

**HACETTEPE ÜNİVERSİTESİ İSTATİSTİK BÖLÜMÜ VERİ MADENCİLİĞİ PROJE ÖDEVİ**

**BERKE YAĞMUR 21322361**

**BÜŞRA ŞENCAN 21622147**

**BANK MARKETİNG DATA SET**

**VERİ HAKKINDA BİLGİ**

Veri madenciliği, geleneksel yöntemlerle anlaşılmayan büyük verilerden anlamlı bilgi çıkarma ve bunu eyleme dökme işlemidir. Bu kapsamda, müşterilerin profillerinin araştırılması sonucu müşterinin bir vadeli mevduata abone olup olmadığı ele alınmıştır. Bu bizim sınıflandırma hedefimizdir.

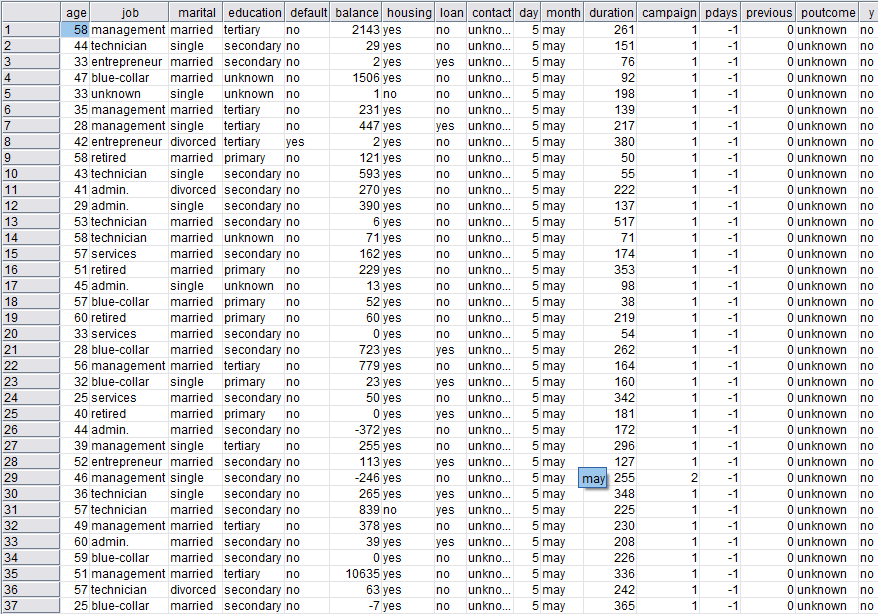
Veri madenciliği; verinin analiz edilmesi, analiz sonucunda ortaya çıkan bilgilerin değerlendirilmesi ve yorumlanmasını sağlayan bir işlem dizisinden oluşmaktadır. Bu işlemler farklı yöntemler ve programlar kullanılarak yapılabilir. Biz çalışmamızda orange, modeller ve clementine adlı programlar üzerindeki algoritmaları kullanacağız. Vadeli mevduat hesabına abone olup abone olmayan müşterilerin hangi özelliklere sahip olduğunu belirlemek amacıyla, sınıflandırma yöntemi kullanılarak ürün önerme, sonucunda da bankacılık sektörüne müşterilerin etkinlik ve aktifliklerini arttırmaya fayda sağlayacak bir çalışma elde etmeye çalışıyoruz.

Veriler, bir Portekiz bankacılık kurumunun doğrudan pazarlama kampanyaları (telefon görüşmeleri) ile ilgilidir. Telefon görüşmelerinden alınan bilgiler doğrultusunda banka müşterisi olup olmama durumu ile ilgilenilir.

**VERİ KÜMESİ İÇERİĞİ**

Bu çalışmada her bir müşteri için 17 tane özellik gözlemlenmiştir. Bunlar; Age, Job, Marital, Education, Default, Balance, Housing, Loan, Contact, Day, Month, Duration, Campaing, Pdays, Previous, Poutcome, Y (target) değişkenleridir.

Ham verimizin önizlemesi;



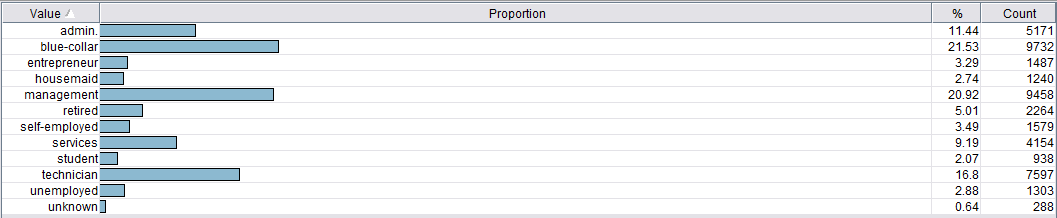
|  |  |  |
| --- | --- | --- |
| **Bağımsız Değişkenler** | **Açıklama** | **Düzeyleri** |
| Age | Yaş | Sürekli |
| Job | Meslek | Kategorik (yönetici, bilinmeyen, mavi yakalı, işsiz, yönetim, hizmetçi, girişimci, öğrenci, serbest meslek, emekli, teknisyen, hizmetler) |
| Marital | Medeni Durumu | Kategorik (evli, bekar, boşanmış) |
| Education | Eğitim Durumu | Kategorik (bilinmiyor, orta, ilköğretim, yükseköğretim) |
| Default | Varsayılan olarak kredi var mı? | İkili (evet, hayır) |
| Balance | Ortalama yıllık bakiye (avro) | Sayısal |
| Housing | Konut kredisi var mı? | İkili (evet, hayır) |
| Loan | Kişisel krediniz var mı? | İkili (evet, hayır) |
| Contact | İletişim türü | Kategorik (bilinmiyor, telefon, hücresel) |
| Day | Son irtibat günü | Sayısal |
| Month | Yılın son iletişim ayı | Kategorik |
| Duration | Son iletişim süresi (saniye) | Sayısal |
| Campaign | Bu kampanya için müşteri ile gerçekleştirilen iletişim sayısı | Sayısal |
| Pdays | Müşteriyle önceki kampanya ile ilgili en son iletişime geçilen tarihten bugüne kadar geçen gün sayısı | Sayısal (-1 = müşteriyle ilk defa iletişime geçildiği anlamına gelir.) |
| Previous | Bu kampayadan önceki kampanya için müşteriyle gerçekleştirilen iletişim sayısı | Sayısal |
| Poutcome | Önceki pazarlama kampanyasının sonucu | Kategorik (bilinmiyor, diğer, başarısızlık, başarı) |
| Y | Müşteri bir vadeli depozito yatırdı mı? | İkili (evet, hayır) |

Y değişkeni hedef değişkenimizdir.

**DEĞİŞKENLER HAKKINDA BİLGİ**

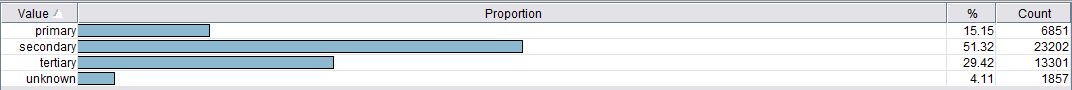
**BAZI DEĞİŞKENLERİN DAĞILIM DURUMU**

1. **MÜŞTERİLERİN MESLEK DAĞILIMI**



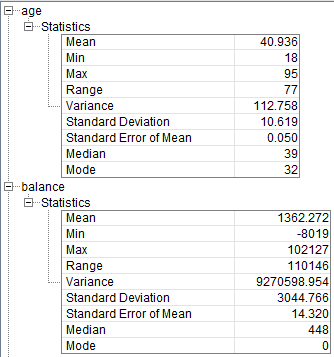
45212 müşteri arasından 13 meslek türünün dağılımları gösterilmiştir.

1. **MÜŞTERİ EĞİTİM DURUMU DAĞILIMI**



45212 müşteri arasında 4 eğitim türünün dağılımları gösterilmiştir.

**ÖZETLEYİCİ İSTATİSTİKLER**



Yıllık bakiye için özetleyici istatistikler;

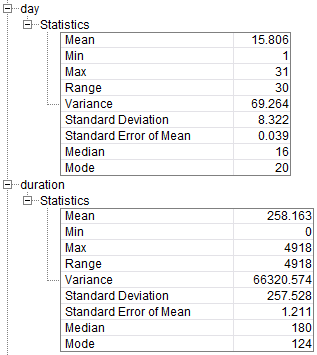
Bu çalışma için yıllık bakiye değeri minimum -8019, maksimum 102127’dir.

Yaş değişkeni için özetleyici istatistikler;

En fazla tekrar eden müşteri yaşı 32’dir.

Müşterilerin yaş aralığının 18 ve 95 arasında olduğunu görüyoruz.

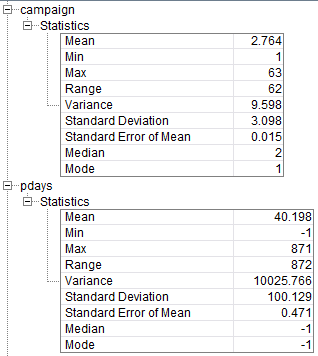
Ortalama yaş yaklaşık 41’dir.



Süre için özetleyici istatistikler;

Son iletişim süresi ortalama 258.163 saniyedir.

Gün için özetleyici istatistikler;

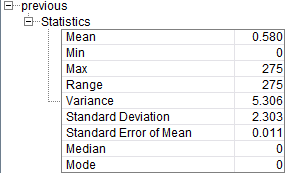


Kampanya için özetleyici istatistikler;

Bu kampanya için müşteriyle gerçekleştirilen iletişim sayısı en az 1, en fazla 63’tür.

Pday için özetleyici istatistikler;

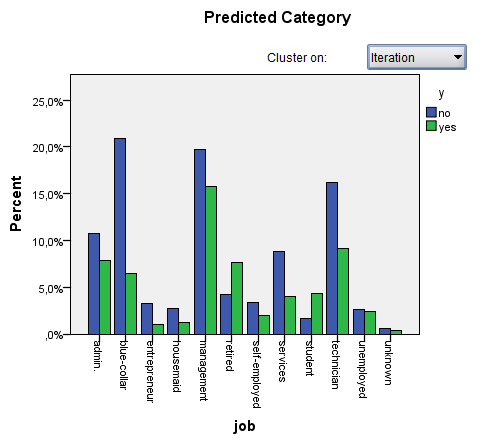
Müşteriyle önceki kampanya için ortalama 40 gün iletişime geçilmemiştir.



Previous için özetleyici istatistikler;

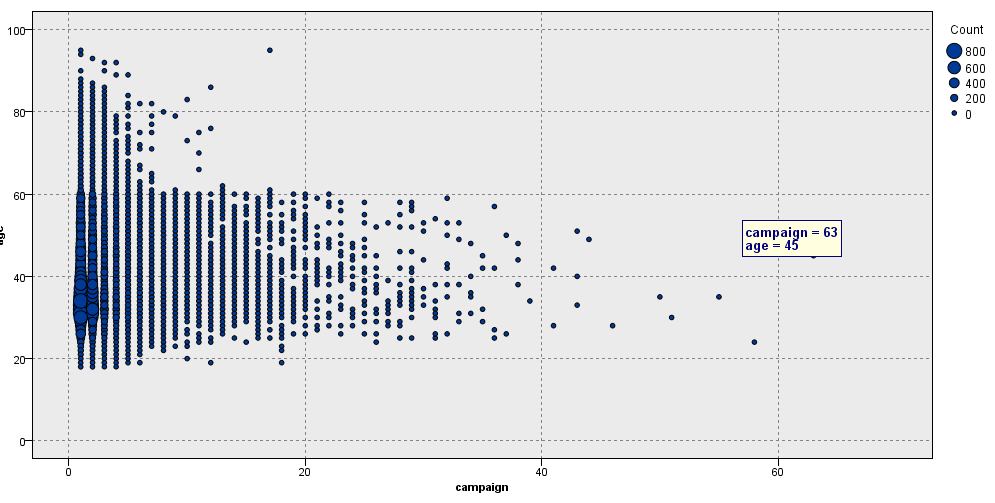
Önceki kampanya için müşteriyle gerçekleştirilen iletişim sayısı ortalama 0.580’dir.

**GRAFİKLER**



Histogram grafiği kampanyalar sonucunda evet ve hayır olarak belirlenen hedef değişkenimiz ( vadeli mevduata abone olup olmama) ile meslek değişkeninin dağılımını gösterir. Grafik baz alındığında yaklaşık %20’lik düzeyde mavi yakalıların vadeli mevduata abone olmadığı, yaklaşık %7’lik oranla vadeli mevduata abone olduğu gözlemlenmiştir.

Yönetim sektöründe bulunan kişilerin ise yaklaşık %15’lik kısmının vadeli mevduata abone olduğunu, yaklaşık %20’lik kısmının vadeli mevduata abone olmadığı gözlemlenmiştir.



Plot grafiği ise yaş değişkeni ile kampanya için müşteriyle gerçekleştirilen iletişim sayısı hakkında bilgi için çizdirilmiştir. Elimizdeki verilere göre 20 ve 60 yaş arasındaki müşteriler ile gerçekleştirilen iletişim sayısı diğer yaş aralıklarına göre daha fazladır. 45 yaşındaki müşterilerle gerçekleştirilen iletişim sayısı en fazla değere sahiptir.

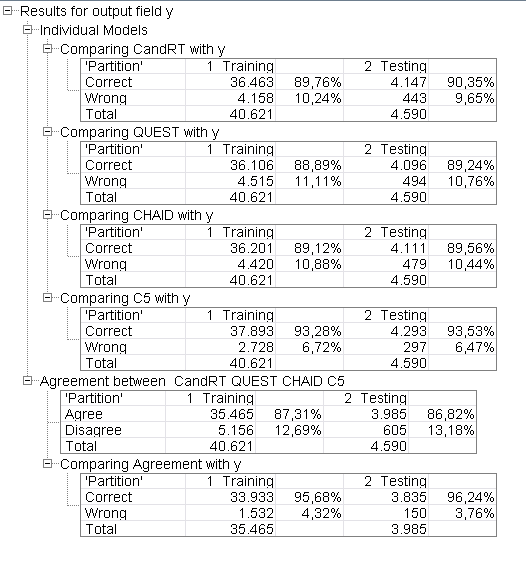
**SINIFLANDIRMA YÖNTEMLERİ**

1. **KARAR AĞAÇLARI**

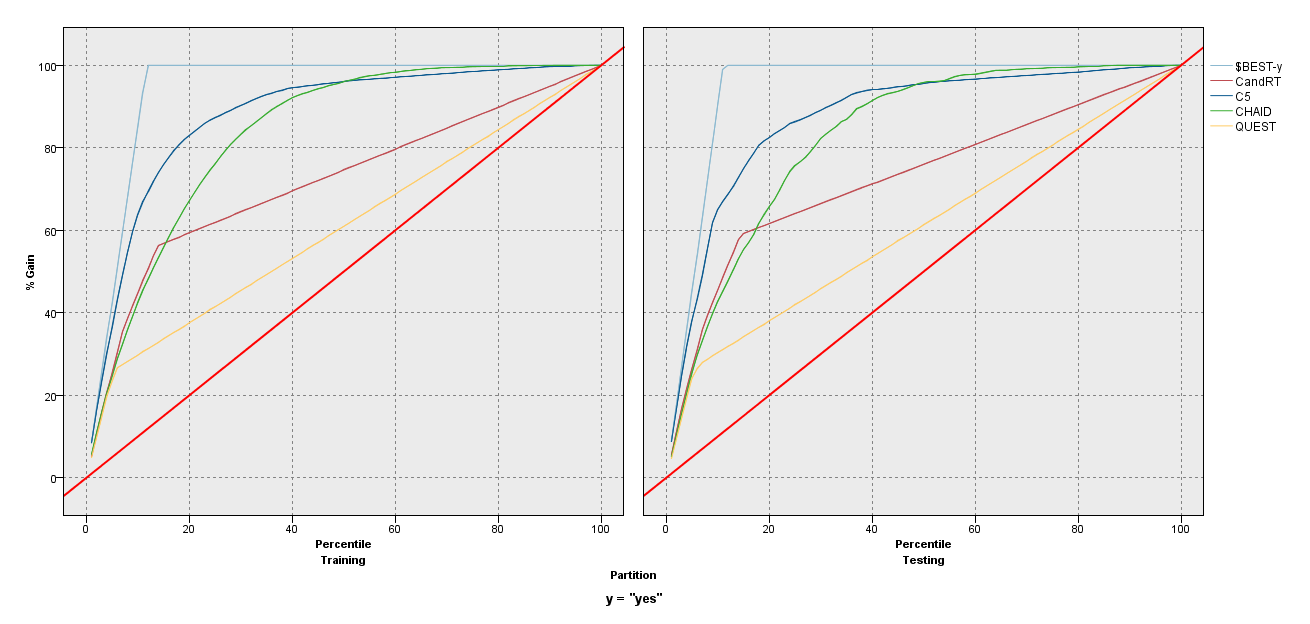
Karar ağaçları bir istatistiksel sınıflandırma algoritmasıdır. Karar ağaçları algoritması tüm veri setini eğitim ve test verisi olarak ayırarak verinin sınıflanmasını iki aşamalı olarak gerçekleştirir.

İlk aşama olan öğrenme aşamasında, önceden bilinen bir takım eğitim verisi modelin oluşturulması amacıyla algoritma tarafından kullanılır. Öğrenilen modelden bir karar ağacı oluşturulur.

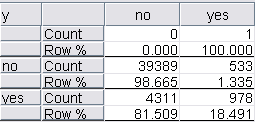
İkinci aşama olan sınıflama aşamasında ise test verileri kullanılarak karar ağacının başarısı belirlenir. Her test verisi örneğinde bilinen sınıf ile model tarafından tahmin edilen sınıf karşılaştırılması yapılır. Modelin doğruluğu, yaptığı doğru sınıflamanın tüm test verisine oranıdır.



En iyi algoritma olarak C5 algoritmasına karar verildi. (Sınıflandırma başarı yüzdesi diğerlerine göre daha yüksek). Bu algoritma ile gelecekteki verinin hangi sınıfa atanacağına dair tahminler gerçekleştireceğiz.

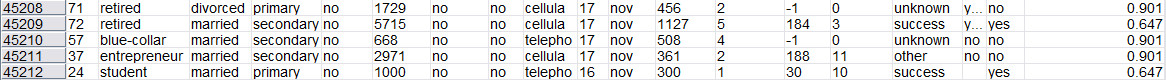


C5 algoritmasını seçmemizin bir diğer sebebi yukarıdaki grafikte gösterildiği gibidir. En iyi çizgisine en yakın karar ağacı algoritması C5’dir.



Yukarıda C5 algoritmasının Konfüzyon matrisi verilmiştir. Bu matriste algoritmanın tahmin ettiği hayır cevapları ile gerçekte de hayır diyenlerin oranı %98.665’dir. Aynı şekilde bu matriste algoritmanın tahmin ettiği evet cevapları ile gerçekte evet diyenlerin oranı %18.491’dir.

Bu nedenle potansiyel bir müşterinin verilerini SPSS kullanarak verimize ekledik. İncelediğimiz araştırma hedef değişkenimizin (vadeli mevduat hesabı açmak veya açmamak) iki sınıfından hangi sınıfa gideceğini gözlemlememiz için yapılan bir araştırmadır.



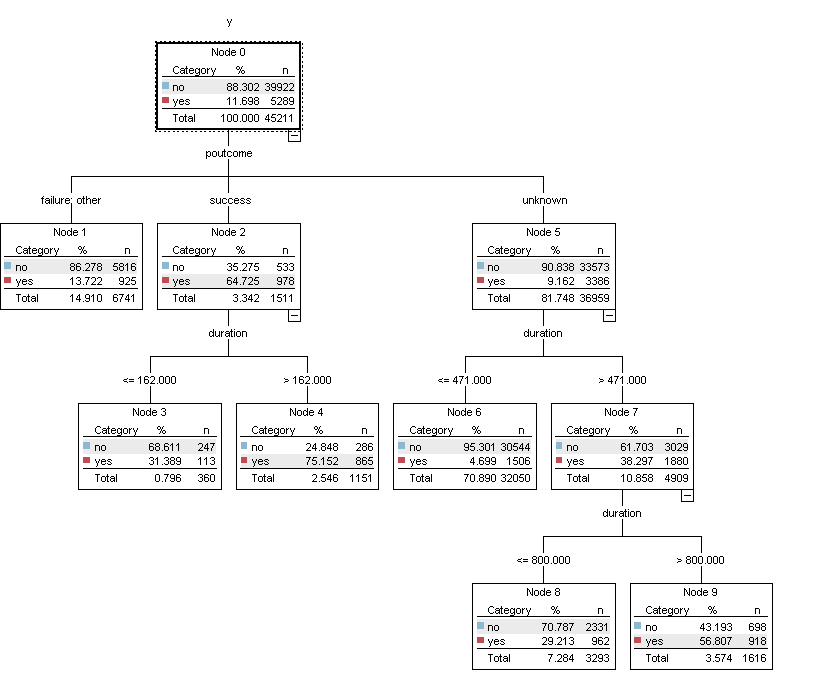
Potansiyel müşterinin girilen verilerine göre vadeli mevduata abone olduğunu gözlemliyoruz.

* 1. **KARAR AĞACI BUDAMA**

Budama işlemi ile karar ağacının sınıflandırma doğruluğunu etkilemeyen kısımlar çıkarılarak daha sade ve anlaşılabilir bir ağaç elde edilir.

Karar ağacı algoritmasında ortaya çıkan sorulardan biri, final ağacının en uygun büyüklüğüdür. Çok büyük bir ağaç, eğitim verilerinin üzerinde çok fazla bir risk oluşturur ve yeni örneklere genellemenin zayıf olmasına neden olur. Küçük bir ağaç örnek alanı hakkında önemli yapısal bilgileri yakalayamayabilir. Bununla birlikte, bir ağaç algoritmasının ne zaman durması gerektiğini söylemek zordur çünkü tek bir ekstra düğüm eklenmesinin hatayı önemli ölçüde azaltıp azaltamayacağını söylemek mümkün değildir. Bu sorun ufuk etkisi olarak bilinir. Ortak bir strateji, her düğüm az sayıda örnek içerene kadar ağacı büyütmek ve ardından ek bilgi sağlamayan düğümleri kaldırmak için budama kullanmaktır.

Bu doğrultuda C5 algoritmasının karar ağacı yapısını budayıp aşağıdaki tabloyu elde ettik.



Gerekli yorumlar;

* Önceki kampanyaya katılmayan müşterilerin %86.278’i vadeli mevduat aboneliğine hayır demiştir.
* Önceki kampanyaya katılım göstermiş müşteriler arasından son iletişim süresi 162 saniyeden büyük olan müşteriler %75.152 oranla vadeli mevduat aboneliğine evet demiştir. Son iletişim süresi 162 saniyeden küçük olan müşteriler ise %68.611 oranla vadeli mevduat aboneliğine hayır demişlerdir.
* Önceki kampanyaya katılım sonucu bilinmeyen müşterilerden, önceki kampanya adına en son iletişime geçilen saniye süresi 471’ten küçük olduğu müşteriler arasında %95.301’lik dilim vadeli mevduata abone olmaya hayır demişlerdir.
* Önceki kampanya sonucu bilinmeyen müşterilerden son iletişim süresi 800 saniyeden küçük olan müşteriler %70.787 oranla vadeli mevduat aboneliğine hayır demiştir.

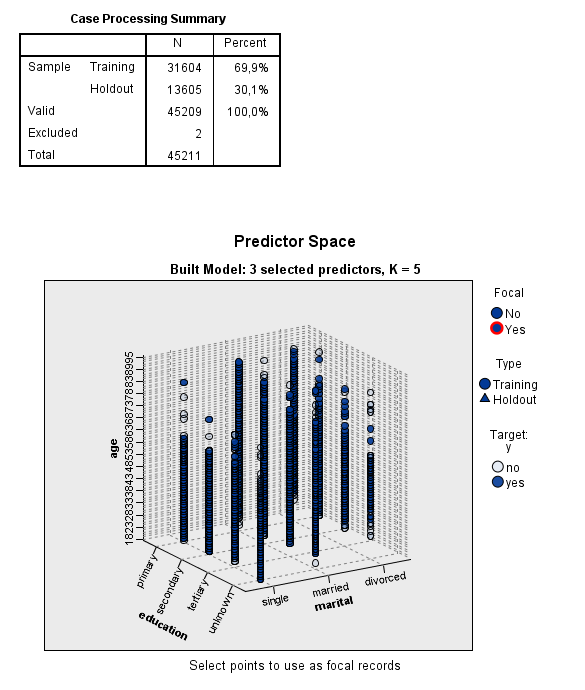
1. **K EN YAKIN KOMŞU**

Bu algoritma, özellikle büyük veri tabanlarında kullanılan, makine öğrenme algoritmaları arasında sıklıkla tercih edilen etkili bir sınıflandırma tekniğidir. K en yakın komşu yöntemi, birçok sınıflandırma probleminde güçlü ve kullanışlı öğrenme ile basit ve etkili çözüm sunan denetimli öğrenme yöntemleri arasında yer almaktadır.

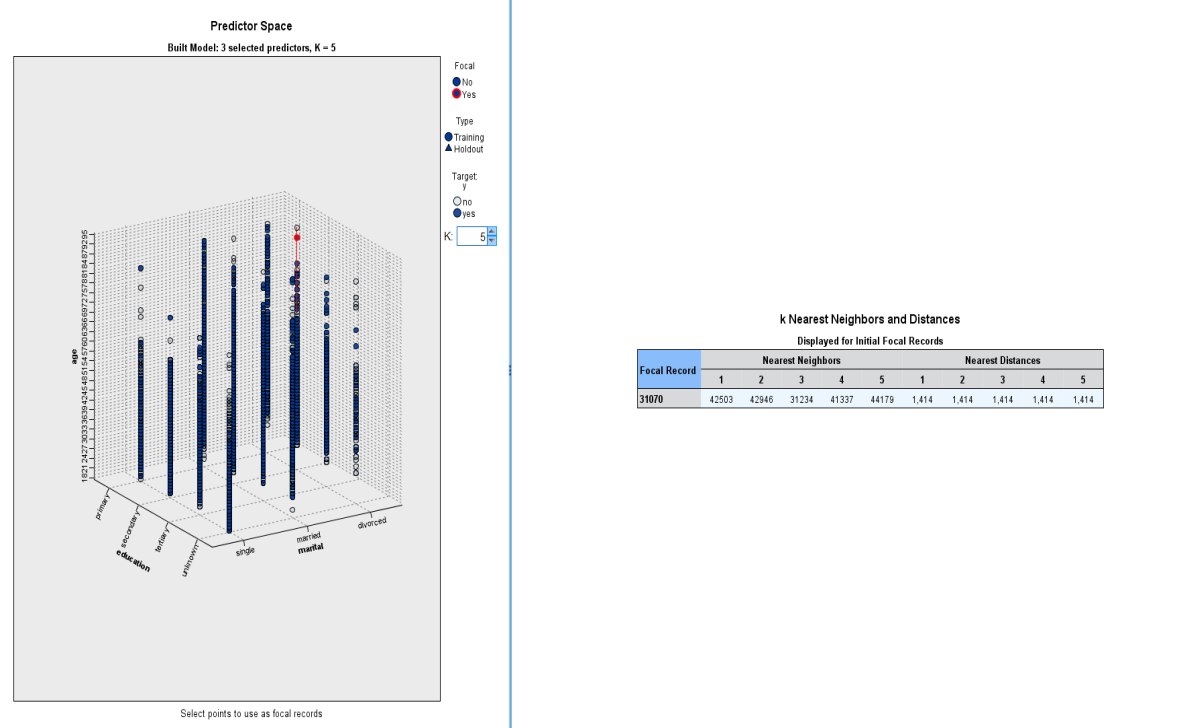
Sınıflandırmada kullanılan bu algoritmaya göre sınıflandırma sırasında çıkarılan özelliklerden, sınıflandırılmak istenen yeni bireyin daha önceki bireylerden k tanesine yakınlığına bakılmasıdır.

Algoritmanın mantığını kısaca şu şekilde özetleyebiliriz;

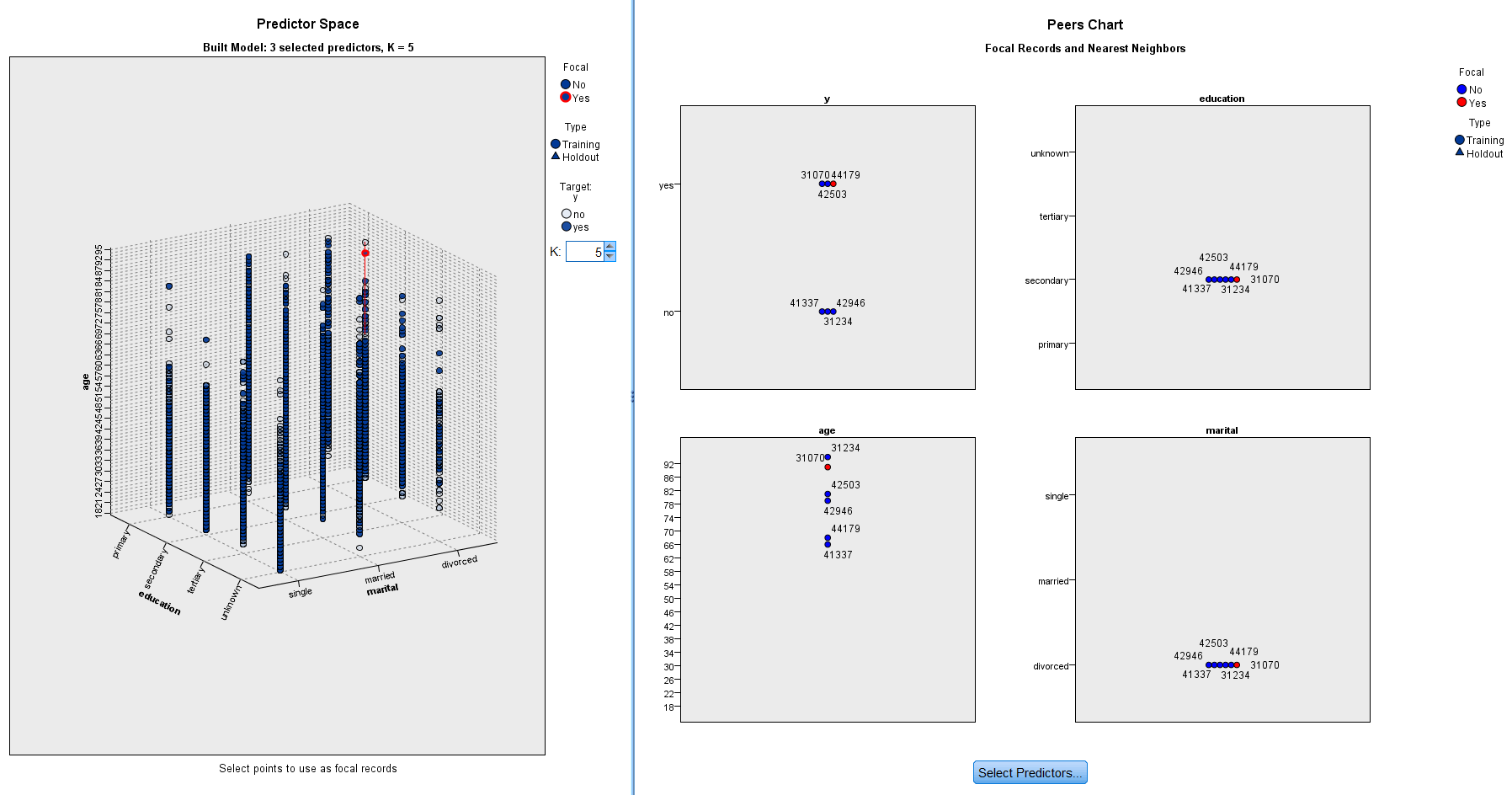
K=5 i ele alırsak, yeni bireyimin hangi sınıfa atanacağına karar verecek iken Öklid uzaklıklarına bakılarak grafikte bireyimize en yakın 5 noktayı ele alırız. Bu noktaya en yakın 5 nokta arasında çoğunluk hangi sınıftaysa yeni bireyimizi de o gruba atarız.



Yukarıda SPSS ile k en yakın komşu uygulamasının çıktısı görülmektedir. K=5 alınmıştır ve değişkenler age,education,marital olarak belirlenmiştir. Grafiğin üstüne çift tıkladığımızda istediğimiz gözlemin yakın komşularını bulacağız.

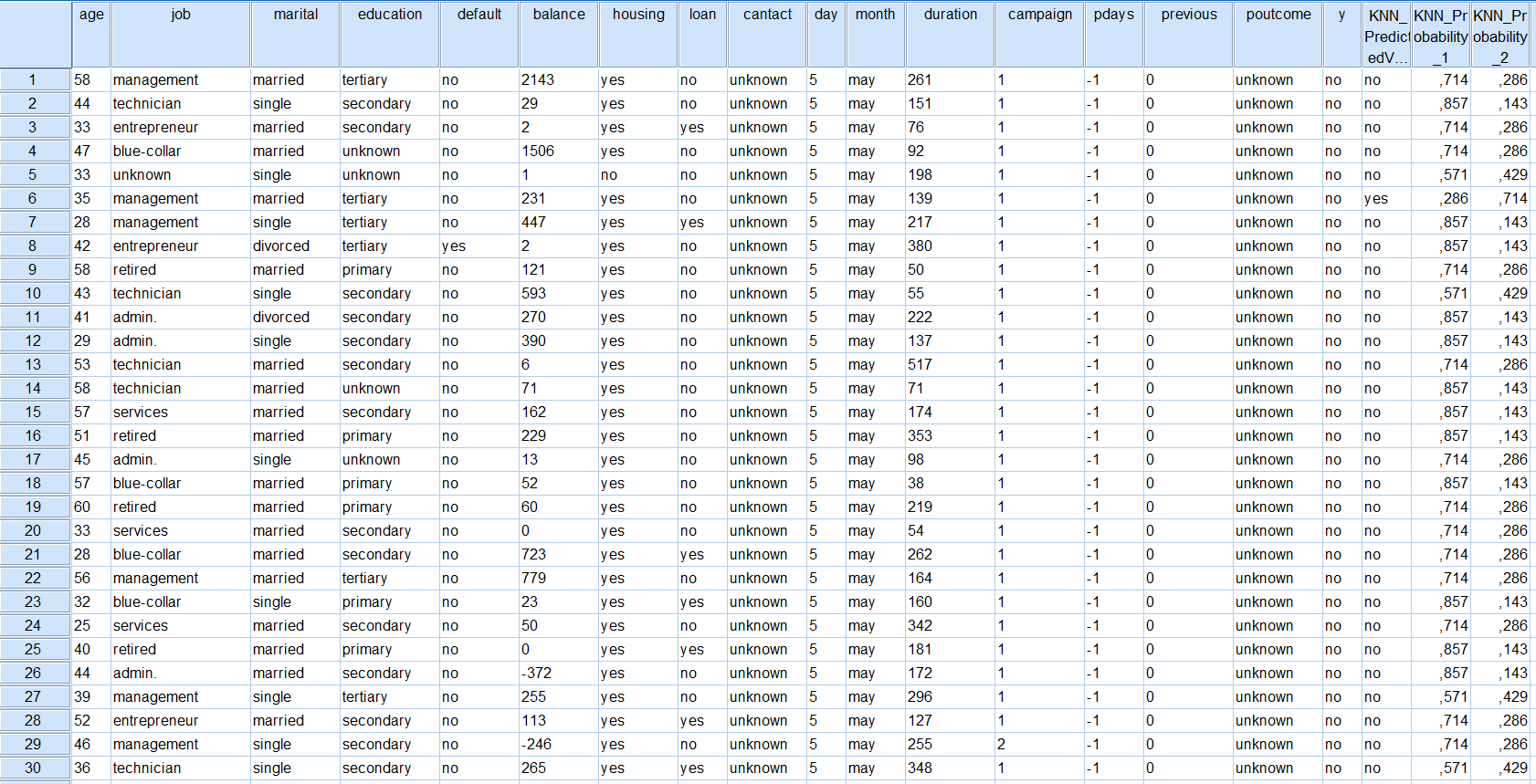


Yukarıda 31070. Gözlemin 5 yakın komşusunun Öklid uzaklarıkları verilmiştir.



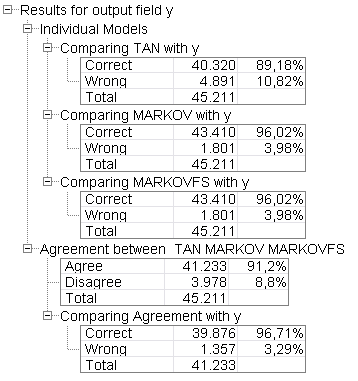
31070 gözleminin komşuları 44179,42503,41337,42968,31234’tür.

31070. gözlemin k en yakın komşu algoritmasına göre 3 komşunun hayır demesiyle birlikte hayır grubunda olacağını söyleyebiliriz.

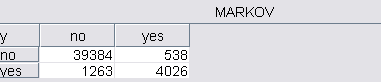


**BAYESCİ AĞLAR**

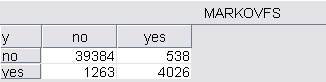
1990’lı yıllarda kullanılmaya başlanan Bayesci ağlar, çok boyutlu veri kümesindeki rastlantı değişkenleri arasındaki olasılıksal ilişkileri kodlayan grafiksel modellerdir. Hem nedensel hem de olasılıksal özelliklere sahip olduklarından, bu ağlar ile veri bilgisi ve uzman görüşü kolaylıkla birleştirilebilir. Bayesci ağlar ile ayrıca, ilgilenilen problemin kesin olmayan tanım bölgesi ile ilgili bilgi temsil edilebildiği gibi, güçlü çıkarsamalar da yapılabilir. İstatistiksel analizlerde Bayesci ağlardan yararlanmak kullanıcıya birçok üstünlük sağlar. Bu üstünlüklerden bazıları; değişkenler arasındaki nedensel ilişkilerin anlaşılmasını sağlamaları, olasılık kuramına dayandığından her zaman tutarlı sonuçlar vermeleri, robust olmaları, uzman görüşünü modellemeye katmaları ve veride kayıp gözlem olması durumunda da güvenilir çıkarsamalar yapmaları olarak verilebilir.



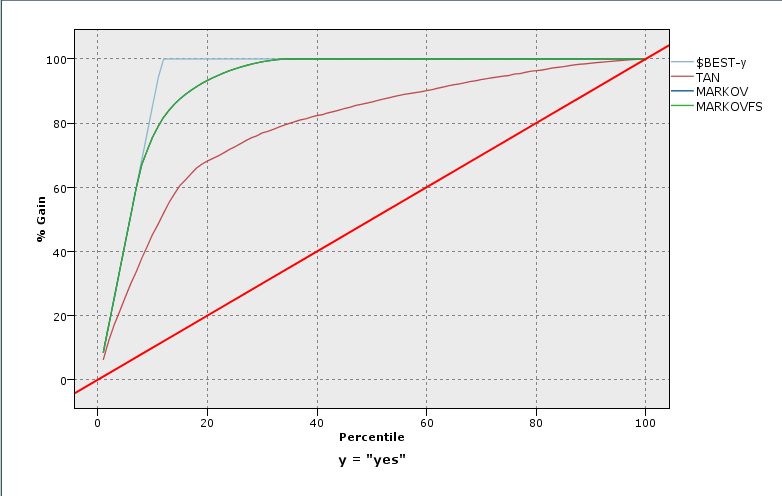
Markov ve markov fs algoritmalarının %96.02 oranla kullanılabileceğini söyleyebiliriz.



Markov’a göre Konfüzyon matrisi yukarıda verilmiştir. Bu matriste algoritmanın tahmin ettiği hayır cevapları ile gerçekte de hayır diyenlerin sayısı 39384 olarak bulunmuştur. Algoritmanın tahmin ettiği evet cevapları ile gerçekte de evet diyenlerin sayısı ise 4026 bulunmuştur.



Markov fs algoritmasına göre Konfüzyon matrisi yukarıdaki gibidir. Görüldüğü gibi markov ile aynıdır.



Markov fs algoritmasının en iyi çizgiye daha yakın olduğunu görüyoruz. Bu sebeple bayesci ağlar arasından markov fs’yi tercih etmeliyiz.

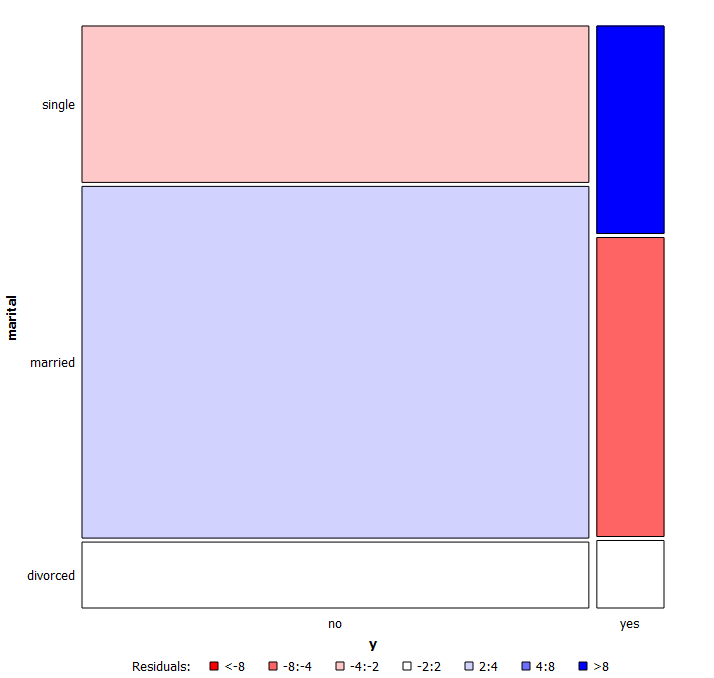
**YAPAY SİNİR AĞLARI**

Yapay sinir ağları biyolojik nöronlardan (sinir hücresi) esinlenerek, beynin çalışma sistemine yapay olarak benzetim çalışmaları sonucunda ortaya çıkmıştır. Genel anlamda insan beynindeki birçok biyolojik nöronun birbirine bağlanması gibi, yapay sinir ağlar; biyolojik nöronun girdi, işlem, çıktı karakteristiğini taklit eden bir çok basit, genellikle adaptif işlem birimlerinin (yapay nöron) değişik etki seviyelerinde, belirli bir bütün işlem yapısını gerçekleştirmek üzere birbirine bağlanması ile oluşturulmuştur.

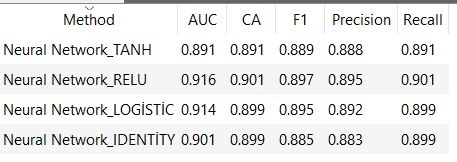
Yapay sinir ağlarında öğrenmek, nöronlar arasındaki ağırlık vektörünün değerini en aza indirgenmesi ile sağlanır. Ağın öğrenmesi Pavlov’un köpekler üzerine yaptığı deneyi ile açıklanabilir: Köpekler Pavlov’un onlara yiyecek göstermesiyle salya akıtırlar. Daha sonra Pavlov köpeklerin kafeslerine bir zil yerleştirir. Zil çaldığında, köpekler salya akıtmaz, çünkü zil ile yiyecek arasında bir bağlantı kuramazlar. Pavlov, köpeklere yiyecek vermeden evvel zili çalarak onları eğitir ve köpekler zil çaldığında yiyeceği görmeseler de salya akıtmaya başlarlar. Eğitilmeden önce salya ile zil arasında bir ilişki yokken, eğitildikten sonra zil ile salya arasında güçlü bir bağ kurulması eğitilen köpeğin nasıl öğrendiğini göstermektedir.

Eğitimin amacı, ağa gösterilen örnekler için doğru çıktıları üretecek ağırlık değerlerini bulmaktır. Ağın doğru aralık değerlerine ulaşması örneklerin temsil ettiği olay hakkında genellemeler yapabilme yeteneğine kavuşması demektir. Ağın bu genelleştirme özelliğine kavuşması işlemine “ağın öğrenmesi” denir.

Ağın öğrenmesi için geri yayılım, ağırlıkları gerçek çıktı ile istenen çıktı arasındaki farkı en aza indirgemek için tekrar tekrar ayarlama prosedürü olarak özetlenebilir.



Grafiğe baktığımızda evli olanların çoğunun kanal aboneliğine hayır dediğini görüyoruz.



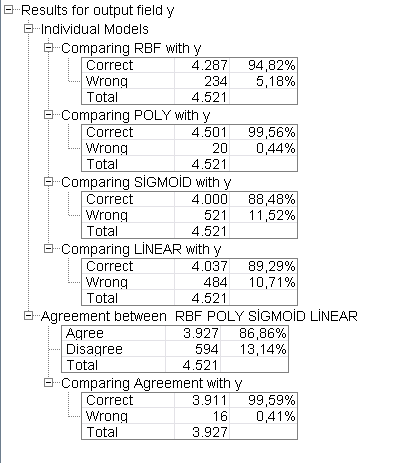
Auc değerlerine baktığımızda Relu algoritmasının diğerlerine göre daha büyük olduğunu görüyoruz. Bu sebeple yapay sinir ağlarında relu algoritmasını kullanmalıyız.

**DESTEK VEKTÖR MAKİNELERİ**

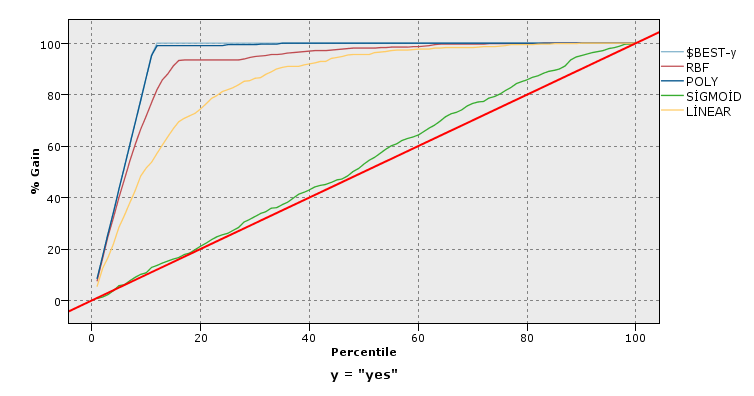
Makine öğrenimi ve veri madenciliği literatüründe, sınıflandırma probleminin çözümüne ilişkin yapılan çalışmalar önemli yer tutmaktadır. Özellikle, bankacılık ve sigortacılık (riskli gruptaki müşterilerin tahmin edilmesi), tıp (hastalık teşhisi), biyoloji (canlı türlerinin sınıflandırılması), kimya (belirli bir hastalık için ilacın etkilerinin belirlenmesi), sosyal medya (spamlerin saptanması), endüstriyel üretim sistemleri (ortaya çıkan kusurlu ürünlerin belirlenmesi) gibi alanlarda sınıflandırma problemleriyle sıkça karşılaşılmaktadır. Dolayısıyla, son yıllarda sınıflandırma problemlerinin çözümü, makine öğreniminin önemli çalışma alanlarından biri olmuştur.

Sınıflandırma problemlerinin çözümü için geliştirilen makine öğrenimi algoritmasının seçiminde dikkat edilecek en önemli kriterlerden biri, algoritmanın genelleme performansıdır. Genelleme performansı, eğitim verisi, bağımsız niteliklerin sayısı/yapısı, model seçimi ve parametre seçimi gibi faktörlere bağlıdır. Tüm bu faktörler göz önünde bulundurulduğunda, veriden hem gizli hem de anlamlı enformasyonun çıkarılması ve doğru bilgiye ulaşma, algoritmanın genelleme başarısıyla doğru orantılıdır. Diğer bir deyişle, algoritmanın genelleme performansı ne kadar iyiyse elde edilen enformasyon da o kadar gerçekçi olacaktır.

Son yıllarda, sınıflandırma problemlerinin çözümü için geliştirilmiş en başarılı makine öğrenimi algoritmalarından biri Destek Vektör Makineleri’dir. Destek Vektör Makineleri, birçok sınıflandırma probleminin çözümünde başarıyla uygulanmış ve genelleme performansı yüksek ve etkin makine öğrenimi algoritmalarından biri olarak literatürdeki yerini almıştır. Destek Vektör Makineleri’nin en önemli avantajı, sınıflandırma problemini kareli optimizasyon problemine dönüştürüp çözmesidir. Böylece problemin çözümüne ilişkin öğrenme aşamasında işlem sayısı azalmakta ve diğer teknik/algoritmalara göre daha hızlı çözüme ulaşılmaktadır. Teknik bu özelliğinden dolayı, özellikle büyük hacimli veri setlerinde büyük avantaj sağlamaktadır. Ayrıca optimizasyon temelli olduğundan sınıflandırma performansı, hesaplama karmaşıklığı ve kullanışlılık açısından diğer tekniklere göre daha başarılıdır.



Destek vektör makineleri arasında poly ve rbf algoritmalarının en yüksek olduğunu görüyoruz. Güvenirlik açısından %100’e bu kadar yakın olması çokta iyi değildir.



Grafikte görüldüğü gibi en iyi çizgiye en yakın olan algoritmalar poly ve rbf’dir.

**KÜMELEME**

Kümeleme Analizi, bir veri matrisinde yer alan ve sade gruplamaları kesin olarak bilinmeyen birim ve değişkenleri birbiri ile benzer olan alt kümelere ayırmaya yardımcı yöntemlerdir.

Kümeleme Analizin de nesneler küme içerisinde çok benzer biçimde fakat kümeler arasında bir o kadar farklı olacak şekilde kümeler. Kümeleme işleminin başarılı olması demek, kümenin geometrik çizimi yapıldığında, nesnelerin küme içerisinde birbirine çok yakın, kümelerin ise birbirinden uzak olmasıdır.

Genel olarak iki tip kümeleme vardır. Bunlar geleneksel ve kavramsal kümelemedir.

Geleneksel kümeleme; nesnelerin geometrik yapılarını baz alarak yaptığımız kümeleme işlemleridir.

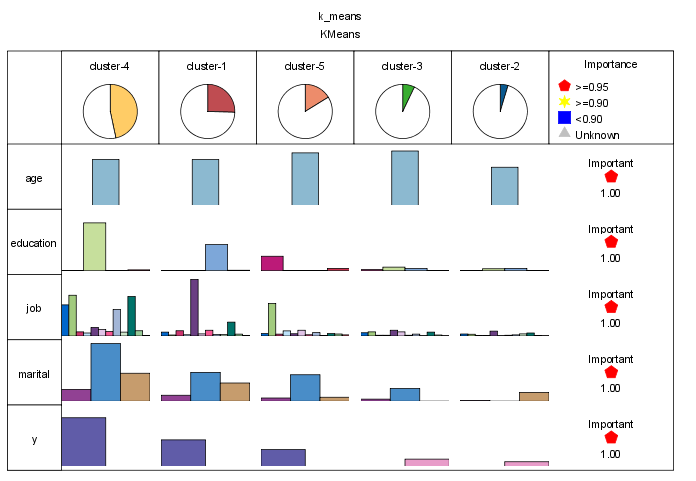
Kavramsal kümeleme ; nesneleri farkılıklarıne ve sınıflandırmada olduğu gibi nesnelerin açıklamalarına göre kümelendiği kümeleme şeklidir.

Kümeleme analizi veriyi anlamlı, yararlı yada hem anlamlı hem de yararlı gruplara(kümelere) ayırır. Eğer amaç anlamlı gruplar ise, bu durumda kümeler verinin doğal yapısını yakalamalıdır.

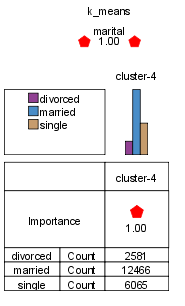
3 farklı kümeleme yöntemi vardır.

1. K ortalama kümelemesi

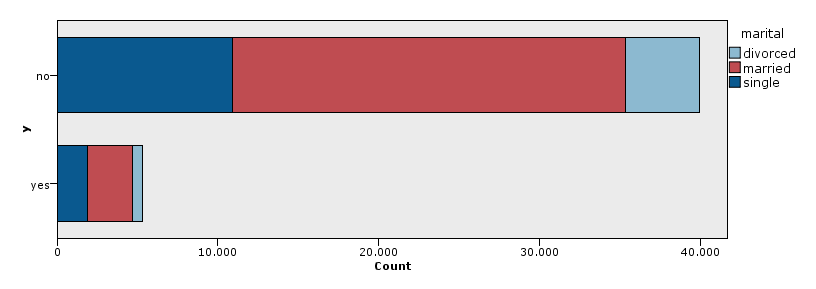
K-Means Clustering, eldeki verileri özelliklerine göre K sayıda kümeye gruplama işlemidir. Gruplama, ilgili kümenin centroid (merkez) değeri ile veri setindeki her nesnenin arasındaki farkın kareleri toplamının minimumu alınarak gerçekleştirilir. K-Ortalamalar Kümelemesin de amaç, gerçekleştirilen bölümleme işlemi sonunda elde edilen kümelerin, küme içi benzerliklerinin maksimum ve kümeler arası benzerliklerinin minimum olmasını sağlamaktır.



Eğitim değişkeni için 4.kümede 2.eğitimin en yüksek olduğunu söyleyebiliriz.



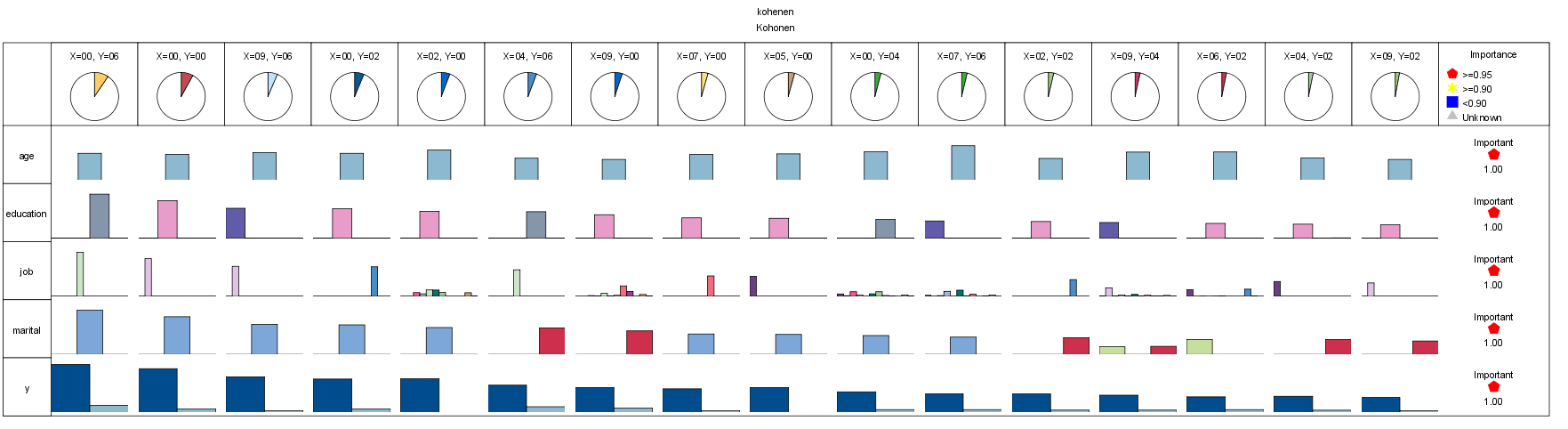
Medeni duruma göre 4.kümede evli olanların yoğun olduğunu gözlemliyoruz.

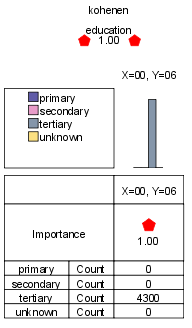


Medeni durumuna göre aboneliğe hayır diyenlerin çoğunun evli olduğunu söyleyebiliriz.

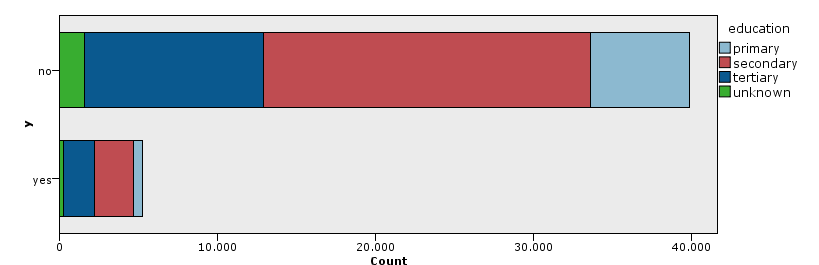
1. Kohenen Kümeleme

Kendini düzenleyen haritalar 1980’ lerde gelitirilen en önemli ağ yapılarından birisidir. Orijinal topolojik ilişkileri koruyarak daha düşük boyutlu (genellikle bir veya iki) çizimle haritalaştırmasıdır.





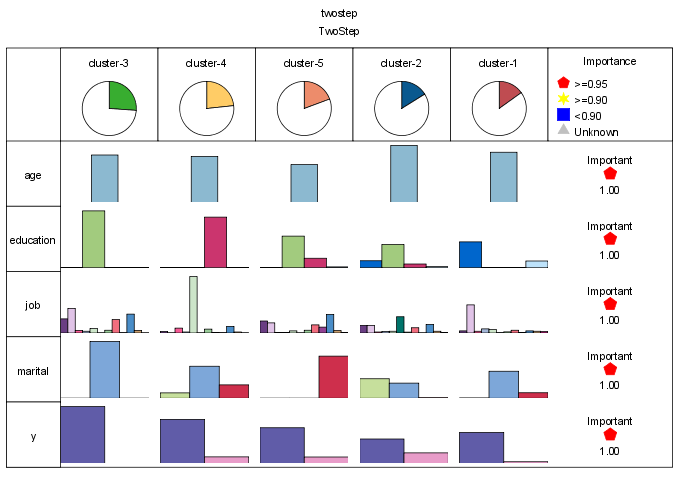
Eğitim durumuna göre x=00,y=06 noktasında yüksek öğrenimin tek olduğunu söyleyebiliriz.

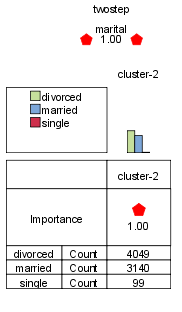


Eğitim durumuna göre aboneliğe hayır diyenlerin çoğunun 2.öğrenim olduğunu söyleyebiliriz.

1. İki aşamalı kümeleme

Bu algoritma, Ward’ın «minimum varyans» yöntemi ile “K-means” yönteminden oluşan bir hibrid yaklaşımdır. Böyle bir karma yaklaşımın avantajı, Ward’ın minimum varyans yönteminin, “K-means” yönteminin gerektirdiği küme sayısını hesaplamasından ileri gelmektedir.





Medeni duruma göre 2.kümede evli olanların sayısı en fazlayken, bekar olanların sayısı en azdır.

KAYNAKÇA

Ders Sunumları

Ders Makaleleri