Module 3 - Libraries, APIs, Datasets & Models

Bu notlar, veri bilimi projelerinde kritik rol oynayan kütüphaneler, API'ler, framework'ler, veri setleri ve modeller hakkında kapsamlı bilgiler sunulmaktadır. Notların hazırlanmasında, IBM Data Science Professional Certificate içeriği, internet kaynakları ve yapay zeka araçlarının sunduğu bilgiler referans alınmış; her bir bileşenin kullanım amaçları, entegrasyon yöntemleri ve avantajları göz önünde bulundurularak düzenlenmiştir.

---

1. Kütüphaneler

Veri bilimi projelerinde, kodun yeniden kullanılabilir ve optimize edilmiş yapılarla geliştirilmesi amacıyla kütüphanelerin kullanımı kritik öneme sahiptir.

1.1. Python Kütüphaneleri

Python kütüphanelerinin, veri hazırlama, analiz, görselleştirme ve modelleme süreçlerinde zaman tasarrufu, topluluk desteği ve güvenilirlik sağladığı vurgulanmaktadır.

- **Genel Özellikler:**Kütüphanelerin çoğunun PyPI üzerinden (pip veya conda yardımıyla) yönetildiği, topluluk desteği ve düzenli güncellemelerle sürdürülebilir geliştirme imkânı sunduğu göz önünde bulundurulmaktadır.

1.2. Bilimsel Hesaplama Kütüphaneleri

- **NumPy:**

  Diziler (ndarray) ve matrisler üzerinde yüksek performanslı matematiksel işlemlerin gerçekleştirilmesi sağlanmakta, pek çok kütüphanenin temelini oluşturmaktadır.

- **Pandas:**

DataFrame yapısı ile veri temizleme, manipülasyon ve analizin kolaylaştırılması, eksik veri işleme ve birleştirme işlemlerinin optimize edilmesi mümkün kılınmaktadır.

1.3. Görselleştirme Kütüphaneleri

- **Matplotlib:**

  Çizgi, çubuk, histogram gibi temel grafiklerin oluşturulması ve düşük seviyede detaylı özelleştirmelerin yapılması sağlanmaktadır.

- **Seaborn:**

  İstatistiksel görselleştirme için şık temalar ve gelişmiş grafik seçenekleri sunulmakta, verinin görsel anlatımının güçlendirilmesi desteklenmektedir.

1.4. Makine Öğrenimi ve Derin Öğrenme Kütüphaneleri

- **Scikit-learn:**

  Regresyon, sınıflandırma, kümeleme gibi temel makine öğrenimi algoritmalarının yanı sıra veri ön işleme ve model değerlendirme adımlarını kapsayan bütüncül fonksiyonların sunulması sağlanmaktadır.

- **TensorFlow:**

  Derin öğrenme modellerinin eğitimi ve karmaşık sinir ağlarının oluşturulması, Keras arayüzü sayesinde hızlı prototipleme imkânı sunulmaktadır.

- **PyTorch:**

  Dinamik grafik yapısı ile araştırma ve prototipleme çalışmalarında esnek kullanım sağlanmakta, model geliştirme süreçlerine kolaylık kazandırılmaktadır.

1.5. Diğer Dillerdeki Kütüphaneler

Python dışındaki dillerde de veri bilimi projelerine katkıda bulunan kütüphaneler bulunmaktadır:

- **Scala:**
  - **Vegas:** İstatistiksel veri görselleştirme desteği sunulmakta,
  - **BigDL:** Apache Spark ile bütünleşik derin öğrenme desteği sağlanmaktadır.

- **R:**
  - **ggplot2:** Zengin ve estetik veri görselleştirme seçenekleri sunulmakta,
  - **Keras** ve **TensorFlow:** R üzerinden derin öğrenme modellerinin geliştirilmesine olanak tanınmaktadır.

---

2. API'ler

API'ler, farklı yazılım bileşenleri arasında iletişim ve entegrasyonu sağlayan tanımlı arayüzler olarak kullanılmaktadır.

2.1. API'nin Tanımı ve Kullanım Amaçları

- **Tanım:**

  API, bir kütüphanenin fonksiyonlarına, web servislerinin özelliklerine veya veritabanlarının erişim yöntemlerine dair tanımlanmış kontratlar olarak değerlendirilmektedir.

- **Kullanım Amaçları:**

  Soyutlama sağlanması, bileşenler arasındaki entegrasyonun bozulmadan sürdürülebilmesi ve farklı dillerde yazılmış uygulamalar arasında iş birliğinin desteklenmesi amaçlanmaktadır.

2.2. API Türleri ve Örnekleri

- **Kütüphane API'leri:**

  Örneğin, Pandas API'si DataFrame'ler üzerinde işlem yapmayı mümkün kılmaktadır.

- **Web Servis API'leri:**
  - **REST API:** JSON formatında veri alışverişi sağlanmakta, modern mikro servis mimarilerinde yaygın olarak kullanılmaktadır.
  - **SOAP API:** XML temelli yapısıyla, daha katı sözleşmeler gerektiren kurumsal uygulamalarda tercih edilmektedir.

---

3. Framework’ler

Framework’ler, belirli alanlarda modüler yapı, standartlaşma ve sürdürülebilir geliştirme imkânı sunan yazılım çatıları olarak değerlendirilmektedir.

3.1. Framework’ün Tanımı ve Kullanım Amaçları

- **Tanım:**

  Belirli modülleri, kütüphaneleri ve kuralları bir arada sunarak projelerin yapılandırılmasını destekleyen iskelet çözümleri sağlanmaktadır.

- **Kullanım Amaçları:**

  Standartlaşma ve modülerlik sağlanarak tekrarlayan yapıların belirli bir çerçevede tutulması, geniş topluluk desteği ve dokümantasyon aracılığıyla sürdürülebilir geliştirme desteklenmektedir.

3.2. Popüler Framework Örnekleri

- **Web Geliştirme:**

  Django (Python), Ruby on Rails (Ruby) ve Spring (Java) gibi framework’ler, veritabanı işlemleri, şablon motoru ve kullanıcı yönetimi gibi ihtiyaçların hızlıca karşılanmasını sağlamaktadır.

- **Büyük Veri ve Veri İşleme:**

  Apache Spark, dağıtık veri işleme yetenekleri ile devasa veri kümelerinin paralel olarak işlenmesine olanak tanımakta; Spark MLlib ve Spark SQL modülleri ile makine öğrenimi ve veri sorgulama işlemleri desteklenmektedir.

- **Makine Öğrenimi ve Derin Öğrenme:**

  TensorFlow Extended (TFX), PyTorch Lightning ve FastAPI + PyTorch gibi çözümler, model eğitimi, dağıtımı ve izleme süreçlerinde entegre yaklaşımlar sunmaktadır.

---

4. Veri Setleri

Veri setleri, model eğitimi ve değerlendirilmesi için temel “yakıt” işlevi görmekte olup, kaliteleri, çeşitliliği ve lisans koşulları dikkatle değerlendirilmelidir.

4.1. Veri Setlerinin Önemi ve Kullanım Amaçları

- **Önemi:**

  Kaliteli ve çeşitli veri setlerinin kullanılması, model performansının artırılmasında temel rol oynadığı vurgulanmaktadır.

- **Kullanım Amaçları:**

  Model eğitimi, test işlemleri ve veri ön işleme süreçlerinin optimize edilmesi desteklenmektedir.

4.2. Veri Sahipliği ve Lisans Türleri

- **Özel Veri (Proprietary):**

  Kurumlara ait, kamuya kapalı veriler; gizlilik ve telif hakları açısından belirleyici koşullar içermektedir.

- **Açık Veri (Open Data):**

  Herkesin erişimine sunulan veri setleri, Community Data License Agreement (CDLA) veya Creative Commons gibi lisanslar altında paylaşılmaktadır.

4.3. IBM Data Asset Exchange (DAX)

IBM Data Asset Exchange, yüksek kaliteli açık veri setlerinin CDLA lisansı kapsamında sunulduğu bir platform olarak değerlendirilmektedir. DAX, görüntü, metin, zaman serisi gibi farklı veri tiplerine erişim sağlanması ve örnek defterler ile veri temizleme adımlarının desteklenmesi açısından avantajlar sunmaktadır.

---

5. Makine Öğrenimi ve Derin Öğrenme

Veri bilimi projelerinde veriden anlam çıkarım ve öngörü oluşturma süreçlerini destekleyen makine öğrenimi ve derin öğrenme algoritmaları, güçlü modellerin geliştirilmesinde kullanılmaktadır.

5.1. Makine Öğrenimi

Makine öğrenimi, verilerdeki desenlerin otomatik olarak öğrenilmesi ve bu desenlerin yeni verilere uygulanması amacıyla kullanılan algoritmalar bütünüdür.

- **Alt Kategoriler:**
  - **Denetimli Öğrenme:** Etiketli veriler üzerinden model eğitimi gerçekleştirilmektedir.
  - **Denetimsiz Öğrenme:** Etiketsiz veriler üzerinde kalıp bulma işlemleri uygulanmaktadır.
  - **Pekiştirmeli Öğrenme:** Ajanın ödül veya ceza alarak öğrenmesi desteklenmektedir.

5.2. Derin Öğrenme

Derin öğrenme, yapay sinir ağları kullanılarak, özellikle görüntü işleme, konuşma tanıma ve doğal dil işleme gibi karmaşık problemlerin çözümünde etkili sonuçlar elde edilmesini sağlamaktadır.

- **Gereksinimler:**

  Büyük veri setleri ve yüksek işlem gücü (GPU, TPU) gerektirdiği dikkate alınmaktadır.

- **Kullanılan Kütüphaneler:**

  TensorFlow, PyTorch ve Keras gibi derin öğrenme kütüphaneleri, modelin katman sayısının artırılması ve öğrenme gücünün optimize edilmesi açısından tercih edilmektedir.

---

6. Model Asset Exchange (MAX)

IBM Model Asset Exchange, önceden eğitilmiş derin öğrenme modellerinin sunulduğu bir depo olarak değerlendirilmektedir. MAX, sıfırdan model eğitilmesi gerekmeksizin hazır modellerin kullanılmasıyla zaman kazandırmakta; Docker konteynerleri ve REST API aracılığıyla mikro servis mimarisi desteği sunulmaktadır.

- **Avantajları:**Hazır modellerin kullanılması sayesinde hızlı prototipleme imkânı sağlanmakta, açık kaynak lisansları ile ticari projelerde de kullanılabilirlik desteklenmekte; Docker ve Kubernetes gibi teknolojilerle ölçeklenebilir dağıtım mümkün kılınmaktadır.

---

7. Ekler

Aşağıda, ilgili kurs içeriğinde yer alan örnek veri tabanı ve kaynak bağlantıları paylaşılmaktadır. Daha kapsamlı öğrenim için kurs sitesinin incelenmesi önerilmektedir.

Hükümet Verileri:

- https://www.data.gov/
- https://www.census.gov/data.html
- https://data.gov.uk/
- https://www.opendatanetwork.com/
- https://data.un.org/

Finansal Veri Kaynakları:

- https://data.worldbank.org/
- https://www.globalfinancialdata.com/
- https://comtrade.un.org/
- https://www.nber.org/
- https://fred.stlouisfed.org/

Suç Verileri:

- https://www.fbi.gov/services/cjis/ucr
- https://www.icpsr.umich.edu/icpsrweb/content/NACJD/index.html
- https://www.drugabuse.gov/related-topics/trends-statistics
- https://www.unodc.org/unodc/en/data-and-analysis/

Sağlık Verileri:

- https://www.who.int/gho/database/en/
- https://www.fda.gov/Food/default.htm
- https://seer.cancer.gov/faststats/selections.php?series=cancer
- https://www.opensciencedatacloud.org/
- https://pds.nasa.gov/
- https://earthdata.nasa.gov/
- https://www.sgim.org/communities/research/dataset-compendium/public-datasets-topic-grid

Akademi ve İş Dünyası Verileri:

- https://scholar.google.com/
- https://nces.ed.gov/
- https://www.glassdoor.com/research/
- https://www.yelp.com/dataset

Diğer Genel Veriler:

- https://www.kaggle.com/datasets
- https://www.reddit.com/r/datasets/

Özel Veri Tabanları ve Kaynakları:

Özel veri kümeleri, belirli kuruluşlara ait olup lisans anlaşmaları kapsamında dağıtılmaktadır.

- **Sağlık:**https://www.sgim.org/communities/research/dataset-compendium/proprietary-datasets
- **Finansal Market Verileri:**https://datarade.ai/data-categories/proprietary-market-data
- **Google Cloud Verileri:**https://cloud.google.com/datasets

---

8. Özet

Bu notlarda, veri bilimi projelerinde kullanılan kütüphaneler, API'ler, framework'ler, veri setleri ve modeller hakkında kapsamlı bilgiler sunulmaktadır.

- **Kütüphaneler:** Proje geliştirme sürecini hızlandıran, yeniden kullanılabilir yapıların desteklenmesiyle zaman tasarrufu ve güvenilirlik sağlanması vurgulanmaktadır.
- **API'ler:** Yazılım bileşenleri arasında soyutlama, esneklik ve entegrasyonun korunması amacıyla kullanılmaktadır.
- **Framework'ler:** Standartlaşma, modülerlik ve sürdürülebilir geliştirme imkânı sağlayarak projelere yapı kazandırılmaktadır.
- **Veri Setleri:** Model eğitimi ve değerlendirilmesinde "yakıt" görevi görmekte olup, kalite, çeşitlilik ve lisans koşulları dikkatle değerlendirilmelidir.
- **Makine Öğrenimi & Derin Öğrenme:** Veriden anlam çıkarım ve öngörü oluşturma süreçlerinde kullanılan algoritmaların desteklenmesi sağlanmakta, uygun kütüphaneler aracılığıyla etkili modeller geliştirilmektedir.
- **DAX ve MAX:** IBM tarafından sunulan, veri ve model paylaşımında projelere hız kazandıran kaynaklar olarak değerlendirilmekte; DAX veri setlerine, MAX ise önceden eğitilmiş modellere erişim imkânı sunmaktadır.

Her proje için uygun bileşenlerin seçilmesi, entegre bir yaklaşımın benimsenmesi ve sürekli öğrenme ile desteklenmesi, veri bilimi yolculuğunun başarılı bir şekilde ilerletilmesinde kritik öneme sahiptir.