Merve Tatlıdil Follow
May 15, 2020 · 6 min read

Veri Türleri ve İstatistiğe Giriş

Herkese merhaba, bu yazımızda veri türlerinden bahsedeceğim. Veri bilimi projelerinde istatistikten de çok yararlandığımız için istatistiğe giriş yapacağız.

İlk yazım olan veri analitiği yazısında veri her işin başı demiştim. Hayatımızın her alanında veri ile yaşadığımızı, sürekli yeni veriler üretildiğini söylemiştim.Tıpkı **petrolün** birçok ekonomik ürün için bir hammadde olması gibi veri de anlamlandırılmış bilgi üretme ve analiz süreçlerinin **hammaddesi** durumundadır. Veri, bilginin yapılandırmayla kayıt altına alınıp, kolay analiz edilebilmesi için bir araya getirilmesidir. Bir veya birden fazla bilgiden oluşan bir kümedir. Veri genellikle araştırma, gözlem, deney, sayım, ölçüm yoluyla elde edilir. İsim, telefon numarası, bir çarpma işlemi sonucu ya da sınıfın yaş ortalaması birer veridir.

Ham veri etkin şekilde bilgi üretme ve analiz için önemli bir hammaddedir. Örneğin : **anketler** ile oluşturulan seçim verileri, bir **oylama** yapıldığında seçim sonucu verileri, bir şey **satın alındığında** (çevrim içi satış kayıtları vb.). gibi. Veri ayrıca **cep telefonları, İnternet, uydu** (GPS verisi gibi) ve birçok farklı teknolojiler tarafından da oluşturulabiliyor.

Gündelik hayatımızda veriyi sıklıkla tablolarda düzenlenmiş buluruz. Veri seti, satırlarında **gözlem birimleri** , sütunlarında ise **değişkenler (öznitelik, attribute, feature da denebilir)** bulunan iki boyutlu bir matristir. Satır ve sütunların kesişim bölgelerine **hücre(cell)** denir. Her bir hücreye, gözlemlerin değerleri yani yapılan ölçmenin sonuçları sayı veya sembol olarak girilir.Bir hücrede herhangi bir sayı veya sembol olmaması, amaçlanan gözlemin bulunmadığı anlamına gelir. Veri setinde yer alan bu tür boşluklar **kayıp veri (missing value)** olarak adlandırılır. Veri setini analiz ederek ondan yeni bilgi -görsel çalışmalar üretmek; karar alma, politika üretme süreci için önemlidir.

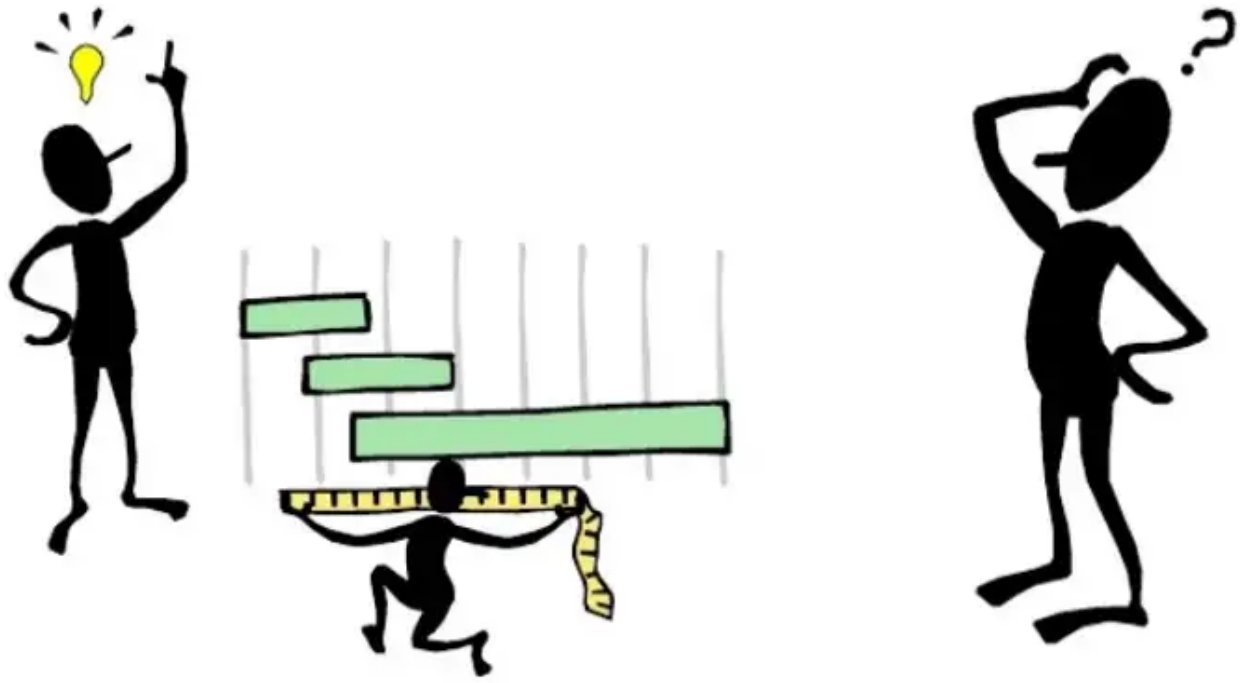
Veri Türleri

Yukarıda tenis topları görüyorsunuz. Tenis topunu gördüğümüzde aklımıza gelen tenis sporudur. Tenis : yani bir spor kategorisi. Bu detay topu bir sınıflandırmaya koymamızda yardımcı oluyor. Fakat bundan daha fazlası var fotoğrafta. Topların "açık yeşil" renkte olduğunu da söyleyebiliriz. Topların hepsinin bir ölçüsü de var, belirli bir sayıdalar ve fiyatları da var. Önemsiz objelerde bile onlarla ilişkili çok sayıda veri bulunur. Yukarıdaki örnekte, kategorik ve numerik farklı veri türleri olduğunu görebiliriz.

Kategorik(Kalitatif, Nitel) Veri : Ölçülemeyen, üzerinden sayısal işlem yapılamayan nesnel verilerdir. Cinsiyet, saç rengi vb. nitel özellikleri belirten veriler.Nitel veriyle veri kümesindeki gözlemlere ilişkin **sıfatlar ya da durumlar** tespit edilir. Eğitim durumu, ev sahibi olup olmama gibi...Kategorik veriler **nominal** ve **ordinal** olmak üzere iki gruba ayrılır :

Sınıflanabilen (Nominal): Bir nominal niteliği ; **gözlemlerin adları, yada sembolleri** olarak düşünebiliriz. Bu değerler **bir kategoriyi veya durumu** temsil eder ve bu nedenle kategorik özellik olarak adlandırılırlar.Sınıfların aralarında hiyerarşik bir yapı yoktur. Araba markası, renk,meslek,il,cinsiyet gibi...

Sıralanabilen (Ordinal): Sıralayıcı nitel veriler aralarında **anlamlı bir sıra veya derecelendirmeye sahip değerler** içerir. Değerleri arasında sıralı bir ilişki bulunur. Akademik unvan, rütbe, öğrenim durumu,sınav notları gibi.



Numerik(Kantitatif,Nicel) Veri : Ölçülebilen, üzerinde aritmetik işlemler yapılabilen veri tipidir. Boy, kilo, hizmet süreleri, hava sıcaklığı, kandaki bir bileşenin miktarı... Nicel veriler **kesikli ve sürekli** olmak üzere **iki şekilde** incelenmektedir :

Discrete (Kesikli) Sayısal Veriler: Sonlu veya sayılabilir bir şekilde sonsuz değerler dizisine sahiptir. **Gözlemlere ait özelliklerin tam sayılarla** ifade edildiği veri kümeleridir. 0, 1, 2, 3 vb. tam sayı değerler...

Continious (Sürekli) Sayısal Veriler: Sonsuz sayıda değer alabilen ve kesirli değerleri de içeren veri setleridir.

7.00, 99.846, 200.1365 vb. değerler sürekli verilere örnek verilebilir

Aralıklı (Interval): Hem sırayı hem de farkı gösterir. Eşit aralıkların eşit mesafeleri temsil ettiği bir ölçek türüdür. Sayılar arasında oransal bir ilişki yoktur.

Sıcaklık bunun için iyi bir örnektir. 25°C ile 30°C arasındaki fark, 40 °C ile 45 °C arasındaki farkla aynıdır.

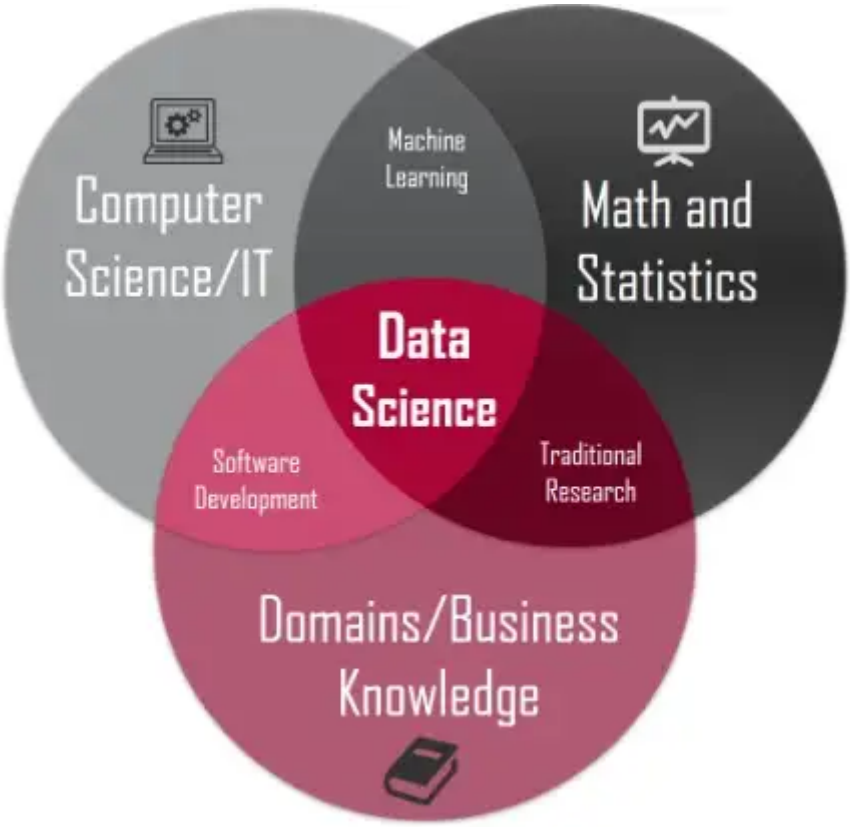
Aralık değişkenleri ile ilgili en önemli nokta, mutlak bir sıfır noktası bulunmamasıdır.Örneğin, 0°C sıcaklığı, havada hiç sıcaklık olmadığı anlamına gelmez!

Oranlı (Ratio): Sırayı, mesafeyi ve anlamlı bir mutlak sıfır değerini gösterir. Bir oran değişkeni 0 değerine sahip olduğunda, değişken tarafından ölçülen miktarın hiçbirinin mevcut olmadığını anlarız.

Örnek olarak, yaş bir oran değişkenidir: 0 yaşında biri henüz doğmamıştır ve 20 yaşında biri 10 yaşında birinin iki katı yaşamıştır.

İstatistiğe Giriş

Veri bilimi projeleri pekçok alan bilgisi ile gerçekleştirilir.Projeyi yapmak için iş bilgisi gereklidir.Çalışacağımız veri seti üzerinde yorum yapabilmek için konu hakkında bilgi sahibi olmamız gereklidir.Aynı zamanda veri setini değerlendirmek için makine öğrenmesi algoritmaları kullanıyoruz bunun için kodlama becerilerimizin olması gerekir.Veri setlerini doğru değerlendirebilmek için doğru algoritmayı seçmemiz gerekir.Bunun için de kişinin algoritmanın nasıl çalıştığını, yazdığı kodla hangi işlemleri yaptığını bilmesi gerekir.Matematik ve istatistik bilgisi varsa algoritmanın arka planını anlamakta zorlanmaz.Durumu daha iyi görmek açısından ilk yazımda paylaştığım aşağıdaki resmi tekrar eklemek istedim.



İstatistik ; verinin ya da bilginin tanımlanması, düzenlenmesi ve yorumlanması için gerekli araçları ve yöntemleri tarif eden bir bilimdir. Verinin belirli özelliklerini

tanımlamak ve onları ortaya çıkaracak sayısal değerleri ifade eder.İstatistik alanında kullanılan “Veri Ölçüleri” yöntemlerini inceleyemeye başlayalım

İstatistik işlemlerinde verileri tanımlamak için kullanılan 3 genel istatistik yöntemi vardır:

· **Merkezi Eğilim Ölçüleri** (Measures of Central Tendency)

· **Dağılım Ölçüleri** (Measures of Spread)

· **Değişkenlik Ölçüleri** (Measures of Variability)

Merkezi Eğilim Ölçüleri

Merkezî eğilim ölçüleri ; bir veri grubunun dağılımında, verilerin etrafında yığılma eğilimi gösterdikleri ve veri grubunu “özetleyen” değerlerdir.

1-Ortalama (Mean ,Average)

Gözlenen değerlerin tümünün toplanıp, gözlem sayısına bölündüğünde elde edilen değere ortalama denir. Aynı zamanda **değişkenin beklenen değeri** olarak da tanımlanır.

-Verideki aşırı değerlere karşı hassastır.

-Normalin üstünde çok büyük bir değeri küçük verilerin olduğu bir ortalamaya alırsanız ortalama üzerindeki etkisi çok belirgin olacaktır.

Elimizdeki sayılar 13,10,15,12,17,13 olsun.Ortalama bulmak için önce bu sayıların toplamını bulur sonra da kaç tane değer varsa ona böleriz.Yani :

(13+10+15+12+17+13) / 6 = 13.33 = ortalama deriz.

2-Medyan (Median)

Veriler küçükten büyüğe dizildiğinde ortadaki değerdir. Bu nedenle **ortanca** olarak da ifade edilmektedir.

-Eğer tek sayıda değer varsa ortadaki değer medyandır.

10, 12 , 13, 15, 17 sayılarımız olsun bunların medyan değeri : 13 olacaktır.

-Eğer çift sayıda değer varsa ortadaki iki değerin ortalamasıdır.

10, 12, 13, 15,17,19 sayılarımız olsun bunların medyan değeri : (13 + 15) / 2 = 14 olacaktır.

*Normalin üstünde çok büyük değerlerden etkilenmediği için ortalama gelir veya maaş hesaplamalarında, aritmetik ortalamaya göre daha doğru sonuçlar verebilmektedir.

3-Mod

Bir sayı dizisinde en çok tekrarlanan değere mod denir.

13, 10, 13, 12, 17, 13, 14 olsun en çok tekrarlayan sayı yani mod = 13 'tür.Çünkü tekrarlanma sayısı (frekansı) = 3'tür.

Dağılım Ölçüleri

Dağılım Ölçüleri; değişkenin sahip olduğu değerlerin birbirinden ne kadar farklı olduğunun bir ölçüsüdür.

1-Değişim Aralığı

Değişim Aralığı = max değer — min değer

10, 12, 13, 15,17,19 sayılarımız olsun bunların değişim aralığı = 19–10 = 9

2-Standart Sapma

Ortalamadan olan sapmaların genel bir ölçüsüdür.

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

3-Varyans (Variance)

Standart sapmanın karesidir.Bir değişkenin varyansı, değerlerin merkezi eğilimden ne kadar farklılaştığını ve her bir değerin bir başkasından ne kadar farklılaştığını tanımlar.Varyans, değişken içinde her bir veri noktası ne kadar değerlidir bunu ifade eder.

· Eğer varyans düşükse birçok veri noktası merkezi eğilim ile benzerdir, bu nedenle her bir veri noktası ölçüm yapılan kavram hakkında çok az farklı bilgi verebilir.

· Eğer varyans yüksekse, her bir veri noktası ölçüm hakkında yeni ve farklı bilgiler sunabilecektir.

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Veriyle kalın.

Kaynak : Datajarlabs

Veri Bilimi Veri Istatistik Makine Ogrenmesi