

Module 1 - Defining Data Science

Bu notlar, IBM Data Science Professional Certificate sertifika programının **What is Data Science** eğitimi üzerinden alınan bilgiler temel alınarak hazırlanmıştır. Notların geliştirilmesinde internet kaynaklarından ve yapay zeka araçlarından da yararlanılmıştır.

1. Veri Biliminin Tanımı

Veri bilimi, disiplinler arası bir alan olarak tanımlanmakta; yapılandırılmış veya yapılandırılmamış verilerden anlamlı bilgiler, öngörüler ve eyleme dönüştürülebilir çözümler elde edilmesi amaçlanmaktadır. Bu alanın temelinde istatistik, bilgisayar bilimi, matematik, alan uzmanlığı ve veri mühendisliği gibi disiplinlerin birleşimi yer almaktadır. Temel amacı, verinin analiz edilmesi, modeller geliştirilmesi ve bu modellerin kullanılmasıyla karmaşık problemlere bilimsel yöntemlerle çözüm üretilmesidir.

Data Science In 5 Minutes | Data Science For Beginners | What Is Data Science? | Simplilearn

🔥 Data Scientist Masters Program (Discount Code - YTBE15) - [https://www.simplilearn.com/big-data-and-analytics/senior-data-scientist-masters-program-training?](https://www.simplilearn.com/big-data-and-analytics/senior-data-scientist-masters-program-training?utm_campaign=X3paOmcrTjQ&utm_medium=DescriptionFirstFold&utm_source=Youtube)

<https://www.youtube.com/watch?v=X3paOmcrTjQ>



What is Data Science?

Want a career in Data Science? Start here → <https://ibm.biz/BdK65F>
Data Science touches almost every operation of a business. See how → <https://ibm.biz/BdPEfu>

🔥 <https://www.youtube.com/watch?v=RBSUwFGa6Fk>



1.1 Temel Bileşenler

Veri biliminin temel bileşenleri şu şekilde özetlenmektedir:

1. Veri Toplama ve İşleme:

- Çeşitli kaynaklardan (sensörler, sosyal medya, veritabanları, IoT cihazları vb.) ham verinin toplanması sağlanmaktadır.
- Veri temizleme, dönüştürme ve depolama süreçleri (ETL: Extract, Transform, Load) uygulanmaktadır.

What is ETL (Extract, Transform, Load)?

Learn more about ETL → <http://ibm.biz/learn-about-etl>

Blog Post: ELT vs. ETL: What's the Difference? → <http://ibm.biz/elt-vs-etl-blog-post>

Check out Cloud Databases on IBM Cloud → <http://ibm.biz/IBMCLOUDDBs>

🔥 <https://www.youtube.com/watch?v=OW5OgsLpDCQ>



2. Veri Analizi ve Modelleme:

- İstatistiksel analiz, makine öğrenmesi (ML) ve yapay zekâ (AI) teknikleri kullanılarak verideki desenlerin keşfedilmesi sağlanmaktadır.
- Tahmine dayalı (predictive) ve açıklayıcı (prescriptive) modeller oluşturulmaktadır.

3. Görselleştirme ve İletişim:

- Elde edilen bulgular, grafikler, raporlar veya etkileşimli panolar vasıtasıyla anlaşılır biçimde sunulmaktadır.
- Teknik olmayan paydaşlara sonuçların aktarılması hedeflenmektedir.

4. Dağıtım ve Üretim:

- Geliştirilen modellerin gerçek dünya senaryolarına entegre edilmesi (API'lar, bulut sistemleri) sağlanmaktadır.
- Model performansının izlenmesi ve sürekli iyileştirilmesi amaçlanmaktadır.

1.2 Süreç ve Adımlar

Veri bilimi süreci, aşağıdaki adımlar üzerinden yürütülmektedir:

1. Sorun Tanımlama:

Organizasyonun çözmek istediği problemin netleştirilmesi en kritik adım olarak kabul edilmektedir.

2. Veri Toplama:

Verinin, yapılandırılmış (veritabanları, tablolar) veya yapılandırılmamış (log dosyaları, sosyal medya, sensör verileri) kaynaklardan toplanması sağlanmaktadır.

3. Analiz ve Modelleme:

Verinin farklı modeller kullanılarak incelenmesiyle, desenlerin, aykırı değerlerin (outliers) ve yeni bilgilerin keşfi gerçekleştirilmektedir.

4. İletişim ve Eylem:

Sonuçların, veri görselleştirme araçları kullanılarak paydaşlara aktarılması ve eyleme dönüştürülebilir önerilerin sunulması sağlanmaktadır.

1.3 Kullanılan Araç ve Teknolojiler

Veri bilimi çalışmalarında aşağıdaki araç ve teknolojilere başvurulmaktadır:

- **Programlama Dilleri:** Python, R, SQL.
- **Makine Öğrenmesi Kütüphaneleri:** TensorFlow, PyTorch, Scikit-learn.
- **Veri Görselleştirme:** Tableau, Power BI, Matplotlib.
- **Büyük Veri Teknolojileri:** Hadoop, Spark, Apache Kafka.
- **Bulut Platformları:** AWS, Google Cloud, Azure.

Bu araçların ve teknolojilerin detayları, ilerleyen notlarda ele alınmış olup, araştırılması gereken konuların incelenmesi kişinin kendisine bırakılmıştır.

1.4 Uygulama Alanları

Veri biliminin uygulama alanları, aşağıdaki sektörlerde ve alanlarda öne çıkmaktadır:

- **İş Zekâsı:** Müşteri davranışı analizi, pazar trendlerinin değerlendirilmesi.
- **Sağlık:** Hastalık teşhisi, ilaç keşif süreçlerinin desteklenmesi.
- **Finans:** Risk yönetimi ve dolandırıcılık tespitinde kullanılmaktadır.
- **Akıllı Şehirler:** Trafik optimizasyonu ve enerji verimliliği sağlanmaktadır.
- **Doğal Dil İşleme (NLP):** Chatbot'lar ve duygu analizi gibi uygulamalarda yer almaktadır.

Veri bilimi, veriden değer yaratma sanatı ve bilimi olarak nitelendirilmektedir. Teknik becerilerin yanında problem çözme yeteneği, merak ve etik sorumluluk da veri biliminin temel gereksinimleri arasında yer almaktadır. Günümüzde veriye dayalı karar alma kültürünün temel taşı olarak kabul edilmektedir.

2. Modül Sözlüğü

Terim	Tanım
Algoritmalar (Algorithms)	Bir problemi çözmek veya bir görevi tamamlamak için adım adım takip edilen talimatlar dizisi olarak tanımlanmaktadır.
Model (Model)	Verilerdeki ilişkileri ve kalıpları temsil ederek tahmin yapma veya karmaşık sistemleri analiz etme amacıyla oluşturulan soyutlama olarak kabul edilmektedir.
Aykırı Değerler (Outliers)	Bir veri kümesinde diğer veri noktalarından belirgin şekilde uzak kalan değerler olarak gözlemlenmekte; anormallikler, hatalar veya farklı bir olgunun göstergesi olabilmektedir.
Nicel Analiz (Quantitative Analysis)	Sayısal verilerin yorumlanması amacıyla matematiksel ve istatistiksel yöntemlerin sistematik kullanımı olarak tanımlanmaktadır.
Yapılandırılmış Veri (Structured Data)	Satır ve sütunlar halinde düzenlenmiş, önceden tanımlı bir şemaya sahip veri olarak ele alınmaktadır.
Yapılandırılmamış Veri (Unstructured Data)	Metin, görüntü, video gibi farklı içerikleri barındıran; önceden tanımlı bir veri modeline sahip olmayan ve geleneksel yöntemlerle analiz edilmesi zor veri türü olarak kabul edilmektedir.

3. Farklı Veri Dosya Türleri ve Formatları

Veri dosyalarının farklı türleri ve formatları aşağıdaki şekilde sınıflandırılmaktadır:

- **Delimited Text Files:**
Örneğin, CSV ve TSV formatlarında verilerin satır bazında metin olarak saklanması; her satırdaki değerlerin bir ayraç (delimiter) ile ayrılması sağlanmaktadır.
- **Microsoft Excel Open XML Spreadsheet (XLSX):**

Microsoft Excel tarafından kullanılan XML tabanlı dosya formatı; bir workbook içinde birden fazla worksheet bulunabilmekte ve her hücrede veri yer almaktadır.

- **Extensible Markup Language (XML):**

Hem insanlar hem de makineler tarafından okunabilir, belirli kurallara sahip işaretleme dili; platformlar arası veri paylaşımını kolaylaştırmaktadır.

- **Portable Document Format (PDF):**

Adobe tarafından geliştirilen, uygulama, donanım ve işletim sistemlerinden bağımsız olarak belgelerin aynı şekilde görüntülenmesini sağlayan dosya formatı; yasal ve finansal belgelerde yaygın olarak kullanılmaktadır.

- **JavaScript Object Notation (JSON):**

Web üzerinden yapılandırılmış veri iletimi için tasarlanmış metin tabanlı, dil bağımsız bir standart olarak kullanılmakta; birçok API ve web servisi tarafından tercih edilmektedir.

3.1 Delimited Text Files Detayları

- Metin dosyaları olarak saklanan bu dosyalarda, satır bazında veriler tutulmakta ve her satırdaki değerler bir ayraç (delimiter) ile ayrılmaktadır.
- Ayraç olarak virgül (,), sekme (tab), iki nokta üst üste (:), dikey çizgi (|) veya boşluk gibi karakterler kullanılabilir.
- **CSV (Comma-Separated Values)** ve **TSV (Tab-Separated Values)** formatları yaygın olarak kullanılmaktadır.
- İlk satır genellikle sütun başlıklarını içermekte, her sütunun farklı veri tiplerine sahip olabilmesi sağlanmaktadır.
- Delimited dosyaların, veri akışında sınırların belirlenmesi ve hemen hemen tüm uygulamalar tarafından işlenebilmesi avantajı bulunmaktadır.

3.2 Microsoft Excel Open XML Spreadsheet (XLSX) Detayları

- Microsoft Excel tarafından kullanılan bu XML tabanlı dosya formatında, **.XLSX** dosyaları workbook olarak adlandırılmakta; bir workbook içinde birden fazla worksheet yer almaktadır.
- Worksheet'ler, satır ve sütunlar halinde organize edilmekte ve her hücrede veri bulunmakta,
- Açık format olması sebebiyle, çoğu uygulama tarafından erişilebilir olması ve Excel fonksiyonlarının tam olarak kullanılabilmesi sağlanmaktadır.

3.3 Extensible Markup Language (XML) Detayları

- Belirli kurallara sahip işaretleme dili olan XML, hem insanlar hem makineler tarafından okunabilir şekilde tasarlanmıştır.
 - İnternet üzerinden veri gönderimi amacıyla kullanılmakta, HTML'ye benzer fakat ön tanımlı etiketlerin bulunmadığı bir yapı sunmaktadır.
 - Platform ve programlama dili bağımsızlığı sayesinde, farklı sistemler arasında veri paylaşımının kolaylaştırılmasına katkı sağlanmaktadır.
-

3.4 Portable Document Format (PDF) Detayları

- Adobe tarafından geliştirilen PDF formatı, uygulama, donanım ve işletim sistemlerinden bağımsız olarak belgelerin aynı şekilde görüntülenmesini sağlamak üzere kullanılmaktadır.
 - Genellikle yasal ve finansal belgelerde tercih edilmekte, form doldurma gibi işlemleri desteklemesiyle öne çıkmaktadır.
-

3.5 JavaScript Object Notation (JSON) Detayları

- JSON, web üzerinden yapılandırılmış veri iletimi için tasarlanmış, metin tabanlı açık standart bir formattır.
- Dil bağımsız olması nedeniyle hemen hemen her programlama dili tarafından okunabilir; kullanım kolaylığı ve tarayıcı uyumluluğu sayesinde tercih edilmektedir.
- API'ler ve web servisleri tarafından verinin paylaşımında yaygın olarak kullanılmaktadır.