Veri madenciliği süreçleri etkinliği sunusu

Hedefler

1. Veri madenciliği süreçlerini incelemek
2. Veri madenciliği modelini uygulamak için veriye yapılması gereken işlemlerin alt yapısını planlamak
3. Veri madenciliği süreçlerinden veri ön işleme için temel matematik işlemleri yapmak.

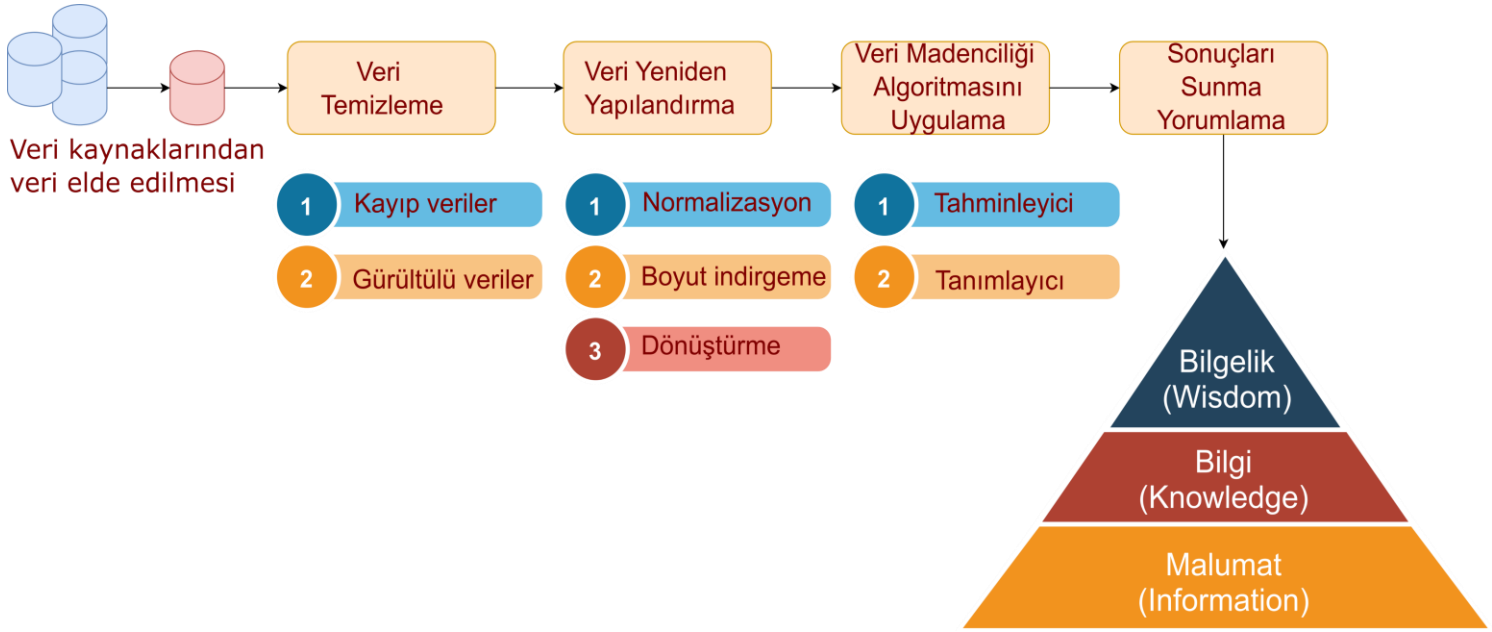
Veri madenciliği süreci

- **Veri madenciliği:** Belli bir amaç için toplanmış verileri işleyerek, daha önce bilinmeyen tahminler veya tanımlar bulmayı amaçlayan çalışma alanı. Veri madenciliği, istatistik, makine öğrenmesi, yapay zeka gibi bir çok alanın ortak olarak çalışıldığı bir alandır.
- Süreç temelde beş adımdan oluşuyor. *Veri Topluyorum* etkinliğinde sürecin 1. adımı incelemiştik. Şimdi süreçte yer alan diğer adımları sırası ile inceleyelim.
- Veri madenciliği sürecinde verinin elde edilmesinden verinin işlenecek hale getirilmesi sürecine ***veri ön işleme*** denir.
- Bu süreç veri madenciliği sürecinin en önemli kısmıdır diyebiliriz.
- Günümüzde veriyi elde etmek çok zor olmasa da, amaca yönelik veri elde etmek zor olabilir. Bunun yanında bulunan veri işlemeye uygun olmayabilir.
- Veri ön işleme süreci veriyi işlemek için hazır hale getirmek olduğundan en önemli süreçtir diyebiliriz.
- Ön işleme süreci, veri temizleme ve veri yeniden yapılandırma kısımlarını sırası ile inceleyelim.

Veri madenciliği süreci

- Veri madenciliği süreci Görsel 1'de görüldüğü gibi çeşitli aşamalardan oluşmaktadır.
- İşlenmemiş veri elde edilip bu verilerden çıkarımlar ve tahminler yapmaya kadar olan süreç, veri madenciliği süreci olarak adlandırılır.
- Bu süreç şöyle sıralanabilir.

1. **Veri toplama:** Gerekli veriyi elde etmek.
2. **Veri temizleme:** Kayıp veya gürültülü veri sorununu çözmek.
3. **Veri yeniden yapılandırma:** Veriyi işlenecek hale getirmek.
4. **Veri madenciliği algoritmasını uygulama:** Veriyi işlemek.
5. **Sonuçları yorumlama:** Veriden çıkarımda bulunmak.



Görsel 1: Veri madenciliği süreci

Veri ön işleme

Veri temizleme

- Veri madenciliği sürecinde, veri setinde bulunan kayıp veya gürültülü veri probleminin çözülmesi aşamasıdır.
- **Kayıp veri:** Hakkında veri toplanan bir varlığa ait özniteliklerin hiçbir değer içermemesi durumu.
- **Gürültülü veri (Noise data):** Veri toplama işlemi esnasında oluşan hatalı, tutarsız değerler veya çok uçta bulunan değerlerdir. Örnek olarak, bir kişinin yaşının yanlışlıkla 450 girilmesi, bir algılayıcı arızasından dolayı hava sıcaklığının 1050 derece ölçülmesi veya boy ortalaması 1.6 metre olan bir grupta, boyu 2.05 metre olan birinin bulunması, verilebilir.

Veri ön işleme

Veri yeniden yapılandırma

- Veri madenciliğinde kullanacağımız model/algoritma yapısına uygun olarak verilerin tekrar düzenlenmesi aşamasıdır. Bu aşamada tecrübe ve veri madenciliği algoritmalarını iyi tanımak çok önemlidir.
- **Normalizasyon:** Bir özniteliğe ait sayısal verilerin istenilen sayı aralığına indirgenmesi işlemidir. Genellikle bu indirgeme 0 ve 1 sayı aralığına yapılır.
- **Boyut indirgeme:** Veri setindeki öznitelik sayısının azaltılmasıdır işlemidir. Bunun yapılabilmesi için en az iki veya daha fazla özniteliği temsil eden değişkenin birlikte hareket ediyor olması gerekir. Bu sayede algoritmaların çalışma zamanlarından, başarımdan ödün vermeden kazanç elde edilebilir.
- **Dönüştürme:** Genellikle sayısal verilerin amaca uygun olarak kategorik veriye dönüştürülmesidir. Bu işlemin amacı bazı veri madenciliği model/algoritmalarının kategorik verilerle daha sağlıklı çalışıyor olmasıdır. Örneğin teoride sonsuz değer alabilecek maaş bilgisinin, belirlenen aralıklarla, “düşük”, “orta”, “yüksek” gibi üç kategoride ifade edilmesidir.

Veri ön işleme

- Veri ön işleme işlemlerini anlamak için Tablo 1'de verilen sayısal bir değişkene ait verilerden yararlanacağız. Bu veriler bir grup sporcuya ait ağırlık verileridir.
- Burada bizi ilgilendiren veri sporcu ağırlığı değişkenine ait veri sütunudur.
- Bu veri ile ilgili takip edeceğimiz işlem sırası:
  A. Veri temizleme
     1. Kayıp veriler sorununun çözülmesi
     2. Gürültülü veri sorununun çözülmesi
  B. Veri yeniden yapılandırma
     1. Normalizasyon
     2. Dönüştürme

Sporcu no	**sporcu ağırlığı(kg)**
sporcu1	65
sporcu2	86
sporcu3	56
sporcu4	78
sporcu5	68
sporcu6	190
sporcu7	51
sporcu8	?
sporcu9	106
sporcu10	90
sporcu11	60
sporcu12	?
sporcu13	96
sporcu14	93
sporcu15	52
sporcu16	40

Tablo 1: Sporcu ağırlığı verileri tablosu

Veri temizleme

Kayıp veri sorununun çözümü için

- Verinin kayıp olduğu kayıtları silmek.
- Kayıp olan verilerin el ile doldurulması.
- Tüm kayıp verilere aynı değeri vermek.
- Kayıp olan verilere merkezi eğilim ölçülerinden birini girmek (ortalama, ortanca veya mod)
- Diğer değişkenlerden yararlanarak değerin tahmin edilmesi.

Biz burada var olan verilerin ortanca değerini kullanmak istiyoruz. Var olan verileri küçükten büyüğe sıraladığımızda:

40,51,52,56,60,65,68,78,86,90,93,96,106,190

Ortanca=(68+78)/2=73 olarak hesaplanır.

Eksik verileri yerine 73 değerini yazıyoruz

Sporcu no	sporcu ağırlığı(kg)
sporcu1	65
sporcu2	86
sporcu3	56
sporcu4	78
sporcu5	68
sporcu6	190
sporcu7	51
sporcu8	?
sporcu9	106
sporcu10	90
sporcu11	60
sporcu12	?
sporcu13	96
sporcu14	93
sporcu15	52
sporcu16	40

Tablo 1: Sporcu ağırlığı verileri tablosu

Veri temizleme

Gürültülü veri sorununun çözümü için

- Veride gürültü olup olmadığını anlamak için çeyrekler aralığı yöntemi sıklıkla kullanılır.
- Var olan verileri küçükten büyüğe sıraladığımızda:

40,51,52,56,60,65,68,73,73,78,86,90,93,96,106,190

Q1=(56+60)/2=58, Q3=(90+93)/2=91.5, IQR=96.5-58=33.5

Alt sınır = Q1-1.5*IQR=58-1.5*33.5=7.75

Üst sınır = Q3+1.5*IQR= 91.5+1.5*33.5=141,75

Bu belirlediğimiz sınırlara göre sporcu6'nın 190 kg ağırlığının gürültülü veri olduğu tespit edilmiştir.

Bundan sonrası için yapılacak en iyi işlem gürültülü kaydın veriden çıkarılması olacaktır.

Sporcu no	sporcu ağırlığı(kg)
sporcu1	65
sporcu2	86
sporcu3	56
sporcu4	78
sporcu5	68
sporcu6	190
sporcu7	51
sporcu8	73
sporcu9	106
sporcu10	90
sporcu11	60
sporcu12	73
sporcu13	96
sporcu14	93
sporcu15	52
sporcu16	40

Tablo 2: Kayıp veriler giderilmiş tablo

Veri yeniden yapılandırma

Normalizasyon

Sayısal değerlerin 0-1 değer aralığına Normalizasyon işlemi için aşağıdaki formül kullanılır:

$$yeni\ deger = \frac{eski\ değer - en\ küçük\ değer}{en\ büyük\ değer - en\ küçük\ değer}$$

Örnek olarak 73 değeri için:

$$yeni\ deger = \frac{73 - 40}{106 - 40} = 0.5$$

Burada sporcu6'ya ait verinin veri setinden kaldırıldığı unutulmamalıdır.

Sporcu no	sporcu ağırlığı(kg)
sporcu1	65
sporcu2	86
sporcu3	56
sporcu4	78
sporcu5	68
sporcu7	51
sporcu8	73
sporcu9	106
sporcu10	90
sporcu11	60
sporcu12	73
sporcu13	96
sporcu14	93
sporcu15	52
sporcu16	40

Tablo 3: Gürültülü veri temizlenmiş tablo

Veri yeniden yapılandırma

Normalizasyon

Tablo 4’te verinin normalize edilmiş değerleri görülmektedir.

Sporcu no	sporcu ağırlığı(kg)	0-1 arası normalize edilmiş değerler
sporcu1	65	0,378787879
sporcu2	86	0,696969697
sporcu3	56	0,242424242
sporcu4	78	0,575757576
sporcu5	68	0,424242424
sporcu7	51	0,166666667
sporcu8	73	0,5
sporcu10	90	0,757575758
sporcu11	60	0,303030303
sporcu12	73	0,5
sporcu13	96	0,848484848
sporcu14	93	0,803030303
sporcu15	52	0,181818182
sporcu16	40	0

Tablo 4: Veri normalize edilmiş halini gösteren tablo

Veri yeniden yapılandırma

Dönüştürme

Sayısal veriyi kategorik veriye dönüştürmek için, belirlediğimiz aralıklarda olan verilere birer kategori ismi vereceğiz.

65 veya daha küçük olan değerleri **zayıf**

65'ten büyük ve 90'dan küçük değerleri **normal**

90 veya daha büyük olan değerleri **kilolu**

Olarak kategorize edeceğiz.

Sporcu no	sporcu ağırlığı(kg)	0-1 arası normalize edilmiş değerler	kategorik değerler
sporcu1	65	0,378787879	Zayıf
sporcu2	86	0,696969697	Normal
sporcu3	56	0,242424242	Zayıf
sporcu4	78	0,575757576	Normal
sporcu5	68	0,424242424	Normal
sporcu7	51	0,166666667	Zayıf
sporcu8	73	0,5	Normal
sporcu10	90	0,757575758	Kilolu
sporcu11	60	0,303030303	Zayıf
sporcu12	73	0,5	Normal
sporcu13	96	0,848484848	Kilolu
sporcu14	93	0,803030303	Kilolu
sporcu15	52	0,181818182	Zayıf
sporcu16	40	0	Zayıf

Tablo 5: Veri Dönüştürme sonuçlarını gösteren tablo

Kaynakça

Akpınar, H. (2014). Data: veri madenciliği veri analizi. Papatya.

Alpaydın, E. (2013). Yapay öğrenme. Boğaziçi Üniversitesi Yayınevi.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Içinde Data Mining: Concepts and Techniques. Elsevier Inc. https://doi.org/10.1016/C2009-0-61819-5

Silahtaroğlu, G. (2008). Kavram ve algoritmalarıyla temel Veri Madenciliği. Papatya.

Görsel kaynakça

Görsel 1: Veri madenciliği süreci