thinktech
STM Teknolojik Düşünce Merkezi

Veri Madeni: Patlama Riski Hâlâ Var

Verilerin az olduğu ve görsel olarak bile bakıldığında bazı çıkarımların yapılabildiği mutlu günlerde bir Excel dosyası üzerinden her şeyi yönettiğimizi düşünürdük. Büyük ölçüde de böyleydi. Dünyanın farklı yerlerinde bu dönem daha eski de olabilir ancak her endüstri kültüründe verilerin bugüne nazaran çok az olduğu böyle bir dönem mevcut. Bir muhasebe kaydına ait fişlerin özenle dosyalanıp klasörlere kaldırıldığı ve her kayda ait "destekleyici belgenin" mevcut olduğu sistematik o dönemlerden bugüne varlığını sürdürüyor.

Bu dönemler geride kaldı. Artık verilerin görsel olarak kontrol edilip yorumlanabildiği çok az alan mevcut. Bunun birkaç nedeni var: İlki veri toplama aygıtlarının, yani kameralar, sensörler, bilgisayarlar, akıllı telefonlar, akıllı saatler, akıllı gözlükler ve her gün daha fazla internete bağlanan nesnelerin interneti (Internet of Things) ordusunun elemenlarının sürekli veri üretiyor olmasıdır. Üretilen bu verileri depolamak için kullanılacak teknolojiler de aynı hızda ilerlerken maliyetler giderek daha azalıyor. Maliyetlerin azalması da insanların ve kuruluşların daha fazla veri toplama ve tutmasını teşvik eden diğer bir unsur.

Veri çağının tek nedeni teknolojik donanım ve yazılımların artması değil kuşkusuz. Veriyi kullanacak olanların subjektif yargıları da bunda etkili. Ölçülemeyen ve kaydedilmeyen olguların yönetilemeyeceğine dair veri-yoğun bir yönetim trendi de kuruluşları her geçen gün daha fazla ve daha farklı veriyi tutmaya itiyor. Bu haliyle bir veri "büyük veri" sınıfına girebilse de bu verileri sadece hızlıca listeleyebilmek heyecan verici bir sonuç olmayabilir.

Madenciliği yapılacak kadar değerli hale gelen veri, temel istatistiği ile içgörü sağlayan veri değil, verisi tutulan varlığın (birey, kuruluş, doğa) kolayca tespit edilemeyen gizli yapılarını keşfetmeye yönelik çeşit ve miktara sahip olan veridir. Bu noktada miktar için kesin olarak belirlenmiş alt veya üst sınır yoktur. Örneğin, terör saldırılarıyla ilgili tutulmakta olan veri tabanı üzerinden terör örgütünün müteakip saldırısının nerede olabileceğine dair bir modelin kurulabildiği bir veri seti bu tanıma giren bir veridir.

Hangi Veri Nasıl Tutulmalı?

Verilerin tutulmasıyla ilgili diğer bir konuysa günümüzün değerli madeni haline gelen verinin belirli bir stratejiye sadık kalınarak tutulmasıdır. Bu strateji ise kuşkusuz organizasyonların amaçlarıyla uyumlu ve gerekli detaylar düşünülerek oluşturulmalıdır. Söz gelimi, bir hastane yönetiminin yatan hasta servislerindeki yoğunluğu analiz ederek çalışanların izinlerini organize etmek istediğini, bunun yanında yeni yapılması gereken altyapı yatırımlarını planlamak ve diğer kaynaklarını da optimize etmek istediğini varsayalım. Bu durumda hemen bir veri bilimcisinden bu analizleri yapmasını istesin, hatta bu analizlerden yola çıkarak gelecek yıllara dair bütçe planmasının yapılmasına karar verilsin. Bu durumda veri bilimci hasta

yatış kayıtları veri tabanlarının mevcut olduğunu ilk duyduğunda memnun olabilir ancak daha sonra bu verileri incelediğinde büyük ihtimalle tüm servislerin tam kapasite çalıştığına şahit olacak ve şu soruyu soracaktır: Kardiyoloji servisi ile nöroloji servisi her dönemde nasıl tam kapasitededir? Bunun neresi optimize edilecektir? Daha sonra kahramanımız, hasta servislerinde dönemsel yoğunluklar olduğunda ilgili personelin nöroloji servisindeki bir hastayı kardiyoloji servisine aldığını fark eder. Üstelik bu yönlendirmeye dair kayıt tutulmamıştır. Bu durumda ilgili veri projesi çöp olmuştur. Üstelik tutulan onlarca yıllık veriye rağmen. İşte bu noktada, "Elimizdeki verilere bakalım, oradan neler çıkar?" şeklindeki bir yaklaşım yerine "Ne yapmak istiyoruz" ve "Elimizdeki kaynaklara göre hangi verileri toplamamız yerinde olur" yaklaşımı tercih edilmelidir.

İstatistik: Sadece Ortalamadan İbaret Değil

Ortalamalarla ilgili ilginç bir metafor vardır: Kafası fırında ayakları ise derin dondurucuda olan bir adamın ortalama vücut sıcaklığı normal sayılabilir. Ortalama kavramı ile ilgili yanılgıları bu kadar iyi açıklayan çok az metafor vardır. Bu talihsiz adam için fırının 75 derece, derin dondurucunun -25 derece olduğu zaman ortalama olarak 25 derece vücut sıcaklığı söz konusudur. Bu sıcaklık ise oda sıcaklığına tekabül eder. Söz konusu insan vücudu olduğunda bu örnekteki ortalama vücut sıcaklığının anlamsızlığını fark etmek kolaydır ancak söz konusu olan yüzlerce finansal parametreyi içeren bir veri seti ise bu kez ortalama vücut sıcaklığı hakkında yorum yapılan kişi bu talihsiz adam değil firmanın sahibi olabilir.

Veri bilimi ile istatistiğin ilişkisi sandığınızdan da fazladır. Veri biliminde verinin ön işlenmesinden (preprocessing) sonra yapılan ve birçok sorunun cevabını barındıran süreçte istatistik çok büyük önem arz eder. Bu ise ortalama değerlerin de içinde yer aldığı tanımlayıcı (descriptive) istatistikten çok daha fazlasıdır.

Akademik amaçlarla veya veri bilim dışındaki istatistik kullanımı, genellikle tanımlayıcı boyutuyla tanınır. Bir veri seti hakkında genel bir kanaate sahip olabilmek için tanımlayıcı istatistiklere bakılır ve süreç biter. Böyle bir durumda karar verici istatistiği kullanan kişidir. Veri biliminde ise istatistiksel analizlerin sonuçlarını karar vericiden önce büyük ihtimalle kodlanan algoritma değerlendirmektedir. Başka bir deyimle algoritma elde edilen bir istatistiksel analiz sonucuna göre veri seti içerisindeki gözlemler üzerine karar vermektedir.

Buna iyi bir örnek olarak uç değerleri yakalayan "outlier detection" metotları gösterilebilir. Bu metotlar da neticede veri seti içerisindeki uç değerleri tanımlamaktadır. Bu noktada bir fark yoktur. Ancak "outlier detection" sonrasında ortaya çıkan sonuçların otomatik olarak veri setinden silinmesi sürecinde artık algoritma daha önceden programlanmış olarak karar verici adına karar vermektedir. Peki neden? Çünkü sürekli güncellenen devasa ve çoğu kez anlaşılmaz veri setleri içerisinde veriye ne yapılacağına dair kararların her birine "insan" karar vericilerin zamanı ve motivasyonu yetmeyecektir.