Veri Bilimi

Yapay Zekâ ve Veri Dersleri Rehberi · Kısım IV: Veri

Soru ve veri | Veri düzenleme | Keşifsel analiz | Görselleştirme | Çıkarım ve anlatı | Komşu dersler

Sercan KÜLCÜ · Giresun Üniversitesi · Bilgisayar Mühendisliği

© Sercan Külcü. Tüm hakları saklıdır. · Sunumun etkileşimli HTML sürümü: https://sercankulcu.github.io/files/ai_courses/sunumlar/Bolum_26_Veri_Bilimi.html

GİRİŞ

Ders Hangi Soruyu Sorar?

> **Veriden doğru soruları sorup karara dönüşen sonuçlar nasıl çıkarılır?**

Veri bilimi, bir soruyla başlayıp veriyi bulmayı, temizlemeyi, çözümlemeyi ve sonucu **karara dönüşecek** biçimde anlatmayı kapsayan süreçtir. İstatistik, programlama ve alan bilgisini birleştirir; veri madenciliği ve makine öğrenmesi bu sürecin araçlarıdır.

Bir bakışta

YAKLAŞIM	GİRDİ	ÇIKTI
soru, veri, analiz, anlatı	ham, dağınık veri	rapor, görselleştirme, model

BAŞARI ÖLÇÜTÜ	MATEMATİK YÜKÜ	VERİ İHTİYACI
kararlara katkı, doğruluk	orta: istatistik	orta ve çok

GİRİŞ

Kısa Tarihçe

1962
Tukey "The Future of Data Analysis" makalesinde veri analizini istatistikten ayrı bir bilim olarak savunur.

1977
Tukey "Exploratory Data Analysis" kitabını yayımlar: önce bak, sonra modelle.

2001
Cleveland "veri bilimi"ni istatistiği genişleten bir akademik alan olarak önerir.

2010
Kaggle kurulur; veri bilimi yarışmaları ve açık veri setleri yaygınlaşır.

2012
Harvard Business Review veri bilimciyi "21. yüzyılın en çekici mesleği" diye tanıtır.

2014
Jupyter projesi IPython'dan ayrılır; kod, grafik ve metin aynı defterde birleşir.

DERS

Temel Kavramlar

Soru tanımı
Ne bilinmek isteniyor, hangi karar verilecek?

Düzenli veri
Her satır bir gözlem, her sütun bir değişken (tidy data).

Veri temizleme
Eksik, hatalı ve tutarsız kayıtları düzeltmek.

Keşifsel analiz
Özet istatistik ve grafikle veriyi tanımak.

Görselleştirme
Doğru grafikle örüntüyü ve sonucu göstermek.

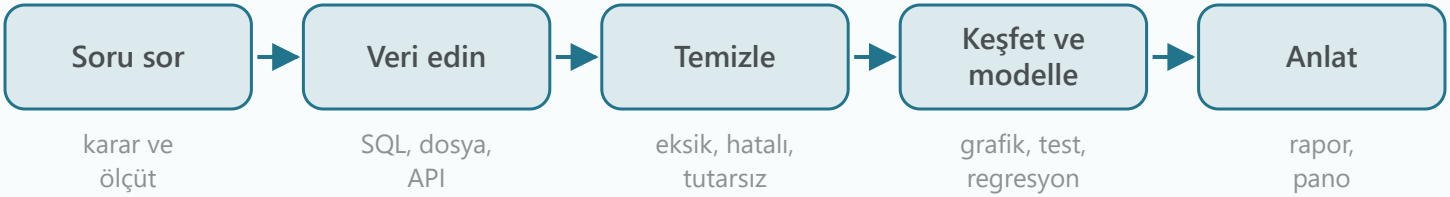
Hipotez testi
Farkın tesadüf olup olmadığını sınamak.

A/B testi
Rastgele deneyle iki seçeneği karşılaştırmak.

Tekrarlanabilirlik
Kod, veri ve sürümle aynı sonucu yeniden üretmek.

DERS

Tipik İş Akışı



Örnek: Bir belediye "otobüs seferi artırılmalı mı?" diye sorar. Kart geçiş kayıtları toplanır, hatalı saatler ayıklanır, hat ve saat bazında doluluk grafiklerle incelenir, yoğunluk regresyonla kestirilir ve sonuç tek sayfalık bir raporla karar vericiye sunulur.

DERS

Derste Neler İşlenir?

1. Veri bilimi süreci ve iyi soru sormak
2. Python ya da R ile veri okuma; düzenli veri
3. Veri temizleme: eksik ve aykırı değerler
4. Keşifsel veri analizi
5. Görselleştirme ilkeleri
6. Örnekleme ve betimsel istatistik
7. Güven aralıkları ve hipotez testleri
8. A/B testi ve deney tasarımı
9. Regresyon ve tahmin modelleri
10. Korelasyon, nedensellik, karıştırıcı değişkenler
11. Veritabanından veri çekmek: SQL
12. Tekrarlanabilir rapor ve sonucu anlatmak

Kazandırdığı beceriler

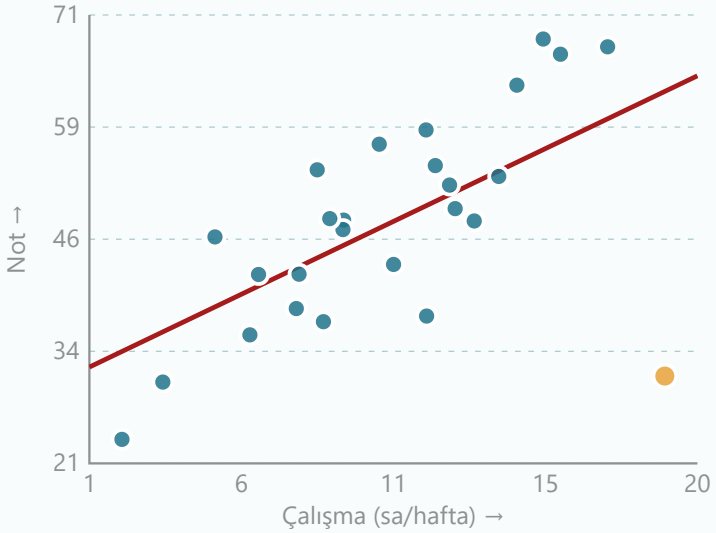
pandas ya da tidyverse | SQL | Matplotlib ve seaborn | hipotez ve A/B testini yorumlamak | Jupyter ile tekrarlanabilir rapor

Tipik bir dönemin akışıdır; sıra ve kapsam üniversiteden üniversiteye değişir.

DENE

Notları Ne Etkiliyor?

Okulun sorusu: "Öğrencilere hangi alışkanlık için öneri verelim?" Elde 26 öğrencinin anket ve not verisi var. Değişken seçip ilişkiye bakın; bir noktaya tıklayarak onu analizden çıkarın.



Yatay eksen (x)

Haftalık çalışma (saat)

Dikey eksen (y)

Dönem notu

Aykırı öğrenci dahil

n = 26 öğrenci
Pearson r = **0,61** (orta, artı)
Doğru: not ≈ 30,3 + 1,71 × çalışma

Turuncu nokta sınavda hastalanan öğrenci. Onunla r = 0,61, onsuz r = 0,84: tek bir kayıt sonucu değiştirebilir, önce veriyi tanıyın.

Veri bilimi bir **soruyla** başlar ve karara varır: keşifsel analiz, temizleme, ilişki ölçme ve sonucu anlatma. Korelasyon neden değildir: çok çalışan öğrenci derse de çok gelir; öneri vermeden önce bunu ayırmak gerekir.

KONUM

Ön Koşullar

Bu dersten önce bilinmesi yararlı olan dersler ve neden gerektikleri.

Ders	Neden gerekir
Olasılık ve İstatistik	hipotez testi, regresyon, görselleştirme
Programlama	Python ya da R ile dosya okuma, döngü ve fonksiyonlar
Veritabanı ve SQL	veriyi sorgulamak, birleştirmek ve gruplamak

KONUM

Bu Ders Sana Göre mi?

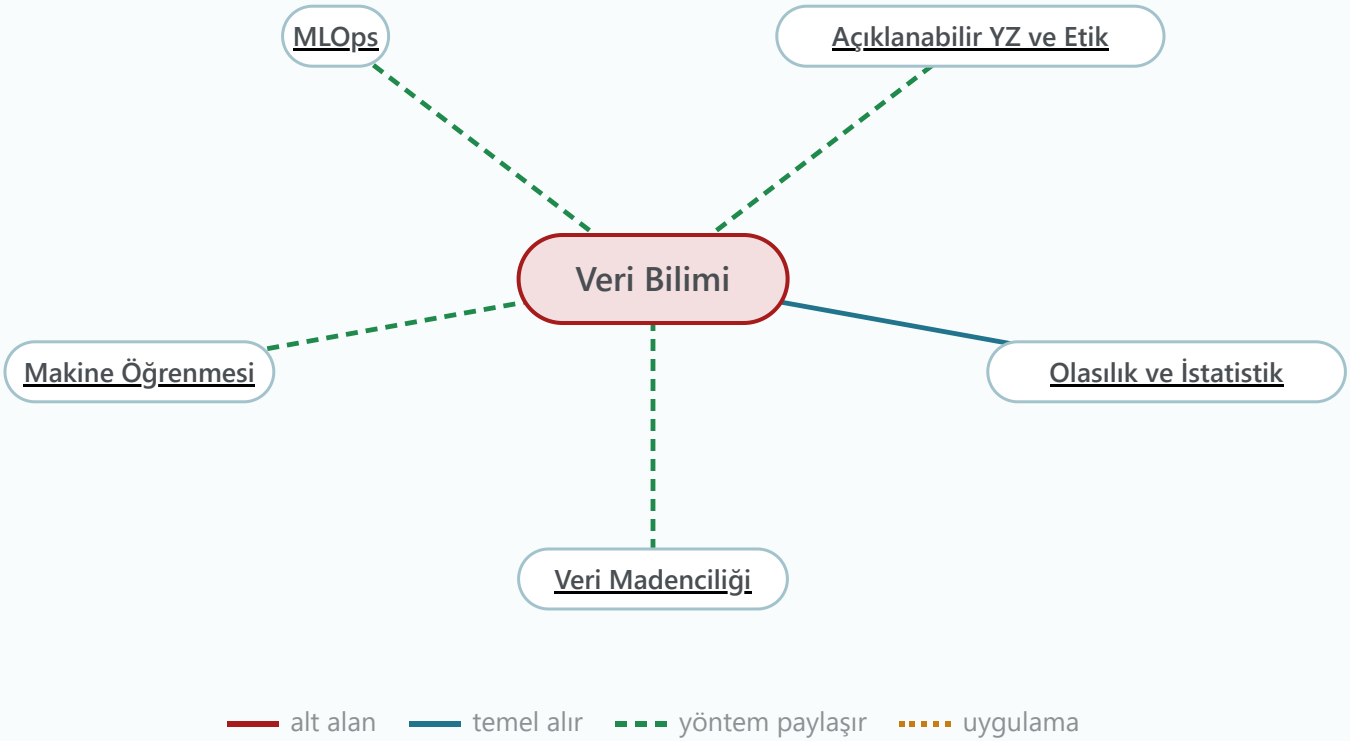
Kendine sor; "hayır" dediklerin için sağdaki sütun nereden başlayacağını gösterir.

Kendine sor	Hayırsa önce
Ortalama, ortanca, standart sapma ve normal dağılımı açıklayabiliyor musun?	Olasılık ve İstatistik
Python ya da R ile bir tablo okuyup sütunlara göre gruplayabiliyor musun?	Programlamaya giriş, ardından pandas ya da dplyr'a kısa bir giriş
SELECT, WHERE ve GROUP BY ile basit bir SQL sorgusu yazabiliyor musun?	Bir veritabanı dersinin SQL bölümü ya da çevrim içi SQL alıştırmaları
Bir grafiğin yanıltıcı olduğunu fark edebiliyor musun?	Şart değil, bunu ders öğretir; merak ediyorsan Tufte'nin kitabı

Çoğuna "evet" diyebiliyorsan derse hazırsın; bir ya da iki "hayır" dönem içinde kapanabilir.

KONUM

Dersin Komşuları



ÖRTÜŞME

Örtüştüğü Dersler

İlişki	Ders	Ortak konu
temel alır	Bu ders Olasılık ve İstatistik dersini temel alır	hipotez testi, regresyon, görselleştirme
yöntem paylaşır	Makine Öğrenmesi ile yöntem paylaşır	tahmin modelleri
yöntem paylaşır	Açıklanabilir Yapay Zekâ ve Yapay Zekâ Etiği ile yöntem paylaşır	veri yanlılığı, adalet ölçütleri
yöntem paylaşır	Veri Madenciliği ile yöntem paylaşır	keşifsel veri analizi
yöntem paylaşır	Makine Öğrenmesi Sistemleri ve MLOps ile yöntem paylaşır	deney takibi, veri sürümleme

Ayrıştığı Noktalar (1/2)

Ölçüt	Veri Bilimi	Veri Madenciliği	Makine Öğrenmesi
Temel soru	Veriden doğru soruları sorup karara dönüşen sonuçlar nasıl çıkarılır?	Büyük veri yığınlarındaki gizli ve işe yarar örüntüler nasıl bulunur?	Bir program örneklerden nasıl genelleme yapar?
Yaklaşım	soru, veri, analiz, anlatı	keşfe dayalı çözümleme	istatistiksel öğrenme
Girdi	ham, dağınık veri	büyük tablo ya da işlem verisi	etiketli ya da etiketsiz örnekler
Çıktı	rapor, görselleştirme, model	kural, küme, aykırı değer	tahmin yapan model
Başarı ölçütü	kararlara katkı, doğruluk	ilginçlik, destek, güven	görülmemiş veride hata
Matematik yükü	orta: istatistik	orta	yüksek: doğrusal cebir, olasılık, optimizasyon
Veri ihtiyacı	orta ve çok	çok	orta ve çok

- **Veri Madenciliği:** Veri madenciliği büyük veride örüntü bulmaya odaklanır; veri bilimi sorudan karara bütün süreci kapsar.
- **Makine Öğrenmesi:** MÖ yeni örnekte iyi tahmin eden modeli kurar; veri bilimi modeli bir soruyu yanıtlayan araçlardan biri olarak kullanır.

Ayrıştığı Noktalar (2/2)

Ölçüt	Veri Bilimi	Olasılık ve İstatistik	Büyük Veri
Temel soru	Veriden doğru soruları sorup karara dönüşen sonuçlar nasıl çıkarılır?	Belirsizlik nasıl ölçülür ve sınırlı veriden nasıl güvenilir sonuç çıkarılır?	Tek bir makineye sığmayan veri nasıl saklanır ve işlenir?
Yaklaşım	soru, veri, analiz, anlatı	olasılık modeli ve çıkarım	dağıtık depolama ve hesaplama
Girdi	ham, dağınık veri	örneklem verisi	çok büyük ve hızlı akan veri
Çıktı	rapor, görselleştirme, model	kestirim, güven aralığı, test sonucu	işlenmiş veri, sorgu sonucu
Başarı ölçütü	kararlara katkı, doğruluk	yansızlık, güç, anlamlılık	ölçeklenme, gecikme, maliyet
Matematik yükü	orta: istatistik	kuramın kendisi	düşük
Veri ihtiyacı	orta ve çok	örneklem	çok fazla

- **Olasılık ve İstatistik:** İstatistik çıkarımın kuramını verir; veri bilimi onu programlama, dağınık gerçek veri ve anlatıyla birleştirir.
- **Büyük Veri:** Büyük veri sistemleri verinin saklanıp işlenmesini sağlar; veri bilimi o veriden anlam ve karar çıkarır.

UYGULAMA

Uygulama Alanları

A/B testi
Yeni tasarım gerçekten daha mı iyi?

Pazarlama analitiği
Hangi kampanya müşteri getiriyor?

Halk sağlığı
Salgın ve hastalık eğilimlerini izlemek.

Spor analitiği
Oyuncu ve takım başarımını ölçmek.

Kamu politikası
Açık veriyle hizmetleri değerlendirmek.

Ürün metrikleri
Kullanıcıyı tutan ya da kaçıran ne?

UYGULAMA

Riskler ve Etik

Nedensellik yanılgısı

Korelasyondan neden çıkarmak yanlış kararlara götürür: dondurma satışı ile boğulmalar birlikte artar, çünkü ikisi de yazın olur.

p-hacking

Yeterince çok test yapılınca bir şey mutlaka "anlamlı" çıkar; hipotez veriye bakmadan önce belirlenmelidir.

Yeniden tanımlama

Adı silinmiş veride posta kodu, doğum tarihi ve cinsiyet çoğu kişiyi yeniden tanımlatmaya yeter.

Yanıltıcı grafik

Kesik bir eksen ya da seçilmiş bir zaman aralığı, aynı veriyle ters bir hikâye anlatabilir.

Ayrıntılı tartışma: Açıklanabilir Yapay Zekâ ve Yapay Zekâ Etiği.

UYGULAMA

Araçlar ve Örnek Veriler

Araçlar ve kütüphaneler

Python | pandas | Jupyter | R ve tidyverse | Matplotlib ve seaborn | SQL | statsmodels

Tanınmış veri setleri

Titanic (Kaggle) | Gapminder | NYC taksi yolculukları | Palmer Penguins | House Prices (Kaggle) | Anscombe dörtlüsü

UYGULAMA

Sık Karıştırılanlar

Yaygın düşünce	Doğrusu
Veri bilimi model eğitmektir.	Zamanın çoğu soruyu netleştirmeye, veriyi bulup temizlemeye ve sonucu anlatmaya gider.
Veri kendi kendine konuşur.	Soru, bağlam ve alan bilgisi olmadan aynı veri yanlış sonuca da götürebilir.
Anlamlı p değeri önemli sonuçtur.	İstatistiksel anlamlılık etkinin büyüklüğünü söylemez; çok sayıda test yanıltıcı sonuç üretir.
Grafik sonucu süsleyen bir eklentidir.	Anscombe dörtlüsü gösterir: özet istatistikleri aynı veriler grafikte bambaşka çıkar.

SONRASI

Bu Dersten Sonra

Bu dersi temel alan ya da genişleten dersler

Bu ders çoğunlukla uç noktadır; öteki dersleri temel alır.

İlgili meslekler

Veri bilimci | Veri analisti | Ürün analisti | İş zekâsı uzmanı | Araştırma analisti

Veri bilimi, veri kısmının çatısıdır. Buradan tahmin modellerini derinleştirmek için makine öğrenmesine, keşif için veri madenciliğine, ölçek için büyük veri sistemlerine ya da adalet ve yanlılık soruları için açıklanabilir YZ ve etiğe geçilebilir.

SONRASI

Kaynaklar

- Wickham, Çetinkaya-Rundel ve Grolemund, *R for Data Science*, 2. baskı, O'Reilly, 2023.
- McKinney, *Python for Data Analysis*, 3. baskı, O'Reilly, 2022.
- James, Witten, Hastie ve Tibshirani, *An Introduction to Statistical Learning*, 2. baskı, Springer, 2021.
- Tukey, *Exploratory Data Analysis*, Addison-Wesley, 1977.

Son

Bölüm 26: Veri Bilimi

← Önceki bölüm · Alanın haritası · Sonraki bölüm →