

Veri Madenciliğine Giriş

1

Veri Madenciliği Nedir?

Veri madenciliği büyük ölçekli veriler arasından "değeri olan" bir bilgiyi elde etme işidir. Bunun anlamı, veri madenciliği bir kurumda üretilen tüm verilerin belirli yöntemler kullanarak var olan ya da gelecekte ortaya çıkabilecek gizli bilgiyi su yüzüne çıkarma süreci olarak değerlendirilebilir. Bu açıdan bakıldığında, veri madenciliği işinin kurumların karar destek sistemleri için önemli bir yere sahip olabileceğini söyleyebiliriz.

2

Veri Madenciliği Nedir?

Veri madenciliği aslında klasik istatistiksel uygulamalara çok benzer. Ancak klasik istatistiksel uygulamalar yeterince düzenlenmiş ve çoğunlukla özet veriler üzerinde çalıştırılır. Veri madenciliğinde ise milyonlarca ve hatta milyarlarca veri ve çok daha fazla değişken ile ilgilenilir. Veri sayısı çok olunca, bazı özel analiz algoritmaların geliştirilmesi gerekmiş, ayrıca verinin saklandığı ortamların da örneğin veri ambarı biçiminde yeniden düzenlenmesini gerekli kılmıştır.

Dr.Eyyüp GÜLBANDILAR
http://web.ogu.edu.tr/egulbandilar

29.09.2022

3

3

Veriyi Bilgiye Dönüştürmenin Yolu

Veriler üzerinde çözümlemeler yapmak amacıyla çeşitli istatistiksel ve matematiksel yöntemler kullanılabilir. Ancak veri sayısı arttıkça sorunlar ortaya çıkacaktır. Özellikle ilişkisel veri tabanları üzerinde bu çözümlemeleri yapmak zorlaşacaktır. Bu tür veriler üzerinde çözümlemeleri yapabilmek için hem yeni veritabanı kavramlarına hem de yeni çözümleme yöntemlerine gereksinim duyulmaktadır. Veriyi yönetmek için "**veri ambarı**" ve verileri çözümleyerek "**yararlı bilgiye**" erişilmesini sağlayan "**veri madenciliği**" kavramları ortaya atılmıştır.

Dr.Eyyüp GÜLBANDILAR
http://web.ogu.edu.tr/egulbandilar

29.09.2022

4

4

Uygulama Alanları

Veri madenciliğinin günümüzde yaygın bir kullanım alanı bulunmaktadır. Örneğin pazarlama, bankacılık ve sigortacılık gibi alanlarda ve elektronik ticaret ile ilgili alanlarda yaygın şekilde kullanılmaktadır.

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

5

5

Pazarlama

- Müşterilerin satın alma alışkanlıklarının belirlenmesi,
- Müşterilerin demografik özellikleri arasındaki bağlantıların bulunması,
- Posta kampanyalarında cevap verme oranının artırılması,
- Mevcut müşterilerin elde tutulması, yeni müşterilerin kazanılması,
- Pazar sepeti analizi,
- Müşteri ilişkileri yönetimi,
- Müşteri değerlendirmesi,
- Satış tahmini,

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

6

6

Bankacılık

- Farklı finansal göstergeler arasında gizli korelasyonların bulunması,
- Kredi kartı dolandırıcılıklarının tespiti,
- Kredi kartı harcamalarına göre müşteri gruplarının belirlenmesi,
- Kredi taleplerinin değerlendirilmesi.

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

7

7

Sigortacılık

- Yeni poliçe talep edecek müşterilerin tahmin edilmesi,
- Sigorta dolandırıcılıklarının tespiti,
- Riskli müşteri örüntülerinin belirlenmesi.

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

8

8

Elektronik Ticaret

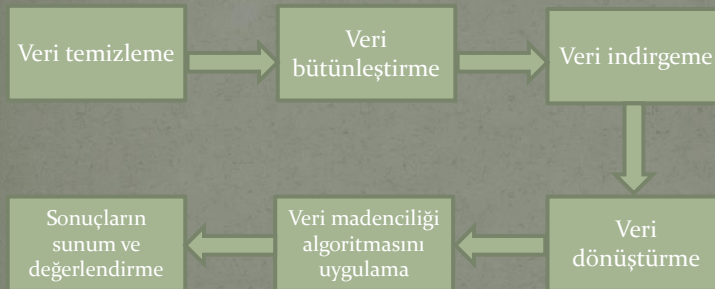
- Saldırıların çözümlenmesi
- e-CRM (Elektronik Müşteri İlişkileri Yönetimi) uygulamalarının yönetimi
- Web sayfalarına yapılan ziyaretlerin çözümlenmesi.

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

9

9

Veri Madenciliği Süreci



Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

10

10

Veri Temizleme

Veritabanında yer alan tutarsız ve hatalı verilere gürültü olarak değerlendirilmektedir. Bu gibi durumlarda verinin söz konusu sorunlardan temizlenmesi gerekecektir. Eksik verilerin yerine yenileri belirlenerek konulmalıdır.

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

11

11

Veri Temizleme

- Eksik değer içeren kayıtlar veri kümesinden atılabilir.
- Kayıp değerlerin yerine bir genel sabit kullanılabilir. Bütün kayıp değerler için aynı sabit kullanılabilir. Ancak bütün değişkenlerde kayıp değerler yerine aynı sabit değer kullanımı sorun yaratacaktır.
- Değişkenin tüm verileri kullanılarak ortalaması hesaplanır ve eksik değer yerine bu değer kullanılabilir.
- Değişkenin tüm verileri yerine, sadece bir sınıfa ait örneklerin değişken ortalaması hesaplanarak eksik değer yerine kullanılabilir.
- Verilere uygun bir tahmin yapılarak, örneğin regresyon ya da karar ağacı modeli kurularak eksik değer tahmin edilebilir ve eksik değer yerine kullanılabilir.

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

12

12

Veri Bütünleştirme

Farklı veri tabanlarından ya da veri kaynaklarından elde edilen verilerin birlikte değerlendirmeye alınabilmesi için farklı türdeki verilerin tek türe dönüştürülmesi yani bütünleştirilmesi söz konusu olacaktır. Eğer veri madenciliği uygulaması için bir veri ambarı altyapısı hazırlanmış ise söz konusu veri bütünleştirme işleminin yapılmış olması gerekmektedir. Ancak böyle bir yapı yoksa söz konusu veri bütünleştirme işleminin doğrudan veri madenciliğine esas oluşturacak veriler üzerine uygulanması gerekecektir.

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

13

13

Veri İndirgeme

- Veriyi birleştirme veya veri küpü
- Boyut indirgeme
- Veri sıkıştırma
- Örneklem
- Genelleme

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

14

14

Veri İndirgeme

Veriyi indirgeme aşamasında verilerin çok boyutlu **veri küpleri** biçimine dönüştürmek söz konusu olabilir. Böylece çözümlemeler sadece belirlenen boyutlara göre yapılır. Veriler arasında bir seçme işlemi yapılarak, gereksiz veriler veri tabanından çıkarılır ve boyut azaltılması sağlanabilir. Veri sıkıştırma aşamasında, büyük veri kümelerinin sıkıştırılarak daha az yer işgal etmeleri sağlanır, örnekleme aşamasında ise, büyük veri topluluğu yerine onu temsil eden daha küçük veri kümelerinin oluşturulması amaçlanır. Genelleme verilerin tek tek değil genel kavramlarla ifade edilmesini sağlar.

Dr.Eyyüp GÜLBANDILAR
http://web.ogu.edu.tr/egulbandilar

29.09.2022

15

15

Veri Dönüştürme

Veriyi bazı durumlarda veri madenciliği çözümlemelerine aynen katmak uygun olmayabilir. Değişkenlerin ortalama ve varyansları birbirlerinden önemli ölçüde farklı olduğu takdirde büyük ortalama ve varyansa sahip değişkenlerin diğerleri üzerindeki baskısı daha fazla olur ve onların rollerini önemli ölçüde azaltır. Ayrıca değişkenlerin sahip olduğu çok büyük ve çok küçük değerler de çözümlemelerin sağlıklı biçimde yapılmasını engeller. Bu nedenle bir dönüşüm yöntemi uygulayarak söz konusu değişkenlerin normalleştirilmesi veya standartlaştırılması uygun bir yol olacaktır.

Dr.Eyyüp GÜLBANDILAR
http://web.ogu.edu.tr/egulbandilar

29.09.2022

16

16

Normalleştirme

Verileri belirli bir sayısal değer aralığında dönüştürmek için normalleştirme yöntemi uygulanır. Bu yöntem, veri içindeki en büyük ve en küçük sayısal değer belirlenerek diğerlerini buna uygun biçimde dönüştürme esasına dayanmaktadır. Söz konusu dönüştürme bağıntısı şu şekilde ifade edilmektedir:

$$S^* = \frac{S - \min}{\max - \min} (\text{yeni_maks} - \text{yeni_min}) + \text{yeni_min}$$

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

17

17

Normalleştirme

X	X*
30	0,0000
36	0,1875
45	0,4688
50	0,6250
62	1,0000

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

18

18

Z-skor Standartlaştırma

Bu yöntem, verilerin ortalaması ve standart hatası göz önüne alınarak yeni değerlere dönüştürülmesi esasına dayanmaktadır. Söz konusu dönüşümlerde şu şekilde bir bağıntıya yer verilir:

$$X^* = \frac{X - \bar{X}}{\sigma_x} \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

19

19

Z-skor Standartlaştırma

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = 44,6$$

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}} = 12,44$$

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

20

20

Z-skor Standartlaştırma

X	X*
30	-1,1735
36	-0,6912
45	0,0321
50	0,4340
62	1,3985

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

21

21

Veri Madenciliği Algoritmasını Uygulama

Veri hazır hale getirildikten sonra konuyla ilgili veri madenciliği algoritmaları uygulanır. Söz konusu algoritmalar sınıflandırma, kümeleme ve birliktelik kuralları konusunda olacaktır.

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

22

22

Sonuçları Sunum ve Değerlendirme

Veri madenciliği algoritması veriler üzerinde uygulandıktan sonra, sonuçlar düzenlenerek ilgili yerlere sunulur. Sonuçlar çoğu kez grafiklerle desteklenir. Örneğin bir hiyerarşik kümeleme modeli uygulanmış ise sonuçlar dendrogram adı verilen özel grafiklerle sunulur.

Dr.Eyyüp GÜLBANDILAR
http://web.ogu.edu.tr/egulbandilar

29.09.2022

23

23

Veri Madenciliği Yöntemleri

Veri madenciliği konusunda çok sayıda yöntem ve algoritma geliştirilmiştir. Bu yöntemlerin bir çoğu istatistiksel tabanlıdır. Bu kitap kapsamında üç ayrı konuda veri madenciliği modelleri tanıtılacaktır. Söz konusu veri madenciliği modellerini temel olarak şu şekilde gruplandırabiliriz:

- Sınıflandırma
- Kümeleme
- Birliktelik kuralları.

Dr.Eyyüp GÜLBANDILAR
http://web.ogu.edu.tr/egulbandilar

29.09.2022

24

24

Sınıflandırma

Sınıflama veri madenciliğinde sıkça kullanılan bir yöntem olup, veri tabanlarındaki gizli örüntüleri ortaya çıkarmakta kullanılır. Verilerin sınıflandırması için belirli bir süreç izlenir. Öncelikle var olan veri tabanının bir kısmı eğitim amacıyla kullanılarak sınıflandırma kurallarının oluşturulması sağlanır. Daha sonra bu kurallar yardımıyla yeni bir durum ortaya çıktığında nasıl karar verileceği belirlenir.

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar>

29.09.2022

25

25

Sınıflandırma

Bir bankanın kredi verdiği müşterilerinin risk durumunu karar ağaçları yardımıyla ortaya koymak istediğini varsayalım. Bu sayede belirli özelliklere sahip müşterilerinden kredi talebi geldiğinde, karar ağacı bilgilerine dayanarak kredi verip vermeme konusunda karar verecektir.

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar>

29.09.2022

26

26

Sınıflandırma

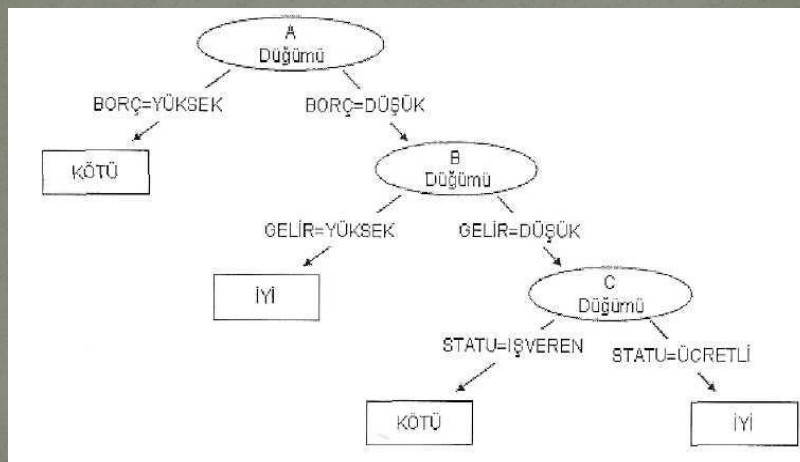
MÜŞTERİ	BORÇ	GELİR	STATÜ	RİSK
1	YÜKSEK	YÜKSEK	İŞVEREN	KÖTÜ
2	YÜKSEK	YÜKSEK	ÜCRETLİ	KÖTÜ
3	YÜKSEK	DÜŞÜK	ÜCRETLİ	KÖTÜ
4	DÜŞÜK	DÜŞÜK	ÜCRETLİ	İYİ
5	DÜŞÜK	DÜŞÜK	İŞVEREN	KÖTÜ
6	DÜŞÜK	YÜKSEK	İŞVEREN	İYİ
7	DÜŞÜK	YÜKSEK	ÜCRETLİ	İYİ
8	DÜŞÜK	DÜŞÜK	ÜCRETLİ	İYİ
9	DÜŞÜK	DÜŞÜK	İŞVEREN	KÖTÜ
10	DÜŞÜK	YÜKSEK	İŞVEREN	İYİ

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

27

27

Sınıflandırma



Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

28

28

Sınıflandırma

KURAL 1:

Eğer BORÇ = YÜKSEK ise RİSK = KÖTÜ;

KURAL 2:

Eğer BORÇ = DÜŞÜK ise ve

Eğer GELİR=YÜKSEK ise RİSK=İYİ;

KURAL 3:

Eğer BORÇ = DÜŞÜK ise ve

Eğer GELİR=DÜŞÜK ise ve

Eğer STATÜ=İŞVEREN ise RİSK=KÖTÜ;

KURAL 4:

Eğer BORÇ = DÜŞÜK ise ve

Eğer GELİR=DÜŞÜK ise ve

Eğer STATÜ=ÜCRETLİ ise RİSK=İYİ;

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar>

29.09.2022

29

29

Kümeleme

Kümeleme, verilerin kendi aralarındaki benzerliklerin göz önüne alınarak gruplandırılması işlemidir. Bu özelliği nedeniyle pek çok alanda uygulanabilmektedir. Örneğin, pazarlama araştırmalarında yaygın biçimde kullanılmaktadır. Bunun dışında desen tanımlama, resim işleme ve uzaysal harita verilerinin analizinde kullanılmaktadır.

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar>

29.09.2022

30

30

Kümeleme

Gözlem	X1	X2
1	1	1
2	2	1
3	4	5
4	7	7
5	5	7

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022 31

31

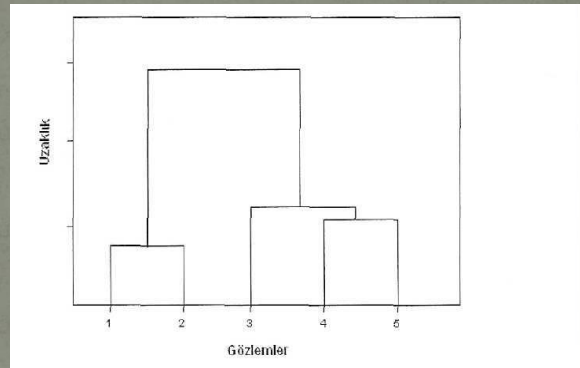
Kümeleme

Kümeler	
Küme 1	(1,2)
Küme 2	(4,5)
Küme 3	(3,4,5)
Küme 4	(1,2,3,4,5)

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022 32

32

Kümeleme

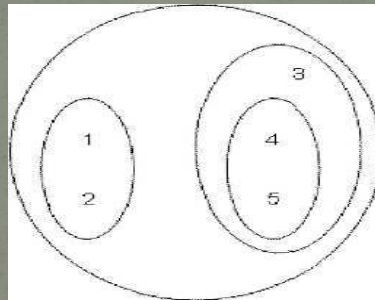


Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

33

33

Kümeleme



Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

34

34

Birliktelik Kuralları

Veritabanı içinde yer alan kayıtların birbirleriyle olan ilişkilerini inceleyerek, hangi olayların eş zamanlı olarak birlikte gerçekleşebileceklerini ortaya koymaya çalışan veri madenciliği yöntemleri bulunmaktadır. Bu ilişkilerin belirlenmesiyle "**birliktelik kuralları**" (association rules) elde edilir.

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

35

35

Birliktelik Kuralları

"Pazar sepet analizleri" adı verilen uygulamalar bu tür veri madenciliği yöntemlerine dayanmaktadır. Bu tür çözümlemelerden hareketle, müşterilerin alışveriş alışkanlıkları belirlenmeye çalışılır.

Pazar sepet analizleri yardımıyla bir müşteri herhangi bir ürünü aldığı anda, sepetine başka hangi ürünleri de koyduğu belirli bir olasılığa göre ortaya konur. Birlikte satın alınan ürünler belirlendiğinde, mağazalarda raflar ona göre düzenlenerek müşterilerin bu tür ürünlere daha kolayca erişimleri sağlanabilir.

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar> 29.09.2022

36

36

Birliktelik Kuralları

Bir mağazada alışveriş yapan müşterilerin alışveriş alışkanlıklarını belirlemek istediğimizi varsayalım. Beş müşterinin alışveriş sepetlerine hangi ürünleri koyduğunu aşağıdaki tablo üzerinde görüyoruz.

Müşteri	Alışveriş sepetindeki ürünler
1	Makama, Yağ, Meyve suyu, Peynir
2	Makarna, Ketçap
3	Ketçap, Yağ, Meyve suyu, Bira
4	Makarna, Ketçap, Yağ, Meyve suyu
5	Makarna, Ketçap, Yağ, Bira

Dr.Eyyüp GÜLBANDILAR
http://web.ogu.edu.tr/egulbandilar 29.09.2022

37

37

Birliktelik Kuralları

$\{\text{Ketçap, Meyve suyu}\} \rightarrow \{\text{Yağ}\} \text{ (s=0.4, c=1.0)}$
 $\{\text{Ketçap, Yağ}\} \rightarrow \{\text{Meyve suyu}\} \text{ (s=0.4, c=0.67)}$
 $\{\text{Yağ, Meyve suyu}\} \rightarrow \{\text{Ketçap}\} \text{ (s=0.4, c=0.67)}$
 $\{\text{Meyve suyu}\} \rightarrow \{\text{Ketçap, Yağ}\} \text{ (s=0.4, c=0.67)}$
 $\{\text{Yağ}\} \rightarrow \{\text{Ketçap, Meyve suyu}\} \text{ (s=0.4, c=0.5)}$
 $\{\text{Ketçap}\} \rightarrow \{\text{Yağ, Meyve suyu}\} \text{ (s=0.4, c=0.5)}$

Dr.Eyyüp GÜLBANDILAR
http://web.ogu.edu.tr/egulbandilar 29.09.2022

38

38

Birliktelik Kuralları

Bu sonuçların her bir satırını şu şekilde yorumlayabiliriz:

- Ketçap ve Meyve suyunu birlikte alanlar mutlaka yağ da alıyorlar.
- Ketçap ve yağ satın alan müşteriler %67 olasılıkla meyve suyu da alıyorlar.
- Yağ ve meyve suyunu birlikte satın alanlar %67 olasılıkla ketçap da alıyorlar.
- Meyve suyu alanlar %67 olasılıkla ketçap ve yağ da satın alıyorlar.
- Yağ alanlar %50 olasılıkla ketçap ve meyve suyu da alıyorlar
- Ketçap alanlar %50 olasılıkla yağ ve meyve suyu da alıyorlar.

Dr.Eyyüp GÜLBANDILAR
<http://web.ogu.edu.tr/egulbandilar>

29.09.2022

39