

# Veri Bilimi ve Büyük Veri Analitiği Eğitimi

Süre: 5 Gün

## Açıklama

Veri Bilimi ve Büyük Veri Analitiği Eğitimi, Python temellerinden büyük veri uygulamalarına uzanan 5 günlük Sanal Sınıf / Online bir programdır. Veri bilimi ve büyük veri alanında becerilerini geliştirmek, veri analizi ve raporlama süreçlerini öğrenmek isteyen profesyonellere yöneliktir.

Program; veri hazırlama, keşifçi veri analizi, görselleştirme, makine öğrenmesi ve PySpark ile büyük veri analitiğini birlikte ele alır. NumPy, Pandas, Matplotlib/Seaborn ve Scikit-Learn başlıkları özgün müfredatta yer alır. Eğitim sırasında uygulamalar için Google Colaboratory ortamı kullanılır.

## Eğitim İçeriği

### Python

#### 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*)

### Veri Bilimi

#### 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, Grublama, Filtreleme, Apply, Pivot Tablolar)
  - Excel ve CSV dosyalardan DataFrame ile 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
- Öznelik Görüntüleme ve Seçme
- Sıralama ve Gruplama
- Öznelikler Üzerinde İşlemler
  - Öznelik Ekleme
  - Öznelik Adı Değiştirme
  - Var Olan Öznelikten Yeni Öznelik Türetme
  - re (regular expression) modülünün Veri Analizinde Kullanımı
  - Öznelik 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

## Makine Öğrenmesi

### Makine Öğrenmesinin Temelleri

- Makine Öğrenmesi Nedir?
- Makine Öğrenmesi ile Derin Öğrenme Farkları Nelerdir?
- 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)
- Özellik Mühendisliği
  - Aykırı Değer Tespiti
  - Veri Temizleme
  - Veri Dönüştürülmesi (encoding scaling)
  - Veri Azaltma
  - Özellik Çıkarma (Feature Extraction)
- Makine Öğrenmesi Modelleri Başarı Değerlendirme Yöntemleri
- Karmaşıklık Matrisi (Confusion Matrix)
  - Accuracy, Recall, Precision
- R2 Score
- F1 Score
- AUC-ROC Curve
- Ortalama Mutlak Hata (Mean Absolute Error, MAE)
- Ortalama Kare Hatası (Mean Squared Error, MSE)

## 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
  - Destek Vektör Makinaları (SVM) ile Sınıflandırma
- Kümeleme Modelleri (*Teori, Model, Tahmin*) (Denetimsiz Öğrenme Uygulaması)
  - K-Means Algoritması ile Kümeleme
    - Elbow ile Optimum Küme Sayısının Belirlenmesi
  - Hiyerarşik Kümeleme
- Boosting (Tahmin) Modelleri (*Teori, Model, Tahmin*)
  - Gradient Boosting

- XBoost
- LightGBM

## Gerçek Zamanlı (Akan Veri Üzerinde) Makine Öğrenmesi Uygulaması

- Bir ML modelinin online bir veritabanı (firebase) üzerinden çekilen veri ile tahmin uygulaması geliştirme

## Makine Öğrenmesi Modellerinin Kayıt Edilmesi ve Başka Uygulamalara Transferi

- Veri Bilimi Modellerinin Farklı Uygulamalarda Kullanma Yöntemleri
- Pickle ile Modelin Transferi
- Joblib ile Modelin Transferi
- m2cgen ile native koda dönüştürülme

## Big Data Analytics

### BIG DATA

- Büyük Veri Nedir? (Big Data)
- Büyük Veri Bileşenleri ve Karakteristik Özellikleri
- Büyük Veri Kullanım Alanları – Gerçek Hayat Örnekleri
- Büyük Veri Çalışanların Sahip Olması Beklenen Özellikler
- Büyük Veri Uzmanlığı İçin Gerekli Yetenekler
- Büyük Veri Teknolojileri ve Araçları
  - Büyük Veri Mimarisinde Kullanılan Teknolojiler
  - Apache Hadoop Ekosistemi
  - Apache Spark Teknolojileri
- Dağıtık Mimarilerde Büyük Veri Uygulama Geliştirme Modeli

### BIG DATA ANALYTICS

- PySpark İle Büyük Veri Uygulama Geliştirme Süreçleri
- PySpark ile Büyük Veride Makine Öğrenmesi Uygulamaları

## Ön Koşullar

Herhangi bir ön koşul bulunmamaktadır.