Module 4 - Introduction to CRISP-DM

Bu notlar, veri madenciliği projelerinde kullanılan CRISP-DM (Cross-Industry Standard Process for Data Mining) metodolojisinin temel aşamalarını hem İngilizce hem de Türkçe terimlerle özetlemektedir. CRISP-DM, veri odaklı kararlar almak için yapılandırılmış, iteratif ve esnek bir süreç sunar. 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. CRISP-DM Nedir? (What is CRISP-DM?)

- **Tanım:**CRISP-DM, Cross-Industry Standard Process for Data Mining, endüstri tarafından kanıtlanmış kapsamlı bir veri madenciliği metodolojisidir.
- **Özellik:**İteratif (iterative) yapıda olup, her aşamada geri bildirim (feedback) alınarak sürecin sürekli iyileştirilmesine olanak tanır.

---

2. CRISP-DM Aşamaları (Stages of CRISP-DM)

CRISP-DM modeli, temel olarak altı aşamadan oluşmaktadır:

1. **Business Understanding (İş Anlayışı):**

   Proje hedeflerinin, iş probleminin ve paydaş beklentilerinin netleştirilmesi. Bu aşama, sürecin temelini oluşturur.

2. **Data Understanding (Veri Anlama):**

Veri kaynaklarının belirlenmesi, verinin toplanması ve keşfedilmesi. Tanımlayıcı istatistikler, univariate analizler ve korelasyon incelemeleri yapılır.

3. **Data Preparation (Veri Hazırlama):**

   Ham verinin temizlenmesi, dönüştürülmesi ve modellemeye uygun hale getirilmesi. Eksik değerlerin giderilmesi ve özellik mühendisliği (feature engineering) bu aşamada gerçekleştirilir.

4. **Modeling (Modelleme):**

   Uygun algoritmalar seçilerek, veriye dayalı modellerin oluşturulması ve eğitilmesi. Hem tanımlayıcı hem de öngörücü modeller bu aşamada geliştirilir.

5. **Evaluation (Değerlendirme):**

   Modelin, iş hedeflerine uygunluğunun ve performansının tanısal ölçümler (diagnostic metrics) ve istatistiksel testlerle değerlendirilmesi.

6. **Deployment (Dağıtım):**

   Modelin, üretim ortamına entegre edilip, gerçek veriler üzerinde uygulanması ve sonuçların paydaşlara sunulması. Dağıtım sonrası, geri bildirim (feedback) ile model sürekli iyileştirilir.

CRISP-DM, John Rollins'in metodolojisinde ayrı aşama olarak ele alınan veri gereksinimleri, veri toplama ve veri anlama süreçlerini tek bir aşamada birleştirmektedir.

---

3. Business Understanding (İş Anlayışı)

- **Amaç:**İş probleminin, hedeflerin ve paydaş beklentilerinin netleştirilmesi sağlanır.
- **Önem:**Bu aşama, tüm metodolojinin yönünü belirler; yanlış anlaşılan bir iş problemi, sonraki aşamalarda istenmeyen sonuçlara yol açabilir.

---

4. Data Understanding & Data Preparation (Veri Anlama & Veri Hazırlama)

- **Data Understanding (Veri Anlama):**
  - Veri kaynaklarının incelenmesi, tanımlayıcı istatistiklerin çıkarılması ve değişkenlerin dağılımının analiz edilmesi.
  - Verinin, çözülmek istenen problemi temsil edip etmediğinin kontrolü yapılır.
- **Data Preparation (Veri Hazırlama):**
  - Verinin temizlenmesi (data cleaning), eksik ve geçersiz değerlerin düzeltilmesi ve dönüştürülmesi (transformation) sağlanır.
  - Özellik mühendisliği (feature engineering) ile veriden modelin performansını artıracak yeni değişkenler türetilir.

---

5. Modeling (Modelleme)

- **Amaç:**Seçilen analitik yaklaşımla, veriye dayalı modeller oluşturulur.
- **Süreç:**
  - Verinin eğitim (training) ve test (testing) setlerine bölünmesi.
  - Algoritmaların uygulanması ve parametre ayarlamaları (parameter tuning) ile modelin optimize edilmesi.
  - İteratif geliştirme (iterative refinement) ile model sürekli iyileştirilir.

---

6. Evaluation (Değerlendirme)

- **Amaç:**Modelin, iş problemini doğru yanıtlayıp yanıtlamadığını ve performansının ölçülmesi.
- **Süreç:**

- Tanısal ölçümler (diagnostic metrics) kullanılarak modelin duyarlılığı (sensitivity) ve özgüllüğü (specificity) hesaplanır.
- ROC (Receiver Operating Characteristic) eğrisi gibi araçlarla modelin performansı görselleştirilir.
- İstatistiksel testler (statistical tests) ile modelin güvenilirliği kontrol edilir.

- **Geri Bildirim:**Değerlendirme sonuçlarına göre model, dağıtım sonrası geri bildirimlerle (feedback) iyileştirilir.

---

7. Deployment (Dağıtım)

- **Amaç:**Modelin, üretim ortamına entegre edilip, gerçek zamanlı veriler üzerinde kullanılabilir hale getirilmesi.
- **Süreç:**
  - Model sonuçlarının, paydaşlar ve son kullanıcılar tarafından anlaşılmasını sağlamak için eğitim ve bilgilendirme yapılır.
  - Uygulama, genellikle sınırlı bir kullanıcı grubu ya da test ortamında başlatılır; sonrasında yaygınlaştırılır.
- **İş Birliği:**Çözüm sahibi, IT, pazarlama ve uygulama geliştiriciler gibi farklı uzmanlık alanlarından kişilerle iş birliği içinde çalışılır.

---

8. İteratif Süreç ve Feedback (Geri Bildirim)

- **İterasyon:**CRISP-DM modeli, her aşamanın tamamlanmasının ardından, elde edilen bulgulara göre önceki aşamalara dönülebilecek döngüsel bir yapıya sahiptir.
- **Feedback (Geri Bildirim):**
  - Kullanıcı ve sistem geri bildirimleri alınarak modelin performansı sürekli iyileştirilir.

- John Rollins metodolojisinde açıkça "Feedback" olarak tanımlanan bu aşama, CRISP-DM'de dağıtım (Deployment) aşmasının ardından gerçekleşir.

---

9. Özet

CRISP-DM (Cross-Industry Standard Process for Data Mining), veri madenciliği projelerinde yapılandırılmış, iteratif ve esnek bir metodolojidir.

- **Business Understanding (İş Anlayışı):** Projenin temel hedeflerinin ve iş probleminin belirlenmesi.
- **Data Understanding (Veri Anlama):** Veri kaynaklarının keşfi, tanımlayıcı analizler ve veri kalitesi kontrolleri.
- **Data Preparation (Veri Hazırlama):** Ham verinin temizlenip, dönüştürülmesi ve özellik mühendisliği ile modellemeye hazır hale getirilmesi.
- **Modeling (Modelleme):** Uygun algoritmalarla model oluşturulması ve eğitilmesi, parametre ayarlamalarıyla optimize edilmesi.
- **Evaluation (Değerlendirme):** Modelin performansının tanısal ölçümler ve istatistiksel testlerle değerlendirilmesi.
- **Deployment (Dağıtım):** Modelin gerçek dünya verileri üzerinde uygulanması ve paydaşlara sunulması.
- **Iterative Process & Feedback (İteratif Süreç ve Geri Bildirim):** Sürecin döngüsel yapısı sayesinde, model ve süreçler sürekli olarak iyileştirilir.

CRISP-DM modeli, iş problemlerinin doğru anlaşılmasından başlayarak, verinin toplanması, işlenmesi, model oluşturulması ve sonuçların uygulanıp geri bildirim alınmasına kadar, tüm adımları kapsayan kapsamlı bir metodoloji sunar. Bu yapı, veri bilimi projelerinin hem bilimsel hem de pratik anlamda başarılı sonuçlar elde etmesine olanak tanır.