

Module 1 - Data Science Methodology

Bu notlar, veri bilimi projelerinde karmaşık problemlerin çözümünde ve veri odaklı kararlar alınmasında rehberlik eden sistematik yöntemler hakkında bir özet sunmaktadır. 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. Metodoloji Tanımı ve Önemi

- **Metodoloji Nedir?**

Belirli bir alanda kullanılan yöntemler sistematiktir. Veri bilimi bağlamında, metodoloji, veri toplama, analiz, modelleme ve sonuçların değerlendirilmesi süreçlerinde atılması gereken adımları belirleyen rehber niteliğindedir.

- **Veri Bilimi Metodolojisinin Rolü:**

- Problemin doğru tanımlanmasını sağlar.
- Hangi verinin nasıl kullanılacağını, hangi analiz yöntemlerinin uygulanacağını ve sonuçların nasıl yorumlanacağını belirler.
- Bilimsel araştırmada tutarlılık ve yeniden üretilebilirlik için yol gösterici rol oynar.

Metodolojiye uymak, doğrudan çözüm üretmeye atlamanın getireceği belirsizlik ve hataların önüne geçilmesinde kritik öneme sahiptir.

2. Standart Veri Bilimi Metodolojisinin 10 Aşaması

Veri bilimi projelerinde uygulanan metodoloji, aşağıdaki 10 aşama üzerinden sistematik bir yaklaşım sunmaktadır.

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

Problemin iş açısından doğru tanımlanması, hedeflerin ve iş gereksinimlerinin belirlenmesi.

2. **Analytic Approach (Analitik Yaklaşım):**

Problemi çözmek için hangi analiz yöntemlerinin kullanılacağına dair stratejinin oluşturulması.

3. **Data Requirements (Veri Gereksinimleri):**

Problemi çözmek için hangi veriye ihtiyaç duyulduğunun, verinin içeriği, formatı ve kaynaklarının belirlenmesi.

4. **Data Collection (Veri Toplama):**

Belirlenen gereksinimlere uygun verinin toplanması.

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

Toplanan verinin incelenmesi, kalitesinin ve özelliklerinin değerlendirilmesi.

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

Verinin temizlenmesi, dönüştürülmesi ve modellemeye uygun hale getirilmesi.

7. **Modeling (Modelleme):**

Seçilen analitik yaklaşım doğrultusunda modellerin oluşturulması ve eğitilmesi.

8. **Evaluation (Değerlendirme):**

Oluşturulan modellerin iş hedeflerine ve performans kriterlerine göre değerlendirilmesi.

9. **Deployment (Dağıtım):**

Modelin üretim ortamına entegre edilerek pratikte uygulanmaya başlanması.

10. Feedback (Geri Bildirim):

Modelin performansının izlenmesi, sonuçların gözden geçirilmesi ve gerekirse iyileştirmeler yapılması.

3. Business Understanding (İş Problemi Tanımlama)

- **Problemin Tanımlanması:**

İş hedefleri ve sorunlar netleştirilerek, "Asıl sorun nedir?" sorusunun doğru cevaplanması sağlanır.

İşveren veya proje sponsorunun hedefi ile, projenin amaçları uyumlu hale getirilir.

- **Hedef ve Amaçların Belirlenmesi:**

Sorunun arkasındaki temel iş hedefleri (örneğin, maliyet azaltma, verimlilik artırma) netleştirilir.

Bu süreç, hangi verilerin kullanılacağı ve analitik yaklaşımın nasıl şekilleneceği konusunda yol gösterici olur.

İş probleminin doğru anlaşılması, metodolojinin diğer aşamalarının da etkili bir şekilde yürütülmesini sağlar.

4. Analytic Approach (Analitik Yaklaşım)

Sorunların çözümüne yönelik analitik yaklaşım, sorulacak temel soruların türüne göre şekillenir. Beş temel soru türü ve karşılık gelen yaklaşımlar şu şekildedir:

- **Descriptive (Betimleyici):**

"Mevcut durum nedir?" sorusuna yanıt arar.

Teknikler: Veri toplama, özetleme, görselleştirme.

- **Diagnostic (Tanısal):**

"Neden böyle oldu?" sorusunu cevaplar.

Teknikler: Drill-down, korelasyon analizi, veri keşfi.

- **Predictive (Öngörücü):**

"Gelecekte ne olacak?" sorusuna yönelik tahminler yapar.

Teknikler: Regresyon analizi, zaman serisi tahmini, makine öğrenimi modelleri.

- **Prescriptive (Önerisel):**

"Ne yapmalıyız?" sorusunu yanıtlar.

Teknikler: Optimizasyon modelleri, simülasyon, karar analizi.

- **Classification (Sınıflandırıcı):**

"Bu veri hangi kategoriye ait?" sorusuna cevap verir.

Teknikler: Karar ağaçları, lojistik regresyon, destek vektör makineleri.

Analitik yaklaşımın belirlenmesi, çözüm sürecinde kullanılacak yöntemlerin ve tekniklerin temelini oluşturur.

5. Data Requirements (Veri Gereksinimleri)

- **Veri Gereksinimlerinin Belirlenmesi:**

Problemi çözmek için hangi verinin gerekli olduğu, verinin içeriği, formatı ve kaynakları netleştirilir.

Bu aşama, veri toplama ve hazırlama süreçlerine geçmeden önce, hangi verinin kullanılacağını planlanmasını sağlar.

- **Örnek Uygulama:**

Örneğin, sağlık sektöründe bir karar ağacı modeli kullanarak hasta yeniden yatış riskini analiz etmek için; hasta seçim kriterleri, klinik geçmiş, tanı ve tedavi bilgileri gibi veriler belirlenir.

Bu veriler, hasta bazında tek bir kayıt oluşturulacak şekilde özetlenir.

Veri gereksinimlerinin doğru belirlenmesi, sonraki veri toplama, hazırlama ve modelleme aşamalarının başarısı için kritik öneme sahiptir.

6. Özet

Bu notlarda, veri bilimi projelerinde metodolojinin tanımı, 10 aşamalı sürecin detaylandırılması, iş probleminin doğru tanımlanması, analitik yaklaşımın belirlenmesi ve veri gereksinimlerinin netleştirilmesi ele alınmaktadır.

- **Metodoloji**, veri bilimi çalışmalarında yapılandırılmış ve tutarlı bir yol haritası sunarak; problemi doğru tanımlama, uygun veri toplama ve analitik yaklaşımlar geliştirme sürecini destekler.
- **10 Aşama**, iş hedefinden geri bildirim aşamasına kadar sistematik bir yaklaşım sunmakta; her aşamada sorulması gereken temel sorular, çözüm sürecinin doğruluğunu ve verimliliğini artırır.

Her proje için doğru metodolojinin uygulanması, veri bilimi çalışmalarının başarıya ulaşmasında ve elde edilen sonuçların güvenilirliğinde belirleyici bir rol oynamaktadır.