**1.4) CRISP-DM Metodolojisi**

Günümüz penceresinden teknoloji devrimine baktığımız zaman dikkatimizi ilk çeken ayrıntı “Veri” olacaktır.

Çığ edasıyla katlanarak büyüyen “Veri” şimdiden insanların işlem ve yetenek gücünü yetersiz duruma getirmiştir.

Yetersizlikten yola çıkan insanlar ise yeni teknolojileri, devrimleri ve yöntemleri beraberinde getirmiştir.

Projeler ve işler büyüdükçe insanlar farklı yollara başvurmuştur.

CRISP-DM de bahsetmeye çalıştığımız yollardan birisidir.

**CRISP-DM Nedir?**

CRISP-DM, iş problemlerinin ve uygulamalarının veri tabanlı çözümlerle verimliliğini arttırmak amacıyla kullanılan standartlaştırılmış bir Veri Bilimi metodolojisidir.

Özetle diyebiliriz ki Veri Bilimi Projelerinin yürütülmesinde izlenen yol haritasıdır.

Tarihsel süreç boyunca SEMMA (SAS Enstitüsü tarafından geliştirilen sıralı adımların bir listesi) ve KDD yöntemlerinin geliştirilmesiyle ortaya çıkan “Veri Bilimi Proje Yöntem Süreci” olarak da bilinmektedir.

Hatta IBM tarafından 2015 yılında CRISP-DM’ in de geliştirilmesiyle ASUM-DM yöntemi ortaya çıkmıştır.

ASUM-DM ise büyük ölçekli, kurumsal projelerde kullanılmak ve proje yönetimi süreçlerine dahil edilmek için ortaya çıkmıştır.

CRISP-DM yüksek seviyeli bir metodoloji olmasından dolayı modelde belirtilen aşamalar, iş ihtiyaçlarını karşılamak için birçok farklı biçimde, sırada ve teknolojide uygulanabilmektedir.

Toplamda 6 aşamadan oluşan modelin şeması aşağıdaki görseldeki gibidir:

metin, daire, diyagram, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

*NOT: Aşamaların sırası kesin değildir ve genellikle farklı aşamalar arasında gidip gelmek gerekebilmektedir.*

**CRISP-DM Adımları**

**1.) İş’i Anlamak (Business Understanding)**

Proje başlangıcında işi anlamak yani mevcut problemi tanımlamak projenin diğer adımları için oldukça önemlidir.

En doğru problem tanımını yapmak proje hedeflerini ve gereksinimlerini belirlemek ve izlenecek yol haritası için ön plan oluşturmaktadır.

Bu aşama veri kaynaklarının ve ihtiyaç duyulacak alanların tanımlanmasını içerse de verilerin herhangi bir analizini içermez.

*Örneğin; Şirketimiz ürün reklamı yapmak istiyor ve hangi müşteriye hangi reklamın gösterileceğini belirlemek istiyor. (Hedef Ürün Pazarlama)*

**2) Veriyi Anlamak (Data Understanding)**

Problem tanımına uygun olarak veri toplanır veya mevcut verinin üzerinden geçilir.

Genelde bu adıma başlanmadan önce problem tanımının doğru yapılması oldukça önemlidir çünkü sık yapılan hatalardan birisi eldeki verinin tamamının gereksiz yere işlenmesidir.

Yine bu aşamada veri ile ilgili problemler de ortaya çıkartılır.

metin, yazı tipi, ekran görüntüsü, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Örneğin; Verinin gürültülü veya kirli olması, eksik veri içermesi gibi problemlerin tespit edilmesi, verinin yapısal / yapısal olmayan veya yarı yapısal verinin tipinin tespit edilmesi adımları bu aşamada gerçekleştirilir.

Gerekli görülürse ilave veri toplanır veya eldeki veriler üzerinden nasıl işlemler yapılarak verinin zenginleştirilebileceğine bu aşamada karar verilir.

**3) Veriyi Hazırlamak (Data Preparation)**

Veriyi hazırlama aşamada veri üzerinde yapılacak işlemlere ve bu işlemlerin hangi yöntemlerle yapılacağına karar verilir.

Bir önceki adımda, veri üzerinde eksik veri tespiti yapıldı, bu tespite göre bir kısmı eksik olan verinin sisteme hiç dahil edilmemesi veya verinin eksik kısımlarının tamamlanması, bu tamamlama sırasında nasıl bir yöntem izleneceği

(bkz. Töhmet (imputation)) gibi kararlar bu aşamada verilir, uygulanır ve çıktılar değerlendirilerek Modelleme daha başarılı hale getirilmeye çalışılır.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

*Örneğin; Doğum tarihlerinin yaşa verilmesi basit bir veri dönüşümü (transformation) veya adres alanında kişinin il — ilçe gibi bilgilerinin çıkarılması veri zenginleştirme (enrichment) işlemleridir.*

*Bu aşamada yapılan çalışmalar bir önceki aşamayla birlikte ele alındığında, literatürde öznitelik mühendisliği (feature engineering) adı verilen bir başlık altında incelenebilir.*

**Feature engineering** (özellik mühendisliği), makine öğrenimi ve veri biliminde, ham veriden model performansını artıracak anlamlı ve tahmin gücü yüksek özellikler (features) oluşturma sürecidir.

İyi bir feature engineering süreci, veri analizinin doğruluğunu artırarak daha güçlü modellerin elde edilmesine yardımcı olur.

* Özellik Seçimi
* Özellik Dönüştürme
* Yeni Özellik Oluşturma

**4) Modelleme (Modelling)**

Modelleme aşamada tanımlanan problem ve veri kaynakları üzerinde bir makine öğrenmesi veya istatistiksel model (Karar Ağaçları, Sinir Ağları vb.) geliştirilir.

Geliştirilen model istenen problem çözümüne yönelik olarak iyileştirilir (optimization)

Bu aşamada genelde modele uygun olarak verinin düzenlenmesi gerekebilir.

yazı tipi, tipografi, hat sanatı, kaligrafi, el yazısı içeren bir resim

Açıklama otomatik olarak oluşturuldu

*Örneğin; bazı modellerin çalışması için dengeli veri veya bazı modeller için aykırı verinin ayıklanması ve hatta başka bazı modeller için de veri dönüşümleri gerekebilir.*

metin, yazı tipi, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

**5) Değerlendirme (Evaluation)**

Değerlendirme aşamasından şimdiye kadar olan adımların genel bir değerlendirmesi yapılır ve aslında ilk adımda (Problem Tanımı Adımı) konulan başarı kriterlerinin ne ölçüde sağlandığı test edilir.

Bunun için deney grupları veya farklı test tekinleri kullanılabilir.

*Örneğin müşterilerine doğru reklamı göstermek için veri bilimi projesi geliştiren bir e-ticaret sitesinde, veri bilimi kullanılan ve kullanılmayan iki grup oluşturarak üzerlerinde A/B testi uygulanabilir.*

*Bazı durumlarda ise saha üzerinde test imkânı olmadığı için geliştirilen yöntemin başarısı bazı metrikler üzerinden ölçülebilir.*

metin, ekran görüntüsü, araba, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Gerçek Fiyat ve Tahmin Edilen Fiyat Arasındaki Fark (RMSE)

Problem tanımlarına göre örneğin kesinlik (accuracy), saflık (purity) gibi değerlere bakılabilir.

Verilen karara göre ürünleştirmeye geçilebilir veya bütün aşamaların tekrar gözden geçirilmesi için ilk aşamaya geri dönülebilir.

**5) Uygulama**

Ürün aşaması: Bu aşamaya kadar, elde edilen çıktılar kurumun çalışma ortamına uygun halde geliştirme sürecine başlanır.

Büyük veri platformuna uygun halde veya kurumda kullanılan programlama dilleri ve veri akışına uygun hale getirilerek sistemin çalışan bir uygulaması geliştirilir.

*Örneğin; bir e-ticaret sitesindeki reklamların veri madenciliği süreçleri ile yönetileceği bir sitenin kodlanmasına başlanır, reklam gösterimlerine karar verilen adımlarda ilgili değişiklikler yazılarak yeni geliştirilen karar sistemi devreye sokulur ve karar sisteminin verdiği sinyallere göre reklamlar gösterilmeye başlanır.*