Giriş Düzeyinde Örnek Bir Veri Madenciliği Projesi-1

Gerçekleştirecek olduğumuz bu projede bir kitapevi firmasının müşterilerinin, Excelde tutulmuş olan verilerinden yola çıkarak; müşterilerini segmente etmek, gruplamak amaçlanmaktadır. Bu amaç doğrultusunda açık kaynak bir yazılım olan Weka programı ve Microsoft Office Excel'in Data Mining Add-in 'i kullanacağız.Projeyi daha kolay uygulayabilmeniz amacıyla 2 makale halinde yayınlamayı uygun gördüm. Bu makalede Weka yardımıyla projemizi gerçekleştireceğiz.
(Projede kullanılmış olan Veri Setlerinin Dosyalarına http://www.iszekam.net/download/Kitapevi.rar ulaşabilirsiniz.) (Dosyanın şifresi :www.iszekam.net dir)

Veri Setinin Tanıtımı
Veri setinin içeriğindeki veriler kategorik ve numerik ifadelerden oluşmaktaydı. Kategorik veriler için eksiklik olmamasına rağmen numerik verilerde olan eksik veriler, diğer bütün değerlerin ortalaması alınarak temizlenmiştir. Dönüştürme işlemleri olarak da numerik veriler belirli aralıktaki gruplara bölünürken, kategorik veriler için 1,2,3.. gibi değerler girilmiştir.

Kitapevi firmasının müşterilerine ait elimizdeki datasetinde gerekli temizleme ve dönüşüm işlemleri gerçekleştirildikten sonra 9 tane niteliğimiz ortaya çıkmıştır. Bunların isimleri ve özellikleri ise aşağıda detaylı olarak açıklanmıştır.

Yaş : Müşterilere ait doğum tarihi bilgisi 3 ana kategoride toplanmıştır. Bunlar; 24 ile 35 yaş arası , 36 ile 45 yaş arası , 45 yaş ve sonrası olarak gruplara ayrılmıştır.
Alışveriş Periyodu: Müşteriler, alışveriş sıklıklarına göre 3 ana kategoride değerlendirilmiştir.Bu kategoriler; 0-6 ay, 7-12 ay, 12 dan daha uzun olarak değerlendirilmiştir.
Churn Durumu : Müşterilerin halen mağazada kayıtlı müşteri olup olmadıklarına göre 2 gruba ayrılmıştırlar. Evet yada Hayır olarak.(Hayır demek halen müşterimiz anlamında)
Semt : Müşterilerin oturmuş oldukları yer (Anadolu Yakası, Avrupa Yakası, İstanbul dışı) sırasıyla 1,2,3 olarak kodlanmıştır.
Harcama Tutarı : Müşterilerin gerçekleştirilmiş oldukları toplam harcamalar 3 grupta toplanmıştır. (100-250, 251-400, 400 + olarak.)
Üyelik Tipi : Kitapevine üye olan müşterilerin üyelikleri Gold, Platin, Gümüş 'ü ifade etmek için 1, 2, 3 ile numaranlandırılmıştır.
Ücretsiz Kiyap : Kitapevinin müşterilerinin promosyon amaçlı ücretsiz kitap alıp,almadıkları Evet,Hayır,Bazen 'in baş harfleriyle kodlanmıştır.
Kitap Türü : Müşterilerin almış oldukları değişik kategorilerde ki kitaplar; Bilimsel için 1 , Bilgisayar için 2, Tıp için 3, Eczacılık için 4,Mühendislik için 5 ile numaralandırılmıştır.

Modelleme :
Elimizde bulunan 100 müşteriye ait yaş, alışveriş periyodu, churn durumu, ücretsiz kitap, üyelik tipi, kitap türü, semt, gibi verilerden yola çıkarak segmentasyon yapmak amaçlanmaktadır. Bu amaç için denetimsiz bir model olarak Kümeleme Analizlerini gerçekleştirip Müşteri Segmentasyonu yapılması planlanmaktadır.(Kümeleme analizinden özet bir şekilde aşağıda bahsetdim ama daha detaylı bilgi için işzekam.net 'e bakabilirsiniz.)

Kümeleme analizi, sınıflandırmada olduğu gibi sahip olunan verileri gruplara ayırma işlemidir. Kümeleme yabancı kaynaklarda clustering ya da segmentation olarak adlandırılmaktadır. Sınıflandırma işleminde, sınıflar önceden belirli iken kümelemede sınıflar önceden belirli değildir. Verilerin hangi gruplara/kümelere, hatta kaç değişik gruba ayrılacağı eldeki verilerin birbirlerine olan benzerliğine göre belirlenir. Belirlenen herbir gruba da küme ismi verilir. Kümeleme analizi biyoloji, tıp, antropoloji, pazarlama, ekonomi ve telekomikasyon gibi birçok ve birbirinden çok farklı alanlarda kullanılmaktadır. (1)
Kümeleme analizi ve algoritmaları 5 ana başlık altında incelenebilir. Bunlar:
ü Hiyerarşik Yöntemler (SLINK Algoritması, CURE Algoritması, CHAMELEON Algoritması, BIRC Algoritması)
ü Bölümlemeli Yöntemler (K-Means Algoritması, PAM Algoritması, CLARA Algoritması,

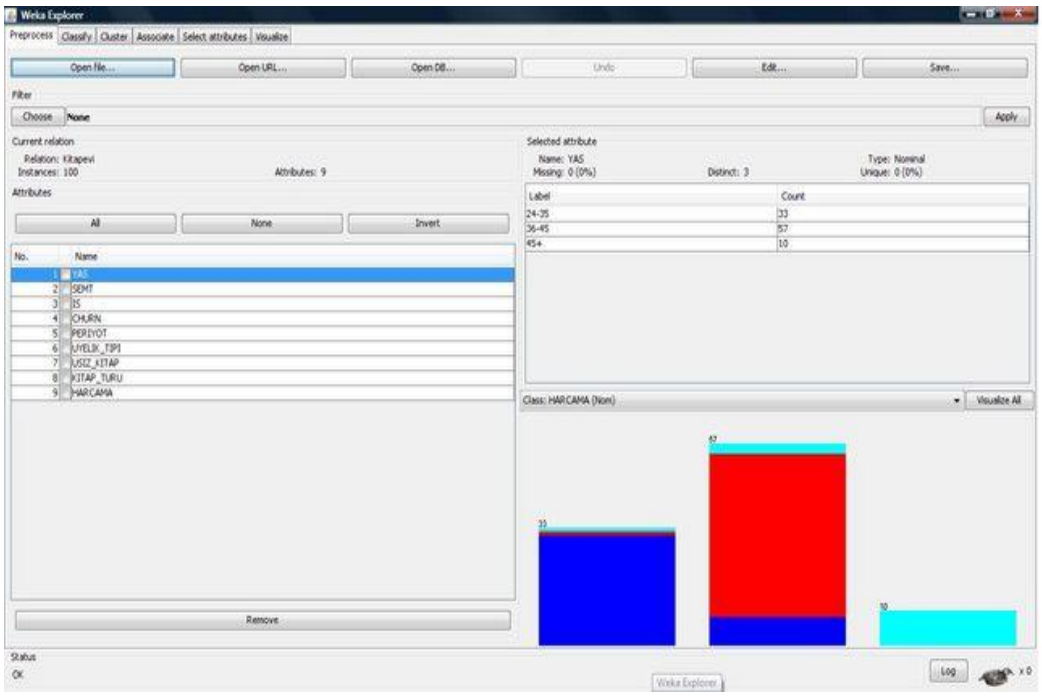
CLARANS Algoritması)
ü Yoğunluğa Dayalı Yöntemler (DBSCAN Algoritması, OPTICS Algoritması, DENCLUE Algoritması)
ü Grid Temelli Algoritmalar (STING Algoritması, Dalga Kümeleme, CLINQUE Algoritması)
ü Genetik Algoritmalar olarak sayılabilir.

Yukarda sayılan algoritmalardan K-Means algoritması, WEKA ve Excel Data Mining Add-In kullanılarak Kümeleme işlemine sokulup Müşteri Segmentasyonu belirlenecektir.

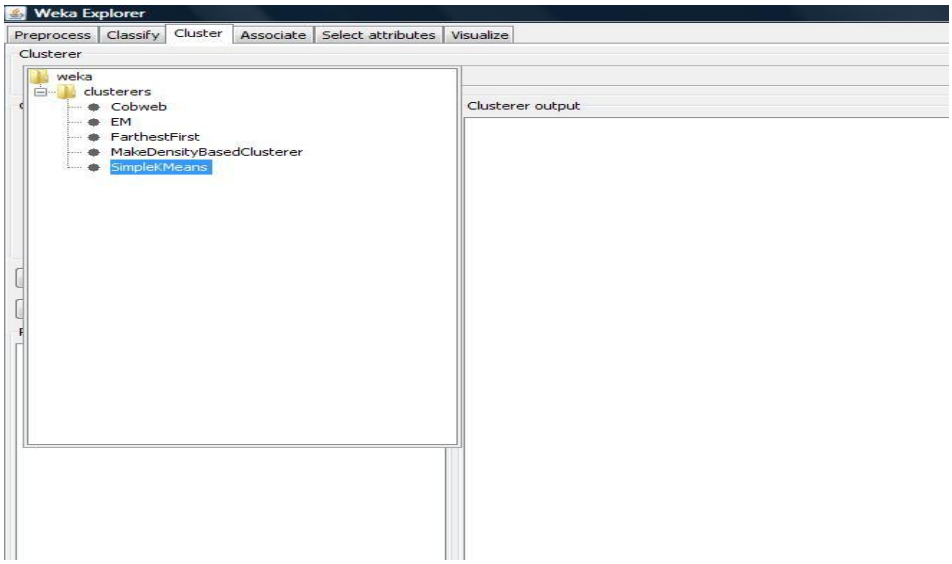
Modelin WEKA Yazılımında Gerçeklenmesi

WEKA açık kaynak kodlu GNU lisanslı bir yazılımdır. Yazılımın çalışabilmesi için bilgisayarda Java 1.4 veya daha sonraki bir sürümün yüklenmiş olması gerekmektedir. Weka yazılımında ARFF (Attribute-Relation File Format) dosyaları kullanılır. ARFF dosyaları, değişken tanımlanmasına izin veren ASCII metin dosyasıdır. ARFF dosyasının başlık kısmında, değişkenler (veritabanındaki herbir kolonun ismi), bunlar arasındaki ilişkiler ve herbir değişkenin türü ve alacağı değer vs bulunur. Veriler @DATA satırından sonra gelir.(Weka da büyük küçük harf duyarlılığı vardır.) Weka Yazılımını ücretsiz olarak buradan indirebilirsiniz.

Bu projede, weka yazılımının ilk açılışında karşımıza çıkan 4 seçenekten Explorer düğmesine tıklandığında karşımıza çıkan kısımda, kümeleme işlemlerinin yapılacağı alanda (Clustering Sekmesi altında) çalışmalarımızı yapacağız.
"Kitapevi.arff" isimli dosyamızı open düğmesine tıklayarak yüklediğimizde bir sonraki sayfadaki ekran görüntüsü karşımıza gelecektir.



Yukardaki Veri Madenciliği görevlerinden (Classify, Cluster, Associate) arasından veri setimizdeki müşteri segmentasyonu görevine en uygun olacağını düşündüğümüz "Cluster" sekmesini seçtiğimizde aşağıdaki gösterilen algoritmaları kullanabileceğimizi görebiliriz.(Burada hiçbir algoritmanın parametreleriyle oynanmamıştır,hepsi default değerlerinde çalışmaktadır.K-Means için default olarak k yani küme sayısı 2' dir)



Şimdi buradaki algoritmaları sırasıyla veri setimize uygulayacağız ve en anlamlı sonucu veren algoritmayı bulup Excel Data Mining'de de kullanıp karşılaştırmalarımızı yapacağız.

Makalemi daha da uzatmamak için K-Means dışında ki 4 algoritmanın sonucunu aşağıda sizinle paylaşıyorum. (Sizin de bu 4 algoritmayı çalıştırıp,sonuçlarını görmenizi tavsiye ederim.)

Cobweb algoritmasının türkçesi örümcek ağı algoritması olduğundan dolayı örümcek ağına çok benzer bir ağ şeması ortaya çıktı.
EM (Expectation Maximazation) Algoritması sonucunda müşterilerimizin 3 gruba (cluster'a) ayrıldığını gözlemliyoruz. (%48,%11,%41 olarak)
FarthestFirst algoritmasının sonucunda veri setimiz 2 gruba ayrılmıştır.(%82 ve % 18 olarak)
MakeDensityBasedClusterer algoritması veri setimizi 2 gruba ayırmıştır. (%43 ve %57 olarak)

Beşinci olarak SimpleKMeans Algoritmasını çalıştırıyoruz. Bu algoritma veri setimizi 2 gruba ayırmıştır. (%47 ve % 53 olarak) Aşağıda bu algoritma için sonucu görebilirsiniz.

```
--- Run information ---

Scheme:       weka.clusterers.SimpleKMeans -N 2 -S 10
Relation:     Kitapevi
Instances:    100
Attributes:   9
              YAS
              SEMT
              IS
              CHURN
              PERIYOT
              UYELIK_TIPI
              USIZ_KITAP
              KITAP_TURU
              HARCAMA
Test mode:    evaluate on training data

=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 256.0

Cluster centroids:

Cluster 0
        Mean/Mode:  24-35 1 E H 7-12 1 H 1 100-250
        Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A   N/A   N/A   N/A
Cluster 1
        Mean/Mode:  36-45 2 E H 12+ 1 H 1 251-400
        Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A   N/A   N/A   N/A

Clustered Instances

0       47 ( 47%)
1       53 ( 53%)
```

Çıkan sonuçları yorumlamak gerekirse K-Means Algoritması 100 adet müşteri bilgimizi aldı ve 2 gruba (cluster yada türkçesiyle kümeye) ayırdı. Birbirinden farklı bu 2 grubtan birincisine ait müşterilerin genel özelliklerine baktığımızda; 24 ile 35 yaş arasında olduklarını ,1 nolu semt de oturduklarını, iş sahibi olduklarını, hala kitapevinin müşterisi olduklarını, 7 ile 12 arasında periyotlarla alışveriş yaptıklarını, 1 nolu üyelik tipinde olduklarını, ücretsiz kitap taleblerinin olmadığını, Bilimsel kitapları daha çok satın aldıklarını ve gelirlerinin 100 ile 250 TL arasında değiştiklerini görebiliriz.

Bu makalede ücretsiz bir veri madenciliği aracı olan WEKA ile İş Zekasının çalışma alanlarından olan müşteri segmentasyonunu gerçekleştirmiş olduk. Bundan sonra ki bağlı makalede gene aynı veri setini ve K-Means algoritmasını kullanarak Excel Data Mining Add-In yardımıyla iş zekası görevimizi gerçekleştireceğiz.

Giriş Düzeyinde Örnek Bir Veri Madenciliği Projesi-2

Bir önceki makalemizde Weka programı yardımıyla elimizdeki veri seti üzerinde müşteri segmentasyonunu gerçekleştirmiştik. Bu makalemizde de aynı veri seti üzerinde aynı görevi Excel Data Mining Add-In yardımıyla gerçekleştireceğiz.

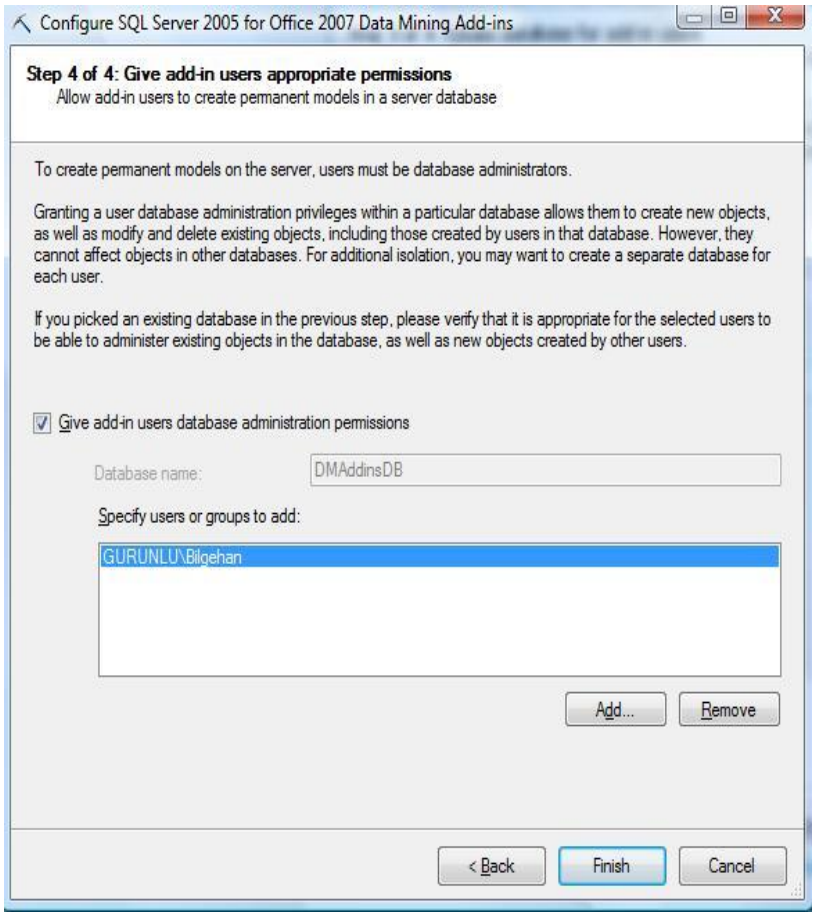
İlk olarak belirtmek isterim ki Excel Add-In aslında analiz çalışmalarını kendisi yürütmemektedir. Excel, SQL Server' a bağlanarak analizi gerçekleştirmekte ve algoritmaların sonucunu son kullanıcının anlayacağı görselikte sunmaktadır. Ayrıca şunu da hatırlatmak isterim ki; iş zekası uygulamalarının son kullancıya aktarılması için SQL Server dışında Oracle başta olmak üzere birçok İş Zekası yada Veri Tabanı yazılımlarının "Excel Add-In" leri bulunmaktadır. (Bu makalede kullanılan SQL Server 2005 Add-Ins 'ine " *http://www.microsoft.com/downloads/details.aspx?FamilyID=7C76E8DF-8674-4C3B-A99B-55B17F3C4C51&displaylang=en* " adresinden ulaşabilirsiniz.)

SQL Server 2005 Data Mining Add-In Kurulumu

Microsoft Office Excel 2007 yazılımı, veri madenciliği görevlerini gerçekleştirmek için DataMining Add-In 'ine sahiptir. Office Excel 2007 yazılımı, Microsoft SQL Server Analysis Servisinin son kullanıcının daha kolay kullanması için 9 adet veri madenciliği modelini uygulama fırsatını sağlamaktadır. Excel ortamında sağlıklı bir şekilde İş Zekası projeleri için Veri Madenciliği görevlerinin gerçekleştirilmesi için SQL Server 2005 Add-In'in kurulum sonrasında kendi SQL Server'ınıza göre ayarlanması gerekmektedir. Aşağıdaki adımları takip ederek öncelikle SQL Server 2005 içi gerekli ayarlamaları gerçekleştirmeliyiz. (Aşağıdaki ayarlar İngilizce Vista ve Office 2007 ye göredir.)

İlk olarak Programs/Microsoft SQL Server Add-Ins/Server Configuration Utility sihirbazını çalıştırıyoruz.

Eğer daha önceden bu işlem için bir database oluşturmamışsanız, "DMAddinsDB" olarak bir database oluşturmasına izin verip, Next butonuna basıyoruz ve hangi kullanıcıya erişim yetkilerini vereceğimiz aşağıdaki ekrana geçiyoruz.

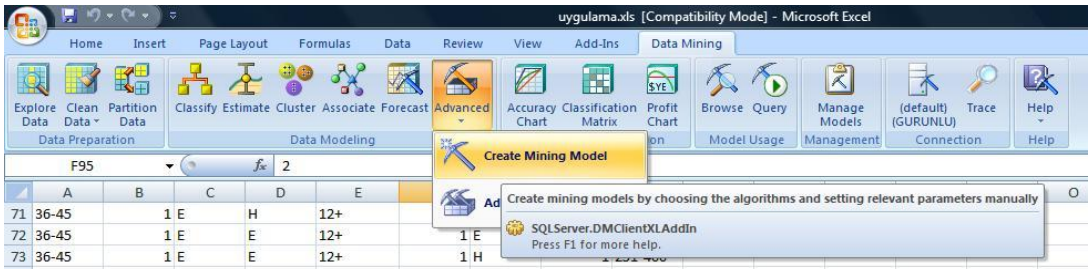


Yukardaki ekranda hangi kullanıcı aracılığıyla Excel 2007'nin SQL Server Analysis Service 'a erişeceğini belirliyoruz. Burada seçilecek kullanıcı Analysis Service'da Admin yetkilerine sahip olmalıdır.
Böylelikle 4 adımda Excel 2007 Data Mining Add-In 'ini sağlıklı bir şekilde kullanabilmek için SQL 2005 Analysis Service için gerekli konfigurasyonları wizard sayesinde gerçekleştirdik.

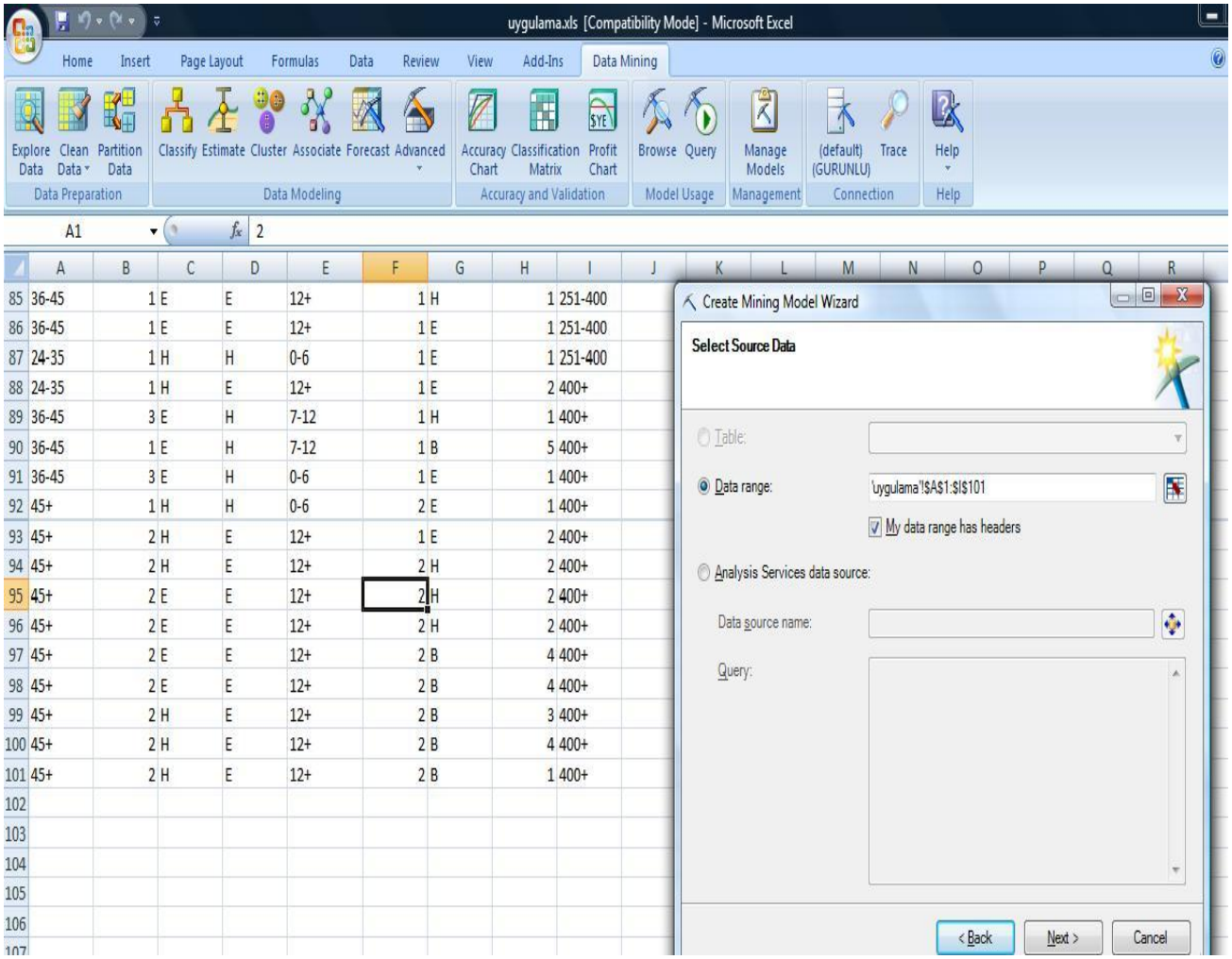
Modelin Excel 2007 DataMining Kullanılarak Gerçeklenmesi

SQL Server 2005 Data Mining Add-In' ni bilgisayarınıza kurduktan sonra Excel 2007 'yi açtığınızda aşağıdaki gibi Data Mining Sekmesi eklenmiş olacaktır.

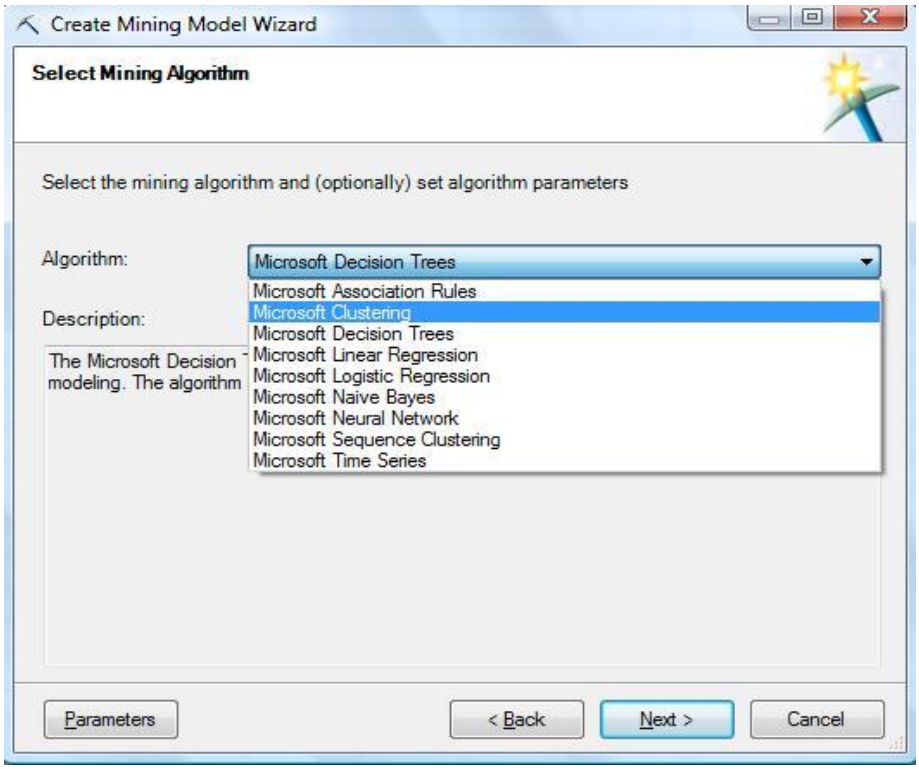
Veri Setimiz üzerinde, WEKA yazılımında gerçekleştirmiş olduğumuz kümeleme işlemlerini Excelde de gerçekleştirebilmek için xls formatındaki verileri excel'de açıyoruz ve daha sonrasında Data Mining Sekmesi altında ki Cluster işlemlerini yapabilmek için Advanced / Create Mining Model düğmesine tıklıyoruz.



Karşımıza Data Mining Cluster Wizard çıkıyor.

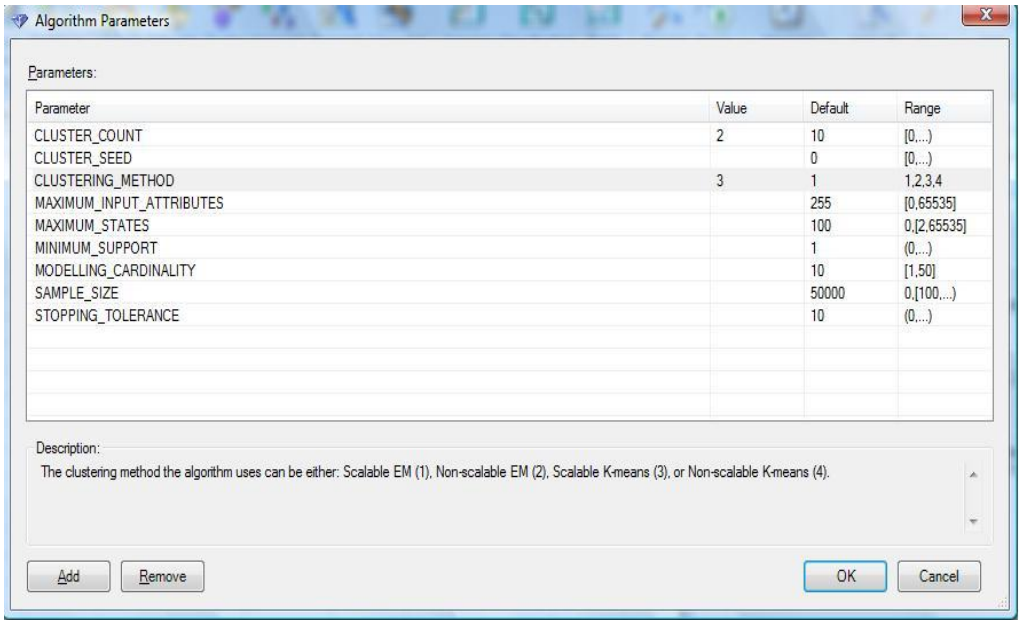


Next diyoruz ve veri madenciliği modelimizi seçeceğimiz ekrana ulaşıyoruz.

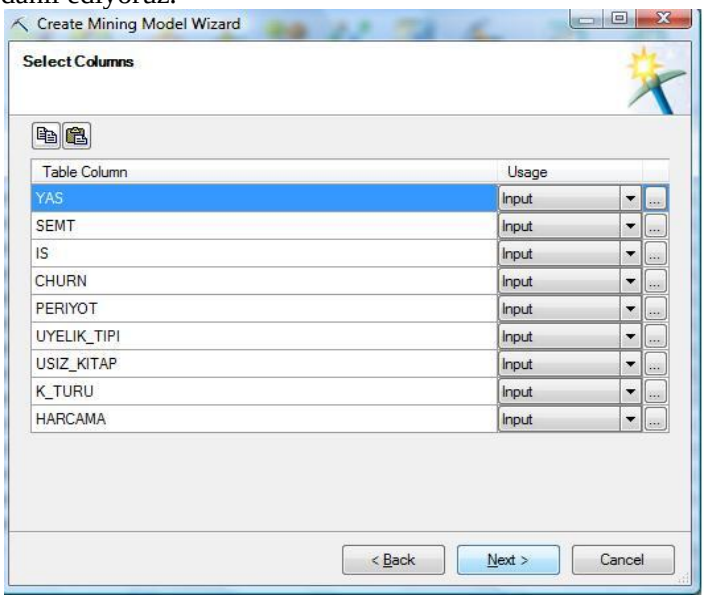


Yukardaki gibi karşımıza çıkan ekranda istediğimiz parametreleri girebilmek amacıyla "Parameters " butonuna tıklıyoruz ve Weka ile aynı parametrelerin geçerli olmasını sağlıyoruz.
Burada hemen şunu belirtmek isterim ki ; Microsoft Clustering Algoritması 4 farklı Clustering Algoritmasından oluşmaktadır ve biz hangi algoritmaya göre kümeleme işlemini gerçekleştireceğimizi belirtmeliyiz. (2)

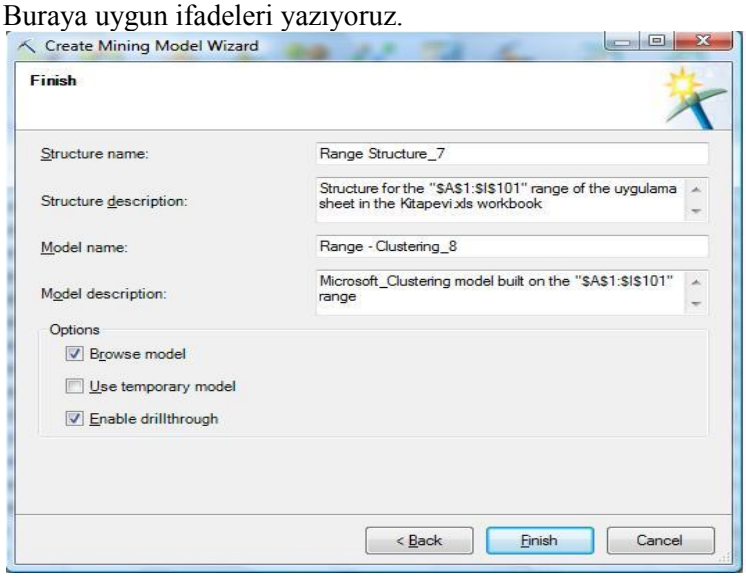
Microsoft Clustering Modeli aşağıda ki 4 algoritmayı desteklemektedir.
1) Scalable EM (default olarak bu seçilidir.)
2)Vanilla (non-scable) EM
3) Scable K-Means (biz bu proje için bunu tercih ediceğiz.)
4) Vanilla (non-scable) K-Means



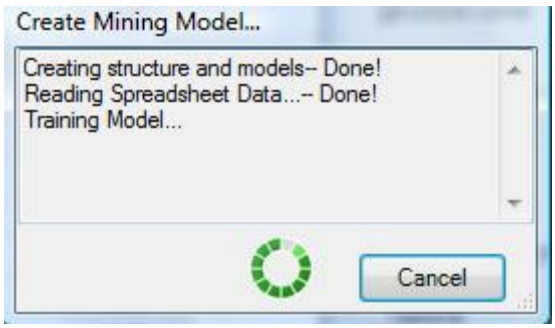
Yukarda Cluster_Count olarak 2 sayısını girerek; 2 kümeye ayırmasını istedik. (Zaten K-Means Algoritmasında ki K bizim belirlediğimiz küme sayısıdır.)
Clusterin_Method olarak da 3 sayısını girerek "Scalable K-means" algoritmasını seçmiş olduk.
Bir sonraki ekranda kümelemeye dahil edeceğimiz, attribute'lar karşımıza çıkıyor. Burada hepsini dahil ediyoruz.



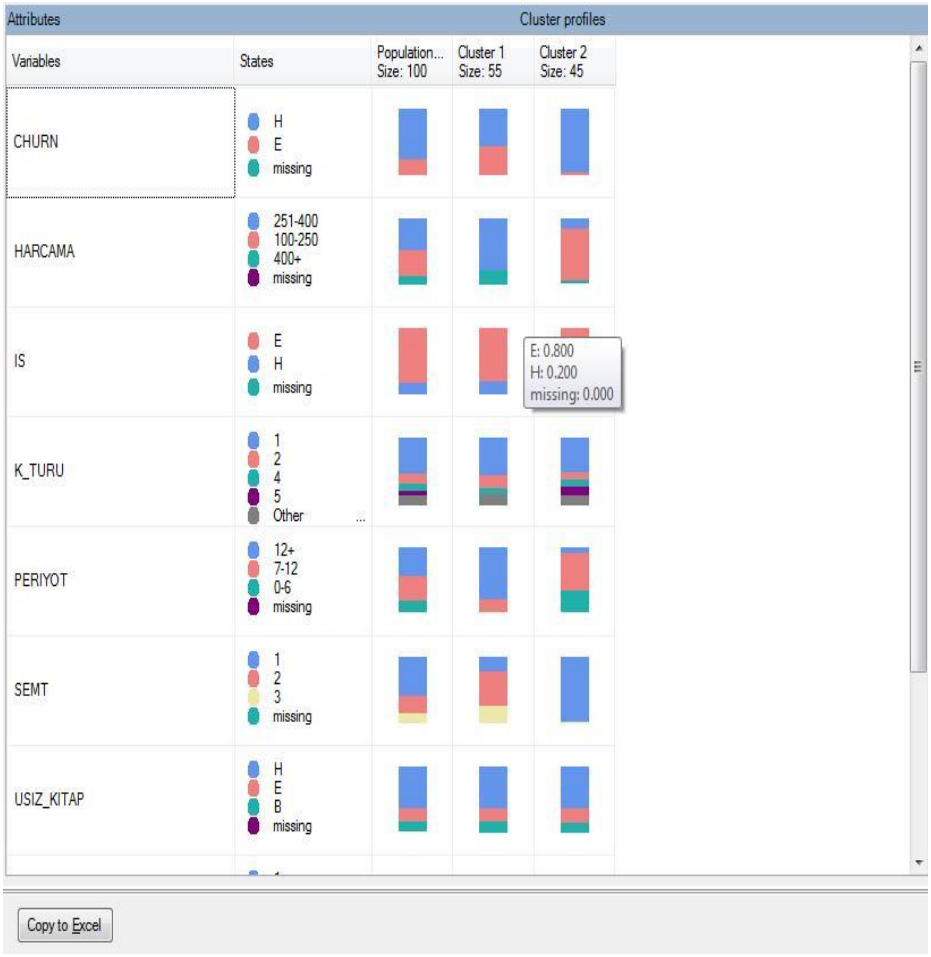
Bir sonraki ekranda gerçekleştirilecek kümeleme işlemine isim ve tanım bilgilerini girmemiz isteniyor. Buraya uygun ifadeleri yazıyoruz.



Tüm değerleri girdikten sonra, Finish butonuna basıyoruz ve kümeleme işlemini başlatıyoruz.



Kümeleme işlemi bittiğinde, müşterilerin % 55 ve % 45 oranında kümelendiği sonucunu görüyoruz.



Sonuçların Karşılaştırılması :

Weka yazılımında, K-Means algoritması ile gerçekleştirilmiş olan kümeleme işleminde müşteriler %47 ve % 53 oranlarında 2 gruba ayrılmıştılar. Buna karşılık Excel Data Mining Add-In kullanılarak yapılan kümeleme işleminde ise %55 ve % 45 gibi bir oranla karşılaşıldı. 100 kişinin verisi üzerinde algoritmayı çalıştırdığımıza göre çıkan % 2 lik farklılık 2 kişiye karşılık geliyor olarak düşünülebilir.Oluşmuş olan bu farklılığın, 1 veya 2 kayıdın Excel tarafından yada Weka tarafından diğer kümeye dahil edilmesinden kaynaklanmış olabilir. Örnek olarak bir kayıdın, birinci küme merkezine olan uzaklığı ikinci küme merkezine olan uzaklığına daha yakınsa o kayıd ikinci kümenin elemanı olarak değerlendirilmiş olmaktadır. Bu da oluşan kümelerin bu gibi durumlarda homojenliğini kayıp edebileceği sonucunu ortaya koymuştur. Outlier değerlerde ki oluşabilecek sapmalara dikkat etmek gerekmektedir.

Bilgehan GÜRÜNLÜ
www.gurunlu.com
bilgehan@gurunlu.com

(1) Kavram ve Algoritmalarıyla Veri Madenciliği –Gökhan Silahtaroğlu - Papatya Yayınları-2008
(2) Data Mining with SQL Server 2005 – Tang/ MacLennan – Wiley Press - 2006