Makine Öğrenmesinde Temel Kavramlar

Makine Öğrenmesinde Temel Kavramlar

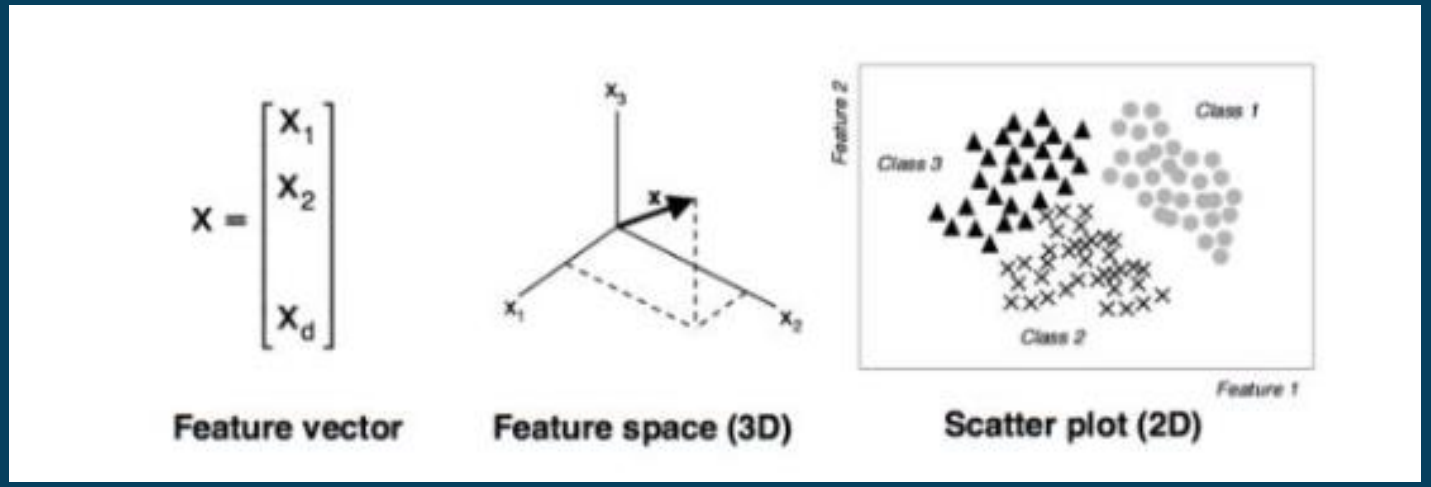
- Makine öğrenmesi, bilgisayarların deneyimlerden öğrenerek ve verilere dayanarak görevleri gerçekleştirmesini sağlayan bir bilgisayar bilimi dalıdır. Bu alanda araştırmacılar, biyolojik öğrenme süreçlerini anlamak için de çalışmaktadır. Özellikle yapay zekanın gelişmesi, robot teknolojileri ve sürücüsüz araçlar gibi alanlarda büyük ilerlemelere yol açmıştır.
- Makine öğrenmesi, genellikle veri setlerinden öğrenen ve otomatik olarak kararlar alabilen algoritmaların ve matematiksel modellerin oluşturulmasını içerir. Bu alanda temel kavramlar arasında tahmin yapma, karar verme, veri analizi, örüntü tanıma ve sınıflandırma gibi işlemler bulunur.
- Makine öğreniminin omurgasını oluşturan disiplinler arasında istatistiksel hesaplama, matematiksel optimizasyon, veri madenciliği ve bilgisayar sistemleri yer alır. Makine öğrenimi, yapay zekanın alt kümesi olarak kabul edilir ve genellikle büyük veri setleri ve yüksek işleme gücü gerektiren karmaşık problemlerin çözümünde kullanılır.

Öznitelik (Özellik) Vektörleri :

- Makine öğrenmesinde, özellik vektörleri, bir nesnenin özellikler adı verilen sayısal veya sembolik özelliklerini matematiksel olarak, kolayca analiz edilebilir bir şekilde temsil etmek için kullanılır. Makine öğreniminin ve kalıp işlemenin birçok farklı alanı için önemlidirler. Makine öğrenimi algoritmaları, algoritmaların işleme ve istatistiksel analiz yapabilmesi için tipik olarak nesnelerin sayısal bir temsilini gerektirir. Özellik vektörleri, doğrusal regresyon gibi istatistiksel prosedürlerde kullanılan açıklayıcı değişkenlerin vektörlerinin eşdeğeridir
- Aşina olabileceğiniz bir özellik vektörüne örnek RGB (kırmızı-yeşil-mavi) renk açıklamalarıdır. Bir renk, içinde ne kadar kırmızı, mavi ve yeşil olduğu ile tanımlanabilir. Bunun için bir özellik vektörü, renk = [R, G, B] olacaktır.
- Bir vektör, tek sütunlu ancak çok satırlı bir matris gibi, genellikle uzamsal olarak temsil edilebilen bir sayı dizisidir. Bir özellik, bir nesnenin bir yönünün sayısal veya sembolik bir özelliğidir. Özellik vektörü, bir nesne hakkında birden çok öğe içeren bir vektördür. Nesneler için özellik vektörlerini bir araya getirmek bir özellik alanı oluşturabilir.

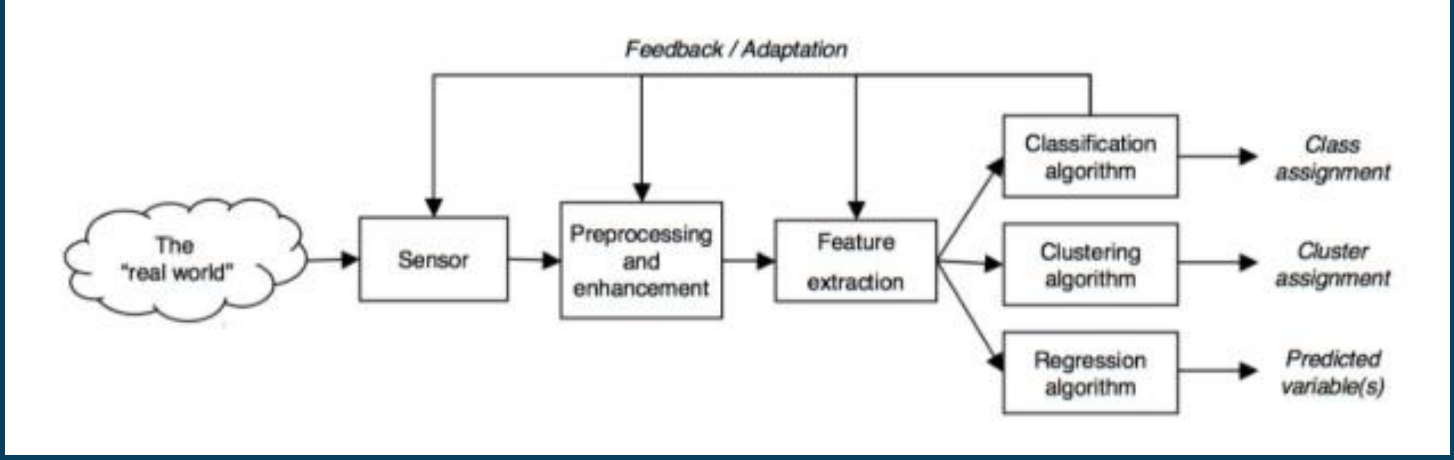
Öznitelik (Özellik) Vektörleri :

- Özellikler, bir bütün olarak, yalnızca bir piksel veya bütün bir görüntüyü temsil edebilir. Ayrıntı düzeyi, bir kişinin nesne hakkında ne öğrenmeye veya temsil etmeye çalıştığına bağlıdır. 3 boyutlu bir şekli, yüksekliğini, genişliğini, derinliğini vb. Gösteren bir özellik vektörüyle tanımlayabilirsiniz.



Öznitelik (Özellik) Vektörleri :

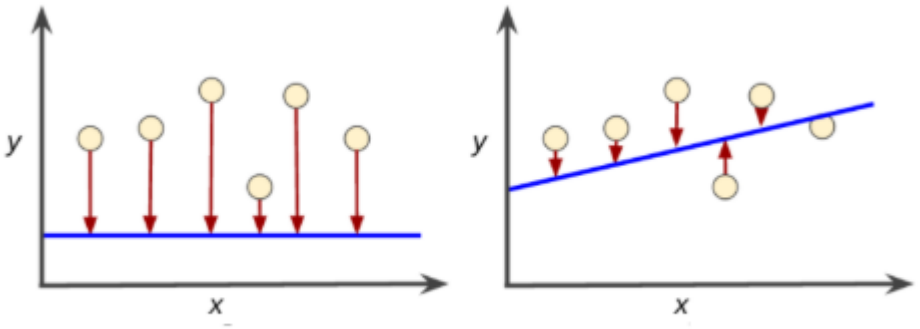
- Özellik vektörleri, nesneleri sayısal bir şekilde temsil etmek için makine öğreniminde yaygın olarak kullanılır. Analiz için idealdirler çünkü öznitelik vektörlerini karşılaştırmak için birçok teknik bulunmaktadır. Örneğin, iki nesnenin özellik vektörlerini karşılaştırmanın basit bir yolu Öklid mesafesini kullanmaktır. Görüntü işlemede, özellikler gradyan büyüklüğü, renk, gri tonlama yoğunluğu, kenarlar, alanlar ve daha fazlası olabilir. Konuşma tanımında, özellikler ses uzunlukları, gürültü seviyesi, gürültü oranları gibi özelliklerdir. İstenmeyen e-posta filtreleme gibi uygulamalarda, özellikler IP konumu, metin yapısı, belirli kelimelerin sıklığı veya belirli e-posta başlıkları olabilir. Özellik vektörleri, makine öğrenmesinde sınıflandırma problemlerinde, yapay sinir ağlarında ve k en yakın komşu algoritmalarında kullanılır.
- Örüntü tanıma süreçlerinde, özellik vektörleri, veri toplama ve verileri anlamlandırma arasında kullanılan araçlardır:



Eğitim Modelleri :

- "Eğitim Model Seçimi", veri setlerini tanımlamak için kullanılan farklı matematiksel modeller arasından seçme işlemidir. Bu seçim istatistik, makine öğrenimi ve veri madenciliği gibi alanlarda uygulanır.
- Makine öğrenmesi modelleri iyi performans gösterebilmek için genellikle çok fazla veriye ihtiyaç duyarlar. Bir makine öğrenmesi modelini eğitirken, temsili veri örneklerini toplamak gerekir. Bu veriler, bir metin topluluğundan, bir resim koleksiyonundan veya bir hizmetin kullanıcılarından elde edilen verilere kadar çeşitlilik gösterebilir.
- Modelin eğitilmesi, tüm ağırlıkların ve etiketli örneklerin değerlerini belirlemek için iyi değerleri öğrenmek anlamına gelir. Denetimli öğrenmede, bir makine öğrenimi algoritması birçok örneği inceleyerek ve kaybı en aza indiren bir model bulmaya çalışır; bu sürece ampirik risk minimizasyonu denir.
- Makine öğrenmesinde "kayıp", kötü bir tahminin cezasını ifade eder. Kayıp, modelin tahmininin ne kadar kötü olduğunu gösteren bir sayıdır. İdeal durumda, modelin tahmini mükemmel olduğunda kayıp sıfırdır; aksi takdirde kayıp daha büyüktür. Bir modeli eğitmenin amacı, tüm örneklerde ortalama olarak düşük kayıplı bir eşik değer bulmaktır.

- Şekilde okların uzunluğu kaybı temsil eder. Mavi çizgiler tahminleri temsil eder



Şekil. Sol modelde yüksek kayıp; doğru modelde düşük kayıp.

Şekil. Sol modelde yüksek kayıp; doğru modelde düşük kayıp.

Eğitim Modelleri :

- Toplu öğrenme: Belirli bir hesaplama programını çözmek için sınıflandırıcılar veya uzmanlar gibi birden çok model stratejik olarak oluşturulur ve birleştirilir. Bu süreç toplu öğrenme olarak bilinir. Toplu öğrenme, bir modelin sınıflandırmasını, tahminini, işlev yaklaşımını geliştirmek için kullanılır. Topluluk öğrenme, daha doğru ve birbirinden bağımsız bileşen sınıflandırıcılar oluşturduğunuzda kullanılır.
- Toplu yöntemler: Topluluk yöntemlerinin iki paradigması şunlardır: Sıralı topluluk yöntemleri Paralel topluluk yöntemleri Bir topluluk yönteminin genel ilkesi, tek bir modele göre sağlamlığı artırmak için belirli bir öğrenme algoritması ile oluşturulmuş birkaç modelin tahminlerini birleştirmektir. Torbalama, istikrarsız tahmin veya sınıflandırma şemalarını iyileştirmek için topluluk içinde bir yöntemdir. Arttırma yöntemi, kombine modelin yanlılığını azaltmak için sırayla kullanılır. Artırma ve Torbalama, varyans terimini azaltarak hataları azaltabilir. Bir öğrenme algoritmasının beklenen hatası, önyargı ve varyansa ayrıştırılabilir. Bir önyargı terimi, öğrenme algoritması tarafından üretilen ortalama sınıflandırıcının hedef işlevle ne kadar yakından eşleştiğini ölçer. Varyans terimi, öğrenme algoritmasının tahmininin farklı eğitim setleri için ne kadar dalgalandığını ölçer. Toplulukta Artımlı Öğrenme algoritması, bir algoritmanın, sınıflandırıcı halihazırda mevcut olan veri kümesinden oluşturulduktan sonra mevcut olabilecek yeni verilerden öğrenme yeteneğidir.

Entropi (Bilgi Kazancı):

- Rassal bir değişkenin belirsizlik ölçütü olarak bilinen Entropi, bir süreç için tüm örnekler tarafından içerilen enformasyonun beklenen değeridir. Enformasyon ise rassal bir olayın gerçeklesmesine ilişkin bir bilgi ölçütüdür. Eşit olasılıklı durumlar yüksek belirsizliği temsil eder. Shannon'a göre bir sistemdeki durum değiştiğinde entropideki değişim kazanılan enformasyonu tanımlar. Buna göre maksimum belirsizlik durumundaki değişim muhtemelen maksimum enformasyonu sağlayacaktır. Shannon bilgiyi bitlerle temsil ettiği için logartimayı iki tabanında kullanımıştır

$$I(x) = \log_2 \frac{1}{P(x)} = -\log_2 P(x)$$

- Shannon'a göre entropi, iletilen bir mesajın tasıdığı enformasyonun beklenen değeridir. Shannon Entropisi (H) adıyla anılan terim, tüm xi durumlarına ait P(xi) olasılıklarına bağlı bir değerdir.

$$H(X) = E(I(X)) = \sum_{1 \le i \le n} P(x_i).I(x_i) = \sum_{i=1}^{n} P(x_i) \log_2 \frac{1}{P(x_i)} = -\sum_{i=1}^{n} P_i \log_2 P_i$$

$$Log_2(P) = \frac{10}{3} Log_{10}(P)$$

$$H(X) = -\frac{10}{3} \sum_{i=1}^{n} P_i \, Log_{10} P_i$$

Entropi (Bilgi Kazancı):

- **Örnek**: Makine öğrenmesi algoritmasının olasılık hesaplanması sonucunda karar vermesi gerekmektedir. İki durum söz konusudur. Birinci durumun olma olasılığı, P1=0.6, olmama olasığı, P2=0.4 ise entropisini hesaplayınız. $Log2\ (0.6) = -0.743$ $Log2\ (0.4) = -4/3$ Entropi, H(x)=0.6*0.743+0.4*4/3=0.979

Kayıp Veri :

- Eğer veride bazı örneklerin bazı özellikleri kayıpsa izlenecek iki yol vardır:
- Kayıp özelliklere sahip örnekler veriden tamamen çıkartılır.
- Kayıp verilerle çalısabilecek sekilde algoritma düzenlenir.
- Eğer kayıplı örneklerin sayısı birinci seçenek uygulanamayacak kadar çoksa ikinci seçenek uygulanmalıdır. Kayıp bilgiye sahip özellik vektörü için kazanç hesaplanırken kayıplı örnekler hariç tutularak bilgi kazancı normal sekilde hesaplanır ve daha sonra F katsayısıyla çarpılır. F, kayıpsız verinin tamamına oranıdır. IG(X) = F.(H(X) − H(V, X)). Kayıp bilgiye sahip özellik vektörü içinde en sık tekrarlanan değerin kayıp bilgi yerine yazılması da önerilen yöntemlerdendir. Eksik değerler ortaya çıktığında veri noktalarını kolayca atamayız. Sınıflandırılacak bir test noktası da eksik değişkenlere sahip olabilir. Sınıflandırma ağaçlarının, eksik değerlerini tamamlamanın güzel bir yolu vardır. Sınıflandırma ağaçları, bir yedek ayrım bularak sorunu çözer. Başka bir değişkene dayalı başka bir ayrım bulmak için, sınıflandırma ağaçları diğer tüm değişkenleri kullanarak tüm bölünmelere bakar ve optimum bölünmeye en çok benzeyen eğitim veri noktaları bölümünü veren birini arar. En iyi bölünmenin sonucunu tahmin etmeye çalışırlar.

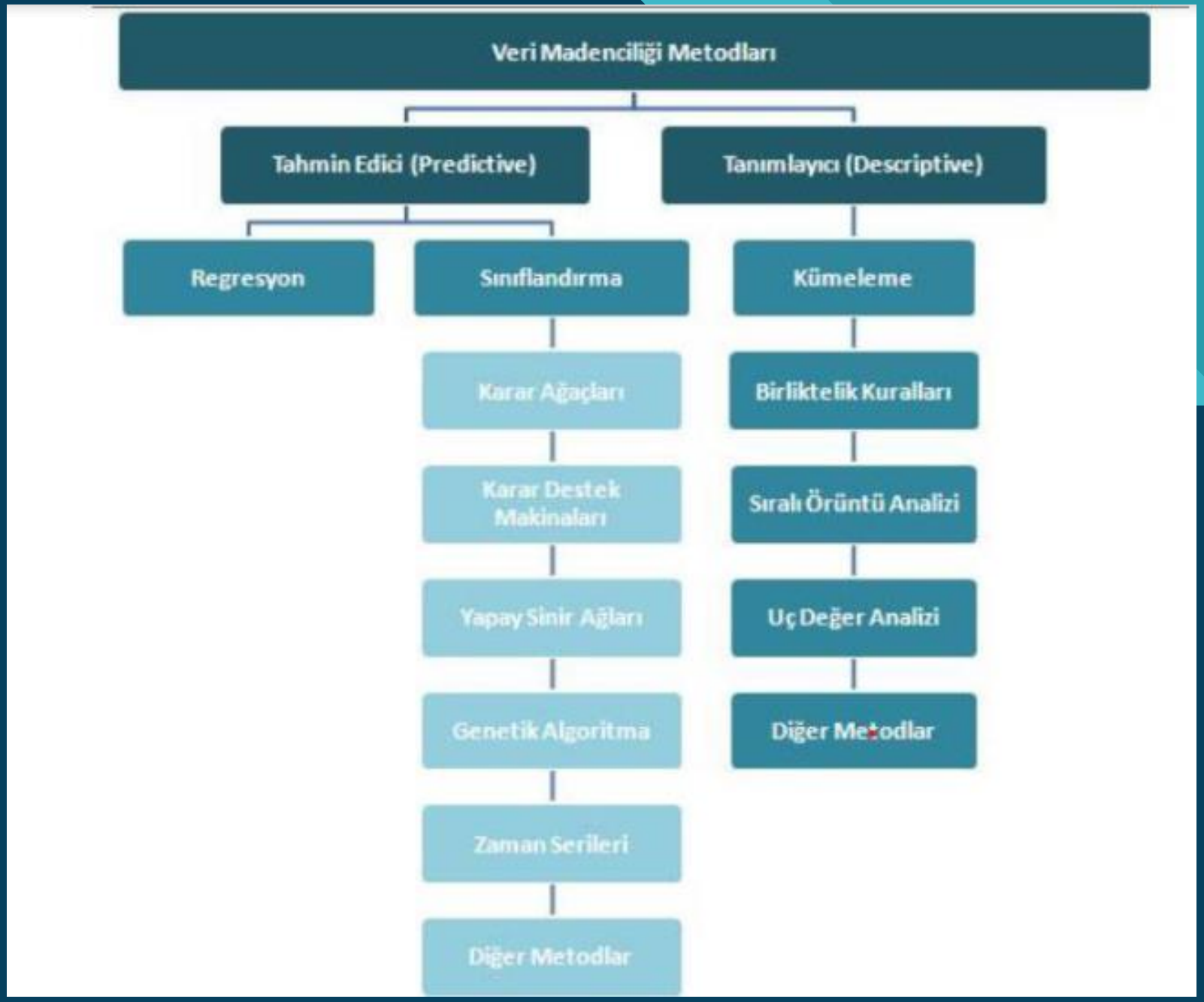
Veri Madenciliği :

- Veri madenciliği, denetimsiz öğrenim yoluyla keşifsel veri analizine odaklanan bir alan olarak tanımlanabilir. Bu alanda, verilerdeki bilinmeyen özelliklerin keşfedilmesi amaçlanır ve genellikle bilgi keşfi analizinin bir adımı olarak görülür.
- Makine öğrenimi ise, bilgisayarlara açıkça programlanmadan öğrenme yeteneği veren algoritmaların incelenmesi, tasarımı ve geliştirilmesiyle ilgilidir. Veri madenciliği, yapılandırılmamış verilerden bilgiyi veya bilinmeyen ilginç kalıpları çıkarmaya çalışırken, makine öğrenimi öngörü yeteneği üzerine odaklanır.
- Bu iki alan genellikle aynı yöntemleri kullanır ve önemli ölçüde örtüşür. Ancak, makine öğrenimi öğrenme verilerinden öğrenilen öngörüye odaklanırken, veri madenciliği verilerde bilinmeyen özelliklerin keşfine odaklanır. Yani, veri madenciliği genellikle veritabanlarında bilgi keşfinin ilk adımı olarak görülür.
- Veri madenciliği, büyük miktardaki veri setlerinden desenlerin, ilişkilerin ve önemli bilgilerin keşfedilmesini sağlar. Kötüye kullanım tespiti ve anormallik tespiti gibi yöntemlerin yanı sıra sınıflandırma yöntemleri ve demetleme yöntemleri de kullanılır. Ayrıca, farklı yöntemlerin birleştirilmesiyle oluşturulan hibrit yöntemler de yaygın olarak kullanılır.

Veri Tabanı Yönetimi:

- Veri tabanları, elde edilen verilerin saklandığı alanlardır ve ilişkili verilere erişimi sağlayarak verinin yönetimini kolaylaştıran yazılım programları içerir. Makine öğrenimi projelerinde veri tabanı yönetim sistemi önemli bir bileşendir. Bu sistemler, verileri sıralar ve anlamlı içgörüler elde etmek için kullanılabilir. Stack Overflow'un 2019 anketine göre, Redis en çok sevilen veri tabanı iken MongoDB en çok aranan veri tabanıdır.
- Veri tabanları, kullanım amaçlarına göre farklı isimler alır:

1. **İlişkisel Veri Tabanları**: Bu tablolar, farklı isimler alan tablolardan oluşur. Her tabloda, her bir kaydın özelliklerinin değerlerini tutan alanlar ve her kayda ait bir tekil anahtar bulunur. Örneğin, bir üniversitenin veri tabanında her bir öğrenci için öğrenci numarası, kayıt yılı, bölüm gibi bilgiler saklanabilir.
2. **İşlemsel Veri Tabanları**: Her bir kaydın bir işlem olduğu varsayılır. Örneğin, bir marketin veri tabanında her bir satış işlemi kaydedilir ve bu veri tabanından belirli bir üründen bugün kaç tane satıldığı gibi bilgiler elde edilebilir.
3. **Zaman Serisi Veri Tabanları**: Bu tablolar, düzenli zaman aralıkları ile elde edilen verilerin saklandığı alanlardır. Örneğin, borsa verileri, stok kontrolleri sonucu alınan veriler, sıcaklık ölçümleri gibi veriler burada saklanabilir.


Veri Madenciliği Metodları
Tahmin Edici (Predictive)
Tanımlayıcı (Descriptive)
Regresyon
Sınıflandırma
Kümeleme
Karar Ağaçları
Karar Destek Makinaları
Yapay Sinir Ağları
Genetik Algoritma
Zaman Serileri
Diğer Metodlar
Birliktelik Kuralları
Sıralı Örüntü Analizi
Uç Değer Analizi
Diğer Metodlar