Ad-Soyad: **No:**

Sivas Cumhuriyet Üniversitesi Mühendislik Fak. Bilgisayar Müh. Böl.

Bil4211 Veri Madenciliği 2018-Güz Ara Sınavı

1. a. Veri madenciliğinde kullanılan veriyi üreten kaynaklar nelerdir? (5 puan)

Veriyi üreten kaynakları genel olarak internet, sensörler, anketler, finans sektörü (bankalar), alışveriş sektörü, tıp, kullanıcı işlemleri gibi düşünebiliriz.

b. Veri madenciliği süreci hangi aşamalardan oluşur; bu aşamalar sonucunda elde edilen bilginin değerli sayılabilmesi için bu bilginin ne gibi özellikler taşıması gerekir? (5 puan)

Veri madenciliği; veri toplama, veri on işleme, veri madenciliği algoritmasi uygulanmasi ve bilgi elde etme aşamalrından oluşur. Elde edilen bilginin işe yarar, tahmin edeilemez ve ilginç olması gerekir.

c. Veri madenciliği günümüzde olduğu gibi gelecektede önemli bir çalışma alanı olacak mıdır? Nedenleriyle yazınız. (5 puan)

Veri çok çesitli kaynaklardan sürekli olarak elde edilir. Gelecektede bankalar, alışveriş merkezleri sensörler vb. gibi alanlar veri üretmeye devam edecektir. Aynı zamanda veri depolama teknolojisinide buna parelel olarak gelişim göstermesi beklenmektedir. Şu halde veri madenciliği veri üretimindeki artış ve veri depolama teknolojilerindeki ilerleme sayesinde gelecekte de önemli bir çalişma alanı olacaktır.

2.

Hasta Ad - Soyad	T.C. Kimlik No	Sigara Alışkanlığı	Kilo	Cinsiyet	Ailede Kanser Görülmesi	Kanser?
Hasta 1	11487612986	1	76	Kadın	1	1
Hasta 2	18934565432	1	48	Kadın		0
Hasta 3	12383470392	1	1002	Erkek		0
Hasta 4	19234335431	1	65	Erkek	0	1

Yukarıda 4 hastanın oluşturduğu bir veri seti verilmiştir. Buradaki veri madenciliği görevi bu veri setini kullanarak kişilerin kanser olup olmadığını tahmin etmektir. Bu veri setine göre aşağıdaki soruları cevaplandırınız.

a. 'T.C. Kimlik No' özelliği bu veri madenciliği görevi için gerekli midir? Çıkarılabilir mi? (5 puan)
b. 'Sigara Alışkanlığı' özelliğini bu veri setinden çıkarsak bir veri kaybı oluşur mu? Neden? (5 puan)
c. Özellik türleri bakımından bu veri setindeki en değerli özellik nedir? Açıklayınız. (5 puan)
d. Bu veri setinde gördüğünüz aksaklıklar, olumsuzluklar nelerdir? Açıklayınız. (5 puan)

a. T.C. kimlik no kanser sınıflandırmasında önemli olmadığı için bu veri madenciliği görevi için gerekli değildi? Veri setinden çıkarılabilir.
b. Sigara Alışkanlığı bütün hastalar için aynıdır, pozitiftir. O halde sınıflandırma için ayırt edici bir özellik değildir. Veri setinden çıkarmak bir veri kaybına yol açmaz.
c. Özellik türleri bakımından en değerli özellik, sayısal (hatta bölüm tipinde) bir özellik olan Kilo özelliğidir.
d. Veri seti kayıp veri içerir. Ailede kanser görülmesi özelliğinde bir cok kayip vardir. Ayrıca gürültü/tutarsızlık mevcuttur. Hasta 3'ün ağırlığı 1002 girilmiştir.

3. Piyasadaki cep telefonlarını A ve B şirketlerinin ürettiğini varsayalım. A şirketi, tüm cep telefonu üretiminin %80'nini gerçekleştirsin. Ayrıca A şirketinin ürettiği cep telefonlarının %5'inin bozuk; B'nin ürettiği cep telefonlarının ise %16'sının bozuk olduğu varsayılsın. Bu durumda bozuk bir cep telefonun B şirketi tarafından üretilmiş olma olasılığı nedir? (10 puan)

$P(A) = 0.8$ bir cep telefonunun A şirketi tarafindan uretilme olasiligi

$P(B) = 0.2$ bir cep telefonunun B şirketi tarafindan uretilme olasiligi

$P(bozuk|A) = 0.05$ A şirketi tarafindan uretilmis bir telefonun bozuk olma olasiligi

$P(bozuk|B) = 0.16$ A şirketi tarafindan uretilmis bir telefonun bozuk olma olasiligi

$$P(B|bozuk) = \frac{P(bozuk|B)P(B)}{P(bozuk|A)P(A) + P(bozuk|B)P(B)}$$

$$= \frac{0.16 \cdot 0.2}{0.05 \cdot 0.8 + 0.16 \cdot 0.2}$$

$$= 0.44$$

4.

		Eğitim Örnekleri						
		E1	E2	E3	E4	E5	E6	E7
Test Örnekleri	T1	1.3	1.1	0.2	0.9	0.1	0.3	0.1
	T2	1	0.9	0.4	0.22	0.2	0.3	0.14
	T3	0.9	1.7	0.99	1.1	0.8	1.6	1
	T4	1.3	1.5	1.1	1.7	0.99	0.98	1

Yukarıdaki tablo T1,T2,T3 ve T4 test örneklerinin; E1,E2,E3,E4,E5,E6 ve E7 eğitim örneklerine olan öklid uzaklıklarını göstermektedir. Test örneklerinden T1 ve T3 pozitif T2 ve T4 negatif sınıftadır. Eğitim örneklerinden E1,E2,E3 ve E4 pozitif sınıfta; E5,E6 ve E7 negatif sınıftadır.

- **a.** Test örneklerini k-en yakın komşu algoritmasıyla, k'yı 3 alarak sınıflandırın. (10 puan)
- **b.** a'da bulunduğunuz sınıf tahminlerine göre sınıflandırıcınızın kesinliğini ölçün. (5 puan)
- **c.** Karışıklık matrisini oluşurun. Gerçek pozitif oranını ve gerçek negatif oranını hesaplayın. (10 puan)

T1' e en yakın 3 komşu: E5,E7 ve E3. Bunların ikisi negatif biri pozitiftir. O halde T1 örnegi icin tahminimiz negatiftir.

T2'ye en yakın 3 komşu: E7, E5 ve E4. Bunların ikisi negatif biri pozitiftir. O halde T2 örnegi icin tahminimiz negatiftir.

T3'e en yakın 3 komşu: E5, E1 ve E3'tür. Bunların ikisi pozitif biri negatiftir. O halde T3 örnegi icin tahminimiz pozitiftir.

T4'e en yakın 3 komşu: E5, E6 ve E7'dir. Bunlarin üçü de negatiftir. O halde T4 örnegi icin tahminimiz negatiftir.

Sonuç olarak T2, T3 ve T4 için sınıf tahminlerimiz doğru; T1 için yanlıştır. (4 tahmini 3 ü doğru)

Kesinlik: $\left(\frac{3}{4}\right) * 100 = \%75$

Karışıklık matrisi:

		Tahmin Edilen Sınıflar	
		Poz.	Neg.
Gerçek Sınıflar	Poz.	1	1
	Neg.	0	2

Gerçek pozitif oranı=$\frac{1}{2}$ Gerçek negatif oranı=$\frac{2}{2} = 1$

5. a. Gradient descent algoritmasi veri madenciliginde ne için kullanılır? Bu algoritmanın veri madenciliginde daha iyi çalışması için nelere dikkat edilmelidir? (5 puan)

b. Gradient descent algoritmasının çalıştığını nasıl anlarsınız? (5 puan)

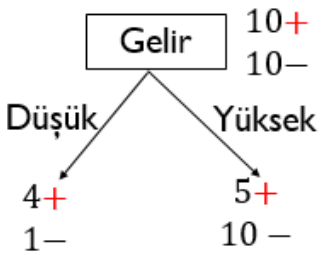
c. Hangi durumda gradient descent çalışmaz? Bu durumda geleneksel çözüm yolu nedir? (5 puan)

a. Gradient descent algoritmasi bir maliyet fonksiyonunun minimum degerini verekn katsayilarin bulunmasinda kullanilir. Bu algoritmanin daha iyi calisabilmesi icin öğrenme orani dikkatli secilmelidir (cok kucuk yada cok buyuk degil) ve veri matrisindeki özelliklerin normalize edilmiş olmasi gerekir.

b. Gradient descent algoritmasinin calisitigini anlamak icin her iterasyon sonucu bulunan yeni katsayilarin maliyet fonksiyonunu bir önceki iterasyona gore dusurdugu gözlemlenmelidir.

c. Birden cok minimum oldugu durumda (lokal minimumlarin var oldugu durumlarda) gradient descent calismaz. Bu durumda bir den fazla kez gradient descent'i calistirmali her defasinda farkli baslangic degerleri secmeliyiz.

6. Bir karar ağacında belirli bir kökte elimizde 20 tane örnek olsun. Bu örneklerin 10 tanesi pozitif sınıfa (+),10 tanesi negatif sınıfa(-) ait olsun. Eğer bu örnekleri aşağıdaki gibi gelir özelliğine göre ikiye ayırırsak entropiyi ne kadar düşürürüz (bilgi kazancımız ne kadar olur) ?



Veri cinsiyete göre yada gelire göre bolunmeden onceki entropi:

$$-\frac{10}{20}log_2\frac{10}{20}-\frac{10}{20}log_2\frac{10}{20}=1$$

Gelire göre veri ikiye ayrilirsa:

Gelirin düşük olduğu durumda entropi: $-\frac{4}{5}log_2\frac{4}{5}-\frac{1}{5}log_2\frac{1}{5}-=0.7219$

Bulunan entropiyi gelirin dusuk olma olasiligi olan $\frac{5}{20}=0.25$ ile carpiyoruz: $0.25\cdot 07219=0.1805$

Gelirin yüksek olduğu durumda entropi: $-\frac{5}{15}log_2\frac{5}{15}-\frac{10}{15}log_2\frac{10}{15}-=0.9183$

Bulunan entropiyi gelirin yuksek olma olasiligi olan $\frac{15}{20}$ ile carpiyiyoruz: $\frac{15}{20}\cdot 0.9183=0.6887$

Gelire gore veriyi ikiye ayirmakla elde edilen yeni entropi: $0.1805+0.6887=0.8692$

Bilgi kazancı=$1-0.8692=0.1308$