanı sistemleriyle ulaşılan veriler tek başına bir anlam ifade etmezken veri madenciliği teknolojisi bu verilerden anlamlı bilgi üretilmede öncü rol oynamaktadır. Aşağıda bazı veri madenciliği tanımlarına yer verilmektedir.

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, otomatik veya yarı otomatik çözüm araçları (tools) ile büyük ölçeklerdeki verinin anlamlı yapılar ve kurallar keşfetmek üzere araştırılması (exploration) ve analiz edilmesidir.” [3].

3. “Veri madenciliği çok büyük tabanları içindeki veriler arasındaki bağlantılar ve örüntüleri araştırarak, gizli kalmış yararlı olabilecek verilerden değerli bilginin çıkarılması sürecidir.” [4].
4. “Veri Madenciliği, büyük veri ambarlarından daha önceden bilinmeyen, doğru ve eyleme geçirilebilir bilgiyi ayrıştırma ve çok önemli kararların alınması aşamasında ayrıştırılan bu bilgiyi kullanma sürecidir.” [5].

Yukarıdaki tanımları toplayıp veri madenciliği kavramına ek bir tanım daha getirilebilir. Veri madenciliği; matematiksel yöntemler yardımıyla, biriken veri yığınları içerisinde bulunan dataların birbirleriyle ilişkisini ortaya çıkartmak için yapılan analiz ve kurulan modeller sonucunda elde edilecek bilgi keşfi sürecidir.

Veri madenciliğinin, disiplinler arası bir teknoloji olarak dört ana başlıktan oluştuğu kabul edilmektedir. Bunlar sınıflama, kategori etme, tahmin etme ve görüntülemedir. Bu dört temel dışında istatistik, makine bilgisi, veritabanları ve yüksek performanslı işlem gibi temelleri de içerir.

3.2 Tarihsel Gelişim

Veri madenciliğinin kavram olarak oluşması 1960’lı yıllara kadar dayanmaktadır. Bu dönemlerde veri taraması (data dredging), veri yakalanması (data fishing) gibi isimler verilmiş ve bilgisayar yardımıyla gerekli sorgulama (query) yapıldığında istenilen bilginin elde edilebileceği düşünülmüştür. Fakat 1990’lar geleneksel istatistiksel yöntemlerinin yerine algoritmik bilgisayar modülleri ile veri analizinin gerçekleştirilebileceğinin kabul edildiği yıllar olmuştur. Veri madenciliğinin tarihsel süreci Tablo1.1 de gösterilmiştir [6].

Gelişim Adımları	Cevaplanan Karar Problemi	Kullanılabilen Teknolojiler	Ürün Sağlayıcıları	Karakteristikler
Veri Toplama (1960'lar)	"Benim toplam karım geçen 5 yılda ne kadardı?"	Bilgisayarlar, Teypler, Diskler	IBM,CDC	Geriye dönük , statik veri dağıtımı
Veri Erişimi (1980'ler)	"İngiltere'de geçen mart ayında birim satışları ne kadardı?"	İlişkisel Veritabanları, SQL, ODBC	Oracle,Sybase, Informix,IBM, Microsoft	Kayıt düzeyinde geriye dönük, dinamik veri dağıtımı
Veri Ambarlama ve Karar Destek Sistemleri (1990'lar)	"İngiltere'de geçen mart ayında birim satışları ne kadardı?"	OLAP, Çok Boyutlu Veritabanı Sistemleri, Veri ambarları	Pilot, Comshare, Arbor,Cognos, Microstrategy	Çoklu düzeylerde, geriye dönük dinamik veri dağıtımı
Veri Madenciliği (Bugün)	"Gelecek ay Boston'daki birim satışlar muhtemelen ne olabilir, niçin?	İleri düzeyde algoritmalar, çok işlemcili bilgisayarlar, büyük veritabanları	Pilot, Lockheed, IBM,SGI, SPSS,SAS, Microsoft vs.	Geleceğe dönük ,proaktif enformasyon dağıtımı

Tablo 3.1: Veri madenciliğinin tarihsel gelişimi

3.3 Kullanım Alanları

Tarihsel süreç, gelişen teknoloji ile veri madenciliğinin işlevliğini etkin bir şekilde sürdürdüğünü göstermektedir. Veriler çok hızlı bir şekilde toplanabilmekte, depolanabilmekte, işlenebilmekte ve bilgi olarak kurumların hizmetine sunulabilmektedir. Günümüzde bilgiye hızlı erişim, firmaların sürekli yeni stratejiler geliştirip etkili kararlar almalarını sağlayabilmektedir. Bu süreçte araştırmacılar, büyük hacimli ve dağınık veri setleri üzerinde firmalara gerekli bilgi keşfini daha hızlı gerçekleştirebilmeleri için veri madenciliği üzerine çalışmalar yapmışlardır. Tüm bu çalışmalar doğrultusunda veri madenciliği günümüzde yaygın bir kullanım alanı bulmuştur. Kısaca veri madenciliğinin kullanılabileceği alanlar aşağıdaki gibidir.

- Perakende/ Pazarlama
- Bankacılık
- Sağlık hizmetleri ve sigortacılık
- Tıp
- Ulaştırma

- Eğitim
- Ekonomi
- Güvenlik
- Elektronik ticaret

3.4 Veri Madenciliği Modelleri

IBM tarafından veri işleme operasyonları için iki çeşit model tanımlanmıştır.

3.4.1 Doğrulama modeli

Doğrulama modeli kullanıcıdan bir hipotez alarak testler yapar ve bu hipotezin geçerliliğini araştırır.

3.4.2 Keşif modeli

Sistem bu modelde önemli bilgileri gizli veriden otomatik olarak elde eder. Veri başka hiçbir aracıya ihtiyaç duymadan yaygın olarak kullanılan modeller, genelleştirmeler ile ayıklanır.

3.5 Veri Madenciliği Uygulamaları İçin Temel Adımlar

Veri madenciliği uygulamalarında sırasıyla takip edilmesi gereken temel aşamalar aşağıda sistematik biçimde verilmiştir.

3.5.1 Uygulama alanının ortaya konulması

Bu ilk adımda veri madenciliğinin hangi alan ve hangi amaç için yapılacağı tespit edilir.

3.5.2 Hedef veri grubu seçimi

Belirlenen amaç doğrultusunda bazı kriterler belirlenir. Bu kriterler çerçevesinde aynı veya farklı veritabanlarından veriler toplanarak hedef (target) veri grubu elde edilir.

3.5.3 Model seçimi

Veri madenciliği probleminin seçimi datalar üzerinden belirlenir. (Sınıflandırma, Kümeleme, Birliktelik Kuralları, Şablonların ve İlişkilerin Yorumlanması v.b.)

3.5.4 Ön işleme

Bu aşamada seçilen veriler ayıklanarak silinir, eksik veri alanları üzerine stratejiler geliştirilir. Veriler tekrardan düzenlenip tutarlı bir hale getirilir. Bu aşamada yapılan işlem data temizleme ve data birleştirme olarak bilinen uyumlandırma işlemidir. Veri birleştirme (bütünleştirme), farklı veri tabanlarından ya da kaynaklarından elde edilen verilerin birlikte değerlendirmeye alınabilmesi için farklı türdeki verilerin tek türe dönüştürülebilmesi demektir

3.5.5 Veri indirgeme

Çözümleme işlemi veri madenciliği uygulamalarında uzun sürebilmektedir. İşlem yapılırken eksik ya da uygun olmayan verilerin oluşturduğu tutarsız verilerle karşılaşılabilir. Bu gibi durumlarda verinin söz konusu sorunlardan arındırılması gerekmektedir. Çözümlemeden elde edilecek sonuçta bir değişiklik olmuyorsa veri sayısı ya da değişkenlerin sayısında azaltmaya gidilir. Veri indirgeme çeşitli biçimlerde yapılabilir.

Veriyi indirirken bu verileri çok boyutlu veri küpleri biçimine dönüştürmek söz konusu olabilir. Böylece çözümler sadece belirlenen boyutlara göre yapılır. Veriler arasında seçme işlemi yapılarak da veri tabanından veriler silinip boyut azaltılması yapılır.

3.5.6 Veri dönüştürme

Verileri direkt veri madenciliği çözümlerine katmak çoğu zaman uygun olmayabilir. Değişkenlerin ortalama ve varyansları birbirlerinden çok farklıysa büyük ortalama ve varyansa sahip değişkenlerin diğer değişkenler üzerindeki etkisi daha fazla olur. Bu nedenle bir dönüşüm yöntemi uygulanarak değişkenlerin normalleştirilmesi ya da standartlaşması uygun yoldur.

3.5.7 Algoritmanın belirlenmesi

Bu aşamada indirgenmiş veriye ve kullanılacak modele hangi algoritmanın uygulanacağına karar verilir. Mümkünse bu algoritmanın seçimine uygun veri madenciliği yazılımı seçilir değilse oluşturulan algoritmaya uygun programlar yazılır.

3.5.8 Yorumlama ve doğrulama

Uygulama sonucunda elde edilen veriler üzerine yorumlama yapılır, bu yorumların test verileri üzerinden doğrulanması hedeflenir. Doğruluğu onaylanan bu yorumlar gizli bilgiye ulaşıldığını göstermektedir. Elde edilen bu bilgiler çoğu kez grafiklerle desteklenir.

3.6 Temel Veri Madenciliği Problemleri ve Çözüm Yöntemleri

Veri madenciliği uygulaması gerektiren problemlerde, farklı veri madenciliği algoritmaları ile çözüme ulaşılmaktadır. Veri madenciliği görevleri iki başlık altında toplanmaktadır.

- 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 açıklanmaktadır.

3.6.1 Karakterize etme (Characterization)

Veri karakterizasyonu hedef sınıfındaki verilerin seçilmesi, bu verilerin genel özelliklerine göre karakteristik kuralların oluşturulması olayıdır.

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

3.6.2 Ayrımlaştırma (Discrimination)

Belirlenen hedef sınıfa karşıt olan sınıf elemanlarının özellikleri arasında karşılaştırma yapılmasını sağlayan algoritmadır. Karakterize etme metodundan farkı mukayese yöntemini kullanmasıdır.

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

3.6.3 Sınıflandırma (Classification)

Sınıflandırma, veri tabanlarındaki gizli örüntüleri ortaya çıkarabilmek için veri madenciliği uygulamalarında sıkça kullanılan bir yöntemdir. Verilerin sınıflandırılması için belirli bir süreç izlenir. Öncelikle var olan veritabanının bir kısmı eğitim amaçlı kullanılarak sınıflandırma kurallarının oluşturulması sağlanır. Bu kurallar kullanılarak veriler sınıflandırılır. Bu veriler sınıflandırıldıktan sonra eklenecek veriler bu sınıflardan karakteristik olarak uygun olan kısma atanır. Sınıflandırma problemleri için "Oracle Data Miner" (ODM)' ın uyumlu olduğu çözüm yöntemleri Naivé Bayes (NB), Karar Destek Vektörleri (SVM), Karar Ağaçları ve Adaptive Bayes Network(ABN)' 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.

3.6.4 Tahmin etme (Prediction)

Kayıt altında tutulan geçmiş verilerin analizi sonucu elde edilen bilgiler gelecekte karşılaşılacak aynı tarz bir durum için tahmin niteliği taşıyacaktır. Örneğin ABC 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.

3.6.5 Birliktelik kuralları (Association rules)

Birliktelik kuralları gerek birbirini izleyen gerekse de eş zamanlı durumlarda araştırma yaparak, bu durumlar arasındaki ilişkilerin tanımlanmasında kullanılır. Bu modelin yaygın olarak Market Sepet Analizi uygulamalarında kullanıldığı

bilinmektedir. 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.

3.6.6 Kümeleme (Clustering)

Yapı olarak sınıflandırmaya benzeyen kümeleme metodunda birbirine benzeyen veri grupları aynı tarafta toplanarak kümelenmesi sağlanır. Sınıflandırma metodunda sınıfların kuralları, sınırları ve çerçevesi belli ve datalar bu kriterlere göre sınıflara atanırken kümeleme metodunda sınıflar arası bir yapı mevcut olup, benzer özellikte olan verilerle yeni gruplar oluşturmak asıl hedeftir. Verilerin kendi aralarındaki benzerliklerinin göz önüne alınarak gruplandırılması yöntemin pek çok alanda uygulanabilmesini sağlamıştır. Örneğin, pazarlama araştırmalarında, desen tanımlama, resim işleme ve uzaysal harita verilerinin analizinde kullanılmaktadır. Tüm bu uygulama alanlarında kullanılması, ODM'nin desteklediği "K-means" ve "O-Cluster" kümeleme yöntemleri ile mümkün kılınmıştır.

3.6.7 Aykırı değer analizi (Outlier analysis)

İstisnalar veya sürpriz olarak tespit edilen aykırı veriler, bir sınıf veya kümelemeye tabii tutulamayan veri tipleridir. Aykırı değerler bazı uygulamalarda atılması gereken değerler olarak düşünülürken bazı durumlarda ise çok önemli bilgiler olarak değerlendirilebilmektedir.

Örneğin 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.

3.6.8 Zaman serileri (Time series)

Yapılan veri madenciliği uygulamalarında kullanılan veriler çoğunlukla statik değildir ve zamana bağlı olarak değişmektedir. Bu metot ile bir veya daha fazla niteliğin belirli bir zaman aralığında, eğilimindeki değişim ve sapma durumlarını inceler. Belirlenen zaman aralığında ölçülebilir ve tahmin edilen/beklenen değerleri karşılaştırmalı olarak inceler ve sapmaları tespit eder.

Örneğin ABC şirketinin Ocak-Haziran 2009 dönemi için önceki yılın satış miktarları göz önünde tutularak bir hedef ortaya konulmuştur. 2008 ve 2009 değerleri karşılaştırmalı olarak incelenerek sapma miktarı belirlenir. ODM her ne kadar çeşitli histogramlarla kullanıcıya görsel destek sağlasa da tam anlamıyla bu tür problemleri desteklememektedir.

3.6.9 Veri görüntüleme (Visualization)

Bu metot, çok boyutlu özelliğe sahip verilerin içerisindeki karmaşık bağlantıların/bağıntıların görsel olarak yorumlanabilme imkanını sağlar. Verilerin birbirleriyle olan ilişkilerini grafik araçları görsel ya da grafiksel olarak sunar. ODM zaman serilerinde olduğu gibi histogramlarla bu metodu desteklemektedir.

3.6.10 Yapay sinir ağları (Artificial neural networks)

Yapay sinir ağları insan beyninden esinlenilerek 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.

3.6.11 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.

3.6.12 Karar ağaçları (Decision trees)

Ağaç yapıları esas itibariyle 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.

3.6.13 Kural çıkarma (Rules induction)

İstatistiksel öneme sahip yararlı "if-else" kurallarının ortaya çıkarılması problemlerini inceler.

4. TEMEL KAVRAMLAR VE MATEMATİKSEL ALTYAPI

Bu bölümde veri kümesine uygulamak üzere seçilen sınıflandırma modeli Geliştirilmiş Bayesian Ağlar' ın (Adaptive Bayes Network) açıklanması için öncesinde bu modelin kullandığı Naivé Bayes yönteminin matematiksel altyapısından bahsedilmiştir.

4.1 Naivé Bayes Yöntemi

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

4.1.1 Temel kavramlar

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 aşağıda gösterilen tablonun formatındadır. Tabloda her satır (her kayıt) 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.

4.1.2 Teori

Naivé Bayes yöntemi ile sınıflandırma koşullu olasılık hesabına dayanmaktadır. Şekil 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.

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.

Öncelikle 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 iken balığa çıktığı gibi, yağmur yağmazken de balığa çıkabilir, yani bir *B* olayına bağlı olarak *A* olayının olma olasılığı

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

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

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

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;

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

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

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

elde edilir. (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 birçok 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 [10].

4.1.3 Örnekler

Örnek 1) *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 4.1a' 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 döndüğü açıklanmıştır. Burada kısaltma amacıyla Gelir=G, Evet=E, Hayır=H şeklinde sembolize edilecektir. Tablo 4.1a verisinden elde edilen her farklı gruptaki kişi sayısı Tablo.4.1b ile gösterilmiştir.

Tablo 4.1a : Gelir-Mülk ilişkisi

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

Tablo 4.1b : Her gruptaki kişi sayısı

Gelir	Ev=E	Ev=H
Düşük	200	130
Orta	100	220
İyi	130	100
Yüksek	110	10

.	.
.	.

Tablo 4.1b yardımıyla sözü edilen olasılıklar (4.8) ifadesi kullanılarak;

$$P(Ev = E \mid Gelir = D) = \frac{P(Ev = E)P(G = D \mid Ev = E)}{P(Ev = E)P(G = D \mid Ev = E) + P(Ev = H)P(G = D \mid Ev = H)} \quad (4.11)$$

$$= \frac{\frac{540}{1000}\frac{200}{540}}{\frac{540}{1000}\frac{200}{540} + \frac{460}{1000}\frac{130}{460}}$$

$$P(Ev = E \mid Gelir = D) = 0.6061$$

$$P(Ev = H \mid Gelir = D) = 1 - \frac{20}{33} = \frac{13}{33} = 0.3939$$

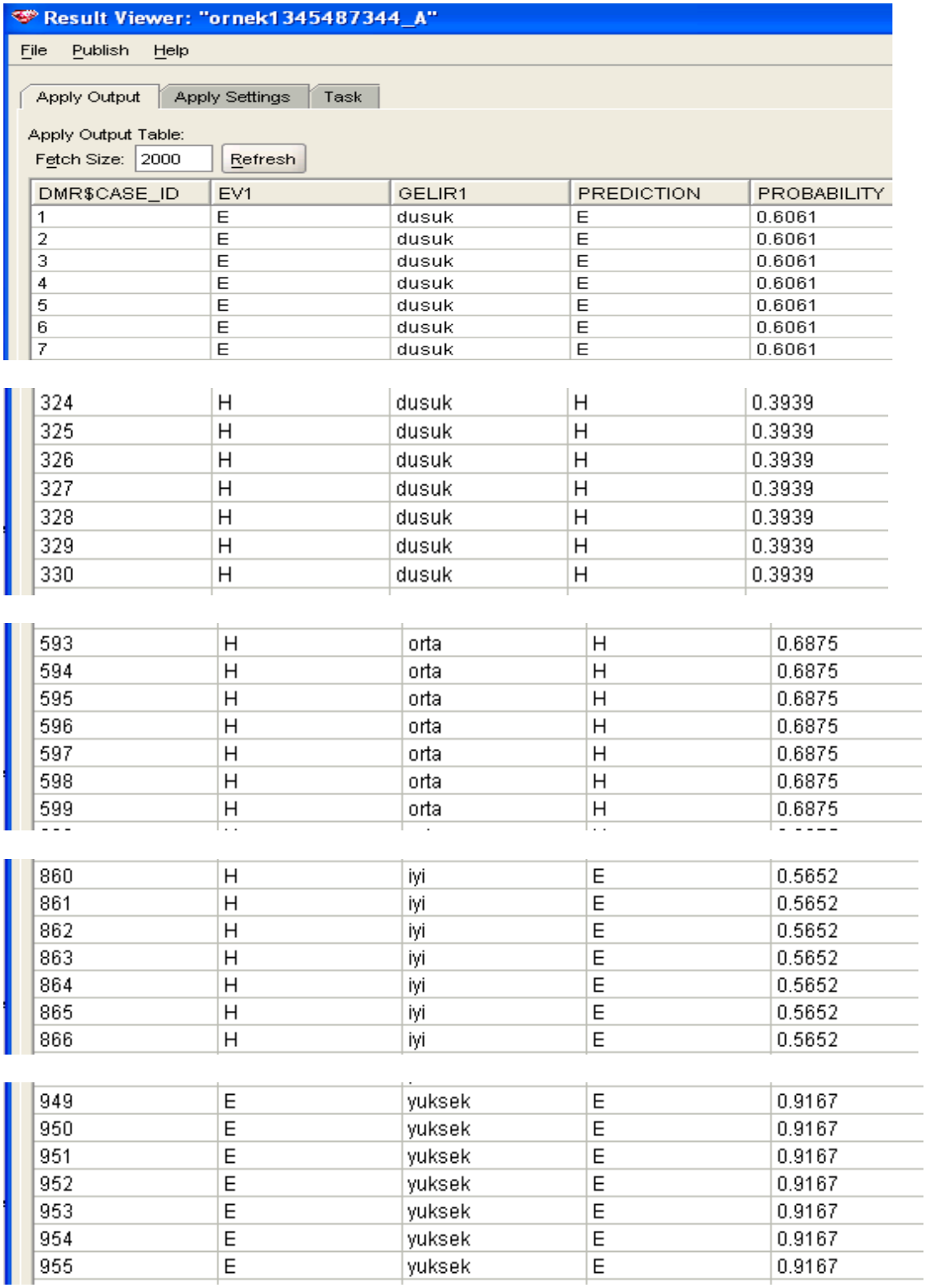
olarak hesaplanabilir. Burada bu sonuçlar çok daha kolay bir şekilde Tablo 4.1b' den de görülmektedir. Fakat hem hedef özelliğin ikiden fazla hem de kestirimci özellik sayısının birden fazla olduğu durumlarda tablodan okuma zorlaşacak ve yukarıdaki formülün uygulanması gerekecektir. Benzer şekilde diğer olasılıklar da hesaplanarak;

$$P(Ev = E \mid Gelir = O) = 0.3125$$
$$P(Ev = H \mid Gelir = O) = 1 - 0.3125 = 0.6875$$
$$P(Ev = E \mid Gelir = \dot{I}) = \frac{13}{23} = 0.5652$$
$$P(Ev = H \mid Gelir = \dot{I}) = 1 - \frac{13}{23} = 0.4348$$
$$P(Ev = E \mid Gelir = Y) = \frac{11}{12} = 0.9167$$
$$P(Ev = H \mid Gelir = Y) = 1 - \frac{11}{12} = 0.0833$$

yazılabilir.

Yukarıdaki hesaplamaların ODM' nin elde ettiği sonuçlarla karşılaştırılabilmesi için Naivé Bayes modeli oluşturulurken Discretize, Sample ve Split adımları atlanmalı, Cost Matrix seçeneği de kaldırılmalıdır. Ayrıca normalde ODM' de model oluşturulurken Apply aktivitesinde kullanılan tablo Build aktivitesinde kullanılandan farklı olmalıdır, çünkü Apply aktivitesindeki amaç yeni veri için tahmin kolonunun oluşturulmasıdır. Fakat burada sadece sonuçların doğruluğunun görülmesi

amaçlandığından Apply aktivitesi de aynı tabloya uygulanmıştır. Aşağıda ODM' nin bu örneğe uygulanması sonucu elde edilen ekran çıktısı verilmiş, sonuçların aynı olduğu gözlenmiştir.



DMR$CASE_ID	EV1	GELIR1	PREDICTION	PROBABILITY
1	E	dusuk	E	0.6061
2	E	dusuk	E	0.6061
3	E	dusuk	E	0.6061
4	E	dusuk	E	0.6061
5	E	dusuk	E	0.6061
6	E	dusuk	E	0.6061
7	E	dusuk	E	0.6061
324	H	dusuk	H	0.3939
325	H	dusuk	H	0.3939
326	H	dusuk	H	0.3939
327	H	dusuk	H	0.3939
328	H	dusuk	H	0.3939
329	H	dusuk	H	0.3939
330	H	dusuk	H	0.3939
593	H	orta	H	0.6875
594	H	orta	H	0.6875
595	H	orta	H	0.6875
596	H	orta	H	0.6875
597	H	orta	H	0.6875
598	H	orta	H	0.6875
599	H	orta	H	0.6875
860	H	iyi	E	0.5652
861	H	iyi	E	0.5652
862	H	iyi	E	0.5652
863	H	iyi	E	0.5652
864	H	iyi	E	0.5652
865	H	iyi	E	0.5652
866	H	iyi	E	0.5652
949	E	yuksek	E	0.9167
950	E	yuksek	E	0.9167
951	E	yuksek	E	0.9167
952	E	yuksek	E	0.9167
953	E	yuksek	E	0.9167
954	E	yuksek	E	0.9167
955	E	yuksek	E	0.9167

Şekil 4.2 : Örnek 1'in ekran çıktıları

ODM, Naivé 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 düşük olanın ev sahibi olma olasılığı 0.6061 olarak hesaplandığı için tahmin evet ve sonucun güvenilirliği 0.6061'dir. Geliri orta olan kişinin ev sahibi olma olasılığı ise 0.3125 olduğu için modelin tahmini hayır ve sonucun güvenilirliği 0.6875 olacaktır. Bu şekilde işleme devam edilerek tüm tablo yeniden oluşturulur.

Tablo 4.2 : Tablo 4.1a' nın yapılan hesaplamalarla elde edilen test sonuçları

Gelir	Ev	Tahmin	Güvenilirlik
Düşük	Evet	Evet	0.6061
Orta	Evet	Hayır	0.6875
Yüksek	Hayır	Evet	0.9167
İyi	Hayır	Evet	0.5652
İyi	Evet	Evet	0.5652
.	.	.	.

Modelin tüm güvenilirliği ise gerçek değerler ile tahmini değerlerin karşılaştırılması sonucu elde edilen aşağıdaki güvenilirlik matrisi ile verilebilir.

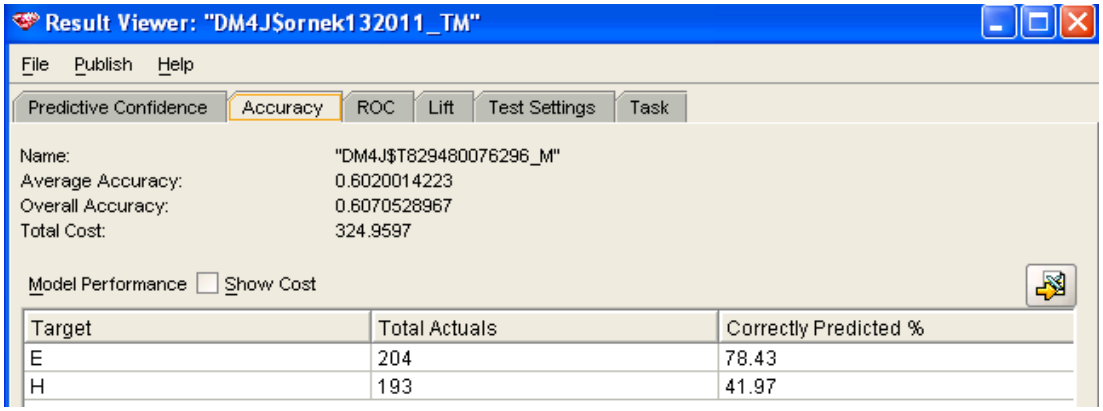
Tablo 4.3 : Güvenilirlik matrisi

	E	H
E	440	100
H	240	220

Tablo 4.3'te 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 440 (doğru), gerçekte evi varken modelin hayır olarak tahmin ettiği kayıt sayısı (yanlış) 100 dür. Dolayısıyla matrisin köşegeni doğru kayıt sayısını, köşegen dışı ise yanlış kayıt sayısını göstermektedir. Buradan modelin doğruluğu

$$\frac{420+220}{440+100+240+220}=0.66$$

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.



Şekil 4.3: Örnek 1'in ODM ile güvenilirliği

Örnek 2) *İki Boyut için:* Yapılan bir araştırma sonucunda 1000 arabanın rengi "Kırmızı" ve "Sari"; tipi "Spor" ve "Klasik"; kökeni ise "Yerli" ve "İthal" olarak ve bu özelliklerdeki arabaların çalınma durumları çalınmışsa "1" çalınmamışsa "0" olarak belirleniyor. Bu durumlar Tablo 4.4 de gösterildiği gibi olsun. Kayıtların dağılımı Tablo 4.5' te verilmiştir.

Tablo 4.4: Model Verisi

Renk	**Tip**	**Köken**	**Çalınma**
Kırmızı	Spor	Yerli	1
Kırmızı	Spor	Yerli	0
Kırmızı	Spor	İthal	1
Sari	Spor	İthal	0
Sari	Klasik	Yerli	1
…	…	…	…

Tablo 4.5: Kayıt Dağılımları

Renk	**Tip**	**Köken**	**Çalınma**	**Kayıt Sayısı**
Kırmızı	Spor	Yerli	1	200
Kırmızı	Spor	Yerli	0	100
Kırmızı	Spor	İthal	1	100
Kırmızı	Spor	İthal	0	0
Kırmızı	Klasik	Yerli	1	0
Kırmızı	Klasik	Yerli	0	0
Kırmızı	Klasik	İthal	1	0

Kırmızı	Klasik	İthal	0	100
Sari	Spor	Yerli	1	0
Sari	Spor	Yerli	0	100
Sari	Spor	İthal	1	100
Sari	Spor	İthal	0	0
Sari	Klasik	Yerli	1	0
Sari	Klasik	Yerli	0	100
Sari	Klasik	İthal	1	100
Sari	Klasik	İthal	0	100

Tablo 4.5 yardımıyla arabaların çalınıp çalınmama olasılıkları (4.10) ifadesiyle belirlenmeye çalışılmıştır. Burada üç belirleyici özellik olduğundan (4.10) ifadesi bu tabloya uygun formda yazılmalıdır. Burada S=Çalınma durumu, C=Renk, T=Tip, O=Köken, S=1=Çalınmış, S=0=Çalınmamış, R=Kırmızı, Y=Sarı, I=İthal, D=Yerli, P=Spor, K=Klasik'i ifade etmektedir.

(4.10) ifadesi örneğe uygulandığında,

$$P(S=0|T=K,O=D,C=Y)=\frac{P(S=0)P(T=K|S=0)P(O=D|S=0)P(C=Y|S=0)}{P(S=0)P(T=K|S=0)P(O=D|S=0)P(C=Y|S=0)} \\ +P(S=1)P(T=K|S=1)P(O=D|S=1)P(C=Y|S=1)$$

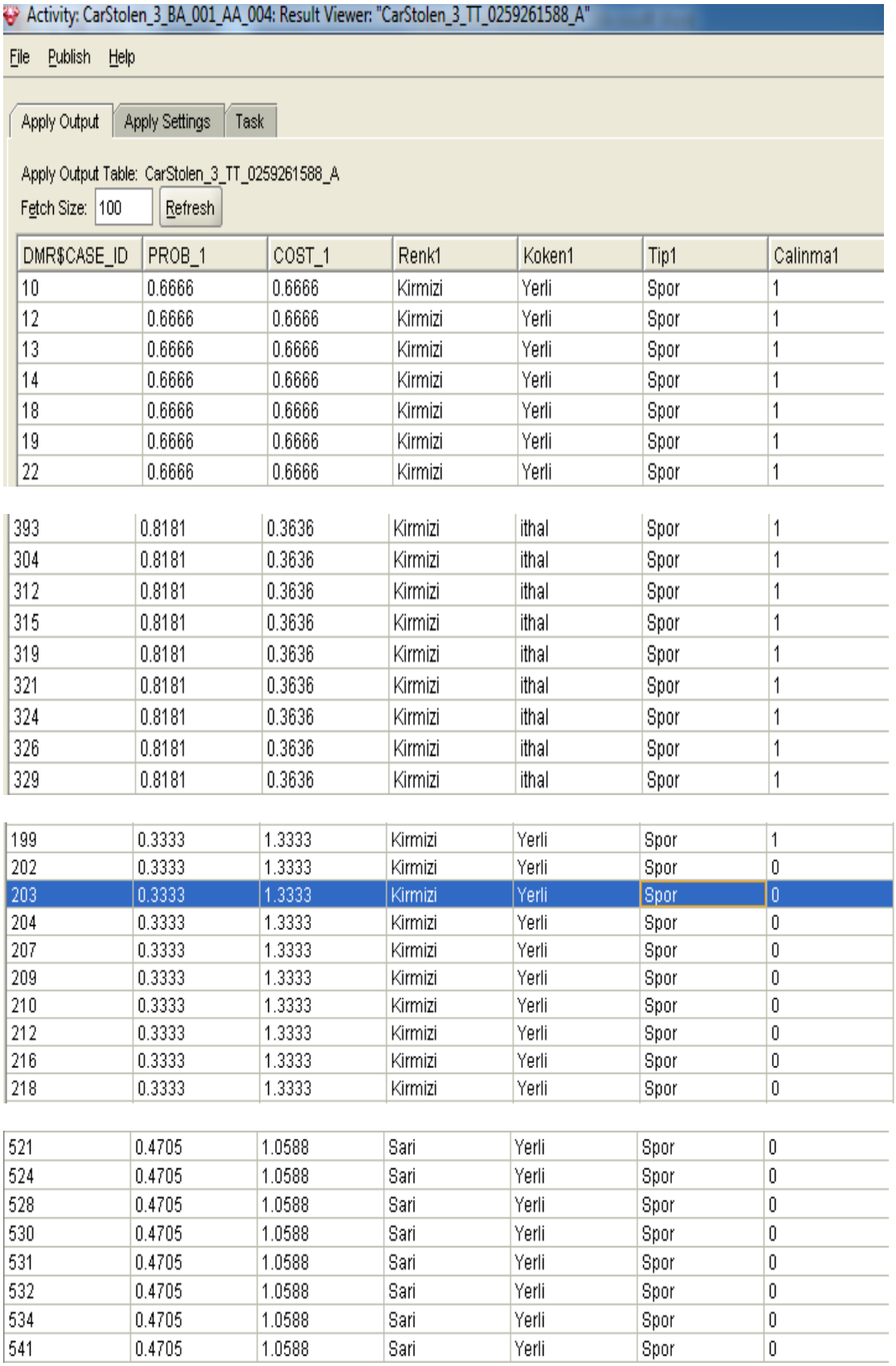
(4.12)

$$=\frac{\frac{500}{1000}\frac{300}{500}\frac{300}{500}\frac{300}{500}}{\frac{500}{1000}\frac{300}{500}\frac{300}{500}\frac{300}{500}+\frac{500}{1000}\frac{100}{500}\frac{200}{500}\frac{200}{500}}=0,8709$$

sonucu elde edilir. Bu sonuca göre rengi sarı, kökeni yerli, tipi klasik olan arabaların çalınmama olasılığı %87.09 bulunur. Benzer şekilde diğer bazı olasılıklar da hesaplanırsa;

$$P(S=1|C=R,T=P,O=D)=0,6666$$
$$P(S=1|C=R,T=P,O=I)=0,8181$$
$$P(S=0|C=R,T=P,O=D)=0,3333$$
$$P(S=0|C=S,T=P,O=D)=0,4705$$

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	PROB_1	COST_1	Renk1	Koken1	Tip1	Calinma1
10	0.6666	0.6666	Kirmizi	Yerli	Spor	1
12	0.6666	0.6666	Kirmizi	Yerli	Spor	1
13	0.6666	0.6666	Kirmizi	Yerli	Spor	1
14	0.6666	0.6666	Kirmizi	Yerli	Spor	1
18	0.6666	0.6666	Kirmizi	Yerli	Spor	1
19	0.6666	0.6666	Kirmizi	Yerli	Spor	1
22	0.6666	0.6666	Kirmizi	Yerli	Spor	1
393	0.8181	0.3636	Kirmizi	ithal	Spor	1
304	0.8181	0.3636	Kirmizi	ithal	Spor	1
312	0.8181	0.3636	Kirmizi	ithal	Spor	1
315	0.8181	0.3636	Kirmizi	ithal	Spor	1
319	0.8181	0.3636	Kirmizi	ithal	Spor	1
321	0.8181	0.3636	Kirmizi	ithal	Spor	1
324	0.8181	0.3636	Kirmizi	ithal	Spor	1
326	0.8181	0.3636	Kirmizi	ithal	Spor	1
329	0.8181	0.3636	Kirmizi	ithal	Spor	1
199	0.3333	1.3333	Kirmizi	Yerli	Spor	1
202	0.3333	1.3333	Kirmizi	Yerli	Spor	0
203	0.3333	1.3333	Kirmizi	Yerli	Spor	0
204	0.3333	1.3333	Kirmizi	Yerli	Spor	0
207	0.3333	1.3333	Kirmizi	Yerli	Spor	0
209	0.3333	1.3333	Kirmizi	Yerli	Spor	0
210	0.3333	1.3333	Kirmizi	Yerli	Spor	0
212	0.3333	1.3333	Kirmizi	Yerli	Spor	0
216	0.3333	1.3333	Kirmizi	Yerli	Spor	0
218	0.3333	1.3333	Kirmizi	Yerli	Spor	0
521	0.4705	1.0588	Sari	Yerli	Spor	0
524	0.4705	1.0588	Sari	Yerli	Spor	0
528	0.4705	1.0588	Sari	Yerli	Spor	0
530	0.4705	1.0588	Sari	Yerli	Spor	0
531	0.4705	1.0588	Sari	Yerli	Spor	0
532	0.4705	1.0588	Sari	Yerli	Spor	0
534	0.4705	1.0588	Sari	Yerli	Spor	0
541	0.4705	1.0588	Sari	Yerli	Spor	0

Şekil 4.4 : Örnek 2'nin ekran çıktıları

4.2 Özelliklerin Önem Sıralaması (Attribute Importance- AI)

ODM' nin kullanıcılarına sunduğu AI, her özelliğin yani her tahmin edici sütunun sonuç sütunu üzerinde etkisini ölçerek elde edilen ölçüm değerlerine göre tahmin edici sütunları önem sırasına göre sıralayan bir uygulamadır. Sınıflandırma modellerinde, özellikle tahmin edici sütun sayısının çok fazla olduğu durumlarda, birçok sütunun hedef sütun üzerinde etkisi olmayabilir hatta bu sütunlar modelin doğruluğuna negatif etki edebilir. Negatif etkisi olan kolonların AI ile belirlenip modelden çıkarılması modelin doğruluğunu artırır.

5. UYGULAMA VE SONUÇLAR

Bu bölümde hisse senetleri alım satım kararlarında kullanılan göstergeler için yapılan modellemenin adımları anlatılmıştır. 22 tane gösterge kullanılmış ve her bir gösterge için 98127 kayıt içeren tabloda hisse senedi ve gösterge verileri ID numaraları ile belirtilmiştir. Göstergeler tahmin edici kolonlar olarak kullanılmış ve gösterge verileri kolonlara yazılırken her bir gösterge için bazı kısa kolon isimleri kullanılmıştır. Bu kolon isimleri ve kolon isimlerinin gösterge olarak karşılıkları aşağıdaki gibidir.

SYMBOL = Hisse senedi sahibi şirketin sembolü

DTYYYYMMDD = Hisse senedi verilerinin ait olduğu tarih

ADI = Toplama-Dağıtım Endeksi (Accumulation-Distribution Index)

CCI = Mal Kanal Endeksi (Commodity Channel Index)

MACD = Hareketli Ortalamaların Birleşmesi-Ayrılması Göstergesi (Moving Average Convergence)

AO = Aroon Osilatörü (Aroon Oscillator)

PO = Fiyat Osilatörü (Price Oscillator)

BB = Bollinger Bantlar (Bollinger Bands)

CMF = Chaikin Para Akım Göstergesi (Chaikin Money Flow)

CMO = Chande Momentum Osilatörü (Chande Momentum Oscillator)

LRI = Doğrusal Regresyon Göstergesi (Linear Regression Indicator)

MG = Momentum Göstergesi

P_ROC = Fiyat Değişim Oranı (Price Rate Of Change)

RMI = Göreceli Momentum Endeksi (Relative Momentum Index)

RSI = Göreceli Güç Endeksi (Relative Strength Index)

SMI = Stokastik Momentum Endeksi (Stochastic Momentum Index)

SO = Stokastik Osilatörü (Stochastic Oscillator)

WS = Wilder'in Düzeltme Göstergesi (Wilder's Smoothing)

WR = Williams'ın %R Göstergesi (Williams's %R)

WAD = Williams'ın Toplama Dağıtım Göstergesi(Williams's Accumulation/ Distribution)

MFI = Para Akım Endeksi (Money Flow Index)

PO2 = İzdüşüm Osilatörü (Projection Oscillator)

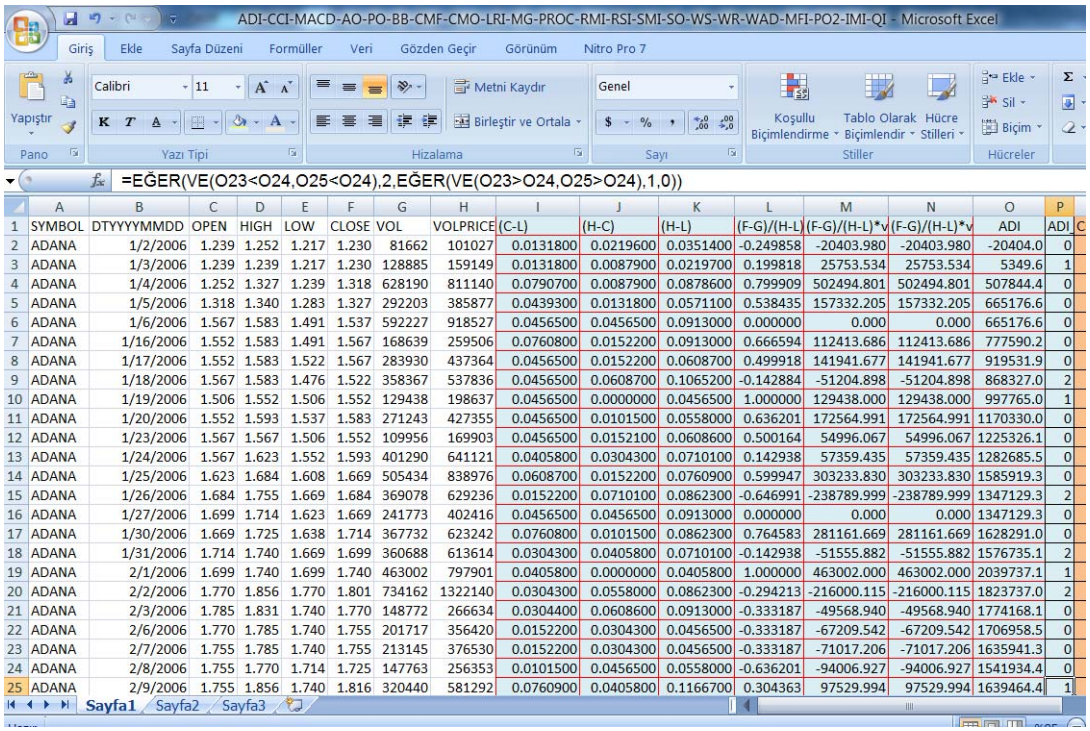
IMI = Gün içi Momentum Endeksi (Intraday Momentum Index)

QI = Q_STICK Göstergesi (Qstick Indicator)

NORMALOLMASIGEREKEN = Hisse senedi bugünkü fiyatının bir sonraki güne oranlayarak kar elde etmek için alım ya da satım kararından hangisinin yapılması gerektiğini belirten kolon.

5.1 Veri Tablosunun Hazırlanması

FOREX şirketinden geçen yılların hisse senetlerinin günlük verileri alınmış ve bu veriler yardımı ile gösterge değerleri Excel programı kullanarak hesaplanmıştır. Gösterge değerleri hesaplandıktan sonra her bir göstergenin özelliğine göre hangi durumlarda al kararı verip hangi durumlarda sat kararı vereceği hesaplanmıştır. Gösterge kararları Excel dosyasına yazılırken "sat kararı" verdiğinde '2' , "al kararı" verdiğinde '1' ve "herhangi bir karar vermediğinde" ise '0' değeri kullanılmıştır. Bazı göstergeler haftalık, on günlük ya da en fazla bir aylık periyotlarla çalıştığı için bu göstergeler periyotlarına bağlı olarak ilk birkaç gün al ya da sat kararı vermemektedir. Bu nedenle bütün göstergelerin aynı anda verdiği kararları görebilmek için ilk 2 aylık veri kayıtlardan çıkarılmıştır. Bazı göstergelerin verdiği kararlar aşağıdaki şekilde görüldüğü gibidir.



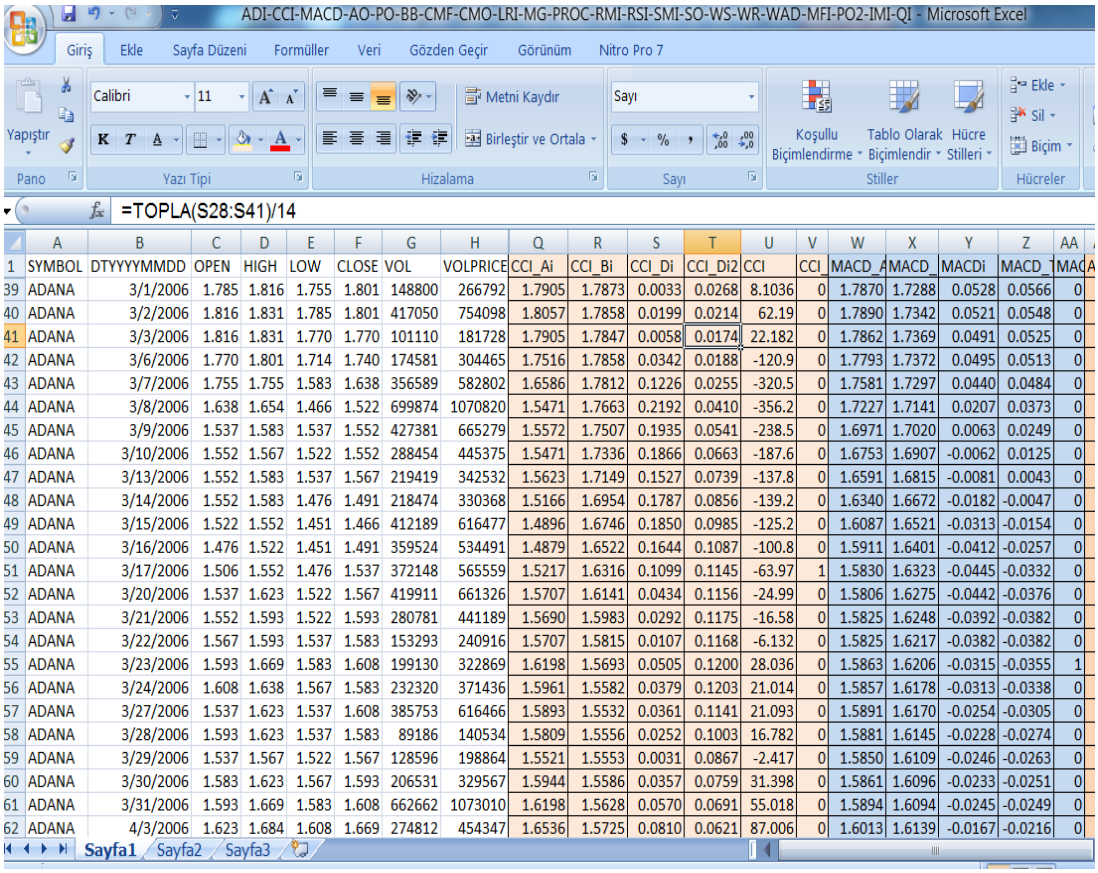
ADI-CCI-MACD-AO-PO-BB-CMF-CMO-LRI-MG-PROC-RMI-RSI-SMI-SO-WS-WR-WAD-MFI-PO2-IMI-QI - Microsoft Excel

=EĞER(VE(O23<O24,O25<O24),2,EĞER(VE(O23>O24,O25>O24),1,0))

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	SYMBOL	DTYYYYMMDD	OPEN	HIGH	LOW	CLOSE	VOL	VOLPRICE	(C-L)	(H-C)	(H-L)	(F-G)/(H-L)	(F-G)/(H-L)*v	(F-G)/(H-L)*v	ADI	ADI_C
2	ADANA	1/2/2006	1.239	1.252	1.217	1.230	81662	101027	0.0131800	0.0219600	0.0351400	-0.249858	-20403.980	-20403.980	-20404.0	0
3	ADANA	1/3/2006	1.239	1.239	1.217	1.230	128885	159149	0.0131800	0.0087900	0.0219700	0.199818	25753.534	25753.534	5349.6	1
4	ADANA	1/4/2006	1.252	1.327	1.239	1.318	628190	811140	0.0790700	0.0087900	0.0878600	0.799909	502494.801	502494.801	507844.4	0
5	ADANA	1/5/2006	1.318	1.340	1.283	1.327	292203	385877	0.0439300	0.0131800	0.0571100	0.538435	157332.205	157332.205	665176.6	0
6	ADANA	1/6/2006	1.567	1.583	1.491	1.537	592227	918527	0.0456500	0.0456500	0.0913000	0.000000	0.000	0.000	665176.6	0
7	ADANA	1/16/2006	1.552	1.583	1.491	1.567	168639	259506	0.0760800	0.0152200	0.0913000	0.666594	112413.686	112413.686	777590.2	0
8	ADANA	1/17/2006	1.552	1.583	1.522	1.567	283930	437364	0.0456500	0.0152200	0.0608700	0.499918	141941.677	141941.677	919531.9	0
9	ADANA	1/18/2006	1.567	1.583	1.476	1.522	358367	537836	0.0456500	0.0608700	0.1065200	-0.142884	-51204.898	-51204.898	868327.0	2
10	ADANA	1/19/2006	1.506	1.552	1.506	1.552	129438	198637	0.0456500	0.0000000	0.0456500	1.000000	129438.000	129438.000	997765.0	1
11	ADANA	1/20/2006	1.552	1.593	1.537	1.583	271243	427355	0.0456500	0.0101500	0.0558000	0.636201	172564.991	172564.991	1170330.0	0
12	ADANA	1/23/2006	1.567	1.567	1.506	1.552	109956	169903	0.0456500	0.0152100	0.0608600	0.500164	54996.067	54996.067	1225326.1	0
13	ADANA	1/24/2006	1.567	1.623	1.552	1.593	401290	641121	0.0405800	0.0304300	0.0710100	0.142938	57359.435	57359.435	1282685.5	0
14	ADANA	1/25/2006	1.623	1.684	1.608	1.669	505434	838976	0.0608700	0.0152200	0.0760900	0.599947	303233.830	303233.830	1585919.3	0
15	ADANA	1/26/2006	1.684	1.755	1.669	1.684	369078	629236	0.0152200	0.0710100	0.0862300	-0.646991	-238789.999	-238789.999	1347129.3	2
16	ADANA	1/27/2006	1.699	1.714	1.623	1.669	241773	402416	0.0456500	0.0456500	0.0913000	0.000000	0.000	0.000	1347129.3	0
17	ADANA	1/30/2006	1.669	1.725	1.638	1.714	367732	623242	0.0760800	0.0101500	0.0862300	0.764583	281161.669	281161.669	1628291.0	0
18	ADANA	1/31/2006	1.714	1.740	1.669	1.699	360688	613614	0.0304300	0.0405800	0.0710100	-0.142938	-51555.882	-51555.882	1576735.1	2
19	ADANA	2/1/2006	1.699	1.740	1.699	1.740	463002	797901	0.0405800	0.0000000	0.0405800	1.000000	463002.000	463002.000	2039737.1	1
20	ADANA	2/2/2006	1.770	1.856	1.770	1.801	734162	1322140	0.0304300	0.0558000	0.0862300	-0.294213	-216000.115	-216000.115	1823737.0	2
21	ADANA	2/3/2006	1.785	1.831	1.740	1.770	148772	266634	0.0304400	0.0608600	0.0913000	-0.333187	-49568.940	-49568.940	1774168.1	0
22	ADANA	2/6/2006	1.770	1.785	1.740	1.755	201717	356420	0.0152200	0.0304300	0.0456500	-0.333187	-67209.542	-67209.542	1706958.5	0
23	ADANA	2/7/2006	1.755	1.785	1.740	1.755	213145	376530	0.0152200	0.0304300	0.0456500	-0.333187	-71017.206	-71017.206	1635941.3	0
24	ADANA	2/8/2006	1.755	1.770	1.714	1.725	147763	256353	0.0101500	0.0456500	0.0558000	-0.636201	-94006.927	-94006.927	1541934.4	0
25	ADANA	2/9/2006	1.755	1.856	1.740	1.816	320440	581292	0.0760900	0.0405800	0.1166700	0.304363	97529.994	97529.994	1639464.4	1

Sayfa1 Sayfa2 Sayfa3

Şekil 5.1: ADI Göstergesinin Excel Görüntüsü



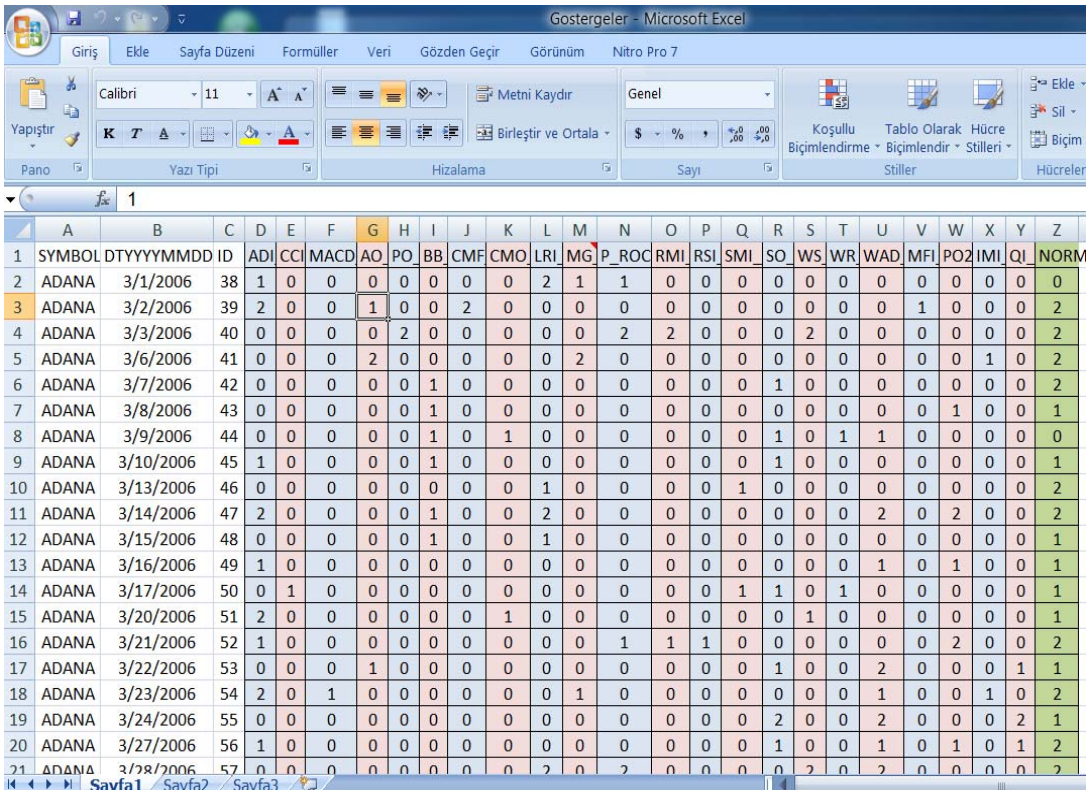
ADI-CCI-MACD-AO-PO-BB-CMF-CMO-LRI-MG-PROC-RMI-RSI-SMI-SO-WS-WR-WAD-MFI-PO2-IMI-QI - Microsoft Excel

=TOPLA(S28:S41)/14

	A	B	C	D	E	F	G	H	Q	R	S	T	U	V	W	X	Y	Z	AA
1	SYMBOL	DTYYYYMMDD	OPEN	HIGH	LOW	CLOSE	VOL	VOLPRICE	CCI_Ai	CCI_Bi	CCI_Di	CCI_Di2	CCI	CCI_	MACD_A	MACD_	MACDi	MACD_1	MACA
39	ADANA	3/1/2006	1.785	1.816	1.755	1.801	148800	266792	1.7905	1.7873	0.0033	0.0268	8.1036	0	1.7870	1.7288	0.0528	0.0566	0
40	ADANA	3/2/2006	1.816	1.831	1.785	1.801	417050	754098	1.8057	1.7858	0.0199	0.0214	62.19	0	1.7890	1.7342	0.0521	0.0548	0
41	ADANA	3/3/2006	1.816	1.831	1.770	1.770	101110	181728	1.7905	1.7847	0.0058	0.0174	22.182	0	1.7862	1.7369	0.0491	0.0525	0
42	ADANA	3/6/2006	1.770	1.801	1.714	1.740	174581	304465	1.7516	1.7858	0.0342	0.0188	-120.9	0	1.7793	1.7372	0.0495	0.0513	0
43	ADANA	3/7/2006	1.755	1.755	1.583	1.638	356589	582802	1.6586	1.7812	0.1226	0.0255	-320.5	0	1.7581	1.7297	0.0440	0.0484	0
44	ADANA	3/8/2006	1.638	1.654	1.466	1.522	699874	1070820	1.5471	1.7663	0.2192	0.0410	-356.2	0	1.7227	1.7141	0.0207	0.0373	0
45	ADANA	3/9/2006	1.537	1.583	1.537	1.552	427381	665279	1.5572	1.7507	0.1935	0.0541	-238.5	0	1.6971	1.7020	0.0063	0.0249	0
46	ADANA	3/10/2006	1.552	1.567	1.522	1.552	288454	445375	1.5471	1.7336	0.1866	0.0663	-187.6	0	1.6753	1.6907	-0.0062	0.0125	0
47	ADANA	3/13/2006	1.552	1.583	1.537	1.567	219419	342532	1.5623	1.7149	0.1527	0.0739	-137.8	0	1.6591	1.6815	-0.0081	0.0043	0
48	ADANA	3/14/2006	1.552	1.583	1.476	1.491	218474	330368	1.5166	1.6954	0.1787	0.0856	-139.2	0	1.6340	1.6672	-0.0182	-0.0047	0
49	ADANA	3/15/2006	1.522	1.552	1.451	1.466	412189	616477	1.4896	1.6746	0.1850	0.0985	-125.2	0	1.6087	1.6521	-0.0313	-0.0154	0
50	ADANA	3/16/2006	1.476	1.522	1.451	1.491	359524	534491	1.4879	1.6522	0.1644	0.1087	-100.8	0	1.5911	1.6401	-0.0412	-0.0257	0
51	ADANA	3/17/2006	1.506	1.552	1.476	1.537	372148	565559	1.5217	1.6316	0.1099	0.1145	-63.97	1	1.5830	1.6323	-0.0445	-0.0332	0
52	ADANA	3/20/2006	1.537	1.623	1.522	1.567	419911	661326	1.5707	1.6141	0.0434	0.1156	-24.99	0	1.5806	1.6275	-0.0442	-0.0376	0
53	ADANA	3/21/2006	1.552	1.593	1.522	1.593	280781	441189	1.5690	1.5983	0.0292	0.1175	-16.58	0	1.5825	1.6248	-0.0392	-0.0382	0
54	ADANA	3/22/2006	1.567	1.593	1.537	1.583	153293	240916	1.5707	1.5815	0.0107	0.1168	-6.132	0	1.5825	1.6217	-0.0382	-0.0382	0
55	ADANA	3/23/2006	1.593	1.669	1.583	1.608	199130	322869	1.6198	1.5693	0.0505	0.1200	28.036	0	1.5863	1.6206	-0.0315	-0.0355	1
56	ADANA	3/24/2006	1.608	1.638	1.567	1.583	232320	371436	1.5961	1.5582	0.0379	0.1203	21.014	0	1.5857	1.6178	-0.0313	-0.0338	0
57	ADANA	3/27/2006	1.537	1.623	1.537	1.608	385753	616466	1.5893	1.5532	0.0361	0.1141	21.093	0	1.5891	1.6170	-0.0254	-0.0305	0
58	ADANA	3/28/2006	1.593	1.623	1.537	1.583	89186	140534	1.5809	1.5556	0.0252	0.1003	16.782	0	1.5881	1.6145	-0.0228	-0.0274	0
59	ADANA	3/29/2006	1.537	1.567	1.522	1.567	128596	198864	1.5521	1.5553	0.0031	0.0867	-2.417	0	1.5850	1.6109	-0.0246	-0.0263	0
60	ADANA	3/30/2006	1.583	1.623	1.567	1.593	206531	329567	1.5944	1.5586	0.0357	0.0759	31.398	0	1.5861	1.6096	-0.0233	-0.0251	0
61	ADANA	3/31/2006	1.593	1.669	1.583	1.608	662662	1073010	1.6198	1.5628	0.0570	0.0691	55.018	0	1.5894	1.6094	-0.0245	-0.0249	0
62	ADANA	4/3/2006	1.623	1.684	1.608	1.669	274812	454347	1.6536	1.5725	0.0810	0.0621	87.006	0	1.6013	1.6139	-0.0167	-0.0216	0

Sayfa1 Sayfa2 Sayfa3

Şekil 5.2: CCI ve MACD Göstergelerinin Excel Görüntüsü



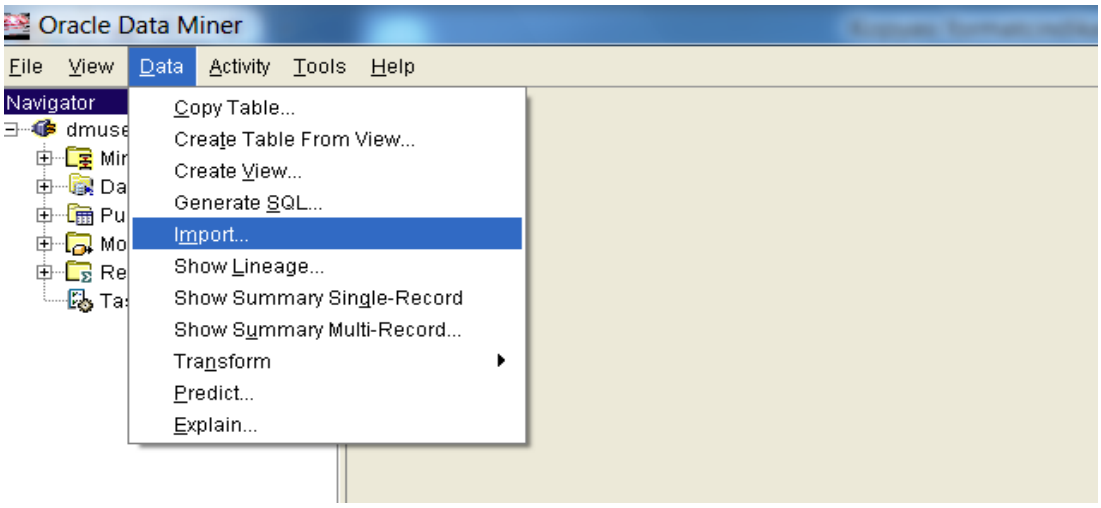
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	SYMBOL	DTYYYYMMDD	ID	ADI	CCI	MACD	AO	PO	BB	CMF	CMO	LRI	MG	P_ROC	RMI	RSI	SMI	SO	WS	WR	WAD	MFI	PO2	IMI	QI	NORM
2	ADANA	3/1/2006	38	1	0	0	0	0	0	0	0	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0
3	ADANA	3/2/2006	39	2	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2
4	ADANA	3/3/2006	40	0	0	0	0	2	0	0	0	0	0	2	2	0	0	0	2	0	0	0	0	0	0	2
5	ADANA	3/6/2006	41	0	0	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	0	2
6	ADANA	3/7/2006	42	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2
7	ADANA	3/8/2006	43	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
8	ADANA	3/9/2006	44	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0
9	ADANA	3/10/2006	45	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
10	ADANA	3/13/2006	46	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	2
11	ADANA	3/14/2006	47	2	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	2	0	2	0	0	2
12	ADANA	3/15/2006	48	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
13	ADANA	3/16/2006	49	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1
14	ADANA	3/17/2006	50	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	1
15	ADANA	3/20/2006	51	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
16	ADANA	3/21/2006	52	1	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	2	0	0	2
17	ADANA	3/22/2006	53	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	0	0	1	1
18	ADANA	3/23/2006	54	2	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	2
19	ADANA	3/24/2006	55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	2	0	0	0	2	1
20	ADANA	3/27/2006	56	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	2
21	ADANA	3/28/2006	57	0	0	0	0	0	0	0	0	2	0	2	0	0	0	0	2	0	2	0	0	0	0	2

Şekil 5.3: Gösterge Karar Verileri

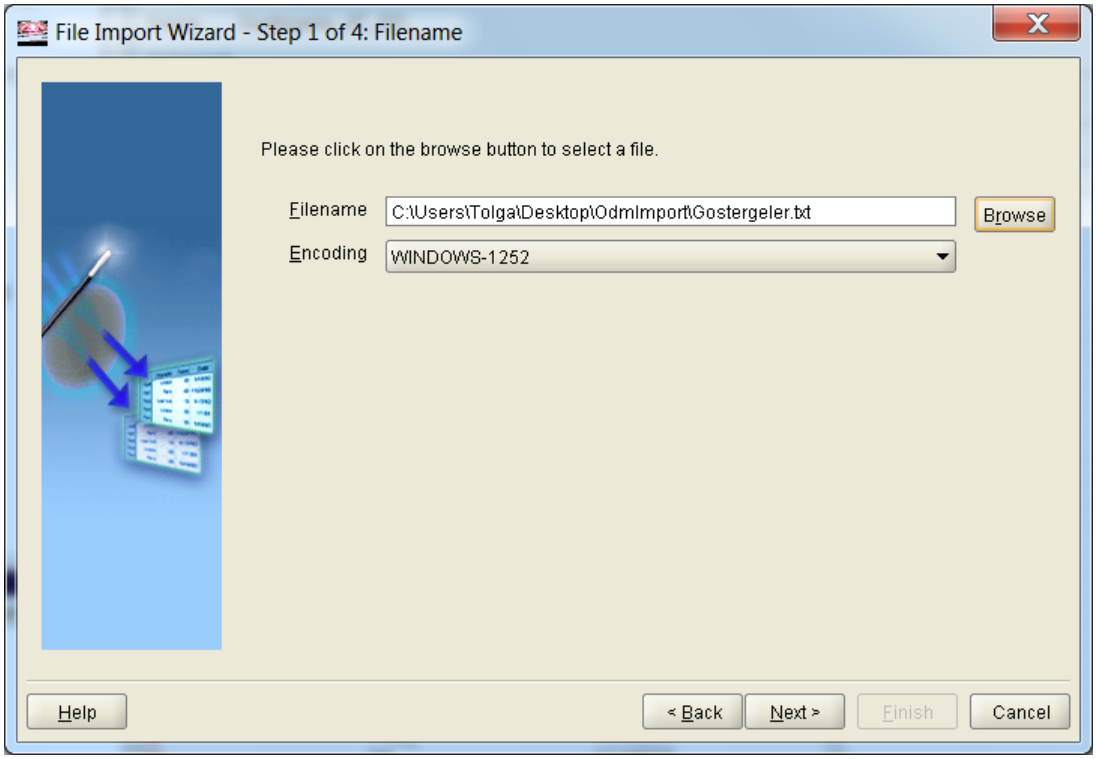
5.2 Model İçin Tablo Oluşturma

Veri tablosu üzerinde yapılan düzenlemelerden sonra problemin yapısının sınıflandırma modeline uygun olduğu belirlenmiş ve problem Oracle Data Miner arayüzünde "Classification" fonksiyon tipi seçilip, çözüm algoritması olarak da "Naivé Bayes" ve "Decision Tree" ile iki farklı yolla çözülmüştür.

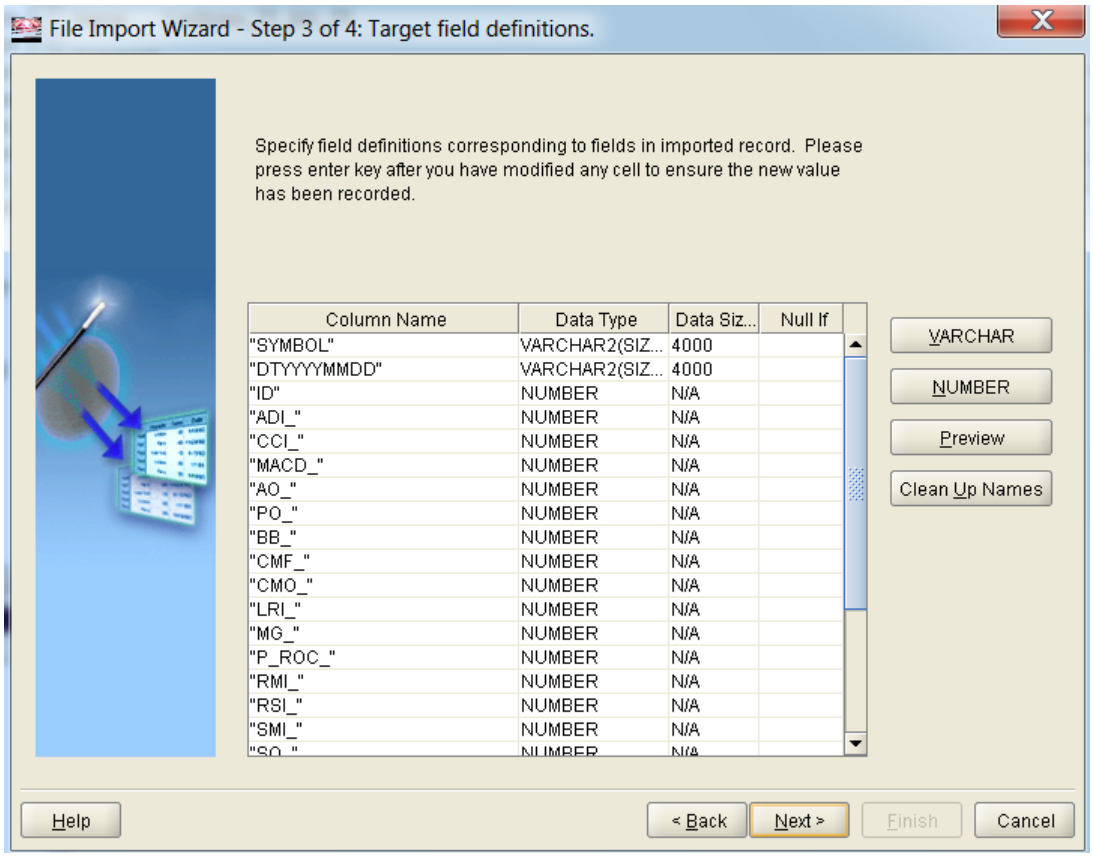
Veri tablosu oluşturulduktan sonra bu dosya metin dosyası haline getirilmiş ve kullanılacak program olan Oracle Data Miner'a aktarılmıştır. Bu aktarım ve tablo oluşturma aşamaları ekran çıktıları ile birlikte aşağıda verilmiştir.



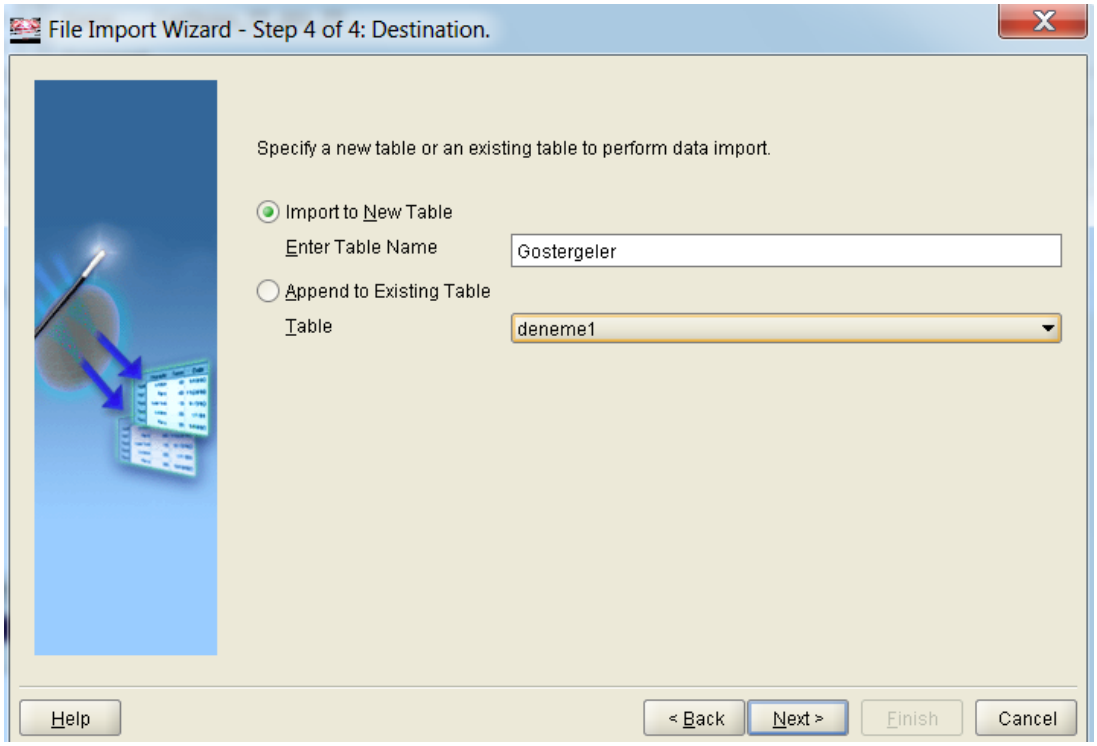
Şekil 5.4: Oracle Data Miner'a veri aktarımına başlangıç



Şekil 5.5: Aktarılacak dosyanın belirlenmesi



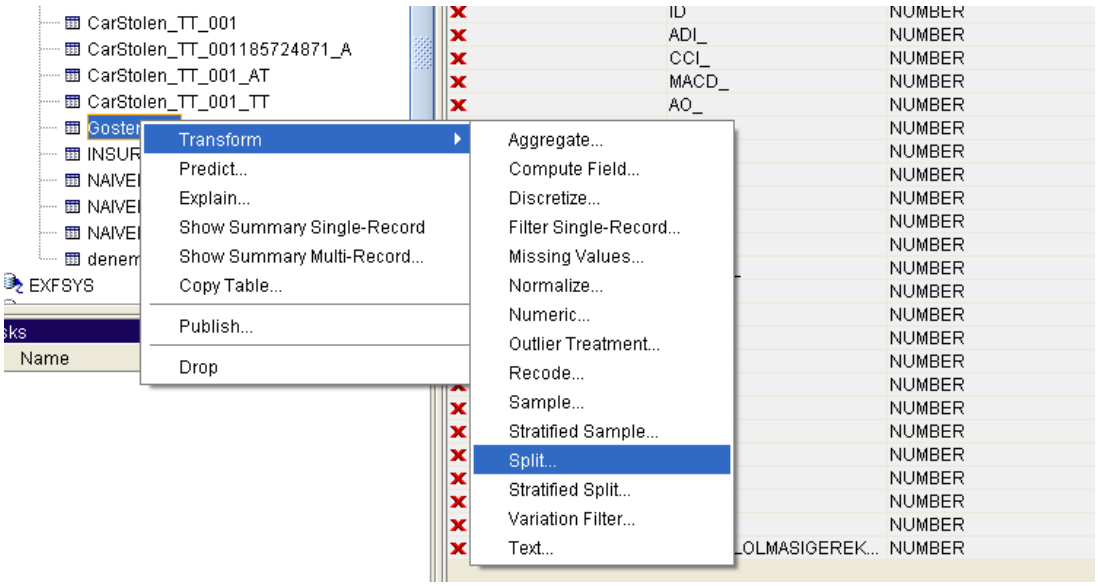
Şekil 5.6: Veri türünün belirlenmesi



Şekil 5.7: Tablo isminin belirlenmesi

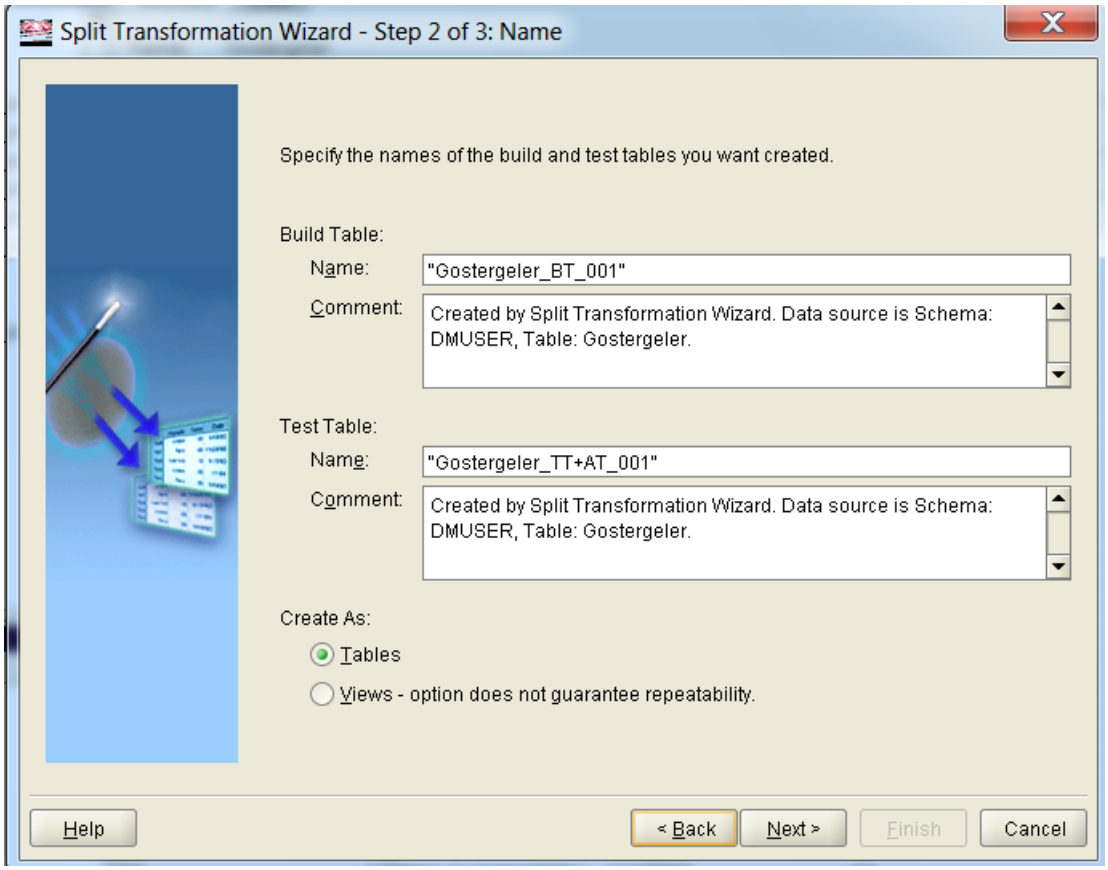
Aktarım işi tamamlandıktan sonra “Gostergeler” isminde bir tablo oluşturulmuştur. Model oluşturulmadan önce verinin bir kısmı model oluşturmak için, diğer kısmı oluşturulan modelin testi için, son kısmı da oluşturulan modelin tablodan seçilen herhangi bir veriye uygulanması için 3 farklı tabloya bölünmesi gerekmektedir. Bu işlemler aşağıda şekiller yardımı ile anlatıl maktadır.

İlk olarak “Gostergeler” tablosu Oracle Data Miner’ın ‘Transform /Split’ seçeneği ile iki farklı tabloya ayrılır.



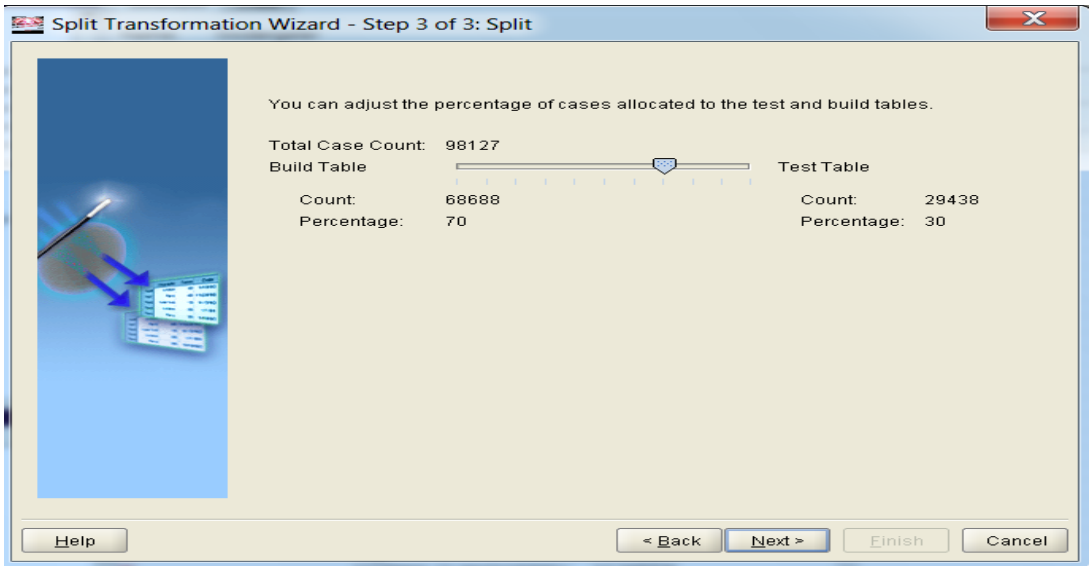
Şekil 5.8: Verinin bölünmesi

Bu adımdan sonra oluşturulacak olan yeni tabloların isimlendirilmesi gerekmektedir. İlk tablo ile model oluşturulmalıdır ve “Gostergeler_BT_001” olarak isimlendirilmiştir. İkinci tablo modelin test edileceği ve uygulama yapılacağı verileri içermektedir ve “Gostergeler_TT+AT_001” olarak isimlendirilmiştir.



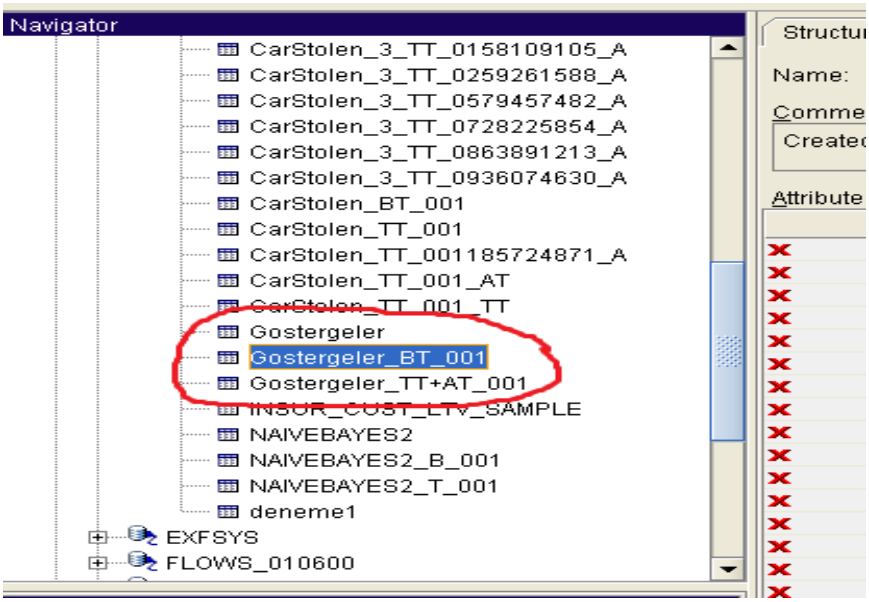
Şekil 5.9: Tablonun isimlendirilmesi

Tablolar isimlendirildikten sonra tabloların hangi oranda ayrılacağı belirlenmelidir. Verilerin %70'i model oluşturmak için %30'u ise test ve uygulama tabloları için ayrılmıştır.



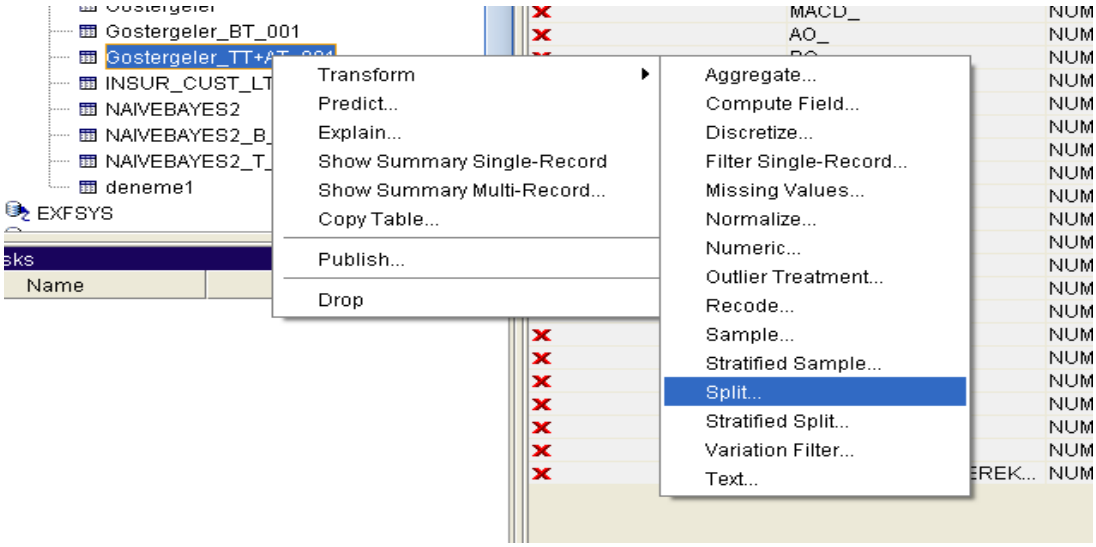
Şekil 5.10: Tabloların veri sayısı seçimi

Bu işlemin sonunda "Göstergeler" tablosu iki farklı tabloya ayrılmıştır. Bu tabloları Oracle Data Miner (ODM) da tablolar bölümünde kolayca görülebilir.

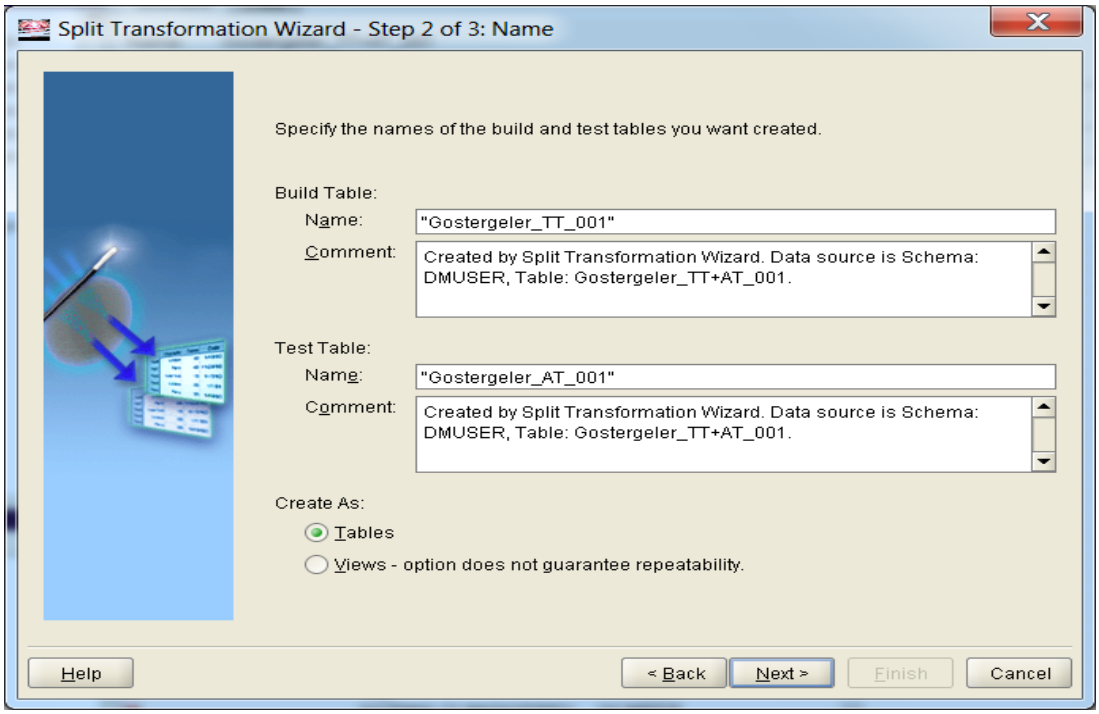


Şekil 5.11: Tabloların görünümü

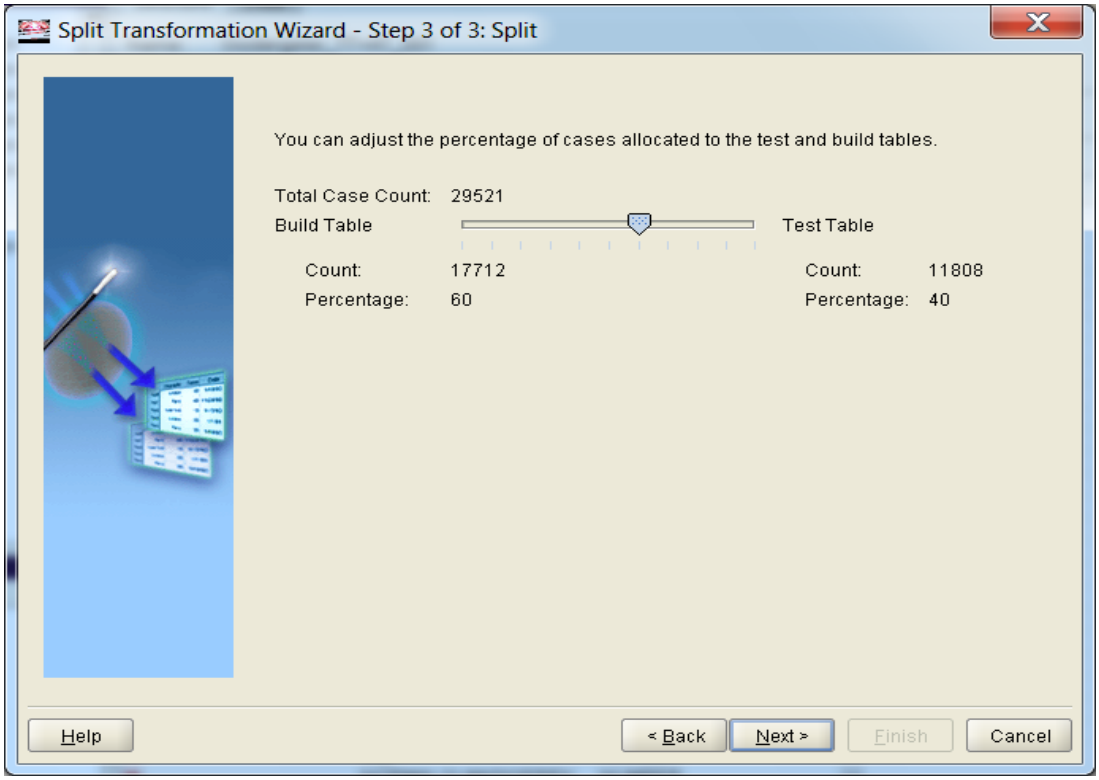
Bu aşamadan sonra "Göstergeler_BT_001" tablosu model oluşturmak ve "Göstergeler_TT+AT_001" tablosu ise modelin testi ve uygulamasında kullanılmak üzere iki farklı tabloya ayrılmıştır (Şekil 5.12).



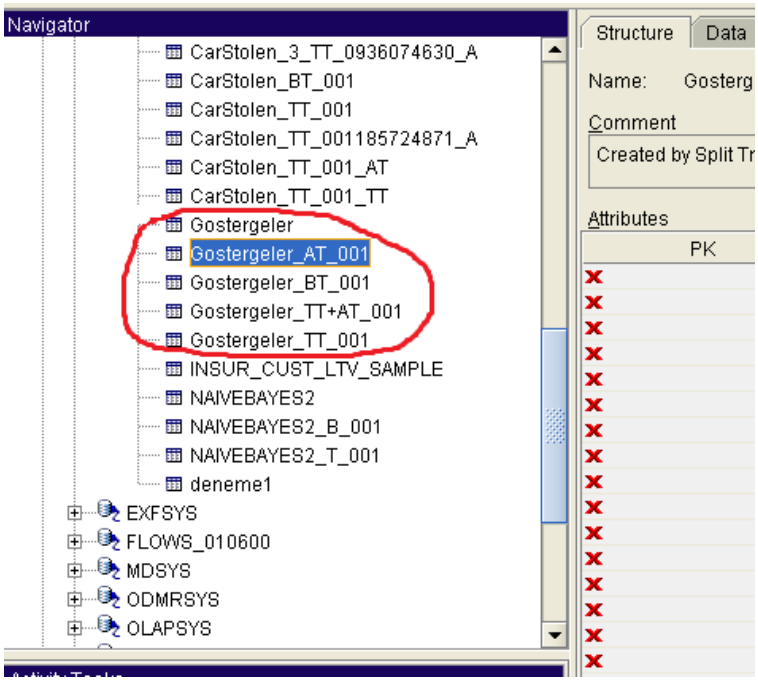
Şekil 5.12: Tabloların ayrılmaya başlanması



Şekil 5.13: Tabloların isimlendirilmesi



Şekil 5.14: Tabloların veri yoğunluğu seçimi



Şekil 5.15: Tabloların görünümü

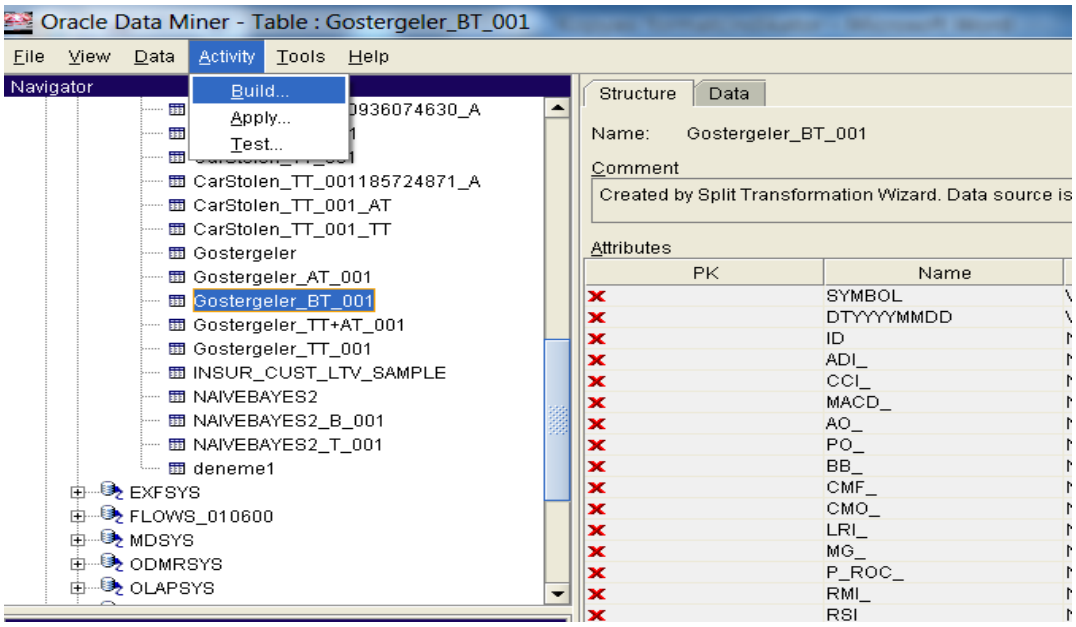
Yukarıda görüldüğü gibi "Gostergeler" tablosu ile birlikte toplam 5 tablo elde edilmiştir. Bunlardan "Gostergeler_BT_001" model oluşturmak için "Gostergeler_TT_001" oluşturulan modelin testi için "Gostergeler_AT_001" ise oluşturulan modelin uygulanması için kullanılmıştır.

5.3 Model Oluşturma

Model oluşturmak için iki farklı algoritma kullanılmıştır. Bunlardan ilki "Naivé Bayes" diğeri ise "Karar Ağaçları" dır.

5.3.1 Naivé Bayes ile Model Oluşturma

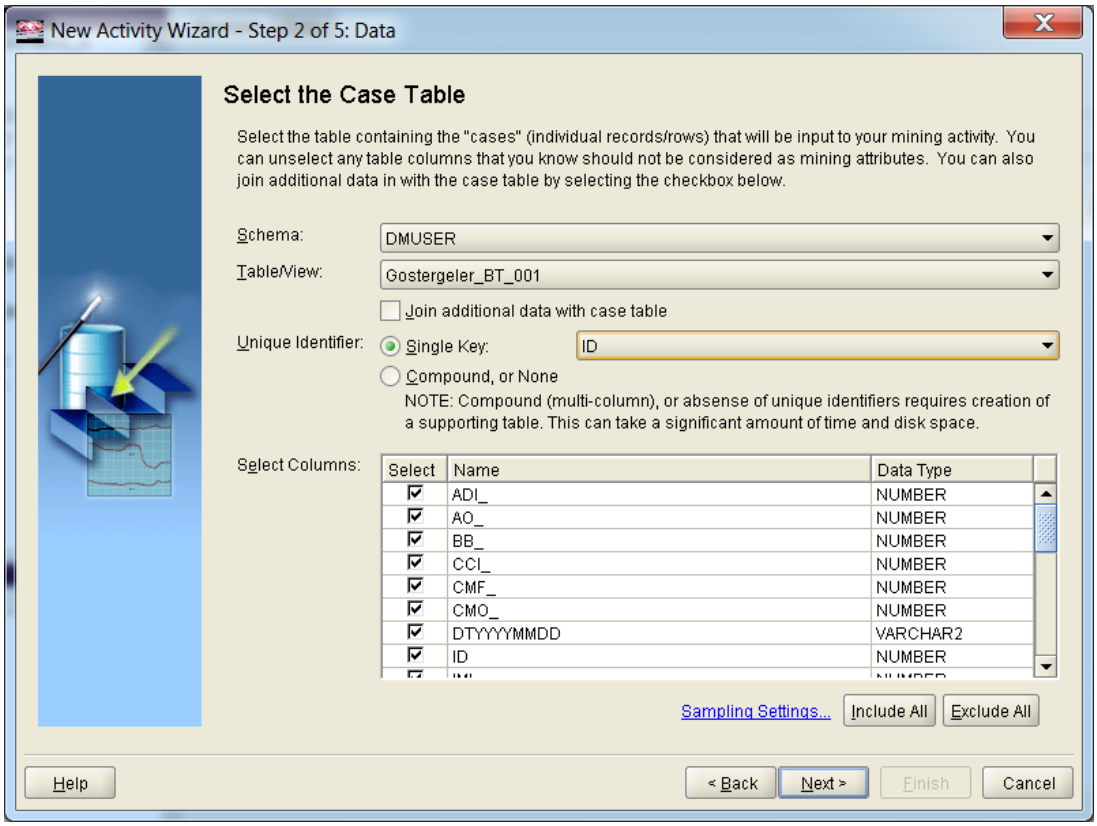
Model oluşturabilmek için ilk olarak "Oracle Data Miner" programından aşağıdaki şekilde görüldüğü gibi "Activity/Build" seçeneği seçilir.



Şekil 5.16 : Model oluşturma 1. aşama

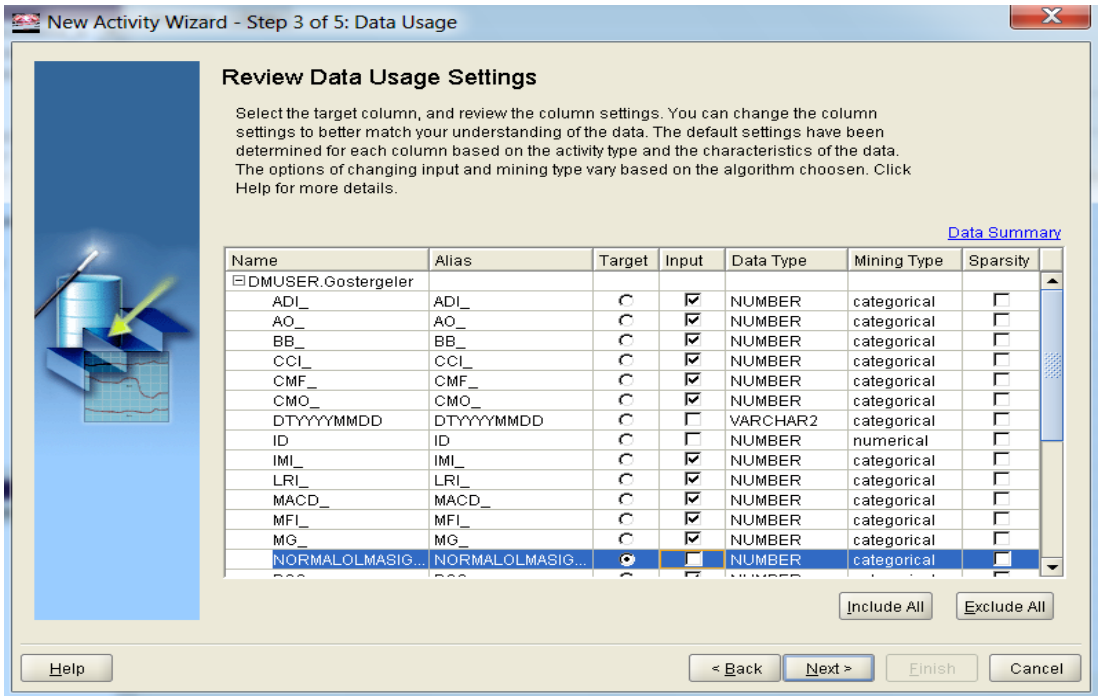
Daha sonra problem sınıflandırma problemi olduğundan çıkan ekranda "Function Type : Classification" ve "Algorithm : Naivé Bayes" seçilmiştir.

Şekil 5.17 : Model oluşturma, 2. aşama



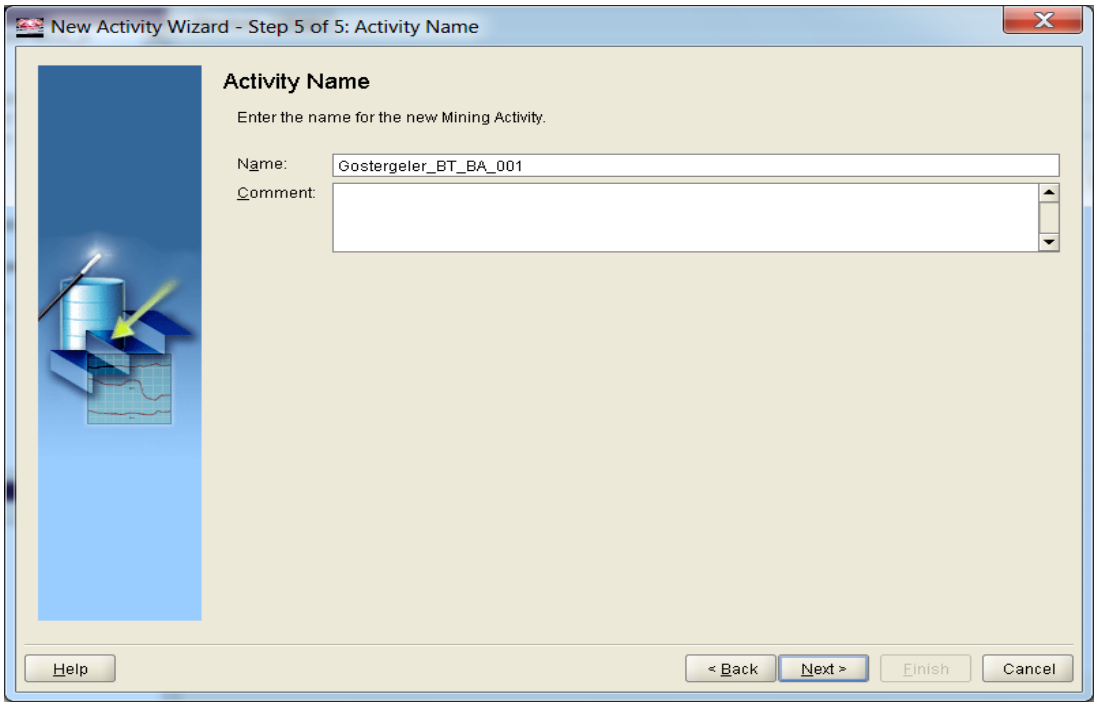
Şekil 5.18 : Model oluşturma, 3. aşama

Model “Gostergeler_BT_001” tablosu üzerinden oluşturulacağı için “Table/View: Gostergeler_BT_001” isimli tablo seçilmiştir. “Unique Identifier” kısmı zorunlu olup, tabloda kayıt ayırt edici sütun “Single Key”yani bu tabloda “ID” sütunu seçilmiştir. Bu modelde kullanılacak sütunlar "AO”, “BB”, “CCI”, “CMF”, “CMO”, “IMI”, “LRI”, “MACD”, “MFI”, “MG”, “NORMALOLMASIGEREKEN”, “PO2”, “PO”, “P_ROC”, “QI”, “RMI”, “RSI”, “SMI”, “SO”, “WAD”, “WR”, “WS”, olarak belirlenmiştir. Burada “DTYYYYMMDD”, “ID” ve “SYMBOL” sütunları sonucu etkilememeleri için modele dâhil edilmemiş fakat yorumlama aşamasında kullanılmak üzere tablodan çıkarılmamıştır.

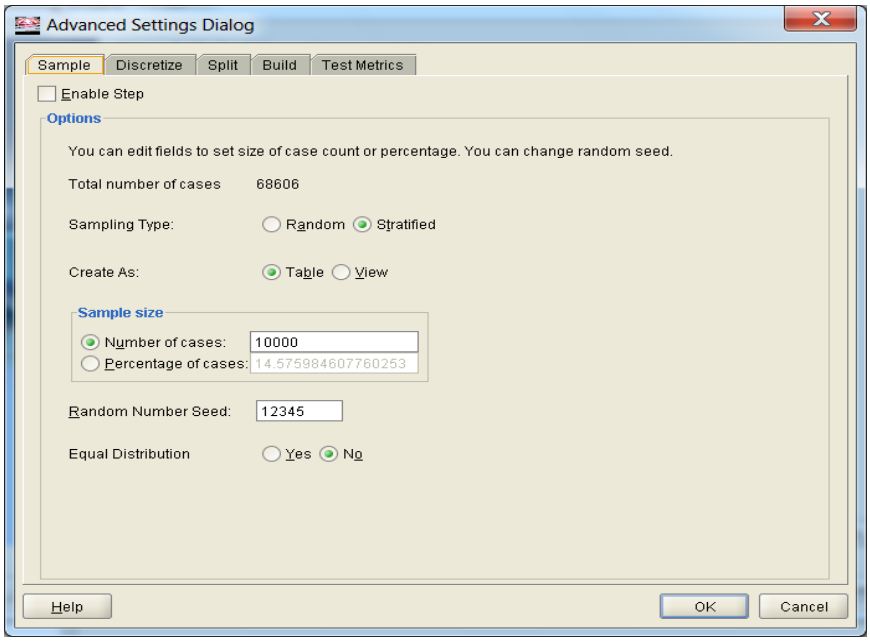


Şekil 5.19: Model oluşturma, 4. aşama

Hedef sütun olarak “NORMALOLMASIGEREKEN” seçilmiş ve tahmin edici özellik olarak seçilmemesi gereken “DTYYYYMMDD”, “ID”, “SYMBOL” ve “NORMALOLMASIGEREKEN” özellikleri girdi olarak alınmamıştır.

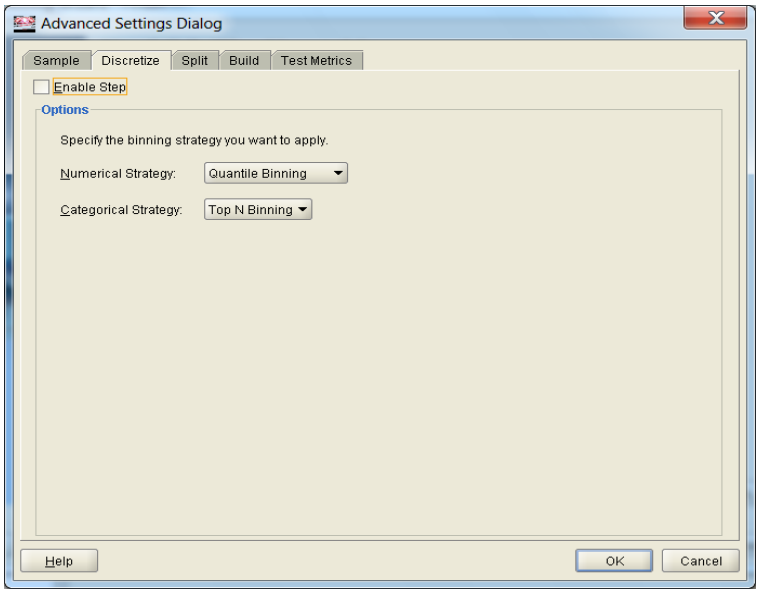


Şekil 5.20 : Modelin ismini belirleme



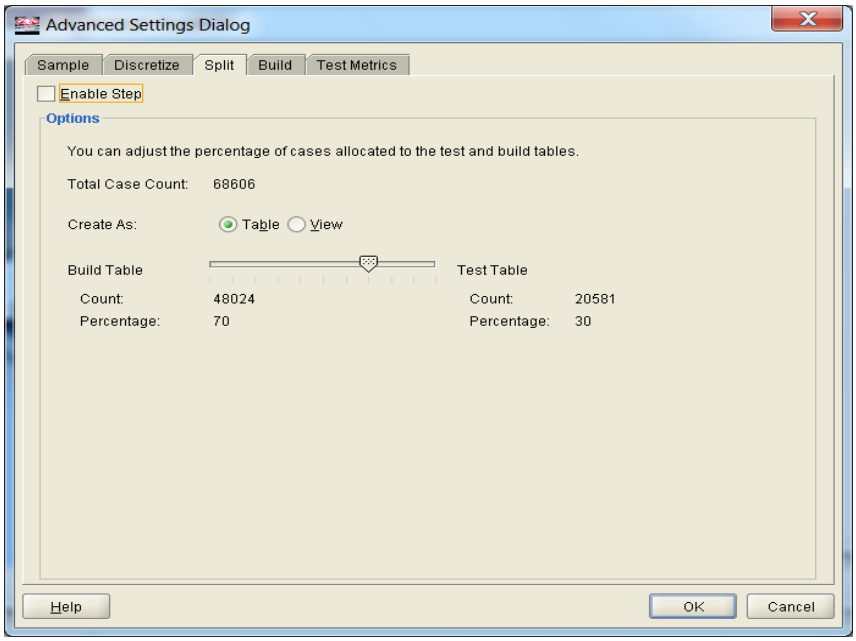
Şekil 5.21 : Gelişmiş ayarlar, “Sample” sekmesi

Son adımda yer alan “Advanced Settings” seçilerek model oluşturma aşamasında yürütülecek adımlar üzerinde gerekli ayarlar yapılmıştır. Tabloda çok fazla sayıda kayıt bulunduğu durumlarda kullanılan ve bu kayıtların içinden bir örneklem kümesi seçmeye yarayan “Sample” özelliği programda varsayılan olarak zaten seçili değildir ve aynı şekilde bırakılır.



Şekil 5.22 : Gelişmiş ayarlar, "Discretize" sekmesi

"Discretize" seçeneği sürekli verileri ayrık kümelere ayırmak için kullanılır. Bizim problemimizde sürekli veriler olmadığından verilerimizi sadece '0', '1' ve '2' den oluştuğundan bu özellik kullanılmayacaktır. Bu nedenle "Discretize" seçili kalmamasına dikkat edilmiştir.

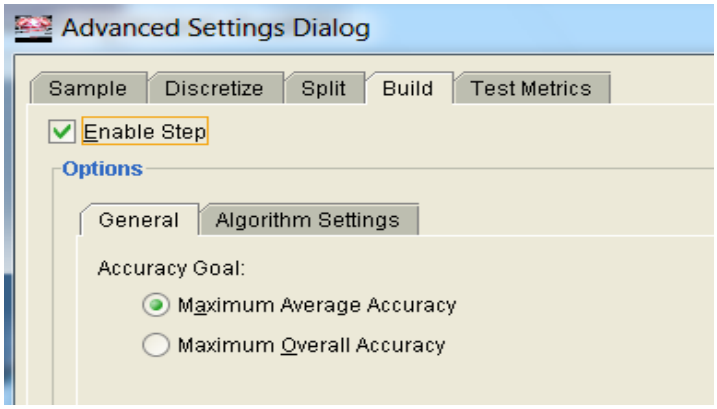


Şekil 5.23 : Gelişmiş ayarlar, "Split" sekmesi

"Split" sekmesi, test için ayrılacak veri yüzdesini ayarlamak içindir. 5.1 bölümünde veri tabloları hazırlanırken verilerin %70'i Build, %30'u Test ve Apply tablolarına, bunun sonrasında %30 luk kısım %60'ı Test ve %40'ı Apply uygulamalarında

kullanılmak üzere bölünmüştü. Bu nedenle bu bölümde "Split" özelliği kullanılmamıştır.

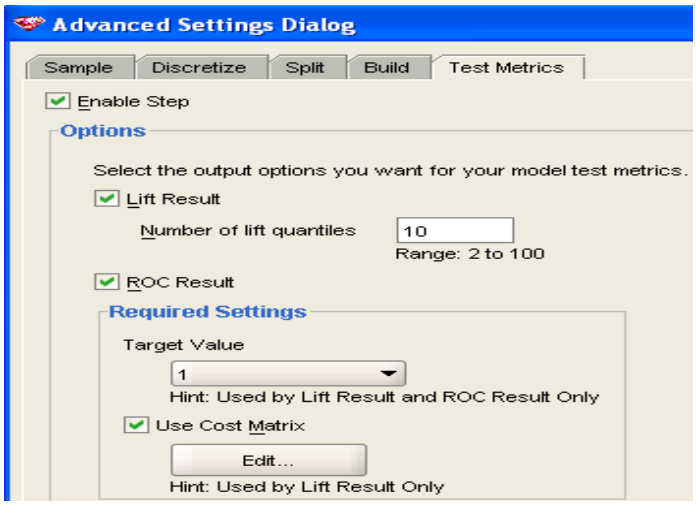
Yukarıdaki tanımlar çerçevesinde, bu uygulama için "Sample", "Discretize" ve "Split" seçeneklerinin kullanılmaması gerektiği açıktır ve bu nedenle modelde bunların seçili olmamasına dikkat edilmelidir.



Şekil 5.24: Gelişmiş ayarlar, "Build/General" sekmesi

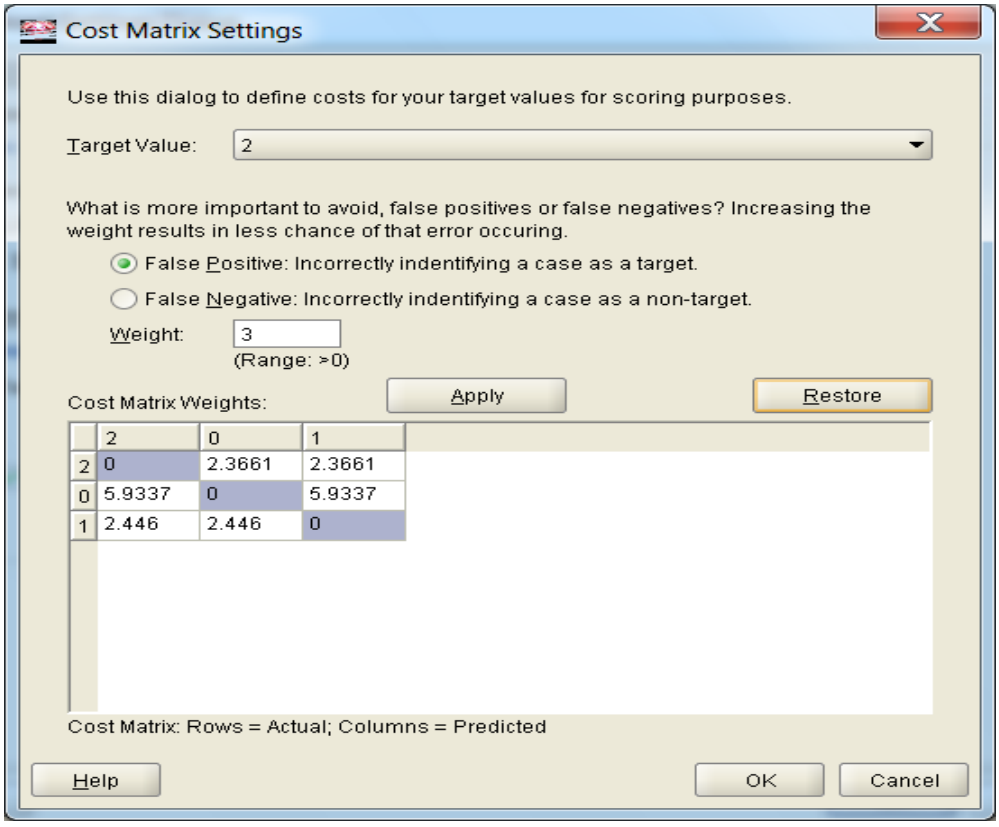
"Build" ayarları iki bölüme ayrılmıştır. "General" sekmesinde, modelin genel doğruluğu artırıcı yönde mi ("Maximum Overall Accuracy"), yoksa her bir hedef değer için yüksek tahminde bulunan ve ortalama doğruluğu artırıcı yönde mi ("Maximum Average Accuracy") olacağı seçilir. Açıktır ki modelin hedef olarak seçilen sütundaki her bir değeri yüksek doğrulukla tahmin etmesi istenecektir. Bu sebeple "Maximum Average Accuracy" seçeneği model kurulurken varsayılan olarak seçilidir ve o şekilde bırakılmıştır [11].

Bölüm 4.2' te anlatılan ve modelin arka planda gerçekleştirdiği "Attribute Importance" fonksiyonuna göre önem sıralaması yapılan sütunlar tahminlemeye katılmaya başlanır. Her bir özellik tahminlemeye katıldığında bir önceki adımla karşılaştırılır ve modelin doğruluğunda bir artış yapıp yapmadığına bakılır. Bu şekilde devam ederek yeni eklenen özelliğin modelin doğruluğunda bir artış yapmadığı görülünce model tamamlanmış olur [11]. Belirtmek gerekir ki bütün bu süreci, ODM arka planda gerçekleştirir, yani kullanıcıya sadece hangi model tipini seçeceğine karar vermek kalır.

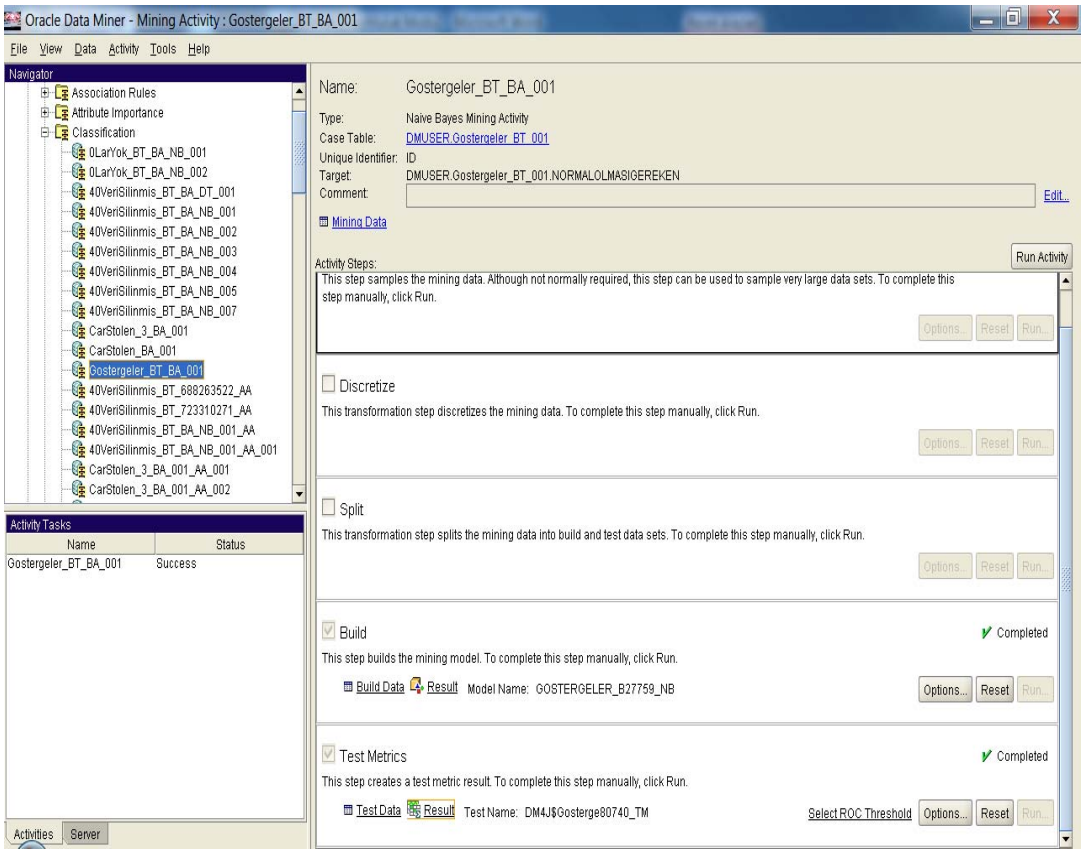


Şekil 5.25 : Gelişmiş ayarlar, "Test Metrics" sekmesi

"Test Metrics" sekmesinde bulunan "Cost Matrix" seçeneği de seçilmelidir, çünkü burada tahmin edilecek olan NORMALOLMASIGEREKEN sütununda '1' ve '2' değerlerinin '0' a göre daha iyi tahmin edilmesi istenmektedir. "Use Cost Matrix" seçeneğinin altındaki "Edit" tuşuna basılarak modelde kullanılan cost matrix Şekil 5.23b'deki gibi görüntülenebilir.



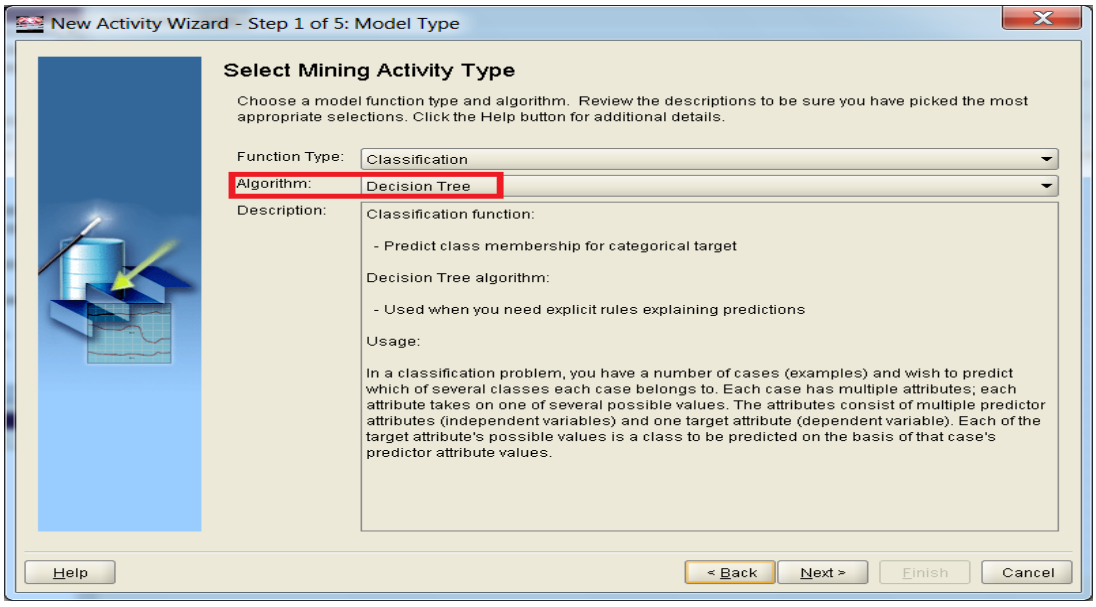
Şekil 5.26 : "Cost Matrix"in görüntülenmesi



Şekil 5.27 : Model oluşturma sürecinin son hali

5.3.2 Karar Ağaçları ile Model Oluşturma

Karar ağaçları ile model oluşturma Naivé Bayes ile model oluşturmaktan birkaç farkla ayrılık gösterir. Bunlardan ilki Şekil 5.16 gösterilmiş olan algoritma seçimi bölümünde "Algortihm=Decision Tree" olarak seçilmesidir.



Şekil 5.28 : Model oluşturmak için algoritma seçimi

Diğer fark ise Şekil 5.20 de gösterilmiş olan "Gelişmiş Ayarlar" menüsünde "Discretize" sekmesinin bulunmamasıdır.

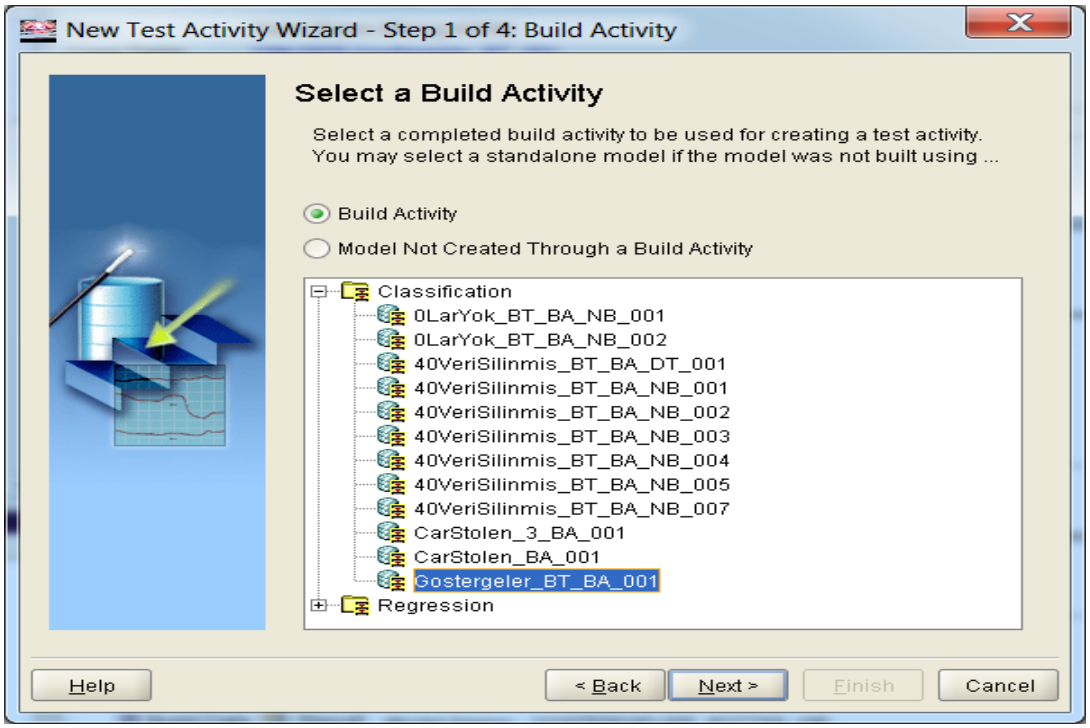
Naivé Bayes yönteminde olduğu gibi Karar Ağaçları algoritmasında da "Sample" ve "Split" seçeneklerinin etkinliği kaldırılır.

5.4 Modelin Testi

Burada ayrılmış olan test verisine, oluşturulan model uygulanmakta ve tahmin edilen değerler gerçek değerlerle karşılaştırılarak, kurulan model için bir güvenilirlik değeri hesaplanmaktadır.

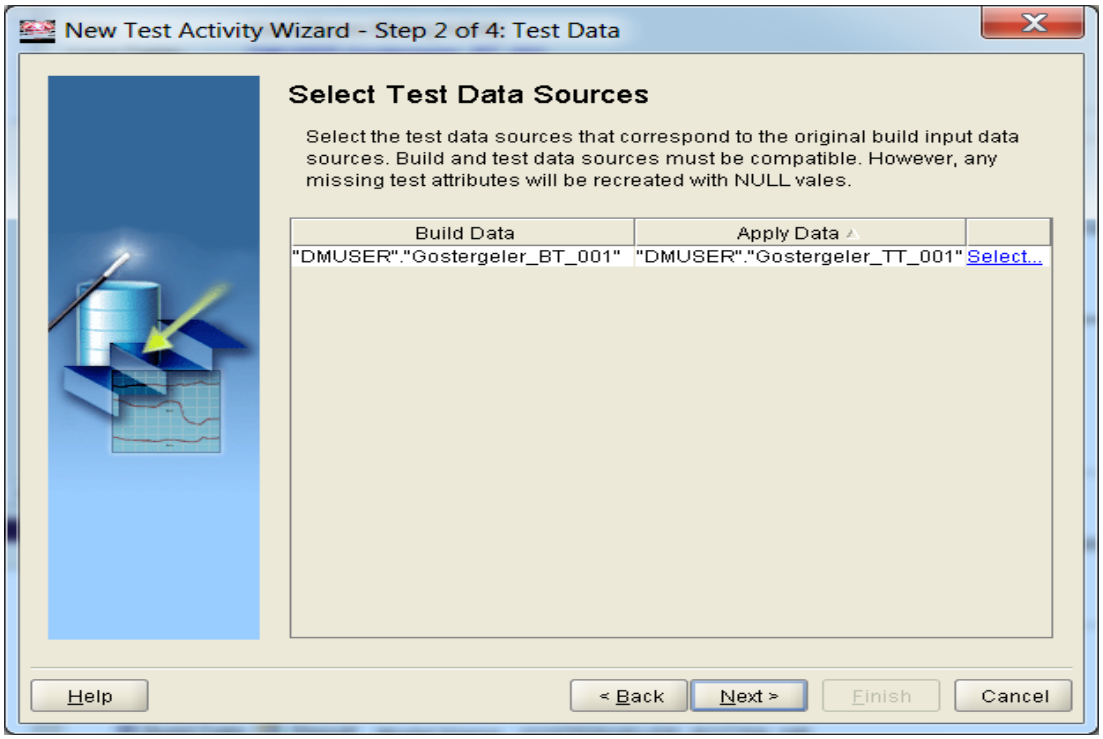
5.4.1 Navié Bayes ile Modelin Testi

Naivé Bayes ile oluşturulan model şekillerle aşağıdaki gibi anlatılmıştır.



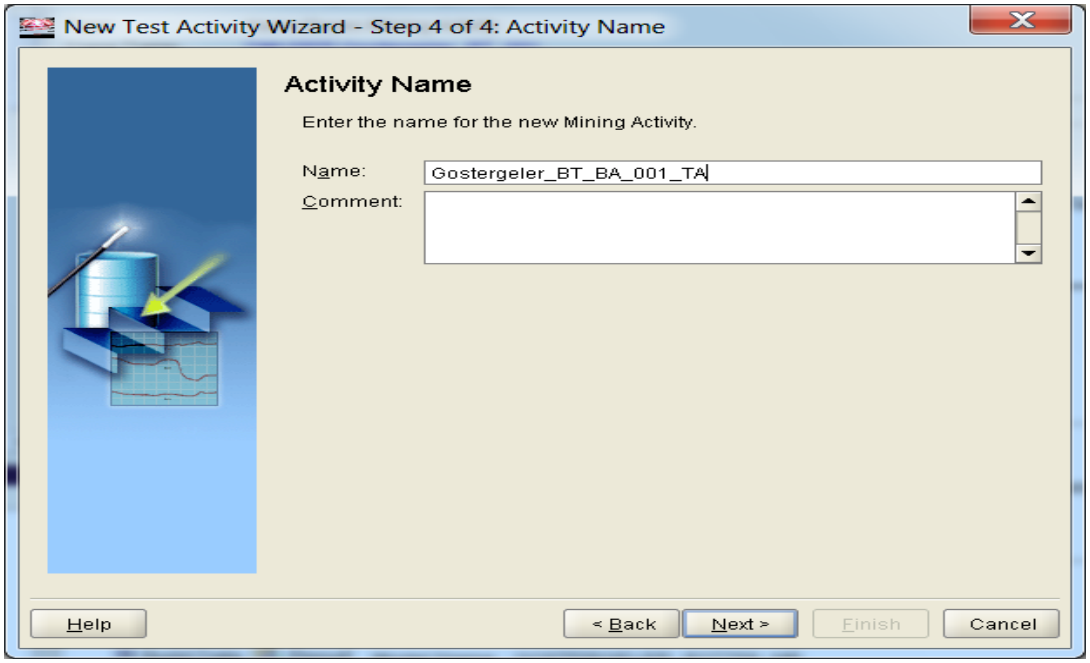
Şekil 5.29 : Test aktivitesi 1. aşama, test edilecek modelin seçimi

İlk aşamada, bir önceki bölümde oluşturmuş olduğumuz model test edilmek üzere seçilmiştir.



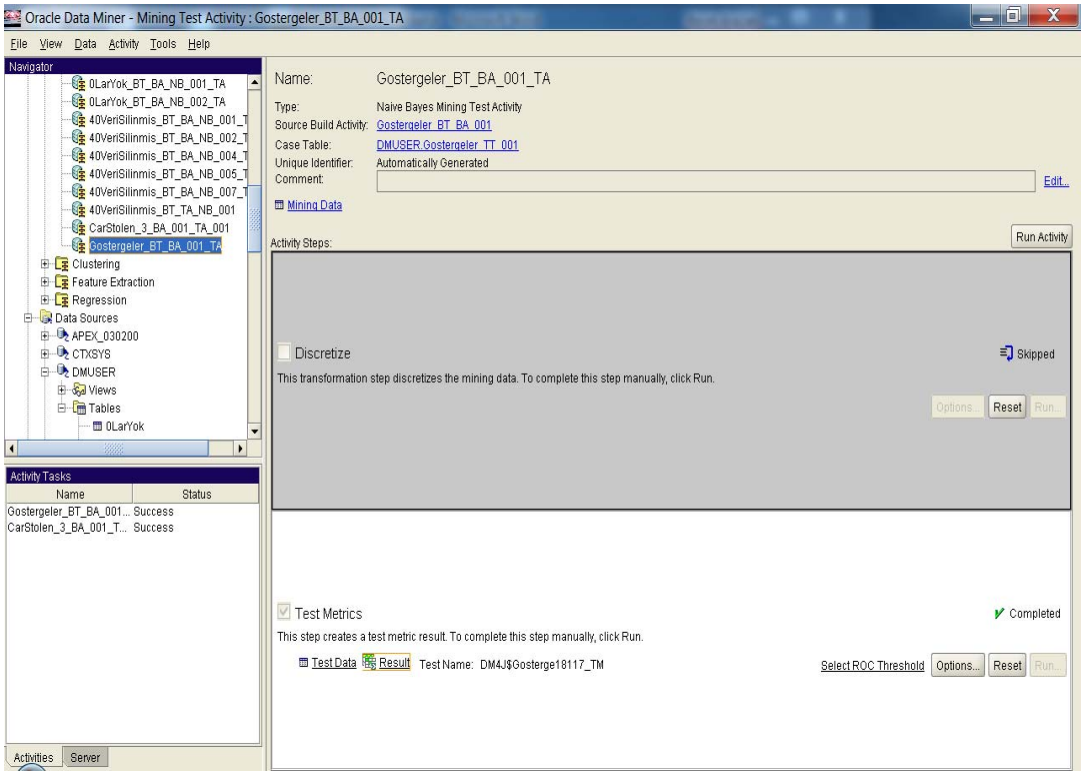
Şekil 5.30: Test aktivitesi 2. aşama, testin uygulanacağı tablonun seçimi

Sonraki aşamada ise oluşturulan modelin, üzerinde test yapılacak tablo olarak "Gostergeler_TT_001" belirlenmiştir. Bu tablo daha önceden "Transform/Split" seçeneği ile oluşturulmuştu.



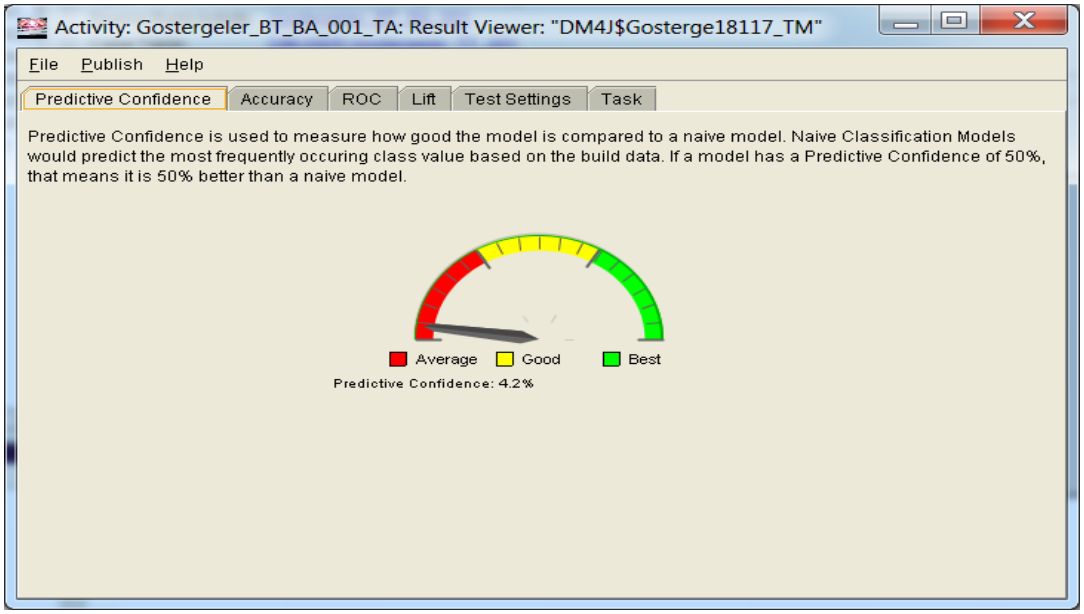
Şekil 5.31 : Test aktivitesi 3. aşama, aktivitenin isminin belirlenmesi

Test aktivitesinin tamamlanmış hali aşağıdaki gibidir.



Şekil 5.32 : Test aktivitesinin son hali

Bu ekranda, “Test Metrics” bölümünün altında yer alan “Result” a tıklanarak test sonuçları aşağıdaki gibi görüntülenebilir.

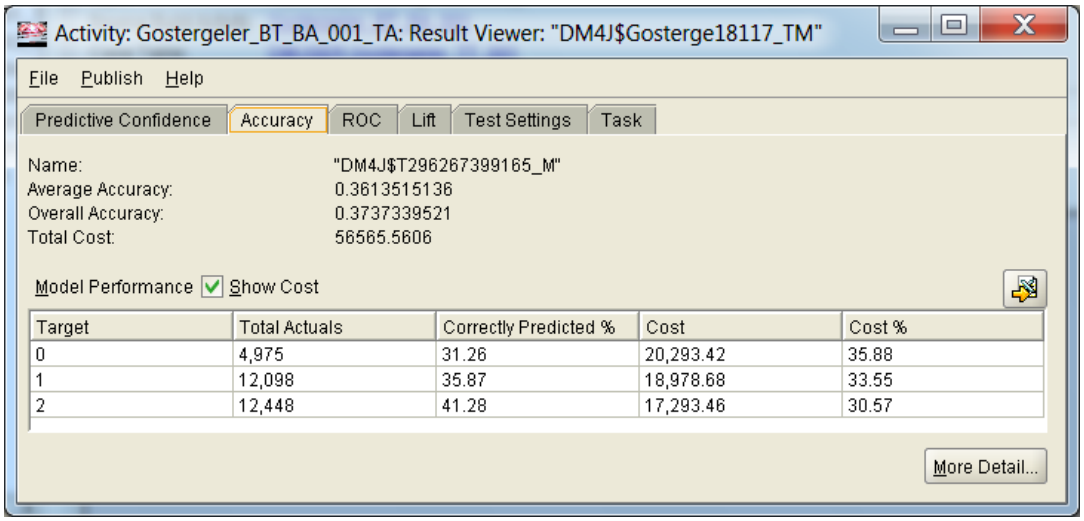


Şekil 5.33: Testin sonucu, “Predictive Confidance” sekmesi

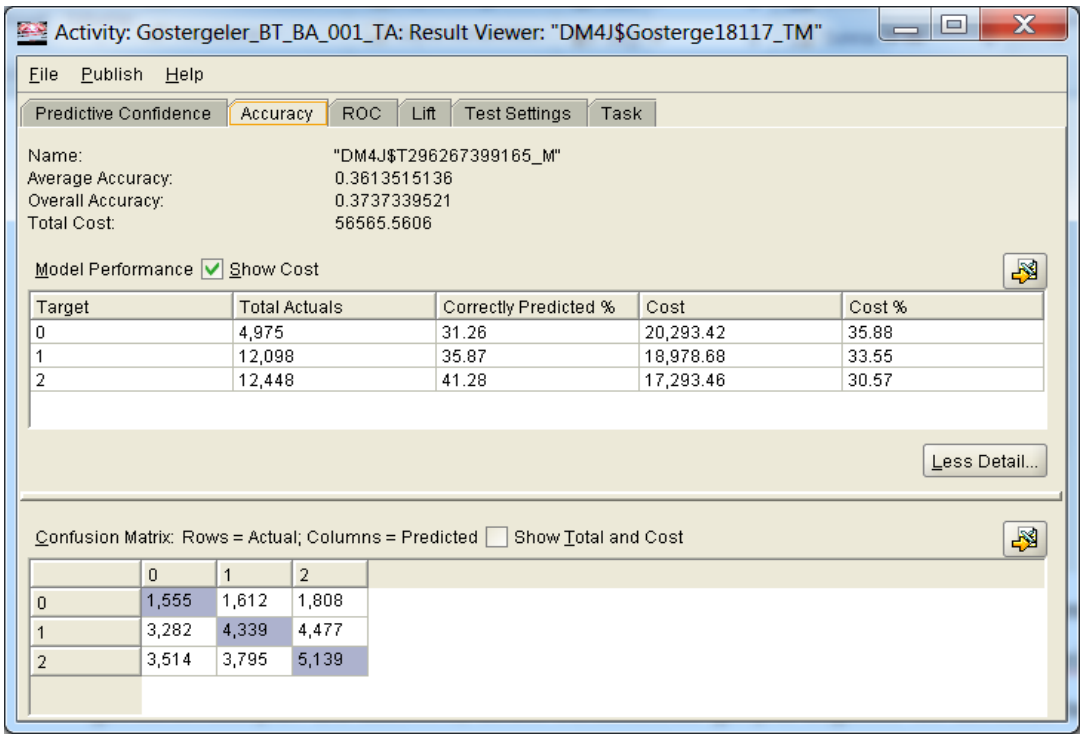
Şekil 5.25’te sonuç kısmının ilk sekmesi olan “Predictive Confidence” görüntülenmiştir. Bu sonuç modelin, bir insanın rastgele yapacağı tahmine karşı ne kadar etkili olduğunun görsel halidir.

Şekil 5.34a: Testin sonucu, “Accuracy” sekmesi

“Accuracy” sekmesine tıklandığında, modelin test tablosuna uygulandığında elde edilen sonucun doğruluk oranları görüntülenir. Tahmin edilmek istenen hedef sütunundaki gerçek değerler bilinmektedir, böylece modelin yaptığı tahminler gerçek değerlerle karşılaştırılabilir. Modelde, “NORMALOLMASIGEREKEN” değeri ‘0’ olan 4975 veri vardır ve model bunların %31.26’sını doğru olarak tahmin etmiştir. “NORMALOLMASIGEREKEN” değeri ‘1’ olan 12098 veri vardır ve model bunların %35.87’sini doğru olarak tahmin etmiştir. Aynı şekilde “NORMALOLMASIGEREKEN” değeri ‘2’ olan 12448 veriden %41.28’inin tahmini doğru olarak yapılmıştır. “Show Cost” butonuna tıklandığında, yanlış tahminin modele vereceği zarar aşağıdaki gibi görüntülenir. Düşük “Cost” değeri, modelin başarılı olduğu anlamına gelir [11].



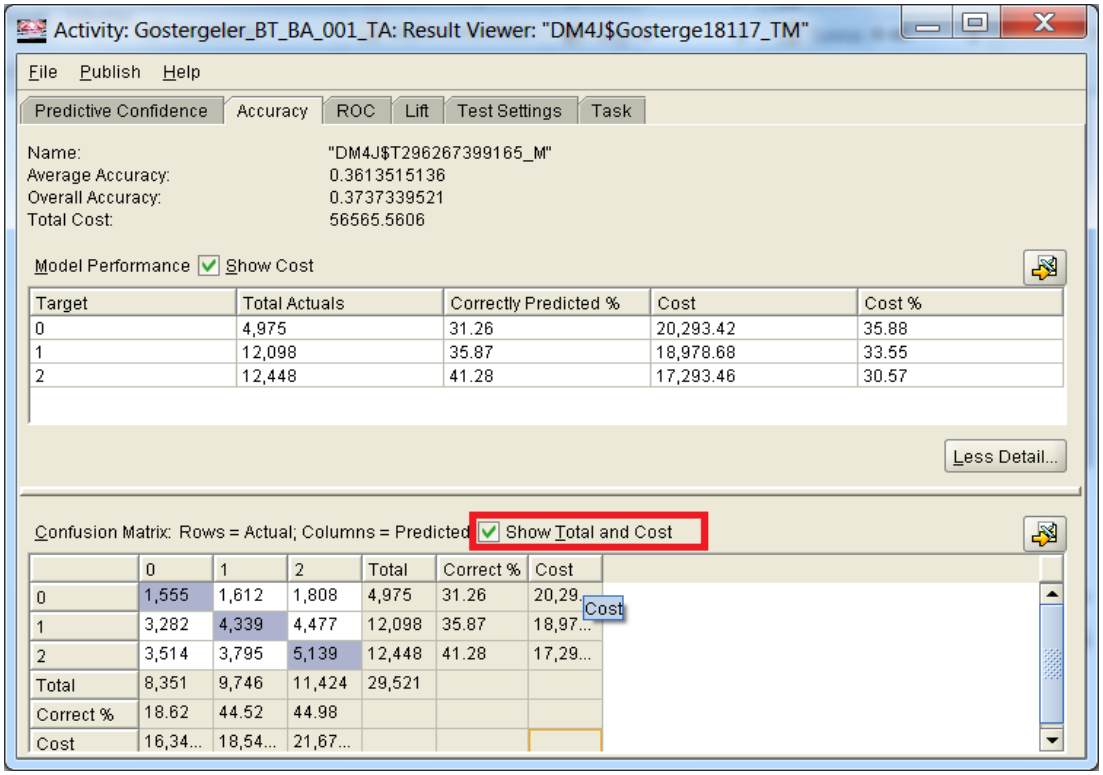
Şekil 5.34b: Testin sonucu, "Accuracy"de cost'un görüntülenmesi



Şekil 5.34c: Testin sonucu, "Accuracy"de güvenilirlik matrisinin görüntülenmesi

Sonraki adımda "More Detail" a tıklanarak güvenilirlik matrisi (Confusion Matrix) görüntülenmiştir. Bu matriste, hedef sütunundaki gerçek değerler ile modelin test tablosuna uygulanarak yapılan tahmin değerlerinin sayısı gösterilmektedir. Test tablosundaki "NORMALOLMASIGEREKEN" in gerçek değerleri bilinmektedir ve bu değerler güvenilirlik matrisinin satırlarındaki değerlerdir. Matrisin sütunları ise modelin yapmış olduğu tahminleri göstermektedir. Örneğin, matrisin sol alt köşesinde yer alan 3514 sayısı gerçek değer '2' iken '0' şeklinde tahmin edilen veri sayısını gösterir. Aynı şekilde matrisin orta solunda yer alan 3282 sayısı gerçek değer

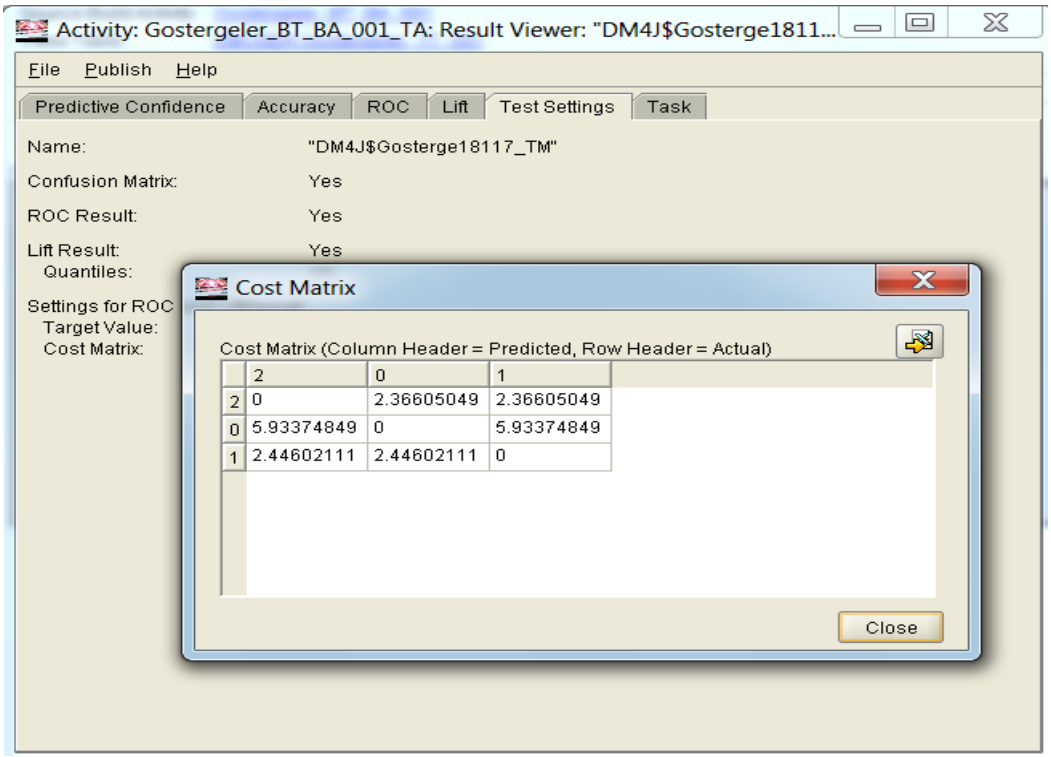
‘1’ iken ‘0’ olarak tahmin edilen veri sayısını vermektedir. Buna benzer olarak sağ üst köşede bulunan 1808 sayısı ise gerçek değer ‘0’ iken ‘2’ şeklinde tahmin edilen veri sayısını göstermektedir [11]. Matrisin köşegenindeki sayılar ise doğru olarak yapılan tahmin sayılarıdır.



Şekil 5.34d: Testin sonucu, “Accuracy”de detaylı güvenilirlik matrisinin görüntülenmesi

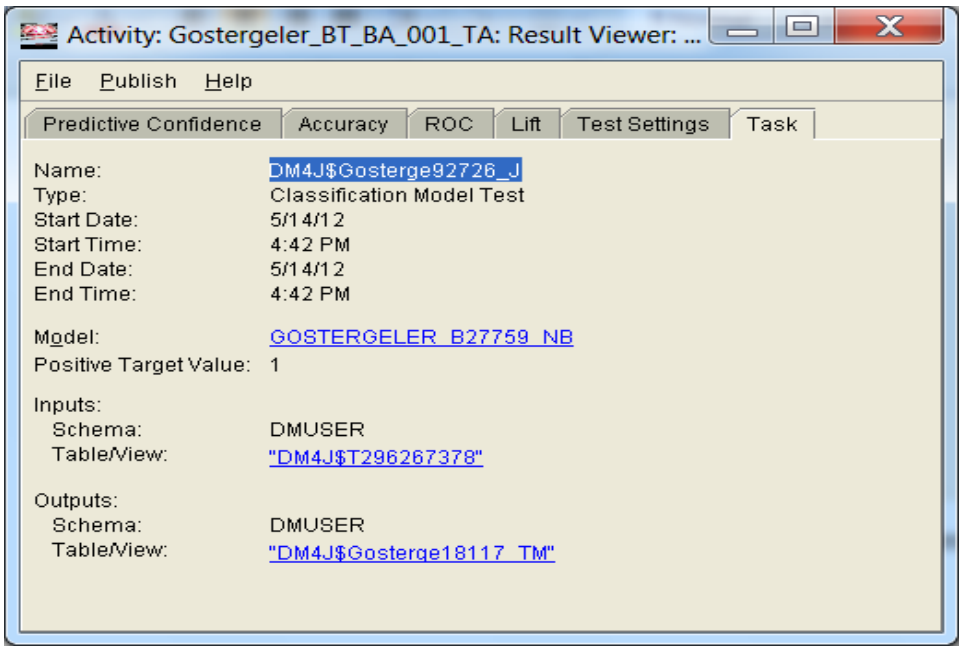
Son olarak, “Show Total and Cost”a tıklanarak güvenilirlik matrisinden elde edilen, kayıt sayıları, doğru tahmin yüzdesi, yanlış tahminin modele vereceği zarar gibi istatistiksel bilgiler görüntülenmiştir [11].

“Test Settings” bölümünde, Şekil 5.29’da görüldüğü gibi modelle ilgili bilgiler yer almaktadır. Buradan cost matrix de görüntülenebilir.



Şekil 5.35 : Testin sonucu, "Test Settings" sekmesi

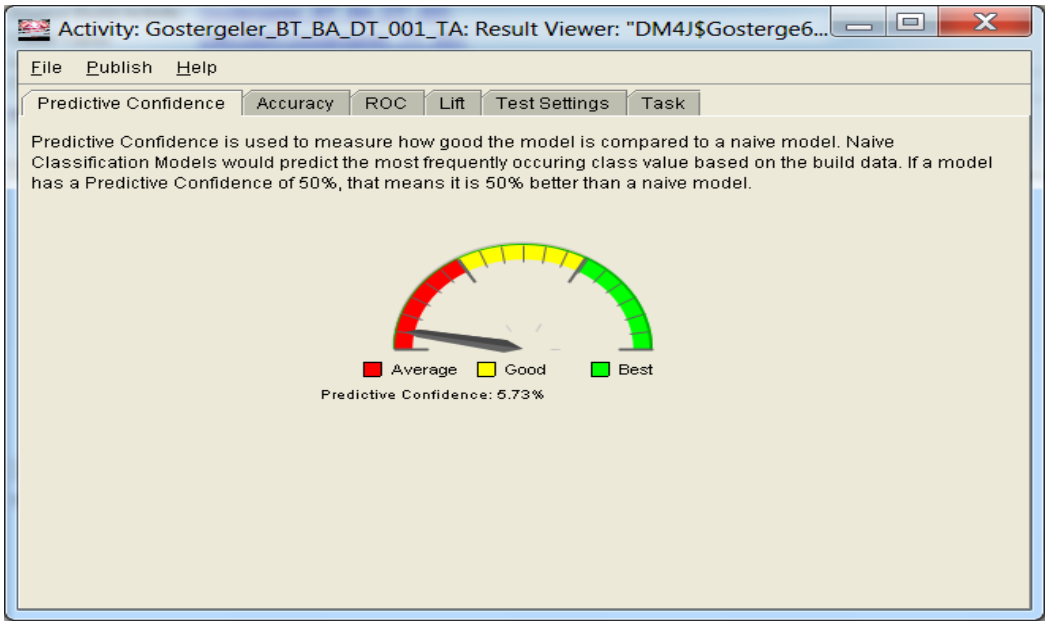
"Task" sekmesinde modelin adı, oluşturulma tarihi, kullanılan tablo gibi bilgiler aşağıdaki gibi görüntülenmektedir.



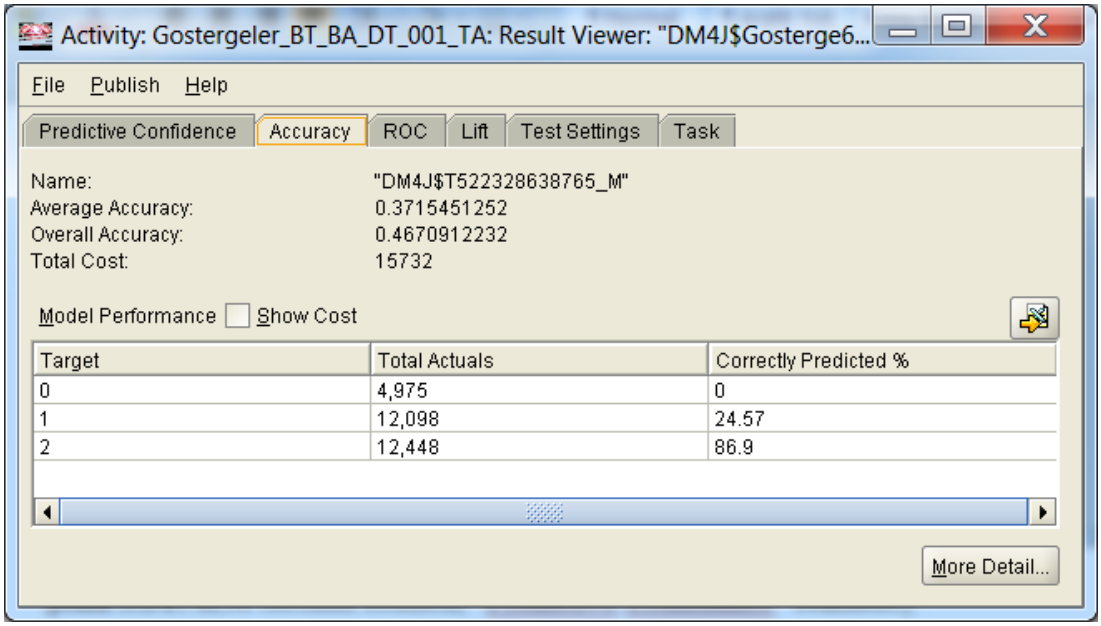
Şekil 5.36 : Testin sonucu, "Task" sekmesi.

5.4.2 Karar Ağaçları ile Modelin Testi

Karar Ağaçları ile modelin testi aynıdır ama algoritma farklılıklarından dolayı elde edilen sonuçlar farklılık gösterir. Bu farklılıklar "Test Metrics" bölümünün altında yer alan "Result" a tıklayarak görüntülenebilir.

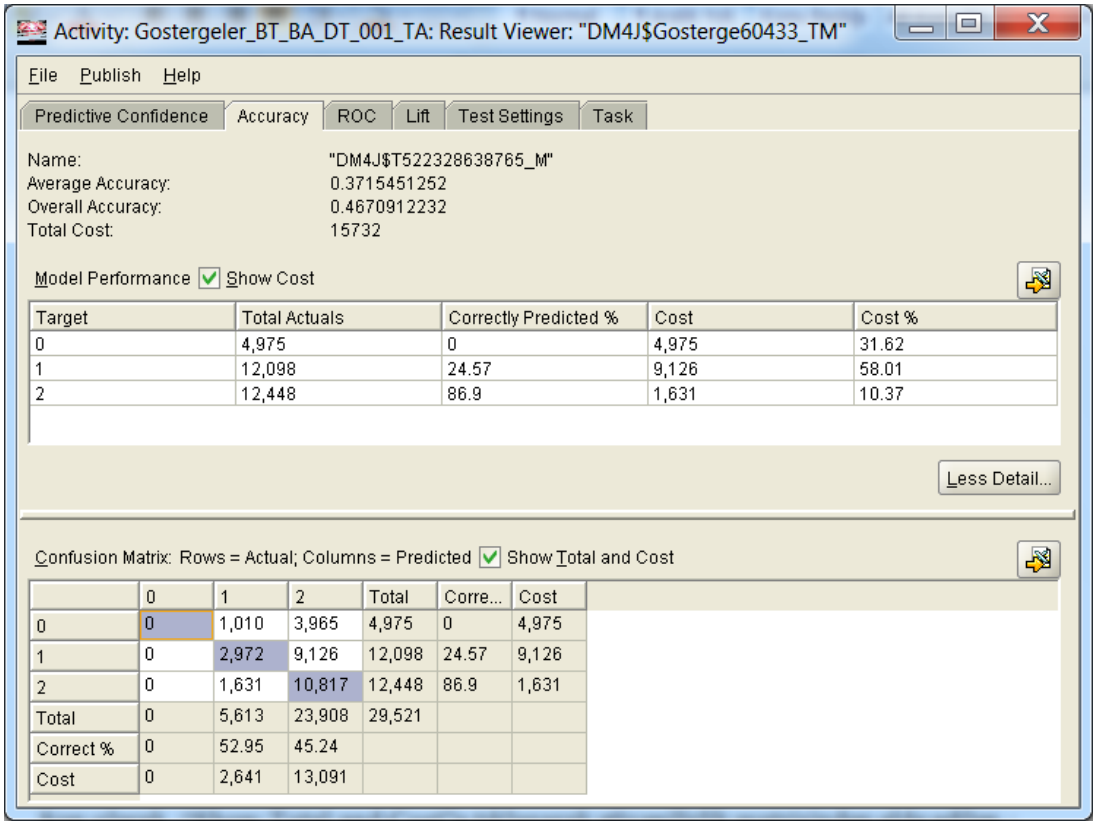


Şekil 5.37: K.A testinin sonucu, "Predictive Confidance" sekmesi



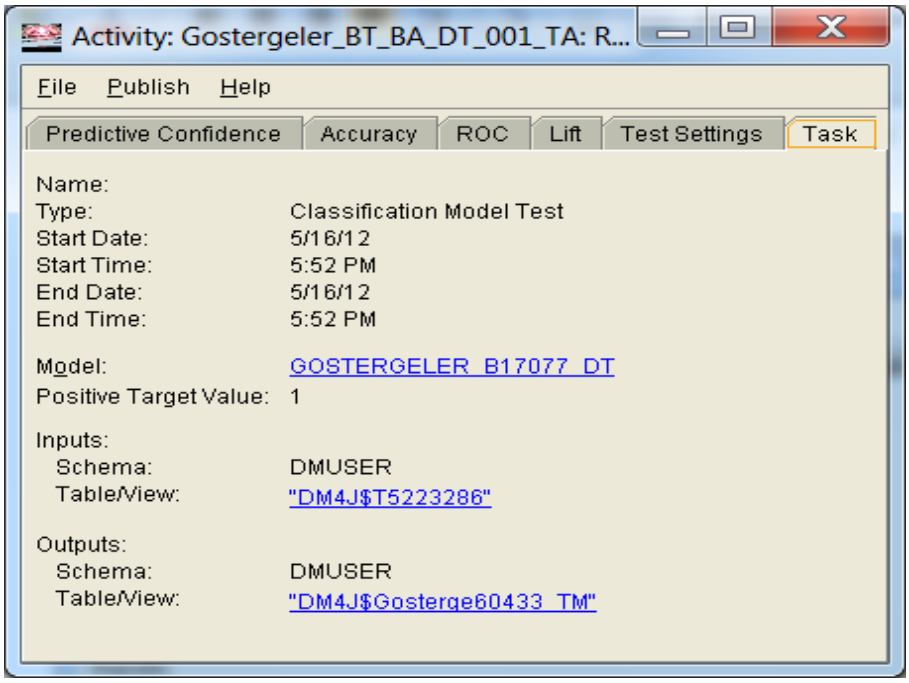
Şekil 5.38: K.A testinin sonucu, "Accuracy" sekmesi

“Accuracy” sekmesine tıklandığında, modelin test tablosuna uygulandığında elde edilen sonucun doğruluk oranları görüntülenir. Tahmin edilmek istenen hedef sütunundaki gerçek değerler bilinmektedir, böylece modelin yaptığı tahminler gerçek değerlerle karşılaştırılabilir. Modelde, “NORMALOLMASIGEREKEN” değeri ‘0’ olan 4975 veri vardır ve model bunların %0’sını doğru olarak tahmin etmiştir. “NORMALOLMASIGEREKEN” değeri ‘1’ olan 12098 veri vardır ve model bunların %24.57’sini doğru olarak tahmin etmiştir. Aynı şekilde “NORMALOLMASIGEREKEN” değeri ‘2’ olan 12448 veriden %86.90’ inin tahmini doğru olarak yapılmıştır. “Show Cost” butonuna tıklandığında, yanlış tahminin modele vereceği zarar aşağıdaki gibi görüntülenir. Düşük “Cost” değeri, modelin başarılı olduğu anlamına gelir [11].



Şekil 5.39: K.A testinin sonucu, “Accuracy”de detaylı güvenilirlik matrisinin görüntülenmesi

“Task” sekmesinde modelin adı, oluşturulma tarihi, kullanılan tablo gibi bilgiler aşağıdaki gibi görüntülenmektedir.

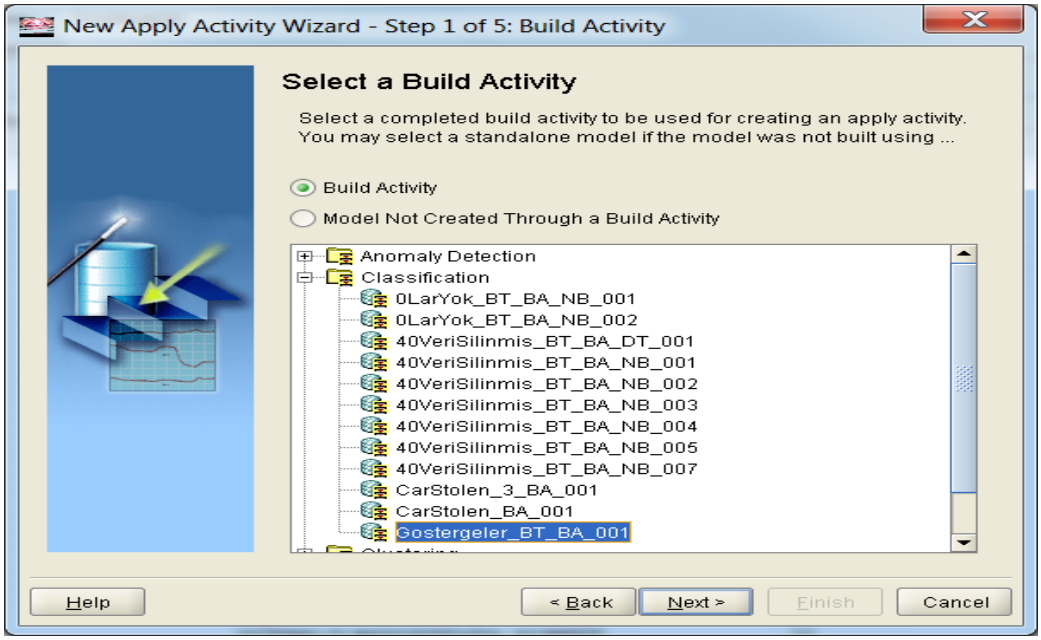


Şekil 5.40 : K.A. testinin sonucu, "Task" sekmesi

5.4 Modelin Uygulanması

Bu aşamada önceki adımda oluşturulan modelin, "Gostergeler" tablosundan uygulama yapmak amacıyla bölünen "Gostergeler_AT_001" tablosuna uygulanması ayrıntılı olarak anlatılmaktadır.

5.4.1 Naivé Bayes ile Modelin Uygulanması

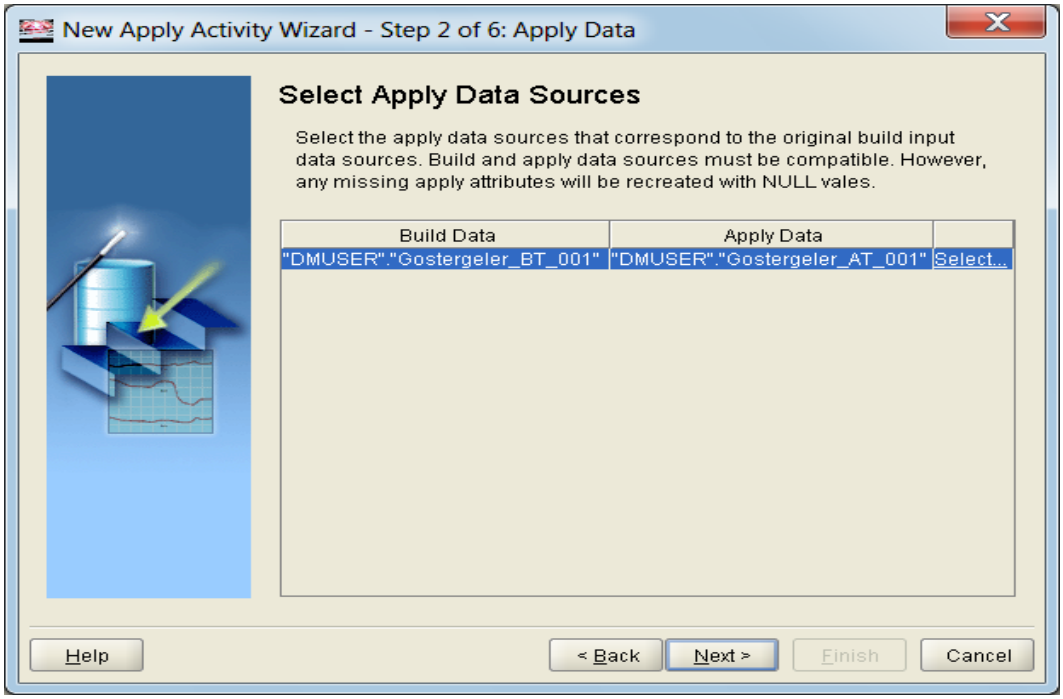


Şekil 5.41 : Uygulama 1. aşama, uygulama yapılacak modelin seçimi

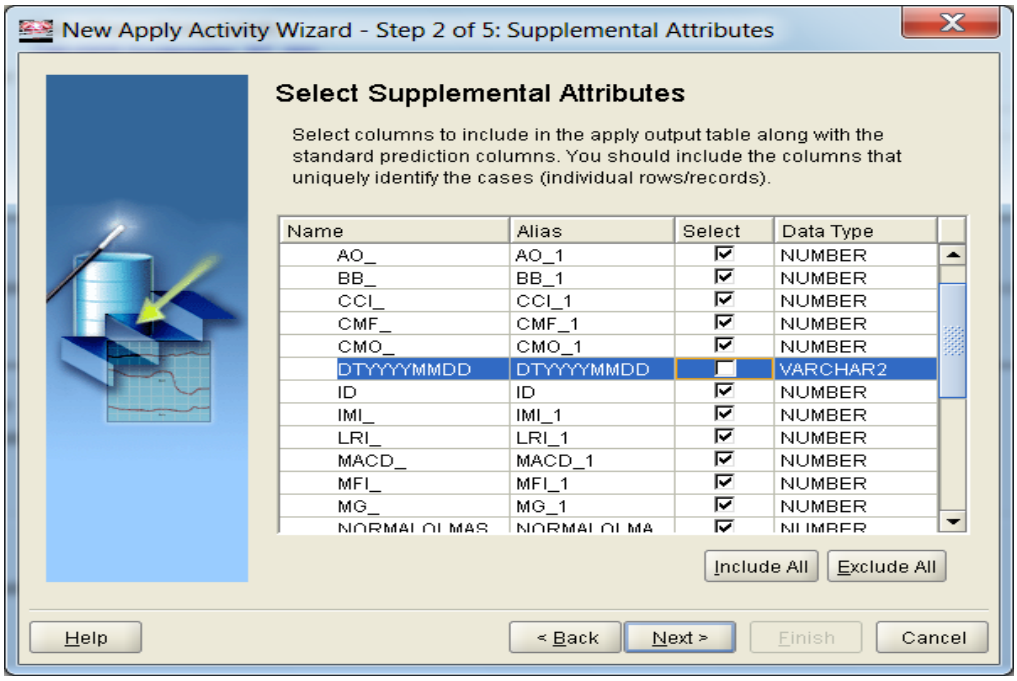
Uygulama için listelenen modellerden, bir önceki aşamada oluşturulan "Gostergeler_BT_BA_001" modeli seçilmiştir ve bir sonraki adımda modelin uygulanacağı tablo Şekil5.32a'daki gibi görüntülenmiştir.

Şekil 5.42a: Uygulama 2. aşama, uygulama yapılacak tablonun seçimi

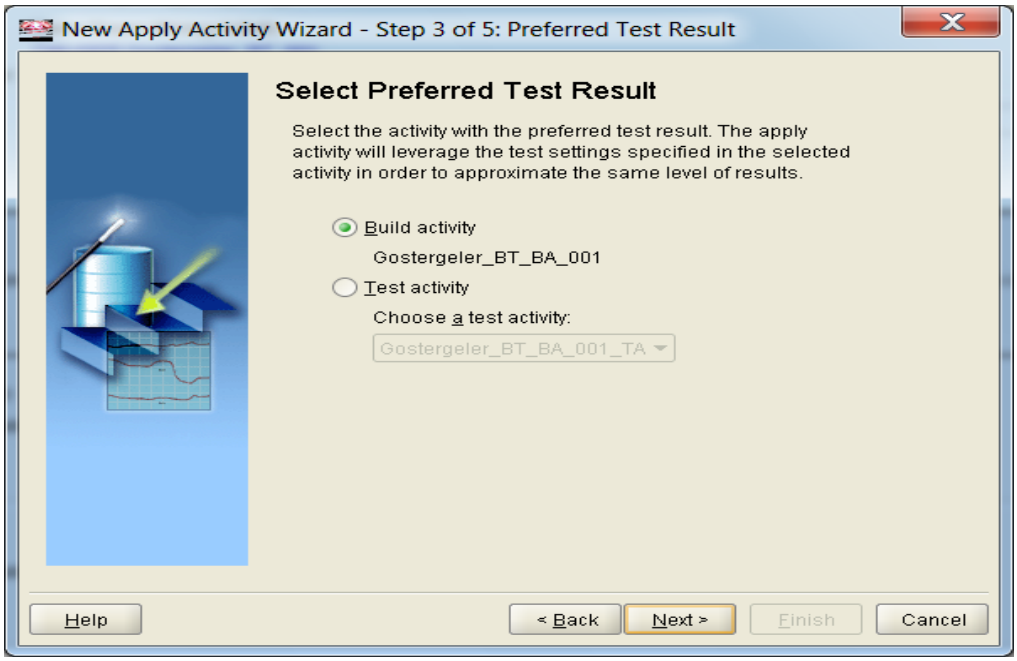
Uygulama sonucunda görüntülenmek istenen sütunlar seçilerek bir sonraki adıma geçilmiştir.



Şekil 5.42b: Seçili tablonun "Apply Data" altında gösterimi

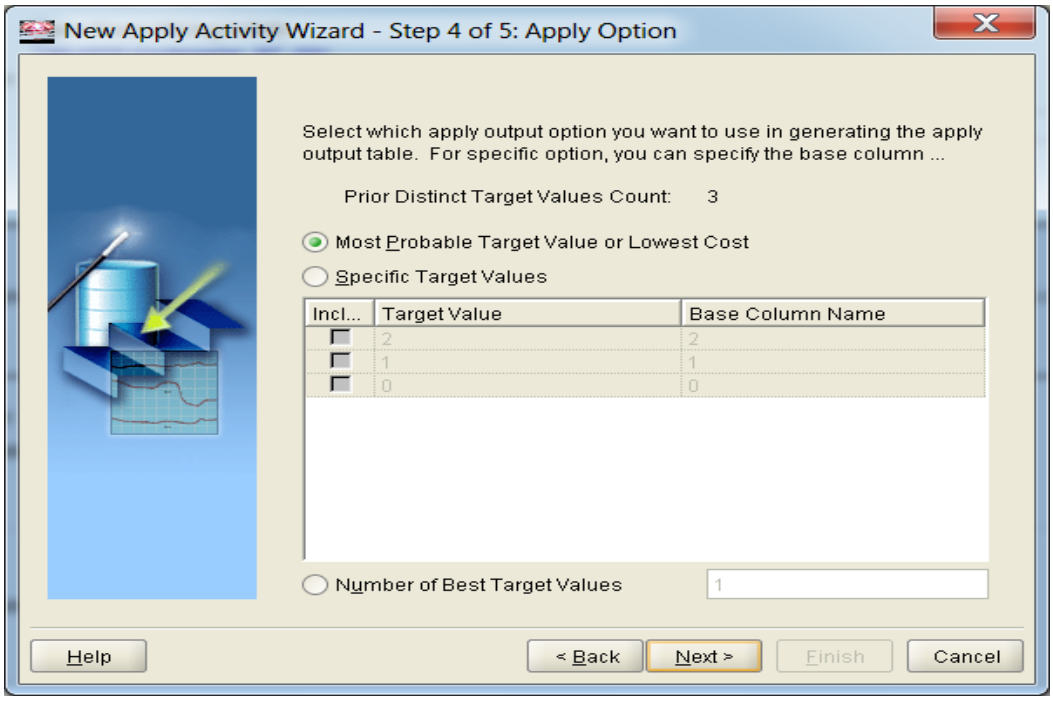


Şekil 5.43 : Uygulama 3. aşama, sonuç tablosunda gösterilecek sütunların seçimi



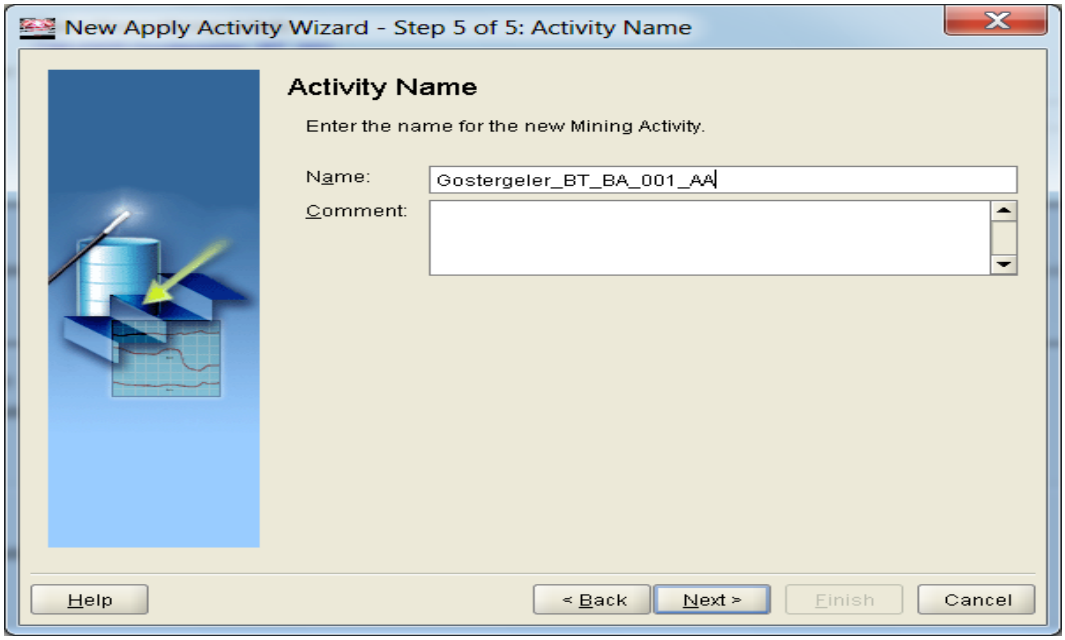
Şekil 5.44 : Uygulama 4. aşama, tahmini yapacak aktivitenin seçimi.

Her bir kayıt için en yüksek olasılığa sahip hedef değerler (0, 1 veya 2) görüntülenmek istendiğinden "Most Probable Target Value or Lowest Cost" seçeneği seçilmiştir. Bize vereceği değer, her bir kayıtta 0, 1 yada 2 değerleri için hangi olasılık değeri daha yüksekse sonuç olarak o değer "Prediction" sütununda görüntülenir.

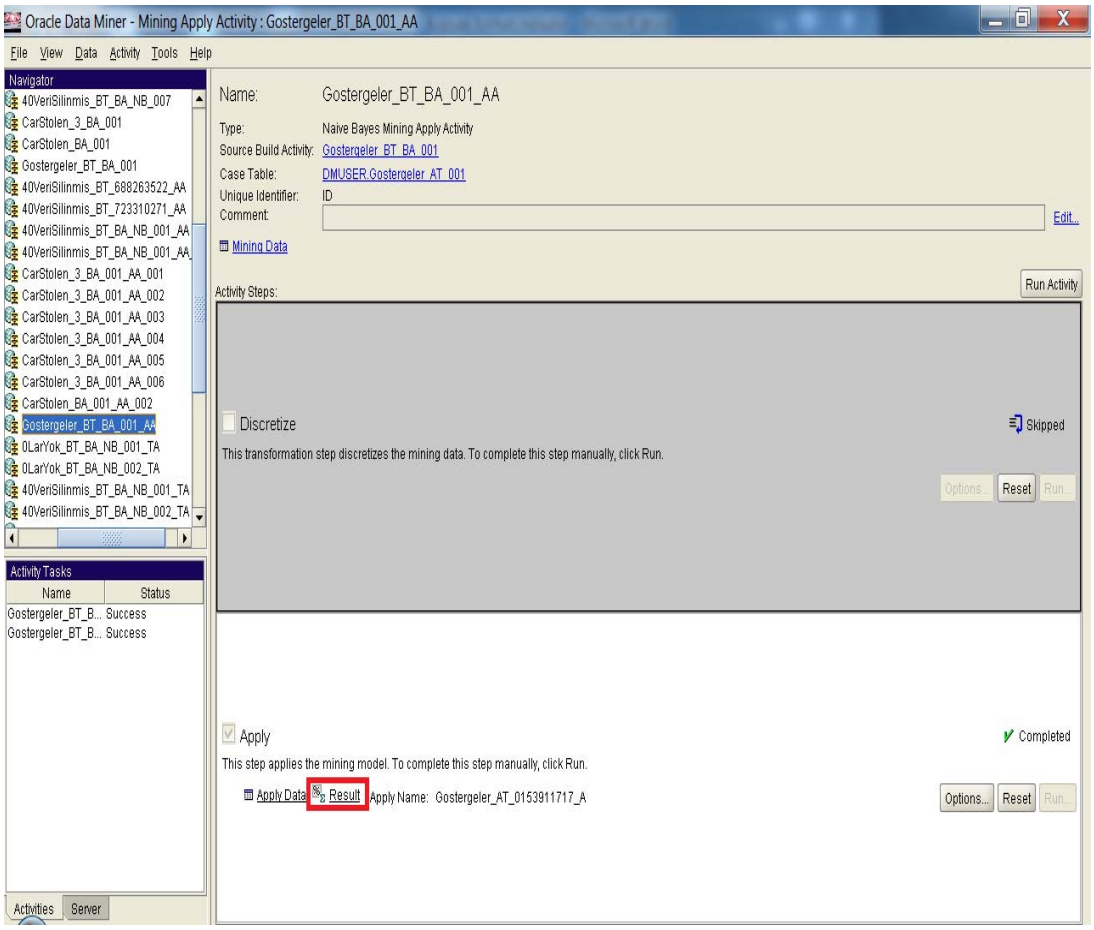


Şekil 5.45 : Uygulama 5. aşama, tahmin yönteminin seçimi

Son adımda uygulamaya isim verilerek uygulama süreci başlatılmıştır.



Şekil 5.46 : Uygulama 6. aşama, aktivitenin isminin belirlenmesi



Şekil 5.47 : Uygulama sürecinin son hali

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

5.4.1 Karar Ağaçları ile Modelin Uygulanması

Bu kısım Naivé Bayes ile modelin uygulanması kısmı ile tamamen aynıdır.

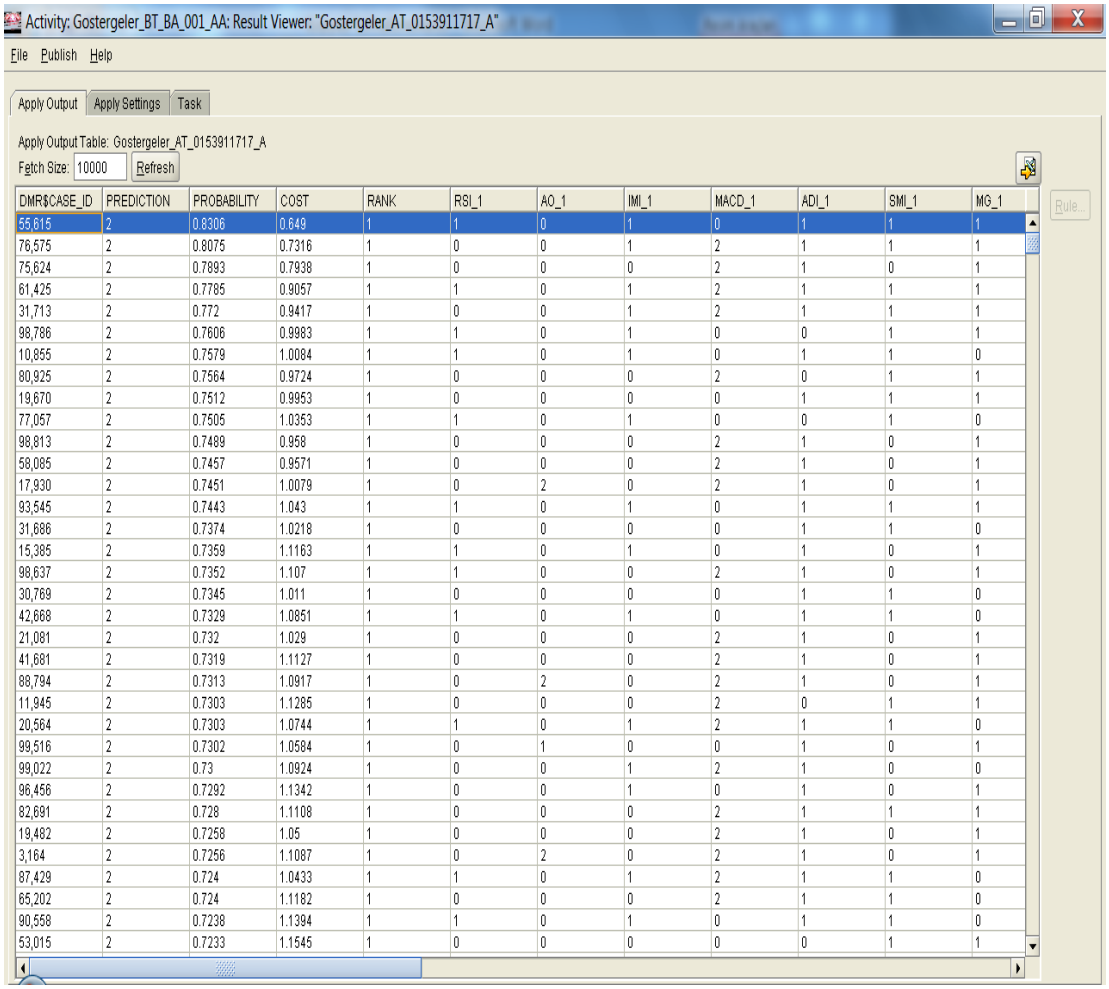
Naivé Bayes ve Karar Ağaçları ile modellerin uygulaması sonuçlar bölümünde farklılık göstermektedir. Bu farklılıklar da "Sonuçlar ve Yorumlar" bölümünde anlatılmıştır.

5.5 Sonuçlar ve Yorumlar

Sonuçlar ve yorumlar kısmı "Naivé Bayes için Sonuçlar ve Yorumlar" ve "Karar Ağaçları için Sonuçlar ve Yorumlar" olmak üzere iki kısımda incelenecektir.

5.5.1 Naivé Bayes için Sonuçlar ve Yorumlar

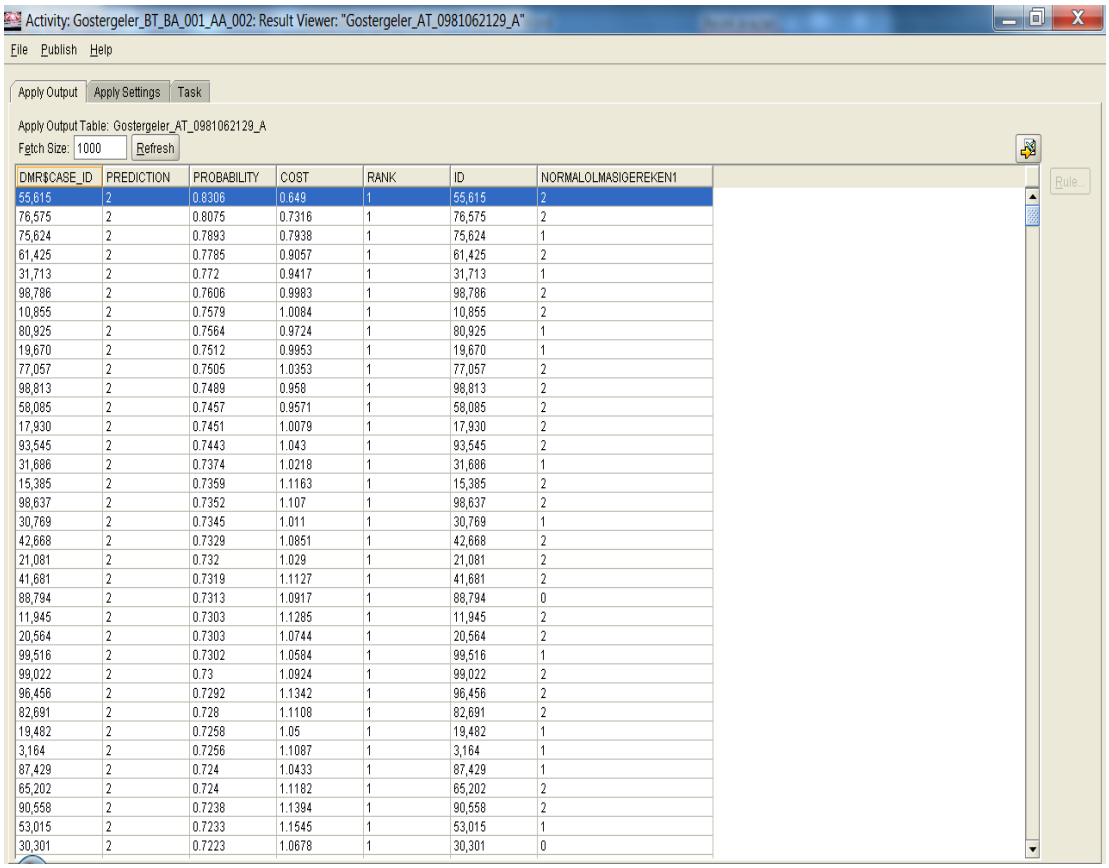
Aşağıda sonuç olarak verilen tüm ekran çıktılarında, tahminleri "Prediction" sütununda, tahminlerin olma olasılığı ise "Probability" sütununda görülmektedir. Diğer sütunlar ise girdi verilerinde de bulunan ve karşılaştırma için buraya da dâhil edilen sütunlardır.



DMR$CASE_ID	PREDICTION	PROBABILITY	COST	RANK	RSI_1	AO_1	IMI_1	MACD_1	ADI_1	SMI_1	MG_1
55,615	2	0.8306	0.649	1	1	0	1	0	1	1	1
76,575	2	0.8075	0.7316	1	0	0	1	2	1	1	1
75,624	2	0.7893	0.7938	1	0	0	0	2	1	0	1
61,425	2	0.7785	0.9057	1	1	0	1	2	1	1	1
31,713	2	0.772	0.9417	1	0	0	1	2	1	1	1
98,786	2	0.7606	0.9983	1	1	0	1	0	0	1	1
10,855	2	0.7579	1.0084	1	1	0	1	0	1	1	0
80,925	2	0.7564	0.9724	1	0	0	0	2	0	1	1
19,670	2	0.7512	0.9953	1	0	0	0	0	1	1	1
77,057	2	0.7505	1.0353	1	1	0	1	0	0	1	0
98,813	2	0.7489	0.958	1	0	0	0	2	1	0	1
58,085	2	0.7457	0.9571	1	0	0	0	2	1	0	1
17,930	2	0.7451	1.0079	1	0	2	0	2	1	0	1
93,545	2	0.7443	1.043	1	1	0	1	0	1	1	1
31,686	2	0.7374	1.0218	1	0	0	0	0	1	1	0
15,385	2	0.7359	1.1163	1	1	0	1	0	1	0	1
98,637	2	0.7352	1.107	1	1	0	0	2	1	0	1
30,769	2	0.7345	1.011	1	0	0	0	0	1	1	0
42,668	2	0.7329	1.0851	1	1	0	1	0	1	1	0
21,081	2	0.732	1.029	1	0	0	0	2	1	0	1
41,681	2	0.7319	1.1127	1	0	0	0	2	1	0	1
88,794	2	0.7313	1.0917	1	0	2	0	2	1	0	1
11,945	2	0.7303	1.1285	1	0	0	0	2	0	1	1
20,564	2	0.7303	1.0744	1	1	0	1	2	1	1	0
99,516	2	0.7302	1.0584	1	0	1	0	0	1	0	1
99,022	2	0.73	1.0924	1	0	0	1	2	1	0	0
96,456	2	0.7292	1.1342	1	0	0	1	0	1	0	1
82,691	2	0.728	1.1108	1	0	0	0	2	1	1	1
19,482	2	0.7258	1.05	1	0	0	0	2	1	0	1
3,164	2	0.7256	1.1087	1	0	2	0	2	1	0	1
87,429	2	0.724	1.0433	1	1	0	1	2	1	1	0
65,202	2	0.724	1.1182	1	0	0	0	2	1	1	0
90,558	2	0.7238	1.1394	1	1	0	1	0	1	1	0
53,015	2	0.7233	1.1545	1	0	0	0	0	0	1	1

Şekil 5.48a : Uygulama sonucu 1

Yukarıdaki tabloda ODM'nin tahminleri "Prediction" ve bu tahminlerin doğru olma olasılığı "Probability" sütununda, veri tablosundaki gerçek değerler ise "NORMALOLMASIGEREKEN" sütununda gösterilmektedir. Diğer sütunlar veri tablosunda kullanılan göstergelerin değerlerini göstermektedir. Gerçek değerler ile tahmin edilen değerleri daha rahat karşılaştırabilmek için Şekil 5.43 sonuç tablosunda gösterilecek verilerin seçimi bölümünde sadece "ID" ve "NORMALOLMASIGEREKEN" sütunları seçilirse aşağıdaki gibi bir sonuç elde edilir.



DMR$CASE_ID	PREDICTION	PROBABILITY	COST	RANK	ID	NORMALOLMASIGEREKEN1
55,615	2	0.8306	0.649	1	55,615	2
76,575	2	0.8075	0.7316	1	76,575	2
75,624	2	0.7893	0.7938	1	75,624	1
61,425	2	0.7785	0.9057	1	61,425	2
31,713	2	0.772	0.9417	1	31,713	1
98,786	2	0.7606	0.9983	1	98,786	2
10,855	2	0.7579	1.0084	1	10,855	2
80,925	2	0.7564	0.9724	1	80,925	1
19,670	2	0.7512	0.9953	1	19,670	1
77,057	2	0.7505	1.0353	1	77,057	2
98,813	2	0.7489	0.958	1	98,813	2
58,085	2	0.7457	0.9571	1	58,085	2
17,930	2	0.7451	1.0079	1	17,930	2
93,545	2	0.7443	1.043	1	93,545	2
31,686	2	0.7374	1.0218	1	31,686	1
15,385	2	0.7359	1.1163	1	15,385	2
98,637	2	0.7352	1.107	1	98,637	2
30,769	2	0.7345	1.011	1	30,769	1
42,668	2	0.7329	1.0851	1	42,668	2
21,081	2	0.732	1.029	1	21,081	2
41,681	2	0.7319	1.1127	1	41,681	2
88,794	2	0.7313	1.0917	1	88,794	0
11,945	2	0.7303	1.1285	1	11,945	2
20,564	2	0.7303	1.0744	1	20,564	2
99,516	2	0.7302	1.0584	1	99,516	1
99,022	2	0.73	1.0924	1	99,022	2
96,456	2	0.7292	1.1342	1	96,456	2
82,691	2	0.728	1.1108	1	82,691	2
19,482	2	0.7258	1.05	1	19,482	1
3,164	2	0.7256	1.1087	1	3,164	1
87,429	2	0.724	1.0433	1	87,429	1
65,202	2	0.724	1.1182	1	65,202	2
90,558	2	0.7238	1.1394	1	90,558	2
53,015	2	0.7233	1.1545	1	53,015	1
30,301	2	0.7223	1.0678	1	30,301	0

Şekil 5.48b : Uygulama sonucu 2

Yukarıda görüldüğü gibi “NORMALOLMASIGEREKEN” sütunu ilk iki satırda ‘2’ değerini almış ve ODM tarafından doğru olarak tahmin edilmiştir. Fakat üçüncü satırda “NORMALOLMASIGEREKEN” değeri ‘1’ iken ODM bu veriyi ‘2’ olarak tahmin etmiştir. Bu sonuçlardan modelin çok güvenilir olmadığı sonucu çıkarılabilir.

	0	1	2
0	1,555	1,612	1,808
1	3,282	4,339	4,477
2	3,514	3,795	5,139

Şekil 5.49: Güvenilirlik matrisi 1

Güvenilirlik matrisinde satırlar gerçek değerleri sütunlar ise tahmin edilen değerleri verdiğinden dolayı güvenirlilik matrisi daha ayrıntılı incelenirse şu sonuçlar çıkarılır.
Gerçek veri ‘0’ iken ODM’nin ‘0’ olarak tahmin ettiği veri sayısı 1555.

Gerçek veri ‘0’ iken ODM’nin ‘1’ olarak tahmin ettiği veri sayısı 1612.

Gerçek veri ‘0’ iken ODM’nin ‘2’ olarak tahmin ettiği veri sayısı 1808.

Gerçek veri ‘1’ iken ODM’nin ‘0’ olarak tahmin ettiği veri sayısı 3282.

Gerçek veri ‘1’ iken ODM’nin ‘1’ olarak tahmin ettiği veri sayısı 4339.

Gerçek veri ‘1’ iken ODM’nin ‘2’ olarak tahmin ettiği veri sayısı 4477.

Gerçek veri ‘2’ iken ODM’nin ‘0’ olarak tahmin ettiği veri sayısı 3514.

Gerçek veri ‘2’ iken ODM’nin ‘1’ olarak tahmin ettiği veri sayısı 3795.

Gerçek veri ‘2’ iken ODM’nin ‘2’ olarak tahmin ettiği veri sayısı 5139.

Bu durumda güvenilirlik matrisi yardımı ile gerçek doğruluk oranı %33,37 olarak şu şekilde bulunur.

$$\frac{1555+4339+5139}{1555+1612+1808+3282+4339+4477+3514+3795+5139}=0.373733$$

Modelde asıl amaçlanan alım ya da satım kararlarının doğru tahmin edilmesidir. Yani “NORMALOLMASIGEREKEN” verisinin ‘0’ değerini aldığı durumların tahmin edilmesi önemli değildir. Bu nedenle ‘0’ ların tahmin edildiği durumlar güvenilirlik matrisinden çıkarıldığında yeni oluşan güvenilirlik matrisi aşağıdaki gibi olur.

	1	2
1	4339	4477
2	3795	5139

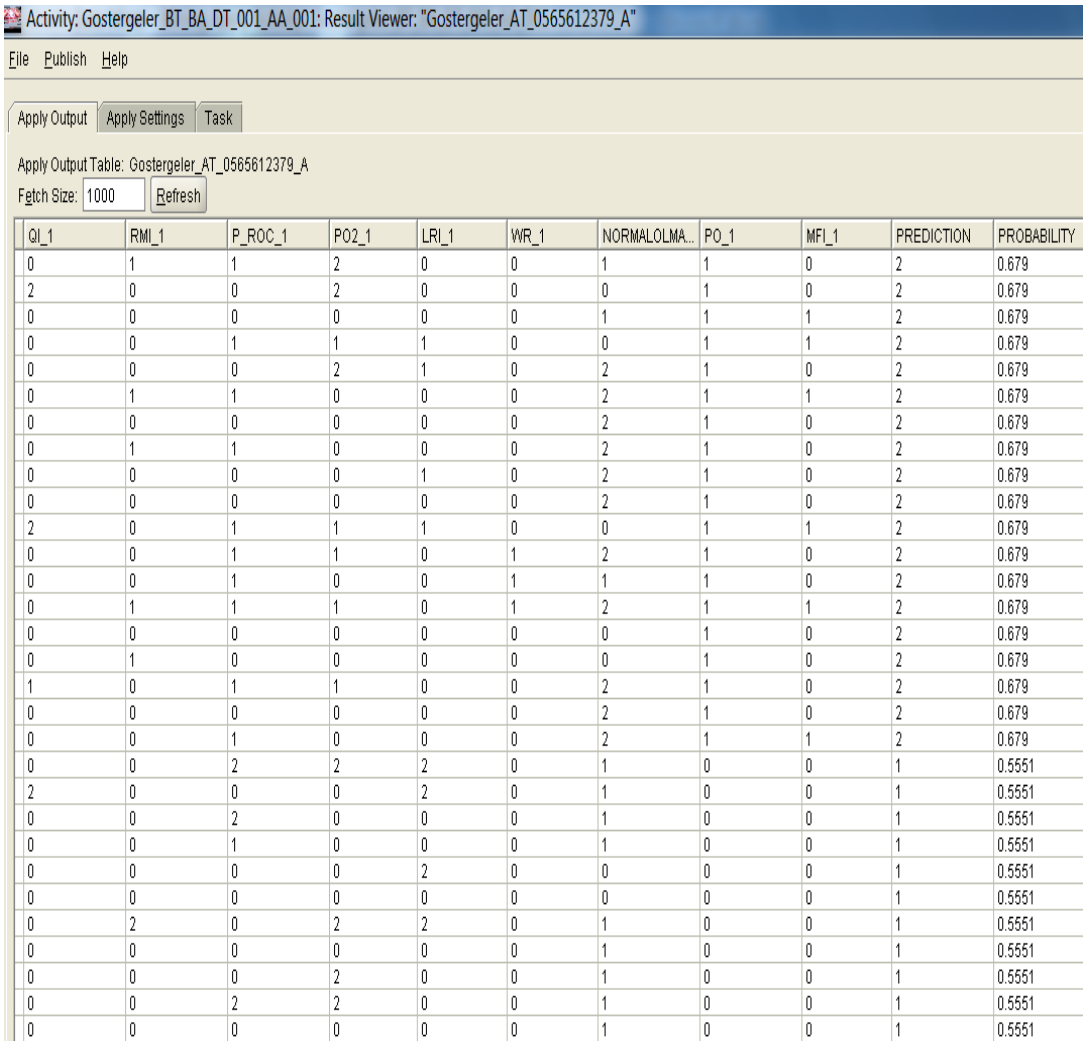
Tablo 5.1: Güvenilirlik matrisi 2

Bu durumda gerçek doğruluk oranı yeniden hesaplanırsa aşağıdaki sonuç elde edilir.

$$\frac{4339+5139}{4339+4477+3795+5139}=0.533971$$

5.5.2 Karar Ağaçları için Sonuçlar ve Yorumlar

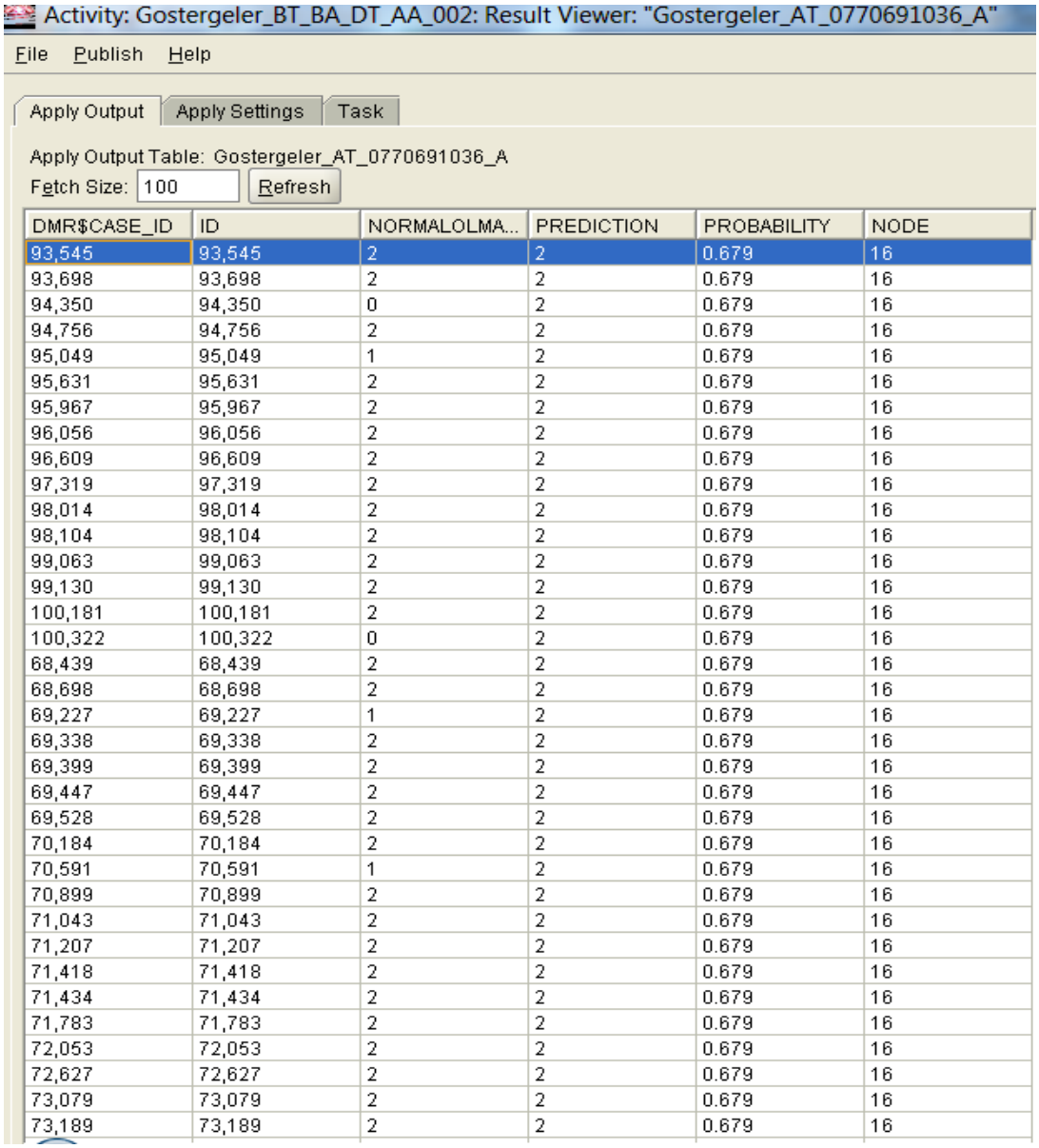
Bu uygulamanın Naivé Bayes uygulamasından farklılıkları aşağıdaki şekillerde gösterilmiştir.



QI_1	RMI_1	P_ROC_1	PO2_1	LRI_1	WR_1	NORMALOLMA...	PO_1	MFI_1	PREDICTION	PROBABILITY
0	1	1	2	0	0	1	1	0	2	0.679
2	0	0	2	0	0	0	1	0	2	0.679
0	0	0	0	0	0	1	1	1	2	0.679
0	0	1	1	1	0	0	1	1	2	0.679
0	0	0	2	1	0	2	1	0	2	0.679
0	1	1	0	0	0	2	1	1	2	0.679
0	0	0	0	0	0	2	1	0	2	0.679
0	1	1	0	0	0	2	1	0	2	0.679
0	0	0	0	1	0	2	1	0	2	0.679
0	0	0	0	0	0	2	1	0	2	0.679
2	0	1	1	1	0	0	1	1	2	0.679
0	0	1	1	0	1	2	1	0	2	0.679
0	0	1	0	0	1	1	1	0	2	0.679
0	1	1	1	0	1	2	1	1	2	0.679
0	0	0	0	0	0	0	1	0	2	0.679
0	1	0	0	0	0	0	1	0	2	0.679
1	0	1	1	0	0	2	1	0	2	0.679
0	0	0	0	0	0	2	1	0	2	0.679
0	0	1	0	0	0	2	1	1	2	0.679
0	0	2	2	2	0	1	0	0	1	0.5551
2	0	0	0	2	0	1	0	0	1	0.5551
0	0	2	0	0	0	1	0	0	1	0.5551
0	0	1	0	0	0	1	0	0	1	0.5551
0	0	0	0	2	0	0	0	0	1	0.5551
0	0	0	0	0	0	0	0	0	1	0.5551
0	2	0	2	2	0	1	0	0	1	0.5551
0	0	0	0	0	0	1	0	0	1	0.5551
0	0	0	2	0	0	1	0	0	1	0.5551
0	0	2	2	0	0	1	0	0	1	0.5551
0	0	0	0	0	0	1	0	0	1	0.5551

Şekil 5.50a : Uygulama sonucu 1

Yukarıdaki tabloda ODM'nin tahminleri "Prediction" ve bu tahminlerin doğru olma olasılığı "Probability" sütununda, veri tablosundaki gerçek değerler ise "NORMALOLMASIGEREKEN" sütununda gösterilmektedir. Diğer sütunlar veri tablosunda kullanılan göstergelerin değerlerini göstermektedir. Gerçek değerler ile tahmin edilen değerleri daha rahat karşılaştırabilmek için Şekil 5.43 sonuç tablosunda gösterilecek verilerin seçimi bölümünde sadece "ID" ve "NORMALOLMASIGEREKEN" sütunları seçilirse aşağıdaki gibi bir sonuç elde edilir.



DMR$CASE_ID	ID	NORMALOLMA...	PREDICTION	PROBABILITY	NODE
93,545	93,545	2	2	0.679	16
93,698	93,698	2	2	0.679	16
94,350	94,350	0	2	0.679	16
94,756	94,756	2	2	0.679	16
95,049	95,049	1	2	0.679	16
95,631	95,631	2	2	0.679	16
95,967	95,967	2	2	0.679	16
96,056	96,056	2	2	0.679	16
96,609	96,609	2	2	0.679	16
97,319	97,319	2	2	0.679	16
98,014	98,014	2	2	0.679	16
98,104	98,104	2	2	0.679	16
99,063	99,063	2	2	0.679	16
99,130	99,130	2	2	0.679	16
100,181	100,181	2	2	0.679	16
100,322	100,322	0	2	0.679	16
68,439	68,439	2	2	0.679	16
68,698	68,698	2	2	0.679	16
69,227	69,227	1	2	0.679	16
69,338	69,338	2	2	0.679	16
69,399	69,399	2	2	0.679	16
69,447	69,447	2	2	0.679	16
69,528	69,528	2	2	0.679	16
70,184	70,184	2	2	0.679	16
70,591	70,591	1	2	0.679	16
70,899	70,899	2	2	0.679	16
71,043	71,043	2	2	0.679	16
71,207	71,207	2	2	0.679	16
71,418	71,418	2	2	0.679	16
71,434	71,434	2	2	0.679	16
71,783	71,783	2	2	0.679	16
72,053	72,053	2	2	0.679	16
72,627	72,627	2	2	0.679	16
73,079	73,079	2	2	0.679	16
73,189	73,189	2	2	0.679	16

Şekil 5.50b : Uygulama sonucu 2

Yukarıda görüldüğü gibi "NORMALOLMASIGEREKEN" sütunu ilk iki satırda '2' değerini almış ve ODM tarafından doğru olarak tahmin edilmiştir. Fakat üçüncü satırda "NORMALOLMASIGEREKEN" değeri '0' iken ODM bu veriyi '2' olarak tahmin etmiştir. Bu sonuçlardan modelin çok güvenilir olmadığı sonucu çıkarılabilir.

	0	1	2
0	0	1,010	3,965
1	0	2,972	9,126
2	0	1,631	10,817

Şekil 5.51: K.A Güvenilirlik matrisi 1

Güvenilirlik matrisinde satırlar gerçek değerleri sütunlar ise tahmin edilen değerleri verdiğinden dolayı güvenirlilik matrisi daha ayrıntılı incelenirse şu sonuçlar çıkarılır. Gerçek veri '0' iken ODM'nin '0' olarak tahmin ettiği veri sayısı 0.

Gerçek veri '0' iken ODM'nin '1' olarak tahmin ettiği veri sayısı 1010.

Gerçek veri '0' iken ODM'nin '2' olarak tahmin ettiği veri sayısı 3965.

Gerçek veri '1' iken ODM'nin '0' olarak tahmin ettiği veri sayısı 0.

Gerçek veri '1' iken ODM'nin '1' olarak tahmin ettiği veri sayısı 2972.

Gerçek veri '1' iken ODM'nin '2' olarak tahmin ettiği veri sayısı 9126.

Gerçek veri '2' iken ODM'nin '0' olarak tahmin ettiği veri sayısı 0.

Gerçek veri '2' iken ODM'nin '1' olarak tahmin ettiği veri sayısı 1631.

Gerçek veri '2' iken ODM'nin '2' olarak tahmin ettiği veri sayısı 10817

Bu durumda güvenilirlik matrisi yardımı ile gerçek doğruluk oranı %46,70 olarak şu şekilde bulunur.

$$\frac{0+2972+10817}{0+2972+10817+1010+3965+0+9126+0+1631} = 0.467091$$

Modelde asıl amaçlanan alım ya da satım kararlarının doğru tahmin edilmesidir. Yani "NORMALOLMASIGEREKEN" verisinin '0' değerini aldığı durumların tahmin edilmesi önemli değildir. Bu nedenle '0' ların tahmin edildiği durumlar güvenilirlik matrisinden çıkarıldığında yeni oluşan güvenilirlik matrisi aşağıdaki gibi olur.

	1	2
1	2972	9126
2	1631	10817

Tablo 5.2: K:A Güvenilirlik matrisi 2

Bu durumda gerçek doğruluk oranı yeniden hesaplanırsa aşağıdaki sonuç elde edilir.

$$\frac{2972+10817}{2972+10817+9126+1631} = 0.562219$$

Sonuç olarak Karar Ağaçları algoritması bu problem için Naivé Bayes'den daha iyi sonuç vermektedir, fakat bu durumda da başarı oranı %56 da kalmıştır. Bu konuda

ODM ile daha kapsamlı araştırmalar ve uygulamalar yapılarak daha verimli sonuçlar elde edilebileceği öngörülmektedir.

KAYNAKLAR

[1] Akgüç Ö., 1991. *Kredi taleplerinin değerlendirilmesi,* p. 1

[2] Aldana, W. A., *Data Mining Industry: Emerging Trends and New Opportunities*, p.11.

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

[4] Cabena P., Hadjnian P., Stadler R., 1998. *Discovering data mining from concept to implementation*, Prentice Hall PTR, NJ.

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

[6] Günak N., 2007. *İleri Teknik Analiz Uygulamaları*, Literatür Yayınları, İstanbul.

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

[8] ORACLE CORPORATION, 2006. *Oracle 10g Release 2 Data Mining Tutorial,* pp. 78, 81, 82, 83, 84, 85, 86, 96.

[9] Özkan, Y., *Veri Madenciliği Yöntemleri*, p.41.

[10] Taft M., Krishnan R., Hornick M., Muhkin D., Tang G., Thomas S., Stengard P., 2005. *Oracle Data Mining Concepts, 10g Release 2,* p. 14, Oracle Corporation, CA.

[11] Url-1 *<http://www.oracle.com/global/tr/solutions/business_intelligence/data-mining.html>*, alındığı tarih 27.04.2010.

[12] Url-2 *<http://www.infora.com.tr/veri_temizleme.html>*, alındığı tarih 27.04.2010.

[13] Yapıcı A.P., Özel A., Ayça C., 2010.*Oracle Data Mıner ile Kredi Ödemeleri Üzerine Bir Veri Madenciliği Uygulaması,* ITU