Module 2 - DS Methodology: From Understanding to Evaluation

Bu notlar, veri bilimi projelerinde karmaşık problemlerin çözümünde rehberlik eden yapılandırılmış metodolojinin temel aşamalarını ele almaktadır. Aşamalar; veri anlama (Data Understanding), veri hazırlama (Data Preparation), modelleme (Modeling) ve değerlendirme (Evaluation) adımlarını içerir. Her aşama, projenin başarısını artırmak amacıyla sorulması gereken temel sorular ve uygulanması gereken işlemleri detaylandırmaktadır. Bu notlar, IBM Data Science Professional Certificate sertifika programının **Data Science Methodology** eğitimi üzerinden, anlatılan bilgiler temel alınarak hazırlanmıştır. Notların geliştirilmesinde internet kaynaklarından ve yapay zeka araçlarından da yararlanılmıştır.

---

1. Data Understanding (Veri Anlama)

Veri anlama aşaması, toplanan verinin, çözülmek istenen problemi temsil edip etmediğini değerlendirme sürecidir.

- **Tanımlayıcı İstatistikler (Descriptive Statistics):**

  Ortalama (mean), medyan (median), minimum, maksimum ve standart sapma (standard deviation) gibi ölçümlerle her değişkenin temel özellikleri ortaya konur.

- **Tek Değişken Analizi (Univariate Analysis):**

  Her bir değişkenin dağılımı, histogramlar (histograms) ve frekans dağılımları kullanılarak incelenir; bu sayede aykırı değerler ve eksiklikler tespit edilir.

- **İkili Korelasyon Analizi (Pairwise Correlation Analysis):**

  Değişkenler arasındaki ilişkinin gücü ölçülerek, yüksek korelasyon (redundancy) durumları belirlenir; böylece gereksiz tekrarlar önlenir.

- **Veri Kalitesi Kontrolleri (Data Quality Checks):**

  Eksik ya da geçersiz değerlerin anlamı değerlendirilir (örneğin, "missing" değerler; "0" veya "bilinmiyor" gibi). Gerekirse, veri toplama (Data Collection) aşamasına geri dönülerek tanımlar güncellenir.

Veri anlama aşaması, sonraki veri hazırlama ve modelleme adımlarının temelini oluşturmakta olup, verinin kalitesinin ve temsil yeteneğinin netleştirilmesinde kritik rol oynar.

---

2. Data Preparation (Veri Hazırlama)

Veri hazırlama aşaması, ham verinin modellemeye uygun hale getirilmesi için temizlenmesi, dönüştürülmesi ve zenginleştirilmesi sürecidir.

- **Veri Temizliği ve Dönüşümü (Data Cleaning and Transformation):**

  Hatalı, eksik veya gereksiz veriler tıpkı taze sebzelerin yıkanması gibi ayıklanır veya düzeltilir.

- **Eksik Değerlerin İşlenmesi (Handling Missing Values):**

  Eksik veriler, uygun şekilde yeniden kodlanır veya gerektiğinde veri setinden çıkarılır.

- **Özellik Mühendisliği (Feature Engineering):**

  Alan bilgisine dayalı olarak, model performansını artıracak yeni değişkenler türetilir. Örneğin, hasta kayıtları gibi çoklu işlemler, tek bir kayıt altında özetlenebilir.

- **Otomasyon (Automation):**

  Tekrarlayan veri hazırlama işlemleri otomatikleştirilerek zaman tasarrufu sağlanır.

- **Case Study Örneği:**

  Sağlık verilerinde, "konjestif kalp yetmezliği" tanımının kapsamının genişletilmesi ve hasta bazında tek kayıt oluşturulması, modelin doğruluğunu artırmak için uygulanmıştır.

Veri hazırlama, verinin doğru ve tutarlı bir biçimde modellenebilmesi için temel adımları içerir. İyi hazırlanmış veri, modelin performansını ve sonuçların güvenilirliğini doğrudan etkiler.

---

3. Modeling (Modelleme)

Modelleme aşaması, hazırlanan veriyi kullanarak problemlere yanıt üretecek modellerin oluşturulması ve eğitilmesi sürecidir.

- **Modelleme Amacı (Purpose of Modeling):**

  Descriptive (tanımlayıcı) modeller, veri içindeki ilişkileri açıklamaya; predictive (öngörücü) modeller ise gelecekteki durumları tahmin etmeye yöneliktir.

- **Eğitim ve Test Setleri (Training and Testing Sets):**

  Veri, modelin eğitilmesi (training) ve doğrulanması (testing) için bölünür; bu, modelin kalibrasyonu ve genel performansının ölçülmesini sağlar.

- **Algoritma Seçimi ve Parametre Ayarlamaları (Algorithm Selection and Parameter Tuning):**

  Farklı algoritmalar denenir; örneğin, karar ağaçlarında misclassification cost (yanlış sınıflandırma maliyeti) ayarlanarak modelin duyarlılığı (sensitivity) ve özgüllüğü (specificity) dengelenir.

- **İteratif Geliştirme (Iterative Refinement):**

  Model oluşturma süreci, gerektiğinde veri hazırlama aşamasına geri dönülerek ve parametreler ayarlanarak sürekli iyileştirilir.

Modelleme, veri bilimi sürecinin kalbini oluşturur. Doğru modelin seçilmesi ve uygun parametre ayarlamalarının yapılması, elde edilecek sonuçların doğruluğunu ve uygulama başarısını belirler.

---

4. Evaluation (Değerlendirme)

Model değerlendirme aşaması, oluşturulan modelin başlangıçtaki iş problemini (business problem) ne kadar doğru yanıtladığını test eder ve optimize edilmesine olanak tanır.

- **Tanısal Ölçümler (Diagnostic Metrics):**

  Modelin true positive rate (gerçek pozitif oranı), true negative rate (gerçek negatif oranı), overall accuracy (genel doğruluk) gibi ölçümlerle performansı değerlendirilir. ROC (Receiver Operating Characteristic) eğrisi gibi araçlar kullanılarak modelin performansı görselleştirilir.

- **İstatistiksel Testler (Statistical Tests):**

  Modelin sonuçlarının güvenilirliği, istatistiksel olarak test edilir; bu, modelin rastlantısallık payını azaltır.

- **Geri Bildirim ve İterasyon (Feedback and Iteration):**

  Değerlendirme sonuçları, modelin yeniden ayarlanması için geri bildirim sağlar. Gerekirse, modelin parametreleri değiştirilir veya veri hazırlama aşamasına geri dönülerek ek düzenlemeler yapılır.

- **Case Study Örneği:**

  Konjestif kalp yetmezliği verileri üzerinde, farklı misclassification cost ayarları denenmiş; ROC eğrisi analiz edilerek optimum model seçilmiştir.

Model değerlendirme, modelin iş hedeflerine uygunluğunu ve genel başarısını belirlemede hayati öneme sahiptir. Doğru değerlendirme, modelin üretim ortamına alınması ve uzun vadeli performansının izlenmesi için temel bir adımdır.

---

5. Özet

Bu notlarda, veri bilimi metodolojisinin temel aşamaları olan **Data Understanding (Veri Anlama)**, **Data Preparation (Veri Hazırlama)**, **Modeling (Modelleme)** ve **Evaluation (Değerlendirme)** detaylandırılmıştır.

- **Data Understanding (Veri Anlama):** Toplanan verinin, çözülmek istenen problemi temsil edip etmediği; tanımlayıcı istatistikler, univariate analizler ve korelasyon incelemeleriyle değerlendirilir.
- **Data Preparation (Veri Hazırlama):** Ham verinin temizlenmesi, dönüştürülmesi ve özellik mühendisliği yoluyla modellemeye hazır hale getirilmesi sağlanır.
- **Modeling (Modelleme):** Hazırlanan veri kullanılarak, tanımlayıcı veya öngörücü modeller oluşturulur; eğitim ve test setleriyle model kalibrasyonu yapılır.
- **Evaluation (Değerlendirme):** Modelin, iş problemini doğru şekilde yanıtlayıp yanıtlamadığı tanısal ölçümler ve istatistiksel testlerle değerlendirilir; geri bildirim mekanizmaları ile sürekli iyileştirme sağlanır.

Doğru metodolojinin uygulanması, veri bilimi projelerinin başarıya ulaşmasında, modelin güvenilirliğinde ve elde edilen sonuçların doğruluğunda belirleyici bir rol oynamaktadır.