2.7. Veri Tabanı Yönetimi

Veri tabanları: Elde edilen verilerin tutulduğu alanlardır. Bir veri tabanı sistemi, birbiri ile ilişkili verilerin birikimini içeren, veriye erişimi sağlayarak veriyi yönetmeye yardımcı olan yazılım programları kümesidir.

Makine öğrenimi projelerindeki en kritik bileşenlerden biri veritabanı yönetim sistemidir. Bu sistemin yardımıyla çok sayıda veri sıralanabilir ve bunlardan anlamlı içgörüler elde edilebilir.

2019 Stack Overflow Survey raporuna göre Redis en çok sevilen veritabanı, MongoDB ise en çok aranan veritabanı.

Veri tabanları kullanım amaçlarına göre farklı isimler alır:

İlişkisel veritabanları, her biri farklı isimler alan tablolardan oluşur. Her tabloda her bir kaydın özelliklerinin değerlerini tutan alanlar ve her kayda ait bir tekil anahtar bulunur. Bir üniversitenin veritabanını ilişkisel veri tabanına örnek olarak verebiliriz. Zira her bir kişi için ayırt edici bir öğrenci numarası, hangi yılda kayıt yaptırdığı, hangi bölümde okuduğu gibi alanlar ile öğrenciye ait bilgiler saklanır. Buradan çeşitli sorgular ile hangi bölümde kaç öğrencinin okuduğu, geçtiğimiz yıl kaç kişinin belli bir bölüme kayıt yaptırdığı gibi soruların cevapları bulunabilir.

İşlemsel veritabanında her bir kaydın bir işlem olduğu varsayılır. Bir marketin veri tabanını düşünecek olursak, her an bir satış yapıldığını ve her bir satışın işlemsel veri tabanında bir kayıt olarak göründüğü varsayılabilir. Bu veritabanından, bugün, ilgilenilen üründen kaç tane satıldığı sorusunun cevabına ulaşılabilir.

Zaman serisi veritabanı düzenli zaman aralıkları ile elde edilmiş (yıllık, haftalık, günlük) verilerin tutulduğu alanlardır. Örnek olarak borsa verilerinin, stok kontrolleri sonucu alınan verilerin, sıcaklık ölçümlerinden elde edilen verilerin depolanması gösterilebilir.

Veri Ambarları:

Veri Ambarları: “Veri ambarları, tüm operasyonel işlemlerin en alt düzeydeki verilerine kadar inebilen, etkili analiz yapılabilmesi için özel olarak modellenen ve tarihsel derinliği olan veri depolama sistematiği olarak tanımlanabilir.” Günlük işlemler sonucu, farklı kaynaklardan toplanan veriler, temizleme dönüştürme, birleştirme gibi işlemlerden geçirilerek, daha önce inşa edilmiş veri ambarının yapısına uygun hale getirilerek veri ambarına aktarılır. Veri ambarları, üzerinde, verilerin yüklenmesi ve erişimi dışında herhangi bir işlem yapılmasına izin vermez. Veri ambarları belirli aralıklar ile güncellenirler. Mimari açıdan veri ambarları üç farklı şekilde olabilir. İlki, işletmelerin farklı kaynaklardan (işletmenin kendi işlemsel veritabanı sistemleri ve dış kaynaklar dâhil olmak üzere) aldıkları tüm verilerin tutulduğu “işletme ambarları”, ikincisi veri üzerinde çalışma yaparak karar alan kişiler için belirli

kurallara göre oluşturulmuş “veri pazarları” , sonuncusu ise işlemsel veri tabanlarının görsel hali olan “ görsel ambarlar” “dır.

Veri madenciliğinde kullanılan modeller:

Tahmin Edici Modeller : Tahmin edici modellerde, sonuçları bilinen verilerden hareket edilerek bir model geliştirilmesi ve kurulan bu modelden yararlanılarak sonuçları bilinmeyen veri kümeleri için sonuç tahmin edilmesi amaçlanmaktadır.

Tanımlayıcı Modeller : Tanımlayıcı modellerde, veri kümesinde bulunan gizli örüntülerin (olayların ve nesnelerin ortaya çıkardığı davranış değişikliklerinin desenleri) tanımlanması amaçlanmaktadır.

Veri madenciliği süreci dört aşama ile tanımlanabilir.

• İlk aşamada problem tanımlanarak veri kaynakları değerlendirilir.

• İkinci aşamada veriler kullanıma uygun hale getirilmek için hazırlanır.

• Arkasından model kurulur ve

• nihai aşamada model değerlendirilerek kullanıma hazır hale getirilir.

Problemin Tanımlanması:

Amaç, işletme problemine verileri kullanarak çözüm getirmek olduğundan, ilk olarak ihtiyaç duyulan şey tam olarak tanımlanmalıdır. Bu problem, işletmenin ayrılmakta olan müşterisinin belirli özelliklerini tanımlayarak ona uygun davranmak olabildiği gibi, kendi kaynaklarını optimum kullanabilmek için yapacağı bir planlamada gelecek dönemdeki harcamalarını tahmin etmek şeklinde de olabilir. “Bu adımda ihtiyaç duyulan şeyin tanımlanması için cevaplanması gereken sorular neyin otomatize edilmeye değer olduğu ve neyin insan içeren süreçlere bırakılması gerektiği, amacın ne olduğu ve hangi performans kriterlerinin daha önemli olduğu, sürecin sonucunda elde edilecek çıktının keşif, sınıflandırma, özetleme gibi şeyler için kullanılıp kullanılmayacağı olabilir.” Problemin tanımlanması durumunda ihtiyaç duyulan iş modelinin kalıbı da belirlenmiş olur.

Verilerin Hazırlanması:

Modelin kurulması için gerekli bilgilerin hazırlandığı aşamadır. Öncelikle toplam, maksimum, minimum değer gibi dağılım ölçüleri; aritmetik ortalama, ağırlıklı ortalama gibi cebirsel ölçüler veya serpilme,dağılma diyagramı gibi grafiksel öğeler kullanılarak verilerin durumu hakkında bilgi edinilir. Verilerde eksik, hatalı, gürültülü bilgi olup olmadığı bu şekilde kontrol edilmiş olur. Eksik değerlerde kaydı dikkate almama, global sabit ile eksik değerleri doldurma, eksik değere o değişkenin ortalama değerini verme, gürültülü değerlerde regresyon ile belirli fonksiyonel kalıba sokma gibi yöntemler ile verilerdeki sıkıntı giderilebilir.

Farklı kaynaklardan gelen, aynı değişkene ait verilerin tiplerinde, alan isimlerinde uyuşmazlık olması halinde gerekli değişikliklere gidilerek tüm verileri bir arada tutabilecek yapı oluşturulmalıdır.

Bazı modellerin gereksinimlerini göz önünde bulundurmak açısından farklı dönüşümlere gitmek de veri hazırlanırken dikkate alınması gereken hususlardan olabilir. Örneğin bazı değişkenlerdeki değerler çok yüksek ise, bu değerleri normalize ederek, uzaklıklar ile çalışan

kümeleme algoritmalarının öğrenme fazını hızlandırarak modelin oluşturulma aşaması için kolaylık sağlanmalıdır.

Değişken sayısının çok yüksek olduğu, hangi değişkenlerin öneminin daha yüksek olduğuna karar verilemediği durumlarda faktör analizi, temel bileşenler analizi gibi yöntemler kullanılarak boyut indirgemeleri yapılmalıdır. Zira bu indirgemeler modele girecek değişken sayısını azaltarak modeli gereksiz bilgilerden ayıklar ve daha sağlıklı bir sonucun çıkmasına zemin hazırlarlar.

Gerektiğinde kategorik değişkenlerde kategori aralıklarını genişleterek kategori sayısını azaltma veya sürekli bir değişkeni kategorik hale getirmek de verinin hazırlanmasında dikkat edilmesi gereken unsurlardandır. Çok kategorili değişkenler duruma göre modelin çalışma süresini ve sürecin performansını olumsuz etkileyebilmektedir.

Modelin Kurulması:

Modelin kurulması aşamasında birçok model denenerek veriyi en iyi temsil eden model seçilir. Verileri temsil eden en iyi modeli bulabilmek için çok sayıda model kurulmalı, en iyi sonucu alana kadar denemeye devam edilmelidir.

Modelin kuruluşu, amacımızın ne olduğuna, problemimizi ne şekilde çözmek istediğimize ve sonucun ne kadar işimize yarar olacağına göre değişebilir. Örneğin görmek istediğimiz gelecek dönemdeki tahmini ciromuz ise, sürekli bir değişkeni tahmin edeceğimiz doğrusal regresyon modelini; müşterilerimizin pasifleşme eğiliminde olup olmadıkları ise kategorik bir değişkeni tahmin edeceğimiz sınıflandırma modelleri olan karar ağaçlarını, yapay sinir ağını veya kategorik değişkenin olasılığını tahmin edeceğimiz lojistik regresyon modelini, hangi ürünlerimizin diğerlerine oranla daha çok beraber alındığı ise birliktelik analizi, beraber alınan bu ürünlerin hangi sırayla alındığı, nedensellikleri ise sıralı örüntü algoritmaları kullanılabilir. Ayrıca müşterilerimizin sahip oldukları alışveriş özelliklerine göre (gelme sıklıkları, uğradıkları mağazalar, satın aldıkları ürünler vb.) belirli gruplara ayırmak için kümeleme algoritmaları kullanılabilir.

Model kurulurken denetimli veya denetimsiz öğrenmeye göre farklı aşamalar uygulanmaktadır. Örneğin sınıflandırma algoritmaları kullanılırken tüm veri kümesi öğrenme

ve test kümesi olarak ayrılmalı; modelin verilerden öğrenerek oluşturulması öğrenme kümesi, doğruluğunun kontrolü ise test kümesi ile gerçekleştirilmelidir.

Kurulan modellerde birbiri ile ilişkili olan veya anlamsız olan değişkenlerin elenmesine dikkat edilmelidir. Amaç bilgi çıkarımı olduğundan ve birbiri ile ilişkili olan değişkenler bize ekstra bilgi vermediğinden, diğerine göre daha anlamlı olan değişkeni modele katmak faydamıza olacaktır.

Modelin Değerlendirilmesi:

Kurulan modellerin karşılaştırılarak veri kümesini en iyi temsil eden modelin seçildiği aşamadır.

Karşılaştırma için, sınıflayıcının tahmin ettiği sınıfların oranını belirten doğruluk oranı kullanılır. Sınıflayıcının doğruluk oranının görece yüksek olması, diğer modellere göre veri kümesini daha iyi ifade ettiğini gösterebilir. Doğruluğun testi için kullanılan geçerlilik

yöntemleri basit geçerlilik yöntemi, çapraz geçerlilik yöntemi, n-katlı geçerlilik yöntemi olarak sıralanabilir.

Basit geçerlilik yönteminde verilerin bir kısmı test verisi olarak ayrılır, kalan kısım üzerinde modelin öğrenimi gerçekleştirildikten sonra ayrılan kısım üzerinde test işlemi yapılır. “Bir sınıflama modelinde yanlış olarak sınıflanan olay sayısının, tüm olay sayısına bölünmesi ile hata oranı, doğru olarak sınıflanan olay sayısının tüm olay sayısına bölünmesi ile doğruluk oranı hesaplanır.” Çapraz geçerlilik yöntemi daha az sayıda veri kümesine sahip olunduğu durumlarda kullanılabilir. Bu yöntemde veri kümesi rastgele seçilerek iki eşit gruba ayrılır, gruplar sırayla öğrenme ve test kümesi yapılarak elde edilen doğruluk oranlarının ortalaması kullanılır. N-katlı geçerlilik yöntemi de çapraz geçerlilik yöntemi gibi küçük veri kümeleri için kullanılmaktadır. Veri kümesi birden fazla gruba ayrılır, bir tanesi test diğerleri öğrenim için kullanılır. Test kümesi değiştirilerek doğruluk oranı hesaplanır ve elde edilen oranların ortalaması kullanılır.

Risk matrisi geçerlilik yöntemlerini görselleştirmek için kullanılabilen bir araç olabilir. Yeni çıkan bir ürünü piyasaya sürmeden önce belli sayıda kişi ile görüşülerek ürünün tutup tutmayacağı konusunda bir araştırma yapıldığını ve ürün hakkındaki fikirleri iyi ya da kötü olarak sınıflandırmak istediğimizi düşünelim. Sonuçta karşılaştıracağımız sınıflandırma algoritmalarının doğruluğunu aşağıdaki şekilde görselleştirebiliriz.