İstanbul
GEDİK
Üniversitesi

İstanbul GEDİK Üniversitesi

Veri Madenciliği

Öğr. Gör. Zeki ÇIPLAK

Verinin Hazırlanması

Verinin Hazırlanması

- Veriyi eğitmeye başlamadan önce, çeşitli ön işleme aşamalarından geçmesi gerekir. Bu aşamaları aşağıdaki gibi özetleyebiliriz.

  - ❑ ***Kayıp Değer Problemi***
  - ❑ ***Aykırı Değer Problemi***
  - ❑ ***Veriyi Dönüştürme***
  - ❑ ***Özellik/Değişken Seçimi***
  - ❑ ***Veriyi Bölme***

- Kayıp ve Aykırı değer problemlerini, önceki derslerimizde ayrıntılı olarak işlemiştik. Şimdi, diğer bahsedilmeyen aşamaları inceleyebiliriz.

Verinin Hazırlanması

Veriyi Dönüştürme: Veri dönüşümlerinde, Normalizasyon ve Standardizasyon işlemlerinin dışında; *Yeniden kodlama* ve *Kukla Nitelik Oluşturma* (One Hot Encoding) yöntemleri de kullanılmaktadır.

Bazı veri madenciliği algoritmalarında, tüm değişkenlerin kategorik olması daha uygundur. O yüzden, veri setindeki sayısal değerlerin, **yeniden kodlama ile** kategorik değişkenlere dönüştürülmesi gerekebilir. Bu her zaman yaşanabilecek bir durum değildir.

En sonunda, veriler algoritma içerisinde kullanılırken, mutlaka **yeniden sayısal değerlere dönüşüm** söz konusudur. Burada sadece sayısal değerlerin yapısı değiştirilmektedir.

Verinin Hazırlanması

Sıcaklık (F)	Kategorik
80	Yüksek
85	Yüksek
90	Yüksek
78	Normal
75	Normal
60	Düşük

Yandaki tabloda görüldüğü gibi, sayısal Sıcaklık değerleri, bazı değer aralıklarına göre kategorik değerlere dönüştürülmüştür.

Verinin Hazırlanması

Kategorik değişkenlerin de, sayısal değerlere dönüşümü söz konusudur. Bu amaçla **Kukla Nitelik/Değişken** tanımlaması (One-hot Encoding) yapılmaktadır.

Sıcaklık
Yüksek
Yüksek
Yüksek
Normal
Normal
Düşük

Sıcaklık_Yüksek	Sıcaklık_Normal	Sıcaklık_Düşük
1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	0	1

Verinin Hazırlanması

Özellik/Nitelik/Değişken Seçimi: Veri setinde bulunan her sütun, veri madenciliği algoritmalarında kullanıma uygun değildir. Bazı nitelikler, sadece etiket amaçlı olup, model oluşturmada kullanılmazlar.

*Bunlardan en bilineni, **ID türü değişkenler**dir. ID benzeri değişkenler, özellikle SQL sorgularında kolaylaştırıcı ve her bir kayıt bilgiyi birbirinden ayrıştırıcı özelliğe sahiptir. Her bir satırın ayrı bir ID bilgisi olduğu için, ID'ye göre sorgulama ile belli bir satır kolaylıkla çağırılabilmektedir.*

Bu şekilde, oluşturulacak model ile ilgisi olmayan tüm kolonların veri setinden çıkarılması, modelin daha doğru sonuçlar vermesini sağlayacaktır.

Verinin Hazırlanması

Öznitelik seçimi ile **veri setini en iyi temsil eden değişkenlerden oluşan, alt bir veri seti** oluşturulmuş olur.

Öznitelik seçimi için kullanılabilecek filtre, sarmal ve gömülü şeklinde birçok yöntem vardır. Filtre yöntemi, en hızlı ve pratik olan çözümdür.

En iyi bilinen filtre yöntemlerinden üçü ise aşağıdaki gibidir:

- ***Relief***
- ***CFS***
- ***Ki-Kare***

Öznitelik Seçimi

Relief: Bu yöntem, her bir değişkene **-1** ile **+1** arasında bir ağırlık değeri ataması yapar. Bu açıdan, bir korelasyon katsayısı gibi kullanılabilmektedir.

CFS: Hedef değişkenle, yüksek korelasyona sahip diğer değişkenlerin alt kümesinin belirlenmesi mantığına dayanır.

Ki-kare (Chi-squared): Kategorik değişkenler arası ilişkinin araştırılmasında da kullanılır. Veri setinde, en büyük ki-kare değerine sahip olan değişken, hedef değişken için en ayırt edici veya en belirleyici değişkendir diyebiliriz.

Buradaki tüm özellik seçimi yöntemlerini, R yazılımındaki hazır kütüphaneleri kullanarak, kolayca gerçekleştireceğiz.

Verinin Hazırlanması

Veriyi Bölme: Veri setimiz, çeşitli ön işleme aşamalarından geçtikten sonra, veri madenciliği algoritmalarında kullanılmak üzere, eğitim ve test verisi olarak ikiye ayrılmak durumundadır.

Eğitim verisi; *veri madenciliği algoritmasının, veriyi temsil eden bir model kurmak üzere eğittiği veridir.*

Test verisi *ise, oluşturulan modelin test edilebilmesi için saklanan veridir. Modelin, test verisi ile denenerek, sağlıklı çalışıp-çalışmadığı görülebilir.*

Eğitim ve test verisi, ön işleme bittikten sonra belirlenir.

Verinin Hazırlanması

Özellikle sınıflandırmaya dayalı veri madenciliğinde, verinin sınıf özelliğinin de dikkate alınması önemlidir.

Örneğin, hedef değişkenimiz ikili bir yapıda olsun. Sınıflandırma sonucunun «**evet**» ya da «**hayır**» çıkması bekleniyor olsun.

Böyle bir durumda, veriyi bölme işlemi sonucunda, eğitim veya test verisindeki hedef değişkenimizin değerleri, sadece «**evet**» veya sadece «**hayır**» değerlerinden oluşmamalıdır**!**

*Tüm **veri setini tam temsil edebilecek şekilde** bölümleme yapılması gerekir.*

Verinin Hazırlanması

Literatürdeki veri madenciliği çalışmalarının birçoğunda, sırayla **eğitim verisi ile test verisinin oranı %80**'e **%20** şeklindedir.

Bu oran **bazı veri madenciliği projelerinde %70 eğitim** verisi ve **%30 test** verisi olacak şekilde de düzenlenebilmektedir.

Bilgi Keşfi sürecinde, bu oranlar sürekli değiştirilerek, modelin başarısı ölçülür ve modelin en başarılı olduğu durumdaki oranlarda karar kılınabilir.

Regresyon
(Regression)

Regresyon

- **Regresyon modeli**; bir değişkende meydana gelen değişimlerin, diğer değişkenler tarafından açıklanabildiği varsayımı ile oluşturulan modeldir.

- Kelime anlamı olarak "**eğri uydurma**" anlamına da gelir.

- **Korelasyon analizi**, herhangi iki değişken arasındaki ilişkinin yönünü ve derecesini gösterirken; **Regresyon analizi** ise, bağımlı bir değişkenin, bağımsız değişkenler tarafından nasıl açıklandığını ifade eder.

- **Bağımlı değişkene örnek** bir evin fiyatı olsun. Evin fiyatı, sadece *oda sayısı* ve *evin alanı*na göre belirleniyorsa, *oda sayısı* ve *evin alanı*, bağımsız değişkenlerdir.

Regresyon

- Bağımlı (yani hedef) değişkende meydana gelen değişimlerin, sadece bir adet bağımsız değişkenle, doğrusal olarak açıklandığı regresyon modeline, "**Basit Doğrusal Regresyon**" denir.

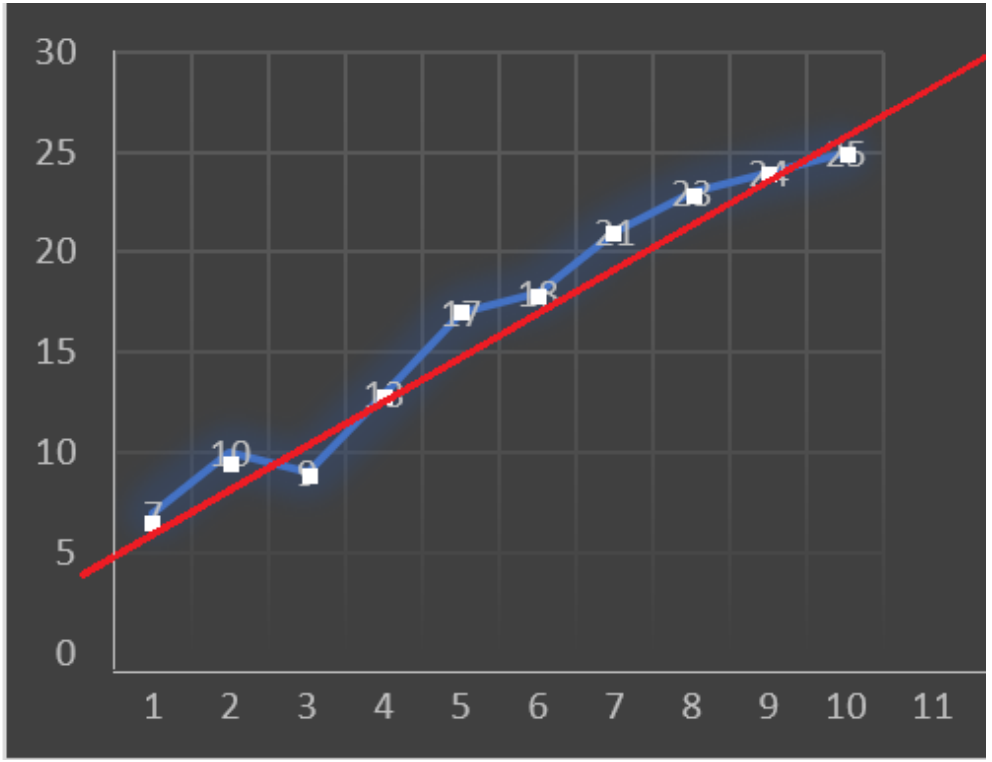
- Bağımlı değişkendeki değişimin, iki veya daha çok bağımsız değişkenle, doğrusal olarak açıklandığı regresyon modeline, "**Çoklu Doğrusal Regresyon**" denir.

- Değişkenler arasındaki ilişkinin doğrusal olmadığı, durumlarda kullanılan regresyon modeline, "**Doğrusal Olmayan Regresyon**" denir.

Regresyon

- Regresyon konusunu basitçe anlatabilmek için, bir değişkenin değerlerinin, indis değerleri ile oluşturduğu grafiği inceleyebiliriz.

	Değişken
1)	7
2)	10
3)	9
4)	13
5)	17
6)	18
7)	21
8)	23
9)	24
10)	25



Regresyon

- Grafikte, tüm değişken değerlerini temsil eden kırmızı çizgi, **regresyon eğrisi**dir. Bu eğrinin geometrik olarak bir denklemi vardır ve değişkende bulunmayan **diğer değerleri tahmin** edebilmemize imkan verir.

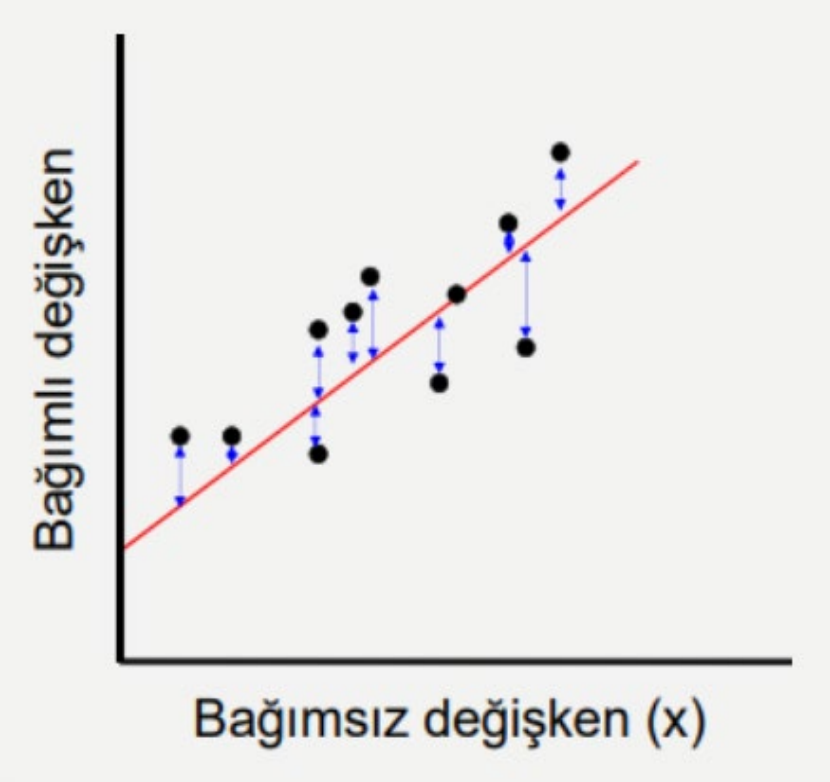
- Örneğin, değişkende toplam 10 değer vardır. **11.** , **12.** veya **n.** indisin alacağı değeri, regresyon eğrisinin denkleminden yola çıkarak "tahmin" edebiliriz.

- Basit doğrusal regresyon denklemi, **ý = ax + b** formundadır. **ý** burada, tahmin edilen değeri temsil eder. a değeri, (Coefficient) indislerin katsayısıdır. Bu katsayı, indisin **ý** değeri üzerinde ne kadar etkili olduğunu gösterir. Etki büyüdükçe, a katsayısı artar. **b** değeri ise **bias terimi**dir. (Intercept) Aynı zamanda regresyon eğrisinin y eksenini kestiği noktadır.

Regresyon

- Önceki grafik örneğinde, indis değerleri (yani **x**) **bağımsız değişken**, diğer değerler ise (yani **y**) **bağımlı değişken** olarak kabul edilir.

- Regresyon eğrisinin (yani modelin), tahmin ettiği değer ile gerçek değer arasındaki farka **hata** denir. Hatalar pozitif olabildiği gibi, negatif de çıkabilmektedir. Eğrinin altında kalan hatalar negatif hatalardır.



Regresyon

- Bu yüzden genelde hataların mutlak değerlerinin ortalaması (**MAE** – **Mean Absolute Error**) veya hataların karelerinin ortalaması (**MSE** – **Mean Square Error**) bulunarak, toplam hata miktarı hakkında bir fikir edinilir.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

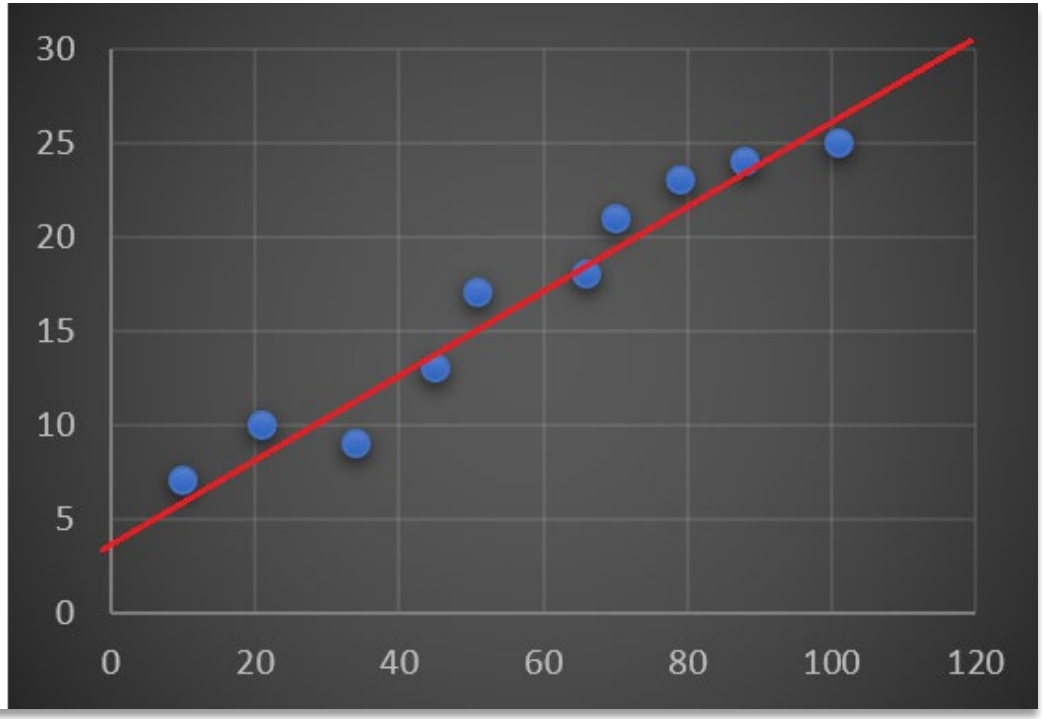
$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

- MSE değerinin karekökü alınarak,
  RMSE (**Root Mean Square Error**) değeri elde edilebilir.

Regresyon

- İndis değerleri (yani **x**) yerine, bir değişkenin değerleri de kullanılabilir. Bu durumda da, yine **basit doğrusal regresyon** elde edilmiş olunur.

X	Y
10	7
21	10
34	9
45	13
51	17
66	18
70	21
79	23
88	24
101	25



Regresyon

X1	X2	Y
10	1	7
21	2	10
34	4	9
45	5	13
51	7	17
66	8	18
70	9	21
79	10	23
88	12	24
101	13	25

Bağımlı değişken **y**, birden fazla **x** değişkenine doğrusal olarak bağlı olabilir.

Böyle bir durumda **Çoklu Doğrusal Regresyon** modeli kullanılır.

Bu modelin regresyon denklemi,

$\acute{y} = w_1.x_1 + w_2.x_2 + b$ formundadır.

w_1 ve w_2; x_1 ve x_2'nin katsayılarıdır.

Regresyon modelinin amacı, en uygun w_1, w_2 ve **b** katsayılarını bulmaktır.

$\acute{y}$ değişkeni, bir evin fiyatı ise, x_1 değişkeni evin m^2 cinsinden alanı ve x_2 değişkeni de evin yaşı olabilir.