Necmettin Erbakan Üniversitesi
Bilgisayar Mühendisliği

Dr. Öğr. Üyesi Murat KARAKOYUN
(mkarakoyun@erbakan.edu.tr)

Ders: Veri Önişlemeye Giriş
[Konu: Eksik Veri Tamamlama]

Sunu İçeriği

- Eksik Veri Nedir?
- Eksik Veri Sebepleri?
- Eksik Veri Kullanma Sonuçları?
- Eksik Veri İşleme Yöntemleri
  - Eksik Veriyi Silme
  - Ortalama (Mean) ile Veri Tamamlama
  - Mod (Mode) ile Veri Tamamlama
  - Medyan (Median) ile Veri Tamamlama

Eksik Veri Nedir?

Bir veri kümesindeki veri örnekleri için bir (veya birden fazla) niteliğin değerinin bilinmemesi durumudur.

Ad	Soyad	Meslek	Yas	Maas
Ahmet	Aslan	Muhasebeci	35	3750
Ayşe	Naz	?	32	4000
Ali	Deniz	Yazılımcı	?	4500
Rıza	Büyük	Öğretmen	26	?
Merve	Kaş	?	45	2500

Eksik Veri Sebepleri?

- Tutarsız veri olması sebebiyle silinmiş olması
- Yanlış anlaşılma nedeniyle veri girilmemiş olması
- Bazı niteliklerin veri giriş esnasında önemsiz kabul edilmiş olması
- Veri üzerinde yapılan bazı güncellemeler sonrası verinin silinmiş olması
- İnsan, donanım veya yazılımsal başka problemlerin olması

Eksik Veri Sonuçları?

Eksik/kayıp verilere sahip veri seti; analiz sayesinde elde edeceğiniz sonuçların **güvenilirliğini**, **tutarlılığını** etkileyecek, hedef kitlenizi temsil eden örneklemin **temsil gücünü** düşürecek ve **yanlış çıkarımlar** yapmanıza neden olacaktır.

Ayrıca birçok makine öğrenmesi algoritması, analiz için kullanılacak veri setinde kayıp değerler olması halinde kullanılamayacaktır.

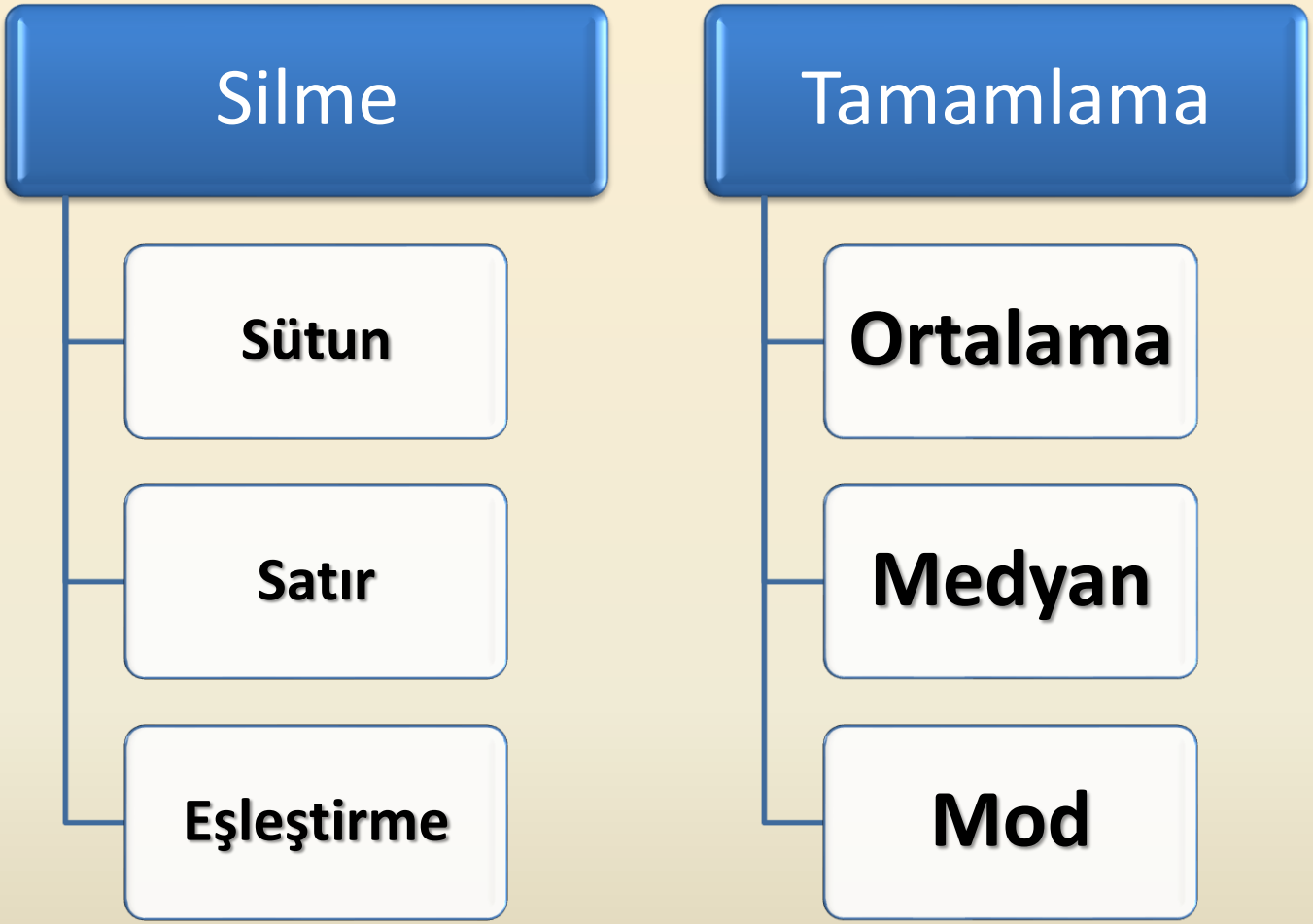
Eksik Veri Nasıl İşlenir?

Yukarıdaki olumsuzlukların yaşanmaması için veri setine herhangi bir model uygulamadan önce kayıp veri analizi yapılarak, elde edilen bulgulardan sonra eksik verinin yarattığı sorunu gidermek için ön işlem uygulanması gerekmektedir.

1. Veri kümesindeki eksik veriler görmezden gelinerek **silinir.**

2. Veri kümesindeki eksik veriler farklı yaklaşımlar ile tamamlanır.

Eksik Veri Nasıl İşlenir?



Silme – Sütun Bazlı Silme

- Herhangi bir niteliğin veri setinden tamamen silinmesidir.

- Niteliğin büyük bir çoğunluğunun (%60 ve fazlası) eksik değerlerden oluşması ve yapılacak analizde önemsiz bir yeri olması durumunda tercih edilebilir.

Silme – Sütun Bazlı Silme

N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	Out
1091262	2	5	3	3	6	7	7	?	1	4
1096800	6	6	6	9	6	?	7	?	1	2
1099510	10	4	3	1	3	3	6	5	2	4
1100524	6	10	10	2	8	10	7	?	3	4
1102573	5	6	5	6	10	1	3	1	1	4
1103608	10	10	10	?	8	1	8	?	1	4
1103722	1	1	1	1	2	1	2	?	2	2
1105257	3	7	7	4	4	?	4	8	1	4
1105524	1	1	1	1	2	1	2	?	1	2

[%11] [%22] [%66]

Silme – Sütun Bazlı Silme

Avantaj:

- Yapılacak/yapılan analize göre daha doğru değişkenlerle çalışma fırsatı verir.

Dezavantaj:

- Silinen niteliğin yapılan/yapılacak analiz için önemsiz olduğundan emin olunması gerekiyor.

Silme – Satır Bazlı Silme

- Kayıp/eksik veri işlemede en yaygın olarak kullanılan yöntem kayıp olan tüm gözlemleri görmezden gelerek, tam olan gözlemlerle ilerlemektir.

- Bunun için de gözlemde (veri kaydında) bir veya daha fazla eksik değer bulunması durumunda tüm satır silinir.

Silme – Satır Bazlı Silme

N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	Out
1091262	2	5	3	3	6	7	7	?	1	4
1096800	6	6	6	9	6	?	7	?	1	2
1099510	10	4	3	1	3	3	6	5	2	4
1100524	6	10	10	2	8	10	7	3	3	4
1102573	5	6	5	6	10	1	3	1	1	4
1103608	10	10	10	?	8	1	8	?	1	4
1103722	1	1	1	1	2	1	2	4	2	2
1105257	3	7	7	4	4	?	4	8	1	4
1105524	1	1	1	1	2	1	2	3	1	2

Silme – Satır Bazlı Silme

Avantaj:

- Uygulanabilirlik açısından basit ve hızlı bir çözümdür.

Dezavantaj:

- Yanlı tahminler üretebilir.
- Örneklem sayısını düşürdüğü için standart hatayı artırır ve testin gücünü düşürür.
- Analiz için kullanılacak modelde fazla nitelik dahil edilmesi durumunda, herhangi bir nitelikteki eksik veri nedeniyle gözlem silineceği için önemli miktarda veri kaybı yaşanabilir.

Silme – Eşleştirme Bazlı Silme

Analiz için kullanılacak değişkenler seçildikten sonra, seçilen değişkenler üzerinde eksik veriler temizlenir.

Silme – Eşleştirme Bazlı Silme

N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	Out
1091262	2	5	3	3	6	7	7	?	1	4
1096800	6	6	6	9	6	?	7	?	1	2
1099510	10	4	3	1	3	3	6	5	2	4
1100524	6	10	10	2	8	10	7	3	3	4
1102573	5	6	5	6	10	1	3	1	1	4
1103608	10	10	10	?	8	1	8	?	1	4
1103722	1	1	1	1	2	1	2	4	2	2
1105257	3	7	7	4	4	?	4	8	1	4
1105524	1	1	1	1	2	1	2	3	1	2

Silme – Eşleştirme Bazlı Silme

Avantaj:

- Satır bazlı silme işlemine kıyasla daha az veri kaybı yarattığı için testin gücü daha yüksektir.

Dezavantaj:

- Farklı analizler, veriden çekilen farklı alt gruplar kullanılarak yapılacağı için sonuçlar tutarlı olmayabilir.
- Eksik ve eksik olmayan veri grupları arasında sistematik bir fark varsa yanlı tahminler üretebilir.

Tamamlama – Ortalama Kullanarak

Bu yaklaşımda eksik verinin olduğu nitelik için ortalama bir değer elde edilerek eksik veri tamamlanır.

1. Nitelik bazlı ortalama kullanma

2. Sınıf + Nitelik bazlı ortalama kullanma

Tamamlama – Ortalama Kullanarak

Nitelik bazlı ortalama kullanma

N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	Out
1091262	2	5	3	3	6	7	7	? 4	1	4
1096800	6	6	6	9	6	?3	7	? 4	1	2
1099510	10	4	3	1	3	3	6	5	2	4
1100524	6	10	10	2	8	10	7	3	3	4
1102573	5	6	5	6	10	1	3	1	1	4
1103608	10	10	10	? 3	8	1	8	? 4	1	4
1103722	1	1	1	1	2	1	2	4	2	2
1105257	3	7	7	4	4	?3	4	8	1	4
1105524	1	1	1	1	2	1	2	3	1	2

M(N5) = 3

M(N7) = 3

M(N9) = 4

Tamamlama – Ortalama Kullanarak

Sınıf + Nitelik bazlı ortalama kullanma

N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	Out
1091262	2	5	3	3	6	7	7	? 4	1	4
1096800	6	6	6	9	6	?	7	? 2	1	2
1099510	10	4	3	1	3	3	6	5	2	4
1100524	6	10	10	2	8	10	7	3	3	4
1102573	5	6	5	6	10	1	3	1	1	4
1103608	10	10	10	?	8	1	8	? 4	1	4
1103722	1	1	1	1	2	1	2	1	2	2
1105257	3	7	7	4	4	?	4	8	1	4
1105524	1	1	1	1	2	1	2	3	1	2

$M(N9)_4 = 4$

$M(N9)_2 = 2$

Tamamlama – Medyan Kullanarak

Bu yaklaşımda eksik verinin olduğu nitelik için veri setinde o nitelik için mevcut medyan değeri hesaplanır ve eksik veriye atanır.

1. Nitelik bazlı medyan kullanma

2. Sınıf + Nitelik bazlı medyan kullanma

Tamamlama – Medyan Kullanarak

Nitelik bazlı medyan kullanma

N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	Out
1091262	2	5	3	3	6	7	7	? 4	1	4
1096800	6	6	6	9	6	?1	7	? 4	1	2
1099510	10	4	3	1	3	3	6	5	2	4
1100524	6	10	10	2	8	10	7	3	3	4
1102573	5	6	5	6	10	1	3	1	1	4
1103608	10	10	10	? 3	8	1	8	? 4	1	4
1103722	1	1	1	1	2	1	2	4	2	2
1105257	3	7	7	4	4	?1	4	8	1	4
1105524	1	1	1	1	2	1	2	3	1	2

Med(N5) = 3 Med(N7) = 1 Med(N9) = 4

Tamamlama – Medyan Kullanarak

Sınıf + Nitelik bazlı medyan kullanma

N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	Out
1091262	2	5	3	3	6	7	7	? 5	1	4
1096800	6	6	6	9	6	?	7	? 2	1	2
1099510	10	4	3	1	3	3	6	5	2	4
1100524	6	10	10	2	8	10	7	4	3	4
1102573	5	6	5	6	10	1	3	1	1	4
1103608	10	10	10	?	8	1	8	? 5	1	4
1103722	1	1	1	1	2	1	2	1	2	2
1105257	3	7	7	4	4	?	4	8	1	4
1105524	1	1	1	1	2	1	2	3	1	2

$Med(N9)_4 = 5$

$Med(N9)_2 = 2$

Tamamlama – Mod Kullanarak

Bu yaklaşımda eksik verinin olduğu nitelik için veri setinde o nitelik için frekansı en yüksek olan değer bulunur ve eksik veriye atanır.

1. Nitelik bazlı mod kullanma

2. Sınıf + Nitelik bazlı mod kullanma

Tamamlama – Mod Kullanarak

Nitelik bazlı mod kullanma

N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	Out
1091262	2	5	3	3	6	7	7	? 3	1	4
1096800	6	6	6	9	6	? 1	7	? 3	1	2
1099510	10	4	3	1	3	3	6	5	2	4
1100524	6	10	10	2	8	10	7	3	3	4
1102573	5	6	5	6	10	1	3	1	1	4
1103608	10	10	10	? 1	8	1	8	? 3	1	4
1103722	1	1	1	1	2	1	2	4	2	2
1105257	3	7	7	4	4	? 1	4	8	1	4
1105524	1	1	1	1	2	1	2	3	1	2

Mod(N5) = 1

Mod(N7) = 1

Mod(N9) = 3

Tamamlama – Mod Kullanarak

Sınıf + Nitelik bazlı mod kullanma

N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	Out
1091262	2	5	3	3	6	7	7	? 5	1	4
1096800	6	6	6	9	6	?	7	? 2	1	2
1099510	10	4	3	1	3	3	6	5	2	4
1100524	6	10	10	2	8	10	7	4	3	4
1102573	5	6	5	6	10	1	3	1	1	4
1103608	10	10	10	?	8	1	8	? 5	1	4
1103722	1	1	1	1	2	1	2	1	2	2
1105257	3	7	7	4	4	?	4	8	1	4
1105524	1	1	1	1	2	1	2	3	1	2

$Mod(N9)_4 = 5$

$Mod(N9)_2 = 2$

Tamamlama – Ortalama, Medyan ve Mod

Avantaj:

- Silme işlemine kıyasla veri kaybı yaşanmaz.

Dezavantaj:

- Tüm eksik verilere sabit bir değer ekleneceği için değişkenin varyans değerini yani değişkenliğini düşürecektir.
- Atama yapılan değişkenin varyans değerini düşürdüğü için, değişkenin dağılımına zarar verecektir.
- Değişkenler arasındaki ilişkileri ihmal ettiği için yöntemin uygulandığı değişkenlerle veri setinde bulunan diğer değişkenler arasındaki korelasyonu düşürür.

Özet

Bu sunu kapsamında bir veri setindeki eksik verilerin hangi sebeplerle oluşabileceği incelendi.

Veri setindeki eksik verilerin hangi yöntemler ile işlenebileceği ele alındı.