TÜRK TIBBİ ONKOLOJİ DERNEĞİ
Hayat için bilimin izinde...

Biyoinformatik ve Karar Verme

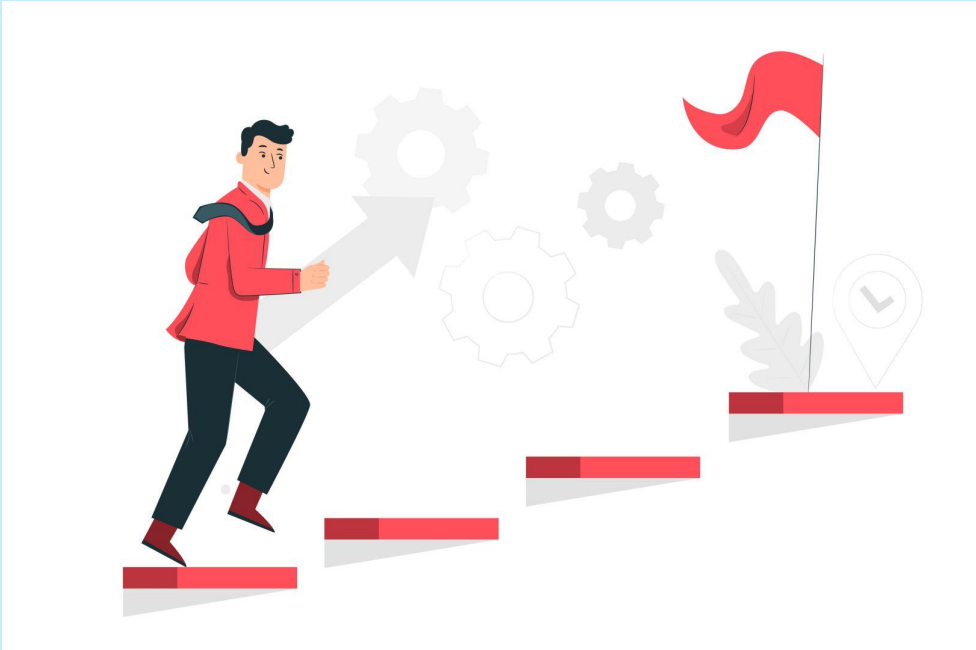
Dr Levent Korkmaz

Medikal Onkoloji Uzmanı
Bilgisayar Mühendisliği Fakültesi Öğretim Üyesi - İstinye Üniversitesi
WisdomEra - CEO

TÜRK TIBBİ ONKOLOJİ DERNEĞİ

Sunum Planı

- Biyoinformatik
  - nedir ?
  - karar vermede izlenen yol nedir ?
- Veri madenciliği
  - nedir ?
  - nerelerde kullanılır ?
  - kimler yapar ?
- Biyoinformatikte veri madenciliği
  - nasıl kullanılır ?



TÜRK TIBBİ ONKOLOJİ DERNEĞİ

Biyoinformatik

Biyoinformatik

- Biyoenformatik,
  - biyolojinin çeşitli dallarını, özellikle moleküler biyoloji,
  - bilgisayar teknolojisi ve veri işleme teknolojilerini
  - bünyesinde barındıran bilimsel disiplin.

Biyoinformatik

- İnsan Genom Projesinin başlangıcından beri biyoinformatiğin deneysel kısmında yayınlanan sonuçların sayısı önemli ölçüde artış gösterdi
- Günümüz araştırma alanı ve teknolojisindeki gelişmelerle, genetik ve proteomik çalışmalardan elde edilen veriler, klinik ve diğer çalışmalardan elde edilen verilerle birleşerek hastalıkların daha iyi anlaşılarak karakterize edilebilmesine sağlamaktadır.

TÜRK TIBBİ ONKOLOJİ DERNEĞİ

Biyoinformatik

- Büyük veri setlerinin artmasıyla birlikte biyolojik problemleri çözme ve **karar verme süreçlerinde veri madenciliği** kritik bir yer almaya başlamıştır.
- Biyoinformatik açısından bilim bugün **direkt olarak veri analiziyle** biyolojik olarak önemli örüntüler tespit etme şeklinde ilerlemektedir.
- Dolayısıyla biyoinformatik hakkında daha detaylı bilgiye **veri madenciliği kavramını ve süreçlerini** dikkatle inceleyerek ilerleyebiliriz.

TÜRK TIBBİ ONKOLOJİ DERNEĞİ

Veri madenciliği

Veri madenciliği - nedir ?

- Büyük ölçekli milyonlarca veriye sahip yazılım sistemlerinden, ihtiyacı karşılayacak değerli verilerin elde edilmesi işlemine ***Veri Madenciliği*** denilmektedir.
- Böylece veriler arasındaki ilişkiler tespit edilebilir.
- Ve gerektiğinde ileriye yönelik doğru tahminlerde bulunmak mümkün hale gelmektedir.
- Ana amaç ***karar destek mekanizmaları*** olarak adlandırılan sistemler için **değerli olan veriyi** belirli yöntemler ve işlem süreçleri sonrası **ortaya çıkarmaktır.**

Veri madenciliği - tarihçesi

Veri madenciliği - tarihçesi

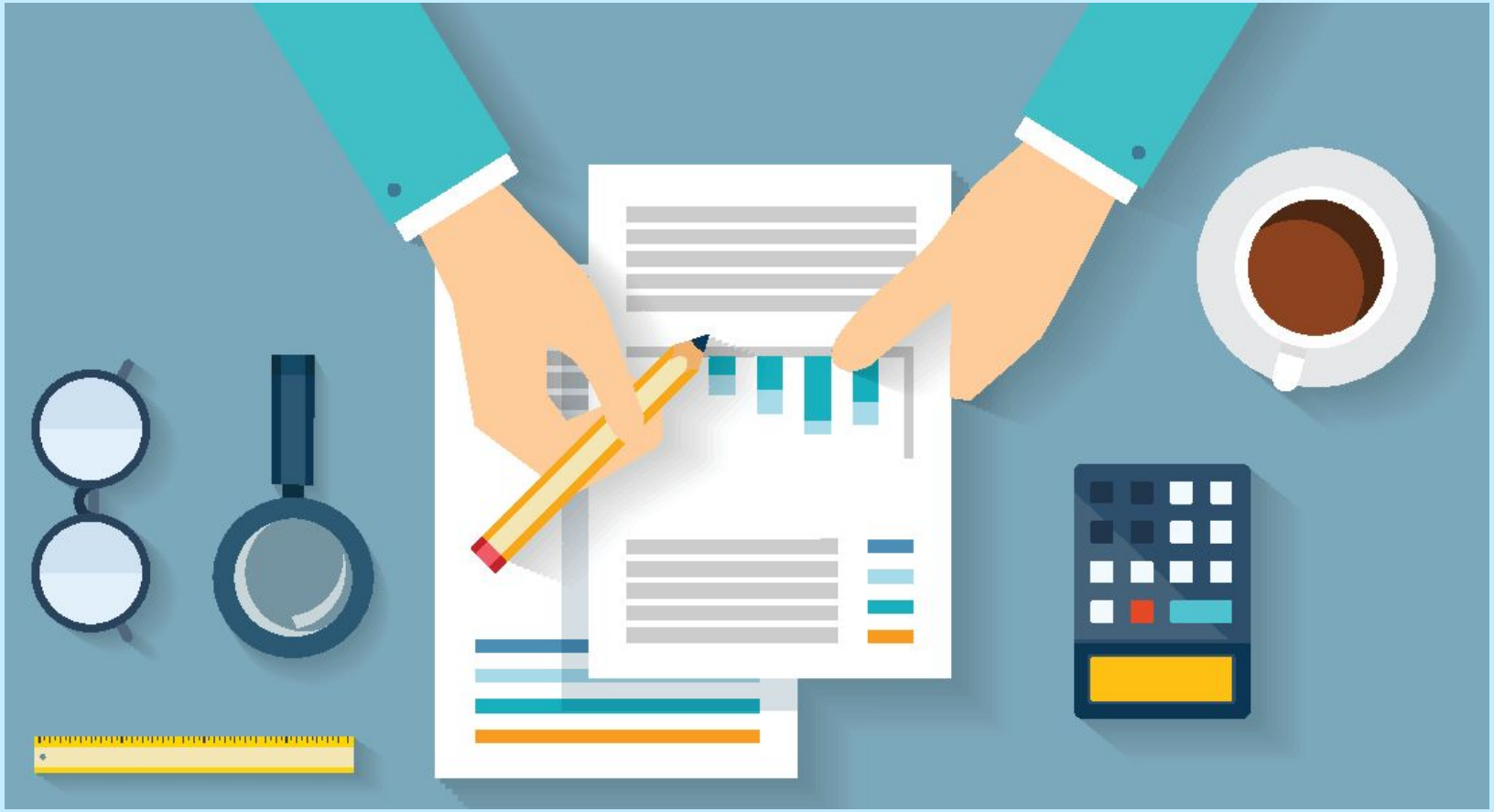
- 1950’lerde ilk bilgisayarlar matematiksel sayımlarda kullanılıyordu
- 1960’larda Veri Koleksiyonları, Veri tabanı kullanımı başladı
- 1970’lerde ilişkisel veri modeli ve ilişkisel RDMS (Relational Database Management System) uygulamaları geliştirildi.
- 1980’lerde İlişkisel RDMS kullanımı yaygınlaşmaya başladı.
- 1990’larda Günlük işlerde derlenen verinin nasıl değerlendirilebileceği sorgulanmaya başladı

Veri madenciliği - tarihçesi

- 1991’de Knowledge Discovery in Real Databases (veritabanlarında bilgi keşfi) tanımı ve kavramları ortaya konuluyor
- 1992’de Veri Madenciliği konusunda yazılımların geliştirilmeye başlanması
- 2000’lerde Veri Ambarları ve Veri Madenciliğinin yaygınlaşması oluyor
- Yapay zeka, makina öğrenmesi ve derin öğrenme
- …..

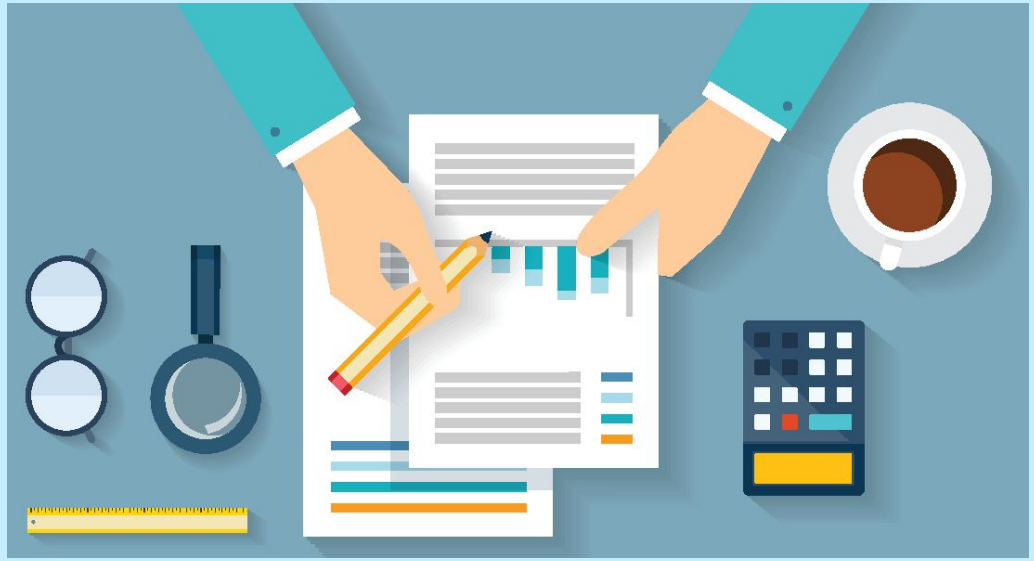
TÜRK TIBBİ ONKOLOJİ DERNEĞİ

Veri madenciliği - Uygulama Alanları



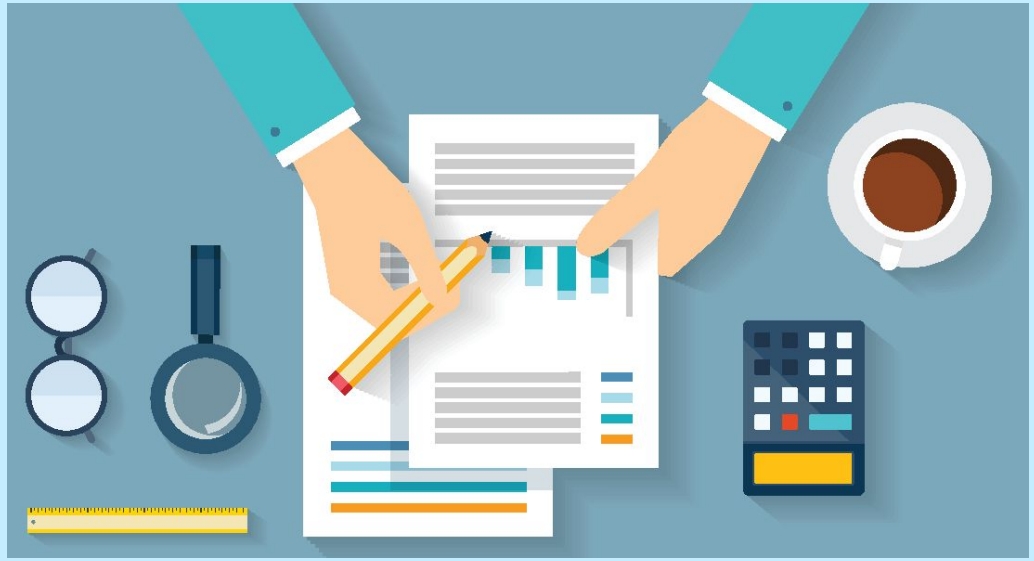
Veri madenciliği - Uygulama Alanları

- Son 10 yılda piyasadaki hemen her alanda çeşitli şekillerde veri madenciliği yapılmaktadır.
- Madencilik her türlü elektronik ortama dayalı işte, pazarlamacılıkta, bankacılık ve sigortacılıkta artık temel bir disiplin haline gelmiştir.
- Örneğin pazarlama alanında kullanılan veri madenciliği sistemi, müşterilerin satın alma alışkanlıklarını tespit ederek bunlara yönelik stratejiler izlemektedir.



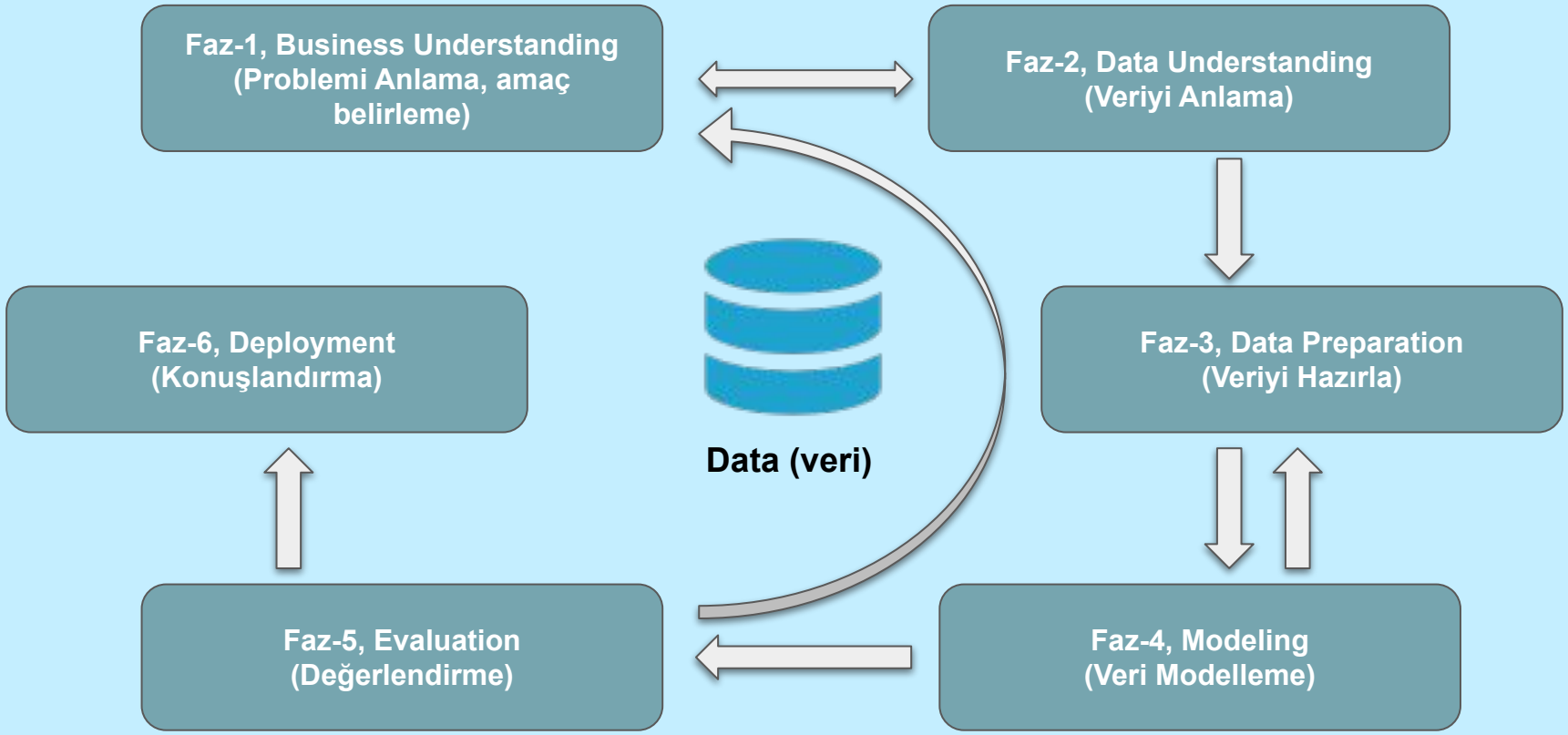
Veri madenciliği - Uygulama Alanları

- Müşterilerin yaş, eğitim, cinsiyet ve lokasyon gibi temel özelliklerinin incelenmesiyle ortaya çıkan satış tahminleri ve pazar sepeti analizleri, sektöre oldukça fayda sağlamaktadır.
- Biyoinformatikte biyolojik verinin analiz edilerek muhtemel veri kümelerini tespit etmek ve karar vermek



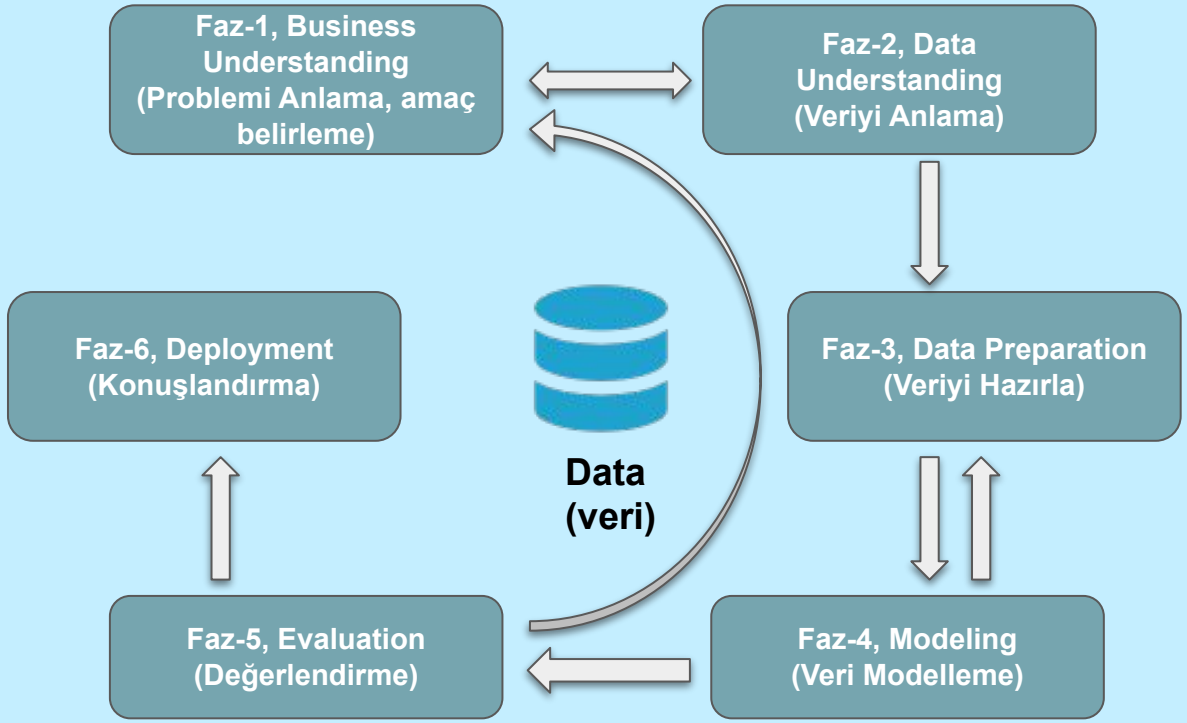
TÜRK TIBBİ ONKOLOJİ DERNEĞİ

Veri madenciliği - Süreç



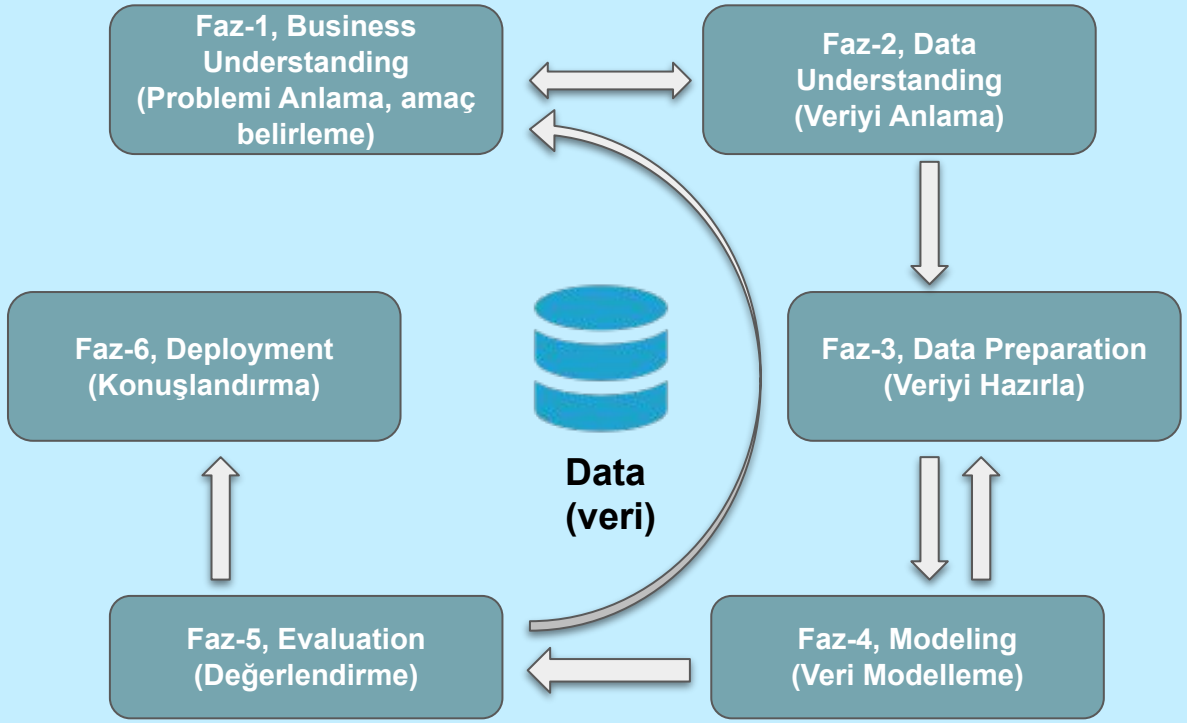
Veri madenciliği - Süreç

- Faz-1, Business Understanding
  - Veri madenciliği projelerindeki en önemli adımdır.
  - Çünkü projenin amacı bu adımda belirlenir.
- Faz-2, Data Understanding
  - Veri toplanır.
  - Veri kalitesi problemleri belirlenir.
  - Veriden ilk görüler çıkarılır.
- Faz-3, Data Preparation
  - Veri toplandıktan sonra toplam veriden alt kümeler oluşturulup oluşturulmayacağı netleştirilir
  - uygun veri setlerinin otomasyonu sağlanır.



Veri madenciliği - Süreç

- Faz-4, Modeling
  - Veri ile modelleme algoritması belirlenir
  - Modellemenin hedef çıktılarına göre farklı algoritmalar mevcuttur.
- Faz-5, Evaluation
  - Oluşturulan modelin test sürecinden geçirilmesi işlemidir.
- Faz-6, Deployment
  - Hazırlanan model analizlere sunulur.
  - İş süreçlerinde kullanılacak hale getirilir.



TÜRK TIBBİ ONKOLOJİ DERNEĞİ

Veri madenciliği - Modeller

Veri madenciliği - Modeller

- Veri madenciliğinde kullanılan modeler
  - Tahmin Edici ve
  - Tanımlayıcı
  - olmak üzere ikiye ayrılmaktadır

Veri madenciliği - Modeller - Tahmin Edici Modeller (Predictive)

- Sonuçları bilinen verilerden hareket ederek
  - bir model oluşturup,
  - sonuçları bilinmeyen veri kümeleri için
  - sonuç değerlerinin tahmin edilmesidir.
    - Sınıflama
    - Regresyon
    - Zaman Serisi Analizi

Veri madenciliği - Modeller - Tanımlayıcı Modeller (Descriptive)

- Karar vermeye
  - rehberlik etmede kullanılabilecek
  - verilerdeki örüntülerin tanımlanmasını sağlamaktadır.
    - Kümeleme Yöntemi
    - Birliktelik Kuralı

Veri madenciliği - Modeller - Birliktelik Kuralları (Association Rules)

- Veri madenciliğinde en iyi bilinen yöntemlerden biridir.
- Büyük veritabanlarında
  - birbiriyle ilişkili değişkenleri ve
  - aralarında bağlantının büyüklüğünü tespit etmek için kullanılan bir yöntemdir.
- Apriori, Carma, Eclat, Sequence, GRI.. birliktelik yönteminde kullanılan algoritmalardır.

Veri madenciliği - Modeller - Sınıflandırma ve Tahmin

- Veri eğilimlerini açıklamak için veri sınıfları oluşturulur.
- Böylece, gelecekteki bir verinin bu sınıflardaki yeri bu modele sorulabilir.
- Sınırlandırma algoritma örnekleri
  - Decision Tree (Karar ağaçları),
  - Random Forest (Rassal ormanlar),
  - Navie Bayes,
  - KNN (K en yakın komşu)

Veri madenciliği - Modeller - Sınıflandırma ve Tahmin

- Örneğin
  - kredi başvurusu yapacak bir müşteriye kredi verilebilirliği,
  - geçmiş bilgilerden hastalık teşhisi,
  - ses tanıma,
  - kullanıcı davranışlarını belirleme.
  - **biyoinformatikte kanser olan vakalardaki genlerin sınırlandırma çalışması**
  - birer sınıflandırma örnekleridir.

Veri madenciliği - Modeller - Kümeleme Analizi (Cluster Analysis)

- Kümelemede amaç
  - dağınık halde duran verileri
  - özelliklerine göre birleştirip
  - işlenebilir hale getirmektir.
- Sınıflandırmaya benzer ama
  - aradaki fark
  - kümelerin ***önceden belirlenmemiş*** olmasıdır.
- Bu analiz sonucu müşteri profili oluşturmak içinde kullanılır.
  - K-Means,
  - K-Metoids kümeleme algoritmalarıdır.

Veri madenciliği - Modeller - Kümeleme Analizi (Cluster Analysis)

- Algoritmalar
  - K-Means,
    - En yaygın kullanılan gözetimsiz öğrenme yöntemidir
    - En eski kümeleme algoritmalarından olan k-means, 1967 yılında J.B. MacQueen tarafından geliştirilmiştir.
    - Her verinin sadece bir kümeye ait olabilmesine izin verir.
    - Bu nedenle, keskin bir kümeleme algoritmasıdır.
    - Eşit büyüklükte küresel kümeleri bulmaya eğilimlidir.

Veri madenciliği - Modeller - Kümeleme Analizi (Cluster Analysis)

- Algoritmalar
  - K-Metoids,
    - algoritmasının temeli, verinin çeşitli yapısal özelliklerini temsil eden k tane temsilci nesneyi bulma esasına dayanır.
    - Temsilci nesne, diğer nesnelere olan ortalama uzaklığı minimum yapan kümenin en merkezi nesnesidir.
    - Bu nedenle, bu bölünme metodu her bir nesne ve onun referans noktası arasındaki benzersizliklerin toplamını küçültme mantığı esas alınarak uygulanır.

Veri madenciliği - Modeller - Kümeleme Analizi (Cluster Analysis)

- Örneğin
  - marketlerde farklı müşteri gruplarının keşfedilmesi ve bu grupların alışveriş örüntülerinin ortaya konması,
  - şehir planlamasında evlerin tiplerine, değerlerine ve coğrafi konumuna göre gruplara ayrılması..
  - **biyoinformatikte işlevlerine göre genlerin sınıflandırılması,**
  - kümeleme örnekleridir.

Veri madenciliği - Modeller - Aykırılık Analizi (Outlier Analysis)

- Verilerin algoritmalar ile kontrol edilerek
  - verilerde aşırı sapma veya
  - aykırı değerlerin bulunma sürecidir.
- Veri madenciliği algoritmaları,
  - sıradışı verileri en aza indirme veya
  - ortadan kaldırmayı amaçlamaktadır.

Veri madenciliği - Modeller - Aykırılık Analizi (Outlier Analysis)

- Örneğin:
  - kredi kartlarının olağandışı kullanımının tespiti,
  - telekomunikasyon servislerinde olağandışı dolandırıcılık tespiti,
  - tıbbı tedavilerde olağandışı sonuçların tespiti
  - **biyoinformatikte sıradışı gen dizilimlerini tespiti**
  - için kullanılabilir.

TÜRK TIBBİ ONKOLOJİ DERNEĞİ

Veri madenciliği - Kimler yapar ?

TÜRK TIBBİ ONKOLOJİ DERNEĞİ

Veri madenciliği - Kimler yapar ?

- Milyonlarca hatta bazen milyarlarca farklı özelliğe sahip veri üzerinde çalışma yapmak,
  - sabır ve dayanıklılık gerektirmektedir.
  - ilgili konuda uzman olmak ve detaylı bilgi sahibi olmak bir diğer önemli noktadır. Bu sağlanırsa yapılan işin kalitesi artacak ve süresi kısalacaktır.

TÜRK TIBBİ ONKOLOJİ DERNEĞİ

Veri madenciliği - Kimler yapar ?

- Önemli düzeyde
  - matematik,
  - istatistik,
  - lineer cebir,
  - optimizasyon bilgisi,
  - modelleme teknikleri ve
  - gelişmiş yazılım becerisine sahip olmak ise olmazsa olmazdır.
- Yazılım dillerinden veri madenciliği için en uygun olanlar R ve Python dilleridir.

TÜRK TIBBİ ONKOLOJİ DERNEĞİ

Biyoinformatik ve Veri madenciliği

Biyoinformatik ve Veri madenciliği

- Veri madenciliğinin biyolojik veri analizi için katkıda bulunduğu noktalar,
  - Heterojen, dağıtılmış genomik ve proteomik veritabanlarının anlamsal entegrasyonu.
  - Çoklu nükleotit dizilerinin hizalanması, indekslenmesi, benzerlik araştırması ve karşılaştırmalı analizi.
  - Yapısal kalıpların keşfi ve genetik ağların ve protein yollarının analizi
  - İlişkilendirme ve yol analizi.

Biyoinformatik ve Veri madenciliği

- Veri madenciliğinin biyolojik veri analizi için katkıda bulunduğu noktalar,
  - Gen bulma,
  - Protein fonksiyon alanı tespiti
  - Fonksiyon motifi tespiti
  - Protein fonksiyon çıkarımı
  - Hastalık teşhisi
  - Hastalık prognozu
  - Hastalık tedavisi optimizasyonu
  - Protein ve gen etkileşim ağı

Biyoinformatik ve Veri madenciliği

- Veri madenciliğinin biyolojik veri analizi için katkıda bulunduğu noktalar,
  - Veri temizleme
  - Protein hücre altı yer tahmini
  - Protein ve DNA dizilerinin analizi
  - Mikrodizi verilerine dayalı kanser sınıflanması
  - Gen expresyon verilerinin kümelenmesi
  - Protein-protein etkileşimlerinin modellenmesi

TÜRK TIBBİ ONKOLOJİ DERNEĞİ

Teşekkürler

Dr Levent Korkmaz
lkorkmaz@wisdomera.io
levent.korkmaz@istinye.edu.tr
www.wisdomera.io