

# Data Science Bootcamp

## Açıklama

Bu beş gün süren atölye çalışması, Apache Spark 2 ve Hadoop ekosisteminin diğer önemli bileşenlerinin büyük ölçeklerde kullanıldığı veri bilimi ve makine öğrenimi iş akışlarını içerir. Atölye çalışması, gerçek dünyadaki iş zorluklarının üstesinden gelinebilmesi için veri bilimi ve makine öğrenimi yöntemlerinin kullanımının önemini gösterir.

Katılımcılar hayali bir teknoloji şirketinin ve veri setlerinin olduğu senaryoları kullanarak, kritik iş kararlarını desteklemek ve iş yapış şekillerini dönüştüren veri ürünleri geliştirmek üzere gerekli bilgileri keşfetmeye çalışırlar. Materyaller, bir dizi kısa sunumlar, etkileşimli gösteriler, kapsamlı uygulamalı alıştırmalar ve tartışmalar aracılığıyla sunulur.

Atölye çalışması, şu anda Python'u tek makinede küçük veri kümelerinde kullanan ve analizlerini ve makine öğrenimi modellerini dağıtık kümelerdeki büyük veri kümelerine genişletmek ihtiyacı duyan veri bilimcileri için tasarlanmıştır. Veri bilimi ve makine öğrenimi bilgisine sahip veri mühendisleri ve geliştiriciler de bu atölye çalışmasını faydalı bulabilir.

## Ön koşullar

Atölye çalışmasına katılanların, Python ile ilgili temel bilgilere sahip olmaları, veri araştırma, veri analizi, istatistiksel veya makine öğrenimi modeli geliştirme gibi konularda deneyimli olmaları gerekir. Hadoop veya Spark ile ilgili bilgi sahibi olunması gerekli değildir.

## Eğitim İçeriği

### Introduction to Python

Python Programlama Diline Bakış

Python Geliştirme Ortamları ve Ayarları

- Anaconda ve Python Geliştirme Ortamı (Spyder, Jupyter Notebook)
- Google Colaboratory (*Eğitim Sırasında Uygulamaların Gerçekleştirileceği Ortam*)

## VERİ BİLİMİ

### Veri Biliminin Temelleri

- Veri Analizi Nedir? Veri Analizi ile Neler Yapılabilir?
- Veri Bilimi Nedir?
- Veri Biliminin Unsurları Nelerdir?
- Veriden Faydalı Bilgi Çıkarma Aşamaları Nasıl İşler? (Veri Analitiği)
  - Tanımlayıcı Analitik (Descriptive Analytics)
  - Teşhis Edici Analitik (Diagnostic Analytics)
  - Tahmine Dayalı Analitik (Predictive Analytics)
  - Öngörüye Dayalı Analitik (Prescriptive Analytics)
- Veri Biliminin Kullanım Alanları

### Veri Bilimi İş Süreci Döngüsünde Nasıl Uygulanır

- CRISP-DM Metodolojisi
  - İş Anlayışı
  - Veriyi Anlamak
  - Verinin Hazırlanması
  - Modelleme
  - Değerlendirme
  - Kullanıma Sokma

### Veri Biliminde Uygulama Geliştirme Temel Aşamaları

### Veri Biliminde Uygulama Geliştirme İçin Kullanılan Araçlar

- Numpy
  - Array - Matris Oluşturma
  - Biçimlendirme İşlemleri (Yeniden Şekillendirme, Birleştirme, Bölme)
  - İndeks İşlemleri
  - Matematiksel İşlemler (Rasgele Sayı İşlemleri vb.)
  - İstatistiksel İşlemler (min, max, mean, std vb.)

- Pandas
  - Seri İşlemleri ( Seri Oluşturma, Özellikler vb.)
  - DataFrame İşlemleri (Oluşturma, Özellikler, Eleman İşlemleri, Birleştirme, Gruplama, Filtreleme, Apply, Pivot Tablolar)
  - Excel ve CSV dosyalardan DataFrame İle Veri Okuma ve Veri Üzerinde İşlemler
- Matplotlib/Seaborn
  - 2 Boyutlu Grafik Kullanımı (çizgi, çubuk, saçılım, histogram, pasta vb.)
  - 3 Boyutlu Grafik Kullanımı
  - Grafikler Üzerinde İşlemler Gerçekleştirme (Başlık ve Eksen Etiketleme, renk, legend vb. tanımlama)

## **Keşifçi Veri Analizi (Exploratory Data Analysis, EDA)**

### **Veri Okuryazarlığı**

- Veri Okuryazarlığı Nedir?
- Veri Okuryazarlığı Temel Kavramlar
  - Popülasyon ve Örneklem
  - Gözlem Birimi
  - Değişken Nedir? Değişken Türleri Nelerdir?
  - Ölçek Nedir? Ölçek Türleri Nelerdir?
  - Merkezi Eğilim Ölçüleri
    - Aritmetik Ortalama, Medyan, Mod, Karttiller
  - Dağılım Ölçüleri
    - Değişim Aralığı, Standart Sapma, Varyans, Çarpıklık, Basıklık
  - Veri Tanımlama
  - Verinin Organize Edilmesi ve İndirgenmesi
  - Verinin Gösterimi
  - Verinin Analizi ve Değerlendirme
  - Dosyadan Veri Seti Yükleme/Okuma
  - Verinin Boyutu Hakkında Bilgi Edinme
  - Veride Örneklem
  - Verideki Data Tipleri

### **Veri Ön İşleme ve Temizleme (Data Preparation and Cleaning)**

- Veri Tanıma
- Öznitelik Görüntüleme ve Seçme
- Sıralama ve Gruplama

- Öznitelikler Üzerinde İşlemler
  - Öznitelik Ekleme
  - Öznitelik Adı Değiştirme
  - Var Olan Öznitelikten Yeni Öznitelik Türetme
  - re (regular expression) modülünün Veri Analizinde Kullanımı
  - Öznitelik Silme
- Gözlemler Üzerinde İşlemler
  - Gözlemlerin Gösterimi (Baştan, Sondan, Rasgele)
  - Gözlem Ekleme
  - Gözlem Silme
- Veri Filtreleme
  - Sözlük ve Liste Kullanımı İle Filtreleme
  - Query İle Filtreleme
- Eksik Veriler
  - Eksik Verileri Tespit Etme
  - Eksik Verileri Silme Yaklaşımları
  - Eksik Verileri Tamamlama Yaklaşımları
    - Sabit Bir Değer İle Tamamla
    - Ortalama İle Tamamla
    - Bir Önceki ve Bir Sonraki Gözleme Ait Veri İle Tamamla
  - Eksik Veriler Üzerinde Oransal İşlemler
- Tekrar Eden Veriler
  - Tekrar Eden Verileri Tespit Etme
  - Tekrar Eden Verileri Temizleme
- Veri Dönüştürme İşlemleri
  - Veriyi Ölçeklendirme ve Normalizasyon,
  - Birleştirme-Toplama (Aggregation)
  - Kategorik Veriler
- Aykırı/Uç Verileri Tespit Etme

## Sayısal Veriler Üzerinde İstatistikî İşlemler

- Dağılım
- Varyans Analizi
- Korelasyon Analizi

## Veri Görselleştirme

- Grafik çizimi (çizgi, çubuk, pasta, ısı/heatmap vb.)

- Grafikler üzerinde işlemler yapma

## Hazır Veri Setleri Üzerinde Veri Analizi Uygulama Çalışmaları

## Hazır Veri Analizi Rapor Oluşturma Araçları

## MAKİNE ÖĞRENMESİ

### Makine Öğrenmesinin Temelleri

- Makine Öğrenmesi Nedir?
- Gerçek Hayat Örnekleri
- Temel Kavramlar ve Terminoloji
  - Problem Türleri (Regresyon, Sınıflandırma)
  - Model
  - Veri Setini Eğitim ve Test Olarak Bölme
  - Aşırı Öğrenme (overfitting)
  - Model Doğrulama
- Öğrenme Çeşitleri
  - Denetimli Öğrenme (Supervised)
  - Denetimsiz Öğrenme (Unsupervised)
  - Takviyeli/Pekiştirmeli Öğrenme (Reinforcement)
- Makine Öğrenmesi Modelleri Başarı Değerlendirme Yöntemleri

### Python İle Makine Öğrenmesi Gerçekleştirme Araçları

- Scikit-Learn Modülü ve Makine Öğrenmesi

### Makine Öğrenmesi Algoritmaları/Modelleri ve Uygulama Geliştirme

- Regresyon Modelleri (*Teori, Model, Tahmin*)
  - Basit Lineer
  - Çoklu Lineer
- Sınıflandırma Modelleri (*Teori, Model, Tahmin*)
  - Lojistik Regresyon ile Sınıflandırma
  - K-En Yakın Komşu (K-Nearest Neighbours - KNN) ile Sınıflandırma
  - Karar Ağaçları (Decision Tree - CART) ile Sınıflandırma
- Kümeleme Modelleri (*Teori, Model, Tahmin*) (*Denetimsiz Öğrenme Uygulaması*)
  - K-Means Algoritması ile Kümeleme