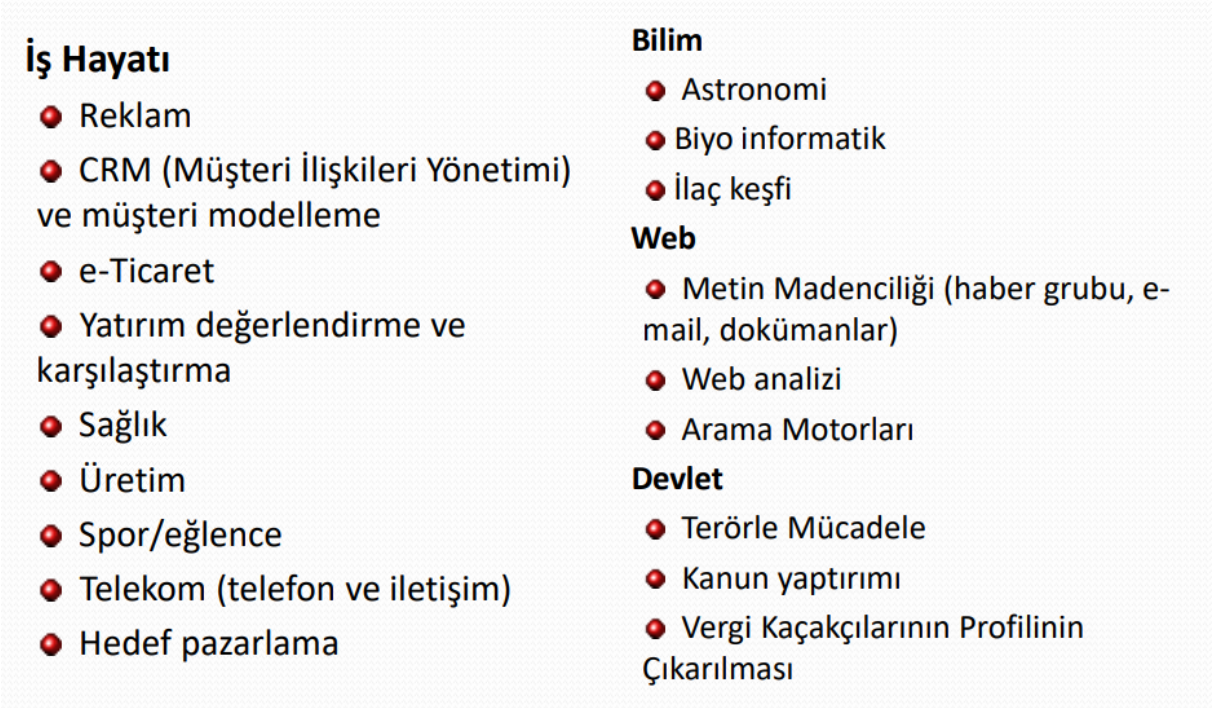
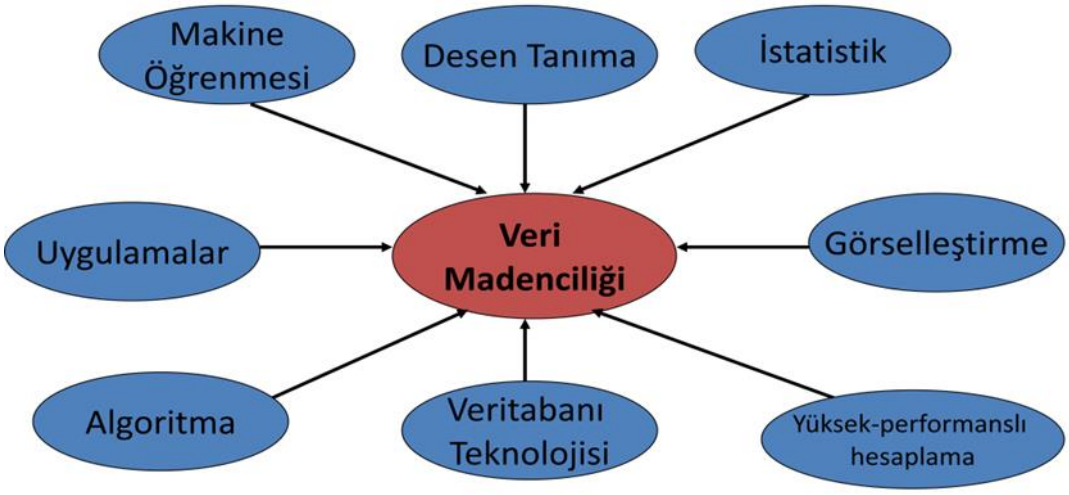
Veri Madenciliği Ders Çalışma

Büyük miktardaki veriler içerisinden önemli olanlarını bulup çıkarmaya Veri Madenciliği denir. Veriler üzerinde çözümlemeler yapmak amacıyla ve veriyi çözümleyip bilgiye ulaşabilmek için veri madenciliği yöntemi ortaya çıkmıştır.

Veri madenciliğinin hedefi düşük düzeyde enformasyon sağlayan veriden, yüksek düzeyde kıymetli bilgiyi açığa çıkarmaktır.





VERİ MADENCİLİĞİ SÜRECİ

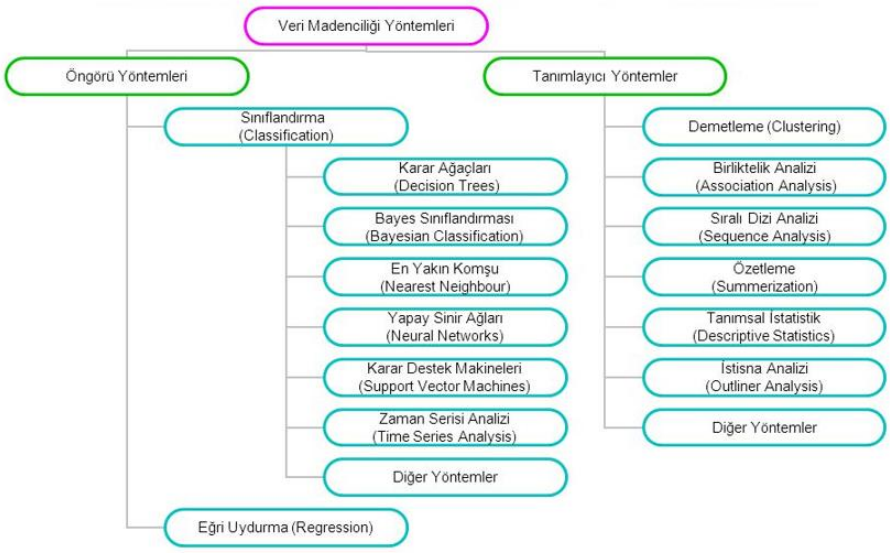
1. Veri temizleme
2. Veri bütünleştirme
3. Veri indirgeme
4. Veri dönüştürme
5. Veri madenciliği algoritmasını uygulama
6. Sonuçları sunum ve değerlendirme

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
30...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

VERİ MADENCİLİĞİ TEKNİKLERİ

- ✓ Sınıflandırma (Classification)
- ✓ Kümeleme (Clustering)
- ✓ Birliktelik Kuralları Keşfi (Assocation Rule Discovery)
- ✓ Sıralı Örüntü Keşfi (Sequential Pattern Discovery)
- ✓ Regresyon (Regression)
- ✓ Sapma Tespiti (Deviation Detection)

VERİ MADENCİLİĞİ YÖNTEMLERİ



Şekil 3: Veri Madenciliği Yöntemleri

1. Veri Madenciliği Teknikleri

1.1. Sınıflandırma

Sınıflandırma algoritmaları, denetimli öğrenmeyi kullanmakta ve sınıf niteliğinin doğru tahminini amaçlamaktadır.

Sınıflandırma, model oluşturma ve model kullanımı olmak üzere iki adımdan oluşmaktadır.

- Model oluşturma aşamasında, her bir örneğin sınıf değişkeni ile belirtilen önceden tanımlanmış bir sınıfa ait olduğu varsayılmaktadır.
- Model kullanımı adımında ise test verisinde yer alan örneklerin bilinen sınıf değeri ile model sonucunda elde edilen sınıf değeri karşılaştırılarak, oluşturulan modelin doğruluğu tahmin edilmektedir.

Model kullanımı adımı sonrasında sınıflandırıcının performansı aşağıdaki özellikler dikkate alınarak değerlendirilmelidir:

- Doğru sınıflandırma başarısı
- Hız
- Kararlı olması
- Ölçeklenebilirlik
- Anlaşılabilir olması
- Kuralların yapısı

1.2. KÜMELEME

Kümeleme algoritmaları, denetimsiz öğrenme algoritmaları olup nesnelerin (örneklerin) belirlenen niteliklere göre gruplara ayrılması amaçlanmaktadır. Kümeleme analizi veri madenciliği projelerinde tek başına kullanılabildiği gibi, diğer algoritmalar için bir önişleme adımı olarak da kullanılabilir.

1.3. Birliktelik Kuralları

Olayların birlikte gerçekleşme durumlarını çözümleyen veri madenciliği yöntemlerine **Birliktelik Kuralları** denir. Örneğin: Marketlerde raf dizaynı yapılırken kullanılır.

1.4. Sıralı Örüntü Keşfi

Zamana bağlı olarak birlikte gelişen olayların tespit edilmesidir. Birliktelik kurallarının çıkarılması tekniğine benzer onun özel bir uygulamasıdır.

1.5. Regresyon

Diğer değişkenlerin değerlerine dayalı olarak bir sürekli değerli hedef değişkenin tahmin edilmesidir. Hedef değişken ile giriş değişkenler arasında doğrusal veya doğrusal olmayan bir ilişki varsayılabilir.

1.6. Sapma tespiti

Sıra dışı anormalliklerin tespit edilmesidir.

2. Veri Ön İşleme (Data preprocessing)

Veri ön işleme; veri madenciliği modelleri kurulmadan önce veri seti üzerinde yapılan bir takım düzeltme, eksik veriyi tamamlama, tekrarlanan verileri kaldırma, dönüştürme, bütünleştirme, temizleme, normalleştirme, boyut indirgeme vb. işlemlerdir.

Birçok veri madenciliği algoritması girdi olarak temiz, kaliteli ve gürültüsüz veri aldığı varsayımına dayanarak çalışır. Oysaki gerçekte böyle bir veri yoktur. Yani oluştuğu esnada mükemmel olan ve hiç ön işleme gerektirmeyen anlamında yoktur.

1. Veri Temizleme
2. Veri Bütünleştirme
3. Veri indirgeme
4. Veri Dönüştürme

2.1. Veri Temizleme

2.1.1. Kayıp Verileri Tamamlama

Kayıp veri, veri madencilerinin kaçışı olmayan kabusu gibidir. Hemen hemen her veri setinde karşılarına çıkar. Çünkü veri tabiatı itibariyle kötüdür (dirty in nature). Kayıp veriyle uğraşırken dikkatli olmak lazım, yanlış bir hareket felaketle sonuçlanabilir. Eksik verilerin bulunduğu satırları çıkarmak bir yöntem olmakla beraber bazı mahsurları vardır. Veriyi bozabilir, değerli verilerin kaybolmasına sebep olabilir, hele kayıp veriler özellikle bir örüntüye sahip ise ciddi sapmalar (bias) oluşturabilir. Kayıp verileri tamamlamak için istatistiksel yöntemleri veya makine öğrenmesi yöntemleri kullanılması çıkarmaktan daha sağlıklıdır. Ancak yazarlar burada kestirip atmış, bence daha uzun durabilirlermiş. Herhangi yenilikçi (novel) bir önerme bulamadım ben burada zaten bilinen şeyler bunlar. Neyse devam edelim okumaya.

2.1.2. Gürültülü Veriyle Uğraşma

Veri doğası icabı kötüdür dedik. Bazı veri madenciliği teknikleri verinin dağılımı konusunda varsayımları vardır. Örneğin regresyon normal dağılım ister. Aksi halde tip-1 hata olasılığı artar. Gürültülü veriyle uğraşma konusunda iki ana yaklaşım vardır. İlki bozuk veriyi düzeltme yöntemleri (data polishing methods). İkinci yaklaşım ise gürültülü veriyi filtrelemek ve eğitim verisi olarak kullanmamak.

2.2. Veri İndirgeme

Bağımsız değişken sayısı çok fazla olduğu durumlarda bağımlı değişkene olan etkiler çok zayıflar ve kurulan modellerin yorumlanabilirliği ve gerçek hayata uygulanabilirliği azalır. Bağımsız değişkenin çokluğuna genelde curse of dimentionality deniyor. Yani Türkçesi çok boyutluluğun laneti. Çok boyutluluk ayrıca hesaplama konusunda da ilave yük getiriyor.

2.2.1. Özellik Seçimi (Feature Selection)

Özellik seçimi, problemi çözmek için gereksiz ve problemin çözümüne etkisi olmayan özellikleri tespit ederek bunları kullanmamaktır. Gereksiz özellikler gereksiz korelasyonlar oluşturur ve modelin genellenebilirliğini zayıflatır. Özellik seçimi ayrıca aşırı öğrenme (overfitting) olasılığını da azaltır, model eğitiminde gereksiz kaynak tüketiminin özellikle ana bellek (bilgisayar rami), önüne geçer. Daha az özellik, daha anlaşılır ve yorumlanır modellerin oluşturulmasını sağlar demiş ve kapatmış konuyu.

2.2.2. Space Transformations

Boyut indirgemenin özellik seçiminden başka yöntemleri de var elbette, örneğin faktör analizi ve ana bileşenler analizi (principal component analysis). Bu ikisi doğrusal yöntemler. Bir de doğrusal olmayanlar var: .LLE ve ISOMAP.

2.2.3. Instance Reduction (IR)

Büyük veri setlerinin veri madenciliği algoritmaları üzerindeki olumsuz etkisini azaltmanın popüler yöntemlerinden birisi IR. Veri boyutunu küçült ama ondan çıkarılacak bilgi kalitesini düşürme felsefesine dayanır.

2.2.4. Instance Selection (IS)

Örneklem seçmek gibi birşey. Klasik usülde evrendeki nesnelerin hepsine ulaşıp veri toplayamadığımız için evreni temsil edebilecek bir örneklem seçiyorduk. Ancak burada evrendeki tüm nesnelere ait veri zaten elimizde. Fazla mal göz çıkardığından hepsini değil de hepsini temsil edecek bir örneklem seçiliyor. Yalnız buradaki fark olay tamamen tesadüfi gelişmiyor, temizleme işlemleri de yapılıyor ve algoritmanın verinin önemli kısımlarına odaklanması sağlanıyor.

2.2.5. Instance Generation (IG)

Instance generation bir bakıma instance selection tersi gibi. Burada da yapay bir veri üretimi var. Ama nerede ve niçin? Eksik yerler, bir alanda temsilcinin olmadığı veya yetersiz olduğu yerler. Yanlış etiketlenmiş verilerin düzeltilmesi bir örnek olarak verilebilir.

2.2.6. Discretization

Veriyi kesikli hale getirme olayı. Örneğin karar ağaçları sürekli değişken kullanmaz. Bu sebeple sürekli değişkenler karar ağacı için kesikli hale getirilir. En çok kullanılan veri ön işleme tekniği. Örneğin yaş değişkeninin çocuk, ergen, genç, orta yaş, yaşlı yapılması. Karar ağaçları gibi bir çok algoritma kesikli değişken istiyor mesela C4.5, Naive Bayes, Apriori. Kesikleştirme veriyi basitleştirme, daha anlaşılır kılma, hızlı ve yüksek doğrulukla öğrenmeyi gibi faydaları var üstelik verinin okunurluğunu artırıyor. Ancak bazı maliyetler var: bilgi kaybı.

3. Eşik Değer

IQR = Q3 – Q1

Eşik Değer(ED) = IQRx1,5

Örnek: 1,2,3,4,5,6,7,8,9,10,11, 100(Üst sınırın üstünde olduğu için Ayrısı veridir)

- Sayıları Sırala, 6:Medyan, 3:Sol Medyan, 9:Sağ Medyan
- Q1 = 3,
- Q3 = 9
- IQR = Q3 – Q1 = 9-3 = 6
- ED = 6*1.5 = 9
- Alt Sınır = Q1 - ED = 3-9 = -6
- Üst Sınır = Q3 + ED = 9+9= 18
- Baskılama Yöntemi ile 100 sayısını 18 e baskılıyoruz, -1 5 diye bir veri gelirse, -15 sayısını -6 ya Baskılıyoruz.

4. Sınıflandırma

4.1. Karar Ağaçları

Karar ağaçları oluşturmak için temel olarak entropiye dayalı algoritmalar, sınıflandırma ve regresyon ağaçları, bellek tabanlı sınıflandırma modelleri biçiminde birçok yöntem geliştirilmiştir.

Her farklı kriter için bir karar ağacı algoritması karşılık gelmektedir. Algoritmaları gruplandıracak olursak;

- Entropiye Dayalı Algoritmalar
- Sınıflandırma ve Regresyon ağaçları
- Bellek tabanlı sınıflandırma algoritmaları

ID3 ve C4.5 algoritmaları entropi tabanlı algoritmalardır.

4.1.1. Entropi

Bir sistemdeki belirsizliğin ölçüsüne entropi denir. Olasılık dağılımına sahip mesajları üreten S kaynağının entropisi şu şekildedir;

$$P=\{p_1,p_2,p_n\}$$

$$H(S) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Karar ağaçlarının oluşturulması esnasında dallanmaya hangi nitelikten başlanacağı önem taşımaktadır. Çünkü sınırlı sayıda kayıttan oluşan bir eğitim kümesinden yararlanarak olası tüm ağaç yapılarını ortaya çıkarmak ve içlerinden en uygun olanı seçerek ondan başlamak kolay değildir.

4.1.2. ID3 Algoritması - Örnek

HAVA	ISI	NEM	RÜZGAR	OYUN
Güneşli	Sıcak	Yüksek	Hafif	Hayır
Güneşli	Sıcak	Yüksek	Kuvvetli	Hayır
Bulutlu	Sıcak	Yüksek	Hafif	Evet
Yağmurlu	Ilık	Yüksek	Hafif	Evet
Yağmurlu	Soğuk	Normal	Hafif	Evet
Yağmurlu	Soğuk	Normal	Kuvvetli	Hayır
Bulutlu	Soğuk	Normal	Kuvvetli	Evet
Güneşli	Ilık	Yüksek	Hafif	Hayır
Güneşli	Soğuk	Normal	Hafif	Evet
Yağmurlu	Ilık	Normal	Hafif	Evet
Güneşli	Ilık	Normal	Kuvvetli	Evet
Bulutlu	Ilık	Yüksek	Kuvvetli	Evet
Bulutlu	Sıcak	Normal	Hafif	Evet
Yağmurlu	Ilık	Yüksek	Kuvvetli	Hayır

4.1.2.1. Adım 1 – Sonuç Kümesinin Entropisi Hesaplanır.

"OYUN" nitelik değerlerinden oluşan küme T kümesi olarak kabul edilecek(Sonuç Sütunu yani).

Burada $C1$ sınıfı 'hayır', $C2$ sınıfı ise 'evet' değerine uymaktadır. $OYUN$ = 14 Adet satır içerir, 5 adet 'hayır' değeri için $C1$ =5, Ve 9 adet 'evet' değeri için $C2$ =9 olur.

Bu sayede; P_1 = 5/14 ve P_2 = 9/14

Oyun Kümesinin Entropisi = $H(OYUN) = -\left(\frac{5}{14}\log_2\frac{5}{14} + \frac{9}{14}\log_2\frac{9}{14}\right)$=0.940

4.1.2.2. Adım 2– Her bir Kümenin Entropisi Hesaplanır.

- $HAVA_{GÜNEŞLİ}$ =5, $HAVA_{YAĞMURLU}$ =5, $HAVA_{BULUTLU}$ =4
- ENTROPİ: $H(HAVA, OYUN) = \frac{5}{14}H(HAVA_{güneşli}) + \frac{4}{14}H(HAVA_{bulutlu}) + \frac{5}{14}H(HAVA_{yağmurlu})$
- Tek tek özelliklerin Entropisi Sonuç Kümesine göre alınır;
- Mesela Güneşli hava 5 adet ve sadece 2 tanesinde Oyun EVET olmuş. Bu yüzden 3/5 = hayır, 2/5 = evet i temsil eder(Aşağıya göre).

$$H(HAVA_{güneşli}) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) = 0.971$$

$$H(HAVA_{yağmurlu}) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0.971$$

$$H(HAVA_{bulutlu}) = -\left(\frac{4}{4}\log_2\frac{4}{4}\right) = 0$$

- Sonra Formülde Yerine Koyulur;

$$H(HAVA, OYUN) = \frac{5}{14}H(HAVA_{güneşli}) + \frac{4}{14}H(HAVA_{bulutlu}) + \frac{5}{14}H(HAVA_{yağmurlu})$$
$$= \frac{5}{14}(0.971) + \frac{4}{14}(0) + \frac{5}{14}(0.971)$$
$$= 0.694$$

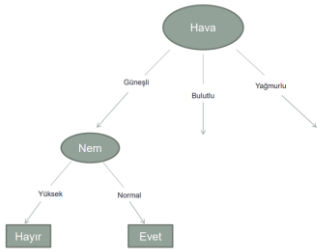
- Sonra Kazanç Ölçütü hesaplanır; Oyun Entropisinden, Hava-Oyun Entropisini çıkartırız.

$$Kazanç(HAVA, OYUN) = H(OYUN) - H(HAVA, OYUN)$$
$$= 0.940\text{-}0.693$$
$$= 0.247$$

- Bu şekilde Her bir küme hesaplanır ve Kazancı en yüksek olan Karar Ağacında yerini alır.

4.1.2.3. Adım 3– Dallanmaya Devam Etmek

- Karar ağacında yerini alan hava kümesinin Niteliklerini Dallara ayırarak yukarıdaki işlemleri tekrar ederiz.
- 

- İlk Güneşliye Göre Bakarsak bu sefer elimizde 5 tane satır olur.
- Önce oyun entropisi sonra diğer entropiler hesaplanır(adım 1 deki ile aynı).
- Kazançlar yeni değerlere göre hesaplanır.
- Ve sonuç güneşliden sonra gelen küme NEM olur
- 

4.1.3. C4.5 Algoritması – Örnek

NİTELİK1	NİTELİK2	NİTELİK3	SINIF
a	70	Doğru	Sınıf1
a	90	Doğru	Sınıf2
a	85	Yanlış	Sınıf2
a	95	Yanlış	Sınıf2
a	70	Yanlış	Sınıf1
b	90	Doğru	Sınıf1
b	78	Yanlış	Sınıf1
b	65	Doğru	Sınıf1
b	75	Yanlış	Sınıf1
c	80	Doğru	Sınıf2
c	70	Doğru	Sınıf2
c	80	Yanlış	Sınıf1
c	70	Yanlış	Sınıf1
c	96	Yanlış	Sınıf1

4.1.3.1. Adım 1– Sayısal Değerleri Metinsel Değerlere Dönüştürme

Algoritma çok basit. Aynı ID3 algoritması gibi ama sayısal değerleri metinsel değerlere dönüştürüyoruz ve ID3 uyguluyoruz.

- NİTELİK2 niteliği 65,70,75,80,85,90,95,96 değerlerine sahiptir.
- Orta Noktasına bakıyoruz 2 tane var çift adet olduğu için o yüzden toplayıp 2 ye bölüyoruz
- Eşik Değeri = (V1 + V2 + 1) /2
- (80 + 85 + 1) / 2 = 83 Eşik Değeri vardır.
- Sonra Nitelik 2 sütununu Eşikten küçük eşit(<=) ve Eşikten Büyük diye 2 farklı metinsel ifadeye dönüştürüyoruz

NİTELİK1	NİTELİK2	NİTELİK3	SINIF
a	Eşit veya küçük	Doğru	Sınıf1
a	Büyük	Doğru	Sınıf2
a	Büyük	Yanlış	Sınıf2
a	Büyük	Yanlış	Sınıf2
a	Eşit veya küçük	Yanlış	Sınıf1
b	Büyük	Doğru	Sınıf1
b	Eşit veya küçük	Yanlış	Sınıf1
b	Eşit veya küçük	Doğru	Sınıf1
b	Eşit veya küçük	Yanlış	Sınıf1
c	Eşit veya küçük	Doğru	Sınıf2
c	Eşit veya küçük	Doğru	Sınıf2
c	Eşit veya küçük	Yanlış	Sınıf1
c	Eşit veya küçük	Yanlış	Sınıf1
c	Büyük	Yanlış	Sınıf1

- Sonrasında ise ID3 Algoritması Çalıştırılır. Entropi vs vs

4.1.3.2. Adım 2– Eğer Eksik veri Var ise

NİTELİK1	NİTELİK2	NİTELİK3	SINIF
a	70	Doğru	Sınıf1
a	90	Doğru	Sınıf2
a	85	Yanlış	Sınıf2
a	95	Yanlış	Sınıf2
a	70	Yanlış	Sınıf1
?	90	Doğru	Sınıf1
b	78	Yanlış	Sınıf1
b	65	Doğru	Sınıf1
b	75	Yanlış	Sınıf1
c	80	Doğru	Sınıf2
c	70	Doğru	Sınıf2
c	80	Yanlış	Sınıf1
c	70	Yanlış	Sınıf1
c	96	Yanlış	Sınıf1

- Eksik değer olan satırı sil.
- Sayısal değerleri metinsele dönüştür
- ID3 devam et

4.2. Twoing Algoritması Örnek

NOT: Bu karar ağacında sadece 2 adet alt çocuk oluşur(2 dallanma oluşur her bir düğümde)

4.2.1.1. Adım 1– Aday Bölünmelerin Bulunması

- Aday bölünmelerin her biri için P_{Sol} , $P_{(j|Sol)}$, $P_{Sağ}$ ve $P_{(j|Sağ)}$ olasılıkları hesaplanır.
- $P_{Sol} = \frac{t_{sol}\text{'daki herbir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı}}{\text{Eğitim kümesindeki kayıtların sayısı}}$ $P_{sağ} = \frac{t_{sağ}\text{'daki herbir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı}}{\text{Eğitim kümesindeki kayıtların sayısı}}$
- $P(j|t_{sol}) = \frac{t_{sol}\text{'daki kayıtların j sayısı}}{t_{sol}\text{'daki herbir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı}}$ $P(j|t_{sağ}) = \frac{t_{sağ}\text{'daki kayıtların j sayısı}}{t_{sağ}\text{'daki herbir nitelik değerinin ilgili nitelik sütunundaki tekrar sayısı}}$
- Sonra Uygunluk Ölçütü hesaplanır
- $$\Phi(s|t) = 2P_{sol}P_{sağ}\sum_{j=1}^{n}|P(j|t_{sol}) - P(j|t_{sağ})|$$
- Φ(s|t) değerleri hesaplandıktan sonra içlerinden en büyük olanı seçilir.

4.2.1.2. Adım 2– ID3 gibi dallanmaya Adım 1 uygulayarak devam et

- Yeni Düğümün dallanmasını adım 1 deki yol ile tekrar et
- Bunu bütün Düğümler için uygula(Sona gelene kadar)

4.2.1.3. Örnek

Müşteri	GELİR	EĞİTİM	SEKTÖR	MEMNUN
1	NORMAL	ORTA	BİLİŞİM	EVET
2	BÜYÜK.	İLK	BİLİŞİM	EVET
3	KÜÇÜK	İLK	İNŞAAT	EVET
4	BÜYÜK	ORTA	İNŞAAT	EVET
5	KÜÇÜK	ORTA	İNŞAAT	EVET
6	BÜYÜK	LİSE	İNŞAAT	EVET
7	KÜÇÜK	LİSE	İNŞAAT	EVET
8	BÜYÜK	ORTA	BİLİŞİM	HAYIR
9	KÜÇÜK	ORTA	BİLİŞİM	HAYIR
10	BÜYÜK	LİSE	BİLİŞİM	HAYIR
1 1	KÜÇÜK	LİSE	BİLİŞİM	HAYIR

Eğitim Kümesi

- ## Adım 1 – Aday Bölünmeleri bul

Aday bölünme (s)	t_{sol}	$t_{sağ}$
1	GELİR=NORMAL	GELİR ∈ {BÜYÜK,KÜÇÜK}
2	GELİR=BÜYÜK	GELİR ∈ {NORMAL,KÜÇÜK}
3	GELİR=KÜÇÜK	GELİR ∈ {BÜYÜK,NORMAL}
4	EĞİTİM=İLK	EĞİTİM ∈ {ORTA,LİSE}
5	EĞİTİ M=ORTA	EĞİTİM ∈ {İLK, LİSE}
6	EĞİTİM=LİSE	EĞİTİ M ∈ {İLK,ORTA}
7	SEKTÖR=BİLİŞİM	SEKTÖR= İNŞAAT
8	SEKTÖR=İNŞAAT	SEKTÖR= BİLİŞİM

Aday bölünmeler/Tablo4

1) GELİR: NORMAL için;
   a. Psol= 1/11= 0.09
   b. P(evet|tsol) = 1/1 = 1
   c. P(hayır|tsol) = 0/1 = 0
2) Gelir: Büyük için;
   a. Psol= 5/11= 0.45
   b. P(evet|tsol) = 3/5 = 0.6
   c. P(hayır|tsol) = 2/5 = 0.4
3) Bu Şekilde TSol için Büyün Aday bölünme satırlarını hesaplıyoruz

Aday Bölünme	t_{sol}'daki kayıt sayısı	P_{sol}	EVET sayısı	HAYIR sayısı	P(EVET|t_{sol})	P(HAYIR|t_{sol})
1	1	0.09	1	0	1.00	0.00
2	5	0.45	3	2	0.60	0.40
3	5	0.45	3	2	0.60	0.40
4	2	0.18	2	0	1.00	0.00
5	5	0.45	3	2	0.60	0.40
6	4	0.36	2	2	0.50	0.50
7	6	0.55	2	4	0.33	0.67
8	5	0.45	5	0	1.00	0.00

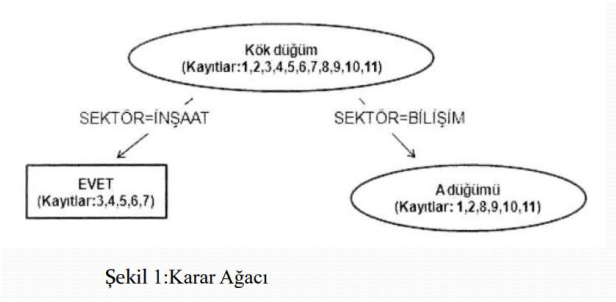
4) Tsağ için hesaplamalar;

Aday Bölünme	$t_{sağ}$'daki kayıt sayısı	$P_{sağ}$	EVET sayısı	HAYIR sayısı	P(EVET|tsag)	P(HAYIR|tsag)
1	10	0.91	6	4	0.60	0.40
2	6	0.55	4	2	0.67	0.33
3	6	0.55	4	2	0.67	0.33
4	9	0.82	5	4	0.56	0.44
5	6	0.55	4	2	0.67	0.33
6	7	0.64	5	2	0.71	0.29
7	5	0.45	5	0	1.00	0.00
8	6	0.55	2	4	0.33	0.67

- ## Adım 2 – Uygunluk Ölçütleri

$$\Phi(s|t)=2P_{sol}P_{sağ}\sum_{j=1}^{n}|P(j|t_{sol})-P(j|t_{sağ})|$$

1) Φ(1|t) için;
   a. Φ(1|t) = 2 * (0.09) * (0.91) * [| (1 - 0.6) | + | (0 - 0.4) |]
   b. Φ(1|t) = 0.13
2) Teker Teker Tüm Aday bölünmelere uygulanır(8 adet var örnekte)
3) Uygunluk ölçütünde en yüksek uygunluk ölçütü Sektörde olduğu için Sektör seçilir.

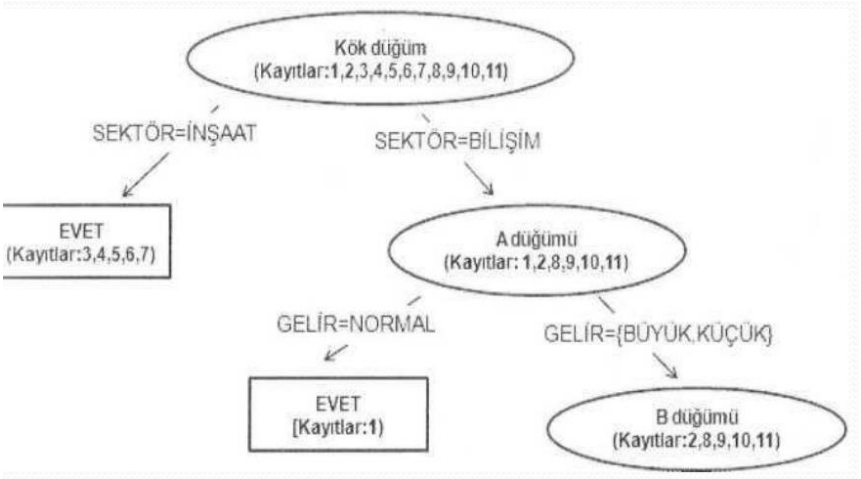


Şekil 1:Karar Ağacı

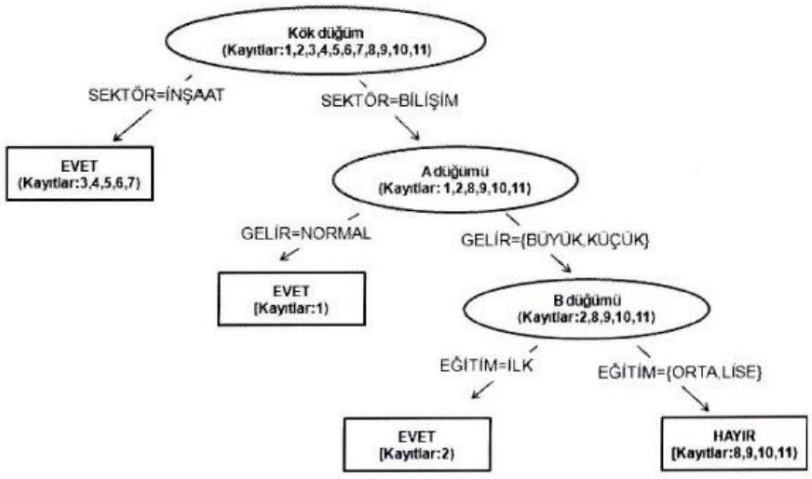
4) Tabloda sektör inşaat hep evet olduğu için o kenarın işi biter.
5) Sektör Bilişime göre adım 1 ve adım 2 tekrarlanır.
6) A düğümü Oluşturulur.

- **Adım 3 – Bütün Düğümler Adım 1 ve 2 ye göre tekrarlanır**

1) A düğümü Oluşturulur.



2) B düğümü Oluşturulur.



4.3. Gini Algoritması Örnek

NOT: Twoing Algoritması gibidir. Bu karar ağacında sadece 2 adet alt çocuk oluşur(2 dallanma oluşur her bir düğümde)

$$Gini_{sol} = 1 - \sum_{i=1}^{k} \left(\frac{L_i}{|T_{Sol}|} \right)^2$$

$$Gini_{sağ} = 1 - \sum_{i=1}^{k} \left(\frac{R_i}{|T_{Sağ}|} \right)^2$$

Bu bağlantıda yer alan ifadeler şu şekildedir:

k	Sınıfların sayısı
T	Bir düğümdeki örnekler
$\|T_{Sol}\|$	Sol taraftaki örneklerin sayısı
$\|T_{Sağ}\|$	Sağ taraftaki örneklerin sayısı
L_i	Sol tarafta *i* kategorisindeki örneklerin sayısı
R_i	Sağ tarafta *i* kategorisindeki örneklerin sayısı

4.3.1. Örnek

Başvuru	EĞİTİM	YAŞ	CİNSİYET	KABUL
1	ORTA	YAŞLI	ERKEK	EVET
2	İLK	GENÇ	ERKEK	HAYIR
3	YÜKSEK	ORTA	KADIN	HAYIR
4	ORTA	ORTA	ERKEK	EVET
5	İLK	ORTA	ERKEK	EVET
6	YÜKSEK	YAŞLI	KADIN	EVET
7	İLK	GENÇ	KADIN	HAYIR

4.3.1.1. Adım 1 – İkili Gruplandırma

Bu eğitim verisi üzerinde *Gini* algoritmasını uygulayabilmek için önce aşağıda belirtilen hesaplamalar yapılır. Bu tabloya göre EVET sınıfına ilişkin olarak EĞİTİM niteliğinin İLK değerinden 1 tane bulunmaktadır. Benzer biçimde (ORTA, YÜKSEK) değerlerinden ise 3 tane bulunmaktadır. Bu şekilde diğer değerler de hesaplanır.

KABUL	EĞİTİM		YAŞ		CİNSİYET	
	İLK	ORTA YÜKSEK	GENÇ	ORTA YAŞLI	KADIN	ERKEK
EVET	1	3	0	4	1	3
HAYIR	2	1	2	1	2	1

4.3.1.2. Adım 2 – $Gini_{sol}$ ve $Gini_{sağ}$ Değerlerinin Hesaplanması

EĞİTİM İçin:

$$Gini_{sol} = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 0.444$$

$$Gini_{sağ} = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 0.375$$

YAŞ İçin:

$$Gini_{sol} = 1 - \left[\left(\frac{0}{2}\right)^2 + \left(\frac{2}{2}\right)^2\right] = 0$$

$$Gini_{sağ} = 1 - \left[\left(\frac{4}{5}\right)^2 + \left(\frac{1}{6}\right)^2\right] = 0.320$$

CİNSİYET İçin:

$$Gini_{sol} = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = 0.444$$

$$Gini_{sağ} = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 0.375$$

4.3.1.3. Adım 3 – Niteliklerin Gini değerinin hesaplanması

c) $Gini_j$ değerlerinin hesaplanması: Elde edilen sonuçlar kullanılarak her bir nitelik için Gini değerleri elde edilir.

$$Gini_{eğitim} = \frac{3(0.444) + 4(0.375)}{7} = 0.405$$

$$Gini_{yaş} = \frac{2(0) + 5(0.320)}{7} = 0.229$$

$$Gini_{cinsiyet} = \frac{3(0.444) + 4(0.375)}{7} = 0.405$$

Sonuç olarak şu şekilde bir tablo elde edilir:

KABUL	EĞİTİM		YAŞ		CİNSİYET	
	İLK	ORTA, YÜKSEK	GENÇ	ORTA, YAŞLI	KADIN	ERKEK
EVET	1	3	0	4	1	3
HAYIR	2	1	2	1	2	1
$Gini_{sol}$ $Gini_{sağ}$	0.444	0.375	0.000	0.320	0.444	0.375
$Gini_j$	0.405		0.229		0.405	

4.3.1.4. Adım 4 – En Küçük Gini değerini seç

En Küçük Gini değeri seçilir ve Dallanmaya adım 1, 2 ve 3 ile devam edilir.

4.4. Bayes Algoritması Örnek

4.4.1. Koşullu Olasılık

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \qquad P(A \cap B) = P(A)P(B \mid A)$$

O halde A ve B gibi iki olayın birbiri ardına gerçekleşme olasılığı, iki olaya ait olasılıkların çarpımına eşittir.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

yazılabildiğinden,

$$P(A \cap B) = P(B)P(A \mid B)$$

biçiminde ifade edilebilir. Bu bağıntı yerine yazılırsa;

$$P(B \mid A) = \frac{P(B)P(A \mid B)}{P(A)}$$

elde edilir.

4.4.2. Sade Bayes Sınıflandırıcısı

$$P(X \mid C_i) = \prod_{k=1}^{n} P(x_k \mid C_i)$$

$$\arg\max_{C_i} \{P(X \mid C_i)P(C_i)\}$$

Sonrasal olasılıkları kullanan yukarıdaki ifade, en büyük sonrasal sınıflandırma yöntemi (Maximum A Posteriori classification = MAP) olarak da bilinir. O halde sonuç olarak, *Bayes* sınıflandırıcısı olarak aşağıdaki bağıntı kullanılabilir.

$$C_{MAP} = \arg\max_{C_i} \prod_{k=1}^{n} P(x_k \mid C_i)$$

4.4.3. Örnek

Başvuru	Eğitim	Yaş	Cinsiyet	Kabul
1	ORTA	YAŞLI	ERKEK	EVET
2	İLK	GENÇ	ERKEK	HAYIR
3	YÜKSEK	ORTA	KADIN	HAYIR
4	ORTA	ORTA	ERKEK	EVET
5	İLK	ORTA	ERKEK	EVET
6	YÜKSEK	YAŞLI	KADIN	EVET
7	İLK	GENÇ	KADIN	HAYIR
8	ORTA	ORTA	KADIN	EVET

Yukarıdaki eğitim kümesini ele alarak, Bayes sınıflandırıcısını kullanmak suretiyle aşağıdaki örneğin hangi sınıfa ait olduğunu belirlemek istiyoruz.

x1:EĞİTİM=YÜKSEK, x2:YAŞ=ORTA, x3:CİNSİYET=KADIN, KABUL = ?

4.4.3.1. Adım 1 – Sınıflara Göre Dağılımlarının Hesaplanması

Olasılıkların yer aldığı tablo

NİTELİKLER	DEĞERİ	KABUL			
		EVET		HAYIR	
		SAYISI	OLASILIK	SAYISI	OLASILIK
EĞİTİM	İLK	1	1/5	2	2/3
	ORTA	3	3/5	0	0
	YÜKSEK	1	1/5	1	1/3
YAŞ	GENÇ	0	0	2	2/3
	ORTA	3	3/5	1	1/3
	YAŞLI	2	2/5	0	0
CİNSİYET	ERKEK	3	3/5	1	1/3
	KADIN	2	2/5	2	2/3

4.4.3.2. Adım 2 – Yeni Gelen Değere göre olasılıkların hesaplanması

Yeni Gelen Veriye Göre; x1:EĞİTİM=YÜKSEK, x2:YAŞ=ORTA, x3:CİNSİYET=KADIN

- **P(X|C1)P(C1) olasılığının hesaplanması(Yani Kabul Evet'e Göre);**

P(x1|C1) = P(EĞİTİM=YÜKSEK|KABUL=EVET)=1/5

P(x2|C1) = P(YAŞ=ORTA|KABUL=EVET)=3/5

P(x3|C1) = P(CİNSİYET=KADIN|KABUL=EVET)=2/5

O halde;

P(x|C1)=P(X|KABUL=EVET)= (1/5)(3/5)(2/5) = 6/125 hesaplanır. Diğer taraftan P(KABUL=EVET) olasılığı şu şekilde elde edilir.

P(C1)=P(KABUL=EVET) = 5/8

Böylece,

P(X|C1)P(C1)=P(X|KABUL=EVET)P(KABUL=EVET) = (6/125)(5/8) = 0.03 elde edilmiş olur.

- **P(X|C2)P(C2) olasılığının hesaplanması(Yani Kabul Hayır'a Göre);**

P(x1|C2) = P(EĞİTİM = YÜKSEK | KABUL = HAYIR) = 1/3

P(x2|C2) = P(YAŞ = ORTA | KABUL = HAYIR) = 1/3

P(x3|C2) = P(CİNSİYET = KADIN | KABUL = HAYIR) = 2/3 bu değerler kullanılarak şu hesaplama yapılır:

P(X|C2) = P(X| KABUL = HAYIR) = (1/3)(1/3)(2/3) = 2/27

Bunun dışında P(KABUL=HAYIR) olasılığı şu şekilde elde edilir:

P(C2) = P(KABUL=HAYIR) = 3/8 olduğundan şu hesaplama yapılabilir:

P(X|C2)P(C2) = P(X|KABUL=HAYIR)P(KABUL=HAYIR) = (2/27)(3/8) = 0.027

4.4.3.3. Adım 3 – En yüksek olanı al

En yük değer yani 0.03 alıyoruz.

$$\arg\max_{C_i}\{P(X \mid C_i)P(C_i)\} = \max\{0.03,\ 0.027\} = 0.03$$

8.3.2. Bayes Sınıflandırıcılarda Sıfır Değer Sorunu

(n + kp) / (d + k) bağıntısı kullanılır. Burada k, 0 ile 1 arasında bir sayıdır. Genellikle 1 tercih edilir. Burada p değeri ise her bir nitelik değerinin muhtemel toplam sayısının bir belirli kısmı olarak seçilir. Eğer bir niteliğin iki muhtemel değeri varsa,

p = 1 / 2 = 0.5 olarak kabul edilebilir. Şimdi P(X | KABUL = HAYIR) koşullu olasılığını yeniden hesaplayalım. K = 1, p = 0.5 olacak biçimde,

$P(X \mid C_2) = P(X \mid KABUL = HAYIR) = (((1 + 0.5) / (3 + 1)) . ((1 + 0.5) / (3 + 1)) . ((2 + 0.5) / (3 + 1))) = 0.0878$

elde edilir.

4.5. EN YAKIN K-KOMŞU ALGORİTMASI(K-NN)

Bu yöntem, sınıfları belli olan bir örnek kümesindeki gözlem değerlerinden yararlanarak, örneğe katılacak yeni bir gözlemin hangi sınıfa ait olduğunu belirlemek amacıyla kullanılır.

1. Uzaklıkları Hesaplamak İçin Öklid Formülünü Kullan

$$D(i,j)=\sqrt[2]{\sum_{k=1}^{p}(x_{ik}-xjk)^2}$$

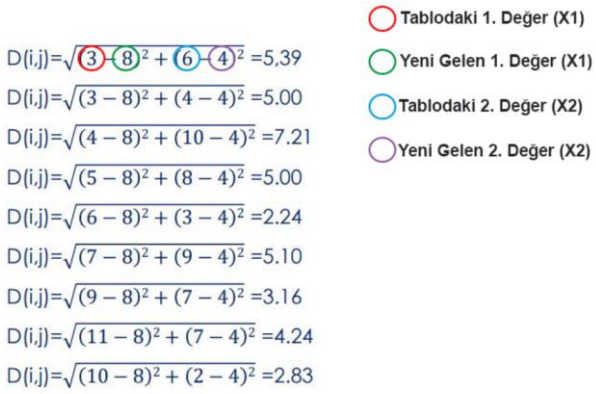
2. Veriler Çok Büyük ise: Normalizasyon kullan.Örnek; Her bir sütunun min ve max değerlerini bul. O sütuna ve yeni gelen noktaya normalizasyon uygula

Gözlem değerlerini (0,1) aralığına göre dönüştürmek için min-max normalleştirme yöntemini uygulayacağız. Bu amaçla tablo5 deki gözlem değerleri için,

$$X^* = \frac{X - Xmin}{X_{max} - Xmin}$$

Burada X* dönüştürülmüş değerleri, X gözlem değerlerini, X_{min} en küçük değerini ve X_{max} en büyük gözlem değerini ifade etmektedir. Bu değerler aşağıdaki tablo üzerinde yer almaktadır.

3. K'nın Belirlenmesi: Algoritmaya başlamadan önce k değerini seçiyoruz. Örnek: k=3 ise yeni gelen değer için en yakın 3 tane komşu arayacağımızı belirtmiş oluruz.
4. Uzaklıkların hesaplanması: Öklid formülünü kullanarak yeni gelen nokta ile tablodaki noktaların verilerinin tek tek uzaklıklarının hesaplanmasıdır. Örnek;



Hesaplanan değerler tablo üzerine yerleştirilir.

5. En küçük Uzaklıkların belirlenmesi: K = 3 ise 3 tane uzaklık seçilir ve bu 4 komşunun sınıflarına bakılır. Çoğunluk hangisi ise o sınıf yeni gelen noktanın sınıfı olur. Örnek;
   a) 1. Komşu: Z Sınıfında(Bu sınıflar iyi, kötü, evet hayır gibi son sınıflandırılma etiketidir)
   b) 2. Komşu: Z Sınıfında
   c) 3. Komşu: W Sınıfında

   Böylece Yeni gelen Noktanın sınıfı Z sınıfı olur.

6. Ağırlıklı Oylama İle Sınıf Seçme

K- algoritması, verilen bir gözleme en yakın komşunun belirlenmesi ve sınıfı yeni bir gözlem değeri için, k gözlem içindeki en fazla tekrar eden sınıfın seçilmesi esasına dayanıyordu. Ancak seçilen bu sınıf, sadece k komşunun göz önüne alınması nedeniyle her zaman uygun olmayabilir. Bu son aşamada k komşu arasında en çok tekrarlanan sınıfı seçme yöntemi yerine ağırlıklı oylama yöntemi uygulanır.

Bu yöntem gözlem değerlerini için aşağıdaki bağıntıya göre ağırlıklı uzunlukların hesaplanması esasına dayanır.

$$d(i,j)' = \frac{1}{d(i,j)^2}$$

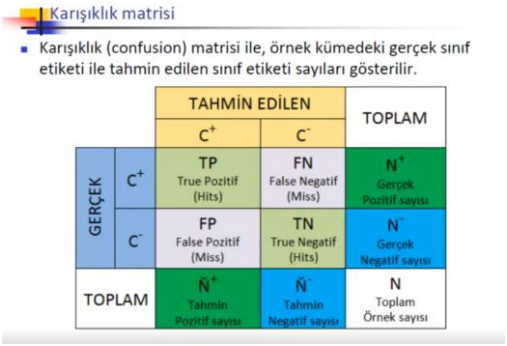
d (i,j) ifadesi i ve j gözlemleri arasındaki Öklid uzaklığıdır. Her bir sınıf değeri için bu uzaklıkların toplamı hesaplanarak ağırlıklı oylama değeri elde edilir. En büyük ağırlıklı oylama değerine sahip olan sınıf değeri yeni gözlemin ait olduğu sınıf kabul edilir.

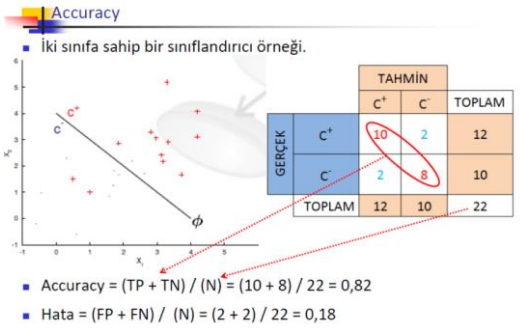
Örneğin; en yakın 1. Sıradaki komşunun Öklid uzaklığını 0.07 bulduk. Yukarıdaki formülü kullanarak her bir komşunun Ağırlıklarını hesaplıyoruz.

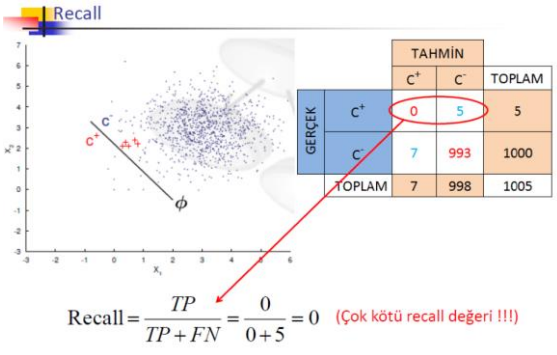
$$d(1,\text{yeni gözlem})' = \frac{1}{0.07^2} = 200$$

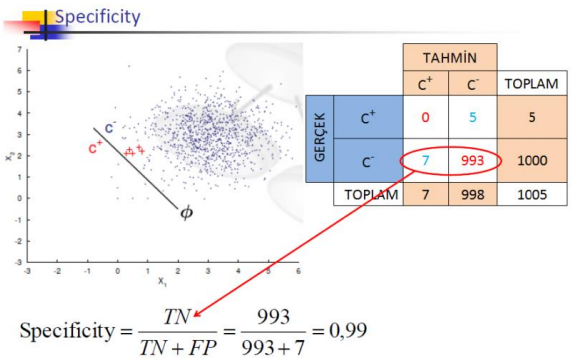
En büyük çıkan ağırlık yeni gelen noktanın sınıfı olur.

7. Yeni tahminin Kontrol Edilmesi: Yeni noktanın sınıfını Z bulduk fakat bunun doğru olup olmadığını kontrol etmemiz gerek.

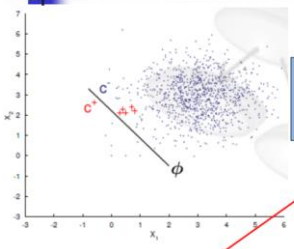








Precision

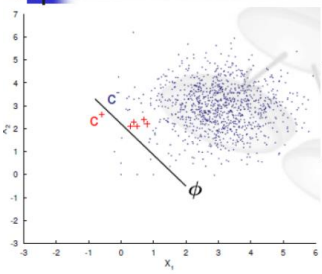


		TAHMİN		
		C^+	C^-	TOPLAM
GERÇEK	C^+	0	5	5
	C^-	7	993	1000
	TOPLAM	7	998	1005

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{0}{0+7} = 0$$

(Çok kötü precison değeri !!!)

F-Score



		TAHMİN		
		C^+	C^-	TOPLAM
GERÇEK	C^+	0	5	5
	C^-	7	993	1000
	TOPLAM	7	998	1005

$$F_1\text{-score} = \frac{2.\text{Precision}.\text{Recall}}{\text{Precision}+\text{Recall}} = \frac{2.0.0}{0+0} = 0$$