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;

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
1	58	management	married	tertiary	no	2143	yes	no	unkno...	5	may	261	1	-1	0	unknown	no
2	44	technician	single	secondary	no	29	yes	no	unkno...	5	may	151	1	-1	0	unknown	no
3	33	entrepreneur	married	secondary	no	2	yes	yes	unkno...	5	may	76	1	-1	0	unknown	no
4	47	blue-collar	married	unknown	no	1506	yes	no	unkno...	5	may	92	1	-1	0	unknown	no
5	33	unknown	single	unknown	no	1	no	no	unkno...	5	may	198	1	-1	0	unknown	no
6	35	management	married	tertiary	no	231	yes	no	unkno...	5	may	139	1	-1	0	unknown	no
7	28	management	single	tertiary	no	447	yes	yes	unkno...	5	may	217	1	-1	0	unknown	no
8	42	entrepreneur	divorced	tertiary	yes	2	yes	no	unkno...	5	may	380	1	-1	0	unknown	no
9	58	retired	married	primary	no	121	yes	no	unkno...	5	may	50	1	-1	0	unknown	no
10	43	technician	single	secondary	no	593	yes	no	unkno...	5	may	55	1	-1	0	unknown	no
11	41	admin.	divorced	secondary	no	270	yes	no	unkno...	5	may	222	1	-1	0	unknown	no
12	29	admin.	single	secondary	no	390	yes	no	unkno...	5	may	137	1	-1	0	unknown	no
13	53	technician	married	secondary	no	6	yes	no	unkno...	5	may	517	1	-1	0	unknown	no
14	58	technician	married	unknown	no	71	yes	no	unkno...	5	may	71	1	-1	0	unknown	no
15	57	services	married	secondary	no	162	yes	no	unkno...	5	may	174	1	-1	0	unknown	no
16	51	retired	married	primary	no	229	yes	no	unkno...	5	may	353	1	-1	0	unknown	no
17	45	admin.	single	unknown	no	13	yes	no	unkno...	5	may	98	1	-1	0	unknown	no
18	57	blue-collar	married	primary	no	52	yes	no	unkno...	5	may	38	1	-1	0	unknown	no
19	60	retired	married	primary	no	60	yes	no	unkno...	5	may	219	1	-1	0	unknown	no
20	33	services	married	secondary	no	0	yes	no	unkno...	5	may	54	1	-1	0	unknown	no
21	28	blue-collar	married	secondary	no	723	yes	yes	unkno...	5	may	262	1	-1	0	unknown	no
22	56	management	married	tertiary	no	779	yes	no	unkno...	5	may	164	1	-1	0	unknown	no
23	32	blue-collar	single	primary	no	23	yes	yes	unkno...	5	may	160	1	-1	0	unknown	no
24	25	services	married	secondary	no	50	yes	no	unkno...	5	may	342	1	-1	0	unknown	no
25	40	retired	married	primary	no	0	yes	yes	unkno...	5	may	181	1	-1	0	unknown	no
26	44	admin.	married	secondary	no	-372	yes	no	unkno...	5	may	172	1	-1	0	unknown	no
27	39	management	single	tertiary	no	255	yes	no	unkno...	5	may	296	1	-1	0	unknown	no
28	52	entrepreneur	married	secondary	no	113	yes	yes	unkno...	5	may	127	1	-1	0	unknown	no
29	46	management	single	secondary	no	-246	yes	no	unkno...	5	may	may	2	-1	0	unknown	no
30	36	technician	single	secondary	no	265	yes	yes	unkno...	5	may	348	1	-1	0	unknown	no
31	57	technician	married	secondary	no	839	no	yes	unkno...	5	may	225	1	-1	0	unknown	no
32	49	management	married	tertiary	no	378	yes	no	unkno...	5	may	230	1	-1	0	unknown	no
33	60	admin.	married	secondary	no	39	yes	yes	unkno...	5	may	208	1	-1	0	unknown	no
34	59	blue-collar	married	secondary	no	0	yes	no	unkno...	5	may	226	1	-1	0	unknown	no
35	51	management	married	tertiary	no	10635	yes	no	unkno...	5	may	336	1	-1	0	unknown	no
36	57	technician	divorced	secondary	no	63	yes	no	unkno...	5	may	242	1	-1	0	unknown	no
37	25	blue-collar	married	secondary	no	-7	yes	no	unkno...	5	may	365	1	-1	0	unknown	no

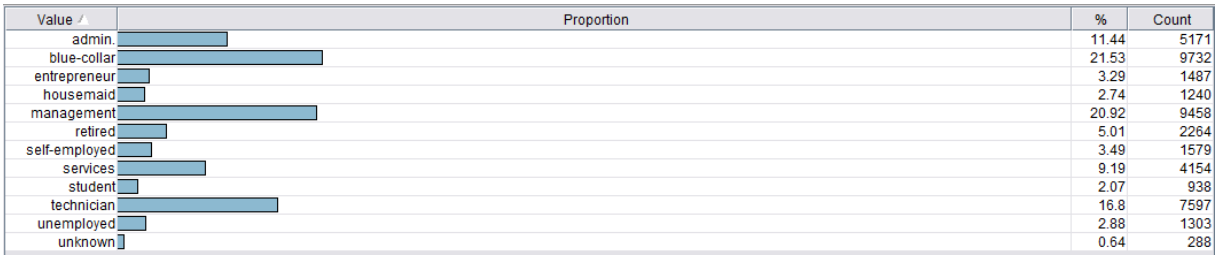
Bağımsız Değişkenler	**Açıklama**	**Düzeyleri**
Age	Yaş	Sürekli
Job	Meslek	Kategorik (yönetici, bilinmeyen, mavi yakalı, işsiz, yönetim, hizmetçi, girişimci, öğrenci, serbest meslek, emekli, teknisyen, hizmetler)
Marital	Medeni Durumu	Kategorik (evli, bekar, boşanmış)
Education	Eğitim Durumu	Kategorik (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

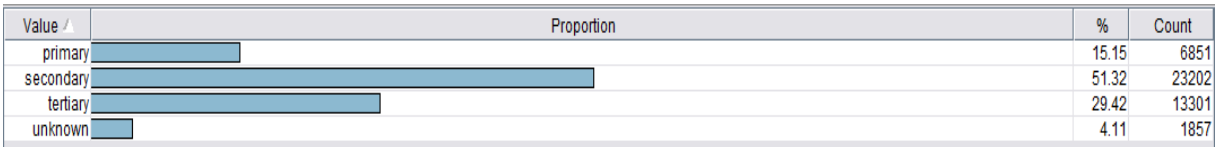
1) MÜŞTERİLERİN MESLEK DAĞILIMI



Value	Proportion	%	Count
admin.		11.44	5171
blue-collar		21.53	9732
entrepreneur		3.29	1487
housemaid		2.74	1240
management		20.92	9458
retired		5.01	2264
self-employed		3.49	1579
services		9.19	4154
student		2.07	938
technician		16.8	7597
unemployed		2.88	1303
unknown		0.64	288

45212 müşteri arasından 13 meslek türünün dağılımları gösterilmiştir.

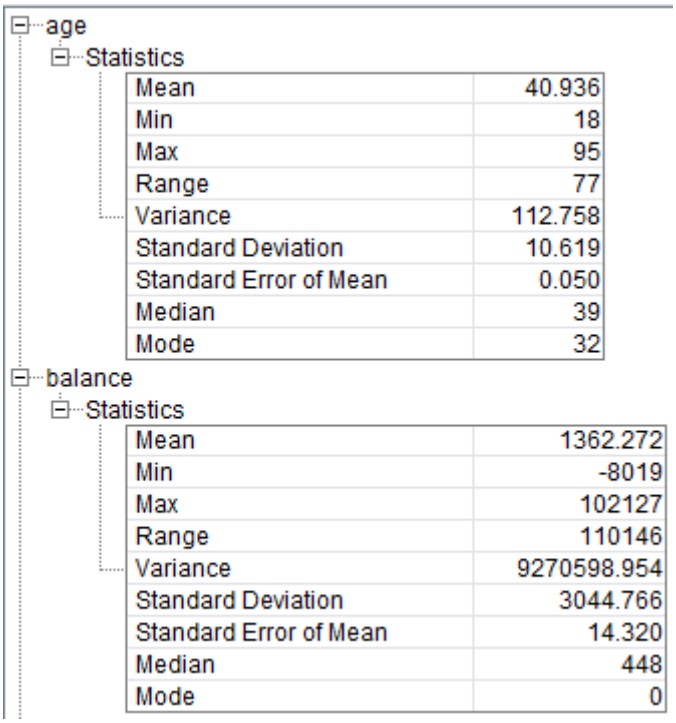
2) MÜŞTERİ EĞİTİM DURUMU DAĞILIMI



Value	Proportion	%	Count
primary		15.15	6851
secondary		51.32	23202
tertiary		29.42	13301
unknown		4.11	1857

45212 müşteri arasında 4 eğitim türünün dağılımları gösterilmiştir.

ÖZETLEYİCİ İSTATİSTİKLER



age

Statistics

Mean	40.936
Min	18
Max	95
Range	77
Variance	112.758
Standard Deviation	10.619
Standard Error of Mean	0.050
Median	39
Mode	32

balance

Statistics

Mean	1362.272
Min	-8019
Max	102127
Range	110146
Variance	9270598.954
Standard Deviation	3044.766
Standard Error of Mean	14.320
Median	448
Mode	0

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.

Yıllık bakiye için özetleyici istatistikler;

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

day

Statistics

Mean	15.806
Min	1
Max	31
Range	30
Variance	69.264
Standard Deviation	8.322
Standard Error of Mean	0.039
Median	16
Mode	20

Gün için özetleyici istatistikler;

duration

Statistics

Mean	258.163
Min	0
Max	4918
Range	4918
Variance	66320.574
Standard Deviation	257.528
Standard Error of Mean	1.211
Median	180
Mode	124

Süre için özetleyici istatistikler;

Son iletişim süresi ortalama 258.163 saniyedir.

campaign

Statistics

Mean	2.764
Min	1
Max	63
Range	62
Variance	9.598
Standard Deviation	3.098
Standard Error of Mean	0.015
Median	2
Mode	1

Kampanya için özetleyici istatistikler;

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

pdays

Statistics

Mean	40.198
Min	-1
Max	871
Range	872
Variance	10025.766
Standard Deviation	100.129
Standard Error of Mean	0.471
Median	-1
Mode	-1

Pday için özetleyici istatistikler;

Müşteriyle önceki kampanya için ortalama 40 gün iletişime geçilmemiştir.

previous

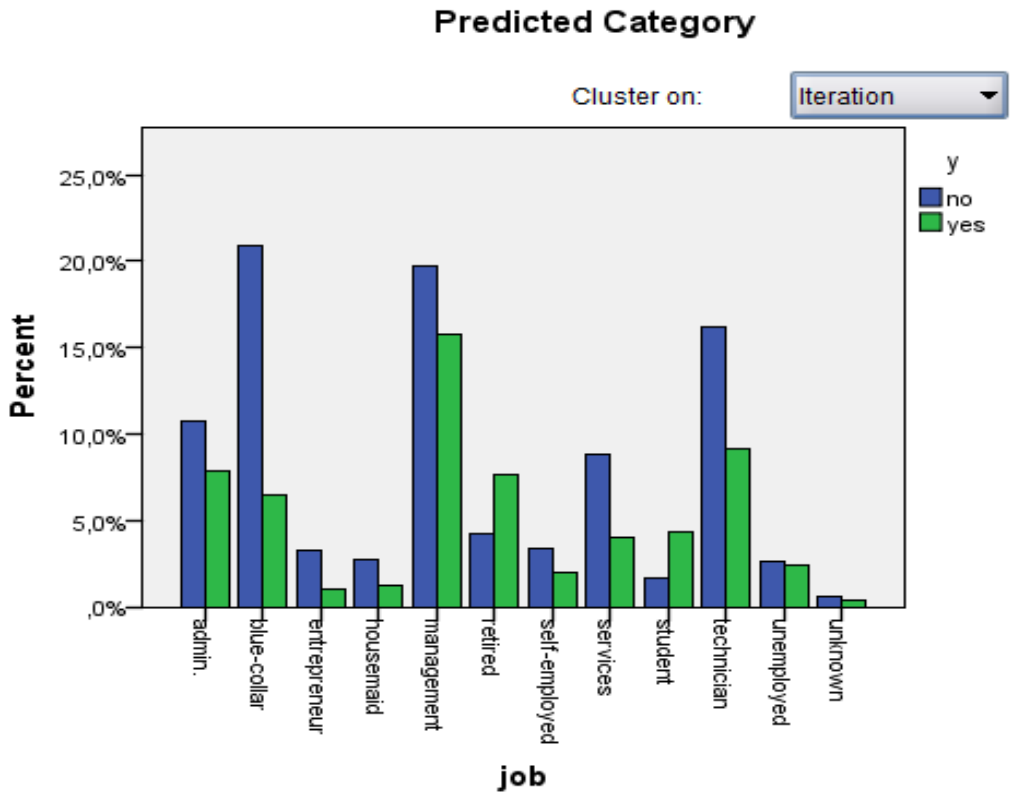
Statistics

Mean	0.580
Min	0
Max	275
Range	275
Variance	5.306
Standard Deviation	2.303
Standard Error of Mean	0.011
Median	0
Mode	0

Previous için özetleyici istatistikler;

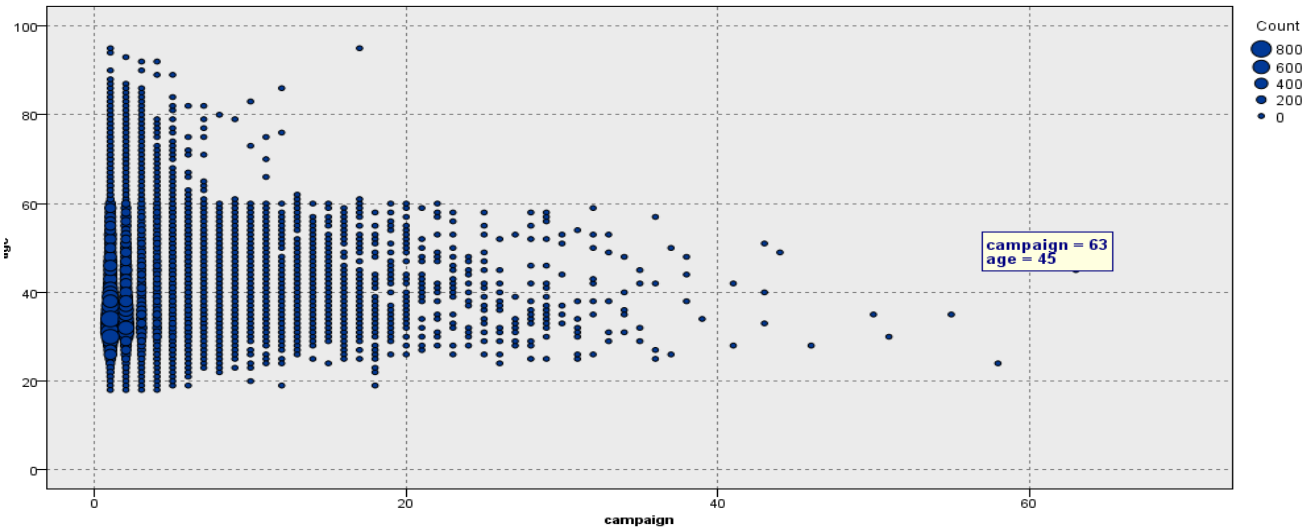
Önceki kampanya için müşteriyle gerçekleştirilen iletişim sayısı ortalama 0.580'dir.

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.

Results for output field y

Individual Models

Comparing CandRT with y

'Partition'	1 Training		2 Testing	
Correct	36.463	89,76%	4.147	90,35%
Wrong	4.158	10,24%	443	9,65%
Total	40.621		4.590	

Comparing QUEST with y

'Partition'	1 Training		2 Testing	
Correct	36.106	88,89%	4.096	89,24%
Wrong	4.515	11,11%	494	10,76%
Total	40.621		4.590	

Comparing CHAID with y

'Partition'	1 Training		2 Testing	
Correct	36.201	89,12%	4.111	89,56%
Wrong	4.420	10,88%	479	10,44%
Total	40.621		4.590	

Comparing C5 with y

'Partition'	1 Training		2 Testing	
Correct	37.893	93,28%	4.293	93,53%
Wrong	2.728	6,72%	297	6,47%
Total	40.621		4.590	

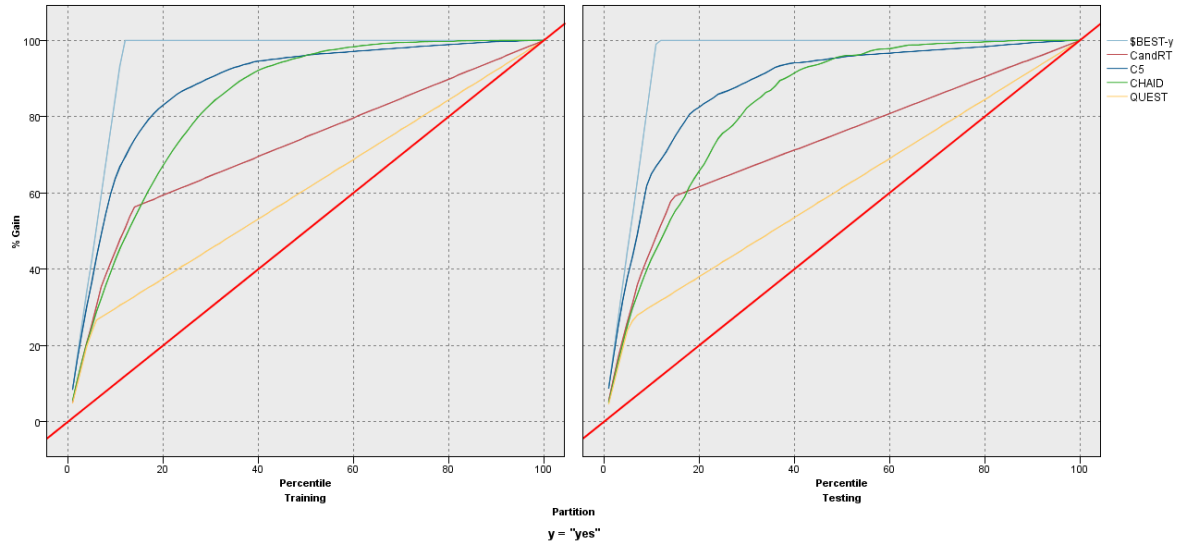
Agreement between CandRT QUEST CHAID C5

'Partition'	1 Training		2 Testing	
Agree	35.465	87,31%	3.985	86,82%
Disagree	5.156	12,69%	605	13,18%
Total	40.621		4.590	

Comparing Agreement with y

'Partition'	1 Training		2 Testing	
Correct	33.933	95,68%	3.835	96,24%
Wrong	1.532	4,32%	150	3,76%
Total	35.465		3.985	

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.

y		no	yes
	Count	0	1
	Row %	0.000	100.000
no	Count	39389	533
	Row %	98.665	1.335
yes	Count	4311	978
	Row %	81.509	18.491

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.

45208	71	retired	divorced	primary	no	1729	no	no	cellula	17	nov	456	2	-1	0	unknown	y...	no	0.901
45209	72	retired	married	secondary	no	5715	no	no	cellula	17	nov	1127	5	184	3	success	y...	yes	0.647
45210	57	blue-collar	married	secondary	no	668	no	no	telepho	17	nov	508	4	-1	0	unknown	no	no	0.901
45211	37	entrepreneur	married	secondary	no	2971	no	no	cellula	17	nov	361	2	188	11	other	no	no	0.901
45212	24	student	married	primary	no	1000	no	no	telepho	16	nov	300	1	30	10	success		yes	0.647

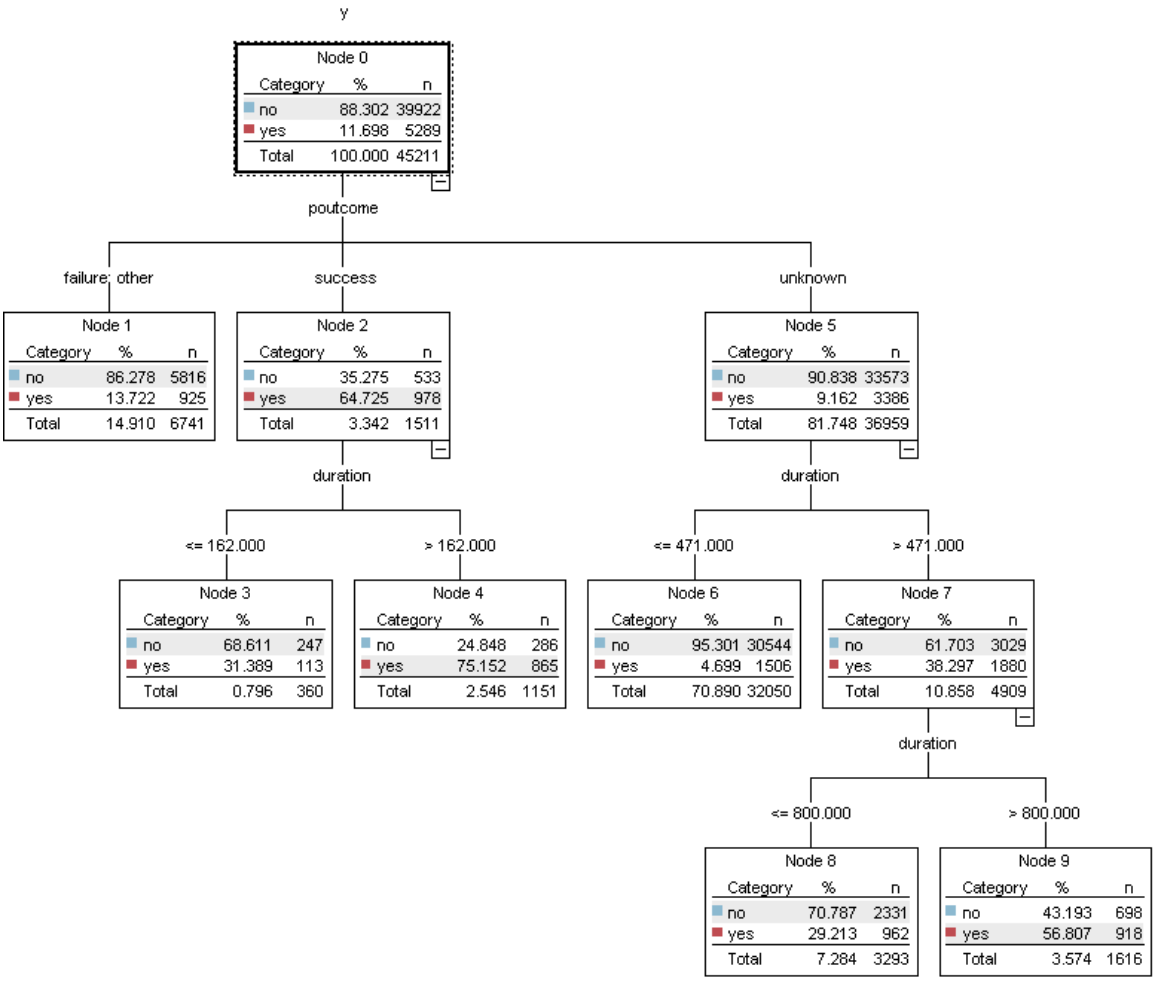
Potansiyel müşterinin girilen verilerine göre vadeli mevduata abone olduğunu gözlemliyoruz.

1.1) 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.

2) 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.

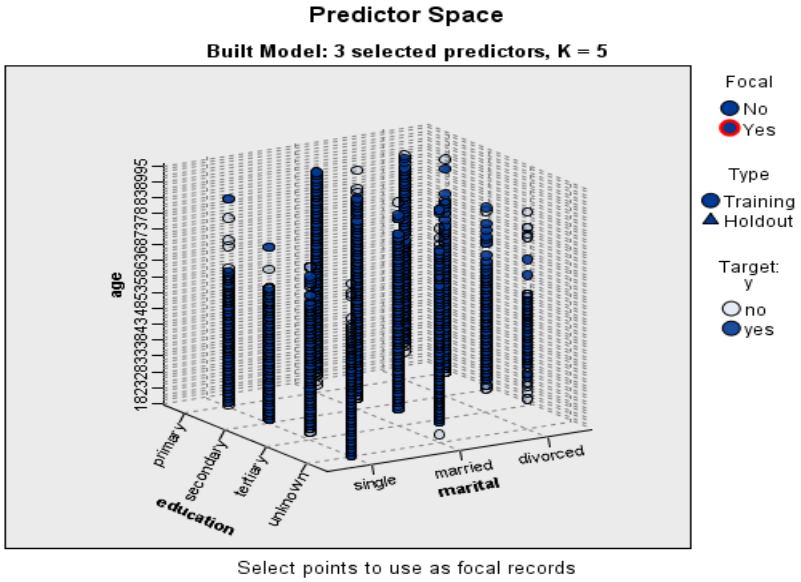
Sınıflandırmada kullanılan bu algoritmaya göre sınıflandırma sırasında çıkarılan özelliklerden, sınıflandırılmak istenen yeni bireyin daha önceki bireylerden k tanesine yakınlığına bakılmasıdır.

Algoritmanın mantığını kısaca şu şekilde özetleyebiliriz;

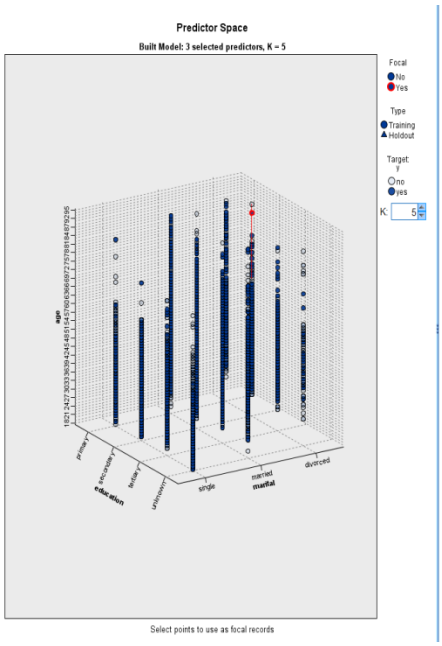
K=5 i ele alırsak, yeni bireyimin hangi sınıfa atanacağına karar verecek iken Öklid uzaklıklarına bakılarak grafikte bireyimize en yakın 5 noktayı ele alırız. Bu noktaya en yakın 5 nokta arasında çoğunluk hangi sınıftaysa yeni bireyimizi de o gruba atarız.

Case Processing Summary

		N	Percent
Sample	Training	31604	69,9%
	Holdout	13605	30,1%
Valid		45209	100,0%
Excluded		2	
Total		45211	



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.

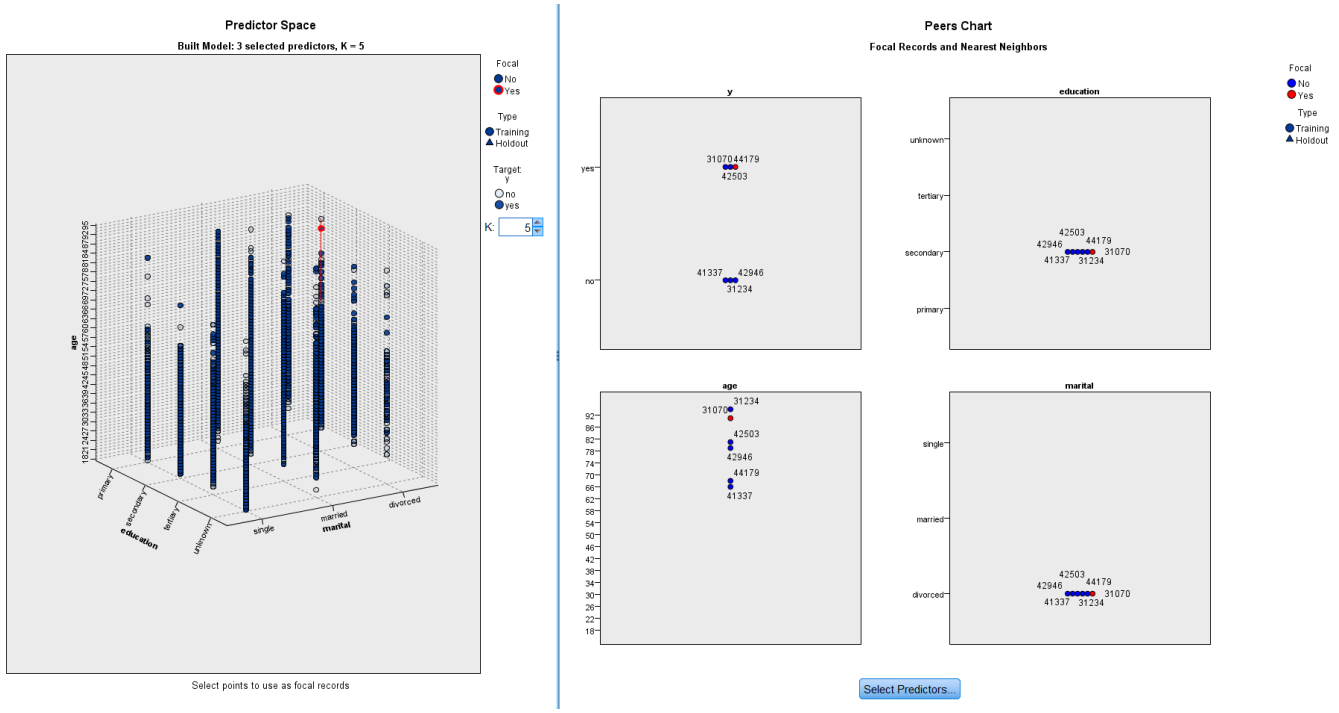


k Nearest Neighbors and Distances

Displayed for Initial Focal Records

Focal Record	Nearest Neighbors					Nearest Distances				
	1	2	3	4	5	1	2	3	4	5
31070	42503	42946	31234	41337	44179	1,414	1,414	1,414	1,414	1,414

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.

	age	job	marital	education	default	balance	housing	loan	cantact	day	month	duration	campaign	pdays	previous	poutcome	y	KNN_PredictedV...	KNN_Probability_1	KNN_Probability_2
1	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no	no	,714	,286
2	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no	no	,857	,143
3	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no	no	,714	,286
4	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no	no	,714	,286
5	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no	no	,571	,429
6	35	management	married	tertiary	no	231	yes	no	unknown	5	may	139	1	-1	0	unknown	no	yes	,286	,714
7	28	management	single	tertiary	no	447	yes	yes	unknown	5	may	217	1	-1	0	unknown	no	no	,857	,143
8	42	entrepreneur	divorced	tertiary	yes	2	yes	no	unknown	5	may	380	1	-1	0	unknown	no	no	,857	,143
9	58	retired	married	primary	no	121	yes	no	unknown	5	may	50	1	-1	0	unknown	no	no	,714	,286
10	43	technician	single	secondary	no	593	yes	no	unknown	5	may	55	1	-1	0	unknown	no	no	,571	,429
11	41	admin.	divorced	secondary	no	270	yes	no	unknown	5	may	222	1	-1	0	unknown	no	no	,857	,143
12	29	admin.	single	secondary	no	390	yes	no	unknown	5	may	137	1	-1	0	unknown	no	no	,857	,143
13	53	technician	married	secondary	no	6	yes	no	unknown	5	may	517	1	-1	0	unknown	no	no	,714	,286
14	58	technician	married	unknown	no	71	yes	no	unknown	5	may	71	1	-1	0	unknown	no	no	,857	,143
15	57	services	married	secondary	no	162	yes	no	unknown	5	may	174	1	-1	0	unknown	no	no	,857	,143
16	51	retired	married	primary	no	229	yes	no	unknown	5	may	353	1	-1	0	unknown	no	no	,857	,143
17	45	admin.	single	unknown	no	13	yes	no	unknown	5	may	98	1	-1	0	unknown	no	no	,714	,286
18	57	blue-collar	married	primary	no	52	yes	no	unknown	5	may	38	1	-1	0	unknown	no	no	,857	,143
19	60	retired	married	primary	no	60	yes	no	unknown	5	may	219	1	-1	0	unknown	no	no	,714	,286
20	33	services	married	secondary	no	0	yes	no	unknown	5	may	54	1	-1	0	unknown	no	no	,714	,286
21	28	blue-collar	married	secondary	no	723	yes	yes	unknown	5	may	262	1	-1	0	unknown	no	no	,714	,286
22	56	management	married	tertiary	no	779	yes	no	unknown	5	may	164	1	-1	0	unknown	no	no	,714	,286
23	32	blue-collar	single	primary	no	23	yes	yes	unknown	5	may	160	1	-1	0	unknown	no	no	,857	,143
24	25	services	married	secondary	no	50	yes	no	unknown	5	may	342	1	-1	0	unknown	no	no	,714	,286
25	40	retired	married	primary	no	0	yes	yes	unknown	5	may	181	1	-1	0	unknown	no	no	,857	,143
26	44	admin.	married	secondary	no	-372	yes	no	unknown	5	may	172	1	-1	0	unknown	no	no	,857	,143
27	39	management	single	tertiary	no	255	yes	no	unknown	5	may	296	1	-1	0	unknown	no	no	,571	,429
28	52	entrepreneur	married	secondary	no	113	yes	yes	unknown	5	may	127	1	-1	0	unknown	no	no	,714	,286
29	46	management	single	secondary	no	-246	yes	no	unknown	5	may	255	2	-1	0	unknown	no	no	,714	,286
30	36	technician	single	secondary	no	265	yes	yes	unknown	5	may	348	1	-1	0	unknown	no	no	,571	,429

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.

Results for output field y

Individual Models

Comparing TAN with y

Correct	40.320	89,18%
Wrong	4.891	10,82%
Total	45.211	

Comparing MARKOV with y

Correct	43.410	96,02%
Wrong	1.801	3,98%
Total	45.211	

Comparing MARKOVFS with y

Correct	43.410	96,02%
Wrong	1.801	3,98%
Total	45.211	

Agreement between TAN MARKOV MARKOVFS

Agree	41.233	91,2%
Disagree	3.978	8,8%
Total	45.211	

Comparing Agreement with y

Correct	39.876	96,71%
Wrong	1.357	3,29%
Total	41.233	

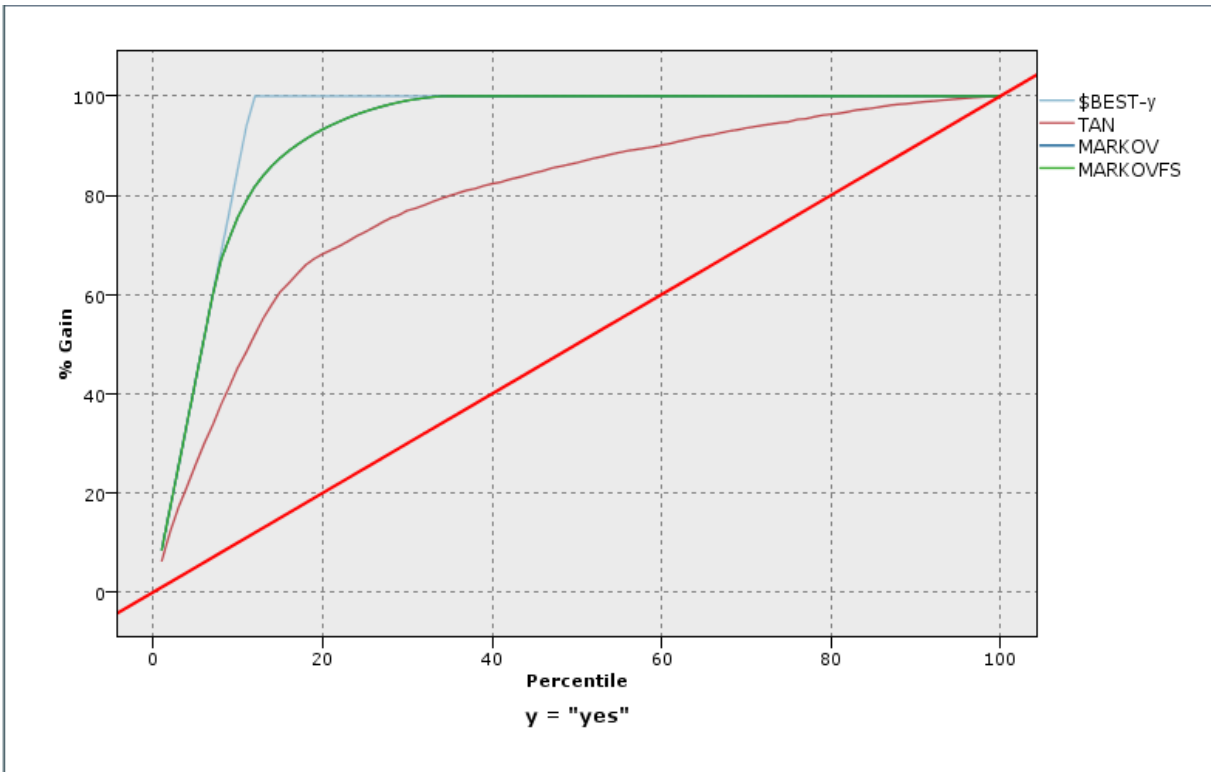
Markov ve markov fs algoritmalarının %96.02 oranla kullanılabileceğini söyleyebiliriz.

	MARKOV	
y	no	yes
no	39384	538
yes	1263	4026

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.

	MARKOVFS	
y	no	yes
no	39384	538
yes	1263	4026

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.

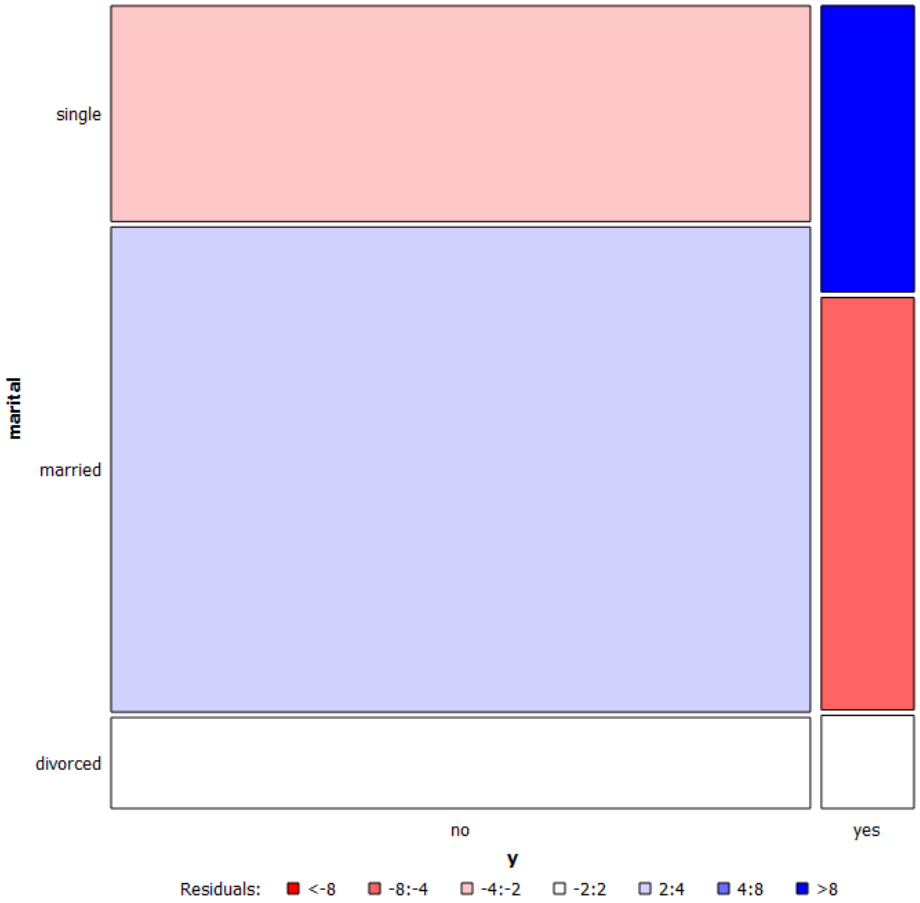
YAPAY SİNİR AĞLARI

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.

Method	AUC	CA	F1	Precision	Recall
Neural Network_TANH	0.891	0.891	0.889	0.888	0.891
Neural Network_RELU	0.916	0.901	0.897	0.895	0.901
Neural Network_LOGİSTİC	0.914	0.899	0.895	0.892	0.899
Neural Network_IDENTİTY	0.901	0.899	0.885	0.883	0.899

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

ROC EĞRİLERİ

ROC eğrileri, testin ayırt etme gücünün belirlenmesi, uygun pozitiflik eşiğinin belirlenmesi, laboratuar sonuçlarının kalitesinin izlenmesi, iki ya da daha fazla teşhis veya laboratuar testlerinin tanı performanslarının karşılaştırılması gibi amaçlarla kullanılmaktadır. Bu yöntem ile aynı zamanda tanı testi ölçütleri de elde edilmektedir. ROC eğrisi yöntemindeki grafiksel yaklaşım verilerin yorumlanmasını kolaylaştırmaktadır. ROC eğrileri ayırt ediciliği göstermekle beraber, farklı testlerin performans açısından karşılaştırılmasında, eğri altında kalan alana (AUC) gereksinim olur. AUC değeri bir tanı testinin doğruluğunu gösteren bir ölçümdür. O halde tanı testine ilişkin performans değerlendirmeleri için eğri altındaki alanın belirlenmesi gerekir.

ROC eğrisi en yüksek doğruluk veren kesim (cut-off) noktasını belirler. Eğri ile duyarlılık ve belirleyicilik arasında optimal bir ilişki ile cut off değerinin saptanmasını sağlar.

Doğruluk (accuracy), ölçülen veya hesaplanan bir değerin gerçek değere veya altın standart (gold standard) değerine yakınlığını belirten bir kavramdır. Bilindiği üzere altın standart klinik çalışmalarda sıkça kullanılan güvenilir bir ölçü kavramı olarak yerleşmiş durumdadır. Örneğin patolojik inceleme materyalleri altın standart olarak kabul edilmektedir. Duyarlılık (precision), ise ayni büyüklüğün tekrarlamalı ölçüm veya hesaplama sonuçlarının birbirine yakınlığını ifade eden bir kavramdır.

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.

Results for output field y

Individual Models

Comparing RBF with y

Correct	4.287	94,82%
Wrong	234	5,18%
Total	4.521	

Comparing POLY with y

Correct	4.501	99,56%
Wrong	20	0,44%
Total	4.521	

Comparing SİGMOİD with y

Correct	4.000	88,48%
Wrong	521	11,52%
Total	4.521	

Comparing LİNEAR with y

Correct	4.037	89,29%
Wrong	484	10,71%
Total	4.521	

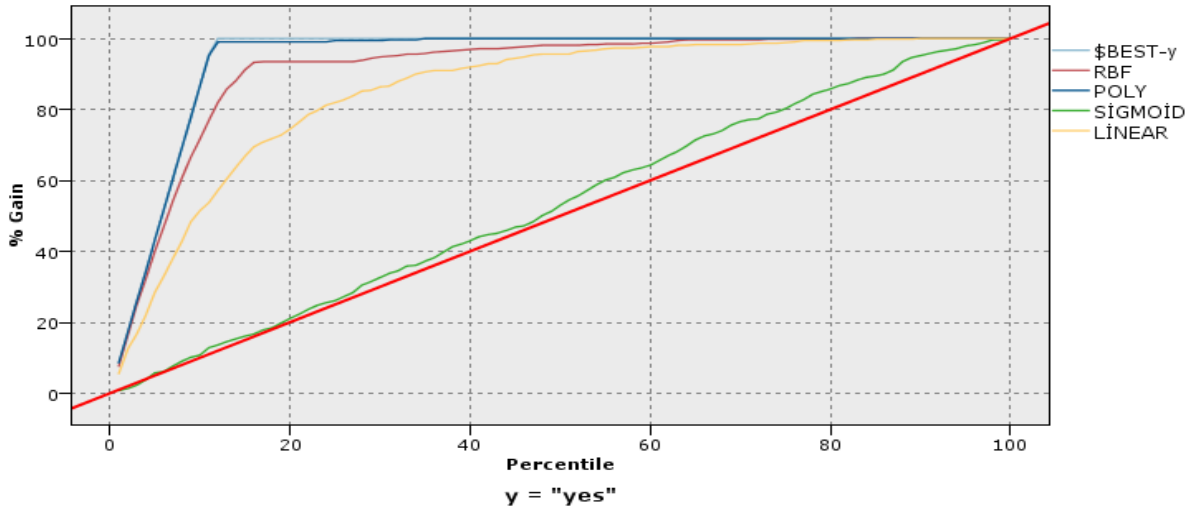
Agreement between RBF POLY SİGMOİD LİNEAR

Agree	3.927	86,86%
Disagree	594	13,14%
Total	4.521	

Comparing Agreement with y

Correct	3.911	99,59%
Wrong	16	0,41%
Total	3.927	

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.

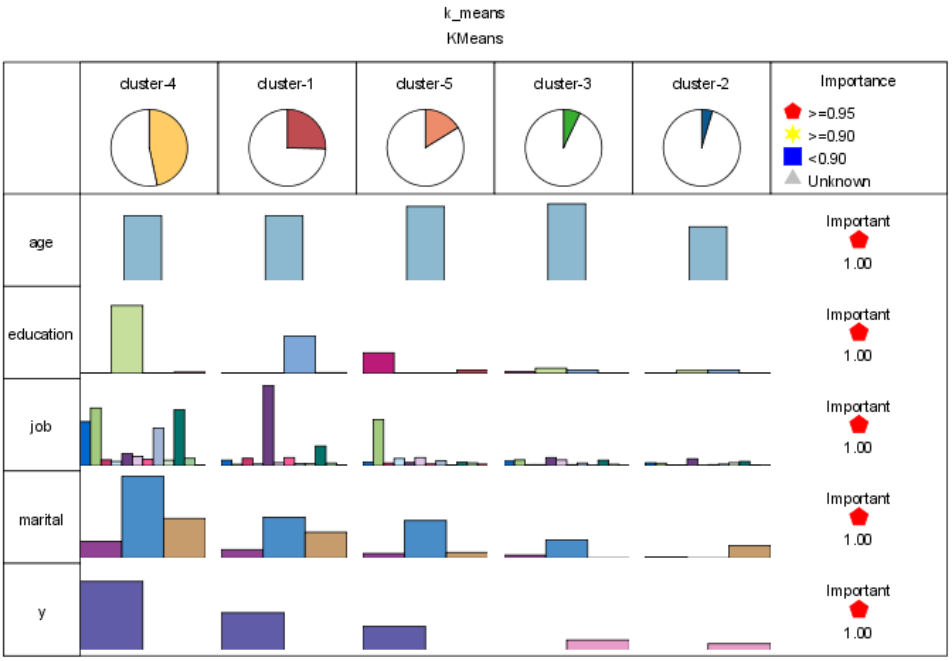
Kavramsal kümeleme ; nesneleri farkılıklarına ve sınıflandırmada olduğu gibi nesnelerin açıklamalarına göre kümelendiği kümeleme şeklidir.

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.

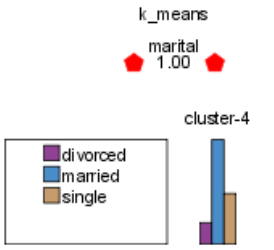
3 farklı kümeleme yöntemi vardır.

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.

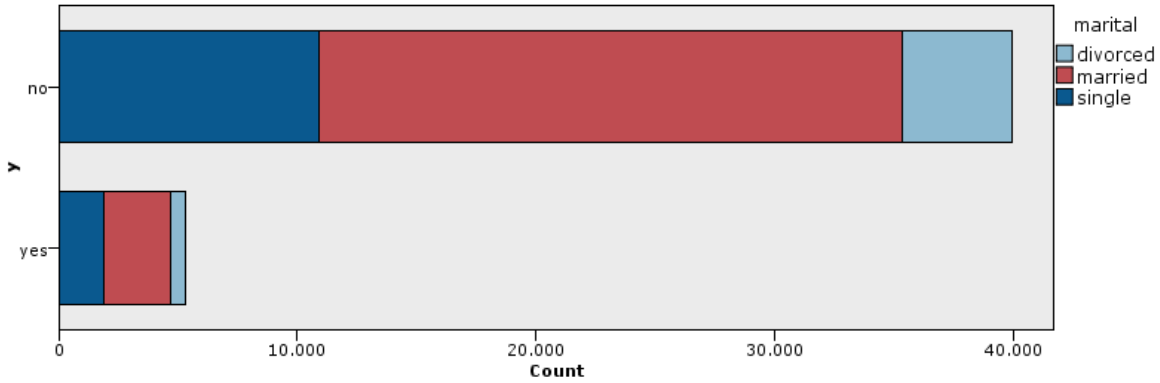


Eğitim değişkeni için 4.kümede 2.eğitimin en yüksek olduğunu söyleyebiliriz.



		cluster-4
Importance		1.00
divorced	Count	2581
married	Count	12466
single	Count	6065

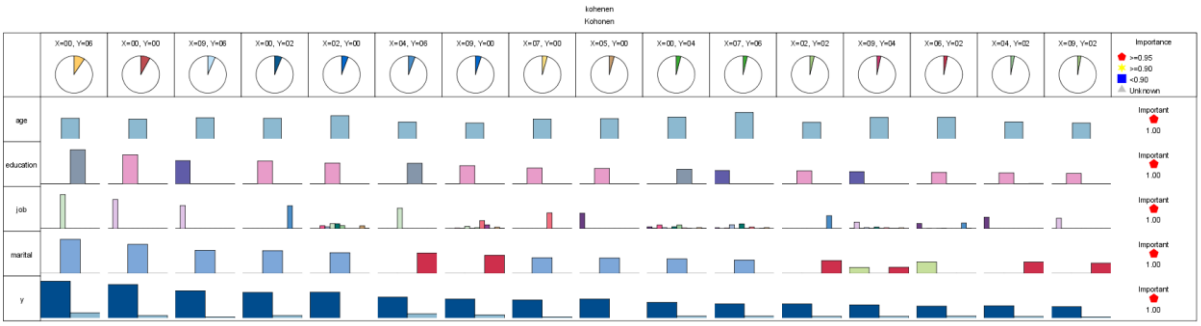
Medeni duruma göre 4.kümede evli olanların yoğun olduğunu gözlemliyoruz.

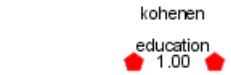


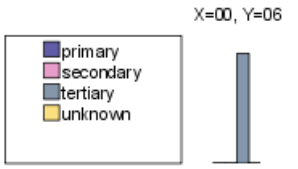
Medeni durumuna göre aboneliğe hayır diyenlerin çoğunun evli olduğunu söyleyebiliriz.

2) 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.

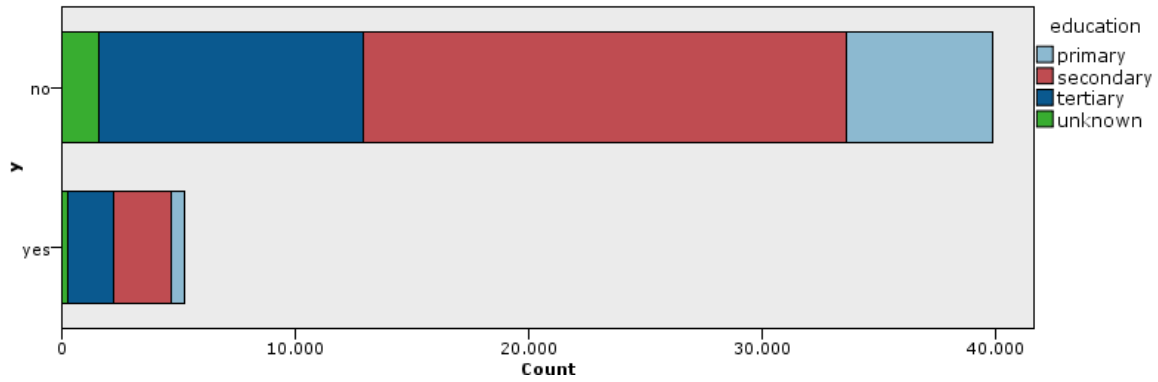






		X=00, Y=06
Importance		1.00
primary	Count	0
secondary	Count	0
tertiary	Count	4300
unknown	Count	0

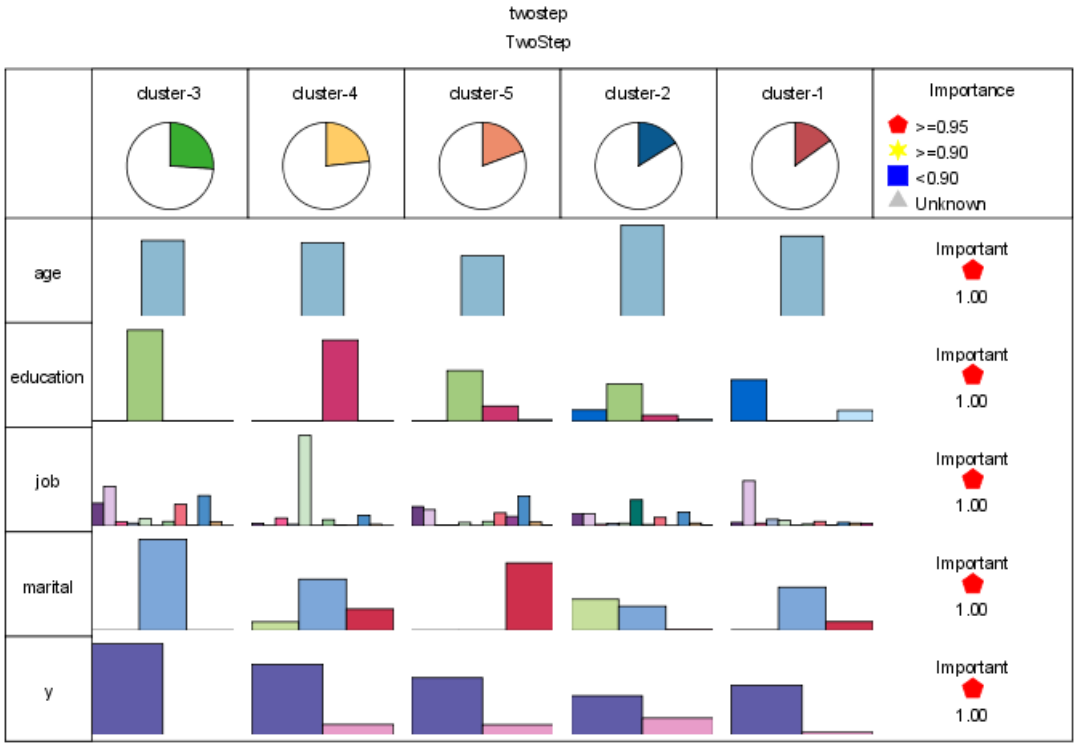
Eğitim durumuna göre x=00,y=06 noktasında yüksek öğrenimin tek olduğunu söyleyebiliriz.



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

3) 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.



		cluster-2
Importance		1.00
divorced	Count	4049
married	Count	3140
single	Count	99

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