Module 1 - Data Science Tools

Bu notlar, veri biliminin temel yapı taşlarını oluşturan araçlar, yöntemler ve platformlar hakkında bir rehber niteliğinde düzenlenmiştir. Notların hazırlanması sırasında, veri bilimi projelerinde yer alan her adımın – veri yönetiminden model dağıtımına kadar – farklı kategorilerde desteklendiği göz önünde bulundurulmuş; IBM Data Science Professional Certificate adresindeki "Tools for Data Science" eğitiminin içeriğinden yararlanılmış, ek olarak internet kaynakları ve yapay zeka araçlarının sunduğu bilgiler referans alınmıştır.

1. Veri Bilimi Araç Türleri ve Genel Bakış

Veri biliminde kullanılmak üzere geliştirilen araçlar üç ana kategori altında incelenmektedir:

- **Açık Kaynak Araçlar:**

  Python, R, Apache Spark, Jupyter Notebook ve GitHub gibi araçların kullanıldığı görülmekte; bu araçlara geniş topluluk desteği, sürekli güncellemeler ve esnek yapı özellikleri kazandırılmaktadır.

- **Ticari (Ücretli) Araçlar:**

  IBM Db2, Microsoft SQL Server, SPSS Modeler ve SAS Enterprise Miner gibi ürünler, büyük ölçekli kurumsal ihtiyaçların karşılanmasında, özel destek ve katma değerli özellikler sunması amacıyla tercih edilmektedir.

- **Bulut Tabanlı Araçlar:**

  IBM Watson Studio, Amazon SageMaker ve Microsoft Azure Machine Learning gibi platformların kullanılmasıyla, donanım ve altyapı kurulum gereksiniminin ortadan kaldırılması, esneklik, ölçeklenebilirlik ve yüksek erişilebilirlik sağlanması amaçlanmaktadır.

Araç seçiminin, projenin ihtiyaçları, ölçeği ve ekibin deneyimi göz önünde bulundurularak yapıldığı; açık kaynak ve bulut tabanlı araçların popülerliğine karşın, büyük kurumlarda ticari platformların sunduğu güvenlik ve destek avantajlarının tercih edilebildiği değerlendirilmiştir.

2. Veri Bilimi Görev Kategorileri

Veri bilimi sürecinde gerçekleştirilen görevler aşağıdaki başlıklar altında sınıflandırılmaktadır:

- **Veri Yönetimi:**

  Verilerin güvenli, ölçeklenebilir biçimde toplanması, depolanması, işlenmesi ve gerektiğinde erişimin sağlanması hedeflenmektedir.

- **Veri Entegrasyonu ve Dönüşümü (ETL – Extract, Transform, Load):**

  Farklı veri kaynaklarından verinin çekilmesi, uygun formatta dönüştürülmesi ve hedef ortama yüklenmesi işlemleri gerçekleştirilmektedir.

- **Veri Görselleştirme:**

  Analiz sonuçlarının grafik, tablo ve harita gibi görsel formatlarda sunulmasıyla, içgörülerin anlaşılır kılınması sağlanmaktadır.

- **Model Oluşturma (Modeling):**

  Makine öğrenimi, derin öğrenme veya istatistiksel yöntemler kullanılarak, veri setlerinden bilgi çıkarılması ve tahmin yapılması amaçlanmaktadır.

- **Model Dağıtımı:**

  Eğitilen modellerin üretim ortamına entegre edilerek, gerçek zamanlı ya da toplu işlem senaryolarında kullanılabilmesi sağlanmaktadır.

- **Model İzleme ve Değerlendirme:**

  Üretim ortamında çalışan modellerin performansının düzenli olarak ölçülmesi, gerektiğinde yeniden eğitilmesi veya iyileştirilmesi amaçlanmaktadır.

Her aşamada kullanılacak araçların farklılık göstermesi sebebiyle, doğru araç kombinasyonunun seçilmesinin projenin başarısında kritik bir rol oynadığı vurgulanmaktadır.

3. Veri Yönetimi

Veri yönetimi, projelerin başlangıç aşaması olarak değerlendirilmektedir. Verilerin saklanma biçimi, formatı, yedeklenmesi ve erişim süreçlerinin nasıl sağlandığı konusu, projenin genel başarısını doğrudan etkilemektedir.

- **Açık Kaynak Veri Yönetimi Araçları:**
  - **MySQL:** İlişkisel veritabanı sistemi; hafif yapısı, topluluk desteği ve ücretsiz olması sayesinde yaygın olarak tercih edilmektedir.
  - **PostgreSQL:** JSON desteği, tam metin arama ve konumsal veri (GIS) özellikleri ile zengin bir ekosistem sunmaktadır.
  - **MongoDB:** Doküman tabanlı NoSQL veritabanı; JSON benzeri yapı (BSON) kullanılarak esnek veri saklama imkânı sağlanmaktadır.
  - **Apache Cassandra:** Dağıtık NoSQL veritabanı; yüksek ölçeklenebilirlik ve hata toleransı özellikleriyle kullanılmaktadır.
  - **CouchDB:** JSON tabanlı belge odaklı veritabanı; replikasyon ve kolay ölçeklenebilirlik özellikleriyle öne çıkmaktadır.
  - **ElasticSearch:** Büyük veri kümelerinde gerçek zamanlı arama ve analiz amacıyla tercih edilmektedir.
  - **Hadoop Distributed File System (HDFS):** Büyük veri işlemleri için dağıtık dosya sistemi olarak kullanılmaktadır.
  - **Ceph:** Nesne, blok ve dosya düzeyinde depolamayı entegre eden açık kaynak dağıtık depolama platformu olarak değerlendirilmektedir.
- **Ticari Veri Yönetimi Araçları:**
  - **IBM Db2, Oracle Database, Microsoft SQL Server:** Büyük kurumsal ortamlarda güvenlik ve müşteri desteği avantajı sağlamak üzere kullanılmaktadır.
  - **IBM Db2 as a Service:** Bulut üzerinden SaaS olarak sunulan veritabanı çözümü tercih edilmektedir.

Veri yönetimi seçiminde, maliyet, hız ve güvenilirlik gereksinimlerinin titizlikle değerlendirilmesi esas alınmaktadır.

---

4. Veri Entegrasyonu ve Dönüşümü (ETL)

Ham verinin analiz edilebilir hale getirilmesi amacıyla, farklı kaynaklardan elde edilen verilerin çekilmesi, dönüştürülmesi ve hedef ortama yüklenmesi işlemleri gerçekleştirilmektedir.

- **Extraction (Çıkarma):**

  Sosyal medya API'leri, veri tabanları, sensörler gibi çeşitli kaynaklardan verinin çekilmesi sağlanmaktadır.

- **Transformation (Dönüştürme):**

  Birim dönüşümü, veri temizleme, normalizasyon ve aykırı değer tespiti işlemleri uygulanmaktadır.

- **Loading (Yükleme):**

  Dönüştürülen verinin, veri ambarı gibi hedef ortamlara yüklenmesi sağlanmaktadır.

Popüler ETL araçları arasında şu araçlar öne çıkmaktadır:

- **Apache Airflow:** İş akışlarının zamanlanması, takibi ve yönetilmesi için kullanılmaktadır.
- **Kubeflow:** Kubernetes üzerinde makine öğrenimi iş akışlarının ölçeklendirilmesi sağlanmaktadır.
- **Apache Kafka:** Yüksek hacimli gerçek zamanlı veri işleme amacıyla dağıtık yayın/abonelik platformu olarak tercih edilmektedir.
- **Apache NiFi:** Web tabanlı arayüzü sayesinde veri akış yönetiminin kolaylaştırılmasına katkı sunmaktadır.
- **Apache Spark SQL:** Büyük veri kümeleri üzerinde dağıtık SQL sorgulama işlemlerinin gerçekleştirilmesini sağlamaktadır.
- **Node-RED:** Sürükle-bırak mantığıyla, düşük kaynak tüketimi ile veri iş akışlarının görsel olarak düzenlenmesine olanak tanımaktadır.

Bu araçların, veri kaynaklarının çeşitlenmesi ve veri boyutlarının artması sebebiyle büyük ölçekli projelerde önem arz ettiği göz önünde bulundurulmaktadır.

---

5. Veri Görselleştirme

Veri bilimi projelerinde elde edilen bulguların etkili bir şekilde sunulabilmesi amacıyla, verilerin grafik, tablo ve harita gibi görsel formatlarda ifade edilmesi sağlanmaktadır. Bu sayede, elde edilen içgörülerin karar vericiler tarafından kolayca anlaşılması hedeflenmektedir.

- **Kod Tabanlı Görselleştirme:**
  - **PixieDust:** Jupyter Notebook ile entegre edilerek etkileşimli görselleştirmeler yapılmaktadır.
  - **Hue:** Hadoop veri kümelerinin sorgulanması ve görselleştirilmesi amacıyla kullanılmaktadır.
  - **Kibana:** Elasticsearch verileri üzerinden görselleştirme ve keşif işlemleri gerçekleştirilmektedir.
  - **Apache Superset:** SQL tabanlı analiz ile geniş görselleştirme seçenekleri sunan modern bir iş zekası uygulaması olarak değerlendirilmektedir.
- **Ticari Araçlar:**
  - **Tableau:** Sürükle-bırak özellikleri, etkileşimli panolar ve zengin görselleştirme seçenekleri ile tercih edilmektedir.
  - **Microsoft Power BI:** Microsoft ekosistemine entegre edilerek kolay kullanım sağlanmaktadır.
  - **IBM Cognos Analytics:** Kurumsal ortamlarda ölçeklenebilir iş zekası çözümleri sunulmaktadır.

Veri görselleştirme sürecinde, sunulan grafiklerin yorumlanması ve hangi içgörülerin vurgulanacağının belirlenmesi amacıyla renk paletleri, etiketler ve tasarım ilkeleri dikkatle uygulanmaktadır.

---

6. Model Oluşturma

Veri bilimi projelerinin temel bileşeni olarak kabul edilen model oluşturma sürecinde, makine öğrenimi ve derin öğrenme modellerinin geliştirilmesi ve

eğitilmesi sağlanmaktadır. Bu süreçte, verilerdeki kalıpların keşfi ve geleceğe yönelik tahminlerin oluşturulması amaçlanmaktadır.

- **Açık Kaynak Model Oluşturma Kütüphaneleri:**
  - **scikit-learn:** Python ortamında, lineer regresyondan karmaşık sınıflandırma yöntemlerine kadar geniş bir yelpaze sunan popüler bir makine öğrenimi kütüphanesi olarak kullanılmaktadır.
  - **TensorFlow** ve **PyTorch:** Derin öğrenme modellerinin geliştirilmesi ve eğitilmesinde, geniş topluluk desteği ve esnek API seçenekleri sunarak tercih edilmektedir.
  - **Spark MLlib:** Apache Spark platformu üzerinde, dağıtık makine öğrenimi algoritmalarının uygulanmasını sağlamaktadır.
- **Ticari Model Oluşturma Araçları:**
  - **IBM SPSS Modeler** ve **SAS Enterprise Miner:** Düşük kodlu veya sürükle-bırak arayüzleri ile istatistiksel modellemenin kolaylaştırılmasını sağlamaktadır.
  - **IBM Watson Machine Learning:** Açık kaynak kütüphanelerin entegrasyonu ile bulut ortamında model oluşturma ve dağıtma imkânı sunmaktadır.

Model oluşturma sürecinde, veri kalitesinin yanı sıra seçilen algoritmanın veriye uygunluğu, özellik mühendisliği ve hiperparametre ayarlamasının model doğruluğunu belirlemede önemli olduğu kabul edilmektedir.

---

7. Model Dağıtımı

Model eğitiminin tamamlanmasının ardından, elde edilen modellerin üretim ortamında kullanılabilmesi için dağıtım süreçleri uygulanmaktadır. Bu süreçte, modellerin gerçek zamanlı veya yığın işleme senaryolarında entegre edilmesi sağlanmaktadır.

- **Açık Kaynak Model Dağıtım Araçları:**
  - **Apache PredictionIO:** Spark ML tabanlı öngörü motorlarının dağıtımının gerçekleştirilmesi amacıyla kullanılmaktadır.

    - **Seldon:** TensorFlow, SparkML, R, scikit-learn gibi çeşitli framework'leri destekleyerek model dağıtımının sağlanmasına olanak tanımaktadır.
    - **TensorFlow Serving:** TensorFlow modellerinin üretim ortamında düşük gecikmeli sunumunun gerçekleştirilmesini sağlamaktadır.
    - **Kubernetes:** Konteyner düzenleme platformu olarak, ölçeklenebilirlik ve güvenlik ihtiyaçlarının karşılanmasında kullanılmaktadır.
- **Ticari Model Dağıtım Çözümleri:**
    - **SPSS Collaboration and Deployment Services:** SPSS araçları ile oluşturulan modellerin kurumsal ortama entegre edilmesini sağlamaktadır.
    - **IBM Watson Machine Learning:** REST API aracılığıyla bulut üzerinde modelin gerçek zamanlı sunumunu gerçekleştirmektedir.
    - **Azure Machine Learning** ve **Amazon SageMaker:** Bulut tabanlı entegre model oluşturma ve dağıtma hizmetleri sunulmaktadır.

Model dağıtım sürecinde, DevOps uygulamaları ve MLOps prensiplerinin benimsenmesiyle sürüm kontrolü ve sürekli entegrasyon süreçlerinin de devreye alınması sağlanmaktadır.

---

8. Model İzleme ve Değerlendirme

Üretim ortamına alınan modellerin performansının zamanla düşme ihtimali dikkate alınarak, model izleme ve değerlendirme süreçleri uygulanmaktadır. Modelin sürekli izlenmesi, gerektiğinde yeniden eğitilmesi veya iyileştirilmesi amaçlanmaktadır.

- **Açık Kaynak İzleme Araçları:**
    - **Prometheus:** Sistem metriklerinin izlenmesi amacıyla kullanılmakta, model performansının takip edilmesi de mümkün kılınmaktadır.
    - **ModelDB:** Makine öğrenimi deneyleri ve modellerle ilgili meta verilerin izlenmesi, saklanması ve sorgulanması sağlanmaktadır.
- **Özel Araçlar:**

    - **IBM AI Fairness 360 (AIF360):** Modellerdeki önyargının tespiti ve azaltılması amacıyla kullanılmaktadır.
    - **AI Explainability 360 (AIX360):** Model kararlarının daha anlaşılır kılınması için çeşitli yöntemler sunmaktadır.
    - **Adversarial Robustness 360 (ART360):** Modellerin adversarial saldırılara karşı dayanıklılığının artırılmasında kullanılmaktadır.
- **Ticari İzleme Araçları:**
    - **IBM Watson OpenScale:** Dağıtılmış modellerin sürekli izlenmesi, adalet ve şeffaflık metriklerinin değerlendirilmesi için entegre bir çözüm sunmaktadır.
    - **Amazon SageMaker Model Monitor:** AWS ekosisteminde model performansının düzenli olarak ölçülmesi ve güncellenmesi sağlanmaktadır.

Veri dağılımı veya yapısındaki değişikliklerin model doğruluğunu etkileyebileceği göz önünde bulundurularak, izleme ve raporlama süreçlerinin sürekli olarak yürütülmesi önem arz etmektedir.

---

9. Kod Varlık Yönetimi (Version Control)

Veri bilimi projelerinde tekrarlanabilirlik ve işbirliği hedeflerine ulaşabilmek amacıyla, kod varlık yönetimi süreçlerinin uygulanması sağlanmaktadır. Bu süreçle, yapılan her değişikliğin kaydı tutulmakta ve ekip içerisindeki ortak çalışma desteklenmektedir.

- **Git:** Sürüm kontrolü için endüstri standardı olarak kullanılmakta, dallanma ve birleştirme işlemleri sayesinde karmaşık projelerde düzen sağlanmaktadır.
- **GitHub, GitLab, Bitbucket:** Git tabanlı bulut platformları olarak, proje yönetimi, konu takibi ve sürekli entegrasyon/dağıtım (CI/CD) araçları ile entegre çalışmaları desteklemektedir.

Bu sayede, her sürümün geri çağrılabilirliği (reproducibility) sağlanmakta ve yapılan değişikliklerin zaman, yer ve sorumluluk açısından izlenebilirliği mümkün kılınmaktadır.

---

10. Veri Varlık Yönetimi

Projedeki verilerin de sürüm kontrol mekanizmasına tabi tutulması, hangi verinin ne zaman değiştiğinin kayıt altına alınması ve erişim kontrollerinin sağlanması amacıyla veri varlık yönetimi süreçleri uygulanmaktadır.

- **Açık Kaynak Araçlar:**
  - **Apache Atlas:** Meta veri yönetimi ve veri soy ağının (data lineage) izlenmesi için kullanılmaktadır.
  - **Egeria:** Farklı platformlarda meta veri paylaşımı ve değişiminin sağlanmasında tercih edilmektedir.
  - **Kylo:** Veri işleme ve varlık yönetimi süreçlerinde açık kaynak desteği sunmaktadır.
- **Ticari Araçlar:**
  - **IBM Information Governance Catalog:** Meta veri ekleme, veri soy ağının izlenmesi ve veri sahiplerinin atanması gibi yönetim görevlerini desteklemektedir.
  - **Informatica Enterprise Data Governance:** Kurumsal veri yönetişimi, politikalar ve veri kalitesi konularında çözümler sunmaktadır.

Bu yaklaşımla, büyük ölçekli kurumlarda veri setleri arasındaki bağlantıların ve regülasyon gereksinimlerinin karşılanması hedeflenmektedir.

---

11. Geliştirme Ortamları (IDEs) ve Çalıştırma Ortamları

Veri bilimcilerin kod geliştirme ve test süreçlerinde kullanılması amacıyla çeşitli geliştirme ortamları (IDE) ile çalıştırma ortamları tercih edilmektedir.

- **Geliştirme Ortamları (IDEs):**
  - **Jupyter Notebook / JupyterLab:** Doküman, kod, çıktı ve görselleştirmenin tek bir sayfada sunulması sağlanmakta; 100’den fazla programlama dilini destekleyen kernel yapısı ile kullanılmaktadır. Bulut tabanlı versiyonları

(örneğin Google Colab, Watson Studio) donanım bağımsız çalışmayı mümkün kılmaktadır.

    - **RStudio:** R dili için kapsamlı bir IDE olarak, veri analizi, istatistiksel modelleme ve görselleştirme işlemlerinde kullanılmaktadır.
    - **Spyder:** Python için hafif bir IDE olup, hızlı prototipleme sürecinde tercih edilmektedir.
    - **Apache Zeppelin:** Jupyter benzeri interaktif not defteri yaklaşımı sunarken, ek olarak kod gerektirmeyen görsel çizim yetenekleri sağlamaktadır.
    - **PyCharm ve Visual Studio Code:** Python ve çoklu dil desteği sunarak, projelerin düzenlenmesi ve yönetilmesinde etkin rol oynamaktadır.
- **Çalıştırma Ortamları (Küme Çalıştırma Ortamları):**
    - **Apache Spark:** Büyük veri analitiği için popüler dağıtık işleme motoru olarak kullanılmaktadır.
    - **Apache Flink:** Gerçek zamanlı veri akış işleme yetenekleriyle tercih edilmektedir.
    - **Ray:** Özellikle büyük ölçekli derin öğrenme projelerinde kullanılmak üzere tasarlanmış yeni nesil sistem olarak değerlendirilmektedir.

---

12. Tam Entegre ve Görsel Araçlar

Farklı seviye veri bilimciler için kullanım kolaylığı sağlanabilmesi amacıyla, sürükle-bırak veya minimal kod yaklaşımlarını benimseyen platformların kullanılması tercih edilmektedir.

- **KNIME:** Zengin sürükle-bırak arayüzü, yerleşik görselleştirme özellikleri ve genişletilebilir mimarisi sayesinde tercih edilmekte; R ve Python kodları ile entegre çalışabilmektedir. Apache Spark ile bağlantı kurulması, büyük veri analitiğine destek sağlamaktadır.
- **Orange:** Hızlı prototipleme ve etkileşimli görselleştirme ihtiyaçlarının karşılanması amacıyla, makine öğrenimi akışlarının sürükle-bırak yöntemiyle oluşturulması sağlanmaktadır.

- **IBM Watson Studio:** Bulut tabanlı, tam entegre bir veri bilimi platformu olarak; Jupyter Notebooks, RStudio, SPSS Modeler gibi bileşenleri tek çatı altında toplanmakta; model dağıtımı, izleme ve yönetişim araçları (örneğin Watson OpenScale) ile entegre çözümler sunulmaktadır.
- **H2O Driverless AI:** Otomatik özellik mühendisliği, hiperparametre optimizasyonu ve modelleme süreçlerinde uzmanlaşmış olup; hem açık kaynak (H2O) hem de ticari sürümü ile kullanılmaktadır.

---

13. Özet

Veri bilimi araç ekosisteminin genişliği ve sürekli gelişim içerisinde olduğu dikkate alınarak, doğru araç seçiminin projelerin başarıya ulaşmasında temel bir rol oynadığı göz önünde bulundurulmaktadır. Aşağıdaki hususlara dikkat edilmesi önerilmektedir:

- **Proje Hedefleri ve Kapsam:**

  Küçük ölçekli prototiplerde Jupyter Notebooks ve açık kaynak veritabanlarının yeterli olabileceği, kurumsal ve düzenlemelere tabi projelerde ise ticari çözümler ile yönetişim araçlarının tercih edilmesinin uygun görüleceği değerlendirilmektedir.

- **Ekibin Deneyimi:**

  Python, R ve SQL gibi programlama dillerine hâkim ekiplerin, açık kaynak kütüphaneler ve IDE’leri kullanması desteklenmektedir.

- **Bütçe ve Maliyet Analizi:**

  Ticari araçların güçlü destek sunması durumunda lisans maliyetlerinin yüksek olabileceği, açık kaynak araçlarda ise destek ve bakım işlemlerinin ekip tarafından üstlenilmesi gerekliliğinin göz önünde bulundurulması önerilmektedir.

- **Ölçeklenebilirlik:**

  Veri hacminin büyük olması durumunda, Apache Spark, Kubernetes veya bulut tabanlı servislerin devreye alınması gerekliliği vurgulanmaktadır.

- **Güvenlik ve Yönetişim:**

Regülasyonlara tabi alanlarda, veri varlık yönetimi ve izlenebilirlik süreçlerinin kritik önem taşıdığı belirtilmektedir.