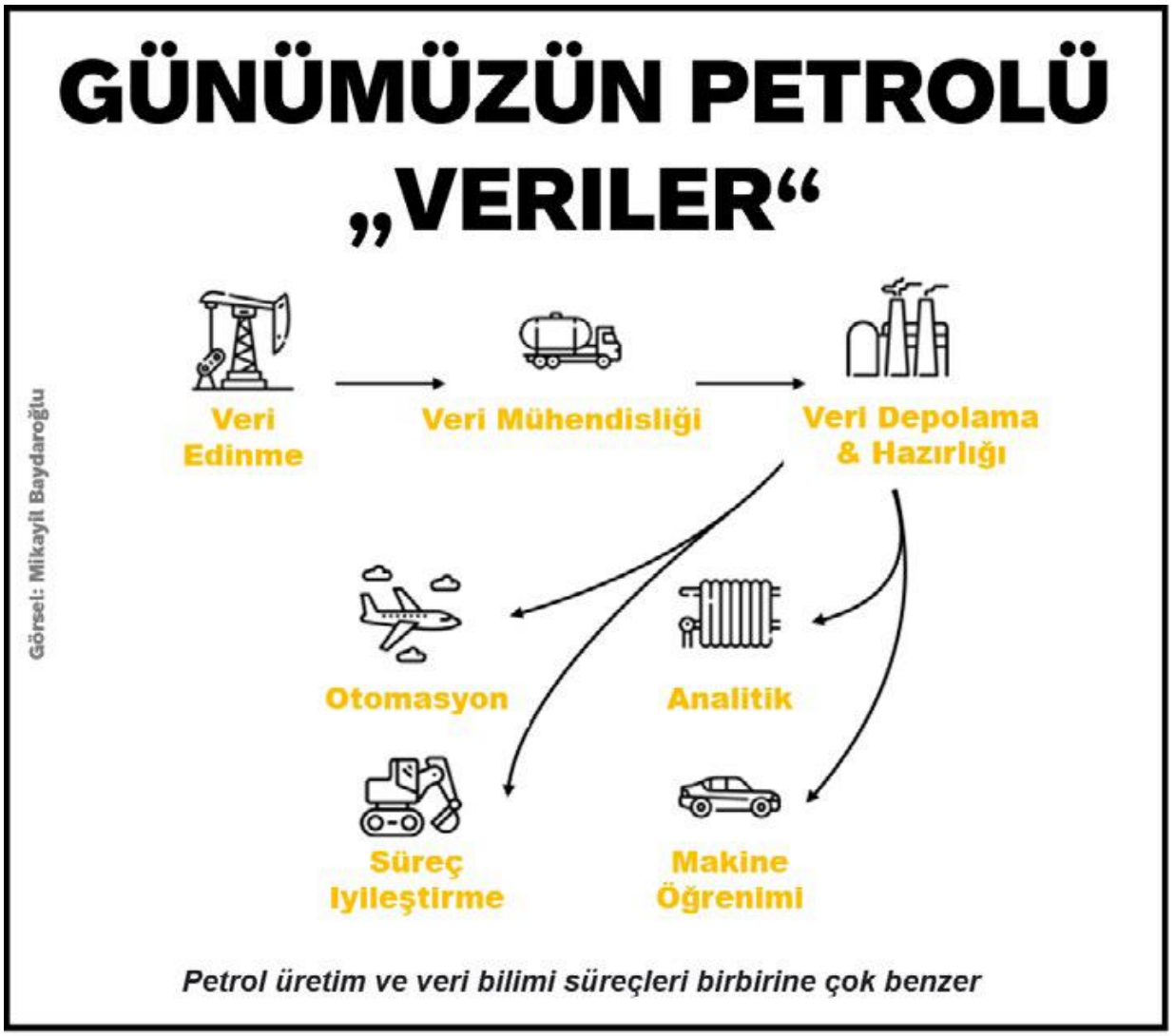
UYGULAMALI VERİ ANALİZİ

HAFTA-1
21.02.2023

Dr.Öğr.Üyesi Ayşe Merve ACILAR

Neden bu dersi açtık?



Bir Veri Analiz Uygulama Örneği

- **86.200** kişiyle yapılan araştırmada Facebook beğenileri üzerine veri analizleri oluşturulmuş.
- Sadece 10 beğeniyi inceleyerek kullanıcıyı iş arkadaşından daha iyi analiz eden sistem,
- 70 beğeniyle kullanıcıyı yakın arkadaşlarından daha iyi tanımlayabiliyor.
- 150 beğeniyle ailesinden daha iyi
- 300 beğeni ve üzeri ile eşinden daha iyi tanımlayabildiği gözlemlenmiş.

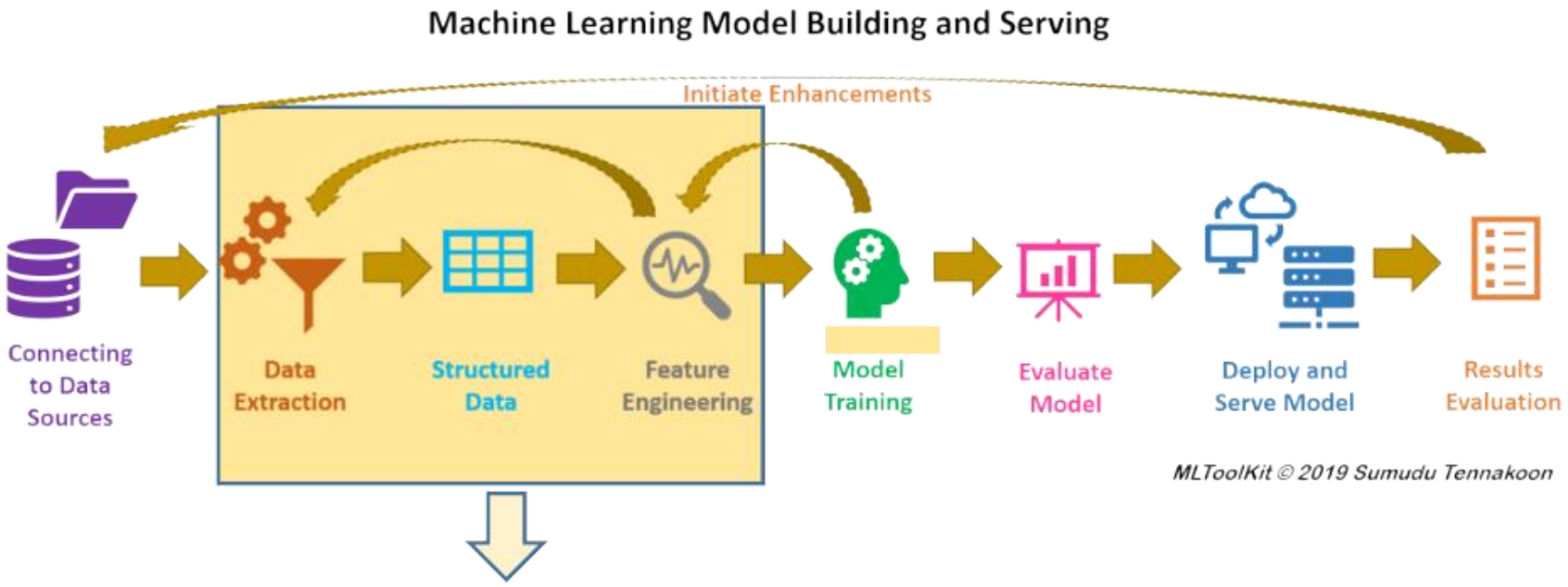
Dersin Amacı

- Öğrencinin veri analizine ait kavramları öğrenmesi
- Bir veri dosyası ile karşılaştığında onu anlamlı hale getirebilecek süreçleri kavraması
- Gerekli istatistiki çıkarımları yapabilmesi
- Veriye uygun modeli oluşturabilmesi
- Edindiği bilgileri gerçek hayat problemlerine uygulayabilmesi amaçlanmaktadır.

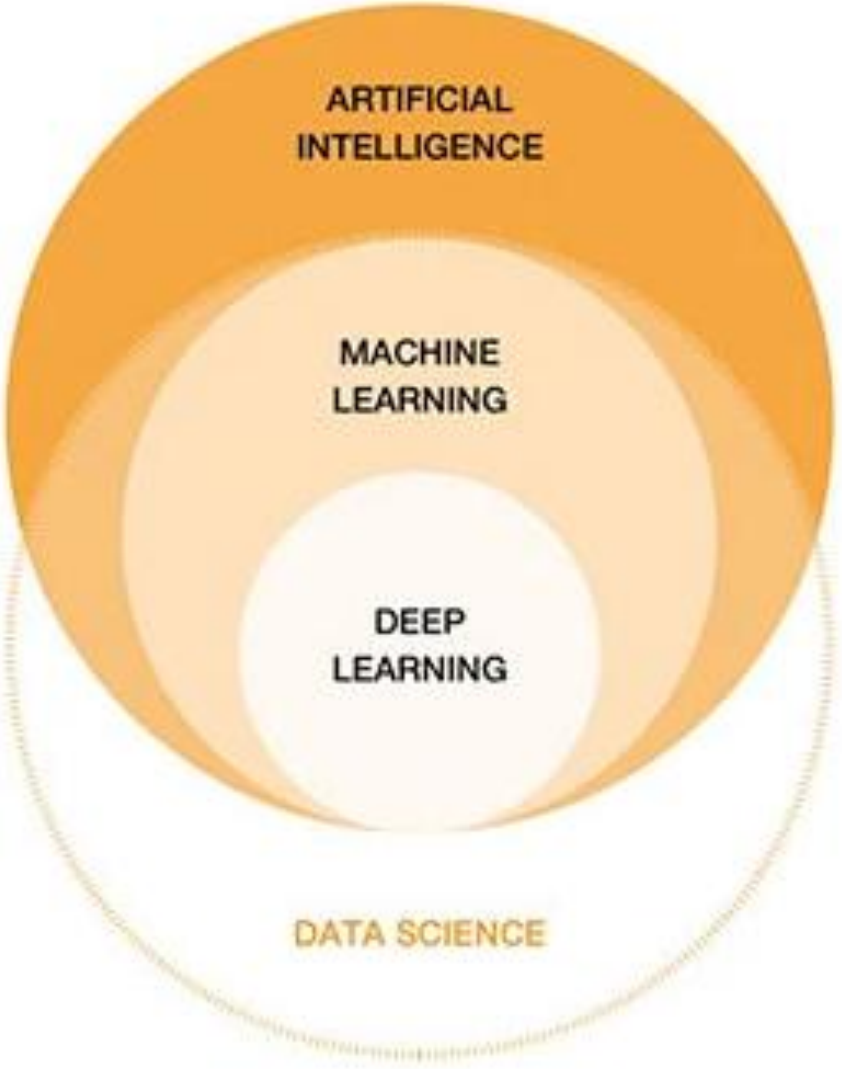
Dersin Konuları

- Veri Analizine Giriş ve Veri Analizi Proje Yönetim Adımları
- Değişken Tipleri ve Özellikleri
- Betimleyici İstatistik
- Çıkarımsal İstatistik
- Olasılık Dağılımları
- Eksik veri analizi
- Ayrık veri analizi
- Kategorik değişken enkodları
- Ayrıklaştırma, Normalizasyon ve Standartizasyon
- Veri görselleştirme
- Özellik Seçimi ve Çıkarımı
- Uygulama Örnekleri

Büyük Resim



Bu derste ilgileneceğimiz kısım



Kaynak:https://www.oracle.com/tr/artificial-intelligence/what-is-artificial-intelligence.html

Veri Nedir?

- *Bir veri, bir hareketin veya etkinliğin kaydıdır; çoğunlukla bir kişi tarafından verilen bir kararın yansımasıdır. Eğer bu karara yönlendiren etkinliklerin sırasını oluşturursan, bundan ders alabilirsin; bu müşterilerin neleri sevip nelerden hoşlanmadıklarını söylemlerinin dolaylı yoludur*"
- "*Veri bilimi bir yorumlama işidir – müşterinin sesini karar vermeye daha uygun bir duruma tercüme ederiz.*"

Airbnb'nin **veri bilimci**lerinden Riley Newman

Kaynak: https://medium.com/airbnb-engineering/at-airbnb-data-science-belongs-everywhere-917250c6beba

VERİ TÜRLERİ
SINIFLANDIRMA

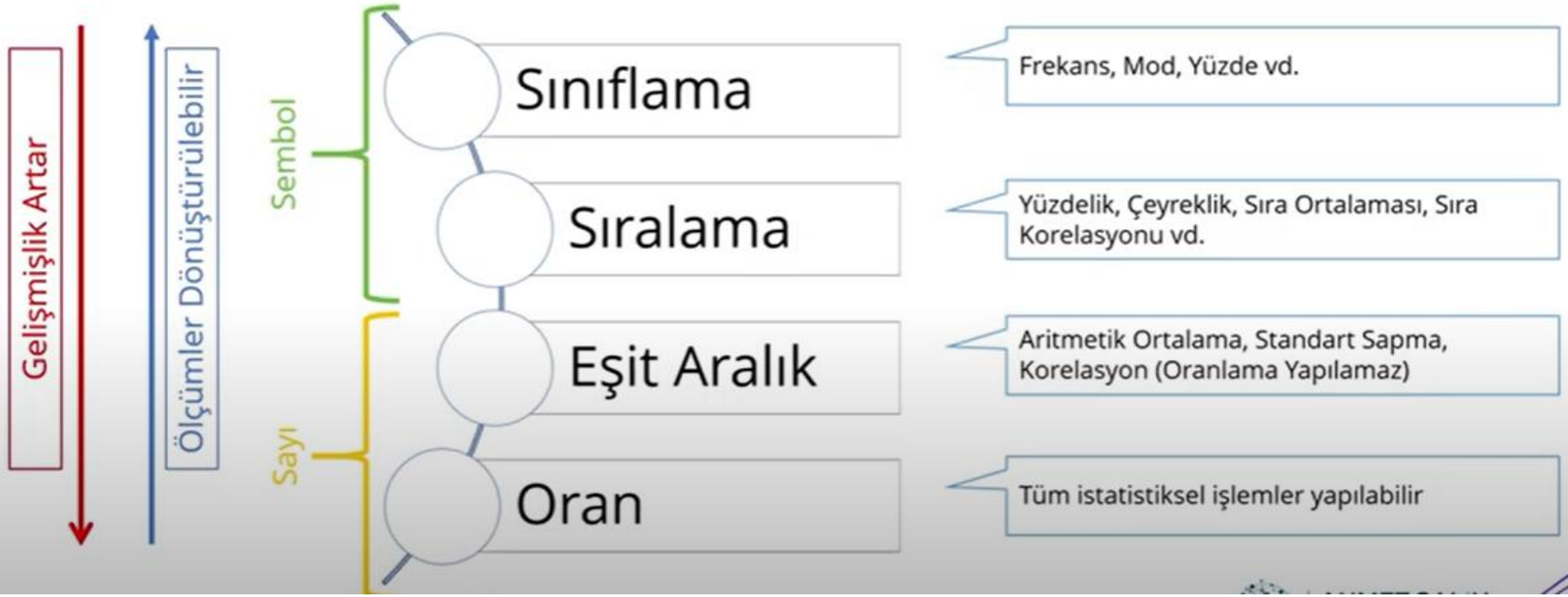
VERİ TÜRLERİ

VERİ	DEĞER	ELDE EDİLMESİNE GÖRE	TÜRÜNE GÖRE	YAPISINA GÖRE
CİNSİYET	E (0); K (1)	OLGUSAL	NİTEL	SINIFLAMA
GELİR DÜZEYİ	DÜŞÜK; ORTA; YÜKTEK	YARGISAL	NİTEL	SINIFLAMA
BAŞARI SIRASI	1; 2; 3; ...	OLGUSAL	NİTEL	SIRALAMA
BAŞARI ALGISI	1; 2; 3; 4; 5	YARGISAL	NİCEL	SAYISAL (EŞİT ARALIK)
SINAV PUANI (0-100)	...	OLGUSAL	NİCEL	SAYISAL (EŞİT ARALIK)
AYLIK GELİR (TL)	...	OLGUSAL	NİCEL	SAYISAL (ORAN)
YAŞ (YIL)	...	OLGUSAL	NİCEL	SAYISAL (ORAN)

VERİ TÜRLERİ
UYARILAR ÖNERİLER

- Yargısal veri yerine olgusal veri elde etmek verinin nesnelliği artırır dolayısıyla veriden elde edilecek bulguların geçerliği ve güvenirliği artar
- Sayısal olarak elde edilebilecek verinin kategorik olarak toplanması bilgi kaybına neden olur (Gelir, Yaş, Kıdem vb.)
- Sayısal olarak elde edilen veriler daha sonra kategorik hale getirilebilir ancak tersi mümkün değildir

VERİ TÜRLERİ
ÖLÇEK DÜZEYİ



VERİ TÜRLERİ
BAĞIMSIZ-BAĞIMLI

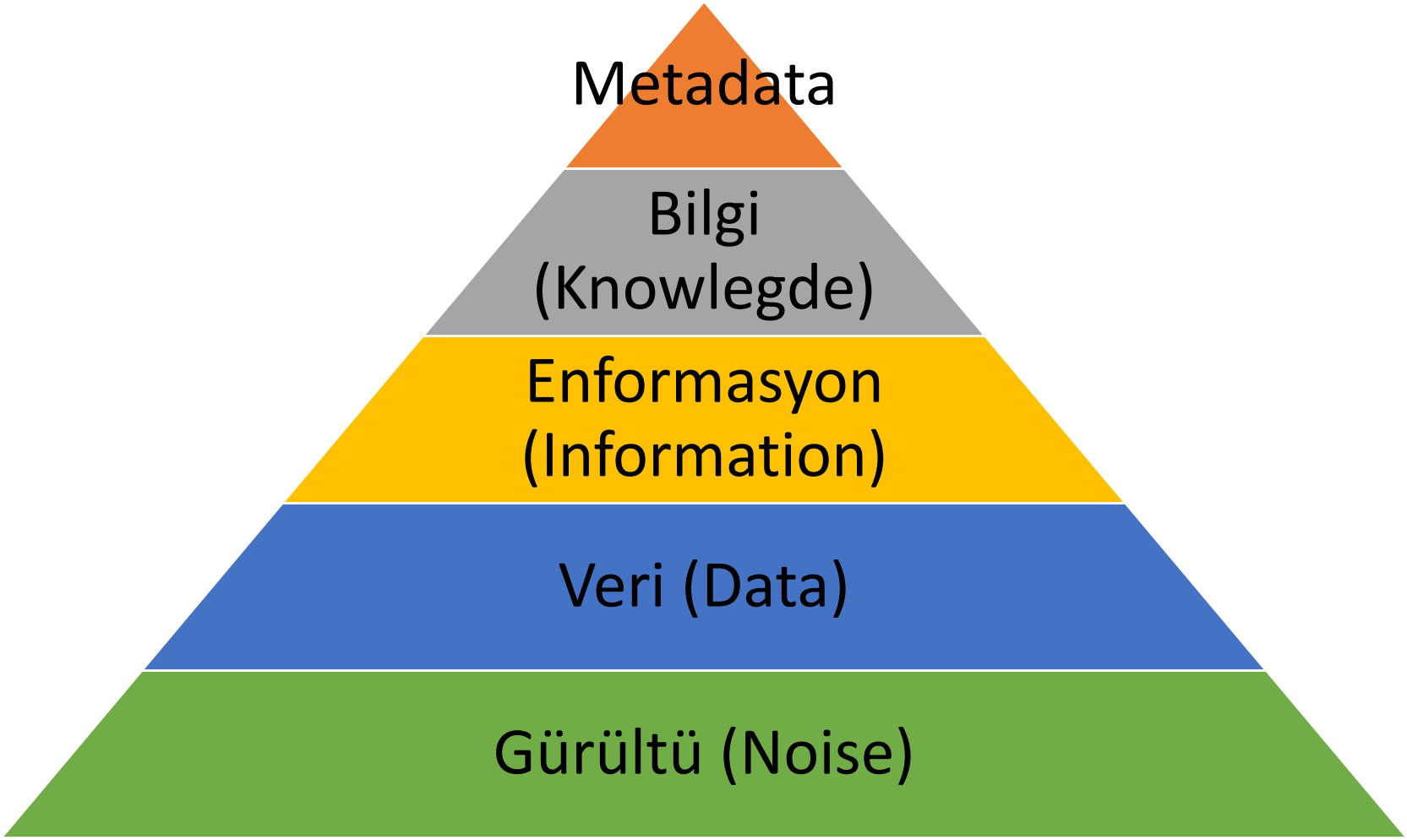
Bağımsız Değişken: Bağımlı değişken üzerinde etkisi olabileceği düşünülen bağımsız olarak değer alabilen değişkendir

Cinsiyet, Kıdem, Ekran Süresi, Sigara Kullanımı vd.

Bağımlı Değişken: Bağımsız değişkendeki değişime göre değer aldığı düşünülen değişkendir.

Başarı, Örgütsel Bağlılık, Göz Hastalıkları, Sağlık Sorunları vd.

Gürültü,Veri, Enformasyon, Bilgi ve Meta Bilgi

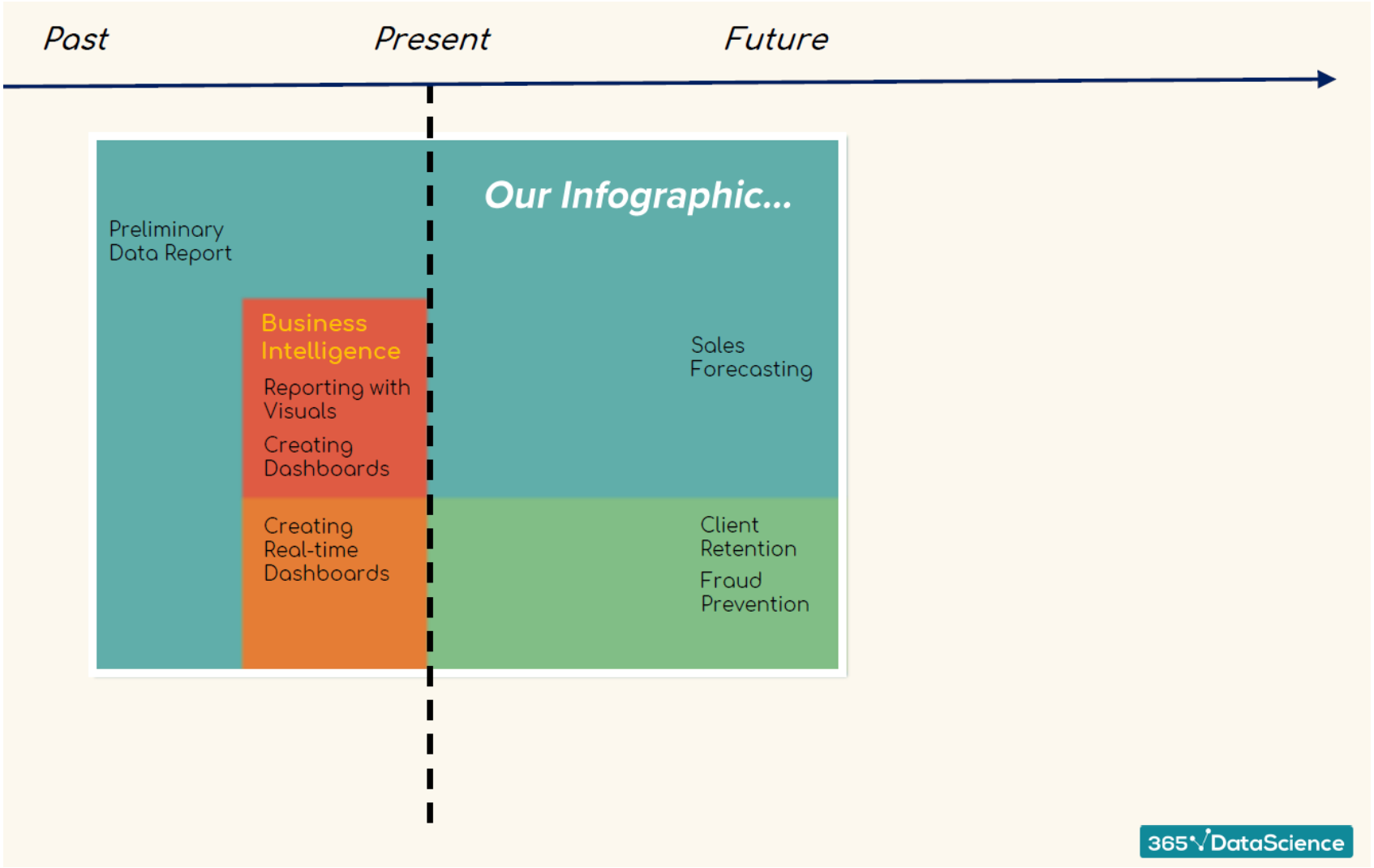


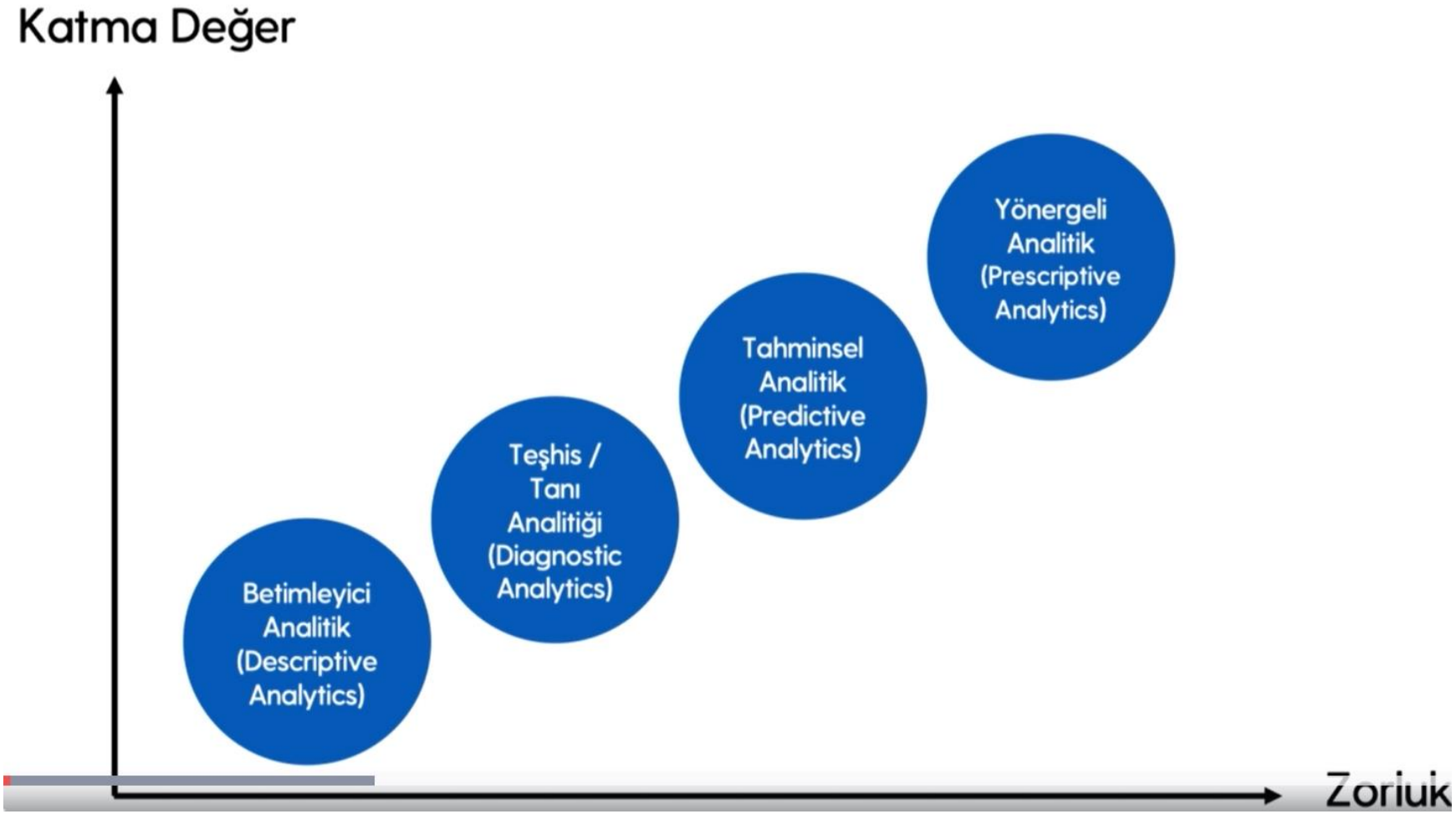
Tanımlar

- Veri: Kavramsal anlamda veri, kayıt altına alınmış her türlü olay, durum, fikirdir.
- Enformasyon (Information): Verilerin ilişkilendirilmiş, düzenlenmiş, anlamlandırılmış veya işlenmiş halidir.
- Bu haliyle enformasyon, potansiyel olarak içinde bilgi barından bir veri halindedir.
- Enformasyonun, bilgiye dönüşmesi, bireyin onu algılaması, özümsemesi ve sonuç çıkarmasıyla gerçekleşir.

Örnek

- 1 2 5 83 8 4178 63 Tüm sayılar: gürültü
- 2353 sayısı: bir veridir (Gürültünün somut Hale gelmesi)
- 2353 bir dahili telefon numarası bir enformasyon (verinin anlamlı hali)
- 2353 Alinin dahili numarası bir bilgi (knowledge)
- Alinin telefonunda iki tane 3 var. Meta data ‘bilgi hakkında bilgi’


Past
Present
Future
Our Infographic...
Preliminary Data Report
Business Intelligence
Reporting with Visuals
Creating Dashboards
Creating Real-time Dashboards
Sales Forecasting
Client Retention
Fraud Prevention
365 DataScience



Kaynak: https://www.veribilimiokulu.com/blog/veri-bilimi-ve-yapay-zeka-nedir/

- **Betimleyici Analitik:** “Ne olmuş?” sorusuna yanıt aranır. Veriyi betimlediğimizde mod, medyan, standart sapma veya görselleştirme teknikleriyle basit raporlar oluşturduğumuzda betimleyici analitik yapmış oluruz. Şirketin ilk üç ay ne kadar ürün sattığını gösteren çizelge yapmak buna örnek olarak verilebilir.
- **Teşhis Tanı Analitiğ:** “Neden, Neden olmuş, Nasıl olmuş?” sorularının yanıtını verir. Betimledikten sonra görmüş olduğumuz durumun neden olduğunu sorgular, yani teşhis tanı analitiği yapmış oluruz.
- **Tahminsel Veri Analitiği:** “Ne olacak?” sorusuna yanıt verir. Geleceksel tahmin yapmak için kullanılır. Satışların ne olacağını tahmin etme örnek olarak verilebilir.
- **Yönergeli Analitik:** “Nasıl olmalı, Ne olmalı?” sorularına yanıt verir. Olasılıkları tahmin ettikten sonra, başarıyı arttırmak için ne yapmalıyım diye soru gelirse, iş-aksiyon kararları alarak başarıyı arttırmaya çalışmaktır.

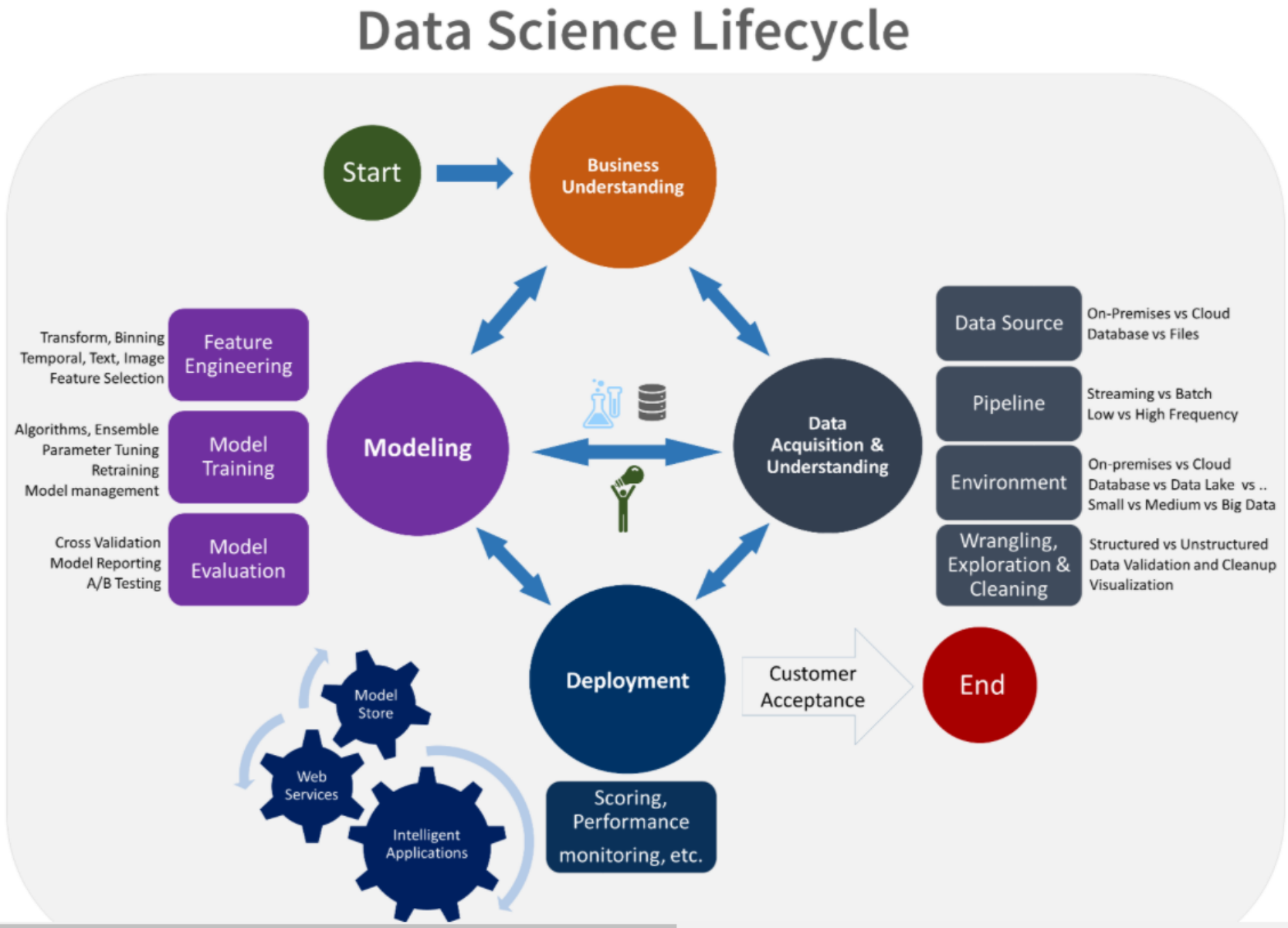
Süreç sonunda beklenen:

Veri Analizinin Bazı Kullanım Alanları

- Pazarlama
- Bankacılık
- Sigortacılık
- Elektronik Ticaret
- Eğitim-Öğretim
- Taşımacılık-Ulaşım-Konaklama
- Finansal servisler
- Sağlık
- Sosyal medya analizi
- Öneri Sistemleri
- Telekominikasyon
- …

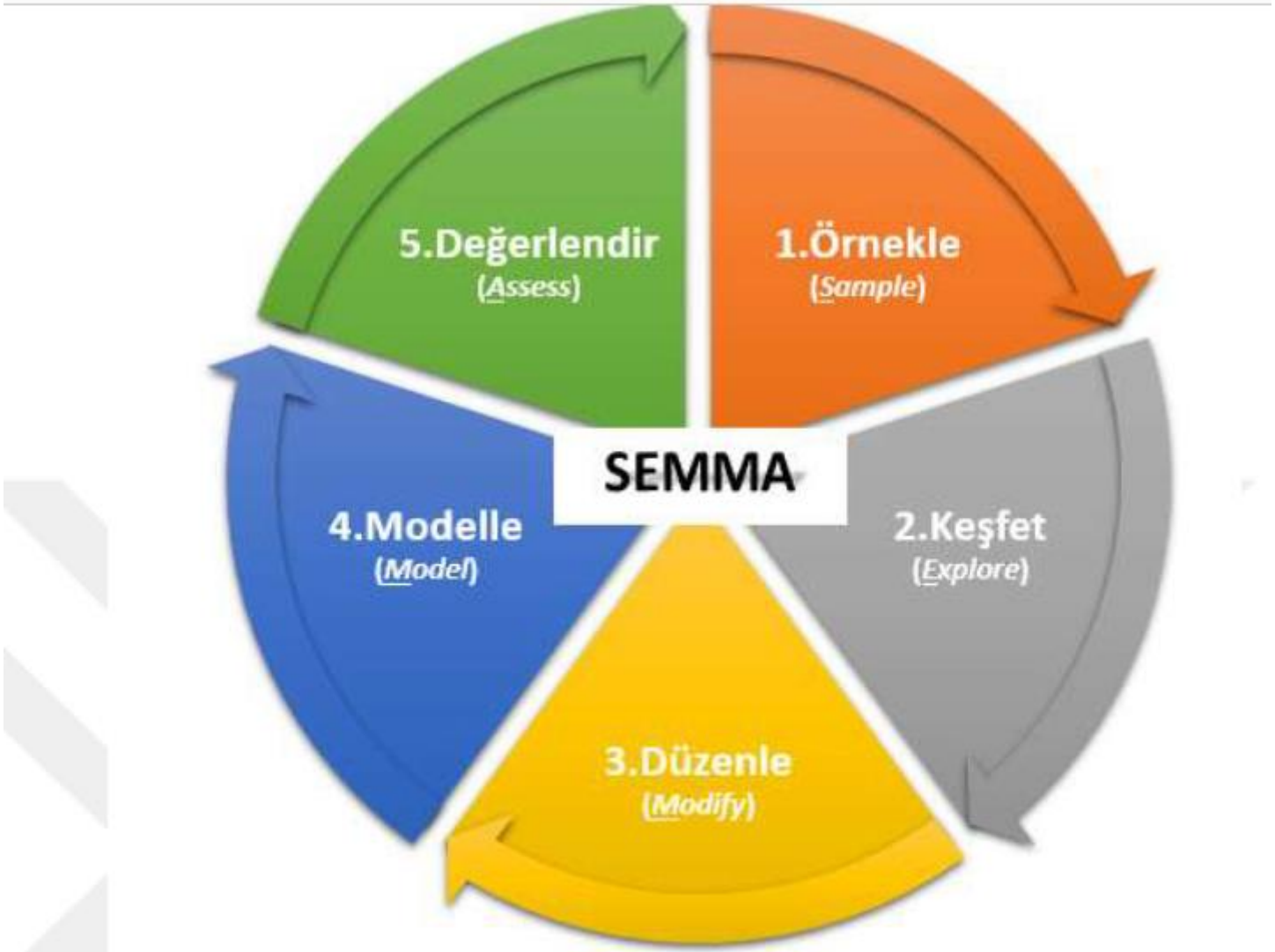
Veri Projesinin Adımları

Microsoft'un dokümanlarında yer alan döngü



Proje Ana Başlıkları		Alt Görev Başlıkları
A.Business Understanding		A1. Çözülmek İstenen Problemin ya da İhtiyaçların Tanımlanması
		A2. İlgili İş Birimleriyle Konunun Tartışılması ve Beklentilerin Belirlenmesi
		A3. Çıktıların, Çözüm Yönteminin ve Başarı Ölçüm Yaklaşımlarının Belirlenmesi
B.Literature and Best Practice Search		B1. Teknik Literatür Taraması ve İlgili Algoritmaların (gerekli ise) Öğrenilmesi
		B2. İlgili problem sektörde uygulanmış mı? Nasıl uygulanmış? Nasıl sonuçlar elde edilmiş?
C.Data Understanding		C1. Projenin dokunacağı olası tüm tabloların incelenmesi ve proje için uygunluğunun araştırılması
		C2. Veriye ulaşım için hangi alt yapıların kullanılacağının ve gerekirse veri taşıma işlemlerinin yapılması
D.Feature Engineering	D1. Tidy Süreci	D1_1. Proje için gerekli olan değişkenlerin tablolardan toplanması
		D1_2. Satır sütun kesişiminde gözlemsel dönüşümün sağlanması
	D2. Keşifci Veri Analizi	D2_1. Veri setinin ruhuna ulaşma: Değişim, dağılım ve korelasyonların kavranması
		D2_2. Veri seti için öngörülen durumların sınanması ve bulguların kaydedilmesi
		D2_3. Tek, çift ve çok değişkenli analizler ile değişkenlerin yapısının kavranması
		D2_4. Pattern Detection ve Veri Görselleştirme
	D3. Veri Manipülasyonu	D3_1. Çok değişkenli aykırı gözlem ve eksik gözlem incelemesi ve bu problemlerin giderilmesi
		D3_2. Kavranan değişim, dağılım ve korelasyonlara göre değişken seçimi yapılması
		D3_3. Modelleme için kullanılacak algoritmalara karar verilmesi ve gerekliliklerinin incelenmesi
E.Modelling		E1. Validasyon Yönteminin Seçilmesi ve Verinin Bölünmesi
		E2. Aday algoritmaların ve uygun değişkenlerin seçilmesi
		E3. Tüm aday modellerin kurulması ve varsayımlarının incelenmesi
F.Model Evaluation		F1. Model Çıktılarının Yorumlanması
		F2. Test hatalarının karşılaştırılması, parametre optimizasyonu ve uygun modelin seçilmesi
		F3. Canlı veri ile modelin test edilmesi
G.Running in Production		G1. Production tarafında çalışacak job'lara ve periyotlarına karar verilmesi
		G2. Job'ların kodlanması ve scheduler yardımıyla planlanması
		G3. Modelin canlıdaki performansının test edilmesi
		G4. Modelin gelir odaklı test edilmesi

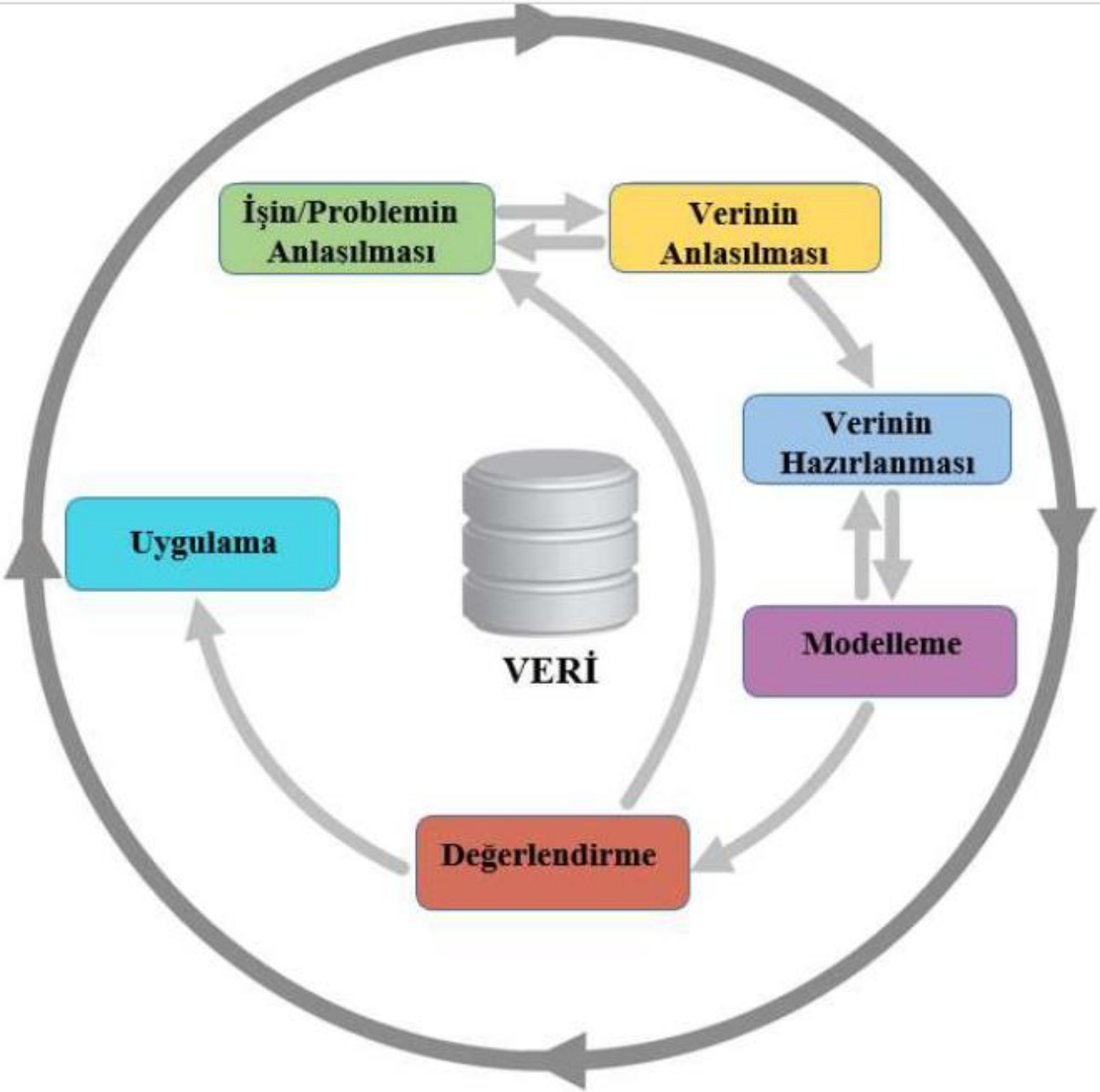
Kaynak: https://www.veribilimiokulu.com/blog/veri-bilimi-proje-dongusu/



Şekil 2.7: SEMMA süreci.

SEMMA, SAS Institute tarafından veri madenciliği sürecini daha anlaşılabilir bir hale getirmek ve bir yol haritası çıkarmak üzere, yine SAS Institute tarafından geliştirilen SAS Enterprise Miner veri madenciliği aracı için oluşturulmuş süreçtir. Bu veri madenciliği aracı için oluşturulması sebebiyle bu sistemin sınırları dışında çalışmaması sürecin dezavantajıdır (Rohanizadeh ve Moghadam, 2009). SEMMA, *Sample* (Örnekle), *Explore* (Keşfet), *Modify* (Düzenle), *Model* (Modelle) ve *Assess* (Değerlendir) şeklinde bir açılıma sahip olup toplamda beş adımlı bir süreçtir.

Süreç örnekle adımı ile başlarken, bu adımda verinin örneklenmesi gerçekleştirilir. Örnekleme işleminde bilgi çıkarımına olanak sağlayacak ve modellerin uygulanmasında performansı düşürmeyecek şekilde bir veri örneğinin seçilmesine önem verilir. Keşfetme verinin anlaşılmasına yönelik verinin özetlendiği, aykırı değerlerin belirlendiği adımdır. Düzenle adımında uygulamaya sokulacak nitelik alanları seçilip, gerekli dönüşüm işlemleri yapılarak bir nevi veri ön işleme gerçekleştirilmektedir. Modelleme aşamasında farklı veri madenciliği yöntemleri kullanılarak modeller elde edilmektedir. Değerlendirme olan sona adımda ise elde edilen modellerin performans göstergeleri karşılaştırılarak doğru modelin seçilmesi ve bu



Şekil 2.8: CRISP-DM süreci.

CRISP-DM (*CRoss-Industry Standard Process for Data Mining*) en çok tercih edilen diğer veri madenciliği süreçlerinden bir tanesi olup ilk sürümü 2000 yılında SPSS, NCR (Teradata), Daimler-Chrysler ve OHRA işbirliği ile geliştirilmiştir.

Yeni veri tiplerinin varlığı, elde edilen sonuçların çeşitli operasyonel sistemler ve mevcut iş süreçleri ile entegrasyonu, geniş ölçekli veri tabanlarının yerinde analizi yerine bu veri tabanlarından çekilecek analitik veri setleri ile işlem yapma, veri madenciliği ve analitik süreçler ile ilgili daha çok sayıda insanın eğitimi gibi konularda iş gereksinimlerinin artması var olan sürümün geliştirilmesi sonucunu doğurmuştur (Mariscal ve diğ., 2010).

İşin Anlaşılması/Problemin Belirlenmesi: Mevcut durumun değerlendirilmesi, hedeflerin belirlenmesi, proje planının oluşturulması işlemleri gerçekleştirilerek, veri madenciliği ile sonuç üretilmek istenilen problem belirlenir.

Verinin Anlaşılması: Birinci adımda belirlenen probleme göre gereksinin duyulan verinin ne olduğu belirlenmeye çalışılır. Buna bağlı olarak ihtiyaç duyulan veri toplanır. Analizlerde kullanılacak veri tek bir kaynağa bağlı olmaksızın farklı veri kaynaklarından elde edilen veri setlerinin bütünleştirilmesi ile de elde edilebilir. Bu adımda verinin anlaşılabilirliği için istatistiksel özetleme, kümeleme, aykırı değer analizi gibi yöntemlerden yararlanılabilir.

Verinin Hazırlanması: Bu adımda verinin ön işlemesi gerçekleştirilmektedir. Veri ön işleme doğru analiz sonuçlarının elde edilebilir olması için oldukça önem arz eden bir aşamadır. Çünkü analizlere ve probleme uygun veri seti, doğruluğu daha yüksek olan sonuçlar elde edilmesini

- **Veri Bütünleştirme:** Veri madenciliği için kullanılacak veri setinin tek bir kaynaktan temin edilmesi gerekmemektedir. Ancak farklı kaynaklardan gelen verinin farklılık göstereceği düşünülürse bu veri kümelerinin bir arada birbiri ile uyumlu hale getirilmesi gerekmektedir. Farklı kaynaktaki verinin tek bir fiziki kaynakta bir araya getirilmesi veri konsolidasyonu, farklı kaynaklarda yer alan verinin kopyasının oluşturulması veri yayınımı ve farklı kaynaklardaki verinin sanal olarak bir araya getirilmesi veri federasyonu olmak üzere veri bütünleştirme bu üç yöntem ile gerçekleştirilmektedir (Akpınar, 2014).

- **Verinin Temizlenmesi:** Daha önce de bahsedildiği üzere çeşitli hatalardan ötürü veri gürültülü, eksik veya tutarsız olabilmektedir. Verinin temizlik aşaması verinin bu tarz olumsuz özelliklerinden kurtularak kalitesi arttırılmış bir veri setine dönüştürülmesi amaçlanmaktadır. Verinin eksik olması özniteliklere karşılık gelen değerlerin olmamasını, gürültülü veri kullanıcı giriş hataları veya sistemsel hatalardan kaynaklı olarak meydana gelen hatalı öznitelik değerlerini ifade etmektedir. Örneğin; bir müşteri veri setinde müşteriye ait yaş öznitelik değeri 250 olarak gözüküyorsa bir insanın yaşı 250 olamayacağından bu bir gürültülü veridir. Benzer şekilde müşterinin yaş öznitelik değeri 25 iken doğum tarihi öznitelik değeri 20.05.1982 ise aynı müşteriye ait birbiri ile ilişkili öznitelikler arasında tutarsızlığın olduğu görülmektedir.

- Veri seti içerisindeki eksik öznitelik değerine sahip kayıtlar veya öznitelik alanları çıkartılabilir.
- Eksik değerler yerine yeni değerler girilebilir.
    - Veriyi analiz edecek uzman tarafından belirlenen değer girilebilir.
    - Sayısal veri tipleri için o özniteliğe ait tüm değerlerin ortalaması, kategorik veri tipleri için en çok tekrar eden değer girilebilir.
    - Entropi tabanlı çalışan karar ağacı yöntemleri kullanılabilir.
    - Öznitelik değer dağılımı çıkartılarak bu dağılıma göre rastgele değer belirlenebilir.

- Maksimum olasılık tahmini, Monte Carlo yöntemleri gibi model tabanlı yaklaşımlar uygulanabilir.

Bir diğer veri temizleme işlemi aykırı değer analizidir. Aykırı değer, veri seti içerisinde belirli bir ölçüye göre diğer öznitelik değerlerin açıkça farklılık gösteren, diğer öznitelik değerlerine göre beklenenin dışında değerlere sahip değer olarak tanımlanmaktadır (Bakar ve diğ., 2006). Aykırı değer analizi ile aykırılık gösteren bu değerlerin ortaya çıkartılması amaçlanmaktadır. Aykırı değer analizi sınıflandırma problemine benzemek ile beraber analiz edilen veri tabanı kayıtları içerisinde aranılan aykırı değer sayısı oldukça az olması ile sınıflandırma probleminden ayrılmaktadır (Petrovskiy, 2003). Aşırı değer analizi, olasılık ve istatistiksel yöntemler, lineer modeller, yakınlık tabalı modeller gibi çeşitli yöntemler aykırı değer analizi için kullanılmak ile beraber bu modellerin hangisinin seçileceği konusunda veri tipi, verinin büyüklüğü, ilişkili aykırı örneklerin varlığı, yorumlanabilirlik gibi özellikler önem arz etmektedir (Aggarwal, 2015).

Veri tutarsızlığı farklı veri kaynaklarından elde edilen verinin hetereojen yapısı ve kaynaklardan gelen tablolardaki uyuşmazlıklar nedeniyle aynı nesne için birbiri ile uyuşmayan değerlerin ortaya çıkması olarak tanımlandığı gibi farklı kaynaklardan gelen veri bütünleştirildikten sonra veriyi tasvir etme noktasında da tutarsızlıkların olabileceği söylenebilir (Anokhin ve Motro, 2001). Veri seti içerisinde tutarsızlığın ortaya çıkartılması ve çözümü için çeşitli yöntemler önerilmek ile beraber en temel düzeyde ortalama, medyan, mod, standart sapma gibi istatistiksel yöntemler ile tutarsızlıkların saptanması mümkündür (Akpınar, 2014).

- **Verinin dönüştürülmesi:** Veri seti içerisinde yer alan her veri nesnesi için o nesneye ait öznitelik değerinin dönüştürülmesi işlemidir. Veri dönüştürme yöntemleri şu şekilde özetlenebilir (Tan ve diğ., 2006; Gezer, 2016):

  - Gürültülü veriden kurtulmak üzere kutulama, regresyon ve kümeleme tekniklerinin kullanılması.
  - Var olan niteliklerin kulanılarak analizlere uygun yeni öznitelik alanlarının oluşturulması.
  - Öznitelik değerlerinin özetlenerek ifade edilmesi (satış rakamlarının günlük yerine aylık ortalamasının alınması, benzer şekilde hava sıcaklılarının günlük yerine haftalık ortalamasının alınması gibi).

- Veri dönüşümünün uygulanması.
- Sayısal değerlere sahip öznitelik alanlarının bu değerlerini kategorik hale getirmek için ayrıklaştırma işleminin yapılması.

Ayrıklaştırma işlemi, sürekli olan öznitelik değerlerinin kategorik değerlere dönüştürülmesi bir başka ifadeyle veri kaybının en az olacak şekilde sürekli olan öznitelik değerlerinin sonlu komşu aralıklarına dönüştürülme işlemidir (Jin ve diğ., 2007). Bir müşteri veri setinde yer alan yaş nitelik alanında her müşteriye ait ayrı ayrı yaş değeri yerine "genç-orta yaşlı-yaşlı" olacak şekilde üç kategori oluşturularak müşterinin yaşı hangi ategoriye denk geliyorsa o kategori değerini yazmak örnek olarak verilebilir (Koçoğlu, 2012).

Veri setinde yer alan öznitelik alanlarına ait değerlerin aralığı değişiklik gösterebilir. Örneğin; bir öznitelik alanının aldığı değerler 0-100000 olabileceği gibi diğer bir öznitelik alanına ait değerler 0-1 arasında değişebilir. Büyük aralıkta değişen değerlere sahip öznitelikler analiz sonuçlarını, özellikle uzaklık ölçüsü kullanan modellerde kendi yönlerinde etkileyebilirler. Dolayısıyla böyle bir durumda geniş aralıkta değerler alan özniteliğin baskınlığını engellemek üzere veri seti üzerinde düzenlemeye gitmek gerekmektedir. Veri seti içerisinde yer alan öznitelik değerlerinin aynı aralığa denk gelecek şekilde dönüştürülmesi işlemine veri dönüştürme (normalizasyon) denir. En sık kullanılan veri dönüşümü yöntemleri doğrusal veri dönüşümü ve z-skor veri dönüşümüdür.

Doğrusal veri dönüşümü işleminde tüm değerler [0-1] aralığına taşınmaktadır. Veri seti içerisindeki en küçük değer 0'a eşitlenirken en büyük değer 1'e eşitlenmektedir. Veri seti içerisindeki diğer değerlerin [0-1] aralığındaki değerlere dönüştürülmesi için denklem (2.1) uygulanmaktadır.

$$x_{\text{normalDeger}} = \frac{x - x_{min}}{x_{maks} - x_{min}} \tag{2.1}$$

Z-skor veri dönüşümü yönteminde ise ortalama değer (μ) ve standart sapma (σ_x) değeri kullanılarak özniteliklere ait tüm değerler aynı aralık içerisinde yer alacak şekilde dönüştürülür. Bu veri dönüşümü yöntemi için denklem (2.2)'den yararlanılmaktadır.

$$x_{\text{normalDeger}} = \frac{x - \mu}{\sigma_x} \tag{2.2}$$

Veri dönüştürme

- Veriyi bazı durumlarda veri madenciliği çözümlemelerine aynen katmak uygun olmayabilir.Değişkenlerin ortalama ve varyansları birbirinden önemli ölçüde farklı olduğu taktirde büyük ortalama ve varyansa sahip değişkenlerin diğerleri üzerindeki baskısı daha fazla olur ve onların önemli rollerini önemli ölçüde azaltır.
- Bu nedenle bir dönüşüm yöntemi uygulanarak söz konusu değişkenlerin normalleştirilmesi veya standartlaştırılması uygun bir yol olacaktır.

- **Verinin İndirgenmesi:** Bazı analiz çalışmalarında tüm veri seti yerine veri setini en iyi tanımlayacak şekilde özet veri seti kullanılmaktadır. İndirgeme işlemi de boyutu yüksek olan veri setlerinin amaca yönelik olarak en iyi şekilde temsili için gerçekleştirilen ön işleme adımıdır (Fayyad ve diğ., 1996). Veri indirgeme için gereksiz değerlerin silinip orijinal verinin karakterinin korunduğu özniteliklerin silinmesi, kayıtların silinmesi ve özniteliklere ait değerlerin silinmesi olmak üzere üç farklı yöntem kullanılmaktadır (Kantardzic, 2011).

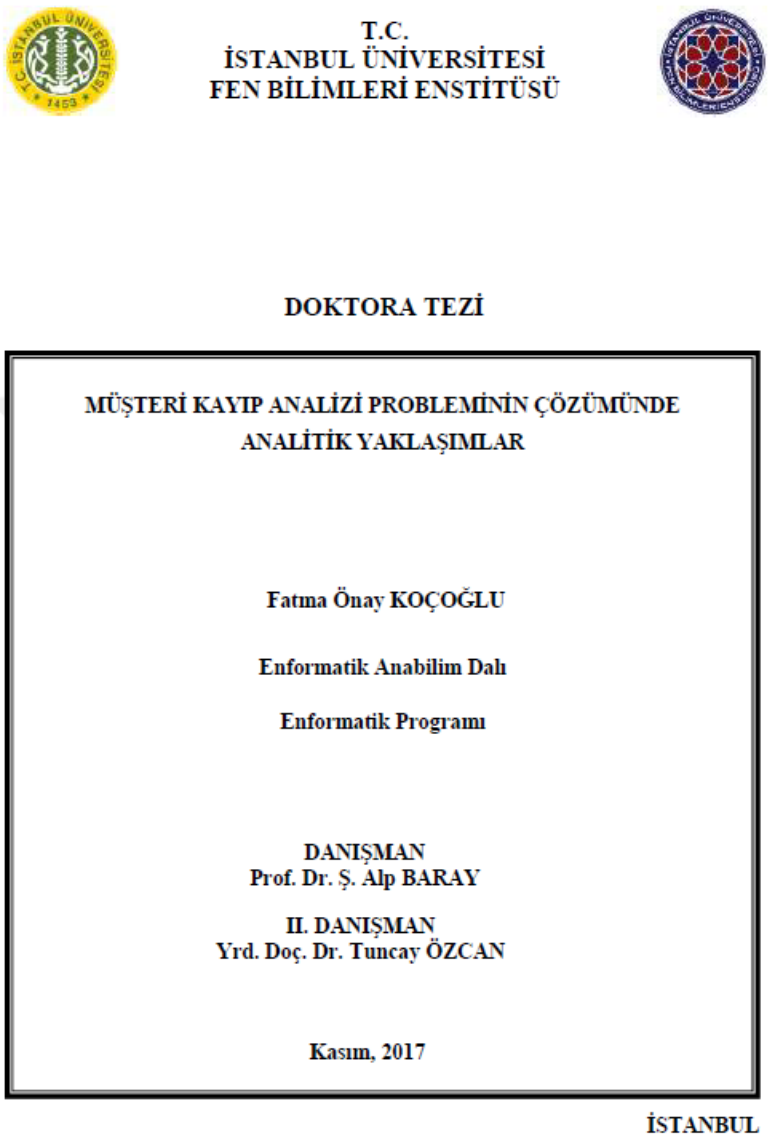
Modelleme: Benzer problemler için farklı veri madenciliği yöntemleri kullanılarak çeşitli modeller oluşturulabilmektedir (Wirth ve Hipp, 2000). Bu aşamada çeşitli algoritmalar kullanılarak modeller oluşturulmakta, modeller uygulanarak parametreleri en iyi şekilde sonuç vermek üzere optimize edilmeye çalışılmaktadır (Azevedo ve Santos, 2008). Kullanılacak yöntemler amaca yönelik olarak sınıflandırma, kümeleme, birliktelik kuralları işlevlerini yerine getirecek şekilde seçilmektedir. Modellerin ileride performans değerlendirilmesinin yapılabilmesi için ayrıca bu aşamada eğitim ve test verileri de oluşturulmakta, çeşitli performans ölçüleri hesaplanmaktadır.

Değerlendirme: Bu aşamada, bir önceki aşamada oluşturulan farklı modellerin, çeşitli performans geçerleme ve değerlendirme yöntem ve ölçülerine göre değerlendirilmesi yapılmaktadır. Hangi modelin en iyi performansı sergilediği veya amaca hangi modelin uygun olduğu değerlendirilerek ilk adımda belirlenen problemin çözümü için önerilen modele karar verilmektedir. Değerlendirme için geliştirilen çeşitli yöntem ve ölçüler olmak ile beraber bu yöntem ve ölçülerin açıklamalarına 2.5 ve 2.6 başlıklarında yer verilmiştir.

- Kaynak: Yukarıdaki metinler

Yanda künyesi verilen tezden

Alınmıştır.



T.C.
İSTANBUL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DOKTORA TEZİ

MÜŞTERİ KAYIP ANALİZİ PROBLEMİNİN ÇÖZÜMÜNDE
ANALİTİK YAKLAŞIMLAR

Fatma Önay KOÇOĞLU

Enformatik Anabilim Dalı

Enformatik Programı

DANIŞMAN
Prof. Dr. Ş. Alp BARAY

II. DANIŞMAN
Yrd. Doç. Dr. Tuncay ÖZCAN

Kasım, 2017

İSTANBUL

Veri İndirgeme

- Veri madenciliği uygulamalarında bazen çözümleme işlemi uzun süre alabilir.Eğer çözümlemeden elde edilecek sonucun değişmeyeceğine inanılıyorsa veri sayısı ya da değişkenlerin sayısı azaltılabilir.
- Veri indirgeme değişik boyutlarda yapılabilir;

  a)Veriyi birleştirme veya veri küpü

  b)Boyut indirgeme

  c)Veri sıkıştırma

  d)Örnekleme

  e)Genelleme

e) Makine öğrenmesi algoritmasını uygulama

- Söz konusu algoritmalar sınıflama, regresyon, kümeleme ve birliktelik kuralları.

Uygulama: Değerlendirme sonucunda farklı modeller arasında en iyi performansa sahip model seçilerek, problemin çözümü ortaya konmaya çalışılmaktadır. Yani modelin aktif olarak kullanılmaya başladığı evredir.

Örnek-Web Madenciliği

- **Örnek:** Web kayıtlarındaki bilgi keşif süreci
-
- Web sitesinin yapısını inceleme
- Verileri seçme: tarih aralığı belirleme
- Veri ayıklama, önişleme: gereksiz kayıtları silme
- Veri azaltma, veri dönüşümü: kullanıcı oturum-ları belirleme
- Algoritma seçme: kümeleme(Örn: k-ortalama, EM, DBSCAN...)
- Model değerlendirme/yorumlama: değişik kullanıcı grupları için sıkça izlenen yolu bulma
- Uygulama alanları: öneri modelleri, kişiselleştirme, ön belleğe alma