Veri Madenciliği ve Bilgi Keşfi

H1. Giriş (Veri Madenciliği Nedir?)

Dr. Öğr. Üyesi Alper Talha KARADENİZ
Samsun Üniversitesi
Yazılım Mühendisliği Bölümü

Veri Madenciliği Nedir?

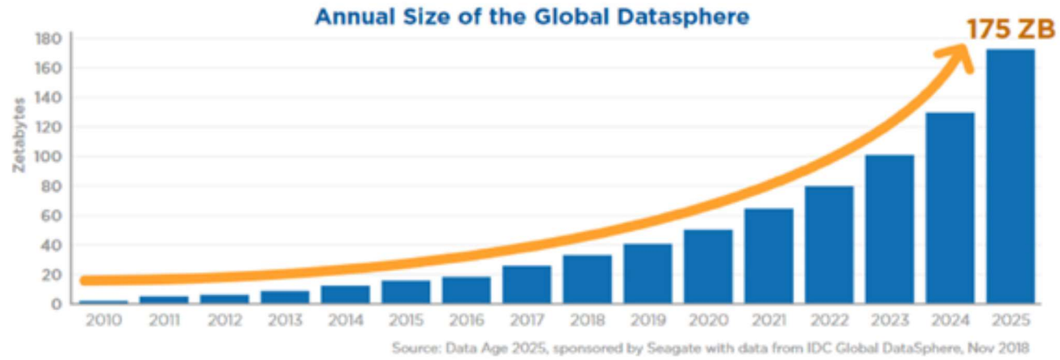
Veri madenciliği, çok büyük veri setlerinden anlamlı bilgiler ve desenler çıkartma sürecidir. İşletmeler, bilimsel araştırmalar, tıp ve e-ticaret gibi alanlarda yaygın olarak kullanılır. Veri madenciliği, veri analizi ve istatistiksel tekniklerin bir kombinasyonu olarak düşünülebilir. Bu alan, yapay zeka, makine öğrenimi ve veri bilimi gibi disiplinlerle de yakından ilgilidir.

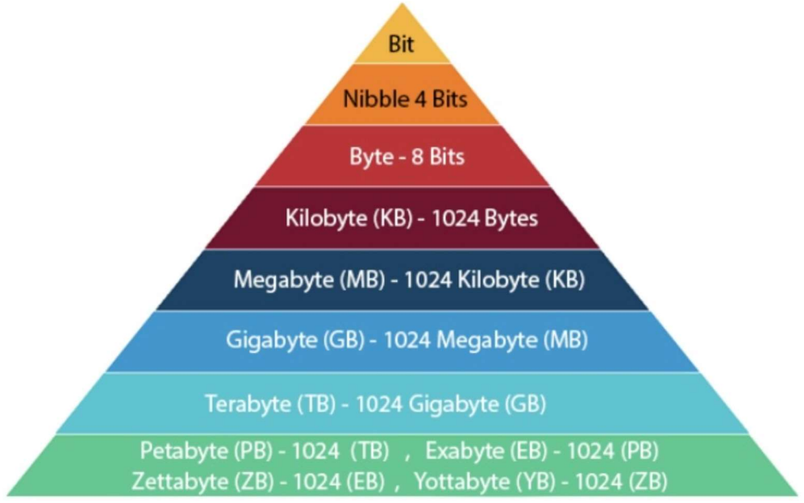
Veri madenciliği, anlamlı örüntüler ve eğilimler bulmak için büyük hacimli bilgilerin incelenmesini ve analiz edilmesini içerir.

Teknolojinin hızla gelişmesi ve yaygınlaşması sayesinde, eskiden yalnızca fiziksel olarak gerçekleştirilen birçok işlem bugün bilgisayar, cep telefonu ya da tablet üzerinden kolayca yapılabilmektedir. Sensörlerden alınan veriler, güvenlik amaçlı kullanılan retina ve parmak izi bilgileri, meteorolojik ve jeofizik gözlemler, tıbbi kayıtlar, banka ödemeleri, alışveriş işlemleri ve hastane randevularının tamamı, birkaç tuşa basılarak çevrimiçi şekilde tamamlanabilmektedir. Bu durum, dijital veri toplama ve saklama yöntemlerinin ne denli yaygın ve hayatımızın vazgeçilmez bir parçası hâline geldiğini açıkça göstermektedir. İnternet üzerinden yapılan her işlem, kullanıcıların kullandığı cihazın (telefon, tablet, bilgisayar vb.) türünden bağımsız olarak, ilgili veri tabanlarında ve sunucularda birikmektedir.

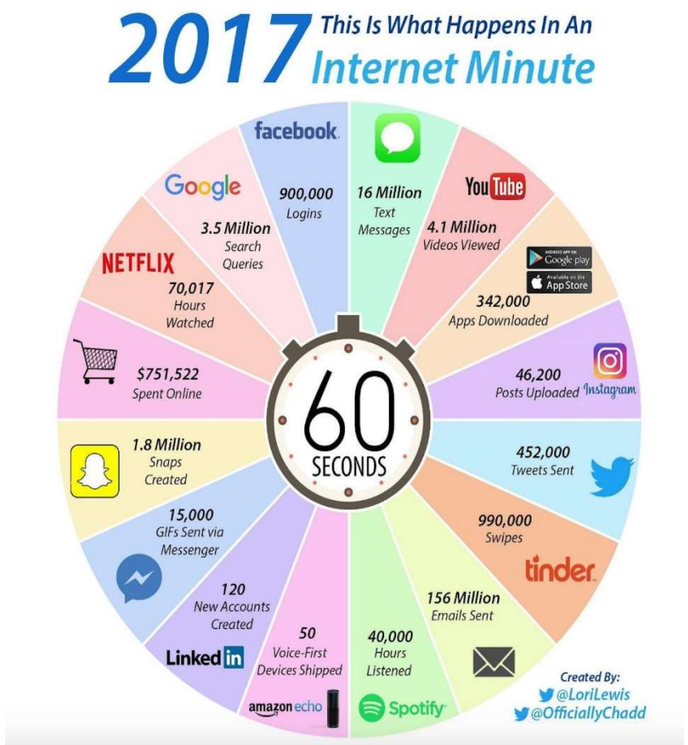
İşletmelerin sunucularında toplanan bu devasa veri yığınlarının analiz edilmesi ve içinden gerçek anlamda faydalı bilgilerin ayrıştırılması sürecine ise **Veri Madenciliği** adı verilmektedir. Bu süreç, kurumların iş stratejilerini belirlemesine, karar verme mekanizmalarını geliştirmesine ve müşterilerine daha iyi hizmet sunmasına yardımcı olacak önemli içgörüler üretmektedir.

IDC (İnternet Veri Merkezi), dijital dünyanın geleceğine yönelik gerçekleştirdiği araştırma çıktısına göre dünya üzerinde gerçek zamanlı veri miktarı 2018 yılında 33 ZettaBytes (ZB) iken 2025 yılına kadar beklenen rakam 175 ZettaBytes.



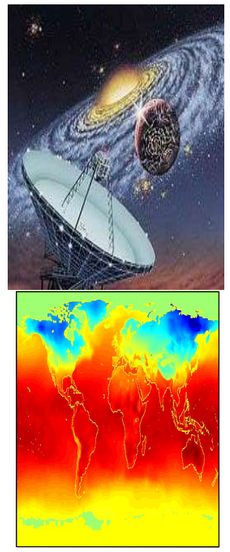


- Web (~50 milyar sayfa) (Google tarafından indekslendi)
- Çevrimiçi sosyal ağlar (Facebook'un 1,86 milyar kullanıcısı var - 2016)
- Öneri sistemleri (Netflix'te 93,8 milyon abone)
- Wikipedia'da 5,33 milyon İngilizce madde, 40 milyon 293 dilde madde ve daha fazlası



Neden Veri Madenciliği Yapıyoruz?

- Çok sayıda veri toplanıyor ve depolanıyor
- Web verileri, e-ticaret departmanda satın almalar/marketler
- Banka/Kredi Kartı işlemleri
- Daha iyi özelleştirilmiş hizmetler sunun
- Veriler muazzam hızlarda (GB/saat) toplanıyor ve depolanıyor
- Uydudaki uzaktan sensörler
- Teleskoplar gökyüzünü tarıyor
- Gen üreten mikrodizilerifade verisi
- Geleneksel teknikler ham veriler için uygulanamaz
- Veri madenciliği bilim insanlarına yardımcı olabilir
- Bilgisayarlar daha ucuz ve daha güçlü hale geldi
- Rekabet Baskısı Güçlü



Bilgi Güçtür. Ancak bilgiyi çıkarmak için veriye ihtiyaç vardır.

- Saklanmış
- Yönetilen
- Ve ANALİZ EDİLMİŞ

Veri madenciliği (verilerden bilgi keşfi)

- İlginç olanın çıkarılması (önemsiz olmayan, örtük,öncedenbilinmeyen ve potansiyel olarak yararlı) büyük miktardaki verilerden elde edilen bilgiler.
- Alternatif isimler; Veritabanlarında bilgi keşfi (madencilik), bilgi çıkarma, veri/desen analizi, bilgi hasadı, iş zekası, vb..

Data Mining is not …

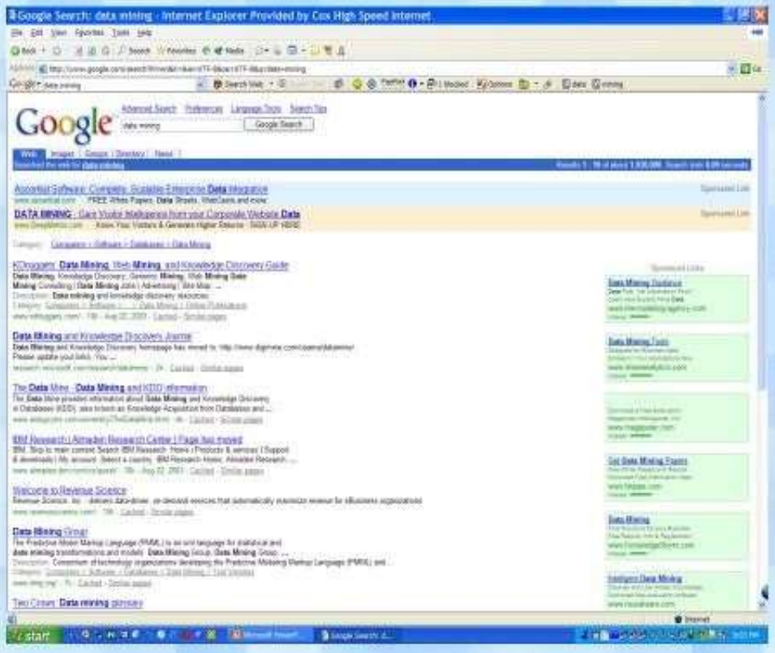
- Generating a histogram of salaries for different age groups

- Issuing SQL query to a database, and reading the reply

Data Mining is not …

- Searching for a phone number in a phone book

- Searching for keywords on Google



Data Mining is …

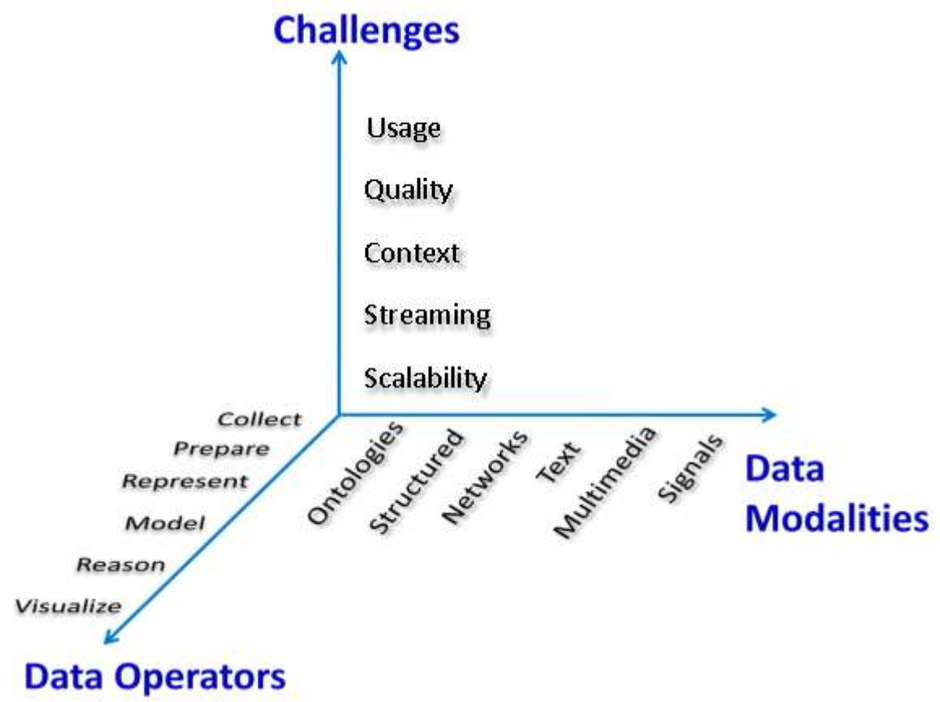
- Finding groups of people with similar hobbies

- Are chances of getting cancer higher if you live near a power line?

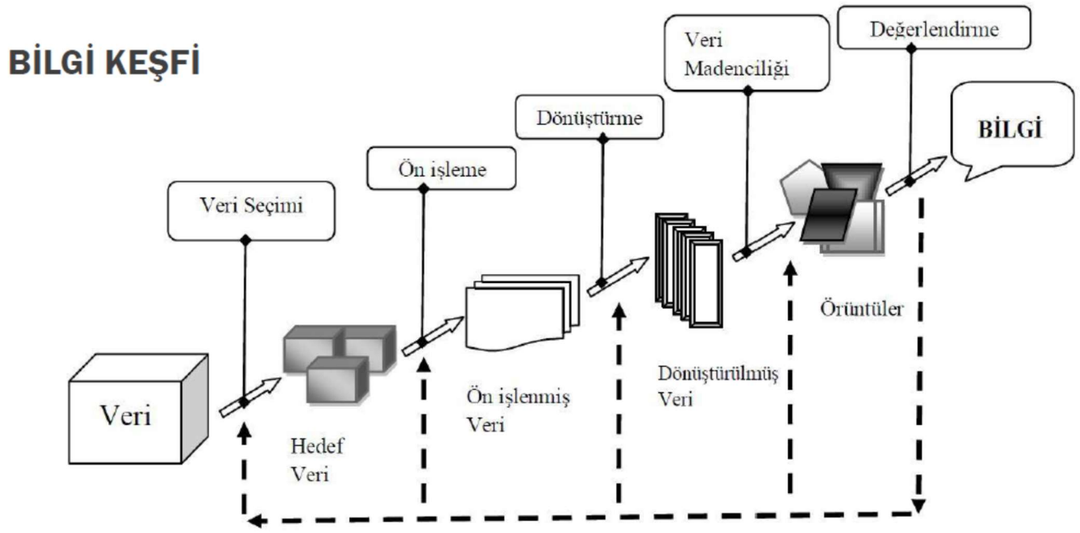
Tipik bir veri madenciliği süreci şu şekildedir:

Hedefinizi tanımlama: Örneğin, müşteri davranışı hakkında daha fazla şey öğrenmek istiyor musunuz? Maliyetleri azaltmak mı yoksa geliri artırmak mı istiyorsunuz? Dolandırıcılığı tespit etmek istiyor musunuz? Veri madenciliği sürecinin başında net bir hedef belirlemek önemlidir.
Verilerinizi toplama: Topladığınız veriler, hedefinize bağlı olacaktır. Kuruluşların tipik olarak birden çok veri tabanında depolanan verileri (ör. müşterilerin işlemler aracılığıyla gönderdikleri bilgiler vb.) vardır.
Verileri temizleme: Verilerin seçildikten sonra genellikle temizlenmesi, yeniden biçimlendirilmesi ve doğrulanması gerekir.
Verileri sorgulama: Bu noktada analistler, istatistiksel analizler yürüterek ve görsel grafikler ile tablolar oluşturarak veriler hakkında bilgi sahibi olurlar. Amaç, veri madenciliği hedefi için önemli olan değişkenleri belirlemek ve bir modele ulaşmak üzere ilk hipotezleri oluşturmaktır.
Bir model oluşturma: Veri madenciliği için farklı teknikler vardır ve bu aşamada amaç, en yararlı sonuçları üretecek bir veri madenciliği yaklaşımı bulmaktır. Analistler, amaçlarına bağlı olarak bir sonraki bölümde özetlenen yaklaşımlardan birini veya birkaçını kullanmayı seçebilirler. Model oluşturma, tekrarlanan bir süreçtir ve bazı modeller verilerin belirli şekillerde biçimlendirilmesini gerektirdiğinden veri biçimlendirmesinin tekrarlanmasını gerektirebilir.
Sonuçları doğrulama: Bu aşamada analistler, bulguların doğru olup olmadığını kontrol etmek için sonuçları incelerler. Sonuçlar doğru değilse modelin yeniden oluşturulup tekrar deneme yapılması gerekir.
Modeli uygulama: Ortaya çıkan içgörüler, sürecin başında tanımlanan hedefi gerçekleştirmek için kullanılabilir.

Verilerle uğraşırken nelere dikkat edilmeli?



- Veri (Data): Kayıt altına alınmış ve henüz işlenmemiş ham değerler (sayılar, metinler, log kayıtları vb.).
- Bilgi (Information): Verinin yorumlanmış ve anlam kazanmış hâli.
- Bilgi Keşfi (KDD- Knowledge Discovery in Databases): Ham verinin toplanması, temizlenmesi,dönüştürülmesi ve madencilik yöntemleriyle işlenmesi sonucu bilgiye dönüştürüldüğü bütün süreç.



Veri Madenciliği vs İstatistik

- Amaçları benzerdir
- Farklı yöntemler kullanırlar
- Veri madenciliğinde, çok sayıda olası hipotez araştırılır
- Veri madenciliği daha keşfedici bir veri analizidir
- Veri madenciliğinde çok daha büyük veri kümeleri vardır – algoritmik/ölçeklenebilirlik bir sorundur

Veri madenciliği vs Makine öğrenmesi

- Veri madenciliği için makine öğrenme yöntemleri kullanılır
  - Sınıflandırma, kümeleme
- Verinin miktarı fark yaratır
  - Veri madenciliği çok daha büyük veri kümeleriyle ilgilenir ve ölçeklenebilirlik bir sorun haline gelir

- Veri madenciliğinin daha mütevazı hedefleri vardır
  - Keşifte insan performansını hedeflemeden çeşitli sıkıcı görevleri otomatikleştirmek
  - Kullanıcıların yerini almak değil, onlara yardımcı olmak

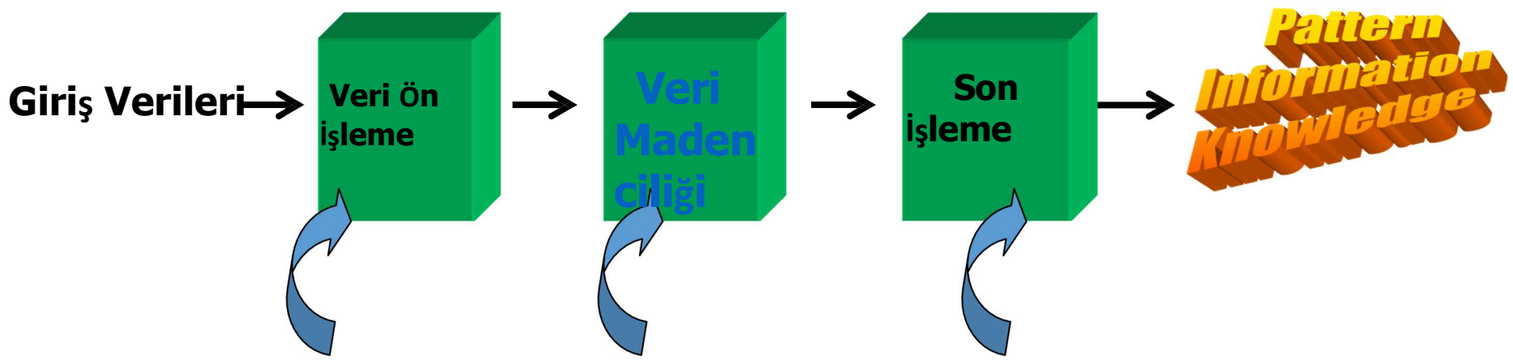
Veri Madenciliğinde;

- Veriler yüksek boyutludur
- Veri bir grafiktir
- Veri sonsuzdur/asla bitmez
- Veriler etiketlendi

Gerçek dünya problemlerini çözmek:

- Öneri sistemleri
- Pazar Sepeti Analizi
- Spam tespiti
- Çift belge tespiti

KDD (Veriden Bilgi Keşfi) Süreci: ML ve İstatistiklerden Tipik Bir Görünüm



Veri Madenciliği Türleri

- Tanımlayıcı modelleme
- Tahmine dayalı modelleme
  - Kuralcı modelleme

Tanımlayıcı modelleme

Bu, başarının veya başarısızlığın ardındaki nedenleri anlamak için geçmiş verilerdeki benzerlikleri veya grupları ortaya çıkarır (örneğin müşterileri ürün tercihlerine veya duygularına göre kategorize etmek). Örnek tekniklere şunlar dâhildir:

- **Birliktelik kuralları:** Bu, aynı zamanda pazar sepeti analizi olarak da bilinir. Bu tür veri madenciliği, değişkenler arasındaki ilişkileri araştırır. Örneğin birliktelik kuralları, hangi ürünlerin en çok birlikte satın alındığını görmek için bir şirketin satış geçmişini inceleyebilir. Şirket bu bilgileri planlama, kampanya ve tahmin için kullanabilir.
- **Kümeleme analizi:** Kümeleme, ortak özellikler paylaşan veri noktalarını alt kümelere ayırarak bir veri kümesi içindeki benzerlikleri belirlemeyi amaçlar. Kümeleme; müşterilerin satın alma davranışına, ihtiyaç durumuna, hayatının evresine veya pazarlama iletişimindeki tercihlerine göre bölümlendirilmesi gibi bir veri kümesi içindeki özellikleri tanımlamak için faydalıdır.
- **Aykırı değer analizi:** Bu model, anormallikleri, yani örüntülere tam olarak uymayan verileri belirlemek için kullanılır. Aykırı değer analizi özellikle dolandırıcılık tespiti, ağ giriş algılaması ve suç soruşturmalarında kullanışlıdır.

Tahmine dayalı modelleme

Bu modelleme, gelecekteki olayları sınıflandırmak veya bilinmeyen sonuçları tahmin etmek için daha derine iner (örneğin, bir kişinin bir krediyi geri ödeme olasılığını belirlemek için kredi derecelendirmesini kullanmak). Örnek tekniklere şunlar dâhildir:

- **Karar ağaçları** : Bunlar, bir dizi kriter listesine dayalı olarak bir sonucu sınıflandırmak veya tahmin etmek için kullanılır. Veri kümesini verilen yanıtlara göre sıralayan bir dizi basamaklı sorunun girdisini istemek için bir karar ağacı kullanılır. Bazen ağaç şeklinde bir görselle gösterilen karar ağacı, verilerde daha derine inerken belirli bir yöne ve kullanıcı girdisine izin verir.
- **Sinir ağları**: Bunlar, düğümlerin kullanımı yoluyla verileri işler. Bu düğümler girdilerden, ağırlıklardan ve bir çıktıdan oluşur. Veriler, insan beyninin işleyişine benzer şekilde, denetimli öğrenme yoluyla eşleştirilir. Bu model, bir modelin doğruluğunu belirlemek için eşik değerler vermeye uygun olabilir.
- **Regresyon analizi:** Regresyon analizi, bir veri kümesindeki en önemli faktörleri, hangi faktörlerin göz ardı edilebileceğini ve bu faktörlerin birbirlerini nasıl etkilediğini anlamayı amaçlar.

- **Sınıflandırma:** Bu, ele alınması gereken belirli bir soru veya zorluğa dayalı olarak veri noktalarının gruplara veya sınıflara atanmasını içerir. Örneğin, bir perakendeci belirli bir ürün için indirim stratejisini optimize etmek isterse kararlarını yönlendirmek için satış verilerine, envanter düzeylerine, kupon kullanım oranlarına ve tüketici davranış verilerine bakabilir.

Kuralcı modelleme

İnternet, e-posta, yorum alanları, kitaplar, PDF'ler ve diğer metin kaynaklarından gelen yapılandırılmamış verilerdeki artışla birlikte, metin madenciliğinin veri madenciliğine bağlı bir disiplin olarak benimsenmesi de önemli ölçüde arttı. Veri analistleri, gelişmiş tahmin doğruluğu için tahmine dayalı modellere dâhil etmek üzere yapılandırılmamış verileri ayrıştırma, filtreleme ve dönüştürme becerisine ihtiyaç duyar.

Veri Madenciliğindeki Büyük Zorluklar

- Veri madenciliği algoritmalarının verimliliği ve ölçeklenebilirliği
- Paralel, dağıtılmış, akış ve artımlı madencilik yöntemleri
- Yüksek boyutluluğun işlenmesi
- Gürültü, belirsizlik ve veri eksikliğinin ele alınması
- Veri madenciliğinde kısıtlamaların, uzman bilgisinin ve arka plan bilgisinin dahil edilmesi
- Desen değerlendirmesi ve bilgi entegrasyonu
- Çeşitli ve heterojen veri türlerinin madenciliği: örneğin, biyoenformatik, Web, yazılım/sistem mühendisliği, bilgi ağları
- Uygulama odaklı ve alan bazlı veri madenciliği
- Görünmez veri madenciliği (diğer fonksiyonel modüllere gömülü)
- Veri madenciliğinde güvenlik, bütünlük ve gizliliğin korunması

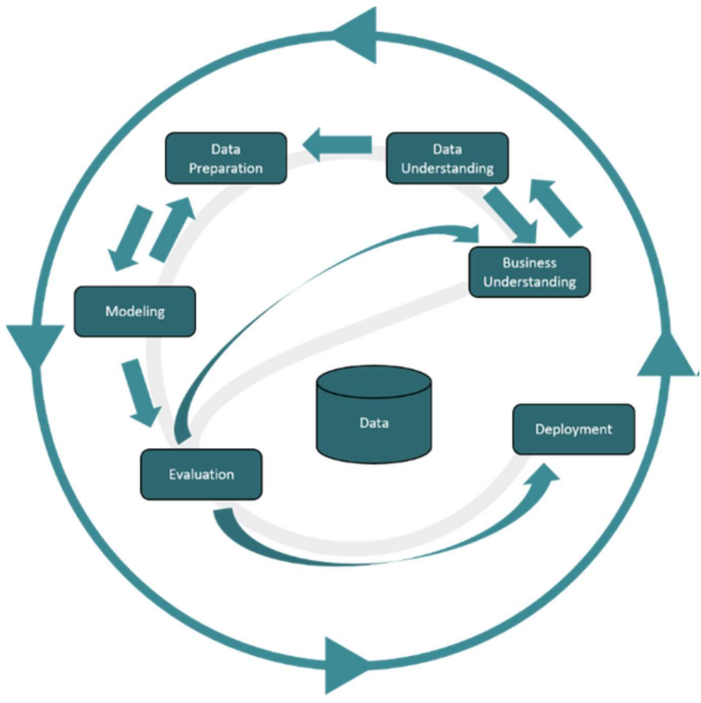
Veri Madenciliği ile ilgili Konferanslar ve Dergiler

- KDD Konferansları
  - ACM SIGKDD Veritabanlarında ve Veri Madenciliğinde Bilgi Keşfi Uluslararası Konferansı (KDD)
  - SIAM Veri Madenciliği Konferansı (SDM)
  - (IEEE) Uluslararası Veri Madenciliği Konferansı (ICDM)
  - Bilgi Keşfi ve Veri Madenciliği İlkeleri ve Uygulamaları Konferansı (PKDD)
  - Pasifik-Asya Bilgi Keşfi ve Veri Madenciliği Konferansı (PAKDD)
- Diğer ilgili konferanslar
  - ACM SIGMOD
  - VLDB
  - (IEEE) ICDE
  - WWW, İMZA
  - ICML, CVPR, NİP'ler
- Dergiler
  - Veri Madenciliği ve Bilgi Keşfi (DAMI veya DMKD)
  - IEEE Trans. Bilgi ve Veri Mühendisliği (TKDE)
  - KDD Keşifleri
  - ACM Trans. KDD'de

Veri Madenciliği Süreç Modelleri

CRISP-DM (Cross-IndustryStandard Process for Data Mining): Veri madenciliği projeleri için yaygın olarak kullanılan ve sektör bağımsız şekilde uygulanabilen bir süreç modelidir.

1. İş Anlayışı (BusinessUnderstanding)
2. Veri Anlayışı (Data Understanding)
3. Veri Hazırlığı (Data Preparation)
4. Modelleme (Modeling)
5. Değerlendirme (Evaluation)
6. Dağıtıma Alma (Deployment)



1. İş Anlayışı (Business Understanding)

Proje Hedefinin Belirlenmesi: Bu aşamada, veri madenciliği çalışmasının iş/kurumaçısından hangi soruları cevaplaması, hangi problemleri çözmesi ya da hangifırsatları ortaya çıkarması gerektiği belirlenir.

Mevcut Durum ve Kısıtların Analizi: Kaynaklar (zaman, bütçe, insan gücü), kullanılacakteknoloji altyapısı, veri gizliliği ve yasal düzenlemeler gibi unsurlar göz önündebulundurulur.

Başarı Kriterlerinin Tanımlanması: Örneğin, “Müşteri segmentasyonu yaparak pazarlama verimliliğini %10 artırmak” veya “Sınıflandırma modelinin doğruluğunu%90’ın üzerine çıkarmak” gibi ölçülebilir hedefler konulur.

Bu aşama, projeye yön veren temel çerçevedir. Eğer iş hedefi net olarak belirlenmezse, sonraki adımlarda yapılacak teknik çalışmaların sağlıklı bir şekilde hedefe ulaşması mümkün olmayabilir.

2. Veri Anlayışı (Data Understanding)

Veri Kaynaklarının İncelenmesi: İş hedefini gerçekleştirmek için hangi veri kümelerininkullanılacağı, bu verilerin nerede saklandığı, formatı (CSV, veritabanı, JSON vb.) veverilerin ne kadar güncel olduğu tespit edilir.
Ön Analiz (Exploratory Data Analysis – EDA): Temel istatistiksel bilgiler (ortalama,medyan, standart sapma), veri dağılımı (histogramlar, kutu grafikleri) ve olasıkorelasyonları içeren hızlı incelemeler yapılır.
Eksik ve Hatalı Değerlerin Tespiti: Verideki boş hücreler (missing values), aykırı değerler(outliers), çakışmalar (duplicates) gibi temizlenmesi veya düzeltilmesi gereken hususlarnot edilir.
Veri anlayışı aşaması, projeye ışık tutacak ilk bulguların elde edildiği kritik bir safhadır. Bu aşamadaki bulgular, sonraki adımlarda hangi yöntemlerin ve tekniklerin kullanılacağına dair fikir verir.

3. Veri Hazırlığı (Data Preparation)

Veri Temizleme (Data Cleaning): Eksik değerlerin yönetimi (silme, ortalama/medyan iledoldurma, KNN vb. yöntemler), aykırı değerlerin tespiti ve gerekiyorsa yenidenbiçimlendirilmesi (transform).
Veri Dönüştürme (Transformation): Veri tiplerini dönüştürme (kategorik-sayısal

vb.),normalizasyon/standardizasyon, metin işleme (tokenization, stop words temizliği),zaman serisi için sıralama gibi işlemler.

Veri Bütünleştirme (Integration): Farklı kaynaklardan gelen verilerin bir araya getirilmesi(örn. veritabanı + Excel + API verileri).

Özellik Mühendisliği (Feature Engineering): Yeni değişkenler (features) üretmek veyagereksiz değişkenleri (noise) elemek. PCA veya t-SNE gibi boyut indirgeme yöntemleride bu safhada gündeme gelebilir.

Bu aşama, “ham” veri ile makine öğrenmesi algoritmaları arasında köprü kurar. Model performansının büyük ölçüde veri hazırlığı kalitesine bağlı olduğunu unutmamak gerekir.

4. Modelleme (Modeling)

Algoritma Seçimi: Projenin iş hedeflerine ve veri yapısına uygun makine öğrenmesi/metotseçimi (sınıflandırma, regresyon, kümeleme, birliktelik kuralları vb.).

Model Eğitimi (Training): Seçilen algoritmaya uygun parametre ayarlamaları (ör. kararağacı derinliği, nitelik sayısı, regularizasyon parametresi), eğitim veri setiyle modelinoluşturulması.

Doğrulama ve Test (Validation & Testing): Modelin performansını ölçmek için doğrulamave test veri setlerinde değerlendirme yapmak. Overfitting veya underfitting gibiproblemlerin incelenmesi.

Hiperparametre Optimizasyonu: Grid Search, Random Search veya daha gelişmişyöntemler (Bayes optimizasyonu vb.) ile modelin en iyi ayarlarının aranması.

Bu aşama, veri madenciliğinin “çekirdek” kısmını oluşturur. Hangi modeli veya modelleri seçeceğiniz, hem veri türüne hem de daha önce tanımladığınız iş gereksinimlerine göre değişecektir.

5. Değerlendirme (Evaluation)

Model Performansının Ölçülmesi: Sınıflandırma için Doğruluk (Accuracy), F1, ROC-AUC;regresyon için MSE, RMSE, MAE gibi metriklerle değerlendirilir.
İş Hedefleriyle Karşılaştırma: Modelin iş gereksinimlerini karşılayıp karşılamadığına bakılır.Örneğin, müşterileri segmentlere ayırma konusunda yeterli başarıya ulaşıldı mı, tahminhatası tolere edilebilir düzeyde mi?
Model Değiştirme veya İyileştirme Kararı: Eğer sonuçlar yetersizse, iş anlayışını veya verihazırlığı aşamasını gözden geçirip modelin yeniden eğitilmesi gerekebilir.
Burada temel amaç, teknik başarı kadar iş hedefleriyle olan uyumu da sağlamaktır. En başarılı model, yalnızca istatistiksel olarak değil, iş değeri üretebilme kapasitesiyle de değerlendirilir.

6. Dağıtıma Alma (Deployment)

Modelin Gerçek Ortama Alınması: Eğitim ve test ortamında başarılı olan modeli, gerçekzamanlı veya toplu veri işleme senaryosunda kullanıma sokma.
Otomasyon ve İzleme: Modelin performansını düzenli olarak izlemek, veri değiştikçe (conceptdrift) modeli güncellemek veya yeniden eğitmek için otomasyon altyapısı kurmak.

Kullanıcı Geri Bildirimi ve Bakım: Son kullanıcıların veya iş birimlerinin modeli nasıl kullandığıtakip edilir, gerekli iyileştirmeler yapılır.

Raporlama ve Paydaş Yönetimi: Model sonuçlarının, metriklerin ve iyileştirme önerilerininpaydaşlara (yöneticiler, iş birimleri, müşteriler) düzenli raporlanması.

Bu aşama, elde edilen veri madenciliği çıktısının pratik iş değerine dönüştüğü noktadır. Model, bir uygulama veya hizmet olarak kurumun operasyonel süreçlerine entegre edilir.

SEMMA (Sample, Explore, Mod ify, Model, Assess) SAS tarafından geliştirilmiş, CRISP-DM'e benzer bir süreç modeli.

Sample: Veriden bir örnekleme yapmak.

Explore: Veri keşfi (ön analiz).

Modify: Veri dönüştürme, özellik mühendisliği.

Model: Uygun algoritmaların denenmesi.

Assess: Performans değerlendirmesi.