

1. Hangisi veri madenciliğinin geçmişte değil de günümüzde popüler olmasının sebeplerinden **DEĞİLDİR**?

- A) Disiplinler arası bir alan olması (Being an interdisciplinary field) +1
- B) Veriye ulaşılabilirlik (Data Availability)
- C) İşlem kapasitesinde artış (Increase in processing capacity)
- D) Veri saklama imkanlarının gelişmesi (Improvements in data storage)
- E) İnternet'in gelişimi (Development of Internet)

2. Hangi veri diğerlerine göre daha **AZ** değerlidir? (Which data is **LESS** valuable compared to others?)

- A) Arama metinleri (Search texts)
- B) Ürün açıklamaları (Product details) +1
- C) Forum mesajları (Forum entries)
- D) Sosyal medya mesajları (Social media entries)
- E) Ürün yorumları (Product comments)

3. Hangisi bir veri madenciliği uygulaması değildir? (Which one is **NOT** a data mining application?)

- A) Kümeleme (Clustering)
- B) Birliklilik analizi (Association analysis)
- C) Veri Çoğaltma (Data augmentation) +1
- D) Anomali analizi (Outlier Analysis)
- E) Sınıflandırma (Classification)

4. Hangi **veri madenciliği tekniği** OLTP'ye daha uygundur? (Which data mining techniques is more available for OLTP?)

- A) SQL Queries +3 (Öncelikle SQL bir veri madenciliği tekniği mi ondan emin misiniz?)
- B) Sınıflandırma (Classification) +1
- C) Veri indirgeme (Data reduction)
- D) Veri dönüştürme (Data transformation)
- E) Birliklilik kuralları (Association rules)

5. Veri madenciliği hangi kaynaklar üzerinde **YAPILMAZ**? (Which data sources **CANNOT** be used for data mining?)

- A) Database
- B) Secure data +4 (Secure data yani şifrelenmiş veya gizli veriler kullanılarak madencilik yapılamaz kullanıcıların izin verdiği ölçüde meta veriler üzerinden yapılır veya deşifre edilmiş datalar üzerinde yapılabilir)

- C)WWW (WWW üzerinde madencilik yapılabilir kesin bilgi web mining olarak geçiyor)
D)Data repositories
E)Data warehouse

6.Veritabanındaki 1 satırlık veri ne ile ifade edilir?(How can we express 1 row of database data?)

- A)Table
B)Tuple +1
C)Attribute
D)Entity
E)Field

7.İlişkisel veri en iyi ile işlenebilir?(Relational data can be best accessed via.....)

- A)Data Selection
B)Data Manipulation
C)Data Aggregation
D)Data Reduction
E)Database Queries +1

8.Veri madenciliği için en uygun veri saklama ortamı hangisidir?

- A)Data Repositories
B)Relational Database (Cevabın bu şık olduğunu düşününler sebebini yazarsa iyi olur)
C)Data Warehouse +4 (grupta veri ambarı denildi) (İlişkisel veritabanları kayıtları bir arada tutmak ve üzerinde birtakım işlemleri (CRUD) yapmak için kullanılan bir ortamdır büyük verileri bir arada barındıramaz.Veritabanı için en uygun ortam bu iş için özel olarak hazırlanan, bir takım aşamalardan geçirilen verileri barındıran veri ambarlarıdır)
<https://www.oracle.com/tr/database/what-is-a-data-warehouse/>
D)Graph Data +1
E)Streaming Data

9.Anomali tespiti neden önemlidir?

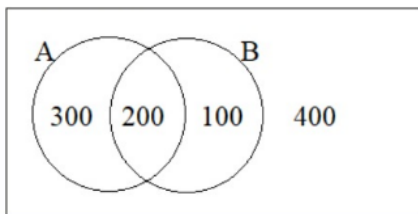
- A)İstatistiğe dayandığı için
B)Eksik verileri tamamladığı için
C)Veri boyutunu azalttığı için
D)Veri dönüşümünü sağladığı için
E)Veri güvenliğini arttırdığı için +3

10.Hangisi bir örüntünün ilgi çekici(interesting) olma ölçütü **DEĞİLDİR?**

- A)Correlation
B)Support
C)Regression +2
D)Lift
E)Confidence

10. $\text{Supp}(K \rightarrow L)$? **Formül: $\text{support}(A \rightarrow B) = P(A \cup B)$**

- A) $P(K|L)$
- B) $P(K \cap L)$
- C) $P(K \cup L) + 2$**
- D) $P(K) + P(L)$
- E) $P(L|K)$



11-12 . sorular için geçerlidir.

11. $\text{Supp}(A \rightarrow B)$? **$\text{Support}(A \rightarrow B) = \text{Support}(B \rightarrow A) \Rightarrow 200/1000 = 0.2$**

- A) 0.8
- B) 0.4
- C) 0.67
- D) 0.2 +2**
- E) 0.6

12. $\text{Conf}(A \rightarrow B)$? **$\text{Confidence}(A \rightarrow B) = 200/500 = 0.4$ $\text{Confidence}(B \rightarrow A) = 200/300 = 0,67$**

- A) 0.33
- B) 0.67
- C) 0.2
- D) 0.6
- E) 0.4 +4**

13. Hangisi danışmansız öğrenme ile ilgilidir?

- A) Decision tree
- B) Lift
- C) Clustering**
- D) Support
- E) Training set

14. Merkezi eğilim ölçütleri, veri madenciliğinde hangi amaçla **KULLANILMAZ**?

- A) İlginç örüntüleri bulma +2
- B) Anomali Tespiti
- C) Veri yumuşatma
- D) Eksik değerleri tamamlama
- E) Gürültüyü giderme

15. Hangisi danışmanlı öğrenme ile ilgili **DEĞİLDİR**?

- A) Classification
- B) Training set
- C) Labeled data
- D) CONFIDENCE ...+1
- E) Decision tree

16. Hangisi veri tabanındaki bir veri alanını ifade **ETMEZ**?

- A) Tuple
- B) Dimension ...+1
- C) Feature
- D) Variable
- E) Attribute 3, 11

17) A: {3, 11, 40, 43, 50, 54, 57, 60, 64, 70, 85, 98}

Outlier Limits = {IQR = Q3 - Q1 Q1 - 1.5IQR = 4 Q3 + 1.5IQR = 100}

Yukarıdaki veri seti için Outlier(lar) hangileri?

- a) 3 +5
- b) 3, 11
- c) hiçbiri...
- d) 85, 98
- e) 98

"Outlier demek Q1 - 1.5IQR = 4 Q3 + 1.5IQR = 100 bu değerlerin dışında kalması demek o yüzden cevap 3

Değişirse sayıları açıklama olsun. Şunun için bile video izledik. İnşallah geçeriz."

Arkadaşlar bu soruyu nasıl çözdünüz bilmiyorum ama bence doğru çözüm bu şekilde:

$Q1 = (40 + 43)/2$, $Q2 = (54 + 57)/2$, $Q3 = (64 + 70)/2$ (Medyan veya çeyrekler illa veri setinde olmak zorunda değildir five-number summary(min,q1,median,q3,max) elde edebilmek için yapıyoruz bu işlemleri.)

$IQR = Q3 - Q1 = 25.5$, $Q1 - (1.5 \cdot IQR) = 3.25$, $Q3 + (1.5 \cdot IQR) = 105.25$ $3 < 3.25$ olduğundan
Cevap : A

<https://www.statisticshowto.com/probability-and-statistics/interquartile-range/#:~:text=The%20interquartile%20range%20formula%20is,%3D%20Q3%20-%20Q1.>

İnşallah geçeriz bu dersi kolay gelsin herkese.

18)A:{30,40,100,120,500,800,900,1000,1100,2000,2700,3000}

Yukarıdaki veri seti için Outlier lar hangileridir?

A)2700,3000 ...+1

B)30,2700,3000

C)3000

D)30,300

E)30

çözüm:

$$Q1 - 1.5 \times IQR =$$

$$Q3 + 1.5 \times IQR =$$

$$IQR = 1100 - 100 = 1000$$

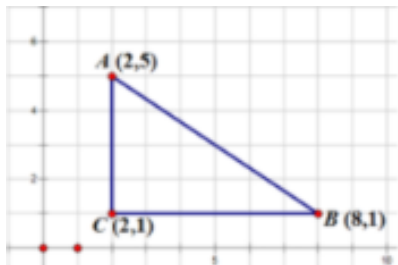
$$100 - 1.5 \times 1000 = -1400$$

$$1100 + 1.5 \times 1000 = 2600$$

$$-1400 < x < 2600$$

outlier: 2700, 3000

19)



A ve B noktaları arasındaki öklit ve supremum uzaklıkları sırası ile hangi şıkta doğru verilmiştir?

A) 7.2 and 6

B) $\sqrt{52}$ and 2

Öklit = Kök içinde $[(5-1)^2 + (8-2)^2] = 7,2$

Supremum = Boyutu en fazla olan yani $5-1=4$ ve $8-2=6$ büyük olan 6'yı

- C)6 and 1015
D)7.2 and 10
E) $\sqrt{52}$ and 4

seiyoruz.

20){500,800,900,1000,1100,2000,2700,3000}

Yukarıdaki veri setine min max normalizasyonu uygularsak 1000 in karşılığı nasıl olur?

- a)0.5
b)0.2 +1 1000-500/3000-500
c)0.1
d)0.4
e)0.25

21){500,800,900,1000,1100,2000,2700,3000}

Yukarıdaki veri setine z- score normalizasyonu uygularsak 1000 in karşılığı nasıl olur?
(Standart sapma: 880)

- A)1.76
B)-0.08
C)-0.57... +1 1000-1500/880
D)0.08
E)0.57

T100:<u1,u2,u3,u4,u5>

T200:<u1,u2,u3>

T300:<u2,u3>

T400:<u1,u3>

T500<u2,u3,u5>

22)Verilen transaction dataset için aşağıdakilerden hangisi **CLOSED FREQUENT ITEMSET** **DEĞİLDİR**? min_sup=2 (Bu tarz soruları nasıl bir mantıkla çözüyoruz?)

- A)<u2,u3,u5>
B)<u3,u5> anladığım üzere closed freq itemset denilince ikili gruplardan min_sup'a değeri eşit olanı seçiyoruz.
C):<u1,u3>
D)<u1,u2,u3>
E)<u2,u3>

İlk önce bütün ürünlerden sepetlerde kaç tane var diye yazıyoruz.

u1-> 3 adet

u2->4 adet min_sup=2 olduğu için 2'den küçük olanları(u4'ü) ikinci aşamaya almıyoruz.

u3->5 adet

u4->1 adet yani u1, u2, u3, u5 ile devam ediyoruz.

u5->2 adet (L1 Kümesi)=u1, u2, u3, u5

İkinci aşamada ise ikili olarak gruplandırıp sepetlerdeki sayılarını yazıyoruz.

u1,u2->2 adet

u1,u3->3 adet

u1,u5->1 adet min_sup=2' den küçük olduğu için bu ikiliyi(u1,u5) çıkarıyoruz.

u2,u3->4 adet

u2,u5->2 adet

..(L2 Kümesi)=u1u2, u1u3, u2u3, u2u5, u3u5

u3,u5->2 adet

Üçüncü aşamada ise üçlü olarak gruplandırıp bitiriyoruz.

u1,u2,u3->2 adet min_sup eşiğini geçen ve sonuç---> u1,u2,u3 ve u2,u3,u5

u2,u3,u5->2 adet (L3 Kümesi)= u1u2u3, u2u3u5

u1,u3,u5->1 adet

u1,u2,u5->1 adet -> bunu 2 sini sayamayız çünkü u1,u5 alt küme olarak almadık L2 de mevcut değil

23)

T100:<u1,u2,u3,u4,u5>

T200:<u1,u2,u3>

T300:<u2,u3>

T400:<u1,u3>

T500<u2,u3,u5>

Verilen transaction dataset için aşağıdakilerden hangisi **MAXİMAL FREQUENT İTEMSET** tir? min_sup=2 (Çözümünü ekleyebilirsiniz? Varsa kaynak da olur)

A)<u2,u3,u4> --> böyle bi grup yok zaten

B)<u2,u3>

C):<u3,u4,u5> --> böyle bi grup yok zaten

D)<u1,u2,u3>.. buradada max freq itemset'te ise üçlü gruplardan min_sup a değeri

E)<u1,u3> eşit olanı seçiyoruz.

24)

T100:<u1,u2,u3,u4,u5>

T200:<u1,u2,u3>

T300:<u2,u3>

T400:<u1,u3>

T500<u2,u3,u5>

min_sup:2; min_conf:50%

All_conf(u1,u3)=? u1 ile u3'ün beraber bulundukları sepet sayısı / tüm sepet sayısı demek

a)80% 3 / 5 = 0.6

b)50%

c)60%...+1

d)67%

e)100%

25)

T100:<u1,u2,u3,u4,u5>

T200:<u1,u2,u3>

T300:<u2,u3>

T400:<u1,u3>

T500<u2,u3,u5>

min_sup:2; min_conf:50%

Kulczynsky(u2,u3)=? **formül: $Kulc(A,B)=\frac{1}{2}*[P(A/B)+P(B/A)]$**

a)60%confid

b)70%

c)90%...+1

d)80%

e)100%

26) Sayısal sınav sonucu verilerini <kötü, orta, iyi> değerlerine dönüştürme işlemine ne ad verilir?

A) Discretization+ 1 example:(youth, adult, or senior)

B) Data Integration

C) Noise reduction

D) Normalization

E) Aggregation

27) Confidence (A1, A2 -> A3) = ? (A1 ve A2 içeren sepetin A3 alma olasılığı)

A) $P(A1, A2 \cup A3)$

B) $P(A1, A2 | A3)$

C) $P(A3 | A1, A2)$

D) $P(A3 \cup A1, A2)$

E) $P(A3 \& A1, A2)$

28) Aşağıdakilerden hangileri işlenmemiş veri ile ilgili problemlerden **DEĞİLDİR?**

A) **Security** +3 **emin misiniz?**

B) Accuracy

C) Completeness

D) Noise

E) Consistency

30) Sayısal sınav sonucu verilerini <kötü,orta,iyi> değerlerine dönüştürme işlemine ne ad verilir?

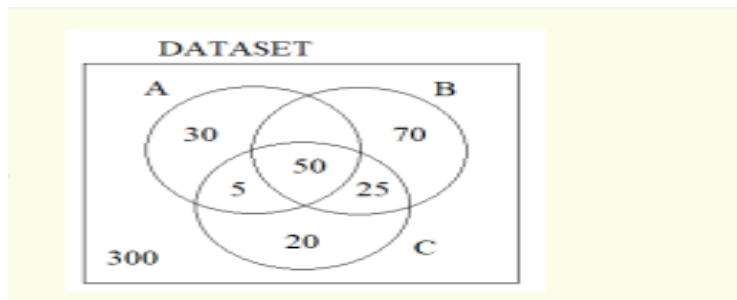
(If you convert numeric exam grades into values <bad, fair, good>, what is the name of this process?)

- A) Discretization
- B) Data Integration
- C) Aggregation
- D) Normalization
- E) Noise reduction

31) Veritabanından veri ambarına veri aktarırken sadece ilgi duyacağımız alanları aktarmak aşağıdakilerden hangisi ile ifade edilebilir?

- A) Data Transformation
- B) Data Cleaning
- C) Data Transferring
- D) Data Selection +1
- E) Data Integration

32)



Supp (C → B) = ?

A) 0.15 +1 support=freq(C,B)/all records = 75/500 = 0.15

- B) 0.52
- C) 0.75
- D) 0.375
- E) 0.44

33) Aşağıdaki değerleri alan bir niteliğin türü nedir?

Values: {0-Extra Small; 1-Small; 2-Medium; 3-Large ; 4-Extra Large }

A) Ordinal

B) Nominal

C) Numeric

D) Continuous

E) Discrete +1

Ordinal veriler de yine **kategorik** veri türündendir. Fakat **değerleri arasında sıralı bir ilişki bulunmaktadır.**

34) {3,7,12,4,6,21,40,30,20} veri setinin midrange'i kaçtır? Yanlış cevap 11.5

A) 21.5 +1 midrange = (min+max)/2=(40+3)/2 = 21.5

B) 11.5

C) 15.9

D) 21

E) 20

35) Hangisi sınıflandırma performansı artırıcı bir etken **DEĞİLDİR**?

A) Class imbalance (emin olmamakla beraber) bence doğru dengesiz veri performans artırmaz diğer seçeneklere netten baktım siz de bakın bi

B) Stratified cross validation

C) All Versus All

D) Ensemble

E) k-fold cross validation

36) TF matrisleri aşağıda verilen 2 dökümanın Cosine benzerliğini bulunuz.

| Document | team | coach | hockey | baseball | soccer | penalty | win | loss | season |
|-----------|------|-------|--------|----------|--------|---------|-----|------|--------|
| Document2 | 1 | 1 | 2 | 0 | 1 | 1 | 1 | 0 | 1 |
| Document3 | 0 | 2 | 0 | 2 | 1 | 0 | 1 | 0 | 0 |

A) $4/\sqrt{10}$

B) $2/\sqrt{10}$

C) $4/\sqrt{48}$

D) $4/48$

E) $2/5$ +2

37) Bir verisetinde 500 satır alışveriş verisi içinde 250 tanesi makarna, 150 tanesi ketçap,

100 tanesi de her iki ürünü içermektedir. min_support=15% ve min_confidence=30% için, Conf(Makarna->Ketçap) ve Lift(Makarna, Ketçap) ikilisi hangi şıklarda doğru verilmiştir?

A) 0.40 and 1.33

Conf(Makarna->Ketçap) = 100/250=0.40

B) 0.66 and 1.33

Lift = P(AUB)/P(A)*P(B) = 0.20 / (0.50 * 0.30) = 1.33

C) 0.66 and 0.75

Makarna = %50 toplam 500 içinde yarısı makarna çünkü

D) 0.20 and 1.33

Ketçap = %30 = 150 / 500 = 300/1000 = 30/100 den %30

E) 0.40 and 0.75

İkisinin old. = %20 = 100 / 500 = 200/1000= 20/100 den %20

38) Bir verisetinde 500 satır alışveriş verisi içinde 250 tanesi makarna, 150 tanesi ketçap, 100 tanesi de her iki ürünü içermektedir. min_support=15%ve min_confidence=30% için, Kulczynsky(Makarna, Ketçap) değerini bulunuz.

A) 0.20

Kulczynsky = ½ [conf(A->B) + conf(B->A)]

B) 0.53

½ [100 / 250 + 100 / 150] = ½ [0,40 + 0,67] = 0,535

C) 0.75

D) 0.40

E) 0.66

39)

| Department | English | Age | SUCCESS |
|------------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |

Class Label : SUCCESS

↓Entropy Table (For.Ex: {Yes, Yes, Yes, No} -> 1|3)

| Division | Entropy | Div. | Entr. | Div. | Entr. |
|---------------|---------|-------|-------|-------|-------|
| 1 1 (3 3) | 1 | 1 3 | 0.811 | 1 5 | 0.65 |
| 1 2 (2 4) | 0.918 | 3 4 | 0.985 | 1 6 | 0.592 |
| 3 2 (6 4) | 0.971 | 3 5 | 0.954 | 2 5 | 0.863 |
| 1 4 (2 8) | 0.722 | 3 7 | 0.881 | | |

Yukarıdaki veri setinde ID3 algoritması için 2.aşamada kök düğümden sonraki gelen düğümü belirlerken Gain(Age) değeri kaçtır?

A) **0.571 (+2)**

B) 0.551

C) 0.42

D) 0.4

E) 0.918

ÇÖZÜM

1.aşama

Gain(Dep)=0,395 -> kök düğüm

Gain(Eng)=0,12

Gain(Age)=0,071

2.aşama

Gain(Eng) = 0,42

Gain(Age)= 0.571

40)

Yukarıdaki veri setinde ID3 algoritması için **kök düğümün belirlediği ilk aşamada**

| Departmer | English | Age | SUCCESS |
|-----------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |

Class Label : SUCCESS

↓ Entropy Table { For.Ex: {Yes, Yes, Yes, No} -> 1|3 }

| Division | Entropy | Div. | Entr. | Div. | Entr. |
|---------------|---------|-------|-------|-------|-------|
| 1 1 (3 3) | 1 | 1 3 | 0.811 | 1 5 | 0.65 |
| 1 2 (2 4) | 0.918 | 3 4 | 0.985 | 1 6 | 0.592 |
| 3 2 (6 4) | 0.971 | 3 5 | 0.954 | 2 5 | 0.863 |
| 1 4 (2 8) | 0.722 | 3 7 | 0.881 | | |

Gain(Department) değeri kaçtır?

A) 0.071

B) 0.486

C) 0.881

D) 0.395 (+2)

E) 0.12

41)

| Departmer | English | Age | SUCCESS |
|-----------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |

Class Label : SUCCESS

↓Entropy Table (For.Ex: {Yes, Yes, Yes, No} -> 1|3)

| Division | Entropy | Div. | Entr. | Div. | Entr. |
|---------------|---------|-------|-------|-------|-------|
| 1 1 (3 3) | 1 | 1 3 | 0.811 | 1 5 | 0.65 |
| 1 2 (2 4) | 0.918 | 3 4 | 0.985 | 1 6 | 0.592 |
| 3 2 (6 4) | 0.971 | 3 5 | 0.954 | 2 5 | 0.863 |
| 1 4 (2 8) | 0.722 | 3 7 | 0.881 | | |

Yukarıdaki veri setinde ID3 algoritması için 2.aşamada kök düğümden sonraki gelen düğümü belirlerken Info(D) değeri kaçtır?

A) 0.971 +3

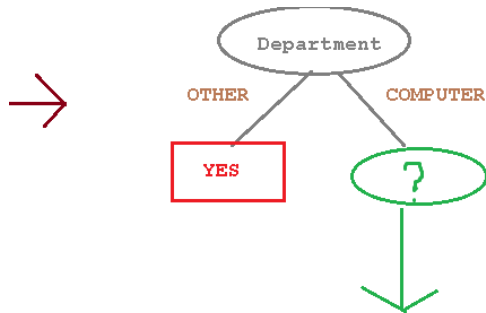
B) 0.918

C) 0.881 +1

D) 0.811

E) 1

Çözüm:



2. AŞAMADA COMPUTER DALINDAN İLERLEYECEĞİZ
(tabloda sadece comp olan kısmı alalım)

| Departmer | English | Age | SUCCESS |
|-----------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |

2 Yes, 3 No için info(d)= 0,971

42)

| Department | English | Age | SUCCESS |
|------------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |

Class Label : SUCCESS

↓Entropy Table (For.Ex: {Yes, Yes, Yes, No} -> 1|3)

| Division | Entropy | Div. | Entr. | Div. | Entr. |
|---------------|---------|-------|-------|-------|-------|
| 1 1 (3 3) | 1 | 1 3 | 0.811 | 1 5 | 0.65 |
| 1 2 (2 4) | 0.918 | 3 4 | 0.985 | 1 6 | 0.592 |
| 3 2 (6 4) | 0.971 | 3 5 | 0.954 | 2 5 | 0.863 |
| 1 4 (2 8) | 0.722 | 3 7 | 0.881 | | |

Yukarıdaki veri setinde ID3 algoritması için **kök düğümün belirlediği ilk aşamada Gain(English)** değeri kaçır?

A) 0,7291

B) 0.1201 Cevap bu şekilde olmalı (+3)

C) 0,486

D) 0,810

E) 0,395

```

a    b    <-- classified as
93   33   |    a = True
7    218  |    b = False

```

Recall=TP/P

43)Verilen Confusion Matrix için **Recall** değerini bulunuz.

A) 73.8% +1

B) 69.9%

C) 93%

D) 26.2%

E) 96.9%

```

a    b    <-- classified as
93   33   |    a = True
7    218  |    b = False

```

Accuracy=(TP+TN)/(P+N)

44)Verilen Confusion Matrix için **Accuracy** değerini bulunuz.

(Find Accuracy for the given Confusion Matrix)

A) 98%

B) 26.5% $(93+7)/(126+225)=100/351$

C) 90.6%

D) 88.6% TP=93, TN=218, P=(TP+FN)=126, N=(FP+TN)=225 => $(93+218)/(126+225) \times 100 =$

Cevap = 88.6

E) 62.1%

45) Hangisi sınıflandırma işleminin bir **ÇIKTISIDIR**?

A) Güçlü ürün setleri (Strong itemsets)

B) Anormal veriler (Outliers)

C) İlginç örüntüler (Interesting patterns)

D) Eğitimi kümesi (Training set)

E) Karar ağacı (Strong itemsets) +1

47) İşleyemeyeceğimiz kadar çok veri içinden belli bir miktarını kullanmak, aşağıdakilerden hangisi ile ifade edilebilir?

(Which of the following expresses using only a portion of a huge dataset that is impossible to process as a whole?)

A) Data transformation

B) Data mining

C) Data selection +1

D) Pattern evaluation

E) Data cleaning

48) Hangi **veri madenciliği tekniği** OLTP'ye daha uygundur?

(Which data mining techniques is more available for OLTP?)

A) Sınıflandırma (Classification)

B) Veri dönüştürme (Data transformation)

C) Biriklilik kuralları (Association rules)

D) Veri indirgeme (Data reduction)

E) SQL Queries +1

49) A:{500,800,900,1000,1100,2000,2700,3000}

Yukarıdaki veri setine min-max normalizasyonu uygularsak 1000'in karşılığı ne olur?

(What is the min-max normalized value of 1000 for the above dataset?)

A) 0,4

B) 0,25

C) 0,1

D) 0,2 +1

E) 0,5

50) Veri tabanındaki **1 satırlık veri** ne ile ifade edilebilir?

(How can we express 1 row of database data?)

A) Tuple

B) Table

C) Attribute

- D) Entity
- E) Field

53) Hangisi bir ratio-scaled attribute **DEĞİLDİR**?

(Which one is NOT a ratio-scaled attribute?)

- A) Height (m)
- B) Speed (Km/h)
- C) Volume (m3)
- D) Page count
- E) Temperature (Fahrenheit) (is fahrenheit ratio or interval? şeklinde aratırsanız doğru cevabın bu şekilde olduğunu görebilirsiniz. Eğer sıcaklık celsius olsaydı o zaman işler farklı olabilirdi.) +1

54) A:{500,800,900,1000,1100,2000,2700,3000}

Yukarıdaki veri setine z-score normalizasyonu uygularsak 1000'in karşılığı ne olur? (Standart Sapma:880)

(What is the z-score normalized value of 1000 for the above dataset? - standard deviation is 880)

- A) 1.76
- B) -0.08
- C) -0.57 +2
- D) 0.08
- E) 0.57

55) T100: <u1, u2, u3, u4, u5>

T200: <u1, u2, u3>

T300: <u2, u3>

T400: <u1, u3>

T500: <u2, u3, u5>

----- min_sup:2 ; min_conf: 50% -----

Hangisi **L2** seti içinde yer almaz ?

Which one is **NOT** listed in **L2** set ?

- A) <u1, u2>
- B) <u1, u3>
- C) <u2, u5>
- D) <u1, u5>
- E) <u3, u5>

56)

T100: <u1, u2, u3, u4, u5>

T200: <u1, u2, u3>

T300: <u2, u3>

T400: <u1, u3>

T500: <u2, u3, u5>

----- min_sup:2 ; min_conf: 50% -----

L1 kümesi hangisini İÇERMEZ ? (Çözümünü de ekleyebilirmisiniz ? Hangi konu nereye çalışmalıyız ya da) 22.soruda çözümü var.

Which one is **NOT** included in **L1** set?

A) u1

B) u2

C) u4 +2

D) u3

E) u5

57) Hangisi sınıflandırma ile ilgili **DEĞİLDİR?**

(Which one is NOT related with classification?)

A) Learning

B) Class label

C) Partitioning +1

D) Training set

E) Accuracy

58) Hangisi danışmanlı öğrenme ile ilgili **DEĞİLDİR?**

(Which one is NOT related with supervised learning?)

A) Farthest Neighbor algorithm

B) k-Nearest Neighbor algorithm

C) ID3

D) Naive Bayes (Naive Bayes unsupervised(danışmansız) bir sınıflandırıcı) +1

E) Gini

DERSTE SORARIM DEMİŞTİ

59)

Bir veri küpünde herhangi bir boyutu daha özetleyerek-az detay ile ifade etmeye ne ad verilir? (Which one is expressing a dimension of a data cube in a more generalized-less detailed manner?)

A) Roll down

- B) Drill down
- C) Hang up
- D) Roll up * +1
- E) Detail in

60)

Nominal niteliklerde eksik veri yerine hangisi kullanılabilir?

(Which measure can be used for missing values of nominal attributes?)

- A) Median
- B) Range
- C) Mode * +1
- D) Mean
- E) Max

61)

A: {30, 40, 100, 120, 500, 800, 900, 1000, 1100, 2000, 2700, 3000}

Yukarıdaki veri seti için Outlier(lar) hangileridir?

(What are the outlier(s) for the above dataset?)

- A) 2700, 3000 +1 (Bu sorunun nasıl çözüldüğünü de eklermisiniz.)
- B) 30
- C) 30, 2700, 3000
- D) 30, 3000
- E) 3000

çözüm : $q_1 = 100$

$q_2 = 800$

$q_3 = 1100$

$IQR = q_3 - q_1 = 1000$

$q_1 - 1.5 \times (IQR) = -1400$

$q_3 + 1.5 \times (IQR) = 2300$

-1400 ve 2300 aralığı dışında kalanlar **cevap** 2700, 3000

q2 neden 800 oldu bilen var mı tam ortadaki değeri almamız gerekmiyor muydu?

12 veride 3.=Q1 6.=Q2 9.= Q3 olur -> Neden bu şekilde alıyoruz derste hoca bu konuda birşey demişmiydi internette araştırınca farklı hesaplanıyor? hoca dedi ders kayıtlarını izle :)

62)

Hangisi sınıflandırmanın uygulama alanlarından **DEĞİLDİR**?

(Which one is NOT an application field of classification)

- A) Loan applications (Borçlanma/Kredi)
- B) Market basket Analysis (Pazar sepet analizi)

C) Weather prediction (Hava tahmini)

D) Fraud Detection * +1

E) Medical Diagnosis (Tıbbi Tahminleme) (Notlarda tıbbi tahminleme için bir sınıflandırma değildir denmiş kontrol edermisiniz)

63)

Bir veriye ait hedef niteliğin tahmin edilmesi işlemine ne ad verilir?
(Which one is the process of predicting the target label of a tuple?)

A) Classification +1

B) Pattern mining

C) Association Rules

D) Clustering

E) Regression

64)

T100: <u1, u2, u3, u4, u5>

T200: <u1, u2, u3>

T300: <u2, u3>

T400: <u1, u3>

T500: <u2, u3, u5>

----- min_sup:2 ; min_conf: 50% -----

Kulczynsky(u2, u3)=? çözümünü yazar mısınız tahminen 1 sepette olmadığı için %90 diye

A) 60%

B) 70%

C) 100%

D) 90% +1

E) 80%

66) Hangisi öbek içi benzerliği yüksek, öbekler arası benzerliği düşük tutma prensibi ile çalışır?

A) Outlier analysis

B) Classification

C) Associate analysis

D) Regression

E) Clustering +1

67) Farklı veri kaynaklarından gelen veride farklı tarih formatları varsa bunları tek formata dönüştürmek gerekir. Bu işlem hangi aşamaya karşılık gelir?

A) Data unification

B) Data selection

C) Data transformation

D) Data Cleaning

E) Data Integration +2 (Notlarda data integration olarak ifade edilmiş kontrol edermisiniz)

68)Pattern evaluation ile kastedilen veri madenciliği uygulaması hangisi olabilir?

(Which one best suites pattern evaluation stage?)

A) Data integration

B) Data reduction

C) Clustering

D) Classification

E) Market Basket Analysis +4

69)Bir veri küpünde herhangi bir boyutu daha detaylı ifade etme işlemine ne ad verilir?

A)Drill down +1

B)Drill up

C)Fill up

D)Roll down

E)Roll up

70)Hangisi bir sınıflandırıcının başarısını değerlendirmede **KULLANILMAZ**?

A)F-score

B)Data integrity

C)Accuracy

D)Robustness

E)Speed

71)Hangisi bir kümeleme algoritması **DEĞİLDİR**?

A)Nearest Neighbor

B)k-Nearest Neighbor (sınıflandırma algoritması)

C)k-Means

D)Single Linkage

E)Farthest Neighbor

72)Aşağıdakilerden hangisi Kümelemede çok büyük veri ile daha kolay çalışır?

A)Grid-based methods

B)Density-based methods

C)Partitioning methods

D)Deep scan methods

E)Hierarchical methods

73)Hangisi bir interval-scaled attribute'tur?

which one is an interval-scaled attribute?

A)Age

B)Distance

C)Power

D)Year +1

E)Weight

74)Hangisi bir sınıflandırıcı modeli test etmek için **EN KÖTÜ** seçenektir?

A)Test set

B) Independent data set

C) Percentage Split from the training set

D) Training set +1

E) Cross validation

75){3,7,12,4,6,21,40,30,20} veri setinin medyanı kaçtır?

A)7

B)20

C)4

D)6

E)12 +1

76) Hangisi bir veri önışleme aşaması **DEĞİLDİR**?

A) Data cleaning

B) Data transformation

C) Data İntegration

D) Data Selection

E) Data retrieving

77) Hangi veri kaynağı geleceğe dair daha az ipucu verir?

A) Hepsiburada search texts

B) Google entries

C) Twitter entries

D) Whatsapp texts +1

E) Hotel comments

78) Hangisi veri ambarı oluşturma aşamalarından **DEĞİLDİR**? (Which one is **NOT** a step of Data Warehouse construction?)

A) Periodic data refreshing *

B) Data loading

C) Data integration

D) Data cleaning

E) Data presentation +1

80) Aşağıdaki değerleri alan bir niteliğin türü nedir? Which type of attribute is the following?
(Bu soru 33.soruyla aynı cevaplar neden farklı lütfen bu tarz soru tekrarları olmasın !?)

Values : { 0-Extra Small ; 1-Small ; 2-Medium ; 3-Large ; 4-Extra Large }

A) Ordinal +2

B) Continuous

C)Numeric

D)Nominal

E)Discrete

Ordinal veriler de yine **kategorik** veri türündendir. Fakat **değerleri arasında sıralı bir ilişki bulunmaktadır.**

81) Bir metin verisini Term Frequency (TF) matrisine dönüştürdüğümüzde hangi dönüşüm yöntemini uygulamış oluruz?(Which transformation method is applied if we convert a text data into Term Frequency (TF) matrix?)

A)Discretization

B)Aggregation

C)Smoothing

D)Attribute construction

E)Concept hierarchy generation

82) t100: <u1, u2, u3, u4, u5>

T200: <u1, u2, u3 >

T300: <u2, u3>

T400: <u1, u3>

T500: <u2, u3, u5>

----- min_sup : 2 ; min_conf : 50% -----

Hangisi **L3** seti içinde **yer alır**?

Which one is **listed** in **L3** set? 22.soruda çözümü mevcut

A)<u2,u3,u5> +2

B)<u1,u3,u5>

C)<u3,u4,u5>

D)<u2,u3,u4>

E)<u1,u2,u5>

83) T100: <u1, u2, u3, u4, u5>

T200: <u1, u2, u3 >

T300: <u2, u3>

T400: <u1, u3>

T500: <u2, u3, u5>

----- min_sup : 2 ; min_conf : 50% -----

Ortak güven hesaplama yöntemi olarak all_confidence kullanıldığında hangisi bir güçlü birliktelik **OLMAZ**?

If all_confidence measure is used for pairs, which of the following is **NOT** a strong association?

A) u2,u3

B) u3,u5

C) u1,u2

D) u1,u3

E) u2,u5

84) Hangi değerlere sahip niteliğin Info Gain değeri en **DÜŞÜKTÜR**?

(Which attribute with corresponding values has the **LOWEST** Info Gain?)

A) A: {10,30,50}

B) A: {10,10,10,10,30,10} Çeşitlilik azaldıkça Gain değeri düşer.

C) A: {10,20,30,40,50}

D) A: {10,15,20,25,30}

E) A: {10,30,30,10,30,10}

85) Hangisi sınıflandırma performansını artırıcı bir etken **DEĞİLDİR**?

Which one is NOT a factor increasing classifier performance?

A) Ensemble

B) Class Imbalance

C) k-fold cross validation

D) Stratified cross validation

E) All Versus All

86) Sınıflandırıcılar hakkında hangisi **YANLIŞTIR**? (Which of the following is **WRONG** about classifiers)

A) Boosting yönteminde her sınıflandırıcının ağırlığı, başarısına göre artabilir. (In boosting method, weight of each classifier may increase according to its success)

B) Random Forest yönteminde birbirinden farklı veri setleri kullanılarak farklı ağaçlar oluşturulur. (In Random Forest method, different trees are generated by using datasets independent from each other)

C) Ensemble modeller tek sınıflandırıcılara göre daha iyi sonuç verir. (Ensemble classifiers are better than one classifier model)

D) Bagging yönteminde her sınıflandırıcının eşit oyu vardır. (In bagging method, each classifier has equal vote)

E) All versus All, One versus All'a göre daha başarılıdır. (All versus All is more successful than One versus All) **doğru**

87) Aşağıdakilerden hangisi Kümelemede Partitioning Methods'a ait özelliklerden **DEĞİLDİR**? Which of the following is **NOT** a property of Partitioning Methods in Clustering?

- A) May use mean or medoid (etc.) to represent cluster center
- B) Cannot correct erroneous merges or splits**
- C) Find mutually exclusive clusters of spherical shape
- D) Distance-based
- E) Effective for small- to medium-size data sets

88) Aşağıdakilerden hangisi Kümelemede çok büyük veri ile daha kolay çalışır? Which of the following runs easier with very big data in Clustering?

- A) Density-based methods
- B) Deep Scan methods
- C) Partitioning methods
- D) Hierarchical methods
- E) Grid-based methods**

89) Aşağıdakilerden hangisi Kümelemede "S" benzeri bir kümenin tespitinde daha verimli çalışır? Which of the following runs more effective to find "S"-shaped clusters?

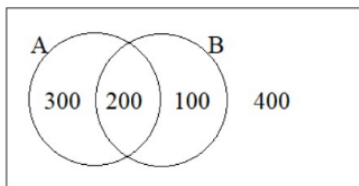
- A) Density-based methods**
- B) Deep Scan methods
- C) Partitioning methods
- D) Hierarchical methods
- E) Grid-based methods

90) Hangisi bir sınıflandırma algoritması **DEĞİLDİR**? (Which one is **NOT** a classification algorithm?)

- A) k-Nearest Neighbor (Bellek tabanlı sınıflandırma)
- B) Naive Bayes (Bayes sınıflandırıcı diye konu başlığı var)
- C) J48 (Bu da c4.5 in weka karşılığı yanı bu da sınıflandırma)
- D) Gini (sınıflandırma)
- E) Farthest Neighbor (kitaptan baktım kümeleme bu sınıflandırma değil)**

Araştırarak baktım ama yine de bir bakın.

91)



Supp(B→A)=?

A) 0.6

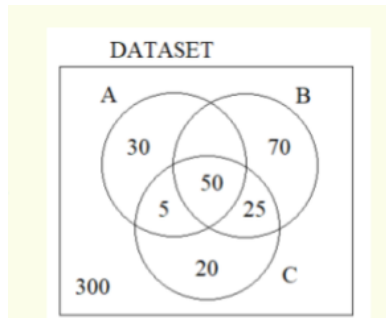
B) 0.33

C) 0.67

D) 0.4

E) 0.2 (200/1000 = 0.2)

92)



Conf(B,C→A)=?

A) 0.32

B) 0.67 (50/75 = 0.67) + 1 Hepsinin olduğu küme / B ve C nin old. küme

C) 0.1

D) 0.59

E) 0.33

93) Verinin genel akışını bozan örneklerin tespitine ne ad verilir? Which one is the name for detecting samples that are inconsistent with general behavior of the data?

A) Outlier detection

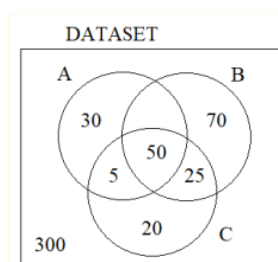
B) Classification

C) Clustering

D) Fraud detection

E) Association analysis

94)



Supp(A,B->C)=?

A) 0.1 (Okun bir önemi yok için tüm verisetinde bulunma oranı $50/500 = 0.1$)

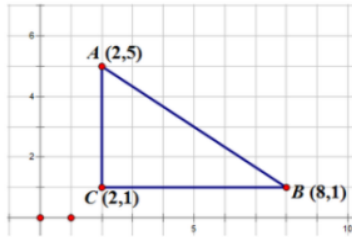
B) 0.25

C) 1.0

D) 0.44

E) 0.5

95)



A ve B noktaları arasındaki Manhattan ve Minkowsky (h=3) uzaklıkları sırası ile hangi şıkta doğru verilmiştir? Which pair is true for Manhattan and Minkowsky (h=3) distances between points A and B ?

A) 10 and 6

B) 7.2 and 6.54

C) 2 and 6

D) 10 and 7.2

E) 10 and 6.54

96)

| Document | team | coach | hockey | baseball | soccer | penalty | score | win | loss | season |
|-----------|------|-------|--------|----------|--------|---------|-------|-----|------|--------|
| Document2 | 1 | 0 | 2 | 0 | 1 | 1 | 0 | 1 | 0 | 1 |
| Document3 | 0 | 1 | 1 | 2 | 1 | 0 | 0 | 1 | 1 | 0 |

TF matrisleri aşağıda verilen 2 dökümanın Cosine benzerliğini bulunuz? For the two documents TF matrices below, what is the Cosine similarity? (Lütfen eksik ve tekrar eden soru olmasın) (Bu sorunun cevabı küsürtlü çıkmıyor mu arkadaşlar)

A) 4/9 (2+1+1)/ 3x3 =4/9 cevap

B) 2/7

C) 4/7 emin değilim

D) 2/9

E) 4/49

97)

T100:<u1,u2,u3,u4,u5>

T200:<u1,u2,u3>

T300:<u2,u3>

T400:<u1,u3>

T500<u2,u3,u5>

Verilen transaction dataset için aşağıdakilerden hangisi **CLOSED FREQUENT ITEMSET** TİR? (min_sup=2)

A)<u1,u5>

B)<u2,u5>

C):<u3,u5>

D)<u2,u3>

E)<u4,u5>

99) Bir verisetinde 500 satır alışveriş verisi içinde 250 tanesi makarna, 150 tanesi ketçap, 100 tanesi de her iki ürünü içermektedir. min_support=15% ve min_confidence=30% için, Conf(Ketçap->Makarna) ve Lift(Ketçap, Makarna) ikilisi hangi şıklarda doğru verilmiştir?

A) 0.4 and 0.75

B) 0.4 and 1.33

C) 0.66 and 1.33 +1

D) 0.2 and 1.33

E) 0.66 and 0.75

100)

T100: <u1, u2, u3, u4, u5>
T200: <u1, u2, u3 >
T300: <u2, u3>
T400: <u1, u3>
T500: <u2, u3, u5>

Ortak güven hesaplama yöntemi olarak **Kulczynsky** kullanıldığında **EN GÜÇLÜ** birliktelik hangisi olur?

A) u1, u2

B) u2, u5

C) u2, u3

D) u1, u3

E) u3, u5

101)

T100: <u1, u2, u3, u4, u5>
T200: <u1, u2, u3 >
T300: <u2, u3>
T400: <u1, u3>
T500: <u2, u3, u5>

Hangi 1-itemset **EN GÜÇLÜDÜR** ?

- A) u5
- B) u2
- C) u3**
- D) u4
- E) u1

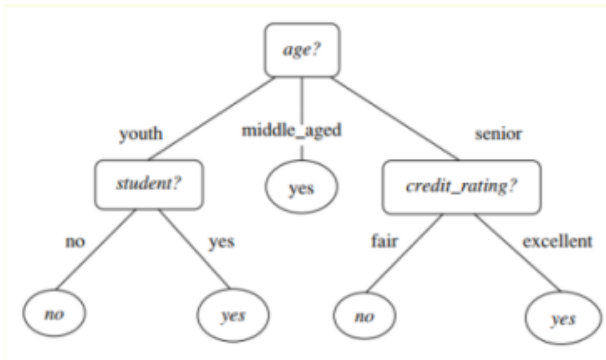
102)

T100: <u1, u2, u3, u4, u5>
T200: <u1, u2, u3 >
T300: <u2, u3>
T400: <u1, u3>
T500: <u2, u3, u5>

Aşağıda verilen 3-itemset'ler içinde hangisi güçlü birliktelik **DEĞİLDİR** ?

- A) u1 -> u2, u3
- B) u2,u3 -> u1
- C) u1,u3 -> u2
- D) u3, -> u1, u2**
- E) u1, u2 -> u3

103)



Yukarıdaki şekle göre [Student?] düğümü ne tür bir düğümdür?

- A) Terminal Node
- B) Root Node
- C) Leaf Node
- D) Branch
- E) Internal Node**

Çözüm:

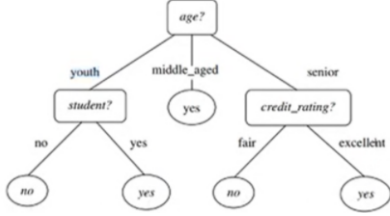
Karar Ağacı (Decision Tree)

kök düğüm->root node(en üstteki düğüm)

yaprak düğüm->leaf node (sonuca ulaştığınız düğüm)

aradaki düğümler -> internal node (student , credit rating)

dallanmalar -> branch (no,yes, fair, excellent)



104) Hangi değerlere sahip niteliğin Info Gain değeri **EN YÜKSEKTİR**?

(Which attribute with corresponding values has the **HIGHEST** Info Gain?)

- A) A:{Karabuk, Ankara, Karabuk, Ankara, Karabuk, Ankara, Karabuk, Ankara}
- B) A:{Ankara, Ankara, İstanbul, İstanbul, İzmir, İzmir}
- C) A:{Karabuk, Ankara, İstanbul, Bursa, İzmir, İzmir}**
- D) A:{Karabuk, İzmir, İstanbul, İzmir, İzmir, İzmir}
- E) :{Karabuk, Karabuk, Karabuk, Karabuk, Karabuk}

105)



Yukarıdaki karar ağacı için budama eşiği $th=0.15$ ile budama işlemi tamamlandığında **A2** düğümünün son durumu ne olur ?

- A)Hiçbiri(None)
- B)Class A
- C)Class B
- D)Bir üst düğüme kadar budanacağı için silinir.
- E)Aynen korunur.(It holds as the same)**

106)

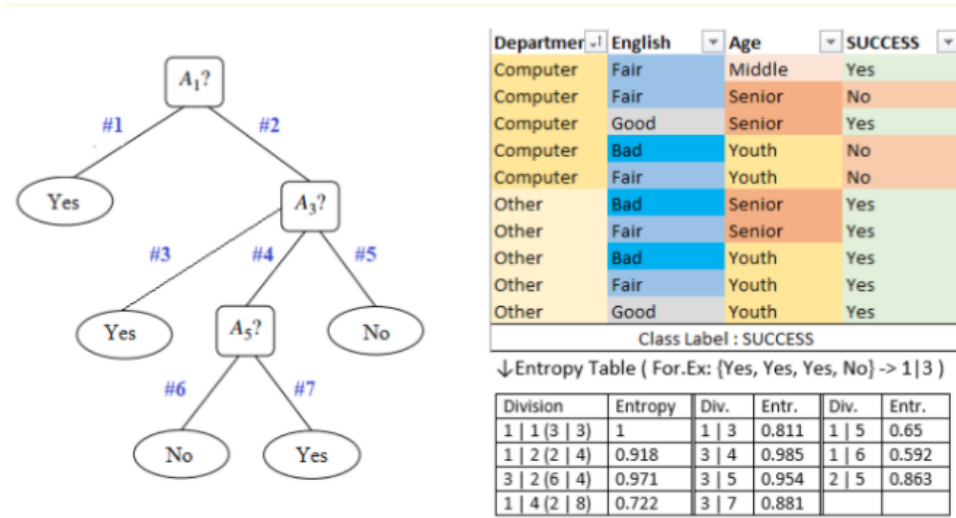
| Departmer | English | Age | SUCCESS |
|-----------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |

Class Label : SUCCESS

Üstteki veri seti için **Gini(Age)** değerini hesaplayınız. Dallanma : [middle&senior | youth]

- A) 0.44
- B) 0.32
- C) 0.4 +1
- D) 0.24
- E) 0.48

107)



ID3 algoritması ile yukarıdaki veri setinden elde edilen karar ağacında **A1 ve A3** düğümleri sırasıyla hangi niteliklerden oluşmaktadır ?

- A) English - Age
- B) Age - Department
- C) Age - English
- D) Department - English
- E) Department - Age

$$\text{Error Rate} = (\text{FP} + \text{FN}) / (\text{P} + \text{N})$$

108)

```
a  b  <-- classified as
93 33 | a = True
7 218 | b = False
```

Verilen Confusion Matrix için **Error Rate** değerini bulunuz.

- A) 9.4%
- B) 73.5%
- C) 37.9%
- D) 11.4% (Formülü uygula yüzdelik olarak göstermek için 100'le çarp)
- E) 2%

109)

```
a  b  <-- classified as
93 33 | a = True
7 218 | b = False
```

Verilen Confusion Matrix için **F-Score** değerini bulunuz.

- A) 82.3% (Formülü uygula yüzdelik olarak göstermek için 100'le çarp)
- B) 77.9%
- C) 77.5%
- D) 80.8%
- E) 79.8%

110)

| TUPLE | X | Y | CLASS |
|-------|---|----|-------|
| i | 1 | 5 | GOOD |
| ii | 2 | 10 | BAD |
| iii | 4 | 10 | BAD |
| iv | 3 | -2 | GOOD |
| v | 1 | -4 | BAD |

| | | | |
|----------------|------------|------------|----------------|
| Tuple : | X:1 | Y:7 | Class=? |
|----------------|------------|------------|----------------|

Yukarıdaki veri setine göre yeni gelen Tuple'ın sınıfını k-en yakın komşu algoritması ve k=3'e göre bulmak istiyoruz. (Uzaklık ölçütü: Manhattan; Normalizasyon: Yok) Ağırlık oylama: Yok

Verilen Tuple'a en yakın **2.nokta ve uzaklığı** hangisidir ?

- A) iii, 6
- B) ii, kök10
- C) iv, 4
- D) iv, 6
- E) ii, 4 +1

111)

| TUPLE | X | Y | CLASS |
|-------|---|----|-------|
| i | 1 | 5 | GOOD |
| ii | 2 | 10 | BAD |
| iii | 4 | 10 | BAD |
| iv | 3 | -2 | GOOD |
| v | 1 | -4 | BAD |

| | | | |
|----------------|------------|------------|----------------|
| Tuple : | X:1 | Y:7 | Class=? |
|----------------|------------|------------|----------------|

Yukarıdaki veri setine göre yeni gelen Tuple'ın sınıfını k-en yakın komşu algoritmasına göre bulmak istiyoruz. Ağırlık oylama: Yok

k'nın hangi değerleri için tahmin edilen sınıf değeri kesinlikle GOOD olur ?

- A) k=1,5 emin değilim
- B) k=1,3,4,5
- C) k=4,5
- D) k=1,2,4,5
- E) k=1,4,5

112) A:{15,25,26,29,29,29,32,32,34,36,39,43,49,62,99}, bu veri seti için herhangi bir smoothing metodu, bin_size=3 parametresi ile uygulandığında hangisi ilk "bin" **olamaz**?

- A)15,22,26
- B)15,26,26
- C)22,22,22
- D)Hepsi olabilir (All are possible)
- E)25,25,25

113) Öğrencilerin ağırlık değerlerinin bulunduğu bir sütunda (Weight) eksik veriler bulunmaktadır. Eksik değerler yerine hangisini yazmak **en kötü tercih** olur?

- A)Midrange(Weight)
- B)0
- C)Mean(Weight)
- D)Median(Weight)
- E)Unknown

114)

T100: <u1, u2, u3, u4, u5>
T200: <u1, u2, u3>
T300: <u2, u3>
T400: <u1, u3>
T500: <u2, u3, u5>
min_sup:2; min_conf:50%

Aşağıda verilen 3-itemset'ler içinde **EN GÜÇLÜ** birliktelik kuralı hangisidir?

(Which one is the STRONGEST association rule among 3-itemsets below?)

- A) u1,u3->u2

- B) $u_3 \rightarrow u_1, u_2$
 C) $u_1 \rightarrow u_2, u_3$
 D) $u_2, u_3 \rightarrow u_1$
 E) $u_1, u_2 \rightarrow u_3$

115)

| TUPLE | X | Y | CLASS |
|-------|---|----|-------|
| i | 1 | 5 | GOOD |
| ii | 2 | 10 | BAD |
| iii | 4 | 10 | BAD |
| iv | 3 | -2 | GOOD |
| v | 1 | -4 | BAD |

| | | | |
|---------|-----|-----|---------|
| Tuple i | X:1 | Y:7 | Class=? |
|---------|-----|-----|---------|

Yukarıdaki soruya göre yeni gelen Tuple'ın sınıfını k-en-yakın komşu algoritması ile bulmak istiyoruz. (Uzaklık ölçütü: Manhattan; Normalizasyon : Yok)

Ağırlıklı oylama: VAR

k'nın hangi değerleri için tahmin edilen sınıf değeri kesinlikle BAD olur?

According to the above dataset, we want to determine the class of the new tuple using k-nearest neighbor classifier. (Distance metric: Manhattan; Normalisation: No)

Weighted Voting: YES

For what k values, the predicted class would exactly be BAD?

- A) $k=2,3,4$
 B) $k=1,2,3$
 C) $k=3$
 D) Hiçbiri (None) emin değilim en mantıklı seçenek bu gibi geldi
 E) $k=2,3$

116)

| Departmer | English | Age | SUCCESS |
|-----------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |

Class Label : SUCCESS

X:{computer,good,youth} kaydının Naive Bayes sınıflandırıcı ile sınıflandırılması esnasında **P(Youth|No)** olasılığını bulunuz.

Find **P(Youth|No)** probability, while classifying tuple X:{computer,good,youth} with Naive Bayes Classifier.

A) 2/3

B) 1

C) 2/5

D) 1/3

E) 3/5

117)

| TUPLE | X | Y | CLASS |
|-------|---|----|-------|
| i | 1 | 5 | GOOD |
| ii | 2 | 10 | BAD |
| iii | 4 | 10 | BAD |
| iv | 3 | -2 | GOOD |
| v | 1 | -4 | BAD |

Tuple : X:1 Y:7 Class=?

Yukarıdaki soruya göre yeni gelen Tuple'ın sınıfını k-en yakın komşu algoritması ve k=3'e göre bulmak istiyoruz.(Uzaklık ölçütü :Manhattan; Normalizasyon: Yok)

Ağırlıklı oylama: Yok

Verilen Tuple'a en yakın 1.nokta ve uzaklığı hangisidir?

According to the above dataset, we want to determine the class of the new tuple using k-nearest neighbor classifier, with k=3.(Distance metric:Manhattan; Normalisation: No)

Weighted Voting: No

Which point is the 1st closest to the given tuple, and its distance?

A) iii, 1

B) i, $\sqrt{10}$

C) i, 2

D) iv, 2

E)iii,1

118)



Yukarıdaki karar ağacı için budama eşiği $th=0.15$ ile budama işlemi tamamlandığında **A4** düğümünün son durumu ne olur ?

- A) Hiçbiri(None)
- B) Class A
- C) Class B
- D) Bir üst düğüme kadar budanacağı için silinir.
- E) Aynen korunur.(It holds as the same)

119) Hangisi bir "Transactional Database" örneği **DEĞİLDİR**?

(Which one is **NOT** a transactional database example?)

- A) Alışveriş sitesindeki tıklamalar (Clicks on a e-commerce website)
- B) Uçuş rezervasyonları (Flight booking)
- C) Market satış bilgileri (Market sales data)
- D) Alışveriş sitesindeki sepete ekleme işlemleri (Add-to-basket actions on a webpage)
- E) Alışveriş sitesine kayıt olma bilgileri(Sign-in data in an e-commerce website)

120)



Yukarıdaki karar ağacı için budama eşiği $th=0.15$ ile budama işlemi tamamlandığında **A5** düğümünün son durumu ne olur ?

- A) Hiçbiri(None)
- B) Class A
- C) Class B
- D) Bir üst düğüme kadar budanacağı için silinir.
- E) Aynen korunur.(It holds as the same)

121)

T100:<u1,u2,u3,u4,u5>

T200:<u1,u2,u3>

T300:<u2,u3>

T400:<u1,u3>

T500<u2,u3,u5>

----- min_sup:2 ; min_conf: 50% -----

Hangisi **C2** seti içerisinde **yer almaz**?

Which one is **NOT** listed in C2 set?

A)<u1,u2>

B)<u3,u4>

C)<u1,u3>

D)<u2,u5>

E)<u1,u5>

122) Hangi veri daha GÜVENİLİRDİR?

Which data is more RELIABLE?

A Forum.donanimhaber.com ürün başlıkları

B) n11.com ürün yorumları

C) Sikayetvar.com kayıtları

D) EFS Tur Otel yorumları

E) Hepsiburada.com ürün yorumları

123) Aşağıdakilerden hangisi Kümelemede Density Based Methods'a ait özelliklerden

DEĞİLDİR? (Soru tam okunmuyor elimden geldiğince yazmaya çalıştım)

Which of the following is **NOT** a property of Density Based Methods in Clustering?

A) Can find arbitrarily shaped clusters

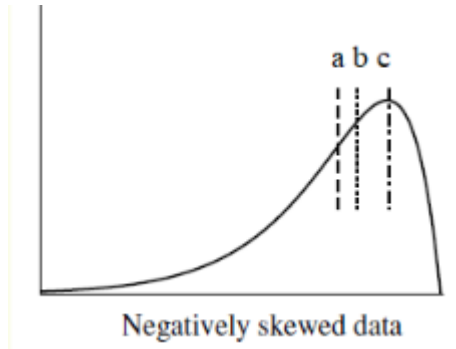
B) May filter out outliers

C) Clustering is a hierarchical decomposition(i.e, multiple levels)

D) Cluster density, Each point must have a minimum number of points within its "neighbourhood"

E) Clusters are dense regions of objects in space that are separated by low density region

124)



Yukarıdaki şekilde a nedir?

(What is a in the above graph?)

A) Median

B) Mode

C) Mean

125)

| Departmer | English | Age | SUCCESS |
|-----------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |

Class Label : SUCCESS

X:{computer,good,youth} kaydının Naive Bayes sınıflandırıcı ile sınıflandırılması esnasında **P(Youth|Yes)** olasılığını bulunuz.

Find **P(Youth|Yes)** probability, while classifying tuple X:{computer,good,youth} with Naive

Bayes Classifier.

A) 3/7

B) 2/7

C) 1/7

D) 2/5

E) 3/5

cevap 3 / 5 olmaz mı ? 5 youth var 3 tanesi Yes

olmaz(kaç tane yes olan youth varsa paya, toplam yes sayısı paydaya yazılıyor)

126)

| Departmer | English | Age | SUCCESS |
|-----------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |

X:{computer,good,youth} kaydının Naive Bayes sınıflandırıcı ile sınıflandırılması esnasında **P(Good|No)** olasılığını bulunuz.

(Sıfır değer durumunda **Laplace correction** uygulayınız)

Find **P(Good|No)** probability, while classifying tuple X:{computer,good,youth} with Naive Bayes Classifier.

A) 1/5

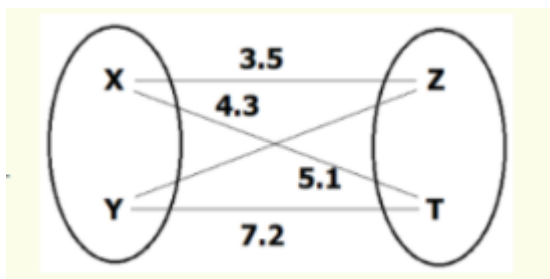
B) 1/4

C) 1/3

D) 1/7

E) 1/6

127)



Hiyerarşik kümelemede **mean distance** kullanılacaksa, yukarıdaki X-Y ve Z-T kümeleri arasındaki uzaklık ne olur?

(For hierarchical clustering, if **mean distance** metric is used, what will the distance between

X-Y and Z-T clusters be?)

A) Bu veriler ile hesaplanamaz (Cannot be calculated with this information) Neden?

B) 5.025

C) 3.5

D) 5.1

E) 7.2

128) Aşağıdaki değerleri alan bir niteliğin türü nedir?

Which type of attribute is the following?

Values: {0-Teacher; 1-Student; 2-Technical Staff; 3-Administration}

A) Nominal

B) Numeric

C) Seminal

D) Discrete

E) Ordinal

129)

| TUPLE | X | Y | CLASS |
|-------|---|----|-------|
| i | 1 | 5 | GOOD |
| ii | 2 | 10 | BAD |
| iii | 4 | 10 | BAD |
| iv | 3 | -2 | GOOD |
| v | 1 | -4 | BAD |

| | | | |
|---------|-----|-----|---------|
| Tuple : | X:1 | Y:7 | Class=? |
|---------|-----|-----|---------|

Yukarıdaki soruya göre yeni gelen Tuple'ın sınıfını k-en-yakın komşu algoritması ve k=3'e göre bulmak istiyoruz. (Uzaklık ölçütü: Manhattan; Normalizasyon : Yok)

Ağırlıklı oylama: VAR

Verilen Tuple'a en yakın 3.nokta ve ağırlıklı uzaklığı hangisidir?

According to the above dataset, we want to determine the class of the new tuple using k-nearest neighbor classifier. (Distance metric: Manhattan; Normalisation: No)

Weighted Voting: YES

Which point is the 3rd closest to the given tuple, and its weighted distance?

A) iv, 0.09

B) i, 0.25

C) iii, 0.17

D) iii, 0.028

E) ii, 0.0625

131)

| Points | X | Y | Initial Clusters |
|--------|----|----|------------------|
| i | 3 | 6 | C1 |
| ii | 5 | 8 | C1 |
| iii | 1 | 8 | C2 |
| iv | 3 | 2 | C2 |
| v | -1 | -1 | C2 |

k-means kümeleme yöntemi,öklit uzaklık ölçütü,ve tabloda belirtilen başlangıç küme atamaları ile çalıştırıldığında ilk aşamada “v” noktasının küme merkezleri M1 ve M2’ye uzaklıkları kaçtır?

A) distance to M1:13,distance to M2:6

B)distance to M1: $\sqrt{6}$,distance to M2: $\sqrt{13}$

C) distance to M1: $\sqrt{89}$,distance to M2: $\sqrt{20} + 1$

D) distance to M1: $\sqrt{89}$,distance to M2:20

E)distance to M1:89,distance to M2:20

Çözüm

$m_1 = (x_1+x_2)/2$, $(y_1+y_2)/2 = (3+5)/2$, $(6+8)/2 = (4,7)$

$v = (-1,-1)$

m_1-v arasındaki uzaklık = $\text{kok}(4-(-1))^2 + (7-(-1))^2) = \text{kok}(25+64) = \text{kok}89$

$m_2 = ((1+3-1)/3, (8+2-1)/3) = (1,3)$

m_2-v arasındaki uzaklık = $\text{kok}(1-(-1))^2 + 3-(-1))^2) = \text{kok}(4+14) = \text{kok}20$

132)

| Points | X | Y | Initial Clusters |
|--------|----|----|------------------|
| i | 3 | 6 | C1 |
| ii | 5 | 8 | C1 |
| iii | 1 | 8 | C2 |
| iv | 3 | 2 | C2 |
| v | -1 | -1 | C2 |

k-means kümeleme yöntemi,öklit uzaklık ölçütü,ve tabloda belirtilen başlangıç küme atamaları ile çalıştırıldığında son küme atamaları nasıl olur?

A) (i,iv,v) and (ii,iii)

B)(i,iv,v) and (ii,iii)

C) (i,ii,iii) and (iv,v)

D)(i,iii) and (ii,iv,v)

E)(i,ii) and (iii,iv,v)

133)

T100: <u1, u2, u3, u4, u5>

T200: <u1, u2, u3 >

T300: <u2, u3>

T400: <u1, u3>

T500: <u2, u3, u5>

----- min_sup : 2 ; min_conf : 50% -----

Hangi 1-itemset en GÜÇLÜSÜDÜR?

Which 1-itemset is the STRONGEST one?

A)u2

B)u3

C)u1

D)u5

E)u4

134)

| Departmen | English | Age | SUCCESS |
|-----------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |

Class Label : SUCCESS

↓ Entropy Table (For.Ex: {Yes, Yes, Yes, No} -> 1 | 3)

| Division | Entropy | Div. | Entr. | Div. | Entr. |
|---------------|---------|-------|-------|-------|-------|
| 1 1 (3 3) | 1 | 1 3 | 0.811 | 1 5 | 0.65 |
| 1 2 (2 4) | 0.918 | 3 4 | 0.985 | 1 6 | 0.592 |
| 3 2 (6 4) | 0.971 | 3 5 | 0.954 | 2 5 | 0.863 |
| 1 4 (2 8) | 0.722 | 3 7 | 0.881 | | |

Yukarıdaki veri setinde ID3 algoritması için kök düğümün belirlendiği ilk aşamada Gain(English) değeri kaçtır?

(For the above data set, what is Gain(English) value for the first stage that root node is determined?)

ilk aşamada : Gain(D)=0.395
 Gain(Eng)=0.12
 Gain(Age)=0.071

A)0,761

B)0,120 +1

C)0,486

D)0,395

E)0,810

135)

| Department | English | Age | SUCCESS |
|------------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |

Class Label : SUCCESS

↓ Entropy Table (For.Ex: {Yes, Yes, Yes, No} -> 1 | 3)

| Division | Entropy | Div. | Entr. | Div. | Entr. |
|---------------|---------|-------|-------|-------|-------|
| 1 1 (3 3) | 1 | 1 3 | 0.811 | 1 5 | 0.65 |
| 1 2 (2 4) | 0.918 | 3 4 | 0.985 | 1 6 | 0.592 |
| 3 2 (6 4) | 0.971 | 3 5 | 0.954 | 2 5 | 0.863 |
| 1 4 (2 8) | 0.722 | 3 7 | 0.881 | | |

Yukarıdaki veri setinde ID3 algoritması için kök düğümün belirlendiği ilk aşamada

Gain(Age) değeri kaçtır?

(For the above data set, what is Gain(Age) value for the first stage that root node is determined?)

A)0,071

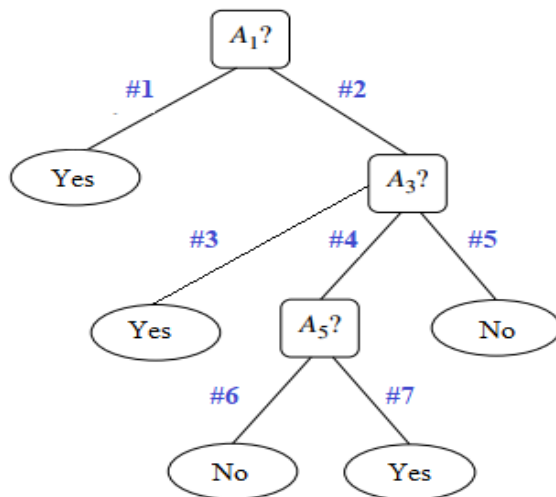
B)0,761

C)0,810

D)0,881

E)0,120

136)



| Department | English | Age | SUCCESS |
|------------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |

Class Label : SUCCESS

↓ Entropy Table (For.Ex: {Yes, Yes, Yes, No} -> 1|3)

| Division | Entropy | Div. | Entr. | Div. | Entr. |
|---------------|---------|-------|-------|-------|-------|
| 1 1 (3 3) | 1 | 1 3 | 0.811 | 1 5 | 0.65 |
| 1 2 (2 4) | 0.918 | 3 4 | 0.985 | 1 6 | 0.592 |
| 3 2 (6 4) | 0.971 | 3 5 | 0.954 | 2 5 | 0.863 |
| 1 4 (2 8) | 0.722 | 3 7 | 0.881 | | |

ID3 algoritması ile yukarıdaki veri setinden elde edilen karar ağacında #1 ve #2 dalları sırasıyla hangi değerlerden oluşmaktadır?

(For the decision tree derived from the above data set with ID3 algorithm, which values do the

branches #1 and #2 correspond to?)

- A)Other - Computer
- B)Fair-Good
- C)Computer-Other
- D)Youth-Senior
- E)Good-Fair

$a_1 = \text{Gain}(\text{department}) = 0,395$ #1= other, #2= computer

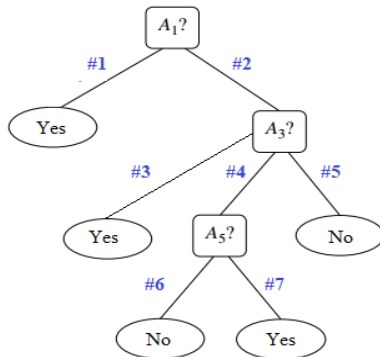
$\text{Gain}(\text{English}) = 0,12$ $A_3 = \text{English}$ #3=good #4=fair #5=bad

#7= middle

#8= senior, youth

$\text{Gain}(\text{Age}) = 0,071$

137)



| Department | English | Age | SUCCESS |
|------------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |

Class Label : SUCCESS

↓ Entropy Table (For.Ex: {Yes, Yes, Yes, No} -> 1|3)

| Division | Entropy | Div. | Entr. | Div. | Entr. |
|---------------|---------|-------|-------|-------|-------|
| 1 1 (3 3) | 1 | 1 3 | 0.811 | 1 5 | 0.65 |
| 1 2 (2 4) | 0.918 | 3 4 | 0.985 | 1 6 | 0.592 |
| 3 2 (6 4) | 0.971 | 3 5 | 0.954 | 2 5 | 0.863 |
| 1 4 (2 8) | 0.722 | 3 7 | 0.881 | | |

ID3 algoritması ile yukarıdaki veri setinden elde edilen karar ağacında A5 ve #6 düğüm-dal ikilisi sırasıyla hangi niteliklerden oluşmaktadır?

- A)Age-Senior
- B)English-Good
- C)Age-Youth
- D)Department-Other
- E)English-Fair

138)

| Department | English | Age | SUCCESS |
|------------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |

Class Label : SUCCESS

X: {computer, good, youth} kaydının Naive Bayes sınıflandırıcı ile sınıflandırılması esnasında $P(\text{Computer}|\text{No})$ olasılığını bulunuz.

- A) 2/5
- B) 2/3
- C) 3/5
- D) 1/4**
- E) 1/3

140)

| Department | English | Age | SUCCESS |
|------------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |

Class Label : SUCCESS

X: {computer, good, youth} kaydının Naive Bayes sınıflandırıcı ile sınıflandırılması esnasında $P(\text{Yes}|X)$ olasılığını bulunuz

- A) 0,012
- B) 0,024**
- C) 0,120
- D) 0,048
- E) 0,036

141-)

| | P1 | P2 | P3 | P4 |
|----|------|------|-----|----|
| P1 | | | | |
| P2 | 3.65 | | | |
| P3 | 2.45 | 5.55 | | |
| P4 | 3.05 | 2.9 | 3.8 | |

yukarıda 4 nokta arası uzaklık matrisinin verildiği veri seti için maximum distance ölçütü kullanılarak hiyerarşik kümeleme yapılacaktır.

Dendrogramın 2. aşaması hangi uzaklık değerine karşılık gelir?

A-)2.45

B-) 3.8

C-) 2.9

D-) 3.05

E-)

Çözüm

| | p1-p3 | p2 | p4 |
|-------|-------|-----|----|
| p1-p3 | | | |
| p2 | 3,65 | | |
| p4 | 3,05 | 2,9 | |

142)

```
a    b    <-- classified as
93  33 |    a = True
7  218 |    b = False
```

Verilen Confusion Matrix için Precision değerini bulunuz

A)86,9%

B)69,9%

C)29,9%

D)93%

E)73,8%

143)

```
a  b  <-- classified as
93 33 |  a = True
 7 218 | b = False
```

Verilen Confusion Matrix için TP değeri nedir?

(What is the TP value for the given confusion matrix?)

A)33

B)7

C)93

D)126

E)218

144)

| TUPLE | X | Y | CLASS |
|-------|---|----|-------|
| i | 1 | 5 | GOOD |
| ii | 2 | 10 | BAD |
| iii | 4 | 10 | BAD |
| iv | 3 | -2 | GOOD |
| v | 1 | -4 | BAD |

| | | | |
|----------------|------------|------------|----------------|
| Tuple : | X:1 | Y:7 | Class=? |
|----------------|------------|------------|----------------|

Yukarıdaki veri setine göre yeni gelen Tuple'ın sınıfını k-en yakın komşu algoritması ve k=3'e göre bulmak istiyoruz. (Uzaklık ölçütü : Manhattan; Normalizasyon : Yok)

Ağırlıklı oylama: Yok

Verilen Tuple'ın sınıfı ve en yakın ilk 3 noktayı sıralı olarak aşağıdaki seçeneklerde bularak işaretleyiniz.

- A)BAD,ii,i,iii
 B)GOOD,i,iii,iv
 C)BAD,i,ii,iii
 D)BAD,ii,iii,iv
 E)GOOD,i,iii,v

145)

| TUPLE | X | Y | CLASS |
|-------|---|----|-------|
| i | 1 | 5 | GOOD |
| ii | 2 | 10 | BAD |
| iii | 4 | 10 | BAD |
| iv | 3 | -2 | GOOD |
| v | 1 | -4 | BAD |

| | | | |
|----------------|------------|------------|----------------|
| Tuple : | X:1 | Y:7 | Class=? |
|----------------|------------|------------|----------------|

Yukarıdaki soruya göre yeni gelen Tuple'ın sınıfını k-en yakın komşu algoritması ve k=3'e göre bulmak istiyoruz. (Uzaklık ölçütü : Manhattan; Normizasyon : Yok)

Ağırlıklı oylama: VAR

Verilen Tuple'ın sınıfı ve toplam ağırlıklı uzaklıkları seçeneklerde bularak işaretleyiniz

- A)GOOD (0.5) - Predicted Class

BAD (0.42)

B)GOOD (0.258) - Predicted Class

BAD (0.0625)

C)GOOD (0.25) - Predicted Class

BAD (0.0905)

D)BAD (0.42) - Predicted Class

GOOD (0.25)

E)BAD (0.0905) - Predicted Class

GOOD (0.008)

146)

| Points | X | Y | Initial Clusters |
|--------|----|----|------------------|
| i | 3 | 6 | C1 |
| ii | 5 | 8 | C1 |
| iii | 1 | 8 | C2 |
| iv | 3 | 2 | C2 |
| v | -1 | -1 | C2 |

k-means kümeleme yöntemi, öklit uzaklık ölçütü, ve tabloda belirtilen başlangıç küme atamaları ile çalıştırıldığında ilk aşamada küme ataması değişecek olan nokta hangisidir, küme merkezleri M1 ve M2'ye uzaklıkları kaçtır?

A)ii -> distance to M1: $\sqrt{2}$; distance to M2: $\sqrt{41}$

B)iii -> distance to M1: $\sqrt{10}$; distance to M2: 5

C)ii -> distance to M1: $\sqrt{2}$; distance to M2: 6

D)iv -> distance to M1: $\sqrt{26}$; distance to M2: $\sqrt{5}$

E)iii -> distance to M1: 10 ; distance to M2: 25

148)

Bir verisetinde 500 satır alışveriş verisi içinde 250 tanesi makarna, 150 tanesi ketçap, 100 tanesi de her iki ürünü içermektedir. min_support=15% ve min_confidence=30% için, All_Confidence(Makarna, Ketçap) değerini bulunuz.

In a dataset of 500 rows, 250 rows include spaghetti, 150 rows include ketchup, while 100 rows include both of them. If min_support=15% and min_confidence=30%, find All_Confidence(Spaghetti, Ketchup) value

A)0,75

B)0,53

C)0,20

D)0,40

$all_conf(A,B)=\frac{sup(A \cup B)}{\max\{sup(A),sup(B)\}}$

E)0,66

149)

T100: <u1, u2, u3, u4, u5>

T200: <u1, u2, u3 >

T300: <u2, u3>

T400: <u1, u3>

T500: <u2, u3, u5>

----- min_sup : 2 ; min_conf : 50% -----

Hangisi C2 seti içinde yer alır? (Yapan olursa çözümünü de ekleyebilirmisiniz?)

Which one IS listed in C2 set?

A)<u2,u4>

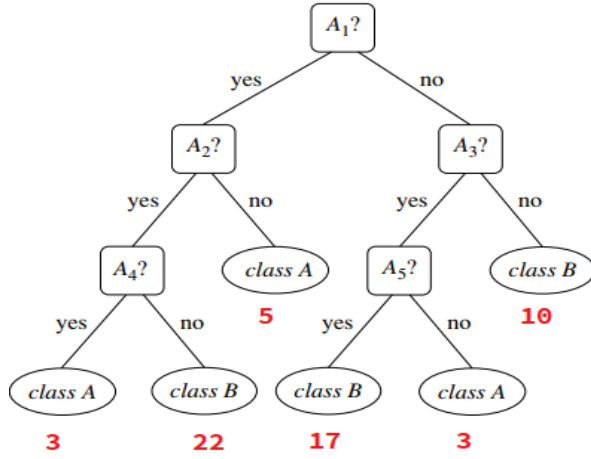
B)<u3,u4>

C)<u1,u4>

D)<u1,u5>

E)<u4,u5>

150)



Yukarıdaki karar ağacı için budama eşiği $th=0.15$ ile budama işlemi tamamlandığında A3 düğümünün son durumu ne olur? **(Çözümünü de ekleyebilir misiniz)**

A)Hiçbiri (None)

B)Class B

C)Bir üst düğüme kadar budanacağı için silinir (Is deleted, since pruning is done towards the upper node)

D)Aynen korunur (It holds as the same)

E)Class A

A3:

önce A5 bakılmalı

classB:17/20:0.85

classA:3/20 :0.15

çıkan değer eşik değerine eşit olduğu için A5 için budama yapılmaz.

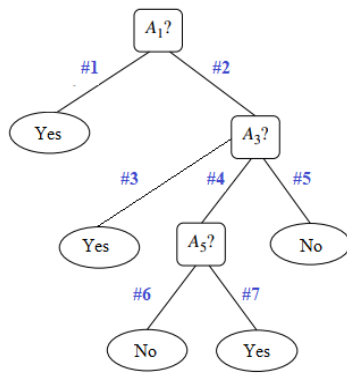
A3 için:

classB:27/30

classA:3/30:0.1

class A nın değeri eşik değerinden küçük olduğu için A3 te budama yapılır.

A3 yerine class B yazılır.(sayısı fazla olduğu için)



| Department | English | Age | SUCCESS |
|------------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |

Class Label : SUCCESS

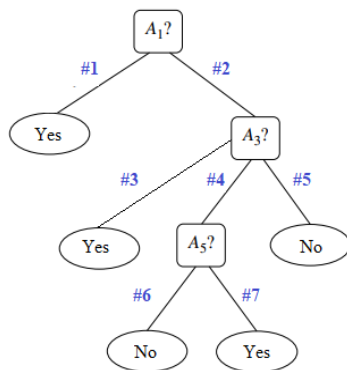
↓ Entropy Table (For.Ex: {Yes, Yes, Yes, No} -> 1 | 3)

| Division | Entropy | Div. | Entr. | Div. | Entr. |
|---------------|---------|-------|-------|-------|-------|
| 1 1 (3 3) | 1 | 1 3 | 0.811 | 1 5 | 0.65 |
| 1 2 (2 4) | 0.918 | 3 4 | 0.985 | 1 6 | 0.592 |
| 3 2 (6 4) | 0.971 | 3 5 | 0.954 | 2 5 | 0.863 |
| 1 4 (2 8) | 0.722 | 3 7 | 0.881 | | |

ID3 algoritması ile yukarıdaki veri setinden elde edilen karar ağacında #5 ve #6 dalları sırasıyla hangi değerlerden oluşmaktadır?

- A) Youth - Fair
- B) Middle - Good
- C) Other - Middle
- D) Computer - Fair
- E) Middle - Computer

152)



| Department | English | Age | SUCCESS |
|------------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |

Class Label : SUCCESS

↓ Entropy Table (For.Ex: {Yes, Yes, Yes, No} -> 1 | 3)

| Division | Entropy | Div. | Entr. | Div. | Entr. |
|---------------|---------|-------|-------|-------|-------|
| 1 1 (3 3) | 1 | 1 3 | 0.811 | 1 5 | 0.65 |
| 1 2 (2 4) | 0.918 | 3 4 | 0.985 | 1 6 | 0.592 |
| 3 2 (6 4) | 0.971 | 3 5 | 0.954 | 2 5 | 0.863 |
| 1 4 (2 8) | 0.722 | 3 7 | 0.881 | | |

ID3 algoritması ile yukarıdaki veri setinden elde edilen karar ağacında #4 ve #7 dalları sırasıyla hangi değerlerden oluşmaktadır?

- A) Computer - Fair
- B) Senior - Good
- C) Youth - Good
- D) Senior - Fair
- E) Youth - Fair

153)

| Department | English | Age | SUCCESS |
|------------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |

Class Label : SUCCESS

X: {computer, good, youth} kaydının Naive Bayes sınıflandırıcı ile sınıflandırılması esnasında $P(\text{Computer}|\text{Yes})$ olasılığını bulunuz

A)2/7

B)2/5

C)3/7

D)1

E)3/5

| | yes | no |
|-------|------|------|
| yes | TP | FN |
| no | FP | TN |
| Total | P' | N' |

154)

```

a    b    <-- classified as
93  33 |    a = True
 7 218 |    b = False

```

Verilen Confusion Matrix için FN değeri nedir?

A)251

B)33

C)218

D)93

E)7

155)

```
a    b    <-- classified as
93   33 |    a = True
 7  218 |    b = False
```

P=tp+fn

Verilen Confusion Matrix için P değeri nedir? (Positive tuples)

- A)133
- B)100
- C)218
- D)93
- E)126

156)

```
a    b    <-- classified as
93   33 |    a = True
 7  218 |    b = False
```

Verilen confusion matrix için hangi sınıfın sınıflandırma performansı daha iyidir, Yüzdesi nedir?

- A)Negative Class, 84.5%
- B)Negative Class, 96.9%
- C)Negative Class, 86.9%
- D)Positive Class, 73.8%
- E)Positive Class, 93% neden acaba !

157)

| <i>TUPLE</i> | <i>X</i> | <i>Y</i> | <i>CLASS</i> |
|--------------|----------|----------|--------------|
| <i>i</i> | 1 | 5 | GOOD |
| <i>ii</i> | 2 | 10 | BAD |
| <i>iii</i> | 4 | 10 | BAD |
| <i>iv</i> | 3 | -2 | GOOD |
| <i>v</i> | 1 | -4 | BAD |

| | | | |
|-----------------------|------------|------------|----------------|
| <i>Tuple :</i> | <i>X:1</i> | <i>Y:7</i> | <i>Class=?</i> |
|-----------------------|------------|------------|----------------|

Yukarıdaki veri setine göre yeni gelen Tuple'ın sınıfını k-en yakın komşu algoritması ve k=3'e göre bulmak istiyoruz. (Uzaklık ölçütü : Manhattan; Normalizasyon : Yok)

Ağırlıklı oylama: Yok

Verilen Tuple'a en yakın 3. nokta ve uzaklığı hangisidir?

- A)i, 4
- B)iv, 6
- C)v,11
- D)ii,3
- E)iii,6

| <i>TUPLE</i> | <i>X</i> | <i>Y</i> | <i>CLASS</i> |
|--------------|----------|----------|--------------|
| <i>i</i> | 1 | 5 | GOOD |
| <i>ii</i> | 2 | 10 | BAD |
| <i>iii</i> | 4 | 10 | BAD |
| <i>iv</i> | 3 | -2 | GOOD |
| <i>v</i> | 1 | -4 | BAD |

| | | | |
|----------------|------------|------------|----------------|
| Tuple : | <i>X:1</i> | <i>Y:7</i> | <i>Class=?</i> |
|----------------|------------|------------|----------------|

Yukarıdaki soruya göre yeni gelen Tuple'ın sınıfını k-en yakın komşu algoritması ve k=3'e göre bulmak istiyoruz. (Uzaklık ölçütü : Manhattan; Normizasyon : Yok)

Ağırlıklı oylama: VAR

Verilen Tuple'a en yakın 2. nokta ve ağırlıklı uzaklığı hangisidir?

- A)iv, 0.09
- B)i, 0.5
- C)iii, 0.028
- D)ii, 0.25
- E)ii, 0.0625**

| <i>TUPLE</i> | <i>X</i> | <i>Y</i> | <i>CLASS</i> |
|--------------|----------|----------|--------------|
| <i>i</i> | 1 | 5 | GOOD |
| <i>ii</i> | 2 | 10 | BAD |
| <i>iii</i> | 4 | 10 | BAD |
| <i>iv</i> | 3 | -2 | GOOD |
| <i>v</i> | 1 | -4 | BAD |

| | | | |
|-----------------------|-------------------|-------------------|-----------------------|
| <i>Tuple :</i> | <i>X:1</i> | <i>Y:7</i> | <i>Class=?</i> |
|-----------------------|-------------------|-------------------|-----------------------|

Yukarıdaki soruya göre yeni gelen Tuple'ın sınıfını k-en yakın komşu algoritması ile bulmak istiyoruz. (Uzaklık ölçütü : Manhattan; Normizasyon : Yok)

Ağırlıklı oylama: VAR

k=4 için hangi nokta veri setinden silinirse sonuç BAD olur?

- A)v
- B)iii
- C)iv
- D)i**
- E)ii

160)

Aşağıdakilerden hangisi Kümelemede Partitioning Methods'a ait özelliklerden DEĞİLDİR?

- A)Cannot correct erroneous merges or splits**
- B)May use mean or medoid (etc.) to represent cluster center
- C)Effective for small- to medium-size data sets
- D)Find mutually exclusive clusters of spherical shape
- E)Distance-based

| Method | General Characteristics |
|-----------------------|---|
| Partitioning methods | <ul style="list-style-type: none"> Find mutually exclusive clusters of spherical shape Distance-based May use mean or medoid (etc.) to represent cluster center Effective for small- to medium-size data sets |
| Hierarchical methods | <ul style="list-style-type: none"> Clustering is a hierarchical decomposition (i.e., multiple levels) Cannot correct erroneous merges or splits May incorporate other techniques like microclustering or consider object "linkages" |
| Density-based methods | <ul style="list-style-type: none"> Can find arbitrarily shaped clusters Clusters are dense regions of objects in space that are separated by low-density regions Cluster density: Each point must have a minimum number of points within its "neighborhood" May filter out outliers |
| Grid-based methods | <ul style="list-style-type: none"> Use a multiresolution grid data structure Fast processing time (typically independent of the number of data objects, yet dependent on grid size) |

161)

| Points | X | Y | Initial Clusters |
|--------|----|----|------------------|
| i | 3 | 6 | C1 |
| ii | 5 | 8 | C1 |
| iii | 1 | 8 | C2 |
| iv | 3 | 2 | C2 |
| v | -1 | -1 | C2 |

k-means kümeleme yöntemi, öklit uzaklık ölçütü, ve tabloda belirtilen başlangıç küme atamaları ile çalıştırıldığında ilk aşamada "i" noktasının küme merkezleri M1 ve M2'ye uzaklıkları kaçtır?

- A)distance to M1: 2; distance to M2: $\sqrt{13}$
B)distance to M1: 4 ; distance to M2: 13
C)distance to M1: 2 ; distance to M2: 13
D)distance to M1: $\sqrt{2}$; distance to M2: 6
E)distance to M1: $\sqrt{2}$; distance to M2: $\sqrt{13} + 1$

162)

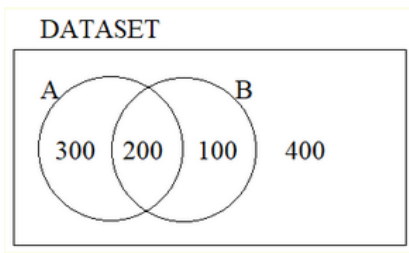
| Points | X | Y | Initial Clusters |
|--------|----|----|------------------|
| i | 3 | 6 | C1 |
| ii | 5 | 8 | C1 |
| iii | 1 | 8 | C2 |
| iv | 3 | 2 | C2 |
| v | -1 | -1 | C2 |

k-means kümeleme yöntemi, öklit uzaklık ölçütü, ve tabloda belirtilen başlangıç küme atamaları ile çalıştırıldığında 2. aşamada birinci küme merkezi M1'i bulunuz. (İlk güncellemeden sonra)

- A)M1 : (3, 11) +1

- B)M1 : (4.5, 11)
C)M1 : (4.5, 7.33)
D)M1 : (3, 7.33) +1
E)M1 : (4, 7)

163)



Conf (B \rightarrow A) = ?

- A) 0.2
B) 0.4
C) 0.33
D) 0.6
E) 0.67

164) Aşağıdaki değerleri alan bir niteliğin türü nedir?

Which type of attribute is the following?

Values : { 0-Teacher ; 1-Student ; 2-Technical Staff ; 3-Administration }

- A) Numeric
B) Ordinal
C) Nominal
D) Discrete
E) Seminal

166)

| Departmer | English | Age | SUCCESS |
|-----------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |

Class Label : SUCCESS

Üstteki veri seti için Gini(English) değerini hesaplayınız. Dallanma : [fair | good&bad]

(Calculate Gini(English) for the above data set) Branching : [fair | good&bad]

- A) 0.44
- B) 0.48
- C) 0.32
- D) 0.24
- E) 0.4

165)

| Departmer | English | Age | SUCCESS |
|-----------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |

Class Label : SUCCESS

X: {computer, good, youth} kaydının Naive Bayes sınıflandırıcı ile sınıflandırılması esnasında $P(\text{No}|\text{X})$ olasılığını bulunuz.

Find $P(\text{No}|\text{X})$ probability, while classifying tuple X: {computer, good, youth} with Naive Bayes Classifier. (Çözümü nedir? Laplace correction yapınca bu sonuçlardan herhangi biri gelmiyor?)

- A) 0.066

- B) 0.033
C) 0.0396
D) 0.055
E) 0.0495

166)

| Points | X | Y | Initial Clusters |
|--------|----|----|------------------|
| i | 3 | 6 | C1 |
| ii | 5 | 8 | C1 |
| iii | 1 | 8 | C2 |
| iv | 3 | 2 | C2 |
| v | -1 | -1 | C2 |

k-means kümeleme yöntemi, öklit uzaklık ölçütü, ve tabloda belirtilen başlangıç küme atamaları ile çalıştırıldığında 2. aşamada birinci küme merkezi M2'yi bulunuz. (İlk güncellemeden sonra)

Applying k-means clustering algorithm with euclidean distance measure and given initial cluster assignments, find second cluster center M2 at the 2nd step. (After the first update)

- A) M2 : (0.66, 0.33)
B) M2 : (2, 1)
C) M2 : (2, 0.5)
D) M2 : (1, 0.5)
E) M2 : (0.66, 0.5)

167)

| Points | X | Y | Initial Clusters |
|--------|----|----|------------------|
| i | 3 | 6 | C1 |
| ii | 5 | 8 | C1 |
| iii | 1 | 8 | C2 |
| iv | 3 | 2 | C2 |
| v | -1 | -1 | C2 |

k-means kümeleme yöntemi, öklit uzaklık ölçütü, ve tabloda belirtilen başlangıç küme atamaları ile çalıştırıldığında ilk aşamada "ii" noktasının küme merkezleri M1 ve M2'ye uzaklıkları kaçtır?

Applying k-means clustering algorithm with euclidean distance measure and given initial cluster assignments, what is the distance of point "ii" to cluster centers M1 and M2 at the first step?

A) distance to M1: 2 ; distance to M2: 9

B) distance to M1: $\sqrt{2}$; distance to M2: $\sqrt{41}$

C) distance to M1: $\sqrt{2}$; distance to M2: 41

D) distance to M1: $\sqrt{2}$; distance to M2: 3

E) distance to M1: 2 ; distance to M2: 41

168) Aşağıdakilerden hangisi Kümelemede Hierarchical Methods'a ait özelliklerden DEĞİLDİR?

Which of the following is NOT a property of Hierarchical Methods in Clustering?

A) Clustering is a hierarchical decomposition (i.e., multiple levels)

B) May consider object "linkages"

C) May use mean or medoid (etc.) to represent cluster center

D) Cannot correct erroneous merges or splits

E) May incorporate other techniques like microclustering

169)

| <i>TUPLE</i> | <i>X</i> | <i>Y</i> | <i>CLASS</i> |
|--------------|----------|----------|--------------|
| <i>i</i> | 1 | 5 | GOOD |
| <i>ii</i> | 2 | 10 | BAD |
| <i>iii</i> | 4 | 10 | BAD |
| <i>iv</i> | 3 | -2 | GOOD |
| <i>v</i> | 1 | -4 | BAD |

| | | | |
|-----------------------|------------|------------|----------------|
| <i>Tuple :</i> | <i>X:1</i> | <i>Y:7</i> | <i>Class=?</i> |
|-----------------------|------------|------------|----------------|

Yukarıdaki veri setine göre yeni gelen Tuple'ın sınıfını k-en yakın komşu algoritmasına göre bulmak istiyoruz. (Uzaklık ölçütü : Manhattan; Normizasyon : Yok)

Ağırlıklı oylama: Yok

k'nın hangi değerleri için tahmin edilen sınıf değeri kesinlikle GOOD olur?

According to the above data set, we want to determine the class of the new tuple using k-nearest neighbor classifier. (Distance metric : Manhattan; Normalization : No)

Weighted Voting : No

For what k values, the predicted class would exactly be GOOD?

- A) k=1,5
- B) k=1,3,4,5
- C) k=1,2,4,5
- D) k=4,5
- E) k=1,4,5

170)

| TUPLE | X | Y | CLASS |
|-------|---|----|-------|
| i | 1 | 5 | GOOD |
| ii | 2 | 10 | BAD |
| iii | 4 | 10 | BAD |
| iv | 3 | -2 | GOOD |
| v | 1 | -4 | BAD |

| | | | |
|----------------|------------|------------|----------------|
| Tuple : | X:1 | Y:7 | Class=? |
|----------------|------------|------------|----------------|

Yukarıdaki soruya göre yeni gelen Tuple'ın sınıfını k-en yakın komşu algoritması ve k=3'e göre bulmak istiyoruz. (Uzaklık ölçütü : Manhattan; Normalizasyon : Yok)

Ağırlıklı oylama: VAR

Verilen Tuple'ın sınıfı ve toplam ağırlıklı uzaklıkları seçeneklerde bularak işaretleyiniz.

According to the above dataset, we want to determine the class of the new tuple using k-nearest neighbor classifier, with k=3. (Distance metric : Manhattan; Normalization : No)

Weighted Voting : YES

Find the class of the given Tuple and total weighted distances from the choices below.

A) BAD (0.0905) - Predicted Class

GOOD (0.008)

B) GOOD (0.5) - Predicted Class

BAD (0.42)

C) GOOD (0.25) - Predicted Class

BAD (0.0905)

D) BAD (0.42) - Predicted Class

GOOD (0.25)

E) GOOD (0.258) - Predicted Class

BAD (0.0625)

171)

| <i>TUPLE</i> | <i>X</i> | <i>Y</i> | <i>CLASS</i> |
|--------------|----------|----------|--------------|
| <i>i</i> | 1 | 5 | GOOD |
| <i>ii</i> | 2 | 10 | BAD |
| <i>iii</i> | 4 | 10 | BAD |
| <i>iv</i> | 3 | -2 | GOOD |
| <i>v</i> | 1 | -4 | BAD |

| | | | |
|----------------|------------|------------|----------------|
| Tuple : | <i>X:1</i> | <i>Y:7</i> | <i>Class=?</i> |
|----------------|------------|------------|----------------|

Yukarıdaki soruya göre yeni gelen Tuple'ın sınıfını k-en yakın komşu algoritması ve k=3'e göre bulmak istiyoruz. (Uzaklık ölçütü : Manhattan; Normizasyon : Yok)

Ağırlıklı oylama: Yok

Verilen Tuple'a en yakın 1. nokta ve uzaklığı hangisidir?

According to the above dataset, we want to determine the class of the new tuple using k-nearest neighbor classifier, with k=3. (Distance metric : Manhattan; Normalization : No)

Weighted Voting : No

Which point is the 1st closest to the given tuple, and its distance?

- A) iii, 1
- B) ii, 2
- C) i, 2**
- D) iv, 2
- E) i, $\sqrt{10}$

172) Veri madenciliğinde niteliklere ait değerler birbirinden çok farklı değer aralıklarında ise veri üzerinde hangi işlem öncelikle düşünülmelidir?

- A) Temizleme
- B) Bütünleştirme
- C) İndirgeme**
- D) Dönüştürme
- E) Nominalleştirme**

173) Dendrogram hangisinde kullanılır?

- A) Sınıflandırma ????
- B) Birliktelik
- C) Kümeleme +2**
- D) Bayes
- E) Hiçbiri

174) Hangisi bir veri madenciliği programı değildir?

- A) Rapid Miner
- B) IKVM
- C) Weka
- D) Data Mining**
- E) Hiçbiri

175) veritabanında gürültüyü gidermek hangi aşama içerisinde yer alır?
(Which stage includes "removing noise"?)

- A) Data Integration
- B) Data Transformation
- C) Data Cleaning**
- D) Pattern Evaluation
- E) Data Selection

176) "Cadde" verisini Şehir, Ülke gibi daha üst seviye verilere dönüştürdüğümüzde hangi yöntemi uygulamış oluruz?

(Which transformation method is used if we convert "street" data into higher-level concepts, like city or country?)

- a) Concept hierarchy generation (emin değilim)
- b) smooting
- c) Normalization
- d) discretization
- e) Aggregation

177) Conf ($X \rightarrow Y, Z$) = ?

- A) $P(Z | X \cap Y)$
- B) $P(Z | X \cup Y)$
- C) $P(Y, Z | X)$
- D) $P(X | Y, Z)$
- E) $P(Y, Z \cap X)$

178) Karar ağacı tabanlı bir sınıflandırma probleminde, A niteliği aşağıdaki değerleri alıyor ise nasıl bir dallanma senaryosu uygulanmalıdır?

(In a decision tree based classification problem, if attribute A holds the following values, what type of branching scenario should be applied?)

A: {50,35,25,40,60,60,100,80,100,80}

- a) $A \leq 40?$
- b) $A \leq 60?$
- c) $A \leq 45?$
- d) $A \leq 50?$
- e) $A \leq 55?$

179)

| Department | English | Age | SUCCESS | | |
|---|---------|--------|---------|-------|-------|
| Computer | Fair | Middle | Yes | | |
| Computer | Fair | Senior | No | | |
| Computer | Good | Senior | Yes | | |
| Computer | Bad | Youth | No | | |
| Computer | Fair | Youth | No | | |
| Other | Bad | Senior | Yes | | |
| Other | Fair | Senior | Yes | | |
| Other | Bad | Youth | Yes | | |
| Other | Fair | Youth | Yes | | |
| Other | Good | Youth | Yes | | |
| Class Label : SUCCESS | | | | | |
| ↓Entropy Table (For.Ex: {Yes, Yes, Yes, No} -> 1 3) | | | | | |
| Division | Entropy | Div. | Entr. | Div. | Entr. |
| 1 1 (3 3) | 1 | 1 3 | 0.811 | 1 5 | 0.65 |
| 1 2 (2 4) | 0.918 | 3 4 | 0.985 | 1 6 | 0.592 |
| 3 2 (6 4) | 0.971 | 3 5 | 0.954 | 2 5 | 0.863 |
| 1 4 (2 8) | 0.722 | 3 7 | 0.881 | | |

Yukarıdaki veri setinde ID3 algoritması için 2. aşamada **kök düğüm**den sonraki gelen düğümü belirlerken **Gain(English)** değeri kaçtır?

(For the above data set, what is **Gain(English)** value for the **second stage** that the node after root node is determined?)

- a) 0.42
- b) 0.4
- c) 0.551
- d) 0.971
- e) 0.571

180)

| Department | English | Age | SUCCESS | | |
|---|---------|--------|---------|-------|-------|
| Computer | Fair | Middle | Yes | | |
| Computer | Fair | Senior | No | | |
| Computer | Good | Senior | Yes | | |
| Computer | Bad | Youth | No | | |
| Computer | Fair | Youth | No | | |
| Other | Bad | Senior | Yes | | |
| Other | Fair | Senior | Yes | | |
| Other | Bad | Youth | Yes | | |
| Other | Fair | Youth | Yes | | |
| Other | Good | Youth | Yes | | |
| Class Label : SUCCESS | | | | | |
| ↓Entropy Table (For.Ex: {Yes, Yes, Yes, No} -> 1 3) | | | | | |
| Division | Entropy | Div. | Entr. | Div. | Entr. |
| 1 1 (3 3) | 1 | 1 3 | 0.811 | 1 5 | 0.65 |
| 1 2 (2 4) | 0.918 | 3 4 | 0.985 | 1 6 | 0.592 |
| 3 2 (6 4) | 0.971 | 3 5 | 0.954 | 2 5 | 0.863 |
| 1 4 (2 8) | 0.722 | 3 7 | 0.881 | | |

Üstteki veri seti için **Gini(Age)** değerini hesaplayınız. Dallanma : [middle&senior | youth]

(Calculate **Gini(Age)** for the above data set) Branching : [middle&senior | youth]

- a) 0.24
- b) 0.4
- c) 0.44
- d) 0.48
- e) 0.32

181)

| Departmer | English | Age | SUCCESS |
|-----------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |

Class Label : SUCCESS

X: {computer, good, youth} kaydının Naive Bayes sınıflandırıcı ile sınıflandırılması esnasında **P(Good|Yes)** olasılığını bulunuz.

Find **P(Good|Yes)** probability, while classifying tuple X: {computer, good, youth} with Naive Bayes Classifier.

- a) 1/2
- b) 1/7
- c) 1
- d) 3/7
- e) 2/7

182)

T100: <u1, u2, u3, u4, u5>

T200: <u1, u2, u3 >

T300: <u2, u3>

T400: <u1, u3>

T500: <u2, u3, u5>

----- min_sup : 2 ; min_conf : 50% -----

Hangisi **C2** seti içinde **yer almaz**?

Which one is **NOT** listed in **C2** set?

A) <u1,u5>

B) <u3,u4>

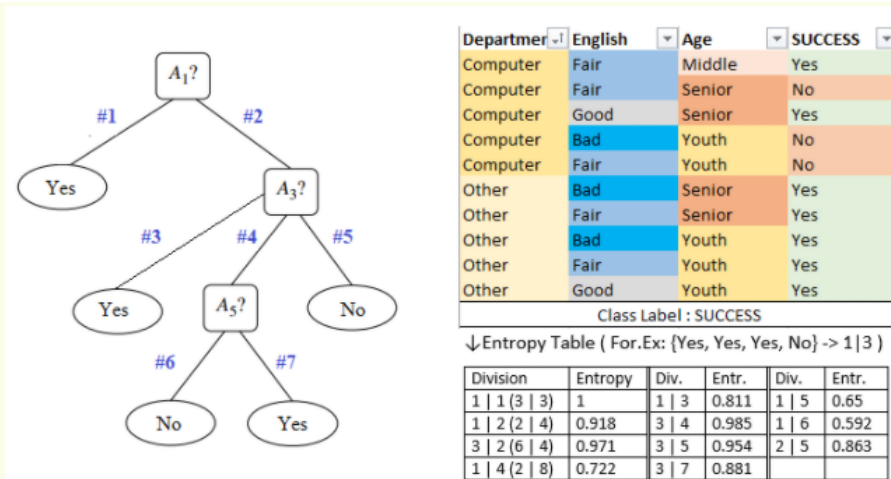
C) <u1,u3>

D) <u1,u2>

E) <u2,u5>

183)

ID3 algoritması ile yukarıdaki veri setinden elde edilen karar ağacında **#3, #4 ve #5** dalları sırasıyla hangi değerlerden oluşmaktadır ?



ID3 algoritması ile yukarıdaki veri setinden elde edilen karar ağacında **#3, #4 ve #5** dalları sırasıyla hangi değerlerden oluşmaktadır?

(For the decision tree derived from the above data set with ID3 algorithm, which values do the branches **#3, #4 and #5** correspond to?)

- ☐ A) Youth - Middle - Senior
- ☐ B) Middle - Youth - Senior
- ☐ C) Senior - Youth - Middle
- ☐ D) Senior - Middle - Youth
- ☒ E) Middle - Senior - Youth

184)

| Departmer | English | Age | SUCCESS |
|-----------------------|---------|--------|---------|
| Computer | Fair | Middle | Yes |
| Computer | Fair | Senior | No |
| Computer | Good | Senior | Yes |
| Computer | Bad | Youth | No |
| Computer | Fair | Youth | No |
| Other | Bad | Senior | Yes |
| Other | Fair | Senior | Yes |
| Other | Bad | Youth | Yes |
| Other | Fair | Youth | Yes |
| Other | Good | Youth | Yes |
| Class Label : SUCCESS | | | |

Üstteki veri seti için **Gini(Department)** değerini hesaplayınız.

(Calculate **Gini(Department)** for the above data set)

- A) 0.44
- B) 0.4
- C) 0.48
- D) 0.24**
- E) 0.16