İSTANBUL KÜLTÜR ÜNİVERTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

Üniversitelerin Adaylar Tarafından Tercih Edilme Desenlerini Veri Madenciliği Yöntemleri ile Belirleyen Bir Model Önerisi

YÜKSEK LİSANS TEZİ
Sinan Ataseven
0409042009

Anabilim Dalı: Matematik – Bilgisayar
Programı: Matematik – Bilgisayar

Tez Danışmanı: Yrd. Doç. Dr. Hikmet ÇAĞLAR

ŞUBAT 2008

İSTANBUL KÜLTÜR ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

Üniversitelerin Adaylar Tarafından Tercih Edilme Desenlerini Veri Madenciliği Yöntemleri ile Belirleyen Bir Model Önerisi

YÜKSEK LİSANS TEZİ

Sinan Ataseven
0409042009

Tezin Enstitüye Verildiği Tarih : 26 Şubat 2008
Tezin Savunulduğu Tarih : 29 Şubat 2008

Tez Danışmanı : **Yrd. Doç. Dr. Hikmet ÇAĞLAR**
Diğer Jüri Üyeleri : **Prof. Dr. Behiç ÇAĞAL**
Yrd. Doç. Dr. Işım DEMİRİZ

ŞUBAT 2008

ÖNSÖZ

Tezimin hazırlanışı sırasında bana yüksek gayretleriyle rehberlik eden sayın hocam Y. Doç. Dr. Hikmet ÇAĞLAR'a, tez konumun uygulama alanları hakkında üstün bilgi ve tecrübeye sahip olan ve çalışmalarımızda benden yardımlarını ve desteğini esirgemeyen sayın hocam Öğr. Gör. Burak KILANÇ'a, yüksek lisans öğrenimim ve asistanlık görevim süresince beraber çalışmaktan büyük onur duyduğum sayın hocam Prof.Dr. Behiç ÇAĞAL'a, beni destekleyen tüm arkadaşlarıma ve beni yetiştiren aileme çok teşekkür ederim.

Sinan Ataseven

Şubat 2008

İÇİNDEKİLER

KISALTMALAR

GUI : Graphical User Interface

OLAP : Online Analytical Processing

ODBC : Open Database Communication

DBMS : Database Management System

ASCII : American Standard Code for Information Interchange

ŞEKİL LİSTESİ

Sayfa

No

Sayfa

No

Sayfa

No

Sayfa

No

ÖZET

Bu çalışmada, üniversite adaylarının tercihlerini belirleyen desenlerin veri madenciliği yöntemleri kullanılarak geliştirilen model önerileri ve uygulamaları sunulmuştur.

SUMMARY

This work presents model suggestions and applications developed by data mining methods, which defines patterns of university student candidates preferences.

1. GİRİŞ

Üniversitelerde yer alan kontenjanların yüksek oranda ve nitelikli adaylar ile dolması, üniversite üst yönetimlerinin temel hedeflerinden biridir. Son yıllarda öncelikle vakıf üniversiteleri olmak üzere bütün üniversitelerin tanıtım faaliyetlerinde temel hedef, bu olmuştur.

Etkin ve verimli bir tanıtım faaliyetinin planlanması ve yürütülmesi ancak üniversite adayları ile ilgili ÖSYM tarafından paylaşılan verilerin Veri Madenciliği yöntemleri ile incelenmesi ve anlamlı sonuçlar üretilmesi ile mümkündür.

Üniversite adaylarının farklı üniversite türlerini tercih ederken yansıttıkları desenlerin saptanması bu çalışmanın temel konusu ve amacıdır.

Bu çalışma kapsamında:

- 2007 ÖSYS verileri düzenlenmiş,
- Modellemeye uygun hale getirilmiş,
- C 5.0 Karar Ağacı ve Lojistik Regresyon yöntemleri ile modeller oluşturulmuş,
- Modellerin tahmin başarıları belirlenmiş,
- Model 2007 ÖSS'de herhangi bir bölüme yerleşememiş aday kitlesi üzerine uygulanmış,
- Geliştirilen model, tanıtım faaliyetlerinin planlanmasında kullanılmak üzere, 2008 ÖSYS verileri üzerinde uygulanmaya hazır hale getirilmiştir.

2. VERİ MADENCİLİĞİ

Bu bölümde ilk olarak Veri Madenciliği hakkında bilgi verilecek ve Veri Madenciliği'nin ilgi alanlarından bahsedilecektir. Veri Madenciliği çalışmalarının bilgisayar ortamında yürütülmesini sağlayan ve günümüzde kullanılan güncel ve gelişmiş yazılımlar teker teker incelenecektir. Veritabanlarında Bilgi Keşfi sürecinden bahsedilecek ve bu sürecin adımlarına değinilecektir.

2.1 Veri Madenciliğine Giriş

2.1.1 Veri Madenciliği Hakkında

Belirli veri yığın kaynaklarını kullanarak, önceden tahmin edilemeyen, geçerli ve uygulanabilir bilgilerin, yeni ilişki ve eğilimlerin keşfedilmesini sağlayan dinamik süreçlerinin tümü, Veri Madenciliği olarak tanımlanabilir.

Yeni bir araştırma disiplini olan Veri Madenciliği; büyük veri tabanlarından ve veri ambarlarından kullanışlı, önceden bilinmeyen, önemli ve sonuçta anlaşılabilir desenlerin keşfedimesini ve çekilmesini sağlar [1].

Geniş içeriğe sahip veri kaynaklarından, anlamlı ve kullanılabilir bilginin otomatik şekilde elde edilmesini hedefler.

Veri Madenciliği; matematik, istatistik, yapay zeka ve veritabanı tekniklerini kullanarak, veriler hakkında kurallar, trendler ve benzerlikler ortaya çıkartıp karar verme aşamasında kullanıcıları bilgilendirerek destek olmayı amaçlar.

Veri Madenciliği, veri tabanlarından ve veri ambarlarından toplanan veriye değer katmak ve bilgi keşfetmek için çok geniş veri havuzlarını tarar [2].

2.1.2 Veri Madenciliği'nin İlgi Alanları

Günümüzde veri madenciliğinin başlıca ilgi alanları olarak aşağıdakiler sayılabilir;

Pazarlama

- Müşteri segmentasyonunda,
- Müşterilerin demografik özellikleri arasındaki bağlantıların kurulmasında,
- Çeşitli pazarlama kampanyalarında,
- Mevcut müşterilerin elde tutulması için geliştirilecek pazarlama stratejilerinin oluşturulmasında,
- Pazar sepeti analizinde,
- Çapraz satış analizleri,
- Müşteri değerleme,
- Müşteri ilişkileri yönetiminde,
- Çeşitli müşteri analizlerinde,
- Satış tahminlerinde,

Bankacılık

- Farklı finansal göstergeler arasındaki gizli korelasyonların bulunmasında,
- Kredi kartı dolandırıcılıklarının tespitinde,
- Müşteri segmentasyonunda,
- Kredi taleplerinin değerlendirilmesinde,
- Usulsüzlük tespiti,
- Risk analizleri,
- Risk yönetimi,

Sigortacılık

- Yeni poliçe talep edecek müşterilerin tahmin edilmesinde,
- Sigorta dolandırıcılıklarının tespitinde,
- Riskli müşteri tipinin belirlenmesinde.

Perakendecilik

- Satış noktası veri analizleri,
- Alış-veriş sepeti analizleri,
- Tedarik ve mağaza yerleşim optimizasyonu,

Borsa

- Hisse senedi fiyat tahmini,
- Genel piyasa analizleri,
- Alım-satım stratejilerinin optimizasyonu.

Telekomünikasyon

- Kalite ve iyileştirme analizlerinde,
- Hisse tespitlerinde,
- Hatların yoğunluk tahminlerinde,

Sağlık ve İlaç

- Test sonuçlarının tahmini,
- Ürün geliştirme,
- Tıbbi teşhis
- Tedavi sürecinin belirlenmesinde

Endüstri

- Kalite kontrol analizlerinde
- Lojistik,
- Üretim süreçlerinin optimizisyonunda,

Bilim ve Mühendislik

- Ampirik veriler üzerinde modeller kurarak bilimsel ve teknik problemlerin çözümlenmesinde

2.2 Veri Madenciliği Yazılımları

Veri Madenciliği aşamalarını içeren genel amaçlı Veri Madenciliği yazılımları incelenmiştir.

2.2.1 Darwin

Sağlayıcı: Oracle Corporation
Kullandığı Teknikler: Karar ağaçları (CART), K-Ortalama, K-En Yakın Komşu, sinir ağları, regresyon
Çalıştığı Platformlar: Windows, Sun Solaris, HP-UX

"Oracle Data Mining Suite" ismiyle de anılan Darwin yazılımı, Windows tabanlı kullanışlı bir arayüz sağlamaktadır. Farklı bir kaç Veri Madenciliği algoritmaları içermektedir. ODBC üzerinde birçok ilişkisel veritabanı sistemine bağlanabilir. Sunucu – İstemci yaklaşımını kullanmaktadır. Ayrıca yapılan çalışmaların C, C++ ve Java kodları export edilebilmektedir.

http://www.oracle.com/technology/documentation/darwin.html

2.2.2 DBMiner

Sağlayıcı: DBMiner Technologies Inc.
Kullandığı Teknikler: Karar ağaçları, K-ortalama
Çalıştığı Platformlar: Windows

DBMiner, Simon Fraser Üniversitesi tarafından bir bilimsel çalışma olarak geliştirilmiştir [4]. Çevrimiçi Analitik İşleme (Online Analytical Mining) özelliğine sahiptir. Bu özellik sayesinde OLAP üzerinden dinamik olarak Veri Madenciliği teknikleri uygulanabilmektedir. Arayüz olarak kullanıcılara GUI veya DMSQL seçenekleri sunulmaktadır.

http://www.dbminer.com

2.2.3 Intelligent Miner

Sağlayıcı: IBM Corporation
Kullandığı Teknikler: Karar ağaçları (modifiye edilmiş CART), K-ortalama, sinir ağları, regresyon
Çalıştığı Platformlar: Windows, Solaris, AIX, OS/390, OS/400

IBM firması tarafından geliştirilen Intelligent Miner, DB2 veritabanları veya düz metin dosyaları üzerinde Veri Madenciliği teknikleri kullanır. Ayrıca ODBC üzerinden diğer ilişkisel DBMS'ler üzerindeki verilere erişebilir. Sunucu – istemci yaklaşımını kullanır ve kullanıcılara veri madenciliği işlemleri için basit bir GUI sunmaktadır. Metin analiz araçları, arama makinası gibi özellikler eklenmiştir.
http://www-306.ibm.com/software/data/iminer/

2.2.4 Enterprise Miner

Sağlayıcı: SAS Institute Inc.
Kullandığı Teknikler: Karar ağaçları (CART ve CHAID), K-en yakın komşu, regresyon, sinir ağları
Çalıştığı Platformlar: İstemci (Windows), Sunucu (Unix, Windows)

Enterprise Miner, SAS firması tarafından geliştirilen Veri Madenciliği çözümleri sunan üründür. Bilgi keşfi sürecini, firmanın kendi belirlediği SEMMA isimli bir süreç ile sağlamaktadır. Simge tabanlı bir GUI ile süreç akışı kullanıcılar tarafından yönetilebilmektedir. Veriden örneklem alma, veriyi görselleştirme, filtreleme, dönüştürme amaçlı araçları içinde barındırmaktadır.

2.2.5 Clementine

Sağlayıcı: SPSS Inc.
Kullandığı Teknikler: Karar ağaçları, sinir ağları, regresyon, K-ortalama, K-en yakın komşu, Naive-Bayes, faktör ve ayırma analizleri, temel bileşenler analizi, kümeleme, kohonen ağları
Çalıştığı Platformlar: Windows, HP/UX, IBM AIX, Sun Solaris

Clementine, SPSS firmasının Veri Madenciliği çözümleri sunan ürünüdür. Kendisini diğer yazılımlardan ayıran en önemli özelliği, veri madenciliğine 1994 yılından itibaren getirmiş olduğu GUI yaklaşımıdır. Açıklayıcı simgelerin kullanımı ile, veri akış süreci yaratılmakta ve işleme konmaktadır. Her simgenin, veri keşfi süreci anlamında, bir işlevi bulunmaktadır. Veriye erişim, verinin hazırlanması, görsellik ve modelleme aşamaları için birbirinden farklı simgelere GUI üzerinden erişilmektedir.

Clementine geniş veri kümeleri üzerinde veri madenciliğini istemci sunucu yaklaşımı ile yapar. Gerektiğinde, veri erişim isteklerini SQL sorgularına dönüştürebilir. Geniş veri tipi desteğine sahiptir. Elde edilen sonuçların değişik biçimlerdeki export seçenekleri bulunmaktadır.

2.3 Veri Madenciliği ile Diğer Disiplinler Arasındaki İlişkiler

2.3.1 Veri Tabanı ile İlişkisi

Veri Madenciliği ile veri tabanı arasında, doğrudan bir ilişki bulunmaktadır. Veri Madenciliğinin kullandığı verilerin girişi veri tabanı tarafından sağlanmaktadır.

Veri tabanı üzerinde, tespit edilmiş örüntülerle ilgili sorgular çalıştırılması söz konusu iken, Veri Madenciliği önceden belirlenemeyen örüntüleri keşfetmeye yönelir.

2.3.2 Makina Öğrenimi ile İlişkisi

Veri Madenciliği, makine öğrenimi ile yakından ilgilidir. Makine öğreniminin ilgilendiği örüntü tanıma çalışmaları, Veri Madenciliğinde örüntü çıkarma ve modelleme amacıyla kullanılmaktadır.

Veri Madenciliğini makina öğreniminden ayıran başlıca özellik; makina öğreniminin gelecekteki olaylarla ilgili öngörücü modeller ve kurallara yoğunlaşmasına karşın, veri madenciliğinin var olan veri üzerinde tanımlayıcı modeller geliştirme ve örüntü çıkarmaya yoğunlaşmasıdır. [5]

Ayrıca; Veri Madenciliği, makina öğreniminden farklı olarak, yüksek yoğunluktaki veri kümeleri ile ilgilenir. Makina öğrenimi için modellerin performansı üzerindeki çalışmalar daha önemlidir. [5]

2.3.3 İstatistik ile İlişkisi

İstatistik biliminin veri kümelerindeki gürültü tespiti, gürültü ayırma ve azaltma teknikleri, Veri Madenciliğinde kullanılmaktadır. Aynı şekilde, istatistiğin olasılık tahminlerine dayanan öngörü teknikleri, Veri Madenciliği sürecinin aşamalarında kullanım alanı bulmaktadır.

2.4 Bilgi Keşfi Süreci

Bilgi keşfi, verinin hazırlanması ile başlayan ve keşfedilen desenlerin değerlendirilmesi ile tamamlanan bir süreçtir. Veri madenciliği, belirli desen keşif algoritmalarının uygulandığı bir bilgi keşfi sürecidir [3].

Tüm süreç aşağıdaki adımlardan oluşur:

- Problemin Tanımlanması
- Verilerin Hazırlanması (Veri Seçimi, Veri Temizlenmesi, Veri Entegrasyonu, Veri Dönüşümü)
- Modelin Kurulması ve Değerlendirilmesi
- Modelin Kullanılması
- Modelin İzlenmesi

2.4.1 Problemin Tanımlanması

Veri madenciliği çalışmalarında başarılı olmanın ilk şartı, uygulamanın hangi işletme amacı için yapılacağının açık bir şekilde tanımlanmasıdır. İlgili işletme amacı işletme problemi üzerine odaklanmış ve açık bir dille ifade edilmiş olmalı, elde edilecek sonuçların başarı düzeylerinin nasıl ölçüleceği tanımlanmalıdır. Ayrıca yanlış tahminlerde katlanılacak olan maliyetlere ve doğru tahminlerde kazanılacak faydalara ilişkin tahminlere de bu aşamada yer verilmelidir.

2.4.2 Verilerin Hazırlanması

Modelin kurulması aşamasında ortaya çıkacak sorunlar, bu aşamaya sık sık geri dönülmesine ve verilerin yeniden düzenlenmesine neden olacaktır. Bu durum verilerin hazırlanması ve modelin kurulması aşamaları için, bir analistin veri keşfi sürecinin toplamı içerisinde enerji ve zamanının % 50 - % 85'ini harcamasına neden olmaktadır.

Verilerin hazırlanması aşaması kendi içerisinde toplama, değer biçme, birleştirme ve temizleme, seçme ve dönüştürme adımlarından meydana gelmektedir.

2.4.2.1 Toplama (Collection)

Tanımlanan problem için gerekli olduğu düşünülen verilerin ve bu verilerin toplanacağı veri kaynaklarının belirlenmesi adımıdır. Verilerin toplanmasında kuruluşun kendi veri kaynaklarının dışında, nufus sayımı, hava durumu, merkez bankası kara listesi gibi veri tabanlarından veya veri pazarlayan kuruluşların veri tabanlarından faydalanılabilir.

2.4.2.2 Değer Biçme (Assessment)

Veri madenciliğinde kullanılacak verilerin farklı kaynaklardan toplanması, doğal olarak veri uyumsuzluklarına neden olacaktır. Bu uyumsuzlukların başlıcaları farklı zamanlara ait olmaları, kodlama farklılıkları (örneğin bir veri tabanında cinsiyet özelliğinin e/k, diğer bir veri tabanında 0/1 olarak kodlanması), farklı ölçü birimleridir. Ayrıca verilerin nasıl, nerede ve hangi koşullar altında toplandığı da önem taşımaktadır.

Bu nedenlerle, iyi sonuç alınacak modeller ancak iyi verilerin üzerine kurulabileceği için, toplanan verilerin ne ölçüde uyumlu oldukları bu adımda incelenerek değerlendirilmelidir.

2.4.2.3 Birleştirme ve Temizleme

Bu adımda farklı kaynaklardan toplanan verilerde bulunan ve bir önceki adımda belirlenen sorunlar mümkün olduğu ölçüde giderilerek veriler tek bir veri tabanında

toplanır. Ancak basit yöntemlerle ve baştan savma olarak yapılacak sorun giderme işlemlerinin, ileriki aşamalarda daha büyük sorunların kaynağı olacağı unutulmamalıdır.

2.4.2.4 Seçim (Selection)

Bu adımda kurulacak modele bağlı olarak veri seçimi yapılır. Örneğin tahmin edici bir model için, bu adım bağımlı ve bağımsız değişkenlerin ve modelin eğitiminde kullanılacak veri kümesinin seçilmesi anlamını taşımaktadır.

Sıra numarası, kimlik numarası gibi anlamlı olmayan ve diğer değişkenlerin modeldeki ağırlığının azalmasına da neden olabilecek değişkenlerin modele girmemesi gerekmektedir. Bazı veri madenciliği algoritmaları konu ile ilgisi olmayan bu tip değişkenleri otomatik olarak elese de, pratikte bu işlemin kullanılan yazılıma bırakılmaması daha akılcı olacaktır.

Verilerin göreselleştirilmesine olanak sağlayan grafik araçlar ve bunların sunduğu ilişkiler, bağımsız değişkenlerin seçilmesinde önemli yararlar sağlayabilir.

Genellikle yanlış veri girişinden veya bir kereye özgü bir olayın gerçekleşmesinden kaynaklanan verilerin (Outlier), önemli bir uyarıcı enformasyon içerip içermediği kontrol edildikten sonra veri kümesinden atılması tercih edilir.

Modelde kullanılan veri tabanının çok büyük olması durumunda tesadüfiliği bozmayacak şekilde örnekleme yapılması uygun olabilir. Günümüzde hesaplama olanakları ne kadar gelişmiş olursa olsun, çok büyük veri tabanları üzerinde çok sayıda modelin denenmesi zaman kısıtı nedeni ile mümkün olamamaktadır. Bu nedenle tüm veri tabanını kullanarak bir kaç model denemek yerine, tesadüfi olarak örneklenmiş bir veri tabanı parçası üzerinde bir çok modelin denenmesi ve bunlar arasından en güvenilir ve güçlü modelin seçilmesi daha uygun olacaktır.

2.4.2.5 Dönüştürme (Transformation)

Kredi riskinin tahmini için geliştirilen bir modelde, borç/gelir gibi önceden hesaplanmış bir oran yerine, ayrı ayrı borç ve gelir verilerinin kullanılması tercih edilebilir. Ayrıca modelde kullanılan algoritma, verilerin gösteriminde önemli rol oynayacaktır. Örneğin bir uygulamada bir yapay sinir ağı algoritmasının kullanılması durumunda kategorik değişken değerlerinin evet/hayır olması; bir karar ağacı algoritmasının kullanılması durumunda ise örneğin gelir değişken değerlerinin yüksek/orta/düşük olarak gruplanmış olması modelin etkinliğini artıracaktır.

2.4.3 Modelin Kurulması ve Değerlendirilmesi

Tanımlanan problem için en uygun modelin bulunabilmesi, olabildiğince çok sayıda modelin kurularak denenmesi ile mümkündür. Bu nedenle veri hazırlama ve model kurma aşamaları, en iyi olduğu düşünülen modele varılıncaya kadar yinelenen bir süreçtir.

Model kuruluş süreci denetimli (Supervised) ve denetimsiz (Unsupervised) öğrenimin kullanıldığı modellere göre farklılık göstermektedir.

Örnekten öğrenme olarak da isimlendirilen denetimli öğrenimde, bir denetçi tarafından ilgili sınıflar önceden belirlenen bir kritere göre ayrılarak, her sınıf için çeşitli örnekler verilir. Sistemin amacı verilen örneklerden hareket ederek her bir sınıfa ilişkin özelliklerin bulunması ve bu özelliklerin kural cümleleri ile ifade edilmesidir.

Öğrenme süreci tamamlandığında, tanımlanan kural cümleleri verilen yeni örneklere uygulanır ve yeni örneklerin hangi sınıfa ait olduğu kurulan model tarafından belirlenir.

Denetimsiz öğrenmede, kümeleme analizinde olduğu gibi ilgili örneklerin gözlenmesi ve bu örneklerin özellikleri arasındaki benzerliklerden hareket ederek sınıfların tanımlanması amaçlanmaktadır.

Denetimli öğrenimde seçilen algoritmaya uygun olarak ilgili veriler hazırlandıktan sonra, ilk aşamada verinin bir kısmı modelin öğrenimi, diğer kısmı ise modelin geçerliliğinin test edilmesi için ayrılır. Modelin öğrenimi öğrenim kümesi kullanılarak

gerçekleştirildikten sonra, test kümesi ile modelin doğruluk derecesi (Accuracy) belirlenir.

Kurulan modelin doğruluk derecesi ne denli yüksek olursa olsun, gerçek dünyayı tam anlamı ile modellediğini garanti edebilmek mümkün değildir. Yapılan testler sonucunda geçerli bir modelin doğru olmamasındaki başlıca nedenler, model kuruluşunda kabul edilen varsayımlar ve modelde kullanılan verilerin doğru olmamasıdır. Örneğin modelin kurulması sırasında varsayılan enflasyon oranının zaman içerisinde değişmesi, bireyin satın alma davranışını belirgin olarak etkileyecektir.

2.4.4 Modelin Kullanılması

Kurulan ve geçerliliği kabul edilen model doğrudan bir uygulama olabileceği gibi, bir başka uygulamanın alt parçası olarak kullanılabilir. Kurulan modeller risk analizi, kredi değerlendirme, dolandırıcılık tespiti gibi işletme uygulamalarında doğrudan kullanılabileceği gibi, promosyon planlaması simülasyonuna entegre edilebilir veya tahmin edilen envanter düzeyleri yeniden sipariş noktasının altına düştüğünde, otomatik olarak sipariş verilmesini sağlayacak bir uygulamanın içine gömülebilir.

2.4.5 Modelin İzlenmesi

Zaman içerisinde bütün sistemlerin özelliklerinde ve dolayısıyla ürettikleri verilerde ortaya çıkan değişiklikler, kurulan modellerin sürekli olarak izlenmesini ve gerekiyorsa yeniden düzenlenmesini gerektirecektir. Tahmin edilen ve gözlenen değişkenler arasındaki farklılığı gösteren grafikler model sonuçlarının izlenmesinde kullanılan yararlı bir yöntemdir.

2.5 Veri Madenciliği Modelleri

Veri madenciliğinde kullanılan modeller, tahmin edici (Predictive) ve tanımlayıcı (Descriptive) olmak üzere iki ana başlık altında incelenmektedir.

Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç değerlerin tahmin edilmesi amaçlanmaktadır.

Tanımlayıcı modellerde ise karar vermeye rehberlik etmede kullanılabilecek mevcut verilerdeki örüntülerin tanımlanması sağlanmaktadır.

Veri madenciliği modellerini gördükleri işlevlere göre,

- Sınıflama (Classification) ve Regresyon (Regression),
- Kümeleme (Clustering),
- Birliktelik Kuralları (Association Rules) ve Ardışık Zamanlı Örüntüler (Sequential Patterns)

olmak üzere üç ana başlık altında incelemek mümkündür. Sınıflama ve regresyon modelleri tahmin edici, kümeleme, birliktelik kuralları ve ardışık zamanlı örüntü modelleri tanımlayıcı modellerdir.

2.5.1 Sınıflama ve Regresyon Modelleri (Öngörüsel Modeller)

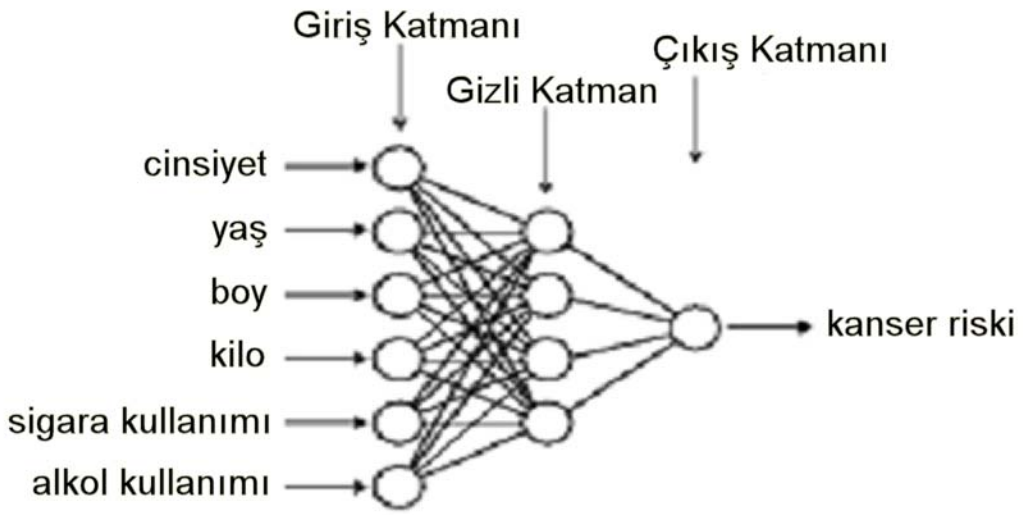
Mevcut verilerden hareket ederek geleceğin tahmin edilmesinde faydalanılan ve veri madenciliği teknikleri içerisinde en yaygın kullanıma sahip olan sınıflama ve regresyon modelleri arasındaki temel fark, tahmin edilen bağımlı değişkenin kategorik veya süreklilik gösteren bir değere sahip olmasıdır. Ancak çok terimli lojistik regresyon (multinomial logistic regression) gibi kategorik değerlerin de tahmin edilmesine olanak sağlayan tekniklerle, her iki model giderek birbirine yaklaşmakta ve bunun bir sonucu olarak aynı tekniklerden yararlanılması mümkün olmaktadır. Sınıflama ve regresyon modellerinde kullanılan başlıca teknikler,

- Karar Ağaçları (Decision Trees),
- Yapay Sinir Ağları (Artificial Neural Networks),
- Doğrusal Regresyon (Linear Regression)
- Lojistik Regresyondur (Logistic Regression)

şeklindedir.

2.5.1.1 Sinir Ağları

İnsan beyninin çalışması temel alınmıştır. En temel birim nöron adı verilen parçalardır. Her bir nöron bir işin basit bir parçasını yürütmeye çalışan bir elemanı olarak da düşünülebilir. Nöronların birbiri ile bağlantısından oluşan ağda birbirileri ile olan etkileşimleri sonucu öğrenme gerçekleşir.



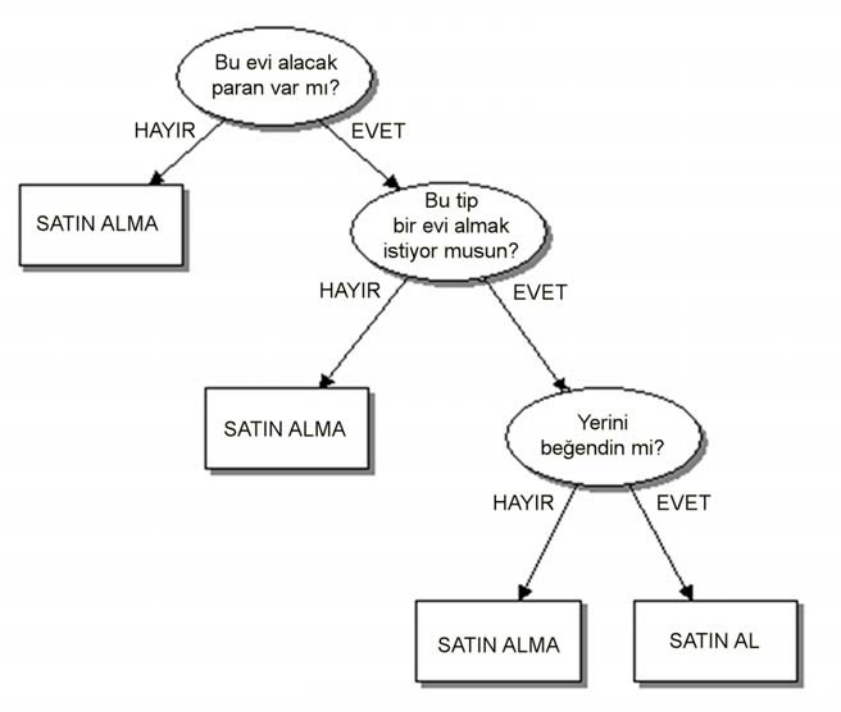
Şekil 2.1: Sinir Ağlarının Yapısı

Input ve Output alanları nümerik yada sembolik (dummy veya binary hale getirilmiş) olabilir. Gizli katman bir önceki katmana bağlı ve outputlarını içeren bir çok nöron içerir. Sinir ağları genel olarak birden fazla gizli katmandan (minimum tutulmaya çalışılır) meydana gelir. Herhangi bir katmandaki nöronlar bir sonraki katmandaki tüm nöronlarla bağlantılıdır.

Sinir ağlarının veri ve sonuçlar arasındaki ilişkiyi öğrenme sürecine ağın eğitilmesi denir. Ağ eğitildikten sonra öğrenme sonucu elde edilen bilgilere dayanılarak yeni verilere ait tahmin ya da karar verilir.

2.5.1.2 Karar Ağaçları

Verinin içindeki sonuçla yada hedef alanla ilişkili farklı segmentleri tanımlamak amacıyla oluşturulan karar ağaçları yada kurallar dizisidir. Model çıktısı her bir kuralı ve nedenlerini açıkca gösterir.



Şekil 2.2: Karar Ağaçları Model Yapısına bir örnek

2.5.1.3 Doğrusal Regresyon

Doğrusal regresyon, sayısal alanları kullanarak, gene sayısal bir sonuç elde etmeye yarayan bir tekniktir. Fakat sayısal olmayan alanlar da, sözde kodlamalarla (dummy-coding) tahminci olarak kullanılabilir.

2.5.1.4 Lojistik Regresyon

Lojistik Regresyon, sayısal olmayan bir çıktının tahmininde kullanılır. Aslında lojistik regresyon, gözlemlerin bağımlı değişkenin hangi kategorisine ait olduğunu tahmin eden bir olasılık fonksiyonudur. Bu prosedür, herhangi bir gözlemi, olasılığı 0.5'ten büyükse bir kategoriye, değilse diğer bir kategoriye atar.

2.5.2 Kümeleme Modelleri

Kümeleme modellerinde amaç, küme üyelerinin birbirlerine çok benzediği, ancak özellikleri birbirlerinden çok farklı olan kümelerin bulunması ve veri tabanındaki kayıtların bu farklı kümelere bölünmesidir. Başlangıç aşamasında veri tabanındaki kayıtların hangi kümelere ayrılacağı veya kümelemenin hangi değişken özelliklerine göre yapılacağı bilinmemekte, konunun uzmanı olan bir kişi tarafından kümelerin neler olacağı tahmin edilmektedir.

Başlıca üç çeşit kümeleme yöntemi vardır:

- Kohonen Ağları
- K-Ortalama Tekniği
- İki Adımlı Kümeleme

2.5.2.1 Kohonen Ağları

Kohonen ağı çıktısı olmayan bir sinir ağı olarak düşünülebilir. Bu tip ağlar veriyi, girdilerin örüntülerine göre bölümlere (segmentlere) ayırmaya yarar.

2.5.2.2 K-Ortalama Tekniği

K-Ortalama Tekniği verideki gruplandırmaları belirlemek için kullanılan hızlı bir yöntemdir. K sayısı oluşması istenen grup sayılarını gösterir ve kullanıcı tarafından belirlenir.

2.5.2.3 İki Adımlı Kümeleme

İki Adımlı Kümeleme yönteminde önce grup sayısı belirlenir. Küme sayısı için bir aralık tespit edildikten sonra, ilk adımda tüm kayıtlar ön kümelere ayrılır. İkinci adımda ise hiyerarşik kümeleme yönteme uygulanır.

2.5.3 Birliktelik Kuralları ve Ardışık Zamanlı Örüntüler

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

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

- Müşteriler bira satın aldığında, % 75 ihtimalle patates cipsi de alırlar,

- Düşük yağlı peynir ve yağsız yoğurt alan müşteriler, %85 ihtimalle diet süt de satın alırlar.

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

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

- İMKB endeksi düşerken A hisse senedinin değeri % 15'den daha fazla artacak olursa, üç iş günü içerisinde B hisse senedinin değeri % 60 ihtimalle artacaktır,

- Çekiç satın alan bir müşteri, ilk üç ay içerisinde % 15, bu dönemi izleyen üç ay içerisinde % 10 ihtimalle çivi satın alacaktır.

SPSS Clementine Yazılımı

Bu bölümde, SPSS firmasının Veri Madenciliği yazılımı olan Clementine programı incelenmiştir. İlk olarak yazılım hakkında genel bir bilgi verilecek olup, daha sonra bu yazılımın Veri Madenciliği ve Bilgi Keşfi süreçlerine yönelik olarak geliştirilen CRISP-DM metadolojisinden bahsedilecektir. Clementine yazılımda veriye erişim, veri kalitesinin incelenmesi, veri manipulasyonu, kümeleme teknikleri ve modelleme yöntemlerinin nasıl yapıldığına değinilecektir.

Clementine, kurumların karşılaştıkları spesifik iş problemleri ile mücadele etmelerinde esnek ve kapsamlı bir çalışma ortamı sağlayan bir veri madenciliği yazılımıdır. Bu yazılımla ilgili genel bir bilgi bu çalışmanın 2.2.5 numaralı bölümünde belirtilmiştir.

2.6 CRISP-DM Metadolojisi

CRISPT-DM Metadolojisi, Veri Madenciliği ve Bilgi Keşfi süreçlerine ortak bir yaklaşım geliştirmek amacıyla, bir Avrupa Birliği projesi içerisinde 4 üyeli bir konsorsiyum tarafından geliştirilmiştir. Değişik endüstri sektörlerine uygulanmak üzere, büyük veri madenciliği projelerini daha hızlı, ucuz, güvenilir ve yönetilebilir kılınması hedeflenmiştir. Bunun yanı sıra, küçük çaptaki Veri Madenciliği araştırmalarının da bu metadolojiden faydalanması mümkündür.

CRISP-DM ismi, Cross Industrial Standart Proccessing for Data Mining (Veri Madenciliği için Çapraz Endüstriyel Süreç Standardı) tanımından türetilmiştir.

CRISP-DM Metadoloji modelini geliştiren konsorsiyum üyeleri aşağıdaki gibidir:

- Teradata
- SPSS
- DaimlerChrysler
- OHRA

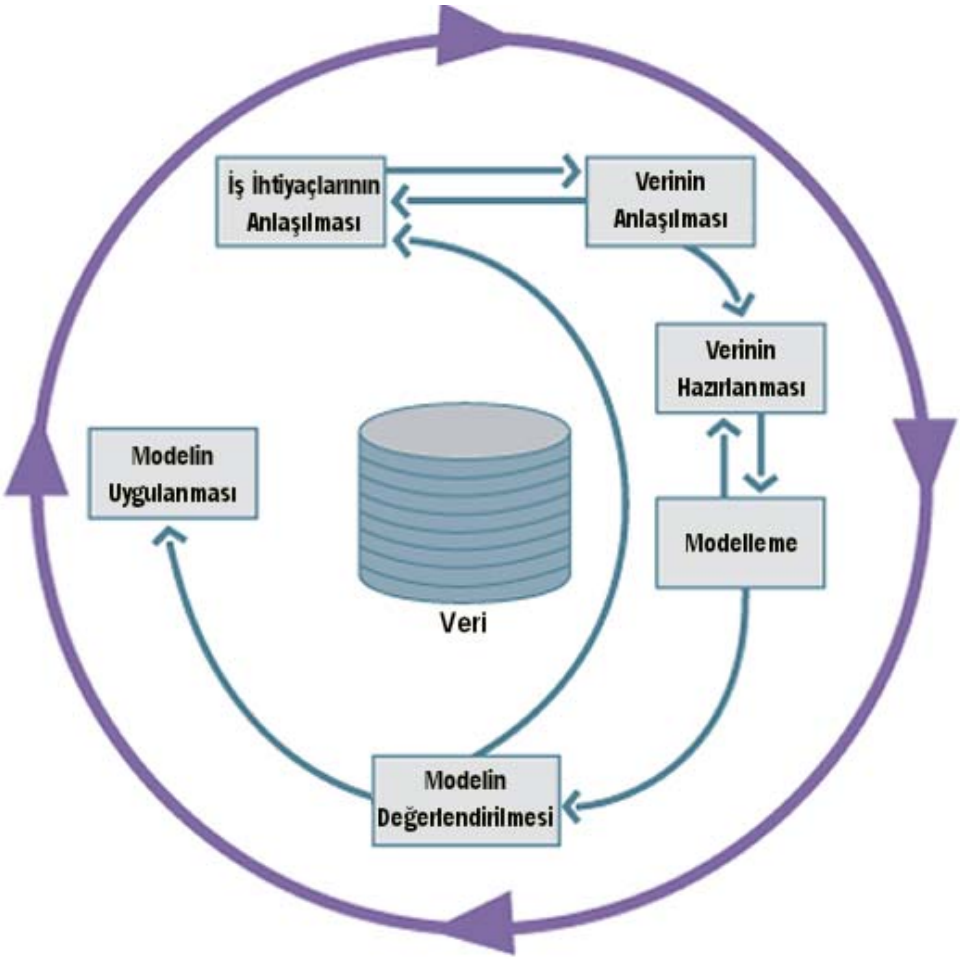
Bu konsorsiyum aynı zamanda, Veri Madenciliği ile ilgilenen 300'ün üzerindeki organizasyonun oluşturduğu bir özel ilgi grubu tarafından da desteklenmektedir.

Konsorsiyum üyesi SPSS firmasının Veri Madenciliği yazılımı olan Clementine, bu metadoloji doğrultusunda geliştirilen süreçe uygun olarak çalışmaktadır.

Veri Madenciliğinde Bilgi Keşfi süreçlerine bir model yaklaşımı olaran geliştirilen CRISP-DM'in süreç adımlar:

- İş İhtiyaçlarının Anlaşılması
- Verinin Anlaşılması
- Verinin Hazırlanması
- Modelleme
- Modelin Değerlendirmesi
- Modelin Uygulanması

şeklinde tanımlanmıştır.



Şekil 2.3: CRISP-DM Süreç Modelinin Aşamaları

2.7 Clementine Yazılımda CRISP-DM

CRISP-DM Süreç Modelinin her bir aşaması için Clementine yazılımının sunduğu değişik araçlar bulunmaktadır.

2.7.1 İş İhtiyaçlarının Anlaşılması

İş problemlerinin irdelenmesi aşamasında iş deneyimi önemlidir. Bu ilk adımda projenin amaç ve gerekliliklerinin iş perspektifi ile anlaşılması, bu bilginin Veri Madenciliği problem tanımı olarak netleştirilmesi ve hedeflere ulaşma amaçlı ilk planların oluşturulması söz konusudur. Clementine ile birlikte opsiyonel olarak lisanslanan uygulama şablonları (CATS: Clementine Application Templates) SPSS'in

farklı Veri Madenciliği projelerine dair ciddi bir iş deneyimini kullanıcılarına aktarmayı amaçlayarak hazırlanmıştır. Bu uygulama şablonları CRM CAT, Web Mining CAT, Telco CAT, Fraud CAT, Microarray CAT dir.

2.7.2 Verinin Anlaşılması

Verinin anlaşılması aşaması veri kaynaklarına bağlanma, veriyi tanıma, verinin kalitesini anlama ve verinin grafiksel olarak incelenmesi, hipotezleri oluşturma amaçlı veri gruplarını değerlendirme aşamalarını içerir. Clementine'ın grafikler ve tablolar üzerinde belli bölgelerin seçimini yapma özelliği bu aşamada önemlidir. Clementine içerisinde yer alan histogram, line plot, point plot, web associaion graphs, statistics, distribution graphs, data audit işlemcileri verinin ön incelenmesinde sıkça kullanılan işlemcilerdendir.

2.7.3 Verinin Hazırlanması

Veri Madenciliği projesinde kullanılacak olan veri setinin modellemeye hazırlanması, modelleme sonrasında yeniden veri üzerinde çeşitli düzenlemelerinin yapılmasını içerir ve veri hazırlama adımı birden fazla tekrarlanabilir.

Clementine'da verinin modellemeye hazırlanması amacı ile çok sayıda metod kullanılmaktadır. Veriye erişim aşamasında Clementine açık bir çözümdür. ODBC uyumlu olan bütün veritabanı verilerine kolayca bağlanabilir ve verilerin formatı değiştirilmeden kullanılabilir. SPSS ve SAS verileri ile serbest ve sabit ASCII formatındaki veriler kolayca alınabilir, Clementine'a entegre olarak kullanılan Text Mining çözümü ile yapısal olmayan yazı tipindeki veriler kullanılabilir. Ayrıca web kayıt verileri kolayca kullanılabilir.

Veri üzerinde temizlik yapma, verinin düzenlenmesi amacı ile çok sayıda işlemci bulunmaktadır. Variable File ve Fixed File işlemcileri ile geçersiz karakterler temizlenir. Kayıtlar ve alanlar üzerinde yapılan işlemler için çok sayıda işlemci bulunmaktadır. Kayıt seçimi ile ilgili olarak sample, merge, sort, aggregate, derive, vb işlemciler kullanılabilir.

2.7.4 Modelleme

Clementine modelleme için zengin bir içerik sunmaktadır. Clementine içerisinde yer alan modelleme yöntemleri 3 ana grup altında toplanmaktadır: Tahmin edici (Predictive), Kümeleyici (Clustering), Birliktelik (Associations) modelleme.

Tahmin edici modellerden

- Sinir Ağları
- Karar Ağaçları: C5.0 ve C&R Tree
- Regresyon
- Lojistik Regresyon
- Ardışıklık Tespiti

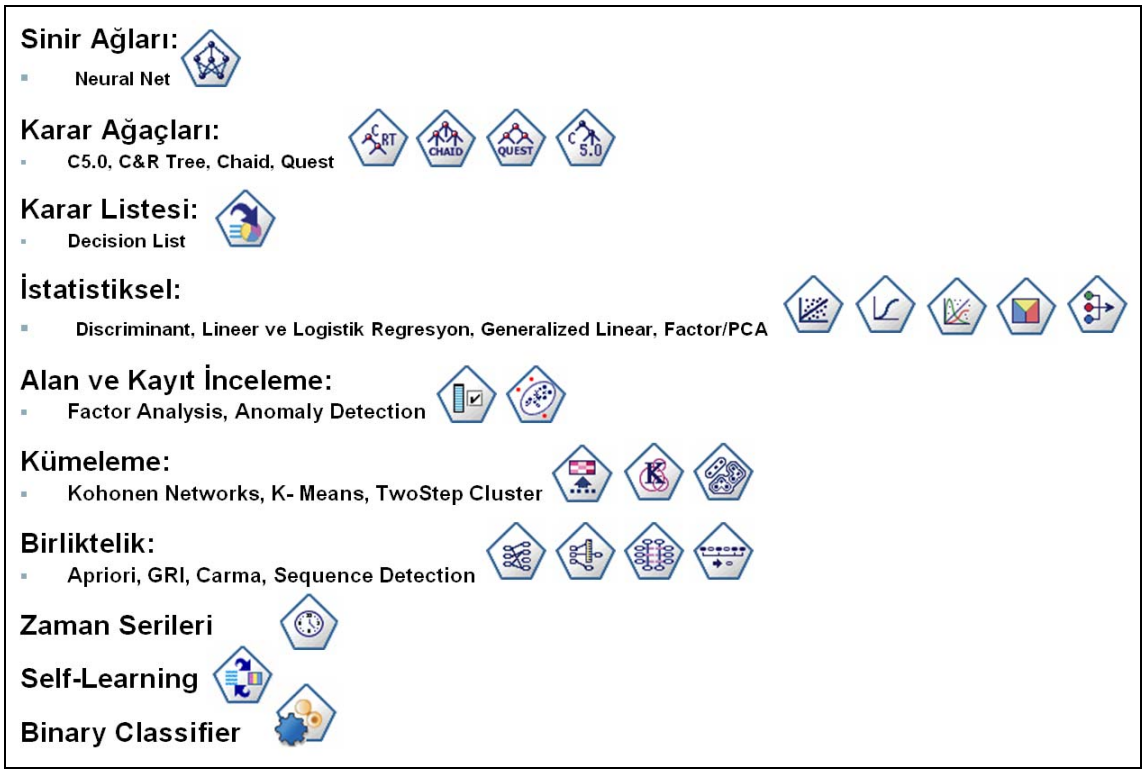
yöntemleri; Kümelemeyici modellerden

- Kohonen Ağları
- K-Ortalama
- İki Adımlı Kümeleme

yöntemleri ve Birliktelik modellerinden

- Apriori
- Gri

yöntemleri için Clementine içerisinde gerekli araçlar bulunmaktadır.



Şekil 2.4: Clementine'da kullanılan Veri Modenciliği Modellerinin Simgeleri

2.7.5 Değerlendirme

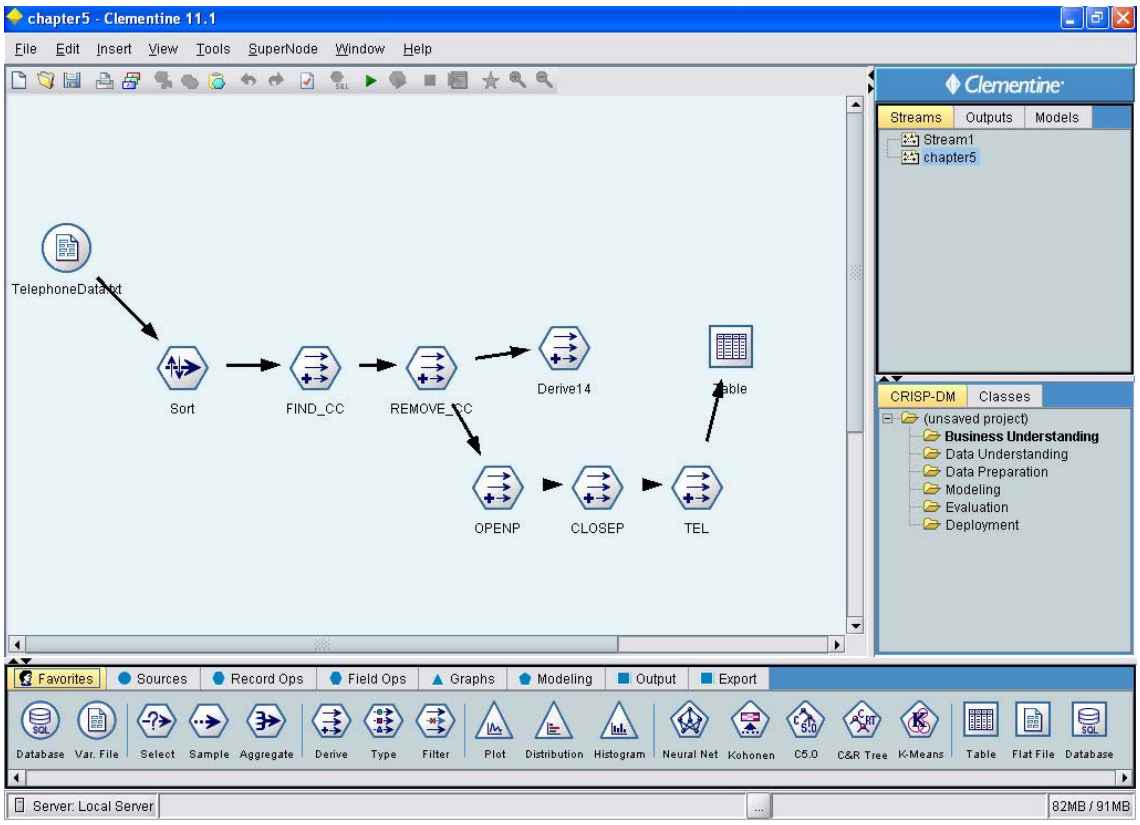
Modeller yaygın kullanıma alınmadan önce sonuçlarının ve güvenilirliklerinin tespit edilmesi aşamasıdır. Gain, Profit, Lift ve Response grafik araçları bu aşama için Clemetine yazılımının sunduğu çözümlerdir.

2.7.6 Uygulama

Geliştirilen modellerin ne şekilde kullanıma alınacağı daha çok iş kullanıcılarının karar vermesi gereken bir husustur. Bu amaçla kullanılabilecek farklı seçimler mevcuttur. Clementine Solution Publisher, Clementine ile birilkte opsiyonel olarak lisanslanabilen, veri erişiminden veri manipulasyonu, skorlama modelleri, vb bütün aşamaların otomatik hale getirilmesini sağlar. CLEO, XML tabanlı bir gereç yardımı ile tahmin edici modellerin çevrimiçi kullanımı amaçlanır. Clementine Batch Mode, Clementine özelliklerinin kullanıcı arayüzü olmaksızın kullanılmasını sağlar.

2.8 Clementine Kullanıcı Arayüzü

Kullanıcının esas çalışma alanı, Stream Canvas olup, bu alan görsel programlama tekniklerini kullanarak veri madenciliği yapmamıza olanak sağlar. Bu çalışma alanına nod adı verilen ve veri üzerinde yapılacak işlemleri niteleyen görsel öğeler eklenerek ve bu öğeler arasında bağlantılar kurularak Veri Madenciliği çalışması yapılır.



Şekil 2.5: Clementine 11.1 Kullanıcı Arayüzü

Kullanıcı arayüzünün alt kısmında paletler bulunmaktadır. Her palet, kendisiyle ilişkili birkaç nod içerir. Örneğin Sources paleti, verileri modele eklemeye yarayan nodları içerir. Nodlar, Stream Canvas'a yerleştirildikten sonra birbirine bağlanarak akımlar (Streams) oluşturulur. Akımlar, nodlardan veri akışını simgeler ve her akım bir çıktı (output) veya modelle sonlanır.

Clementine penceresinin sağ üst köşesinde üç tip yönetici alt pencere vardır: Streams, Outputs ve Models. Akımları açmak, saklamak, adlarını değiştirmek ve silmek için Streams tabı kullanılır. Clementine'ın çıktıları (grafik ve tablolar) Outputs tabında saklanır. Models tabı, Clementine'da oluşturulan modelleri sağlamak için kullanılır. Modeller direk olarak Browse seçeneğiyle görüntülenebilir ya da Stream Canvas'ta bulunan akımlara eklenebilir.

Sol alt köşede, Veri Madenciliği çalışmalarının organize edildiği Projects penceresi bulunur. CRISP-DM tabı, akımları, çıktıları, ve dip notları CRISP-DM aşamalarına uygun olarak düzenlememizi sağlar. Classes tabı, oluşturduğumuz nesnelerin kategorilerine uygun olarak düzenlenmesini sağlar. Nesneler aşağıdaki kategorilere eklenir:

- Streams
- Nods
- Models
- Tables, Graphs, Reports
- Other (Clementine ile ilgisiz dosyalar)

Menü çubuğu 8 nesneden oluşur:
File: Kullanıcının akım ya da projeleri oluşturması, açması ya da kaydetmesine yarar.

Edit: Kullanıcının klasik düzenleme işlemlerini (kopyalama, yapıştırma gibi) gerçekleştirmesine olanak sağlar.

Insert: Kullanıcının bir nodu kanvasa eklemesini sağlar.

View: Kullanıcının istenilen elemanları görüntülemesine ya da gizlemesine yarar.

Tools: Kullanıcının Clementine'ın çalıştığı ortamı düzenlemesini sağlar.

Supernode: Kullanıcının supernod ismi verilen yoğunlaştırılmış akımlarla ilgili oluşturma, kaydetme gibi işlemleri yapmasını sağlar.

Window: Kullanıcının istenilen pencereleri görüntülemesini ve kaptmasını sağlar.

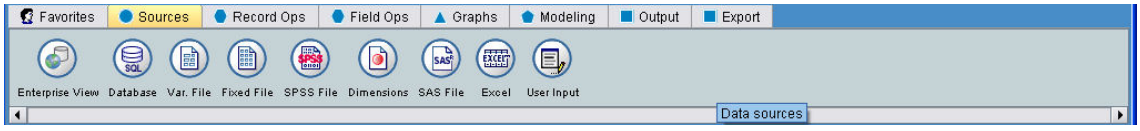
Help: Clementine yazılımının yardım dokümanlarına erişilmesini sağlar.

2.9 Veriye Erişim

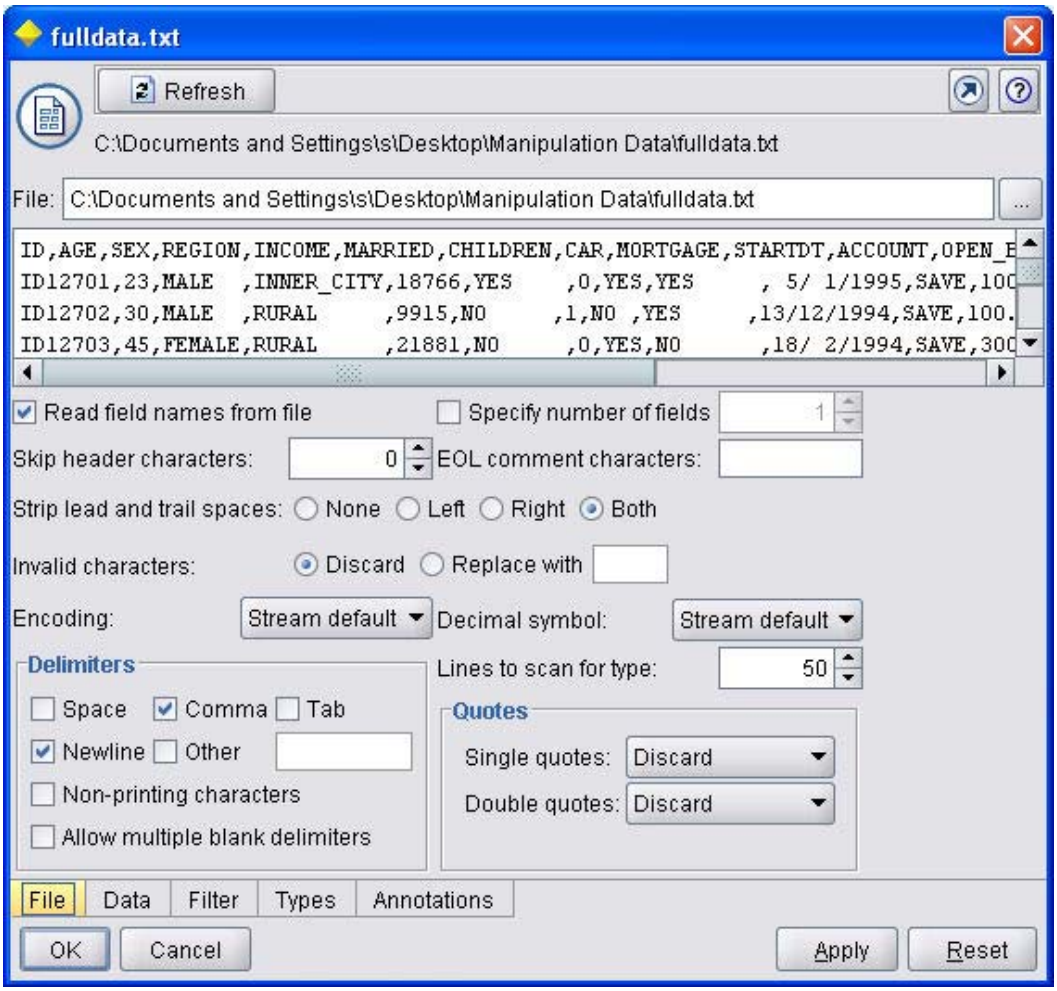
2.9.1 Clementine Veri Erişim İşlemcileri

Clementine, değişik formatlardaki dosyaları Sources paletindeki nodlar sayesinde okuyabilir. Metin dosyalarını okumak için Var.File ve Fixed File nodları kullanılır.

SPSS ve SAS veri dosyaları SPPS File ve SAS File nodları kullanılarak direk olarak okunabilir. Database nodu kullanılarak ODBC protokolunu destekleyen tüm veritabanları (Excel, Oracle, Access, SQL Server, vb.) okunabilir.



Şekil 2.6: Clementine Arayüzünde Veri Erişim İşlemcileri



Şekil 2.7: Fixed File Nodu ile Fixed ASCII Formatındaki Veri Dosyasının Okutulma Penceresi

2.9.2 Clementine Veri Erişiminde Temel Kavramlar

Clementine, Veri Madenciliği yapılmak üzere verilere erişim aşamasında, belirli bazı kavramlar üzerinde çalışmaktadır. Populasyon, örneklem ve değişken kavramları Clementine içerisinde verilerin ifadesi amacıyla kullanılmaktadır.

2.9.2.1 Populasyon

Araştırmacının ilgilendiği ve ortak özelliklere sahip birimlerden oluşan topluluğun tamamıdır. Bir şirketin tüm çalışanları ya da şirketin bütün müşterileri populasyona örnek olabilir.

2.9.2.2 Örneklem

Populasyon özellikleri hakkında bilgi edinmek için populasyondan seçilen alt grub örneklem olarak adlandırılmaktadır. Örneklem bilgilerinden yola çıkarak istatistiksel tekniklerle populasyon hakkında genelleme yapılır.

2.9.2.3 Değişken:

Populasyonda bulunan birimlerin ortak özelliklerinin her birine "değişken" denir.

Değişkenler bağımlılık durumlarına göre ikiye ayrılır;

- Bağımlı Değişkenler
- Bağımsız Değişkenler

Değişkenler Clementine'da ölçüm seviyelerine göre 4 tiptedir,

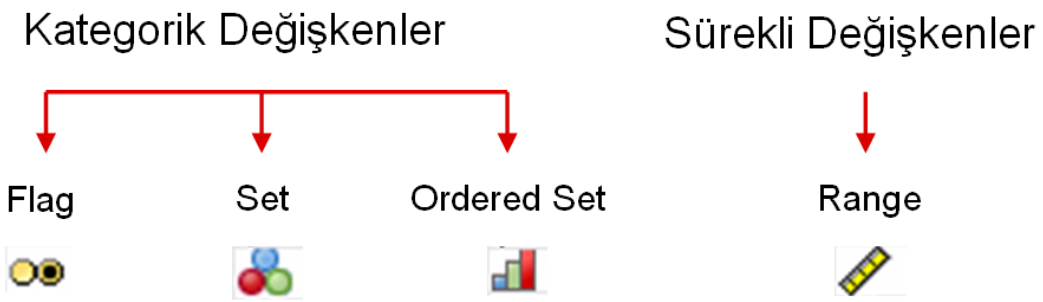
- Flag
- Set
- Ordered Set
- Range

2.9.2.3.1 Bağımlılık Durumlarına Göre Değişkenler

Değeri, diğer değişkenler tarafından bir fonksiyonla açıklanan veya tahmin edilen değişkene Bağımlı Değişken denir. Bağımlı değişkenin değerlerini bir fonksiyonla açıklayan değişkenlere Bağımsız Değişkenler denir.

Kişinin eğitim durumuna bakarak gelir durumunu tahmin etmek, bu duruma örnek olarak gösterilebilir.

2.9.2.3.2 Ölçüm Seviyelerine Göre Değişken Tipleri



Şekil 2.8: Ölçüm Seviyelerine Göre Değişken Tipleri

Flag: Evet/Hayır veya 1, 2 gibi iki farklı değer alabilen alanlara denir.

Set: Bekar, evli, boşanmış, dul gibi ikiden fazla değer alabilen nominal alanlara denir.

Ordered Set: Yüksek, orta düşük gibi ikiden fazla değer alabilen ordinal (sıralı) alanlara denir.

Range: 0-100 veya 0.75-1.25 gibi bir aralık içerisinde bütün değerleri alabilen sayısal alanlara denir. Integer, real veya date/time formatında olabilir.

2.9.2.3.3 Diğer Ölçüm Seviyeleri

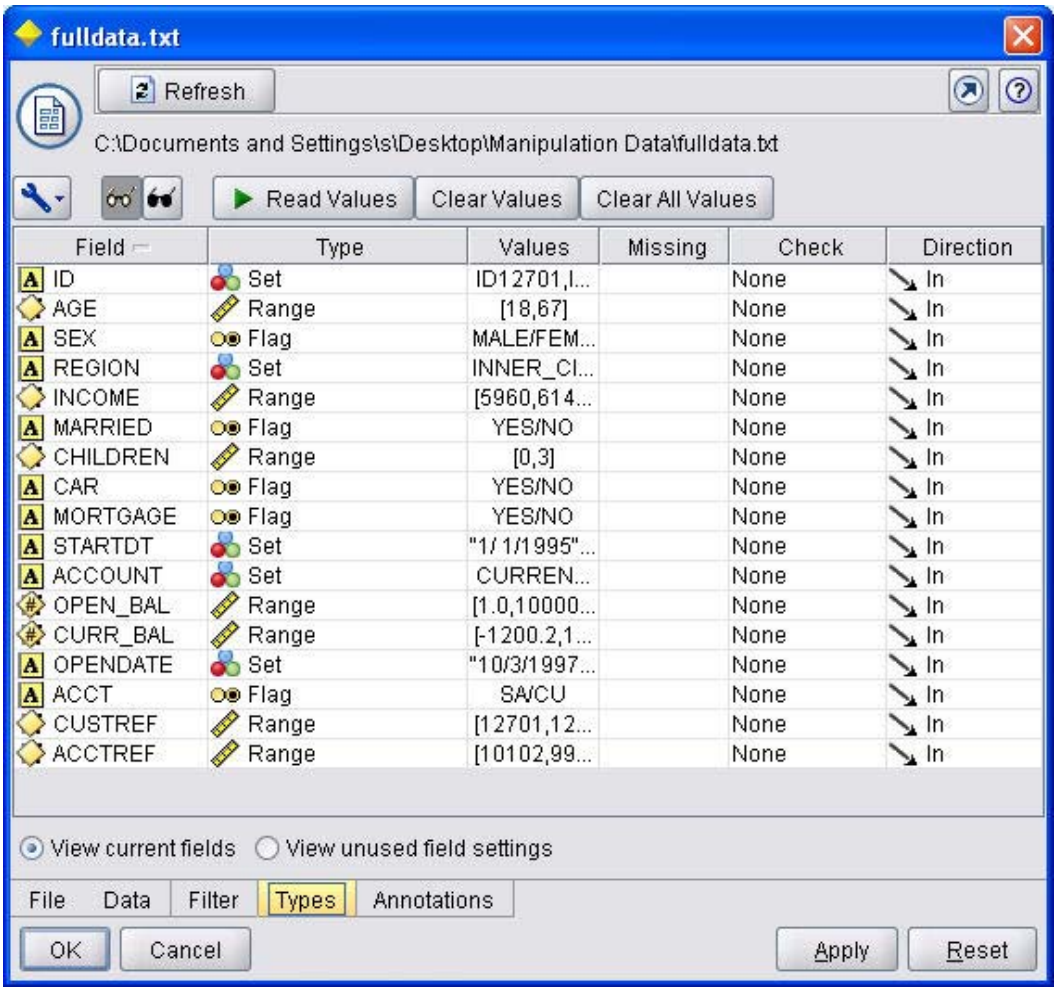
Discrete: Kaç farklı değer aldığı tahmin edilemeyen sözel verilere denir. Distinct veriler okunduktan sonra aldıkları değerlere göre:

- Set
- Flag
- Typless

olarak belirlenir.

Typeless: Flag, Set, Ordered Set ya da Range alan türlerinden hiçbirine uymayan alanlara denir. Kategori sayısı çok fazla olan set alanlar da "Typeless" atanır.

Maximum set size değeri 250'dir. Typless atanan alanları direction değerleri "None" olarak belirlenir.



Şekil 2.9: Clementine Arayüzünde Değişken Tiplerinin Kullanımı

2.9.2.4 Veri Depolama (Storage) Türleri

Clementine, veri işlemleri için aşağıdaki veri depolama türleri ile işlem yapmaktadır:

String: Sözel alan türü. String alanlar rakam da içerebilir ancak hesaplamalarda kullanılamaz.

Real: Ondalıklı sayıları da içerebilen alan türü (tam sayılarla sınırlı değildir)

Integer: Değerleri tam sayılar olan alan türü.

Time: Süre olarak ölçülmüş zaman bilgisini içeren alan formatıdır.
Örnek: 1 saat, 26 dakika, ve 38 saniye; 01:26:38 formatında gösterilir.

Timestamp: Süreden çok günün belirli bir saatini temsil eden zaman formatıdır.
Örnek: Saat 9'u 4 geçe 09:04:00 formatında gösterilir.

Date: Tarih değerlerini içeren alan formatıdır.
Örnek: 26 Eylül 2005 tarihi 2005-09-26 formatında gösterilir.

Real, Time, Timestamp ve Date formatları Stream Options diyalog kutusundan değiştirilebilir.

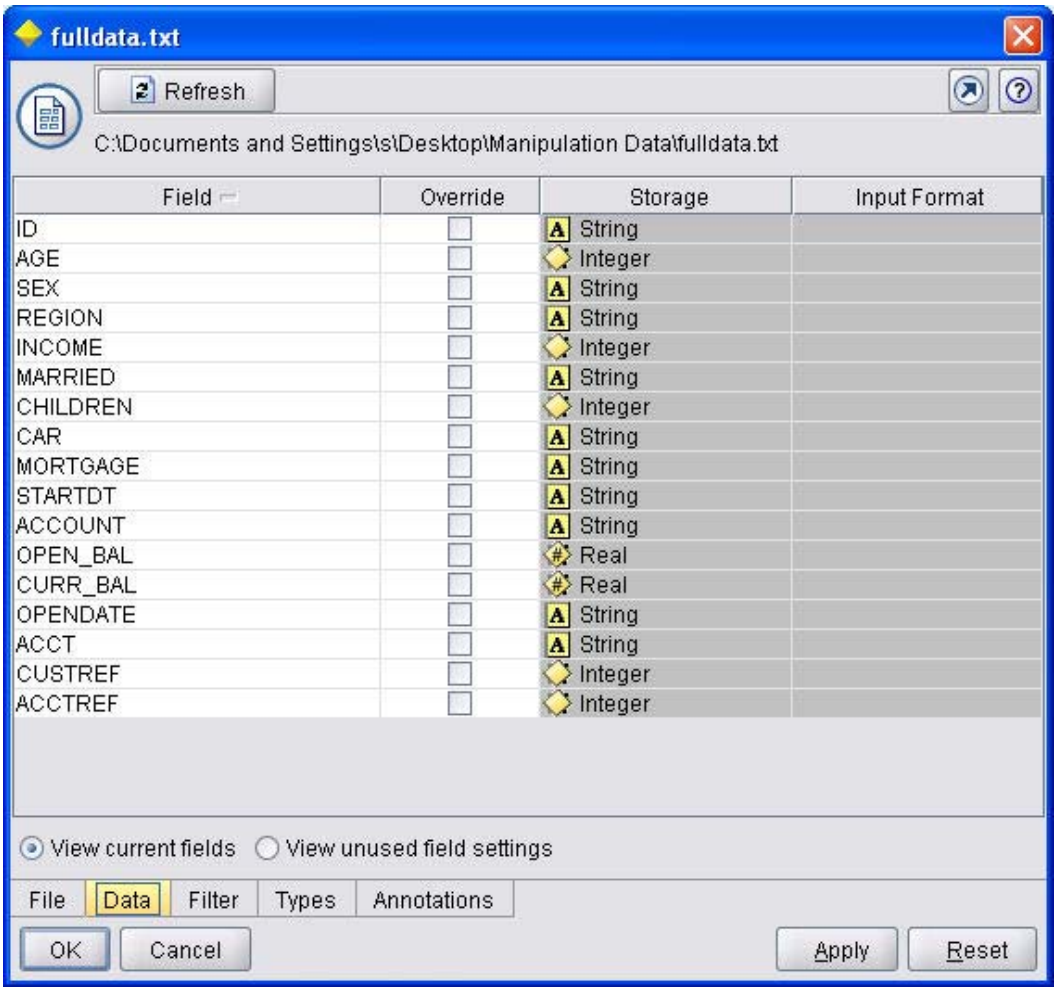
Eğer alanın storage türü ile bilgi var ancak değerler hakkında bilgi yok ise bu alanlara kısmi belirli alan adı verilir. İki tane kısmi belirli alan türü vardır. Bunlar;

Discrete: Alanın sözel bir alan olduğu bilinmektedir ancak alan değerleri bilinmemektedir.

Range: Alanın numerik olduğu bilinmektedir ancak alanın değerleri bilinmemektedir.

Eğer alanları storage'ı, türü, değerleri biliniyorsa bu alana tam belirli alan denir. Tam belirli alan türleri;

- Flag
- Set,
- Ordered Set
- Range (Hem kısmi hem de tam belirli alanlar için kullanılmaktadır)



Şekil 2.10: Clementine Arayüzünde Veri Depolama Türlerinin Kullanımı

Stream'in çalıştığı süreçte belirsiz alan türleri giriş değerleri baz alınarak, yarı belirli hale gelir. Tüm veri type işlemcisinden geçtiği anda bütün alanlar tam belirli hale gelir (Alan değerleri de belirlenmiş olur).

Eğer akışın çalışması yarıda kesilir ise alanlar kısmi belirli olarak kalır. Veriler bir kere tam belirli hale geldikten sonra akışın o noktasında artık veriler statiktir. "Type" işlemcisinden önce yapılacak herhangi bir değişiklik (akış baştan çalıştırılsa bile) alanların değerlerini etkilemeyecektir. Alanları değerlerini değiştirebilmek için <Read>, ya da <Read+> değerlerinin seçilmesi gerekmektedir.

2.10 Veri Kalitesinin İncelenmesi

Veri kümeleri, hemen her zaman hatalı ya da eksik değer içerirler. Bu yüzden veri madenciliğine başlamadan önce, verinin kalitesinin belirlenmesi gereklidir. Veri kalitesi ne kadar iyiyse elde edilecek sonuçlar da o kadar kesin olacaktır.

2.10.1 Kayıp Veriler

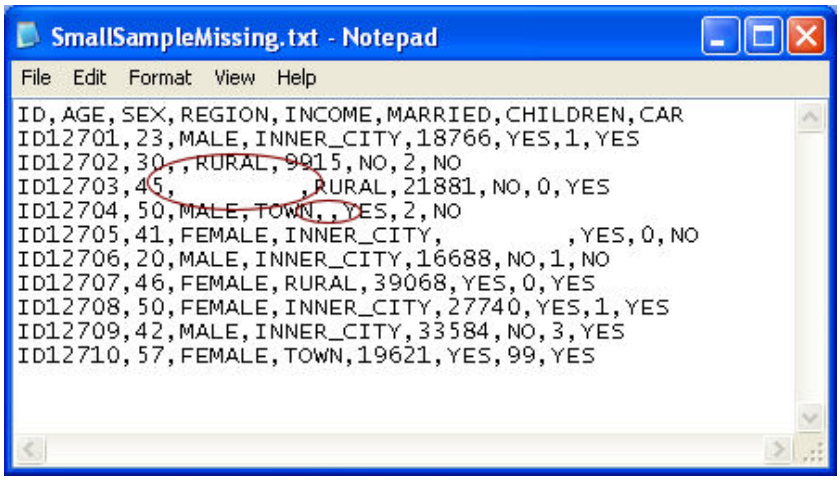
Clementine'da eksik veriler çeşitli gösterimlere sahiptir. Örneğin bir alan tamamen boş olabilir. Clementine'da bu tip alanlar sembolikse white space, sayısalsa null value olarak adlandırılır. Ayrıca sayısal bir alan sayısal olmayan bir karakter içerebilir. Clementine bu durumu da null value olarak adlandırır. Son olarak, veri girilirken önceden belirlenmiş bazı değerler eksik ya da yanlış bilginin gösterimi için kullanılmış olabilir. Clementine bu tip kodlamalara value blank adını verir.

Nulls (System Missing Values $null$): Sayısal alanlardaki boş değerlerdir.

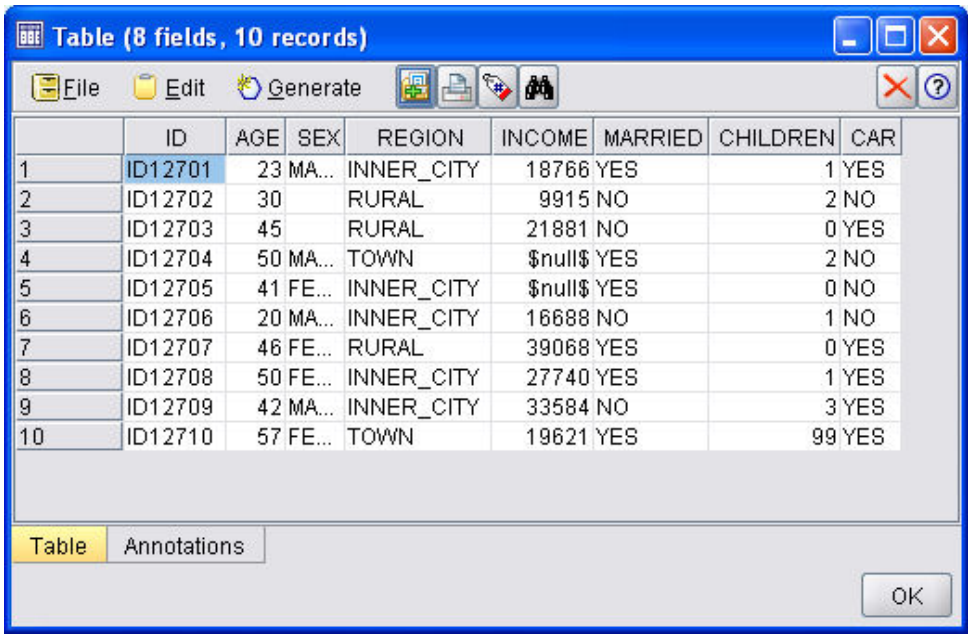
Blanks (User-defined missing values): Önceden tanımlanmış "geçersiz" yada "bilinmiyor" gibi anlamlara gelen (ör. 99, -1) özel kayıp değerlerdir.

Empty String: Sözel alanlardaki boş yani hiç bir değer girilmemiş anlamına gelir.

White Space: Sözel alanlardaki görünür olmayan karakter/lere (space) denir.



Şekil 2.11: Bir ASCII veri dosyasında bulunan kayıp değerler



	ID	AGE	SEX	REGION	INCOME	MARRIED	CHILDREN	CAR
1	ID12701	23	MA...	INNER_CITY	18766	YES	1	YES
2	ID12702	30		RURAL	9915	NO	2	NO
3	ID12703	45		RURAL	21881	NO	0	YES
4	ID12704	50	MA...	TOWN	$null$	YES	2	NO
5	ID12705	41	FE...	INNER_CITY	$null$	YES	0	NO
6	ID12706	20	MA...	INNER_CITY	16688	NO	1	NO
7	ID12707	46	FE...	RURAL	39068	YES	0	YES
8	ID12708	50	FE...	INNER_CITY	27740	YES	1	YES
9	ID12709	42	MA...	INNER_CITY	33584	NO	3	YES
10	ID12710	57	FE...	TOWN	19621	YES	99	YES

Şekil 2.12: ASCII veri dosyasınındaki kayıp değerlerin Clementine arayüzündeki görüntüsü

Clementine'da, "Type" işlemcisinde yer alan -"Check" kolonu- limitlerin dışında gözlemler belirlendiğinde ne yapılacağını belirler. Yapılacak işlemler aşağıdaki gibidir:

None: "Check" ayarı kapalı anlamına gelir. Varsayılan olarak bütün alanlar None'dır.

Nullify: Limitlerin dışında kalan değerleri null ($null$) olarak değiştirir.

Discard: Geçersiz değerler bulunur ve geçersiz değere sahip gözlem veriden çıkarılır.

Warn: Verinin okutulduğunda bütün geçersiz gözlemler sayılır ve bulunan sayı "Stream Properties" diyalog kutusunda raporlanır.

Abort: Belirlenen ilk geçersiz değer ile akışın çalışması durdurulur. Hata "Stream Properties" diyalog kutusunda raporlanır.

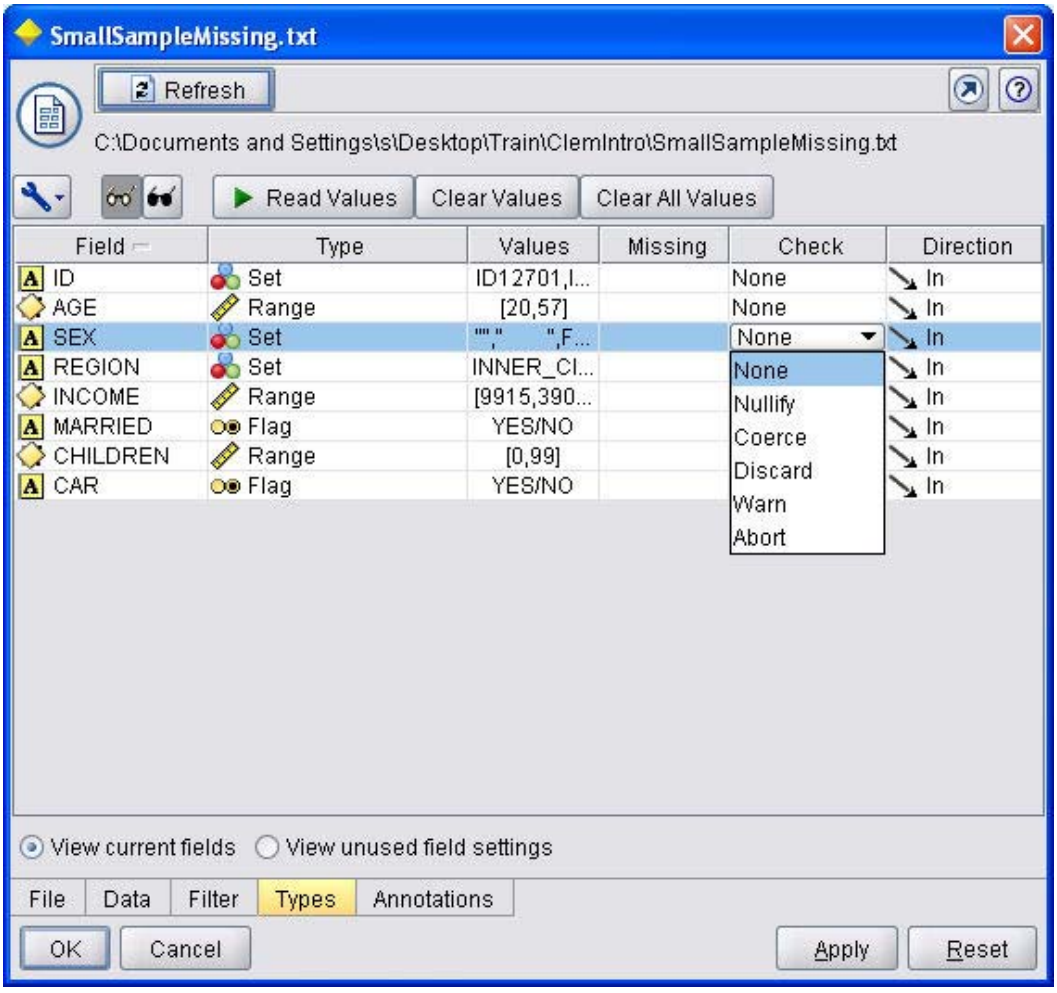
Coerce: Bütün alanlar tarandığında belirlenen limitlerin dışında kalan gözlemler aşağıdaki kurallara göre doldurulur:

- Flag değerler için; True ve False değerlerinde farklı her değer False değerine dönüştürülür.

- Set değerler için; set değerler listesinde yer alan birinci değere dönüştürülür.

- Range alanlarda üst limitin üzerindeki değerler üst limit değeri ile, alt limitin altındaki değerler alt limit değeri ile değiştirilir.

- Range alanlarda bulunan Null değerler alanın orta nokta değeri ile doldurulur.



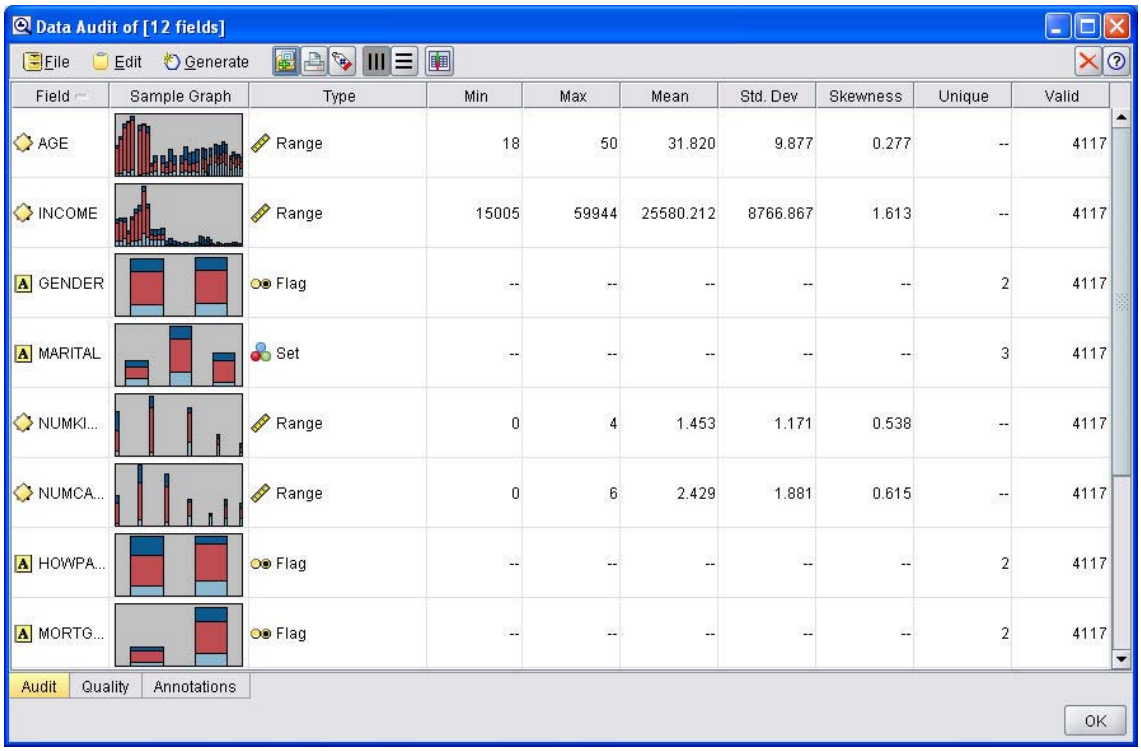
Şekil 2.13: Clementine Arayüzünde Type işlemcisi

Clementine'da veri kalitesi Data Audit Node, Plot Node ve Histogram Node isimli nodlar ile incelenmektedir.

2.10.1.1 Data Audit Node

Bir veri seti tam veriler içermesine karşın anormal değerler içerebilir. Data Audit nodunu kullanarak, her alan hakkında detaylı bilgi elde edebiliriz.

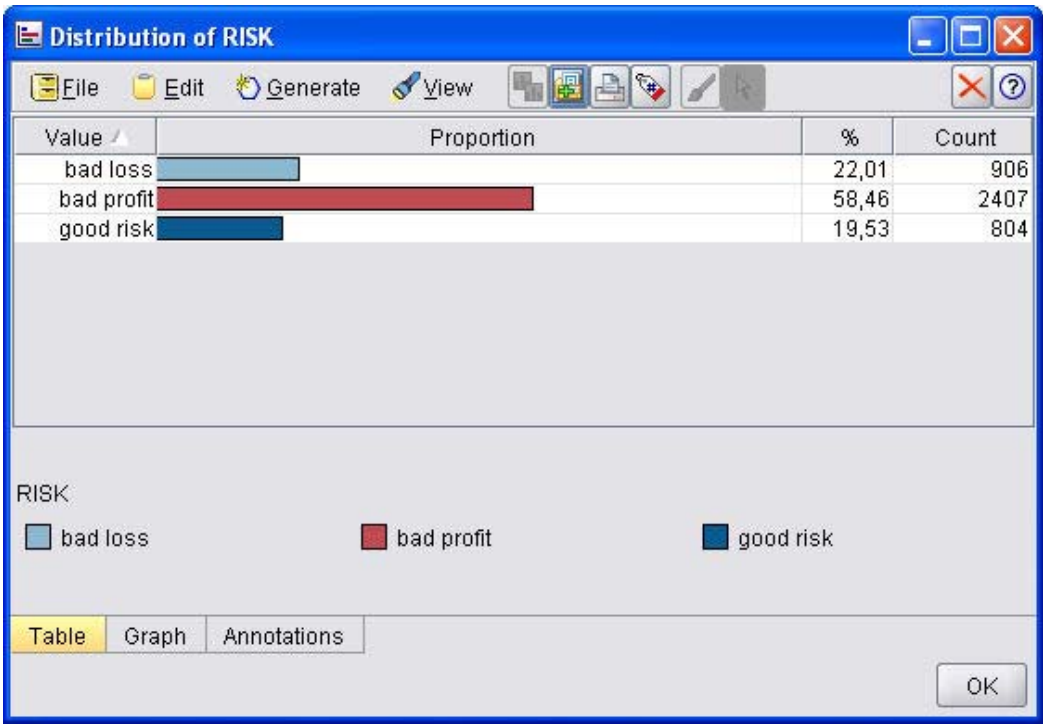
Data Audit nodundaki her satırın bir alana karşılık geldiğini; her sütunun da grafik, tip bilgisi ve istatistiksel özetler içerir. Bir alanın tipi Range ise histogramların, Set ya da Flag ise çubuk grafikler kullanılır.



Şekil 2.14: Bir veri setinin veri kalitesinin Data Audit Node ile incelenmesi

2.10.1.2 Plot Node

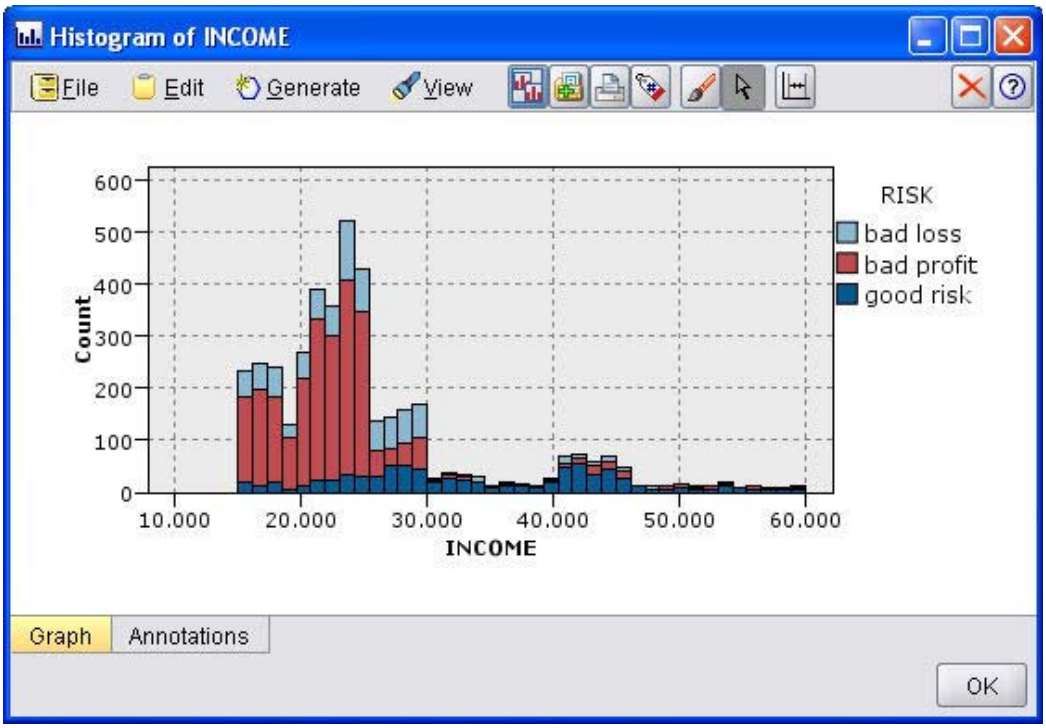
Plot Node, sembolik alanların dağılımlarının incelenmesinde kullanılır. Data Audit çıktısının her bir sembolik alanı için, bir dağılım grafiği (distibution plot) çizilebilir.



Şekil 2.15: Plot Node ile sembolik bir alanın dağılım grafiğinin görüntülenmesi

2.10.1.3 Histogram Node

Histogram Node, nümerik alanların dağılımlarının incelenmesinde kullanılır. Data Audit çıktısının her bir nümerik alanı için, bir histogram grafiği çizilebilir. Oluşan histogram, sayısal alanların içerdikleri değerlerin frekanslarını görmemize olanak sağlar.



Şekil 2.16: Histogram Node ile nümerik bir alanın histogram grafiğinin görüntülenmesi

2.11 Veri Manipulasyonu

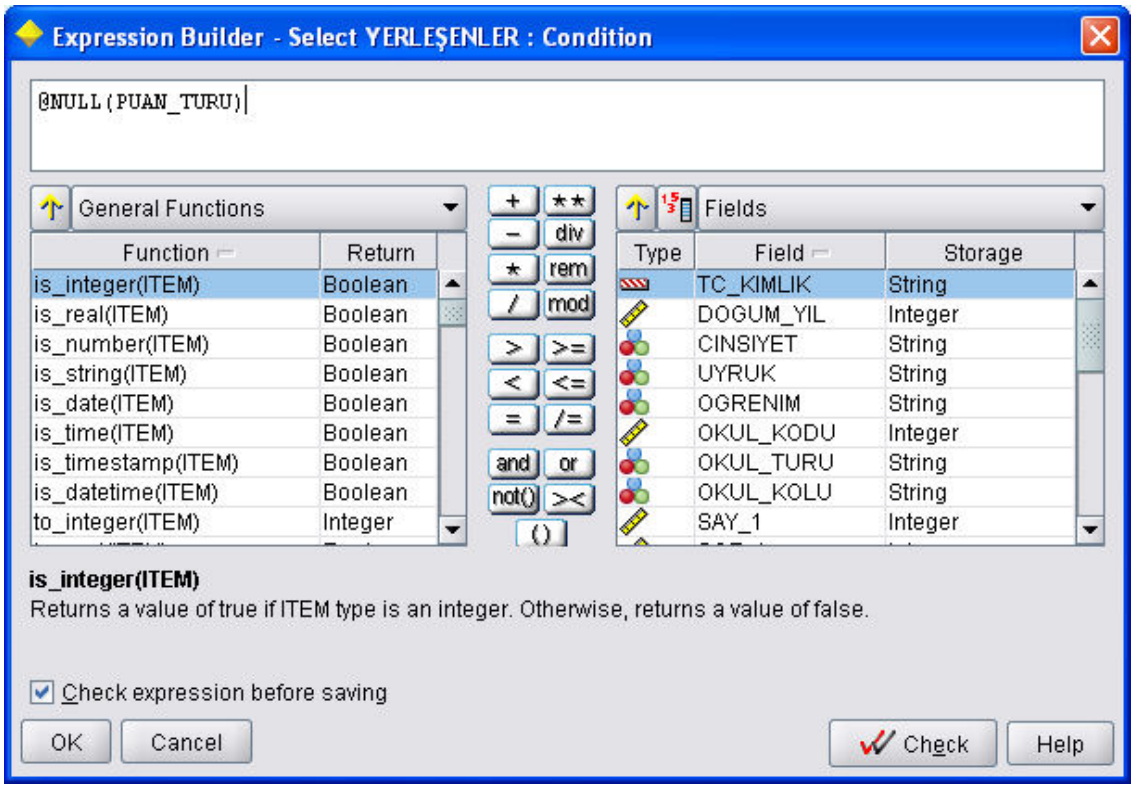
Bu bölümde, verilerin Clementine içerisine entegre edilmiş veri manipulasyon teknikleri ile nasıl veri madenciliğine uygun hale getirileceğini inceleyeceğiz. Clementine, kayıtların manipulasyonu için Record Ops paletini, alanların manipulasyonu için ise Fields Ops paletini içerir. Bu paletlerin içerdiği nodların bir kısmında veri manipulasyonu için CLEM programlama dilinin kullanılması mümkündür.

2.11.1 CLEM Programlama Dili

CLEM (Clementine Language for Expression Manipulation) stream boyunca ilerleyen verinin manipulasyonu için kullanılan Clementine'a özel bir dildir. CLEM dilinin Derive, Select, Filter, Balance ve Report nodlarının içinde kullanılması ile aşağıdaki işlemler gerçekleştirilir:

- Koşulların karşılaştırılması ve hesaplanması

- Yeni alanların oluşturulması
- Raporların oluşturulması



Şekil 2.17: Clementine içerisinde "Expression Builder" ile CLEM dilinin kullanımı

2.11.2 Kayıt İşlemleri (Record Operations) Paleti

Kayıt işlemleri paleti, kayıtların sıralanması, seçilmesi, birleştirilmesi, eklenmesi ve dengelenmesi için çeşitli nodlar içerir.

Şekil 2.18: Clementine'da "Record Ops" paletinin görünümü

2.11.2.1 "Select" Nodu

Select nodu, bir grup kaydın önceden belirlenmiş bir koşula bağlı olarak seçilmesi için kullanılır.

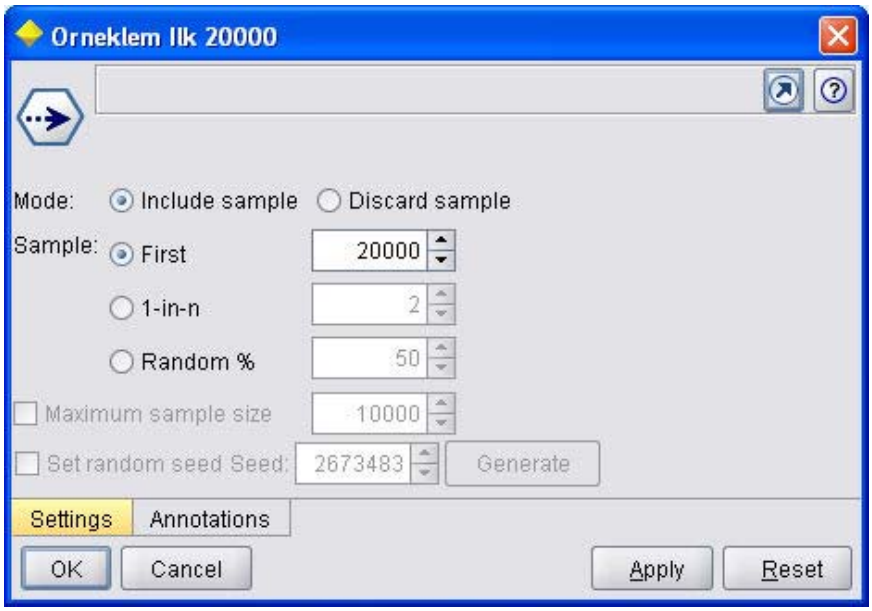
Belirlenen kriterlere uymayan kayıtlar elenir, kriterlere uyan kayıtların akış içerisinde devam etmesine izin verilir.

Şekil 2.19: "Select" nodunun görüntüsü

2.11.2.2 "Sample" Nodu

Bu nodun kullanımı ile veri setindeki kayıt sayısı sınırlanarak örneklem seçim işlemi yapılabilir. Veri Madenciliğinde asıl veriden örneklem seçilmesinin nedenleri aşağıdaki gibidir:

- Performansı arttırmak
- Veriyi eğitmek ve test etmek amaçlı iki parçaya ayırmak



Şekil 2.20: "Sample" nodunun görüntüsü

Sample nodunun ayarlarında kullanılan parametrelerin açıklaması aşağıdaki gibidir:

Mode: Seçilen kaydın örneklemde kullanılmasını (Include sample) ya da dışarıda bırakılmasını (Discard Sample) sağlar.

First: Veri setinin belirli bir sayıdaki ilk kayıtlarını örneklem olarak seçmeye yarar.

1-in-n: Her n kayıttan bir tanesinin seçilmesini sağlar. Örneğin n=2 ise her 2 kayıttan bir tanesi seçilir.

Random %: Kullanıcının belirlediği oranda tesadüfi olarak örneklem seçilir.

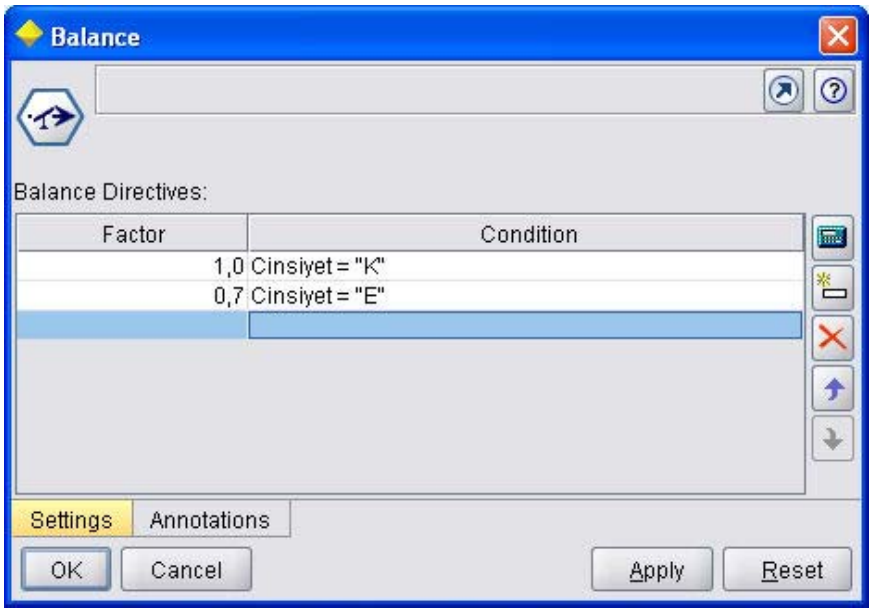
Maximum sample size: Maksimum örneklem genişliği isteğe bağlı olarak bu parametre alanı ile belirlenebilir.

Set random seed: Rakamın değiştirilmesi ile farklı bir örneklem seçilmesi sağlanır.

2.11.2.3 "Balance" Nodu

Balance nodu kullanılarak veri setindeki dengesizlik giderilebilir. Veri setinde iki farklı değeri olan bir alan ile hedef değişkenin tahmin edilmesi istendiğinde, bu alan değerlerinden bir tanesinin veri setinde bulunma sayısı diğerinden daha fazla ise, tahmini modelleme yöntemleri bu fazla bulunan değeri öğrenecek ve modelin performansı bu nedenle daha düşük olacaktır. Balance nodu ile verinin dengelenmesi durumunda, bu değerler yaklaşık olarak eşit orana getirilir ve böylece performansı yüksek tahminler yapılabilir.

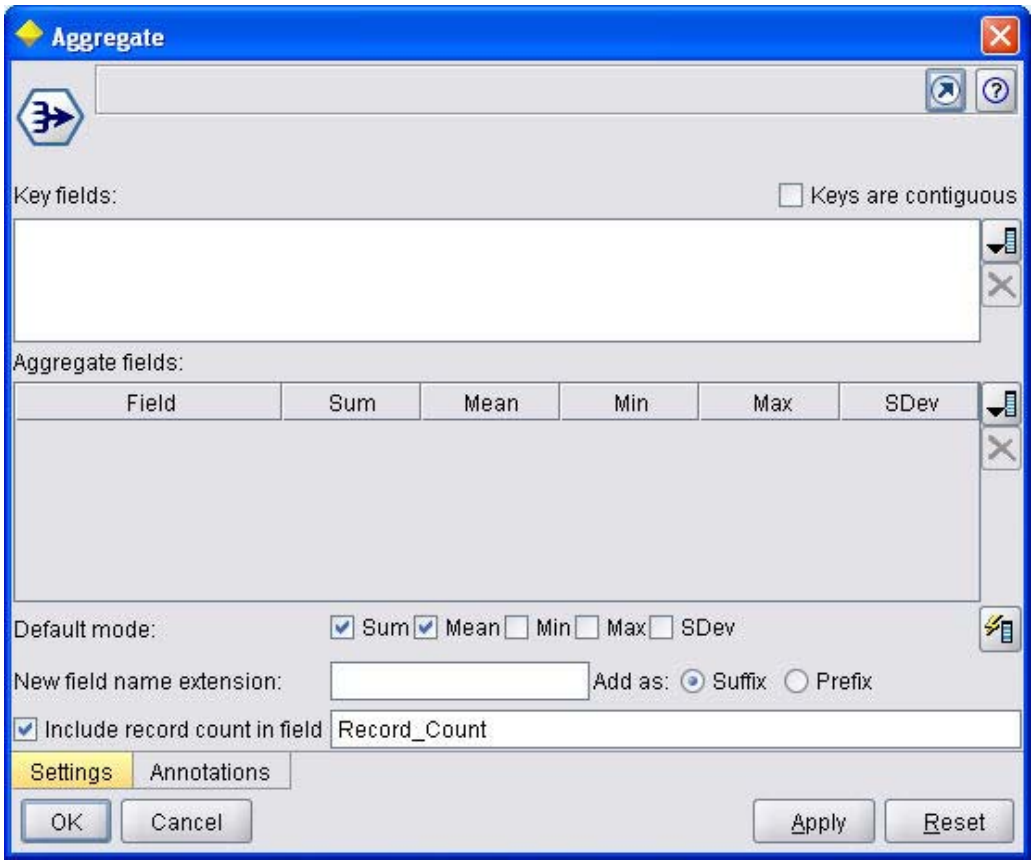
Bu nodun ayarlarlarında bulunan "Factor" kolonuna 1.0'den küçük bir sayı girildiğinde koşula uyan kayıtların veri seti içindeki oranı düşürülür. 1.0'den büyük bir sayı girildiğinde ise o oranda dublike kayıtlar yaratılarak ilgili koşula denk gelen kayıtların oranı artırılır. Balance nodu kullanılması ile veri sıralamasının değişmesi söz konusudur.



Şekil 2.21: "Balance" nodunun görüntüsü

2.11.2.4 "Aggregate" Nodu

Aggregate nodu, veri setindeki kayıtlara ait özel bilgileri ile yeni bir veri seti yaratılması amacıyla kullanıılır. Bu nodu kullanmadan önce verinin temizlenmesi ile, kayıp değerlerin içerdiği önemli bilginin kaybolması engellebilir.



Şekil 2.22: "Aggregate" nodunun görüntüsü

Aggregate nodunun ayarlarında kullanılan parametrelerin açıklaması aşağıdaki gibidir:

Key Fields: Seçilen alanlardaki değerlerin detayında "Aggregate Fields" kısmında tanımlanan alanların özetini çıkarır.

Aggregate Fields: Özet bilgilerin hesaplanacağı alanlar bu kısımdan seçilir.

Sum: Toplam hesaplanır

Mean: Ortalama hesaplanır

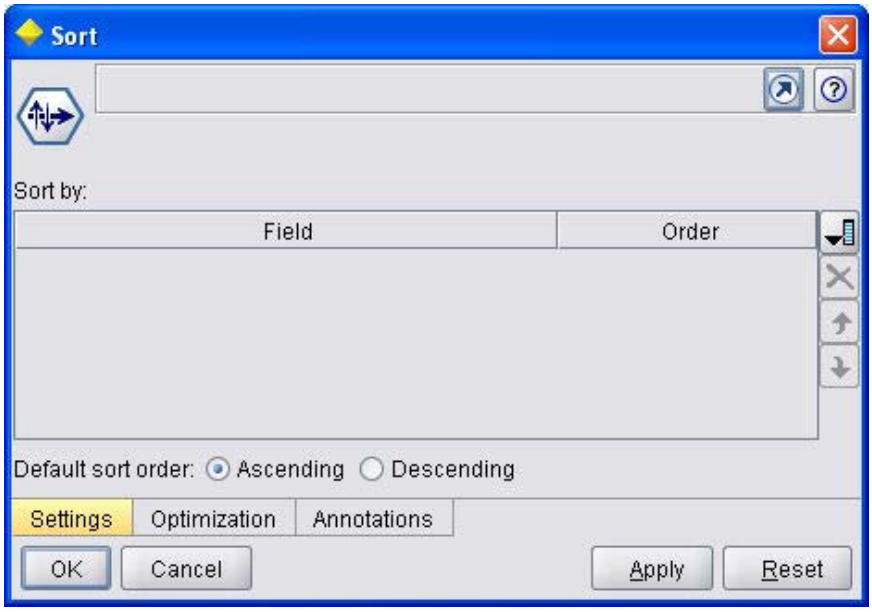
Min: Minimum değer hesaplanır

Max: Maksimum değer hesaplanır

SDev: Standard sapma hesaplanır

2.11.2.5 "Sort" Nodu

Sort nodu veri setinin bir ya da birden fazla alana göre sıralanmasını sağlar.



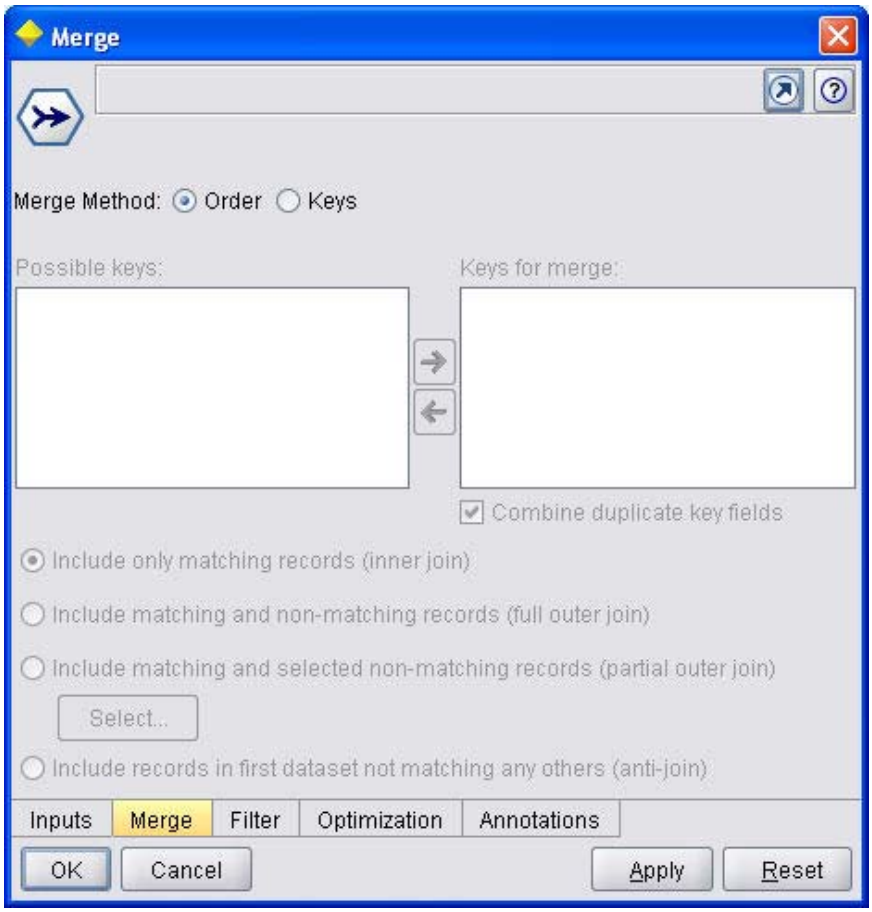
Şekil 2.23: "Sort" nodunun görüntüsü

2.11.2.6 "Merge" Nodu

Merge nodu ile farklı kaynaklardan gelen kayıtları birleştirerek tek bir kayıt oluşturulur. Clementine'da birleştirme iki türlü yapılır:

Merge by order: Farklı kaynaklardan gelen verileri kayıt sırasına göre eşleştirir. Bu seçeneğin kullanılması durumunda verilerin sıralanmış olmasına dikkat edilmelidir.

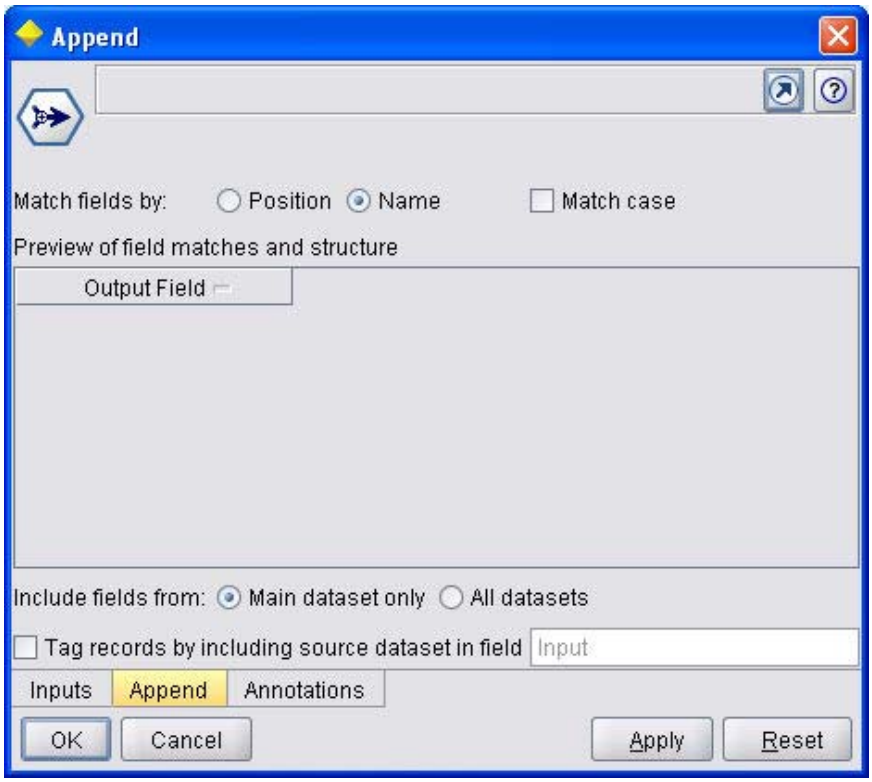
Merge using a key field: Farklı kaynaklardan gelen verilerin eşleştirilmesini belirlenen alana göre yapar.



Şekil 2.24: "Merge" nodunun görünümü

2.11.2.7 "Append" Nodu

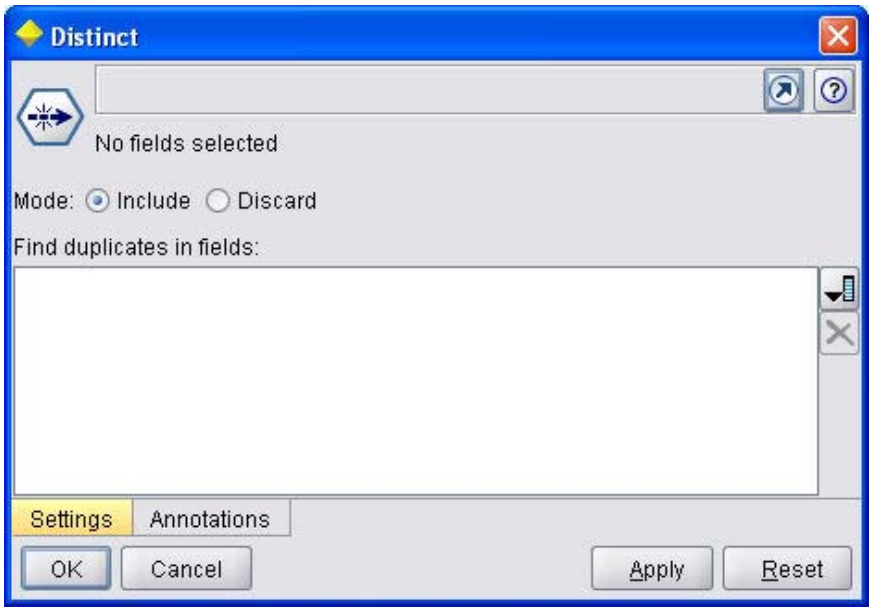
Append nodu, çok sayıda girdiye izin verir. Veri kaynaklarından gelen tüm kayıtları, birinci veri kaynağından başlayarak okur ve sonraki nodlara iletir. Kaynakların sırasını, Append noduna bağlanma zamanları belirler. İlk bağlanan nod, main dataset (esas veri kümesi) olarak adlandırılır ve varsayım olarak Append nodundan çıkan birleştirilmiş verinin formatını belirler.



Şekil 2.25: "Append" nodunun görünümü

2.11.2.8 "Distinct" Nodu

Distinct nodu ile kullanıcı tarafından belirlenen alanlar temel alınarak tekrarlanan kayıtlar kontrol edilir ve tekrarlayan kayıtlar içinde sadece birinci kayıt ya da birinci hariç tüm kayıtlar elenir.



Şekil 2.26: "Distinct" nodunun görünümü

2.11.3 Alan İşlemleri (Field Operations) Paleti

Alan işlemleri paleti, veri alanlarının yönetimi ile ilgili nodları içerir. Bu nodlar aşağıdaki gibidir:

Filter: Belirli veri alanlarının bu akıştan kaldırılmasını sağlar

Field Reorder: Veri alanlarının, diyalog kutularındaki ve veri akışındaki sıralarını değiştirmek için kullanılır.

Derive: Yeni veri alanlarının oluşturulması için kullanılır.

Reclassify: Set ya da Flag tipli verilerin yeniden sınıflandırılmasını sağlar.

Binning: Nümerik bir veri alanı otomatik olarak gruplara ayrılarak bu grupları işaret eden set tipinde yeni alan oluşturulmasını sağlar.

Filler: Belirtilen koşula uyan kayıtların belirlenen bir değerle doldurulmasını sağlar.

Set to Flag: Set tipindeki bir alanın kategorilerini işaret eden flag tipinde yeni alanlar yaratılması için kullanılır.

History: Genellikle zaman serileri gibi dizi şeklindeki verilerde kullanılır. Önceki kayıtları kullanarak yeni alanlar yaratır.

3. Üniversitelerin Adaylar Tarafından Tercih Edilme Desenlerinin Belirlenmesi

Bu bölümde, üniversitelerin adaylar tarafından tercih edilme desenlerinin belirlenmesi amacına yönelik olarak, Clementine içerisinde CRISP-DM metadolojisine uygun olarak öncelikle iş ihtiyaçlarının tanımlandığı "Problemin Tanımlanması" aşaması üzerinde çalışılacaktır. Daha sonra veri manipulasyon işlemlerinin yapıldığı "Verilerin Hazırlanması" adımı üzerinde çalışılacak ve ardından "Modelin Kurulması ve Değerlendirilmesi" aşamasında gerekli modeller hazırlanacaktır. Son olarak "Modelin Kullanılması" aşamasında geliştirilen modeller kullanılacak ve diğer kullanım alanlarından bahsedilecektir.

3.1 Problemin Tanımlanması

Üniversitelerde yer alan kontenjanların yüksek oranda ve nitelikli adaylar ile dolması üniversite üst yönetimlerinin temel hedeflerinden biridir. Son yıllarda öncelikle vakıf üniversiteleri olmak üzere bütün üniversitelerin tanıtım faaliyetlerinde temel hedef, bu olmuştur.

Etkin ve verimli bir tanıtım faaliyetinin planlanması ve yürütülmesi ancak üniversite adayları ile ilgili ÖSYM tarafından paylaşılan verilerin Veri Madenciliği yöntemleri ile incelenmesi ve anlamlı sonuçlar üretilmesi ile mümkündür.

Üniversite adaylarının farklı üniversite türlerini tercih ederken yansıttıkları desenlerin saptanması bu çalışmanın temel konusu ve amacıdır.

İstanbul Kültür Üniversitesi AR-GE Merkezi, yıllardır ÖSYM'nin açıkladığı ham verileri anlamlı sonuçlar çıkarmak amacıyla işlemektedir.

Geçmiş yıllarda yapılan çalışmalar çerçevesinde aşağıdaki parametrelerin kurulacak modelde girdi değişkenleri olarak kullanılmasına karar verilmiştir.

- Puan
- OBP
- Uyruk
- Öğrenim Durumu
- Okul Türü
- Okul Kolu
- Adayın geldiği il ile yerleştiği üniversitenin bulunduğu ilin eşitliği
- Cinsiyet

3.2 Verilerin Hazırlanması

3.2.1 ÖSYM Verileri Hakkında

ÖSYS'de herhangi bir puan türünde 160 puanı geçip tercih yapma hakkı kazanan adayların verisi, ÖSYM tarafından üniversitelerin kullanımına açılmaktadır.

Üniversitelere yerleştirme işlemleri tamamlandıktan sonra da, yerleşen adayların verileri üniversiteler ile paylaşılmaktadır.

Bu iki veritabanı kullanılarak 160 barajını geçen adayların hangi üniversite ve bölümlere yerleştiği tespit edilmiştir.

İKÜ AR-GE Merkezinin hazırladığı bu veritabanı, her aday için:

- TC_KIMLIK
- DOGUM_YIL
- CINSIYET
- UYRUK
- OGRENIM:
- OKUL_KODU
- OKUL_TURU
- OKUL_KOLU
- SAY_1
- SOZ_1
- EA_1
- DIL

- SAY_2
- SOZ_2
- EA_2
- ADRES_IL
- PUAN_TURU
- MEZUNIYET_YILI
- OBP
- UNI_TURU
- UNI_2
- UNI_SEHIR
- FAKULTE
- BOLUM ADI
- BURS
- IKINCI

alanlarını içermektedir.

Yapılan çalışmanın kuvvetli yönlerinden bir tanesi, modellemede kullanılan verinin popülasyonun tamamından oluşmasıdır. Bu özellik, modelin gerçeği yansıtma yeteneğini en üst düzeye çıkarmaktadır.

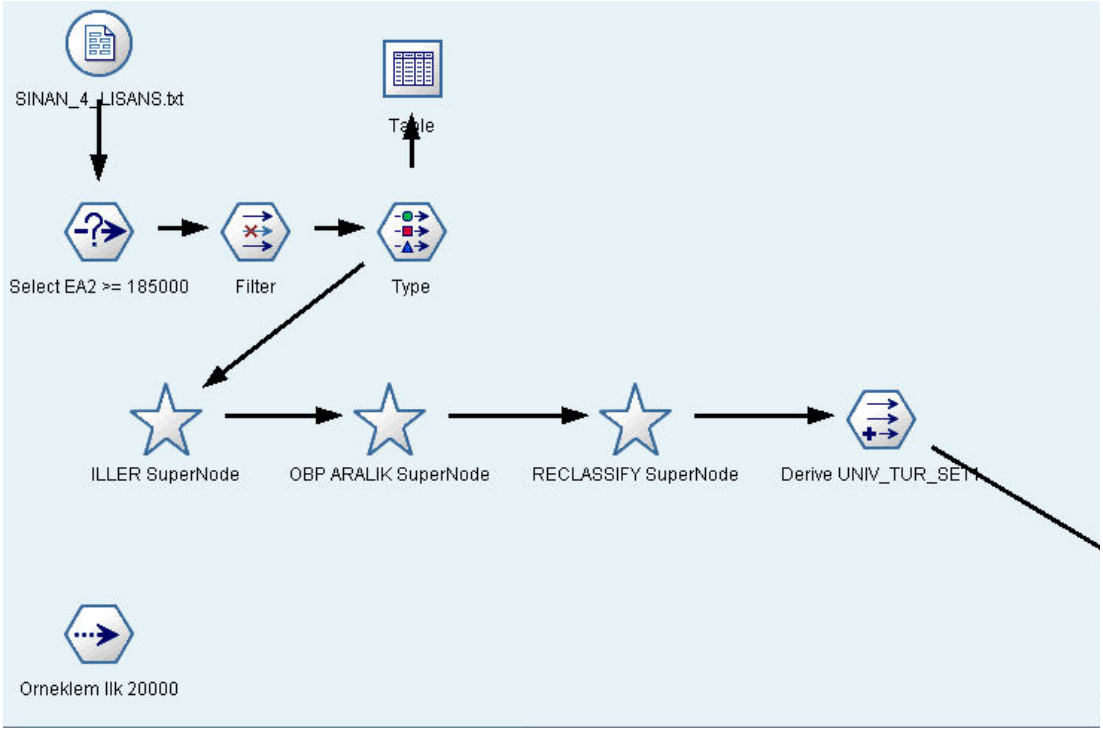
3.2.2 Veri Manipulasyonu

ÖSYM tarafından tüm popülasyon verileri eksiksiz ve düzgün olarak sağlandığından, aynı şekilde İKÜ AR-GE merkezi tarafından veri üzerinde eksiksiz ve düzgün eşleştirmeler yapıldığından, veri kalitesi üzerinde bir çalışma yapılmasına gerek olmamıştır.

Field	Type	Values
TC_KIMLIK	Typeless	
DOGUM_YIL	Range	[48,98]
CINSIYET	Set	1,2
UYRUK	Set	1,2,3
OGRENIM	Set	1,2,3,4,5,6,7,8
OKUL_KODU	Range	[10012,990009]
OKUL_TURU	Set	11017,11025,...
OKUL_KOLU	Set	1018,1024,10...
SAY_1	Range	[141636,3000...
SOZ_1	Range	[148783,3000...
EA_1	Range	[168105,3000...
DIL	Range	[0,300000]
SAY_2	Range	[0,300000]
SOZ_2	Range	[0,300000]
EA_2	Range	[185001,3000...
ADRES_IL	Set	0,1,2,3,4,5,6,7,...
PUAN_TURU	Set	7,8,9,10,11,12...
MEZUNIYET_YILI	Set	0,1980,1984,1...
OBP	Typeless	
UNI_TURU	Set	"","1,00","2,00"...
UNI_2	Set	"","ABANT IZZ...
UNI_SEHIR	Set	"",ADANA,ADIY...
FAKULTE	Set	"","Adana Sagl...
BOLUM ADI	Typeless	
BURS	Set	"","0,00","1,00"
IKINCI	Set	"","0,00","1,00"

Şekil 3.1: Clementine Type işlemcisinde düzenlenmiş 2007 ÖSYS verilerinin görüntüsü

Clementine içerisinde amaçlanan modellemeye uygun olarak veri üzerinde gerekli manipulasyon işlemleri yapılması amacıyla, akışa manipulasyon işlemleri ile başlanmıştır.

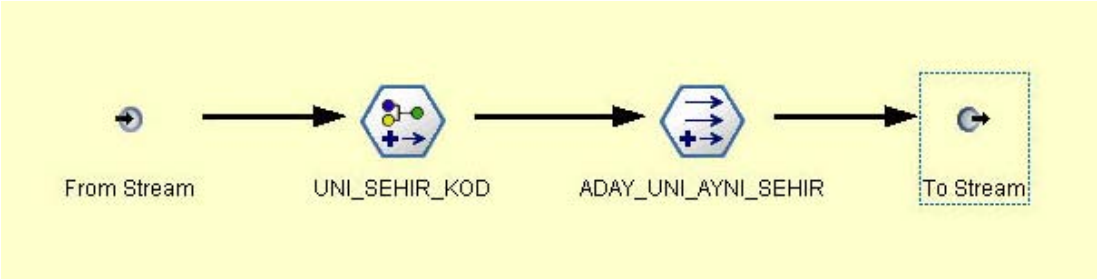


Şekil 3.2: Manipulasyon işlemlerinin Clementine arayüzündeki akış görünümü

İlk olarak tüm verilere "Var.File" nodu kullanılarak erişilmiştir. Daha sonra modellemesi yapılacak puan türünde, 4 yıllık lisans öğrenimine hak kazanmanın koşulu olan 185.000 puanın üzerinde puan alan adayların seçiminin yapılması için ilgili Select nodu kullanılmıştır.

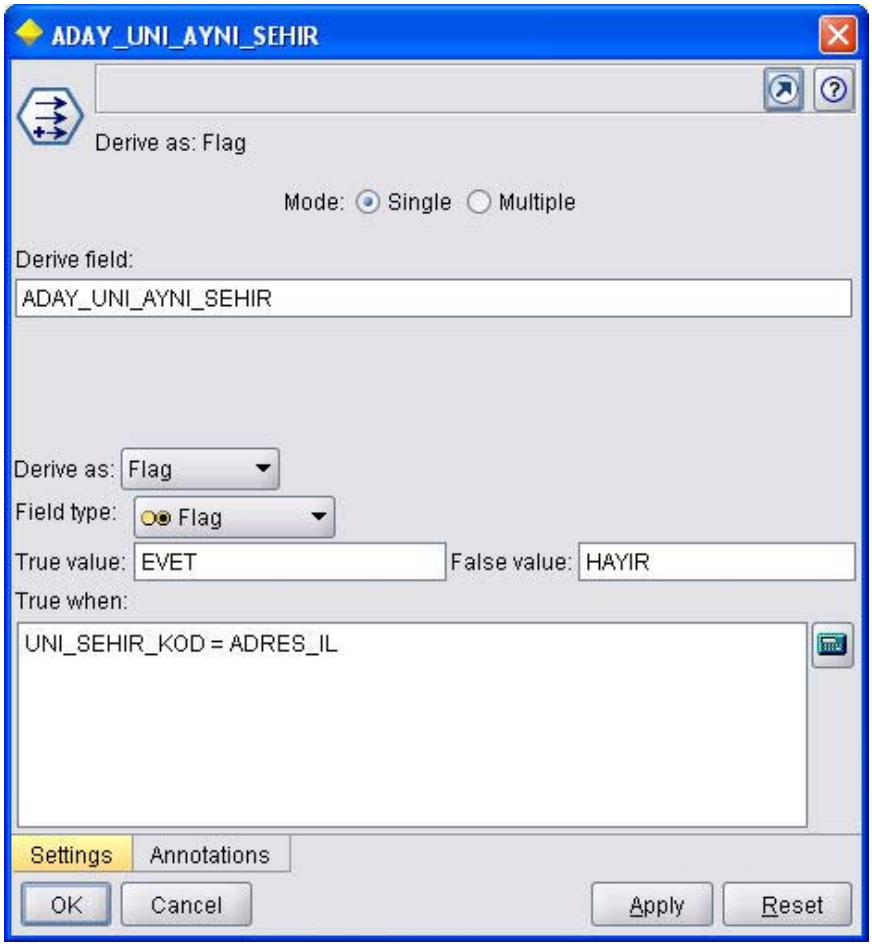
Manipulasyon işlemlerinin performansını yüksek tutmak açısından, veriden örneklem olarak ilk 20.000 kayıt seçilmiş ve bu kayıtlar baz alınarak veri manipulasyonu yapılmıştır.

Filter nodu kullanılarak, modelleme ve manipulasyon işlemlerinde kullanılmasına gerek görülmeyen veri alanları ayıklanmıştır. Type işlemcisi ile ilk veri okuması yapılıp, "ILLER" isimli supernod içerisindeki işlemler ile, adayın kendi ilindeki üniversiteye yerleşip yerleşmediği ile ilgili bilgi verecek yeni bir alanın oluşturulması amaçlanmıştır.



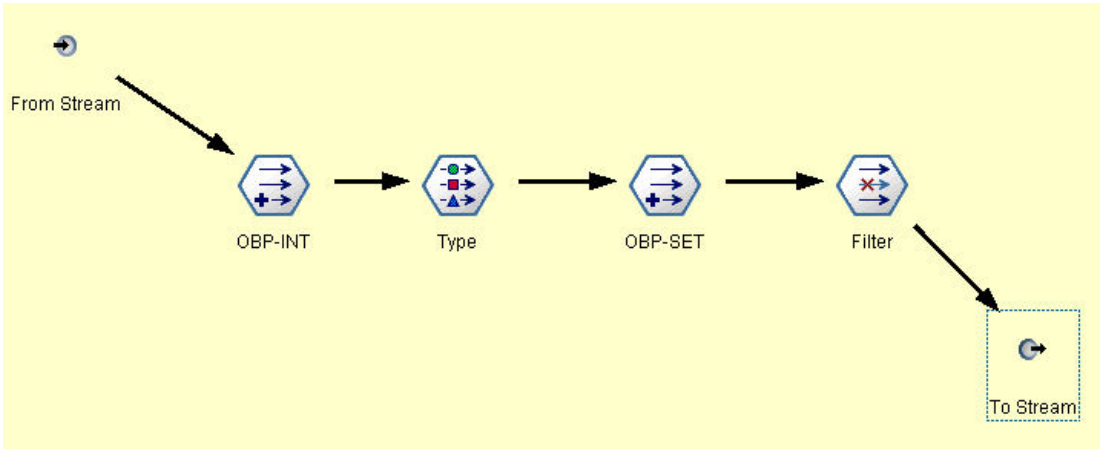
Şekil 3.3: "ILLER" supernodunun görüntüsü

Bu supernodun içersinde öncelikle Reclassify işlemcisi ile veritabanından gelen UNI_SEHIR alanındaki il isimleri, karşılaştırma yapılabilmesi için plaka numaraları ile değiştirilmiştir. Hemen ardından Derive işlemcisi ile ADAY_UNI_AYNI_SEHIR isimli, adayın kendi ilindeki üniversiteye yerleşip yerleşmediği konusunda bilgi veren yeni alan oluşturulmuştur.



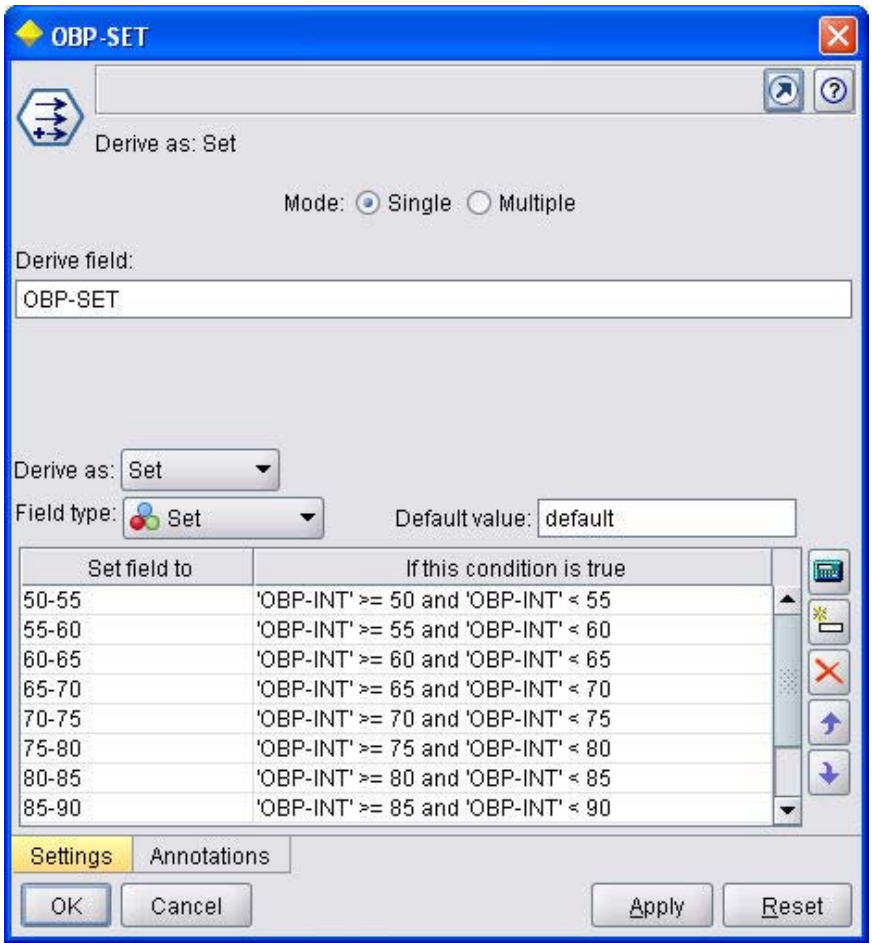
Şekil 3.4: Derive işlemcisi ile ADAY_UNI_AYNI_SEHIR alanının oluşturulması

Daha sonra, OBP supernodu içerisinde, adayların Ortaöğrenim Başarı Puanlarının segmentasyonunun hazırlanması amaçlanmıştır.



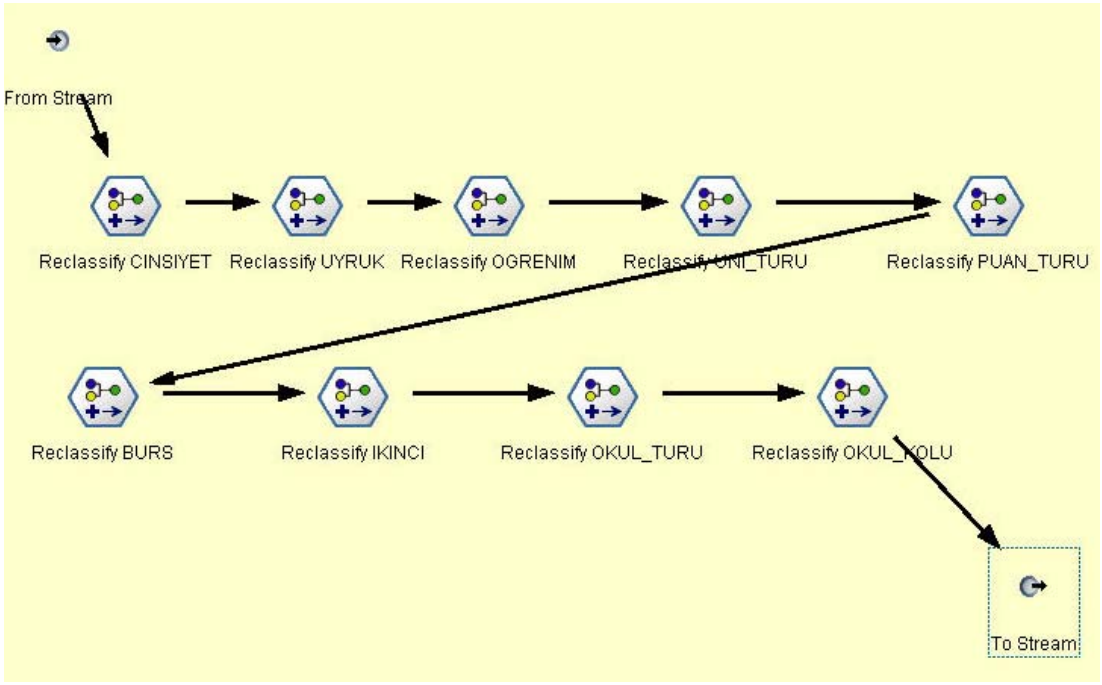
Şekil 3.5: OBP supernodunun görüntüsü

OBP supernodunun içerisinde öncelikle OBP alanı, üzerinde eşitsizlik karşılaştırma işlemlerinin yapılabileceği sayısal verilere dönüştürülmüştür. Type işlemcisi ile bu yeni alan okutulup, "OBP-SET" işlemcisi ile kullanılması düşünülen segmentasyon yapılmıştır.



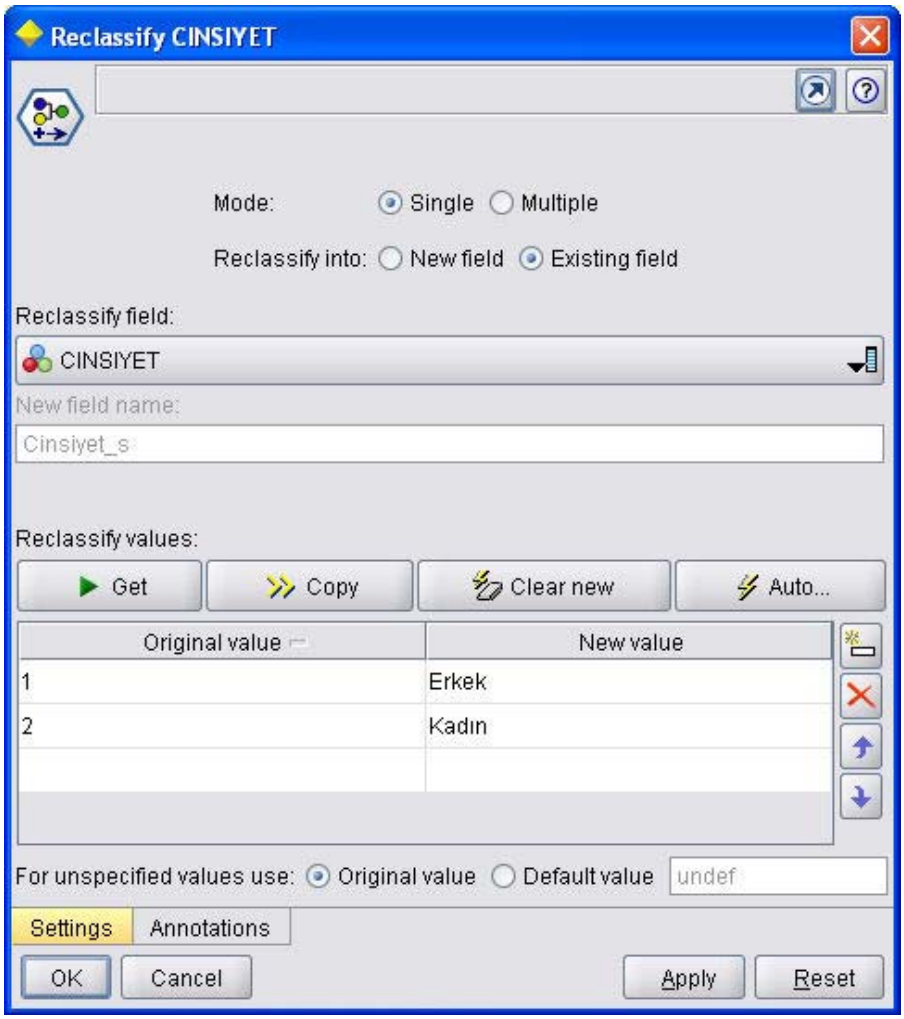
Şekil 3.6: Derive işlemcisi ile OBP-SET alanının oluşturulması

Daha sonra "RECLASSIFY" isimli supernod ile geriye kalan diğer veri manipulasyonlarının yapılması amaçlanmıştır.

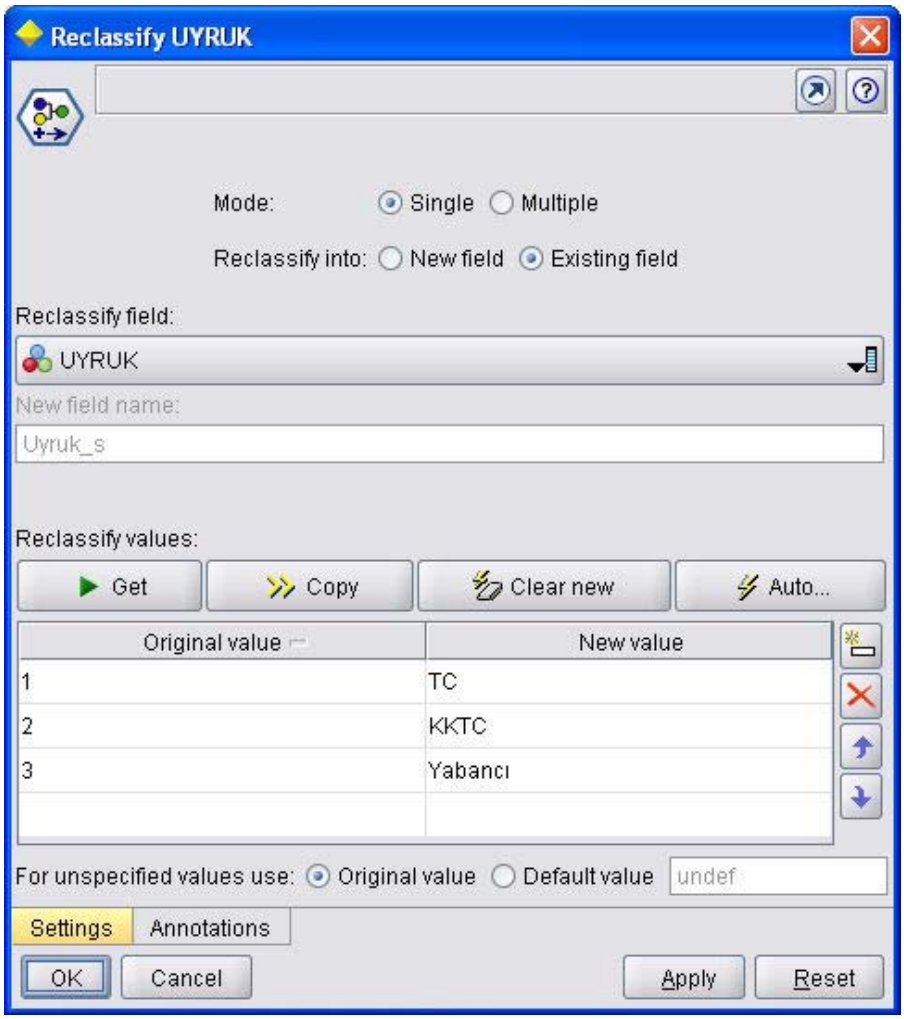


Şekil 3.7: RECLASSIFY supernodunun görüntüsü

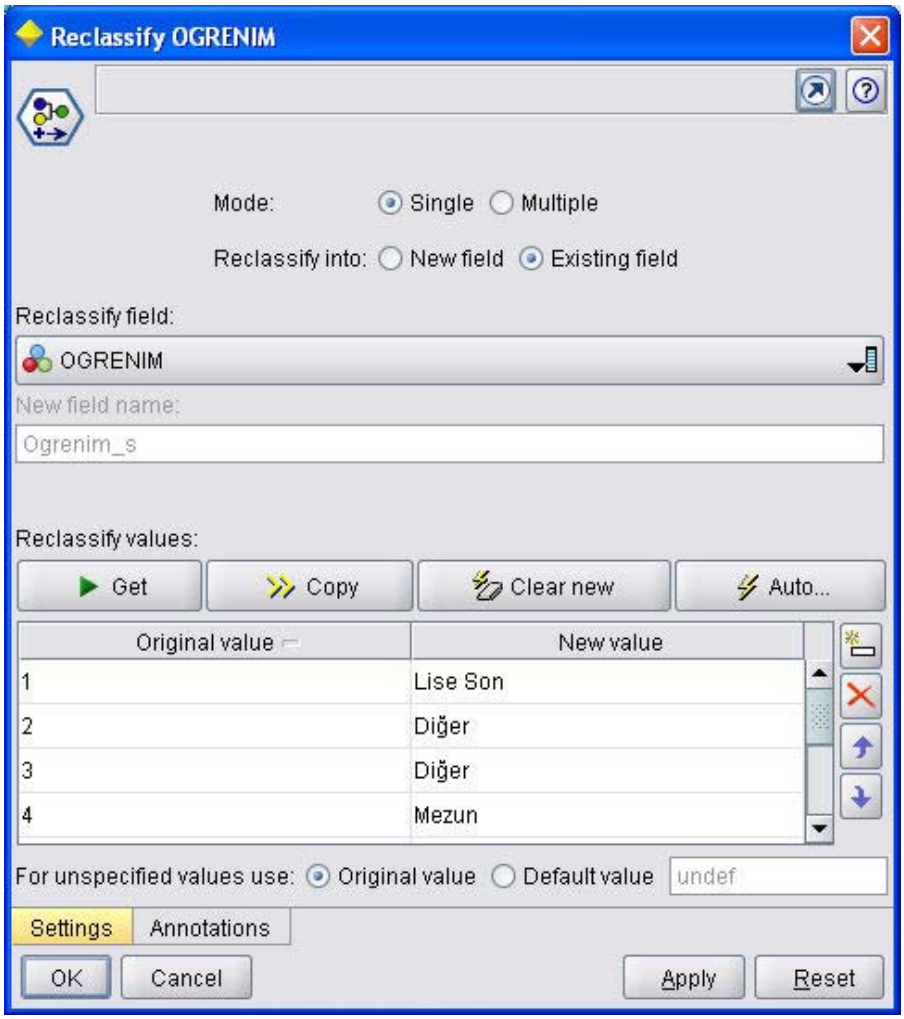
CINSIYET, UYRUK, OGRENIM, UNI_TURU, PUAN_TURU, BURS, IKINCI, OKUL_TURU, OKUL_KOLU veri alanları üzerinde, anlamlı sonuçların görüntülenmesi amacıyla, Reclassify işlemcileri kullanılarak gerekli manipulasyon işlemleri yapılmıştır.



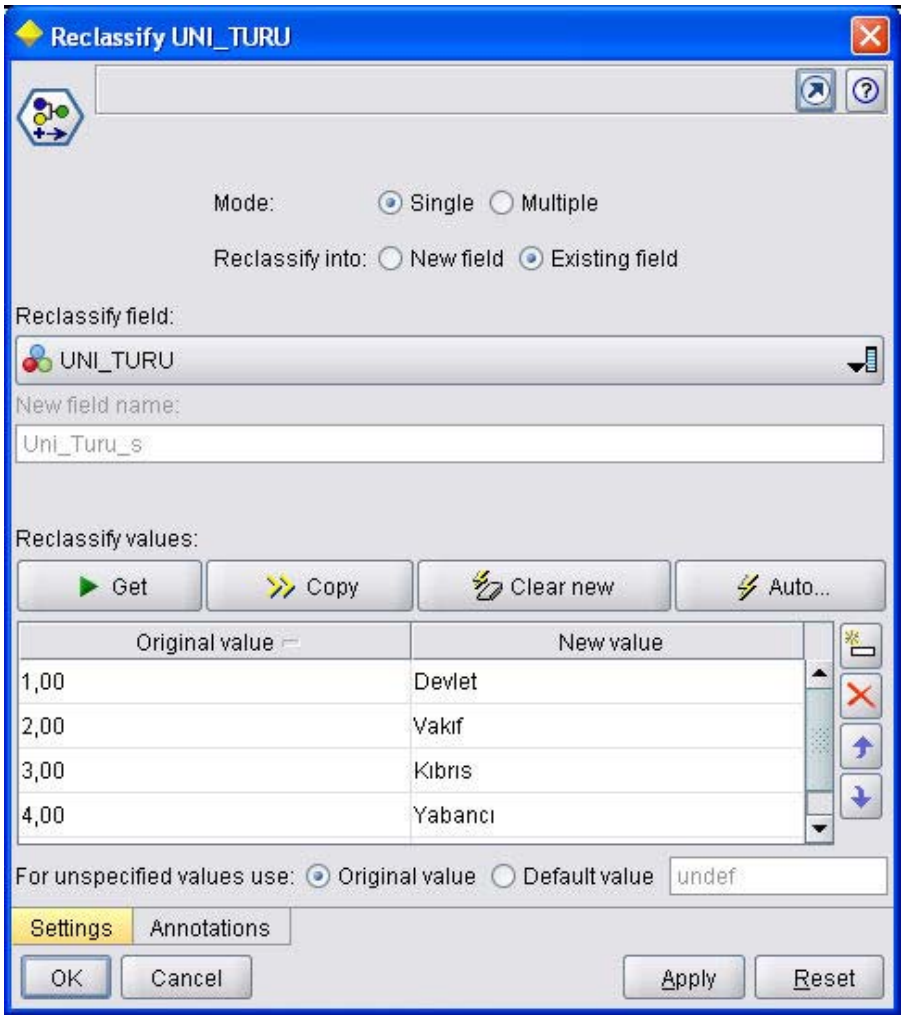
Şekil 3.8: Reclassify işlemcisi ile CINSIYET alanının manipulasyonu



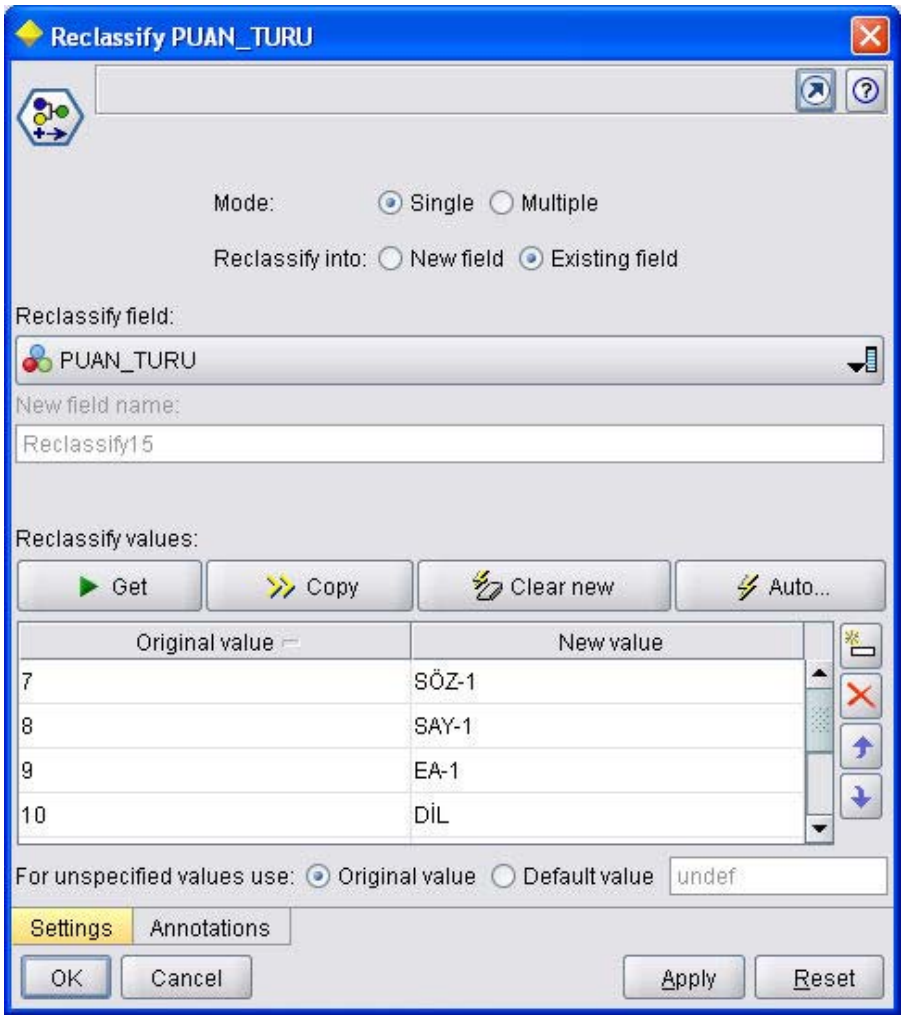
Şekil 3.9: Reclassify işlemcisi ile UYRUK alanının manipulasyonu



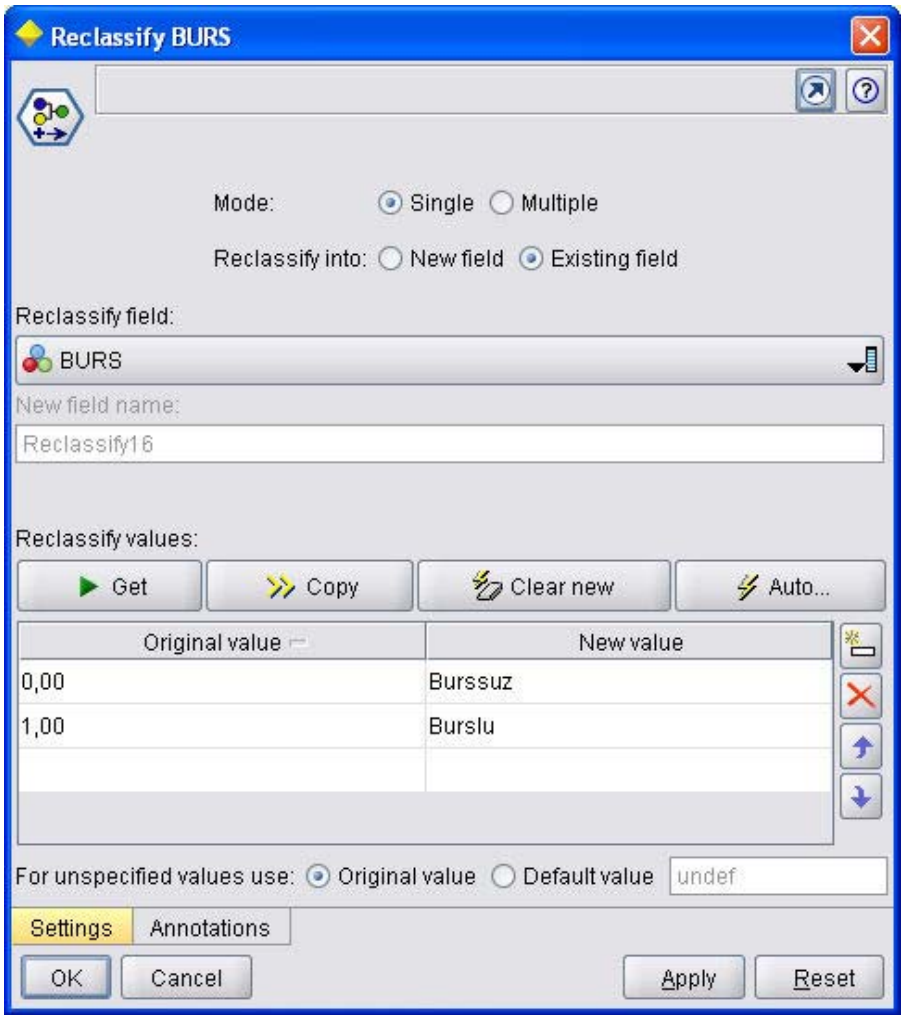
Şekil 3.10: Reclassify işlemcisi ile OGRENIM alanının manipulasyonu



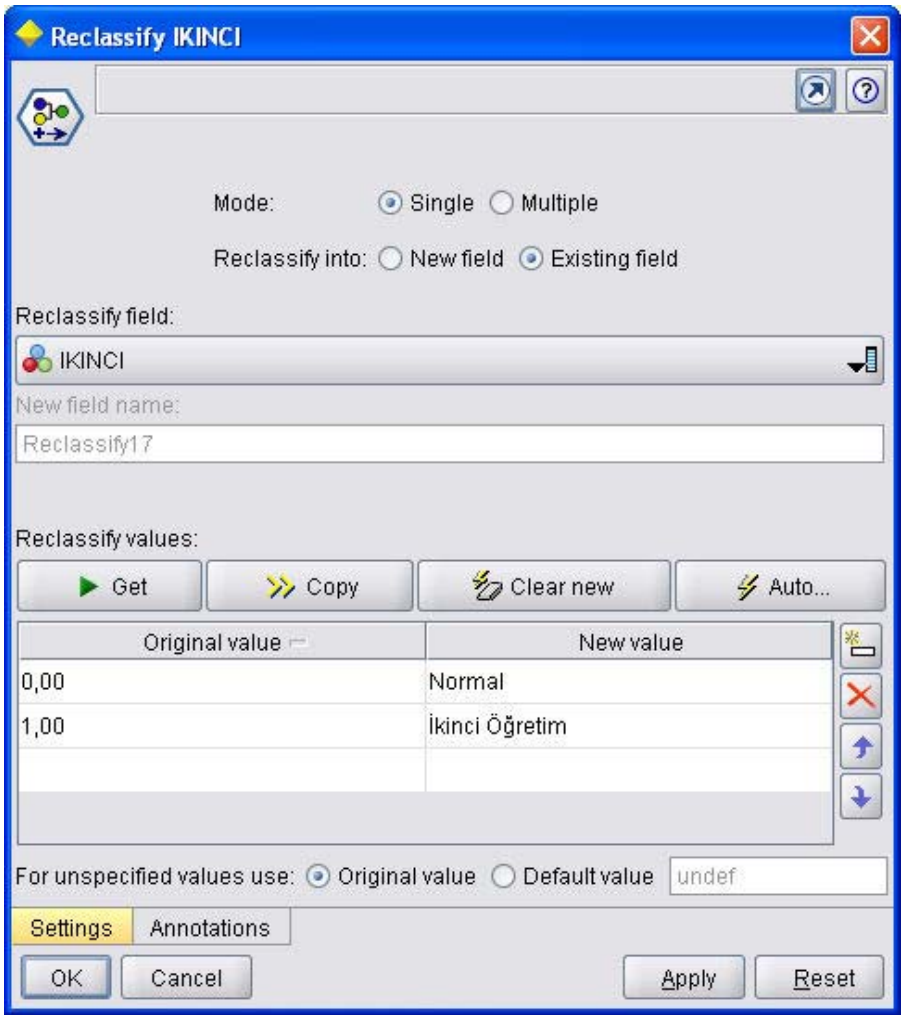
Şekil 3.11: Reclassify işlemcisi ile UNI_TURU alanının manipulasyonu



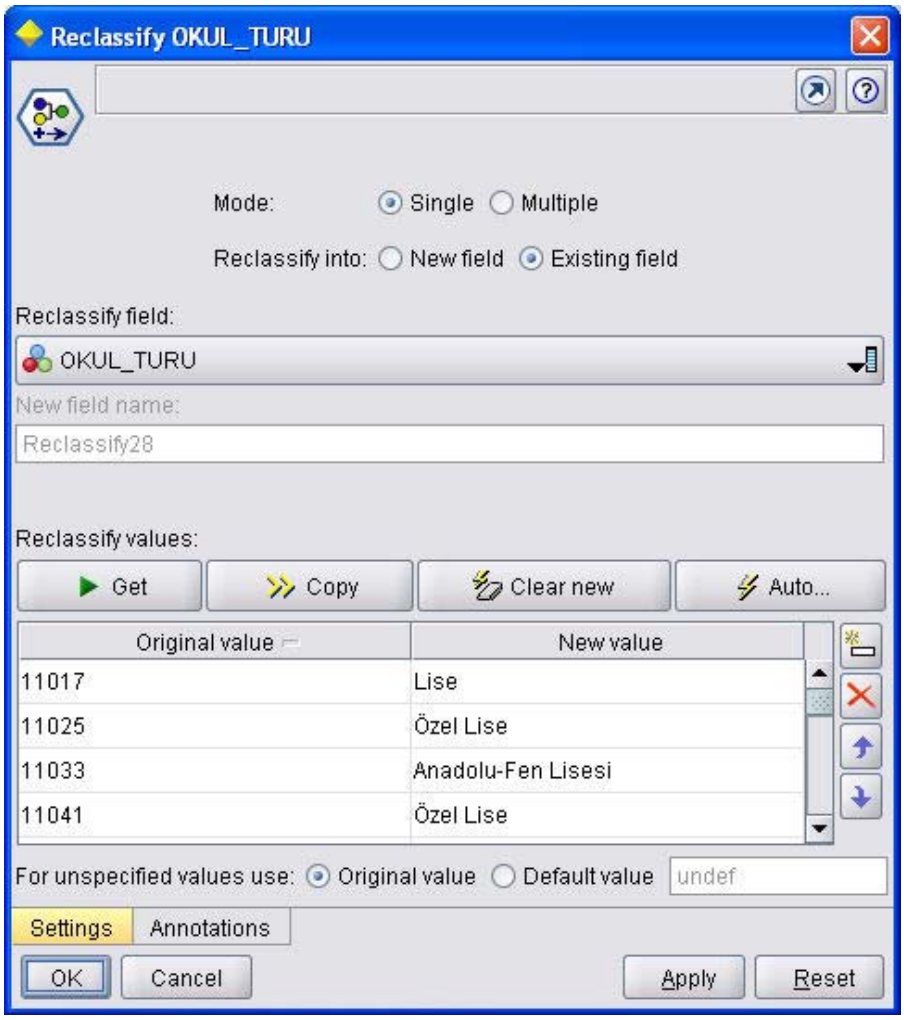
Şekil 3.12: Reclassify işlemcisi ile PUAN_TURU alanının manipulasyonu



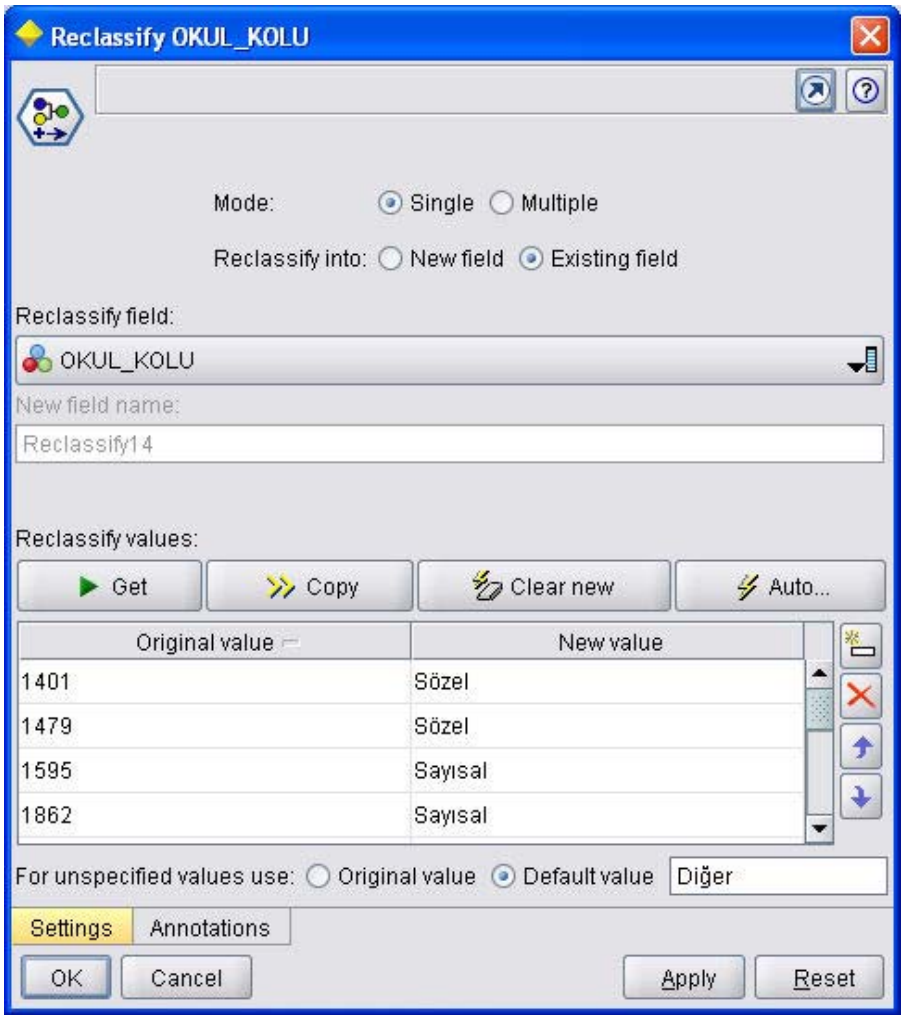
Şekil 3.13: Reclassify işlemcisi ile BURS alanının manipulasyonu



Şekil 3.14: Reclassify işlemcisi ile IKINCI alanının manipulasyonu



Şekil 3.15: Reclassify işlemcisi ile OKUL_TURU alanının manipulasyonu

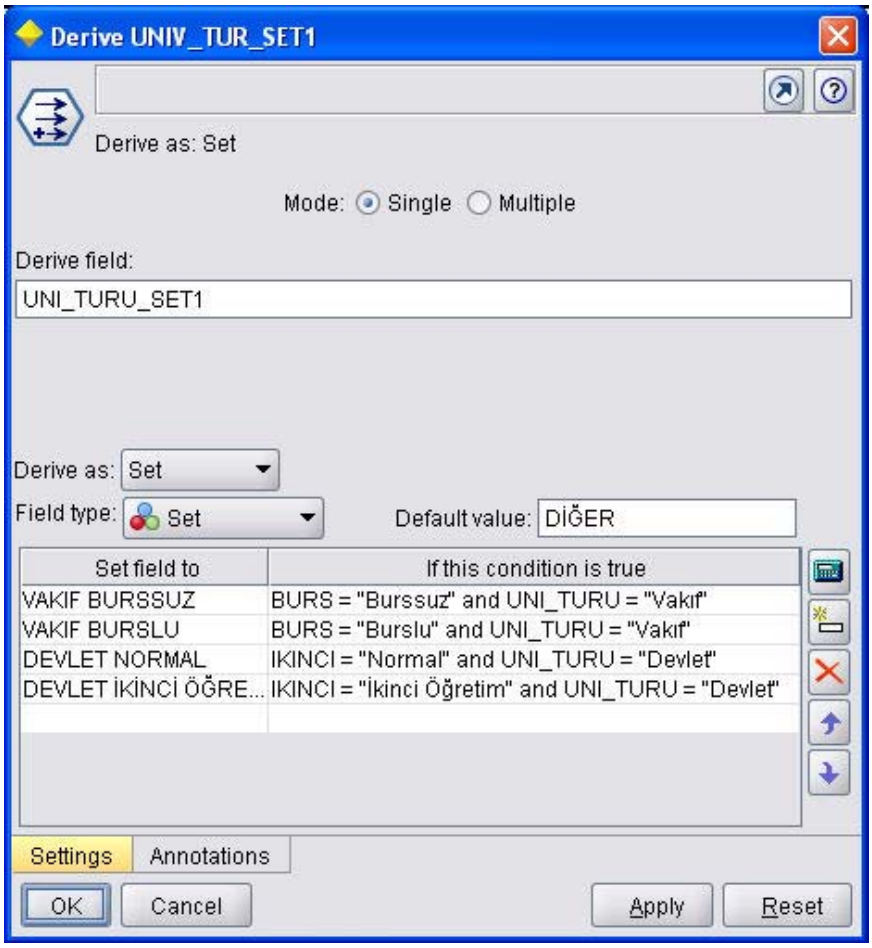


Şekil 3.16: Reclassify işlemcisi ile OKUL_KOLU alanının manipulasyonu

Son olarak, UNIV_TUR_SET1 isimli, hedef alanımızı oluşturacak, adayların tercih ettikleri üniversite bölüm türünü belirleyen alan Derive işlemcisi ile oluşturulmuştur.

Bu alanın alabileceği değerler aşağıdaki gibidir:

- DEVLET NORMAL
- DEVLET İKİNCİ ÖĞRETİM
- VAKIF BURSLU
- VAKIF BURSSUZ
- DİĞER



Şekil 3.17: Derive işlemcisi ile UNIV_TUR_SET1 alanının oluşturulması

Bu işlemin sonucunda, gerekli tüm manipulasyon işlemleri tamamlanmış olup, bir sonraki adım olan Modelin Kurulması aşamasına geçilmiştir.

3.3 Modelin Kurulması ve Değerlendirilmesi

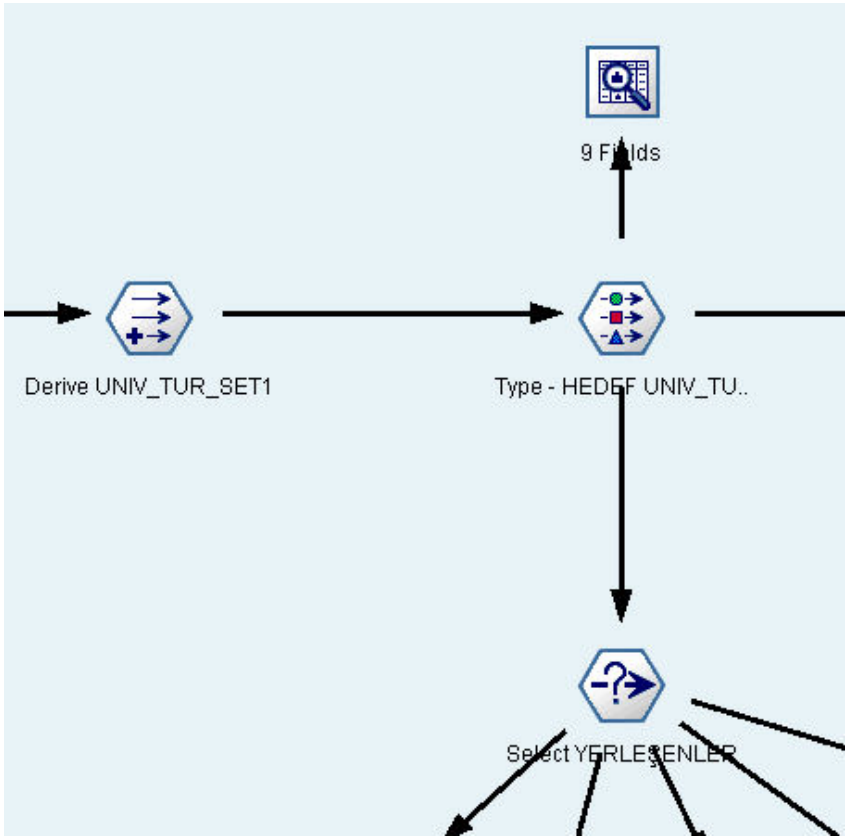
Modelin kurulması için C5.0 Karar Ağacı ve Lojistik Regresyon modelleme yöntemleri kullanılmıştır.

Model, 3 puan türü (EA2, SAY2, SÖZ2) için ayrı ayrı kurulmuştur.

3.3.1 EA2 Puan Türü

EA2 puan türü ile yerleştirilen adayların tercihleri ile ilgili modellemenin yapılması için, öncelikle akışın başında Select nodu ile EA2 puan türünden 185.000 puanın üstünde puan alan adaylar seçilmiştir.

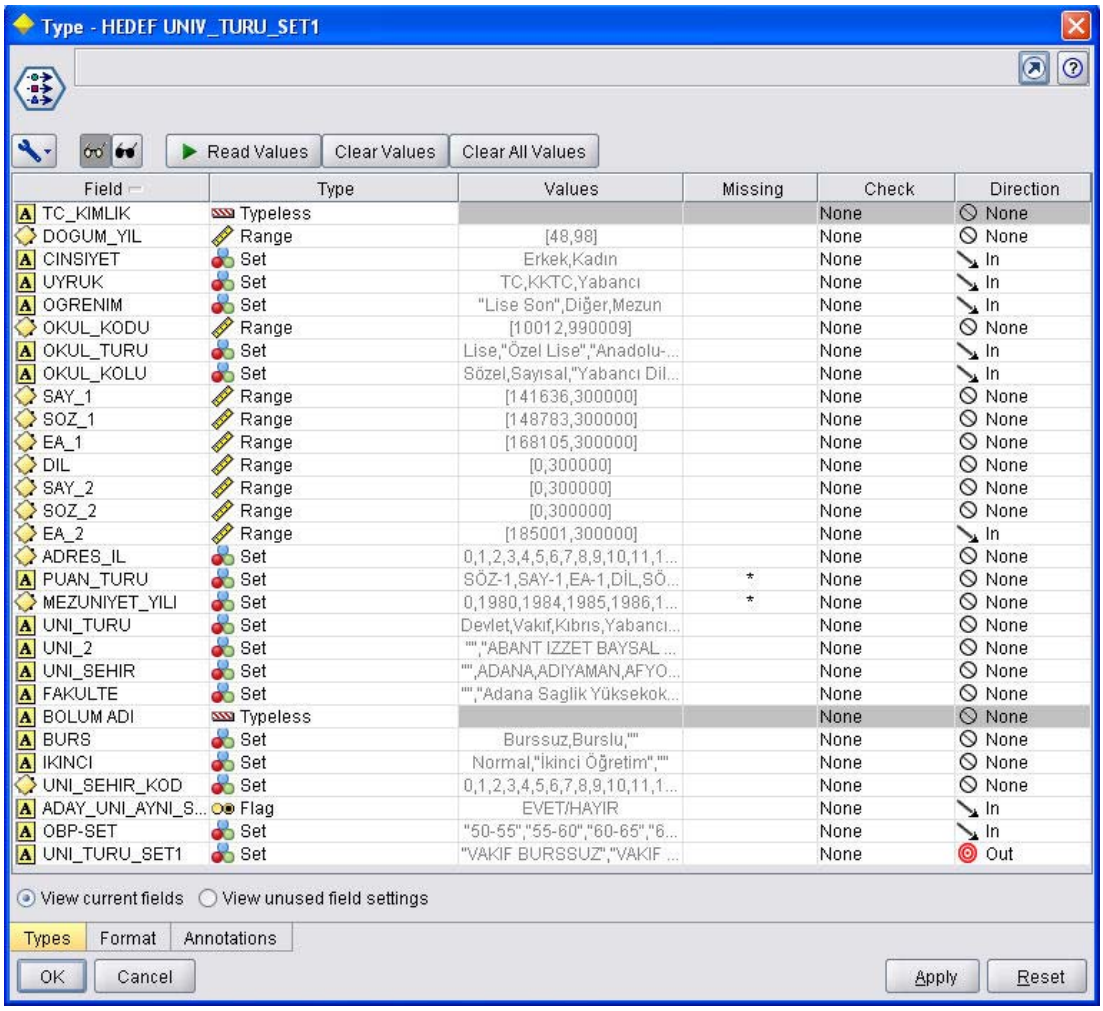
Akışın veri manipulasyon işlemlerinin bittiği noktadan sonra Type işlemcisi kullanılarak gerekli girdi alanları ve hedef alan olan UNIV_TUR_SET1 alanı belirlenmiştir.



Şekil 3.18: Girdi ve hedef alanlarının belirlendiği Type işlemcisinin akış alanındaki görünümü

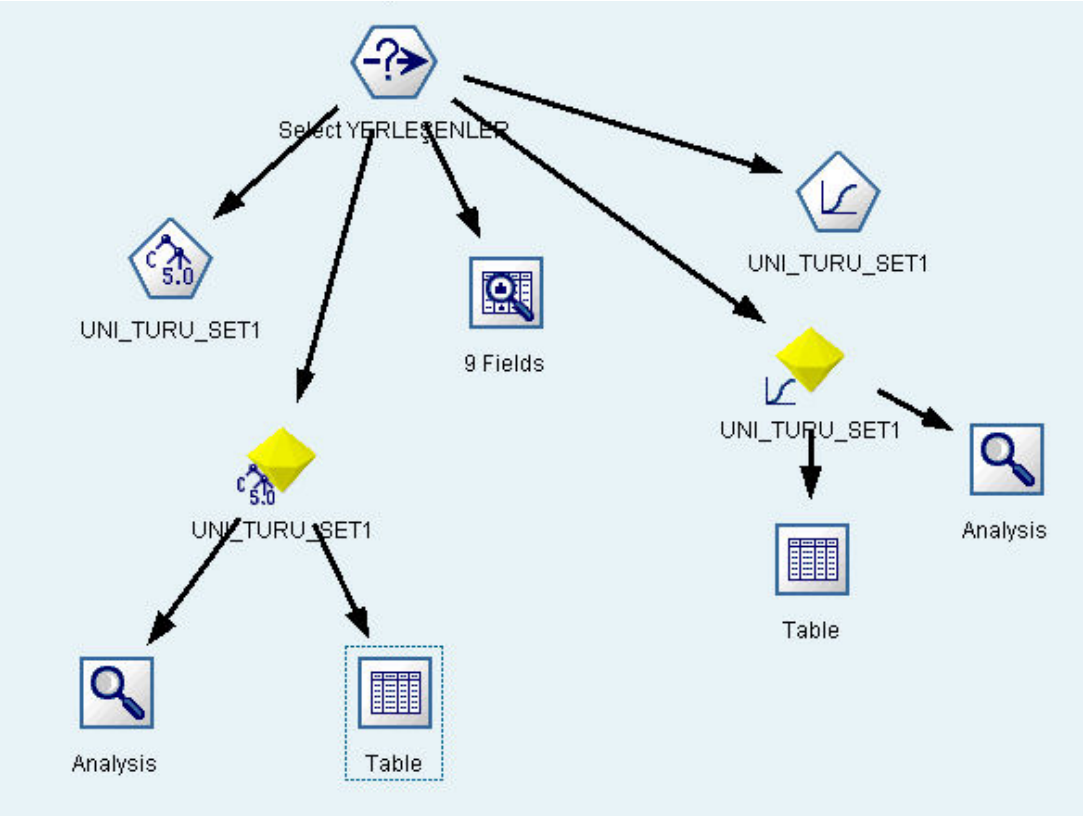
Bu Type işlemcisi içerisinde belirlenen girdi değişken alanları aşağıdaki gibidir:

- Puan
- OBP
- Uyruk
- Öğrenim Durumu
- Okul Türü
- Okul Kolu
- Adayın geldiği il ile yerleştiği üniversitenin bulunduğu ilin eşitliği
- Cinsiyet



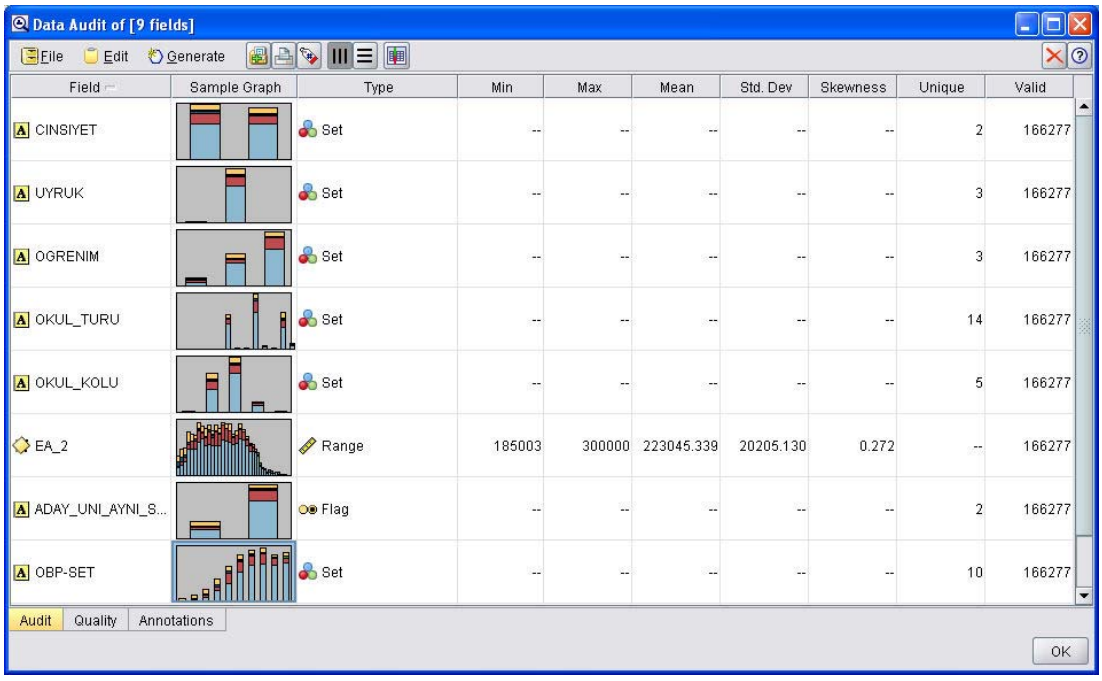
Şekil 3.19: Girdi ve hedef alanlarının belirlendiği Type işlemcisinin görünümü

Type işlemcisinin ardından, Select işlemcisi kullanılarak, EA-2 puan türünde herhangi bir bölüme yerleşen tüm adayların seçimi işlemi yapılmıştır.



Şekil 3.20: EA-2 puan türünde yerleşenler adayların modellenme akışı

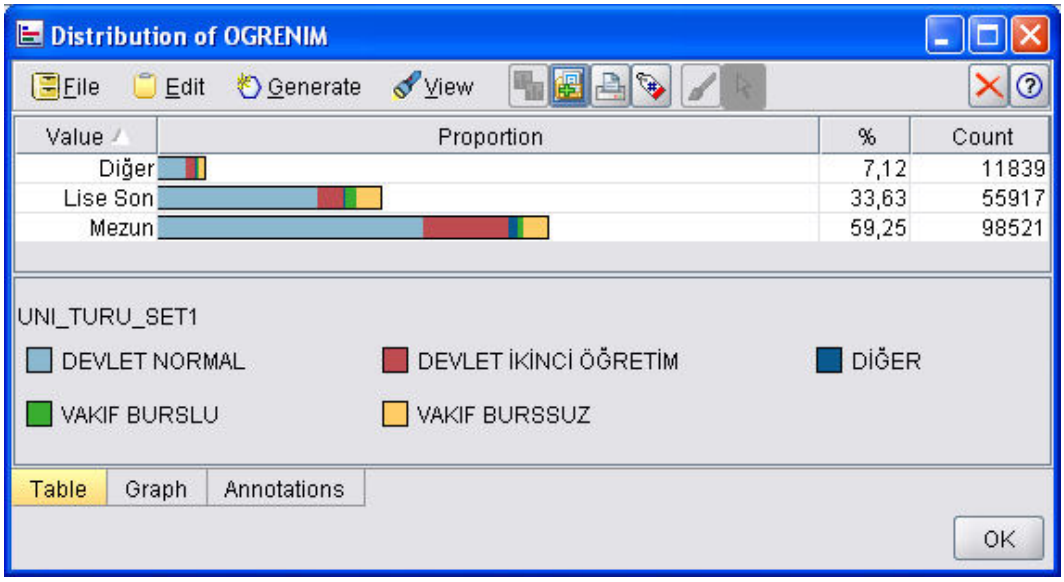
Data Audit nodu kullanılarak, EA-2 puan türünde herhangi bir bölüme yerleşen adayların hedef alan doğrultusundaki dağılımları incelenmiştir.



Field	Sample Graph	Type	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
CINSIYET		Set	--	--	--	--	--	2	166277
UYRUK		Set	--	--	--	--	--	3	166277
OGRENIM		Set	--	--	--	--	--	3	166277
OKUL_TURU		Set	--	--	--	--	--	14	166277
OKUL_KOLU		Set	--	--	--	--	--	5	166277
EA_2		Range	185003	300000	223045.339	20205.130	0.272	--	166277
ADAY_UNI_AYNI_S...		Flag	--	--	--	--	--	2	166277
OBP-SET		Set	--	--	--	--	--	10	166277

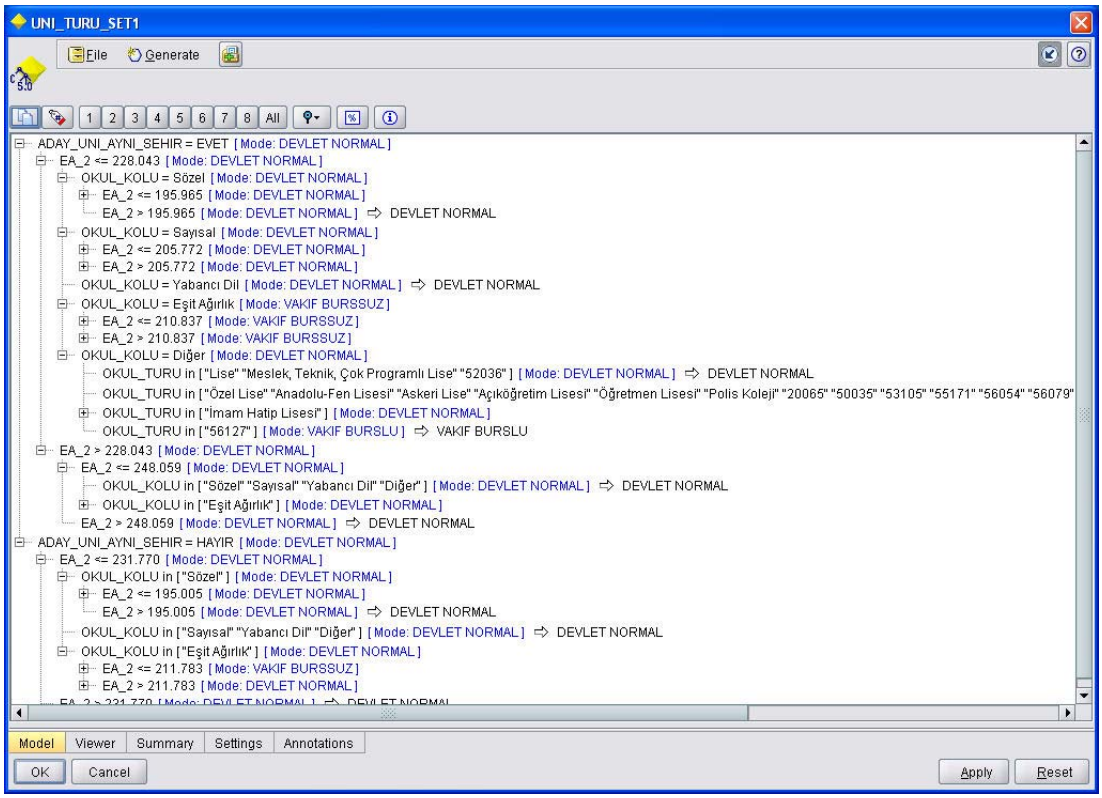
Şekil 3.21: Data Audit işlemcisi ile EA-2 puan türünde yerleşenlerin dağılımlarının görüntüsü

Bu inceleme esnasında, herhangi bir alan için hedef doğrultusundaki dağılımın daha ayrıntılı incelenmesi mümkündür.

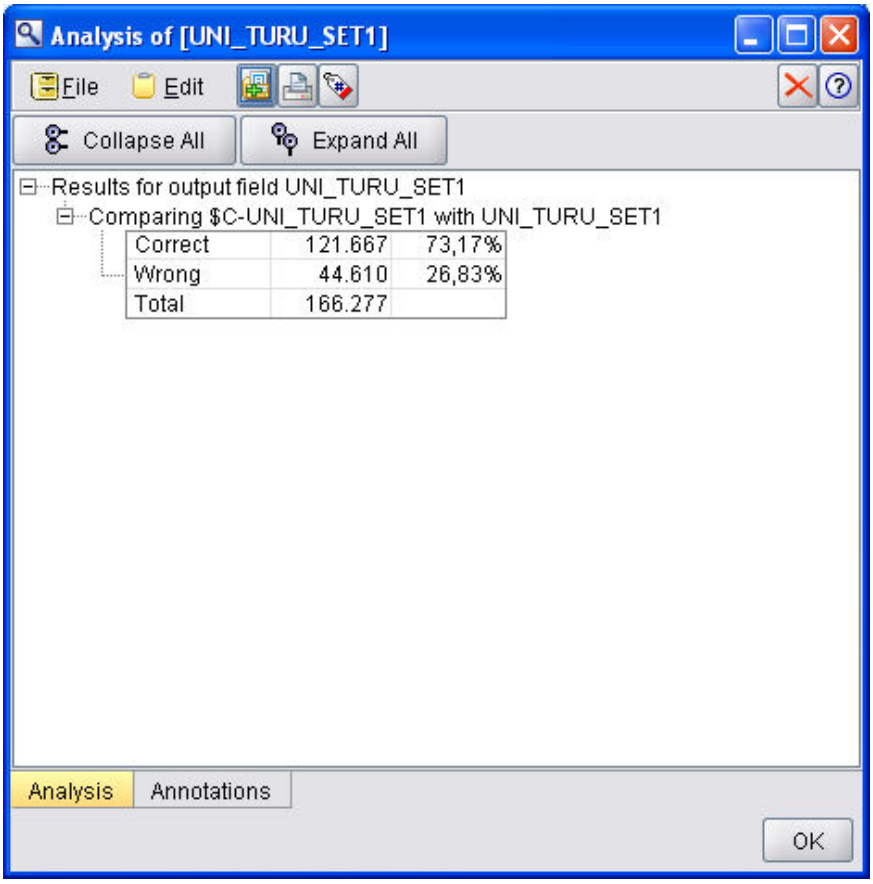


Şekil 3.22: EA-2 puan türü için OGRENIM alanının hedef alan doğrutusundaki dağılımı

Select işlemcisinden gelen veriler C 5.0 karar ağacı modelleme algoritması ile eğitilmiş, ardından bu modelin başarısı yine aynı veri kümesi için değerlendirilmiştir.

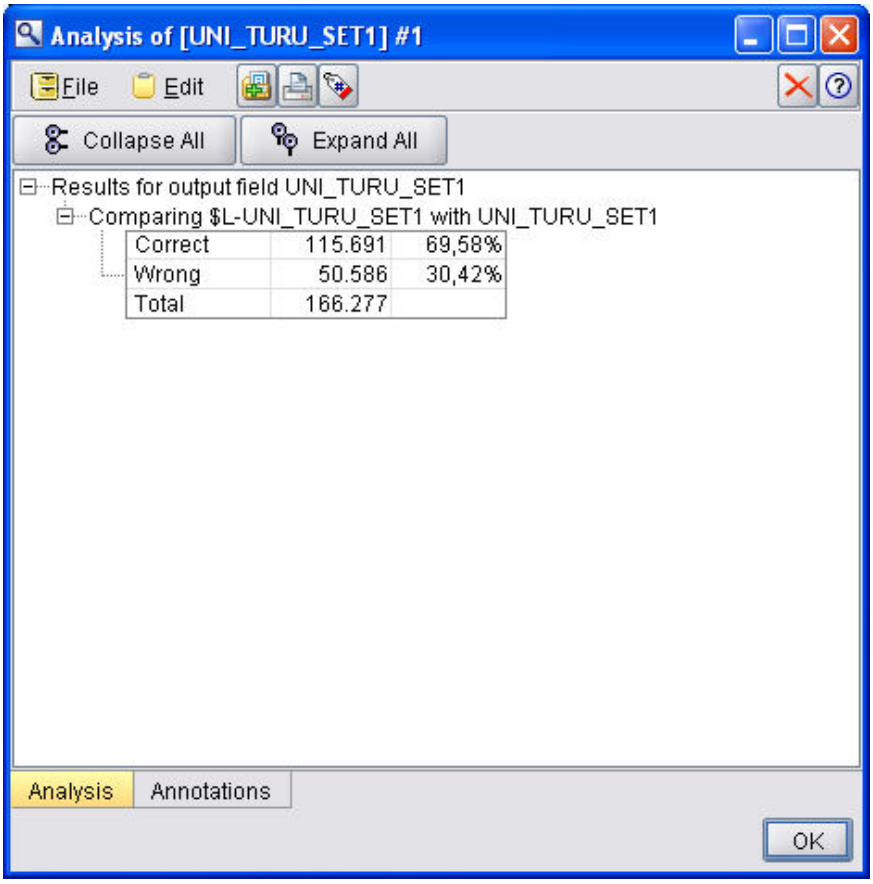


Şekil 3.23: EA-2 Puan türü için C 5.0 Karar Ağacı model yapısının görüntüsü



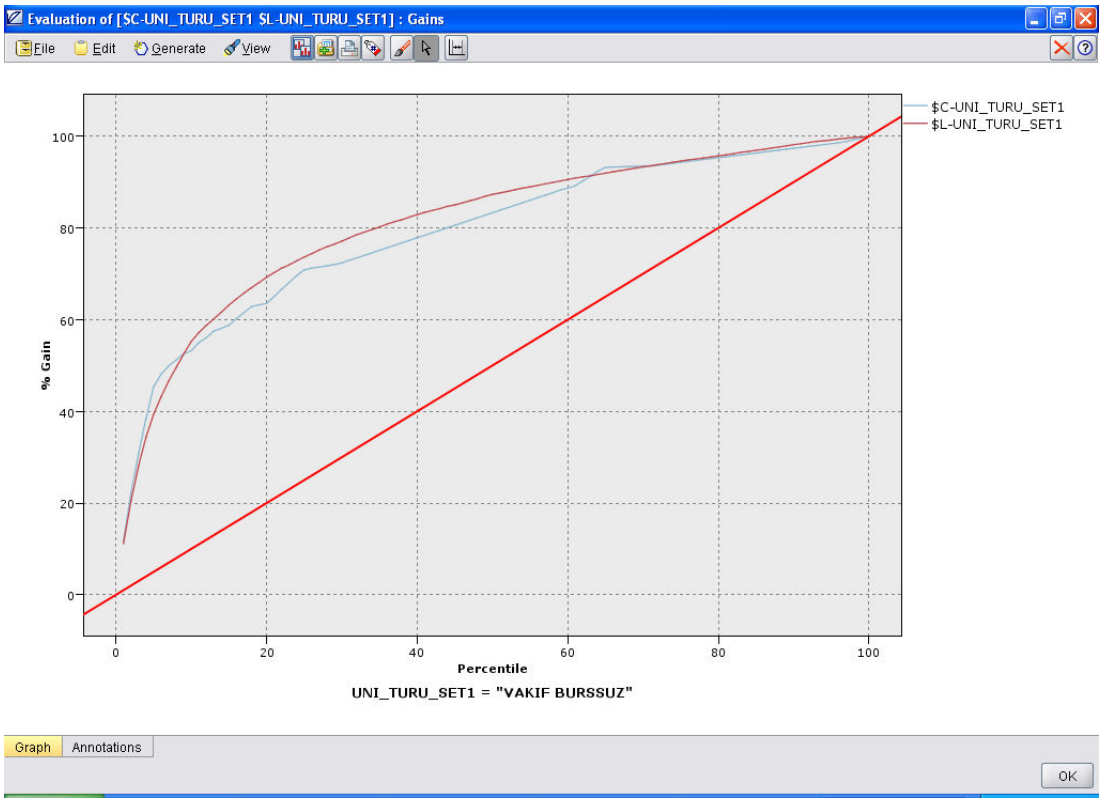
Şekil 3.24: EA-2 Puan türü için Analysis işlemcisi ile C 5.0 modelinin başarısının görüntülenmesi

Aynı şekilde, Select işlemcisinden gelen veriler bu sefer lojistik regresyon modelleme algoritması ile eğitilmiş, ardından yine bu modelin başarısı aynı veri kümesi için değerlendirilmiştir.



Şekil 3.25: EA-2 Puan türü için Analysis işlemcisi ile Lojistik Regresyon modelinin başarısının görüntülenmesi

Son olarak bu iki modelin tahmin başarısı Evaluation grafiği çizdirilerek değerlendirilmiştir.

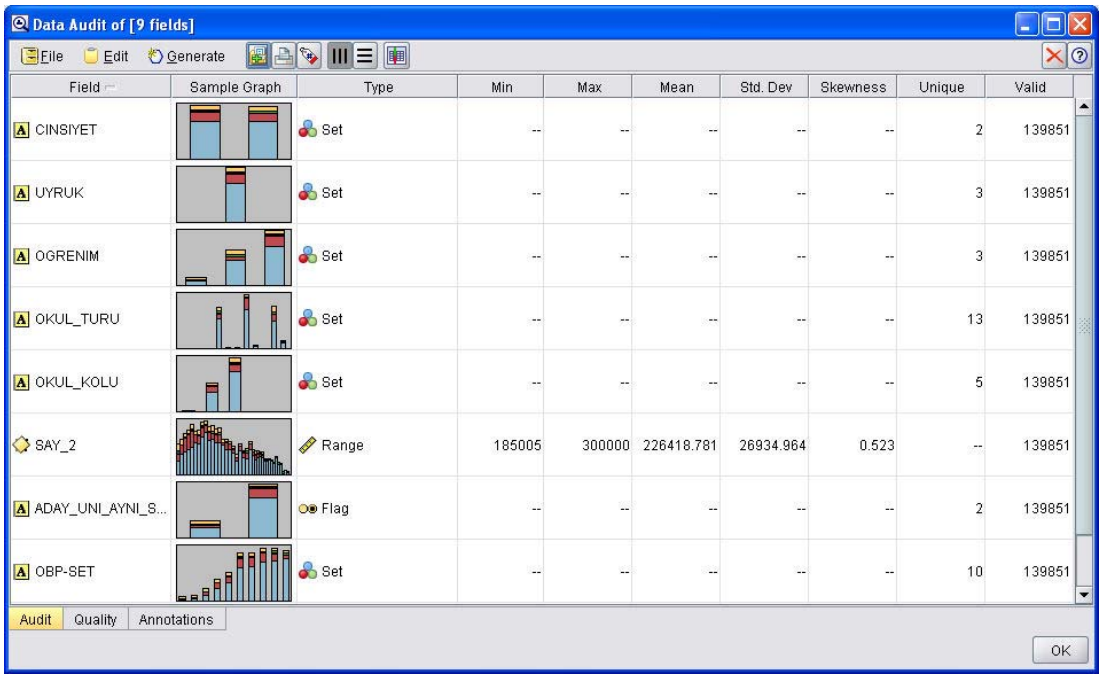


Şekil 3.26: EA-2 Puan türü için her iki modelin Evaluation grafiğinde görüntülenen başarısı

3.3.2 SAY-2 Puan Türü

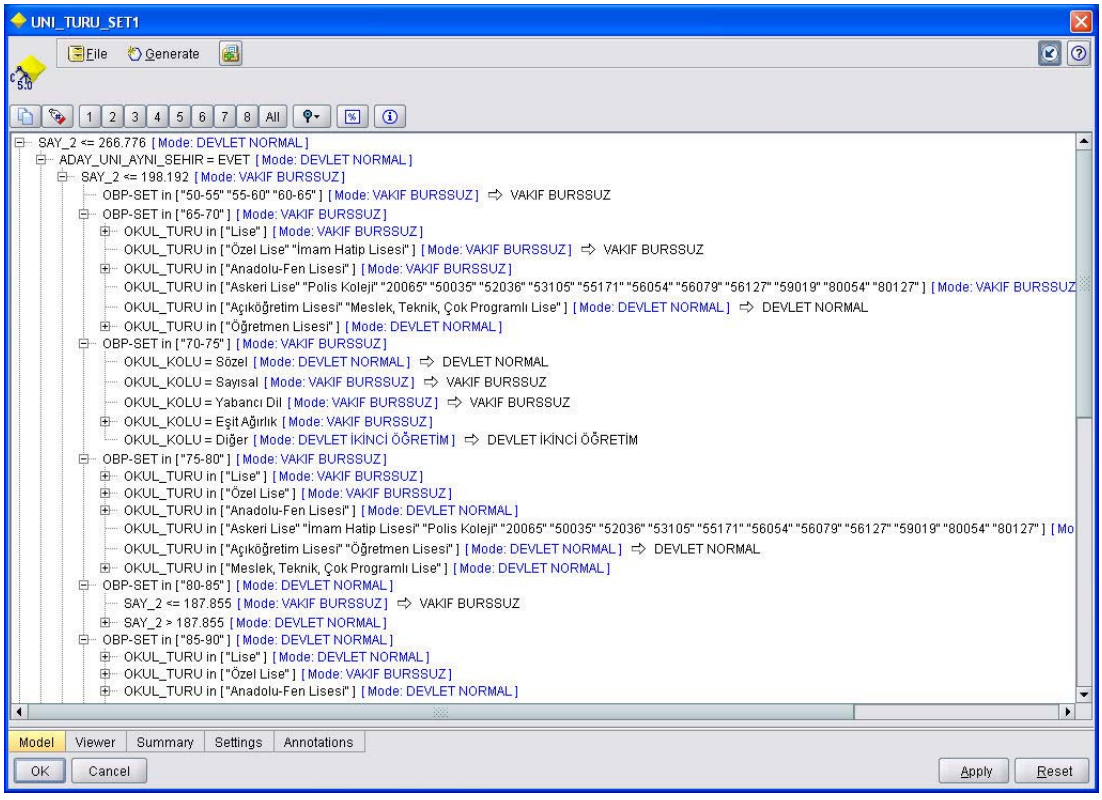
SAY-2 puan türü ile yerleştirilen adayların tercihleri ile ilgili modellemenin yapılması için, öncelikle akışın başında Select nodu ile SAY-2 puan türünden 185.000 puanın üstünde puan alan adaylar seçilmiştir.

Data Audit nodu kullanılarak, SAY-2 puan türünde herhangi bir bölüme yerleşen adayların hedef alan doğrultusundaki dağılımları incelenmiştir.

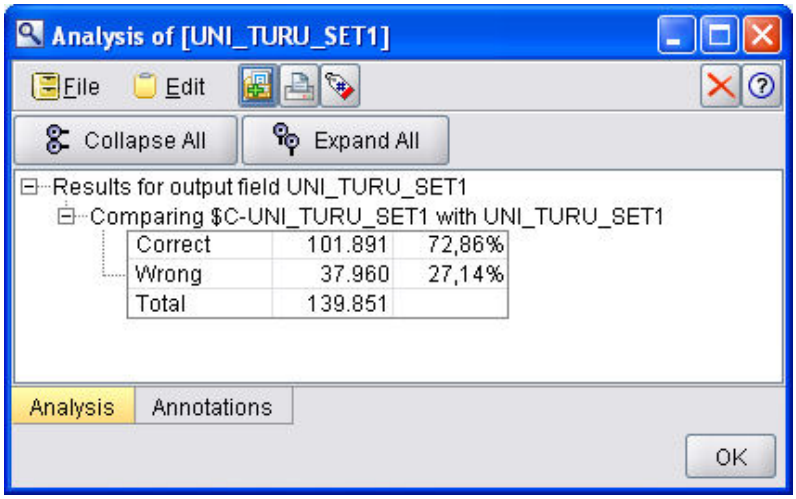


Şekil 3.27: Data Audit işlemcisi ile SAY-2 puan türünde yerleşenlerin dağılımlarının görüntüsü

Select işlemcisinden gelen veriler C 5.0 karar ağacı modelleme algoritması ile eğitilmiş, ardından bu modelin başarısı yine aynı veri kümesi için değerlendirilmiştir.

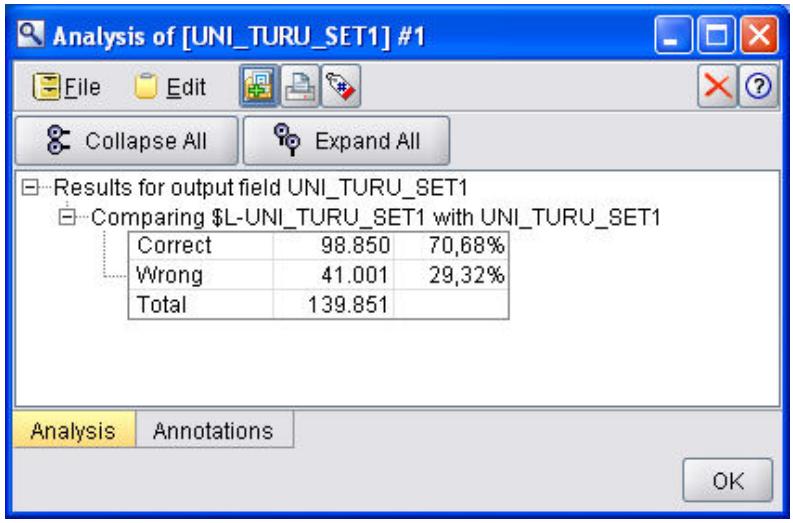


Şekil 3.28: SAY-2 Puan türü için C 5.0 Karar Ağacı model yapısının görüntüsü



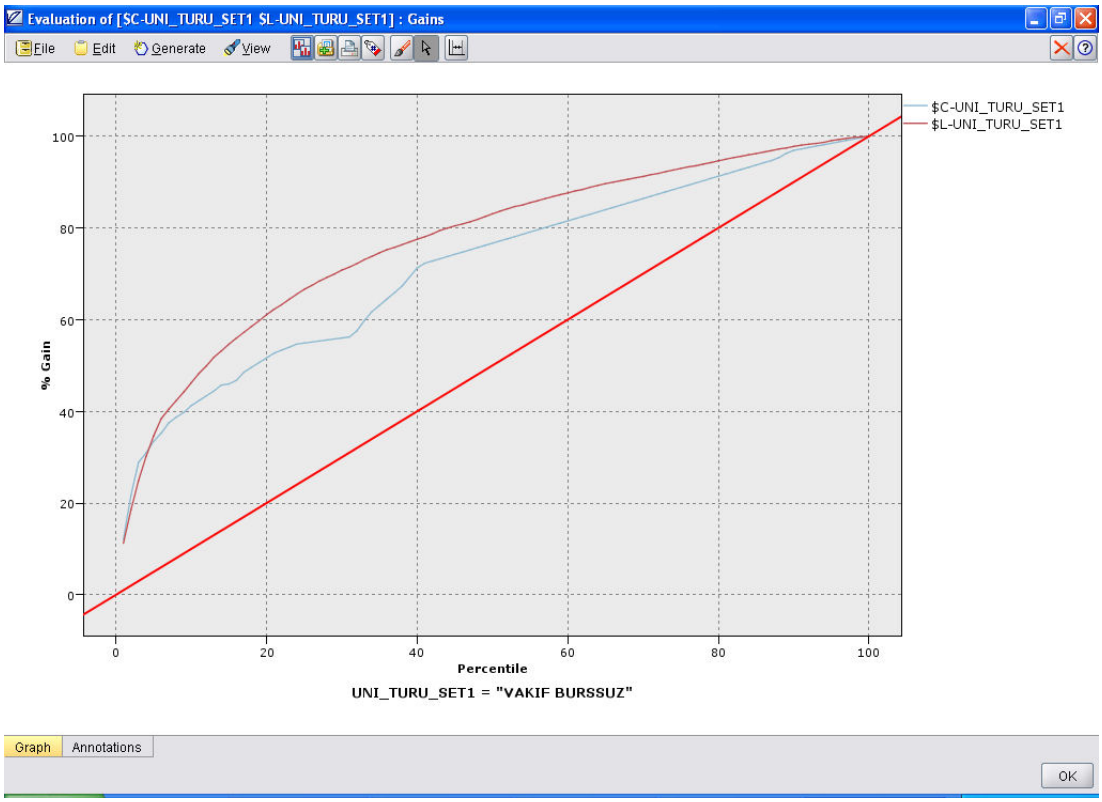
Şekil 3.29: SAY-2 Puan türü için Analysis işlemcisi ile C 5.0 modelinin başarısının görüntülenmesi

Aynı şekilde, Select işlemcisinden gelen veriler bu sefer lojistik regresyon modelleme algoritması ile eğitilmiş, ardından yine bu modelin başarısı aynı veri kümesi için değerlendirilmiştir.



Şekil 3.30: SAY-2 Puan türü için Analysis işlemcisi ile Lojistik Regresyon modelinin başarısının görüntülenmesi

Son olarak bu iki modelin tahmin başarısı Evaluation grafiği çizdirilerek değerlendirilmiştir.

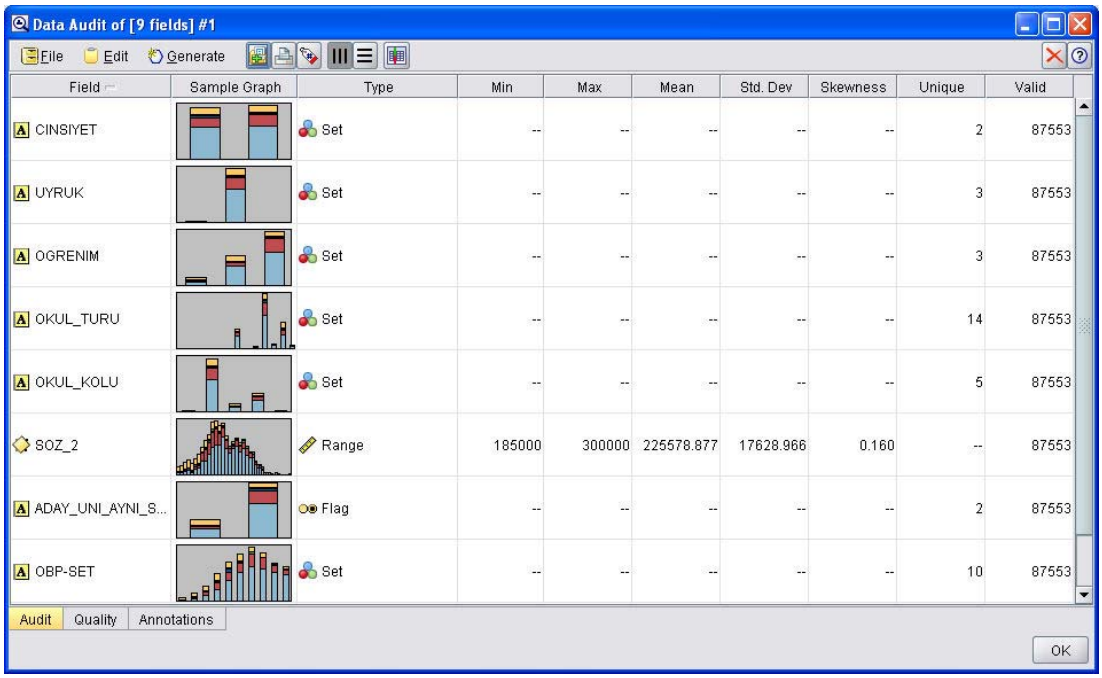


Şekil 3.31: SAY-2 Puan türü için her iki modelin Evaluation grafiğinde görüntülenen başarısı

3.3.3 SÖZ-2 Puan Türü

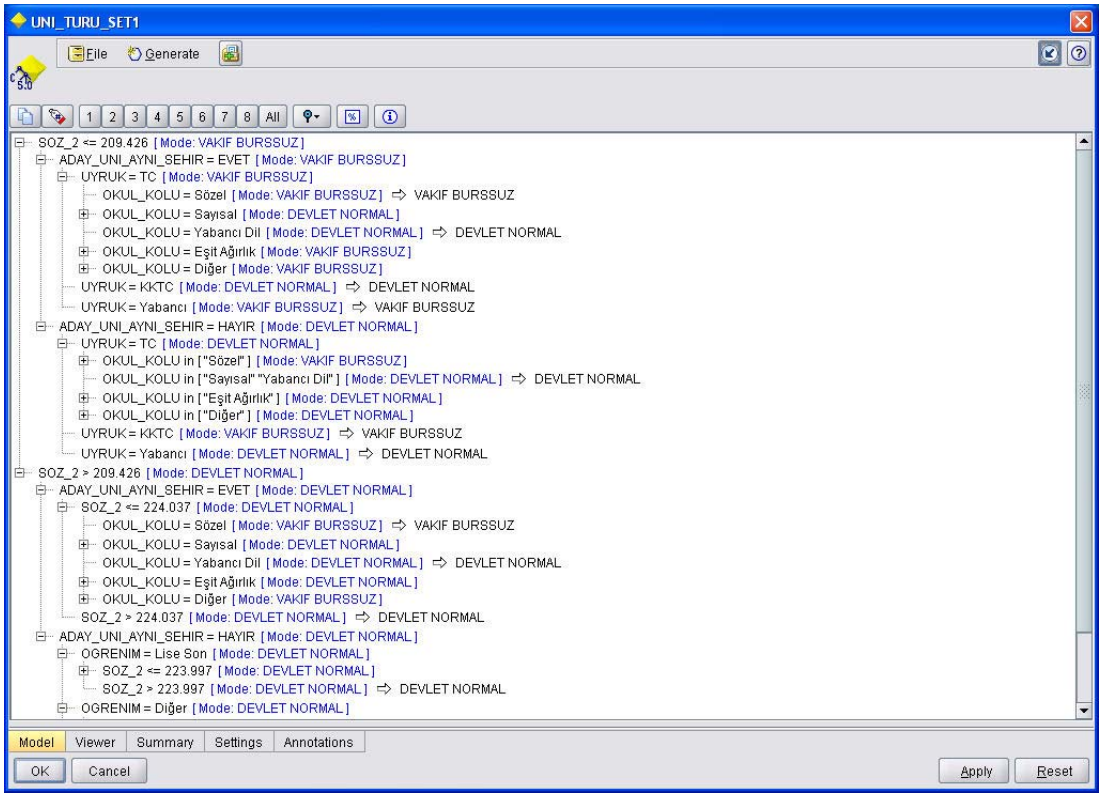
SÖZ-2 puan türü ile yerleştirilen adayların tercihleri ile ilgili modellemenin yapılması için, öncelikle akışın başında Select nodu ile SÖZ-2 puan türünden 185.000 puanın üstünde puan alan adaylar seçilmiştir.

Data Audit nodu kullanılarak, SÖZ-2 puan türünde herhangi bir bölüme yerleşen adayların hedef alan doğrultusundaki dağılımları incelenmiştir.

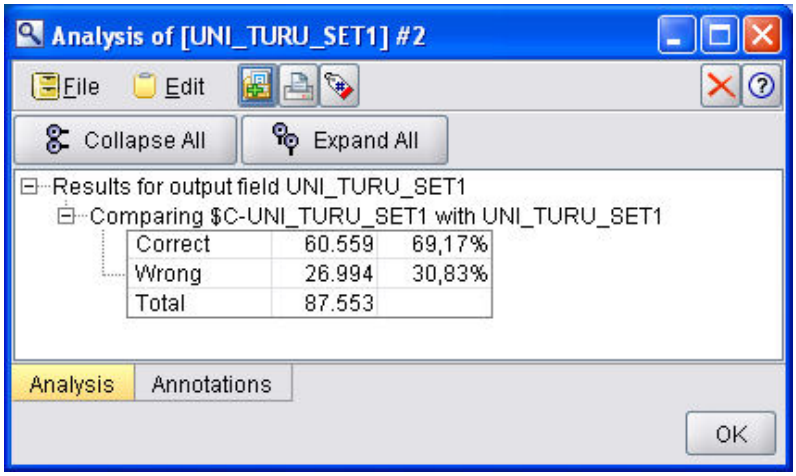


Şekil 3.32: Data Audit işlemcisi ile SÖZ-2 puan türünde yerleşenlerin dağılımlarının görüntüsü

Select işlemcisinden gelen veriler C 5.0 karar ağacı modelleme algoritması ile eğitilmiş, ardından bu modelin başarısı yine aynı veri kümesi için değerlendirilmiştir.

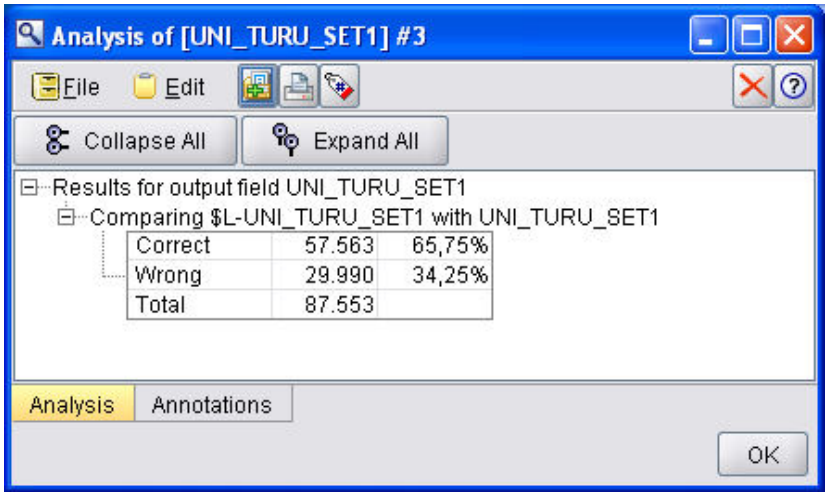


Şekil 3.33: SÖZ-2 Puan türü için C 5.0 Karar Ağacı model yapısının görüntüsü



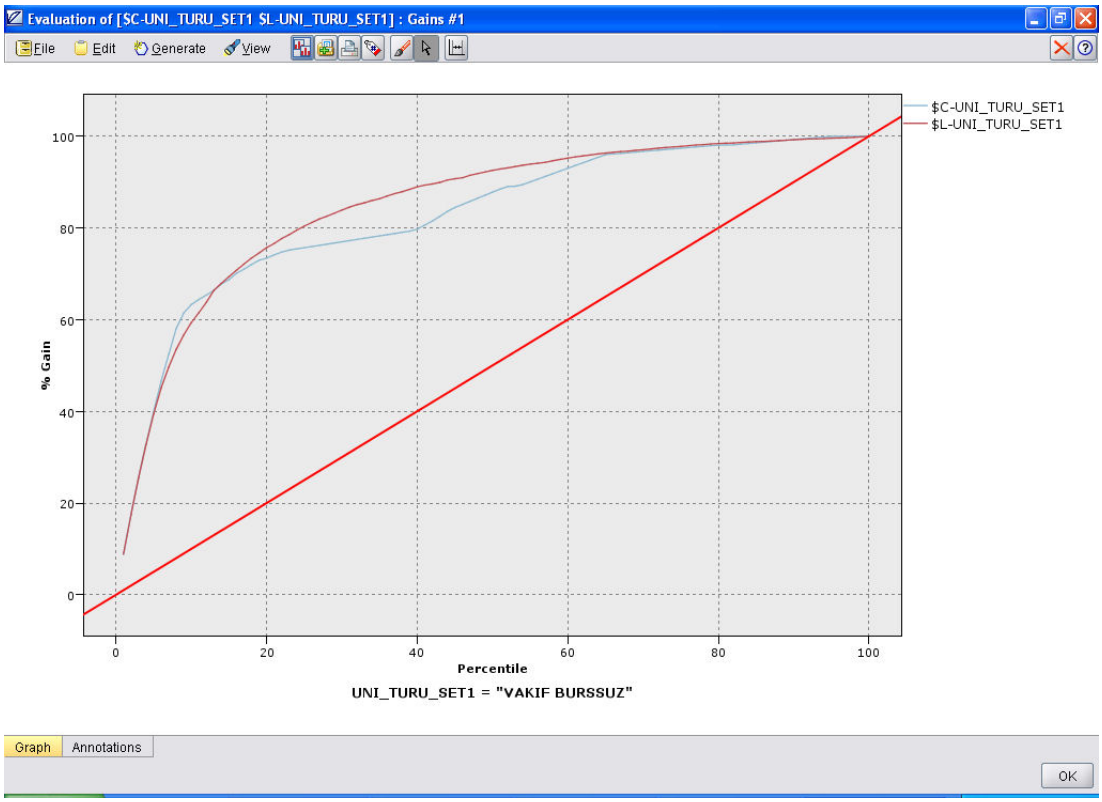
Şekil 3.34: SÖZ-2 Puan türü için Analysis işlemcisi ile C 5.0 modelinin başarısının görüntülenmesi

Aynı şekilde, Select işlemcisinden gelen veriler bu sefer lojistik regresyon modelleme algoritması ile eğitilmiş, ardından yine bu modelin başarısı aynı veri kümesi için değerlendirilmiştir.



Şekil 3.35: SÖZ-2 Puan türü için Analysis işlemcisi ile Lojistik Regresyon modelinin başarısının görüntülenmesi

Son olarak bu iki modelin tahmin başarısı Evaluation grafiği çizdirilerek değerlendirilmiştir.

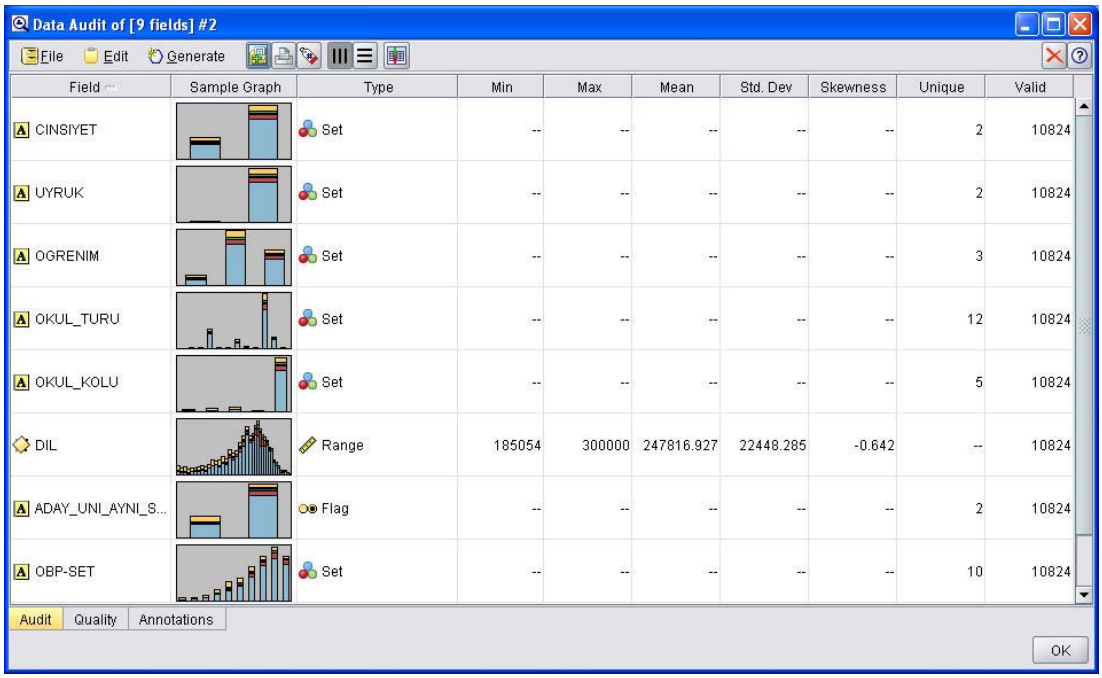


Şekil 3.36: SÖZ-2 Puan türü için her iki modelin Evaluation grafiğinde görüntülenen başarısı

3.3.4 DİL Puan Türü

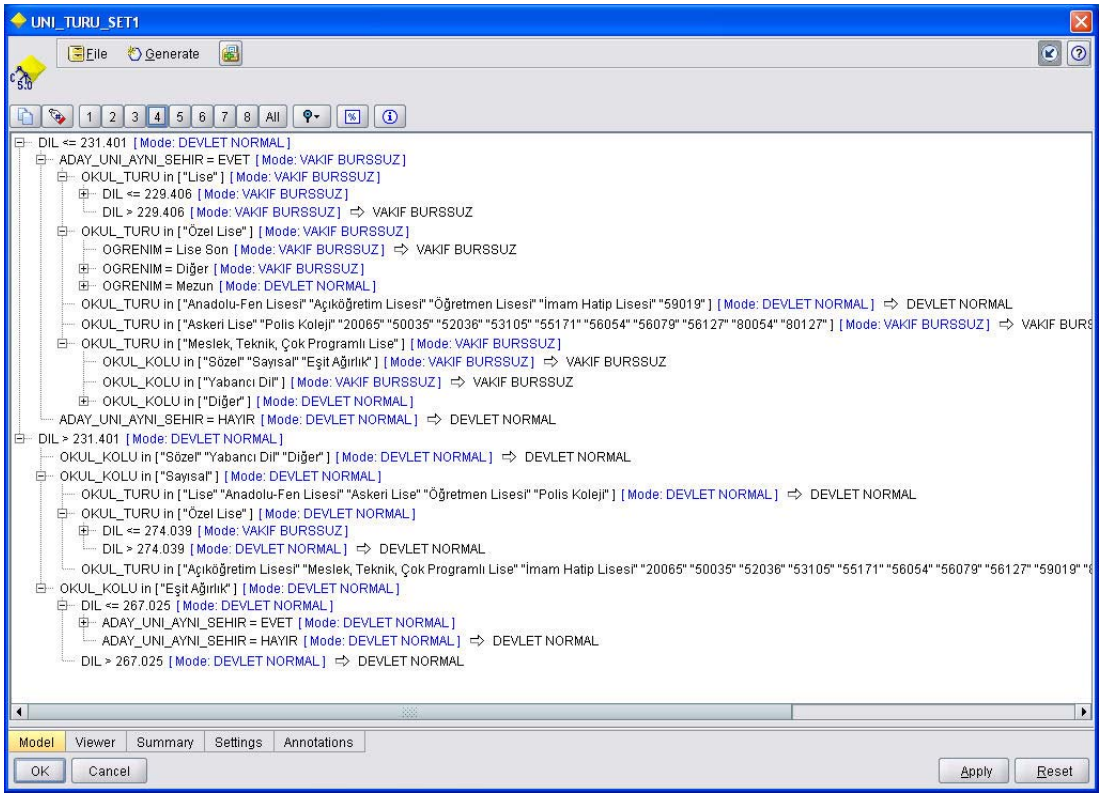
DİL puan türü ile yerleştirilen adayların tercihleri ile ilgili modellemenin yapılması için, öncelikle akışın başında Select nodu ile DİL puan türünden 185.000 puanın üstünde puan alan adaylar seçilmiştir.

Data Audit nodu kullanılarak, DİL puan türünde herhangi bir bölüme yerleşen adayların hedef alan doğrultusundaki dağılımları incelenmiştir.

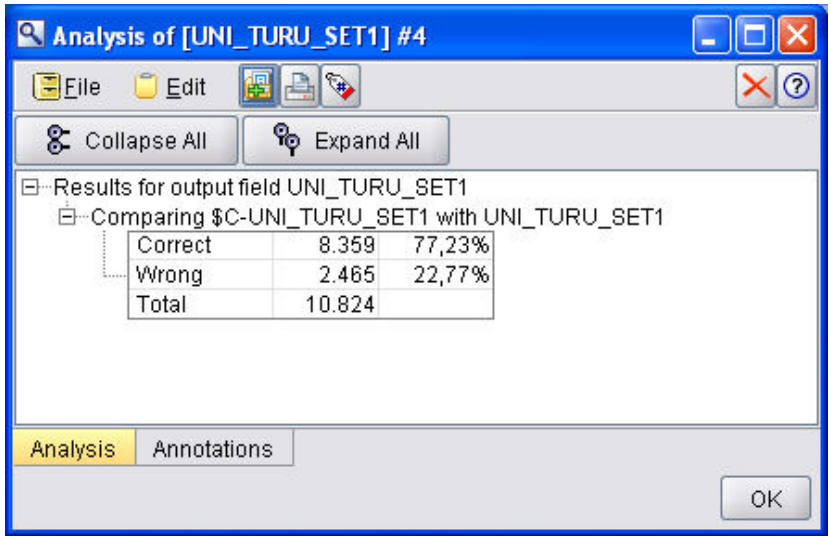


Şekil 3.37: Data Audit işlemcisi ile DİL puan türünde yerleşenlerin dağılımlarının görüntüsü

Select işlemcisinden gelen veriler C 5.0 karar ağacı modelleme algoritması ile eğitilmiş, ardından bu modelin başarısı yine aynı veri kümesi için değerlendirilmiştir.

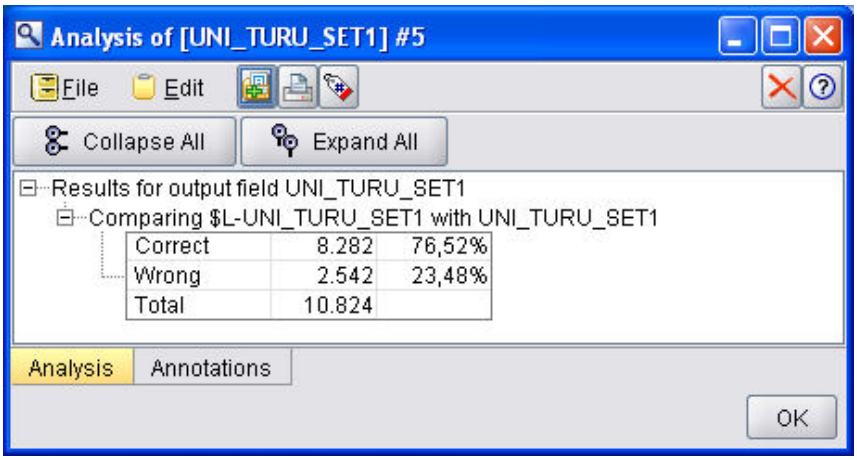


Şekil 3.38: DİL Puan türü için C 5.0 Karar Ağacı model yapısının görüntüsü



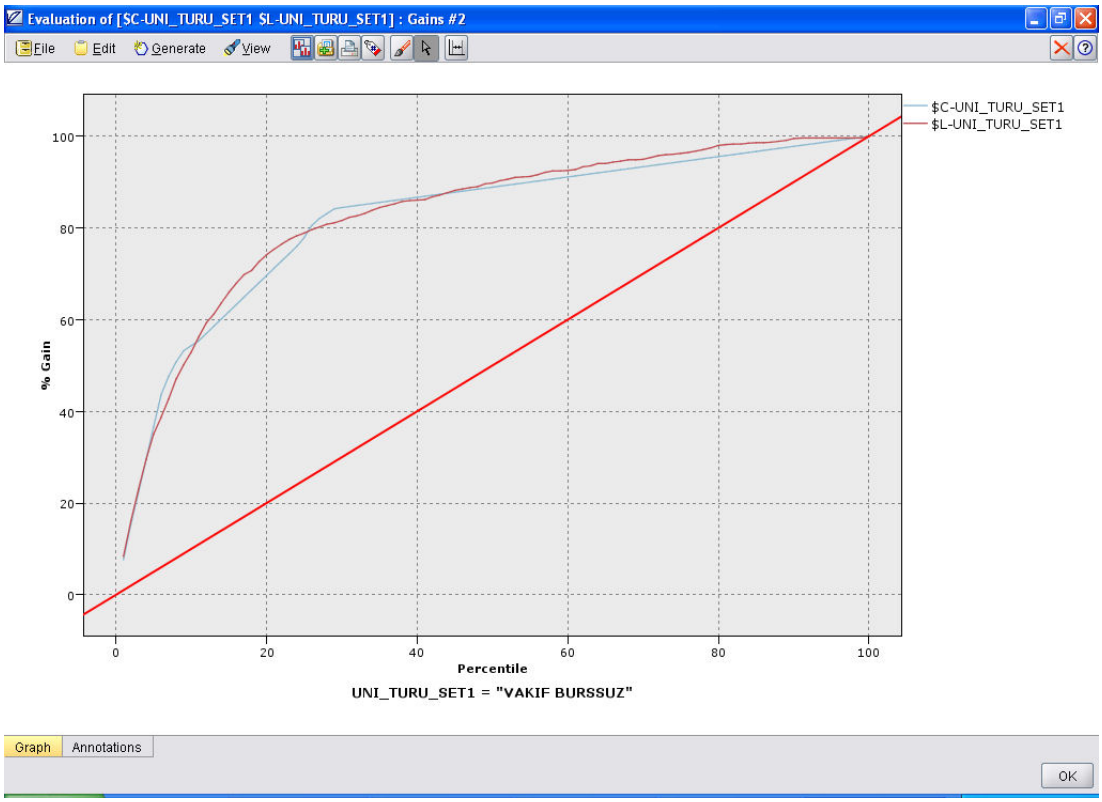
Şekil 3.39: DİL Puan türü için Analysis işlemcisi ile C 5.0 modelinin başarısının görüntülenmesi

Aynı şekilde, Select işlemcisinden gelen veriler bu sefer lojistik regresyon modelleme algoritması ile eğitilmiş, ardından yine bu modelin başarısı aynı veri kümesi için değerlendirilmiştir.



Şekil 3.40: DİL Puan türü için Analysis işlemcisi ile Lojistik Regresyon modelinin başarısının görüntülenmesi

Son olarak bu iki modelin tahmin başarısı Evaluation grafiği çizdirilerek değerlendirilmiştir.

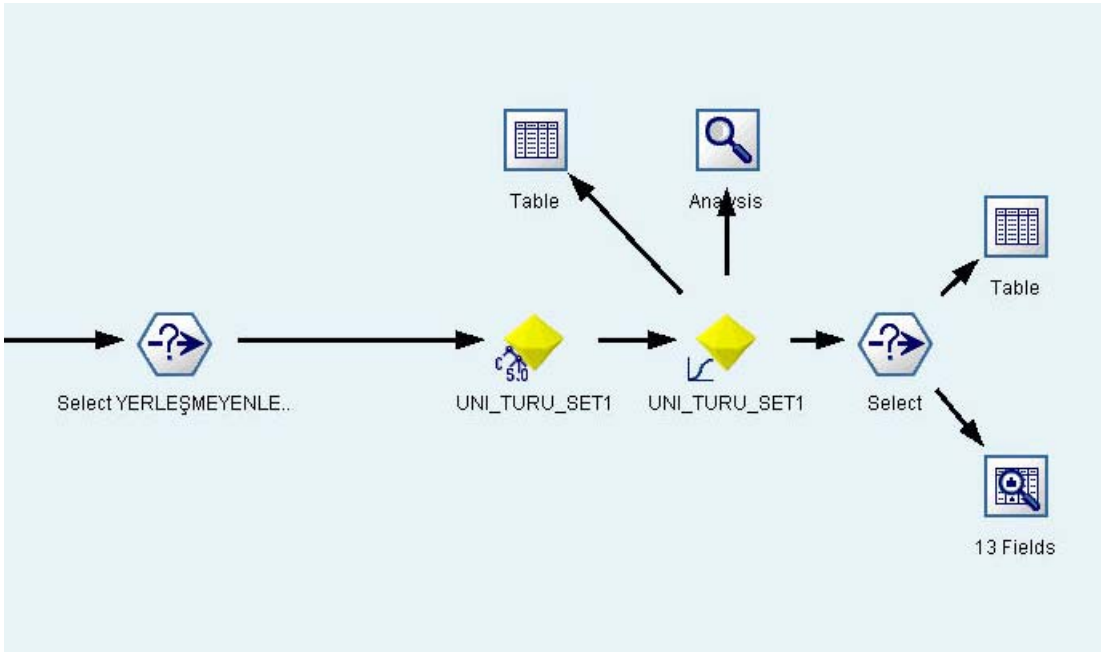


Şekil 3.41: DİL Puan türü için her iki modelin Evaluation grafiğinde görüntülenen başarısı

3.4 Modelin Kullanılması

Yerleşmeyen adayların oluşturduğu veri kümesine bu model uygulanarak 2008 yılında bu adayların ÖSS'ye girmesi durumunda hangi yönde tercih yapacakları tahmin edilmiştir.

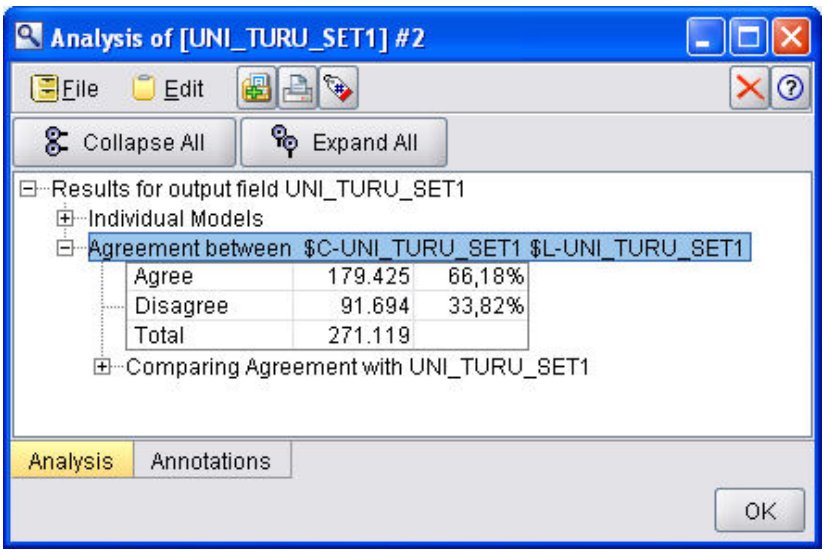
Modelin yerleşmeyen adaylar üzerinde kullanılması için belirlenen akış, hedef alanın belirlendiği Type işlemcisinden sonra devam etmektedir.



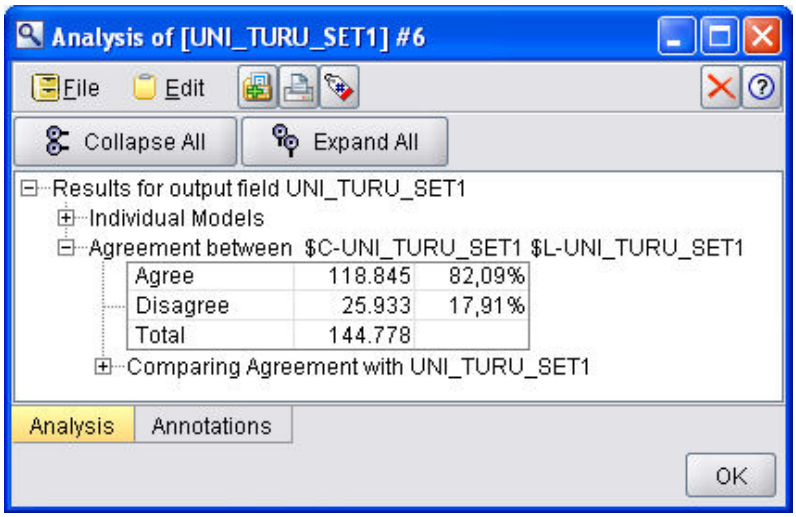
Şekil 3.42: Yerleşmeyen adaylar üzerinde modelin kullanılması

İki model, arka arkaya yerleşmeyen adayların bulunduğu veri kümesine uygulanmıştır. Böylece yerleşmeyen bu adayların, hangi üniversite türü tercihine yatkın olduklarına dair bir tahmin yürütülmüştür.

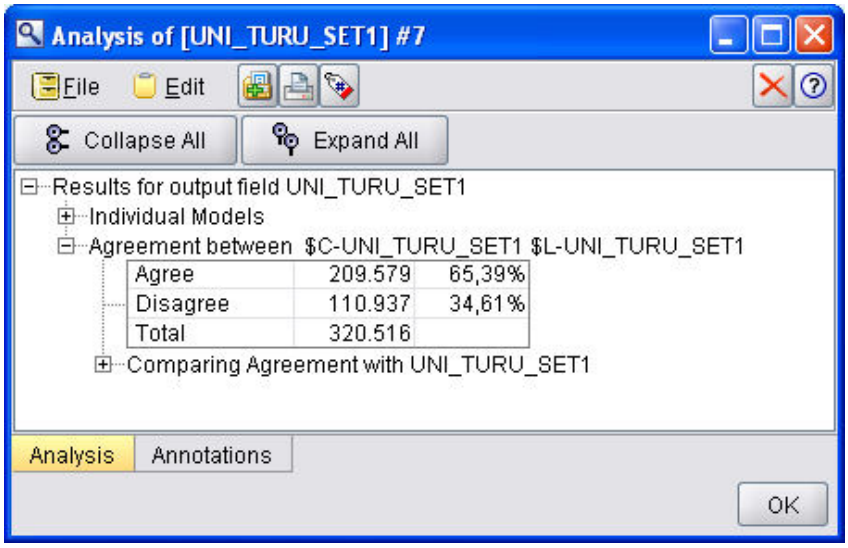
İki modelin de hedef alan değeri konusunda hemfikir oldukları kayıtların sayısı Analysis işlemcisi kullanılarak görüntülenmiştir.



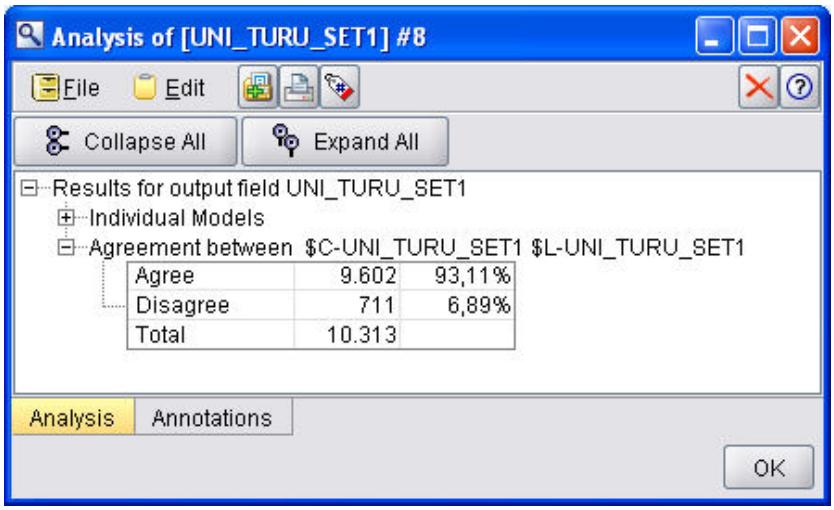
Şekil 3.43: Analysis işlemcisi ile her iki modelin hemfikir oldukları kayıtların oranının görüntüsü (EA-2 Puan Türü İçin)



Şekil 3.44: Analysis işlemcisi ile her iki modelin hemfikir oldukları kayıtların oranının görüntüsü (SAY-2 Puan Türü İçin)

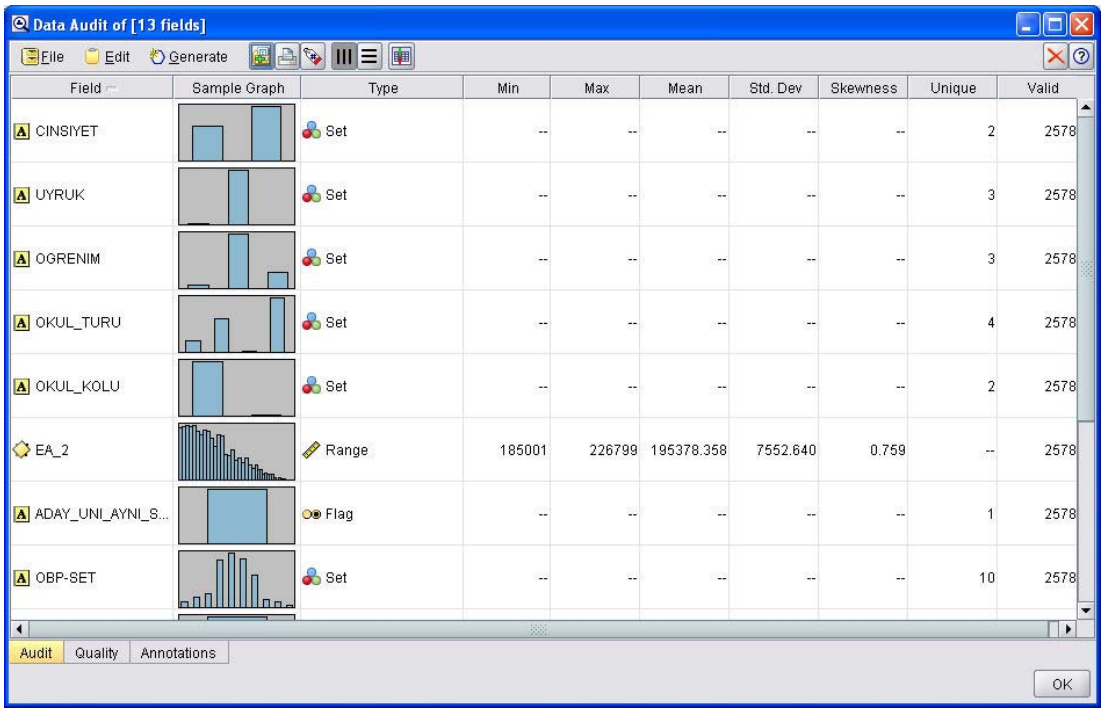


Şekil 3.45: Analysis işlemcisi ile her iki modelin hemfikir oldukları kayıtların oranının görüntüsü (SÖZ-2 Puan Türü İçin)



Şekil 3.46: Analysis işlemcisi ile her iki modelin hemfikir oldukları kayıtların oranının görüntüsü (DİL Puan Türü İçin)

Modellerin hemen ardından, bir Select işlemcisi kullanılarak, her iki modelin de üniversite türü tercihini "VAKIF BURSSUZ" olarak tahmin ettiği adaylar seçilmiştir. Bu adayların kayıtları Table işlemcisi ile listelenerek incelenmiş, Data Audit işlemcisi ile de hedef değişkene yönelik giriş alanlarının yatkınlıkları gözlemlenmiştir.



Şekil 3.47: EA-2 Puan türü için hedef alanında "VAKIF BURSSUZ" bulunması tahmin edilen kayıtların dağılımı

Bu yaklaşım, 2008 ÖSS verilerinin üniversitelerle paylaşılmasının ardından, tanıtım faaliyetlerinin planlanmasında doğrudan kullanılabilir.

4. SONUÇ

Bu çalışmanın sonucunda, üniversite adaylarının farklı üniversite türlerini tercih ederken yansıttıkları desenler, Veri Madenciliği modellerinin kullanılması ile tespit edilmiştir.

Her puan türü için, o puan türünde yerleşmeye hak kazanmış (İlgili puan türünde 185.000 üzeri puan almış) adaylar üzerinde modellerin uygulaması yapılarak, yerleşememiş adayların üniversite yatkınlıkları ile ilgili tahmin sonuçları elde edilmiştir.

Üniversite adaylarının, üniversite türü tercih davranışlarının modellenerek tahmin edilmesi, üniversite yönetimini ilgilendiren, hem adayların davranışlarını tanıma, hem de etkin ve verimli tanıtım faaliyetinin planlanması ve yürütülmesi anlamında önemli sonuçlar ortaya çıkarmıştır.

Bu çalışma ile, 2007 senesinde vakıf üniversitelerinin burslu ve burssuz bölümlerine yerleşen adayların veri madenciliği yöntemleri ile modellenmesi tamamlanmış olup, bu modeller 2008 ÖSS sınavının yapılmasının ardından ÖSYM tarafından üniversiteler ile paylaşılacak verilere uygulanmaya hazırdır.

Bu çalışmanın devamı olarak, belirlenen modellerin açıklanacak bu verilere uygulanması ile, adayların üniversite tercih döneminde, daha iyi hedef seçimleri ile adaylara ulaşılması mümkün olacaktır.

Ayrıca, bir başka geliştirme olarak, bu çalışmada elde edilen modellerin her sene oluşacak yeni veriler ile kendini geliştirmebilmesi sözkonusudur.

Sınav sistemi değişikliklerinin, adayların üniversite tercihlerini nasıl etkileyebileceği ile ilgili model geliştirme çalışması da, bu çalışmanın devamında yapılabilir.

ÖSYM tarafından adayların yerleştirilmesinin ardından açıklanan tercih listeleri verileri kullanılarak, adayların üniversitelere yerleşme desenlerinin yanı sıra, üniversiteleri tercih etme sıraları girdileri kullanılarak tercih desenleri, bu çalışmanın devamında geliştirilebilir.

KAYNAKLAR

[1] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.

[2] Morzy, Mikolaj, "Advanced database structures for effective association rule mining", phD, Pozna´n University of Technology, 2004

[3] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.

[4] Jiawei Han, et al DBMiner: "DBMiner: A System for Data Mining in Relational Databases and Data Warehouses", Proc. CASCON\'97: Meeting of Minds, Toronto, Canada, November 1997.

[5] W. J. Frawley, G. P-Shapiro, C.J. Matheus: Knowledge Discovery in Databases: An Overview. In: G. P-Shapiro, W. J. Frawley (Ed.): Knowledge Discovery in Databases, 1991.

[6] Akpınar Haldun, "Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği", İstanbul Üniversitesi, İşletme Fakültesi: www.isletme.istanbul.edu.tr/akpinar (Erişim, 10 Şubat 2008)

[7] SPSS Türkiye Clementine Eğitim Ders Notları

Internet Kaynakları:
http://www.bilgiyonetimi.org/cm/pages/mkl_gos.php?nt=538

ÖZGEÇMİŞ

1981 yılında Düzce`de doğdu. 2000 yılında öğrenimine başladığı İstanbul Kültür Üniversitesi Matematik-Bilgisayar Bölümünden 2004 yılında mezun oldu. Bu yıldan itibaren İstanbul Kültür Üniversitesi Matematik - Bilgisayar Bölümünde Araştırma Görevlisi olarak çalışmaktadır. Ayrıca 2007 yılında İstanbul Kültür Üniversitesi Araştırma – Geliştirme Merkezi'nin yürüttüğü bir proje olan Doğru Tercih Proje Ekibi içerisinde Bilişim Sorumlusu olarak görev almaktadır.