Veri Bilimi Okuryazarlığı e-Kılavuzu

Zafer Demirkol
Yapay Zeka Mentörü ve Yazar

VERİ BİLİMİ OKURYAZARLIĞI E-KILAVUZU

Günümüzde Veri Bilimi ve Yapay Zeka oldukça popüler konulardır. Bu alanlar her geçen gün hayatımıza daha fazla katkı sağlıyorlar. Verinin büyük bir hızla büyümesi, veriyi toplama, saklama, yönetme, analiz ve kullanmaya yönelik araçları, yöntemleri de geliştirmemizi zorunlu kılmıştır. Bunlara yönelik çalışmalar "Veri Bilimi" çatısı altında yapılmaktadır.

Büyük veri ve buna dayalı Yapay Zekanın getirdiği paradigma değişimi bütün alanları etkilemiş ve etkilemeye devam edecektir. Yeni meslekler oluşmuş, gelecekte de bambaşka meslekler oluşacaktır.

Günümüzde veriyi anlamak, buna dayalı iş yapmak son derece değerli bir beceri haline gelmiştir.

İşte bu kılavuz; "veriyi", günümüz veri bilimindeki kavramları, teknikleri anlamanıza yardımcı olacaktır.

Bu kılavuzu anlayabilmeniz için teknik bilgiye ihtiyacınız yoktur. İçerik hazırlanırken her alandaki profil hedef alınmıştır. Sıkıcı teknik açıklamalar, formüller veya ifadeler kullanılmamaya özen gösterilmiştir. Konuların temelinin, mantığının anlaşılmasına yönelik hazırlanmıştır. Bunun için pek çok çizim-görsel kullanılmıştır.

Bu kılavuzun amacı kariyerinizi oldukça etkileyecek veriye dayalı çalışma teknolojilerinin, kavramlarının anlaşılması ve bir içgörü sağlanmasıdır. Ayrıca Veri Bilimi konusunda çalışmak isteyenlere, bu alandaki temel kavramları anlatarak bir başlangıç noktası sunmaktır. Veri Biliminde kullanılan en yaygın teknikler, yöntemler ve kavramlar kılavuzun içeriğini oluşturmaktadır.

Zafer Demirkol

ZAFER DEMİRKOL

Programlamaya 1985 yılında 16 yaşında başladı. Televizyona bağlanan bir Sinclair bilgisayar ile ilk “merhaba dünya” programını yazdı. 18 yaşında Boğaziçi Üniversitesinde Quick Basic kurslarına katıldı. Yıldız Teknik Üniversitesi Elektrik Mühendisliğini kazandı, okurken sürekli kod yazmaya devam etti. Üniversitedeyken karmaşık sayılarla 4 işlem yapan bir program, animasyon, teklif hazırlama gibi uygulamalar geliştirdi.

Üniversite bittikten sonra değişik firmalarda çalıştı. 1999 yılında Genel Müdür olarak çalıştığı firmadan ayrılarak Avustralya’da “ New South Wales” Üniversitesinde “English for Business Communication” eğitimi aldı. Bu yıllarda hızla gelişen Web teknolojilerini yakından takip ederek, değişik uygulamalar geliştirdi.

Avustralya’da Üniversitede çalışma gruplarında Web yazılımları üzerine olan bilgisini arttırdı. ASP (Active Server Pages) ile burada tanıştı. Türkiye’ye döndüğünde ilk ASP kitabını yayınladı. Kitap ilk çıktığı yıl 5 ve toplamda 14 baskı yaparak satış rekorları kırdı. Daha sonra ASP.NET teknolojisine ait kitapları yayınlandı. Şu ana kadar toplam 12 kitabı yayınlanmıştır.

“Çocuklar için Kodlama” kitabı 4. Baskıya ulaştı ve İngilizceye çevrildi. Türkiye’de Çocuklara kodlama öğretmeye yönelik çalışmaların başlangıç ve rehber kitabı olmuştur.

Son kitabı “Herkes için Yapay Zeka” oldukça ilgi görmüş ve kategorisinde “En Çok Satanlar” listesinde yer almaktadır.

Profesyonel kariyerini pek çok proje ve yazılım firmasında çalışarak geçirdi. “PC World”, Byte, “Mobile Life” dergilerinde köşe yazarlığı yaptı. Yazılım üzerine onlarca makale yayınladı. 3 dönem Microsoft MVP (Most Valuable Professional) seçildi.

Yeditepe, Maltepe, Bahçeşehir ve Medipol üniversitelerinde, “Bilgisayar Mühendisliği”, “Bilişim Teknolojileri”, “Bilgisayar Öğretmenliği”, “Bilişim Yönetim Sistemleri” fakültelerinde değişik dönemlerde “Nesneye Yönelik Programlama”, “Web Programlama”, “İleri Web Tasarımı”, “Veritabanı Programlama”, “Mobil Programlama”, “Python”, “Yapay Zeka” dersleri verdi ve vermeye devam etmektedir.

36 yıllık programcılık tecrübesini kitapları, makaleleriyle paylaştığı gibi Profesyonel eğitimlerle de genç programcılara aktarmakta. Özel sektör, kamu ve devlet kuruluşlarına değişik yazılım teknolojileri konusunda onlarca eğitim düzenlemiş, düzenlemektedir. Profesyonellere yönelik düzenlediği eğitim konuları: ASP.NET MVC, JavaScript, Mobil Hybrid Programlama, HTML5, JQuery, CSS3, iOS swift, AngularJs, Python ve “Yapay Zeka Teknolojileri” dir.

Son yıllarda Yapay Zeka çalışmalarına odaklandı. Çalışmalarını Linkedin, Medium ve Twitter’dan takip edebilirsiniz.

Avrupa Yapay Zeka Birliği Üyesidir.

DİJİTAL TEKNOLOJİ GELİŞTİRİCİLER HAKKINDA

Bilgi toplumuna dönüşüm yolunda önemli çalışmalar yürüten Türkiye Bilişim Vakfı ve ileri analitik ve yapay zeka firması SAS olarak global vizyonumuzu, tecrübelerimizi Türkiye'nin gelişmesine katkı sunmak üzere paylaşmak istiyoruz. Çünkü, Türkiye'nin daha refah bir ülke haline gelmesi ve rekabette öne çıkması için şimdinin ve geleceğin ihtiyacı olan, teknolojinin gelişimi ile ortaya çıkan yeni meslek alanlarında yetişmiş insan açığının bir an önce kapatılması gerektiğine inanıyoruz. Dijital Teknoloji Geliştiriciler Programı, bu vizyonda gerçekleştirilecek projeler üretmek üzere hayata geçirilmiştir.

https://dtg.org.tr/

İÇİNDEKİLER

Bölüm 1: Veri Nedir? Nasıl Elde Edilir?	**9-17**
Veri Nedir?	9
Veri Seti (Kümesi)	9
Veri Tipleri	10
Yapılandırılmış – Yapılandırılmamış Veri	11
Veriyi Nereden Temin Ederiz?	12
Verilerin Saklanması	13
Verilerle Çalışma - Veri Bilimi İşlemleri	13
Problemin Tanımı	14
Verinin Temini	14
Bölüm 2: Verinin Hazırlanması ve Analizi	**14-24**
Verinin Hazırlanması	14
Verinin Keşfi / Analizi	14
Descriptive Analysis - Açıklayıcı Analiz	15
Diagnostic Analysis - Teşhis Analizi	15
Predictive Analysis - Tahmine Dayalı Analiz	16
Prescriptive Analysis - Kuralcı Analiz	17
Keşifsel Veri Analizi (Exploratory Data Analysis – EDA)	17
Verinin Anlaşılması Açıklanması	18
Verinin Temizlenmesi – Ön işlenmesi	18
Verinin Birleştirilmesi	20

Veri Dönüşümü	21
Min-Max Normalleştirmesi	21
Normalleştirmenin Faydaları	22
Bölüm 3: Veri Görselleştirme	**25-45**
Veri Görselleştirme	25
Karşılaştırma Grafikleri	25
Çizgi Grafikler	26
Bar Grafikler	27
Radar Grafikleri	28
İlişki Grafikleri	28
Scatter Plot	29
Belirli Aralık Dağılımları (Histogram) ile Scatter Grafikleri	30
Kabarcık (Baloncuk) Grafikleri	31
Correlogram – Korelasyon Matrisleri	32
Heatmap	33
Kompozisyon (Bileşim) Grafikleri	36
Pasta Grafikleri	36
Donut Grafiği	36
Yığın Bar Grafiği (Stacked Bar Chart)	37
Yığın Alan Grafiği (Stacked Area Chart)	38
Venn (Küme) Diagramları	39

Dağılım Grafikleri	39
Histogram	40
Density (Yoğunluk) Grafikleri	42
Kutu Grafikleri (Box Plot)	43
Keman (Violin) Grafikleri	45
Bölüm 4: Temel İstatistik	**46-57**
Temel İstatistik	46
Tanımlayıcı İstatistik	47
Merkezi Eğilim	47
Aritmetiksel Ortalama	48
Medyan	49
Mod	49
Merkezi Eğilimin Değerlendirilmesi	50
Yayılım Eğilimi	50
Açıklık (Aralık)	51
Ortalama Mutlak Sapma	52
Varyans - Standart Sapma	53
Örneklem Varyansı – Standart Sapması	54
Normal Dağılımı	55
Eğrilik - Çarpıklık	55
Simetrik Dağılım	55

Pozitif Çarpıklık	56
Negatif Çarpıklık	56
Çıkarımsal İstatistik	57
Bölüm 5: Makine Öğrenimi	**57-67**
Yapay Zeka - Makine Öğrenimi	57
Makine Öğrenme Tipleri ve Algoritmaları	58
Lineer Regresyon Algoritması	59
Polinom Regresyon	60
Karar Ağaçları	61
Destek Vektör Makineleri (Support Vector Machines)	61
Lojistik Regresyon	62
En Yakın Komşular Algoritması	63
Naive Bayes Algoritması	64
Kümeleme Algoritmaları – K-Means Kümeleme Algoritması	65
Günümüz Yapay Zeka Teknoloji ve Disiplinleri	66
Geleneksel Makine Öğrenimi – Yapay Sinir Ağları/Derin Öğrenme	67

VERİ NEDİR?

Veri bilginin en küçük birimi, parçasıdır. Veri latince "Datum" kavramından gelir. Veriler bir araya gelip bilgiyi oluştururlar. Aslında veri dediğimiz kavram, gerçeğin bir soyutlamasıdır. Veri tek başına bir şey ifade etmez, bir anlam, bilgi oluşturmaz. Veriler bir bilgi bütününün parçası olabildikleri gibi başka verilerle kıyaslanarak da bilgi döndürürler.

Örneğin Notebook'unuzun ağırlığının 220gr olduğu verisi tek başına bir şey ifade etmez ama, çantanızın veya taşıdığınız diğer şeylerin ağırlıkları ile kıyasladığınızda size bir bilgi verir.

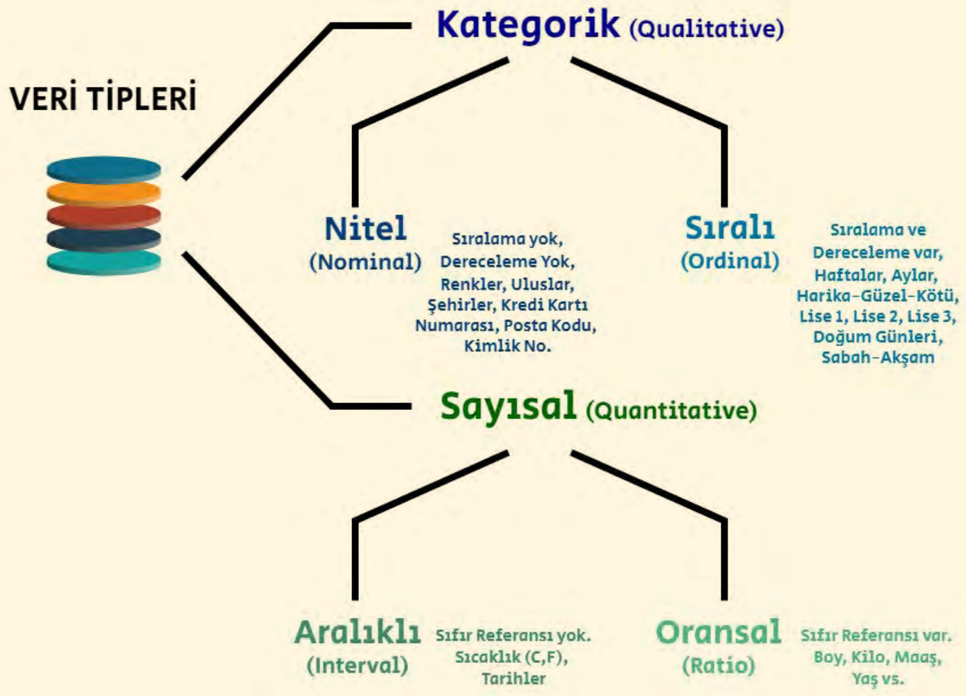
VERİ SETİ (KÜMESİ)

Varlıklar veya olgular verilerle ifade edilirler. Her bir varlık da bir takım nitelikteki (özellik, değişken) verilerin bir araya gelmesiyle tanımlanır. Örneğin bir araba marka, model, uzunluk, motor hacmi vs. gibi birçok nitelikle tanımlanır. Arabalara dair bir veri seti oluşturmak istersek, bu niteliklere sahip verilerden oluşan varlıkların(arabaların), her birine ait kayıtlarını(satırlar) oluşturmalıyız. Tanım biraz uzun oldu ancak aşağıdaki şemadan bu durum daha net anlaşılıyor. Daha kısa ifade etmek istersek, bir veri seti veya kümesi belirli niteliklerden oluşan bir grup varlığın bir araya gelmesinden oluşur.

Varlıklar satırlarla ifade edilir ve nitelikleri sütunlardan oluşur. İlerleyen bölümlerde göreceğimiz gibi bu aynı zamanda yapılandırılmış veriyi de tanımlar.

VERİ TİPLERİ

Farklı nitelikteki verilere veri tipleri denir. Veriler "sayısal" ve "kategorik" olmak üzere iki ana nitelikte olabilir. Bu iki ana nitelikte kendi içinde alt kategorilere ayrılır.



Kategorik veriler, ölçülemez, sayılarla ifade edilmez ve üzerinde aritmetiksel işlem yapılamaz. Esas olarak kelimeler, resimler ve sembollerden oluşurlar. Bazı kategorik veriler sıralı olabilir. Ancak bu sıralama sayısal değerlere göre yapılmaz. Ülke, şehir isimleri, renkler, büyük, küçük, orta, cinsiyet vs. Kategorik verilere örnektir.

Sayısal veriler, nicel bir tanım sunar ve sayılarla ifade edilirler. Ölçülebilirlerdir. "Kaç", "Ne kadar"

"Ne sıklıkla" sorularına cevap verirler.

Sayısal veriler de kendi içinde temel olarak, "Aralıklı" ve "Oransal" olarak ikiye ayrılır. Aralık sayısal veriler ile "Oransal" veriler hemen hemen aynıdır. Aralarındaki fark, "Oransal" veriler "sıfır" noktasında başlayarak devam eder ve hiçbir zaman sıfırın altına düşmez. Örneğin, "yaş", "boy" verilerinin sıfırın altında olması mümkün değildir, dolayısı ile bunlar, oransal verilerdir. Ayrık veriler sıfır ve eksi değerleri de gösterebilir.

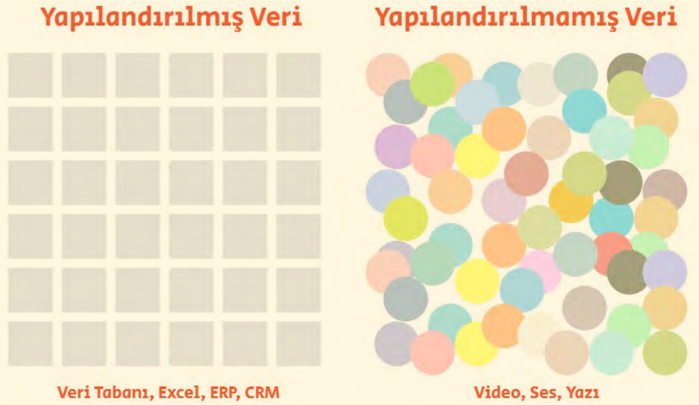
Oransal verilerin sıfır gibi bir referans noktası olduğu için çarpma ve bölme işlemlerine tabii olabilir. Ancak ayrık verilerin böyle bir referansı olmadığından sadece toplama ve çıkarma işlemleri yapılabilir.

Bunun tipik bir örneği Santigrat derece birimine dayalı sıcaklık ölçümüdür. 30 santigrat derece 15 santigrat dereceden 15 derece daha fazladır diyebiliriz ama iki katıdır diyemeyiz. Çünkü santigrat derecenin dayandığı bir mutlak sıfır noktası yoktur. Eksi değerler de alabilir ve bunun bir bitiş referansı yoktur.

İŞLEM	KATEGORİK		SAYISAL	
	Nitel (Nominal)	Sıralı (Ordinal)	Aralıklı (Interval)	Oransal (Ratio)
Eşitlik	✓	✓	✓	✓
Sıralama		✓	✓	✓
Ekle-Çıkar			✓	✓
Çarp-Böl				✓
Mod	✓	✓	✓	✓
Medyan		✓	✓	✓
Ortalama			✓	✓
Geometrik Ortalama				✓
Frekans Dağılımı	✓	✓	✓	✓

YAPILANDIRILMIŞ – YAPILANDIRILMAMIŞ VERİ

Pek çok veri tipi vardır. Başka şekilde söylemek gerekirse veriler pek çok şekilde kategorize edilir. Yukarıda bunlardan kategorik ve sayısal olarak tiplere ayrılmış olanını inceledik. Diğer bir sınıflandırmaysa "Yapılandırılmış – Yapılandırılmamış" verilerdir.



Yapılandırılmış veriler, organize edilmiş veriler olarak da bilinir. Bunlar belirli bir formatta saklanan verilerdir. Yukarıdaki bölümde yapısını gösterdiğimiz veri setleri, yapılandırılmış verilerin yapısıdır. Genellikle satır ve sütunlardan oluşurlar. Veritabanı tablosu veya hesap tablosu gibi özel programlar kullanılarak bir tablo halinde tutulabildikleri gibi, boşluk veya özel karakterlerle ayrılmış metin dosyaları olarak da saklanırlar.

Metin Dosyası

gelirVerileri.csv – Not Defteri

Dosya Düzen Biçim Görünüm Yardım

```
sehir,yas,tecrube,gelir,ev
is,20,1,3,y
is,40,18,10,v
is,37,14,7,v
an,51,26,7,y
an,49,28,10,v
an,39,20,6,v
iz,47,28,9,v
iz,29,8,6,y
iz,32,15,9,v
```

Hesap Taplosu Veya Veri Tabanı Tablosu

sehir	yas	tecrube	gelir	ev
is	20	1	3	y
is	40	18	10	v
is	37	14	7	v
an	51	26	7	y
an	49	28	10	v
an	39	20	6	v
iz	47	28	9	v
iz	29	8	6	y
iz	32	15	9	v

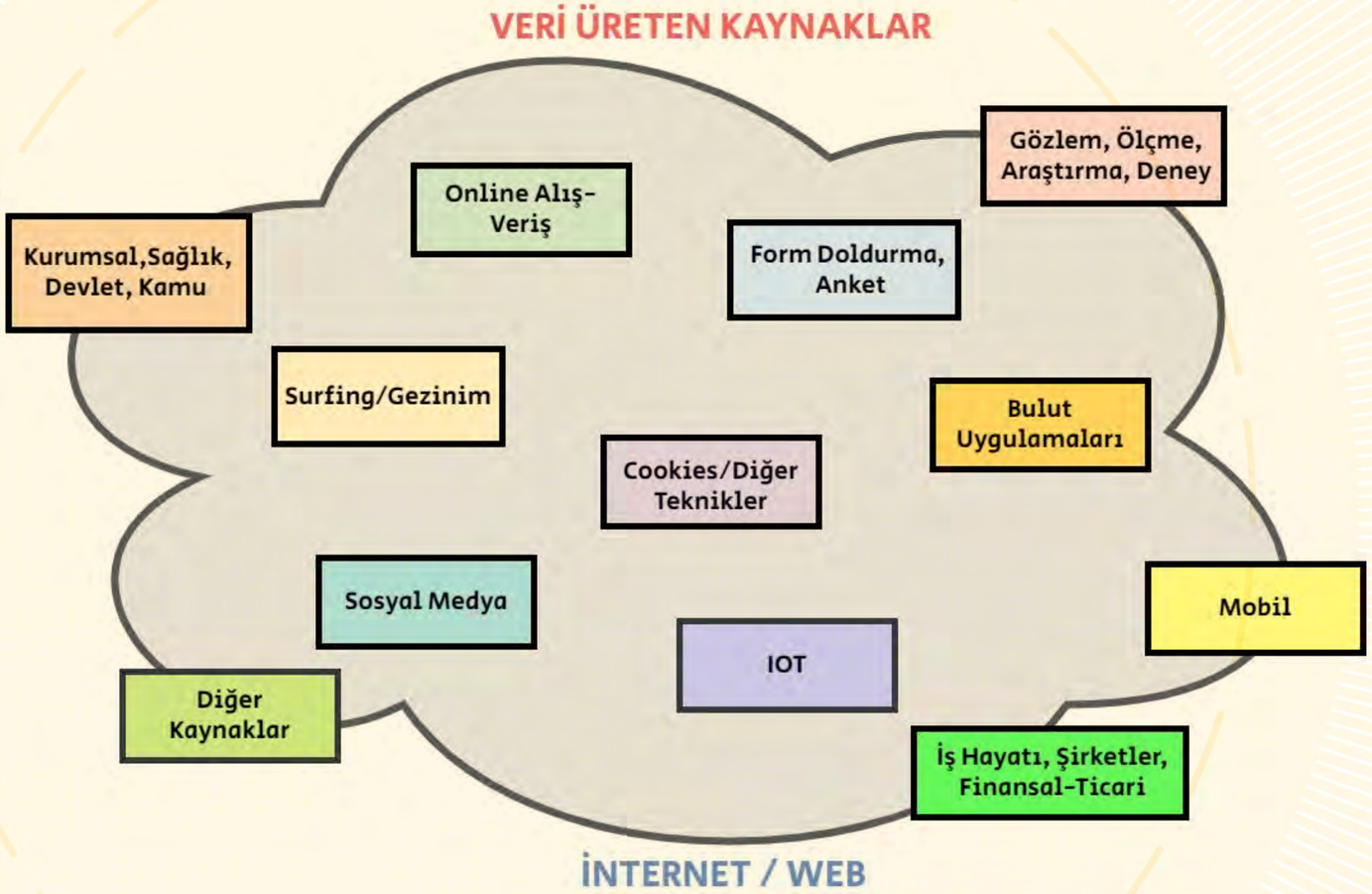
Örneğin yukarıdaki metin dosyasındaki veri saklama formatı virgülle ayrılmış veriler olarak organize edilmiştir. Ama virgül yerine boşluk, "|", ";" gibi özel karakterler konarak da veriler ayrıştırılabilir ve organize edilebilir. Bu tip veriler genellikle Veri Bilimi, Geleneksel makine öğrenimi alanlarının temel veri yapısıdır.

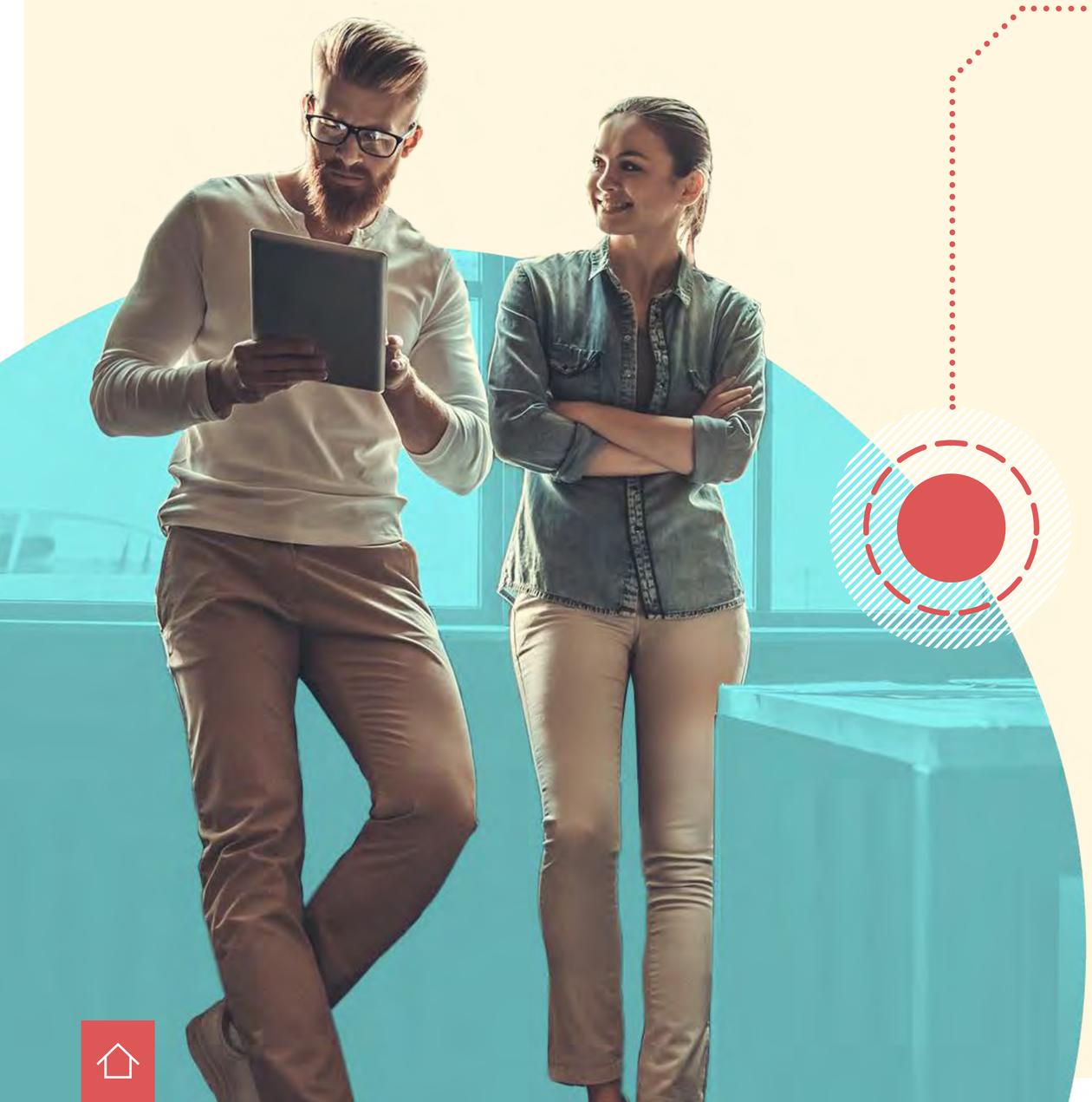
Yapılandırılmamış veriler, organize edilmemiş verilerdir. Bunlar serbest halde bulunurlar ve herhangi bir standart organizasyon hiyerarşisine sahip değildir. Resim, video, ses, serbest formda yazılmış metinler yapılandırılmamış veri tipine girerler. Bu veriler günümüzde ağırlıklı olarak Yapay sinir ağları ve Derin öğrenme algoritmalarıyla işlenirler ve günümüz yapay zekasının temelini oluştururlar.

VERİYİ NEREDEN TEMİN EDERİZ?

Ölçülebilen, gözlemlenebilen, görülebilen her şey veridir. Verileri kendiniz toplayabileceğiniz gibi hazır kaynaklardan da temin edebilirsiniz. Hazır kaynaklarda, ücretli ve ücretsiz olarak veri sunabilirler. Çalıştığınız konuya ait verilerin toplanmasında genellikle bu işte uzman ticari firmalardan yardım alırsınız. Eğitim amaçlı ve jenerik konularda pek çok açık veri kaynağı mevcuttur. Veriler değişik kaynaklardan gelebilir; En başta internet. İnternet veri kaynaklarına altyapı sunar. Örneğin anket veya form doldurularak elde edilen verilerin pek çoğu internet, web altyapısını kullanır. Bunun haricinde, alışveriş tercihlerimiz, ziyaret ettiğimiz siteler, sosyal medya hareketlerimiz hep veri üretirler. Mobil telefonlarımızdaki kullanımlarımız, IOT cihazlarının sensörleri, kameralar veri üreten kaynaklardır. Ölçme ve gözleme dayalı edindiğimiz veriler, araştırmalar veri üreten alanlardır.

Bunun yanında firmaların, devlet-kamu kuruluşlarının, sağlık sektörünün, ticari-finans alanının, akademik alanın ürettiği veriler vardır. Bunların bir kısmı internet üzerinde herkese açık olabileceği gibi, bir kısmı da özel kullanıma aittir. Veri Bilimi, Makine öğrenimi, istatistik için çokça veri gerekir ve bugünkü yapay zeka teknolojileri çok fazla veri ile çalışır. Bunun için günümüzde veri çok kıymetlidir ve büyük firmalar veri temini için pek çok yöntemi kullanırlar.





VERİLERİN SAKLANMASI

Yapılandırılmış veriler basit bir metin dosyasından, veri tabanı tablosuna kadar pek çok yöntemle saklanabilir. Yapılandırılmamış veriler genelde bir resim, video veya ses dosyası şeklinde, bilgisayarların saklama alanlarında bu formatlara özel dosya uzantıları ile saklanırlar. Eğer bu veriler bir veri kaynağı şeklinde sunulacaksa, hızlı ve yüksek kapasiteli sunucularda, saklanırlar. Büyük verinin saklanması, yönetimi için farklı teknikler ve altyapılar kullanılır.

VERİLERLE ÇALIŞMA - VERİ BİLİMİ İŞLEMLERİ

Verilerle çalışma "Veri Bilimi" alanında incelenir. Veri Bilimi verilerle yapılabilecek olası bütün işlemleri tanımlar ve metotlar sunar. Veri biliminin amacı, karar alma süreçlerini verilere dayandırarak, geliştirmektir.

PROBLEMİN TANIMI

Bir veri bilimi çalışmasının ilk aşamasında ne yapmak istediğinizi, hangi sorulara cevap aradığınızı iyi belirlemelisiniz. Bir veri bilimi çalışması genellikle ekip çalışması olur ve bu ekipteki herkesin rolünün belirlenmesi gerekir. Projenin amacı, analizlerin nasıl gerçekleşeceği, hangi kaynakların ve araçların kullanılacağı bir zaman çizelgesinin oluşturulması gibi ön çalışmalar yapılmalıdır.

VERİNİN TEMİNİ

2. aşama verilerin teminidir. Verilerin temininde temel iki yöntem kullanabilirsiniz. Birinci yöntemde verileri bizzat kendiniz toplarsınız. Ancak bu zahmetli ve uzun bir süreç demektir. Günümüzde hazır veriler sunan/satan pek çok 3. Parti şirketler vardır. Bu verilerin bir kısmı ücretli olduğu gibi bir kısım veri setleri de ücretsiz olarak sunulmaktadır. Bu konuyla ilgili "Verileri Nereden Temin Edebiliriz" başlıklı bölümü incelemenizi öneririm.

VERİNİN HAZIRLANMASI

Veri bilimi çalışması için 3. Aşama verilerin hazırlanmasıdır. Çalışmanız için elde ettiğiniz veriler çoğunlukla hemen kullanıma hazır değildir. Öncesinde veriler üzerinde bir takım işlemler yapmanız gerekebilir. Bu ön işlemler son derece önemlidir. Veri bilimi çalışması sonucunda bir rapor veya verilere dayalı bir model çıkar. Sağlıklı veri girdisi, sağlıklı bir rapor veya model çıktısı demektir. Bunun için verilerimizi "temizlememiz", amacımız için "uyarlamamız" ve "dönüştürmemiz" gerekebilir.

Bu işlemler için farklı yöntem ve metotlar kullanılır. Örneğin veri temizlemede, eksik verilerin tamamlanması, tekrarlı verilerin tespiti, yazım hatalarının düzeltilmesi, gerçekte imkansız olabilecek verilerin tespit edilmesi, veri girişi esnasında yapılabilecek hataların tespiti ve düzeltilmesi gibi işlemler yer alır.

Bazen farklı kaynaklardan gelen verilerin birleştirilmesi, uyarlanması, alanlarının düzenlenmesi veya dönüştürülmesi gerekebilir. Bütün bu işlere verinin hazırlanması denir.

VERİNİN KEŞFİ / ANALİZİ

Veriyi anlamaya yönelik yapılan analizlerdir. Veri keşfi ve analizine yönelik pek çok yöntem vardır. Bunlar

Descriptive Analysis - Açıklayıcı analiz

Diagnostic Analysis - Teşhis Analizi

Predictive Analysis - Tahmine Dayalı Analiz

Prescriptive Analysis - Kuralcı Analiz

Exploratory Data Analysis – Keşifsel Veri Analizi

Bütün bu yöntemleri bir birinden kesin bir şekilde ayırmak doğru değildir. Birbirlerini destekleyen, birbirlerinin üzerine inşa edilen yöntemlerdir. Bunların karışımından ve olası yeni yöntemlerin dahil edilmesinden farklı analiz yaklaşımları da mümkündür.

Şimdi kısaca bu analiz tiplerine bakalım.

DESCRIPTIVE ANALYSIS - AÇIKLAYICI ANALİZ

En temel analiz tipi olan “Açıklayıcı Analiz” geçmiş verilere bakarak “neler olduğunu” anlamaya yönelik yöntemler sunar. Veri kümelerinin tanımlanmasına ve anlaşılmasına yardımcı olur. Bu analizde genelde verinin sunulduğu gösterge tabloları kullanılır. Verilerin özetlenmesini ve analiz edilmesini kolaylaştıran farklı yöntemleri de kullanır. Nüfus sayımı ve bunun sunumu Açıklayıcı analize bir örnektir. Henüz teşhis veya tanımlayıcı bir analiz yapılmamasına rağmen bunlar için bir altyapı sunar. Veya müşterilerin satın alma tercihlerine göre sınıflandırılması ve sunulması yine Açıklayıcı analiz yöntemine bir örnektir.

DIAGNOSTIC ANALYSIS - TEŞHİS ANALİZİ

“Neden” sorusunu bulmak için yapılan analizlerdir. Keşif için veya bir şeyin “neden” olduğunu belirlemek için kullanılır. Açıklayıcı analizlerden gelen sonuçların nedenlerini bulmak için detaylara iner. Ağırlıklı olarak korelasyonlar, detaya inme, veri keşfi, veri madenciliği gibi tekniklerle karakterize edilir. Örneğin bir kargo şirketinin belirli bir hattının daha yavaş çalışmasının nedenlerinin araştırılması. Veya bir pazarlama faaliyetinin diğerine göre neden daha başarılı olduğunun analizi bunlara örnek gösterilebilir.

PREDICTIVE ANALYSIS - TAHMİNE DAYALI ANALİZ

"Ne olması muhtemel" sorusuna cevap vermeye çalışır. Geçmişteki verileri kullanarak gelecek hakkında tahminler yapar. Bu kategoride "Açıklayıcı" ve "Teşhis Analiz" lerine göre tahminlerde bulunulur. Tahmine dayalı analitik, şirketlere verilere dayalı eyleme geçirilebilir öngörüler sağlar.

Bu tür analiz, gelecekteki bir sonucun olasılığı hakkındaki tahminleri içerir.

Bu analiz istatistiksel modellemeyi kullanır. Tahmine dayalı analitiğin temeli olasılıklara dayandığından hiçbir istatistiksel sonucun geleceği %100 doğrulukla "tahmin edemeyeceğini" bilmek önemlidir. Satış tahminleri, risk değerlendirmesi gibi uygulamalar bu analiz kategorisine örnek gösterilebilir.

PRESCRIPTIVE ANALYSIS - KURALCI ANALİZ

Ne yapmalı, nasıl gerçekleştirebiliriz sorularına cevap arar. Tahmine dayalı analize benzer ancak ondan farklı olarak eylem/aksiyon için süreçleri de tanımlar. Eylem adımları Yapay Zeka gibi büyük veri ve işlem gerektiren süreçlerden sonra oluşturulabilir ve hatta eyleme geçirilebilir. Eylemler insan müdahalesi olmadan optimize edilebilir ve güncellenebilir.

Veri Analizi Tipleri

KEŞİFSEL VERİ ANALİZİ (EXPLORATORY DATA ANALYSIS – EDA)

Yukarıda saydığımız analiz tiplerinin yanı sıra, özellikle son yıllarda Yapay Zeka, Makine Öğreniminin gelişmesiyle Keşifsel veri analizi ön plana çıkmıştır. Genellikle makine öğrenimi modellemesi öncesi verilere ait bir içgörü oluşturabilmek için yapılan analizlerdir.

Böylece modele gönderilen verilerin bir sağlaması yapılır ve geçerliliği denetlenmiş olur ayrıca modelin çıktıları bu analize göre daha sağlıklı değerlendirilebilir. Bazen veriler ve makine öğrenimi süreci arasında tutarsızlıklar olabilir, bu durumdan kaçınmak için makine öğrenimi modellemesi için kullanılan verilerin daha iyi tanınması gerekir, işte bunun için makine öğrenimi modellemesi öncesi Keşifsel veri analizi yapmak son derece sağlıklıdır. Öte yandan daha iyi bir modelleme için hangi özelliklerin kullanılması gerektiği hakkında da içgörü sağlar.

Keşifsel veri analizi veriler arasındaki ilişkileri tespit etmeye, farklı açılardan analiz etmeye yarayan yöntemleri kapsar, bu yapısıyla yukarıdaki "teşhis analizi" yöntemine çok benzer. Sonuçlar genellikle görselleştirme ve özetleme olarak sunulur.

Bunun yanında eksik, hatalı verilerin tanımlanması ve düzeltilmesi gibi veri ön işleme yöntemlerini de kullanır.

Temel Adımları şekilde görüldüğü gibidir.

VERİNİN ANLAŞILMASI AÇIKLANMASI

Bu aşama mevcut veri setinin anlaşılmasına yöneliktir. Hangi alanların olduğu, kaç satırdan oluştuğu, alanların neyi ifade ettiği gibi veri setine ait yapıları anlamaya yöneliktir. Bunun yanında veri setini anlamaya yönelik istatistiki verilere de bakılır. Alanların ortalamaları, standart sapmaları, maksimum, minimum değerleri gibi istatistiki bilgiler edinilir. Bu aşamada çalışılacak veri seti mümkün olduğu kadar kavranmaya, yapısı anlaşılmaya çalışılır.

VERİNİN TEMİZLENMESİ – ÖN İŞLENMESİ

Elimizdeki veri setini anlamaya ve keşfetmeye yönelik bir takım işlemler yapmadan önce, veri setindeki verilerin çoğunlukla düzenlenmesi, bir takım ön işlemlerin yapılması gerekir. Bu işlemler şemada gösterildiği gibi:

- Temizleme
- Birleştirme
- Dönüştürme
- İndirgeme

Başlıkları altında incelenebilir.

Veri Temizleme / Ön-İşleme

Verilerin temizlenmesi: Eksik verilerin tamamlanması, "göze batan", hatalı, "anlamsız" verilerin düzgünleştirilmesi, tutarsızlıkların giderilmesi gibi işlemleri kapsar.

	sehir	yas	tecrube	gelir	ev
2	is	20	1	3	y
3	is	40	18	10	v
4	is		7	6	y
5	is	33	10	7	v
6	is	45	25	8	v
7	is	55	3 6	17	v
8	is	31	12	8	y
9	is	24	2	3	y
10	is	61	25	15	598
11	an	37	z	7	v
12	an	45	15	18	y
13	an	51		7	y
14	an	49	28	10	v
15	an	39	20	6	v
16	an	30	10	4	y
17	iz	47	28	9	v
18	iz	29	8	6	y
19	iz		15		v
20	iz	53	25	13	y
21	iz	27	10	5	12
22	iz	23	6	4	y
23	iz	62	40	12	v
24	is	37	17	15	v
25	an	20	2	2	
26	iz	54	35	10	v
27	is	52	33	15	v
28	an	62	40	13	v
29	iz	25	5	4	y
30	an	30	10	7	y

Örneğin şöyle bir veri setimiz olsun. Burada 4 ve 19. Satırdaki "yaş" verileri eksik. 13. Satırdaki tecrübe, 19. Satırdaki "gelir" ve 25. Satırdaki "ev" verileri eksik. Bunun yanında "tecrübe" alanında "3 6" şeklinde yanlış yazılmış veri "z" olarak girilmiş bir karakter verisi mevcut. "ev" alanında da "598" ve "12" değerleri girilmiş. Oysa burası "v" (var) veya "y" (yok) değerleri girilmesi gereken alanlar.

Veri setiminiz eksikliklerinin tamamlanması, hatalarının düzeltilmesi gerekir. Bu işlemleri yapmanın değişik yöntemleri ve araçları vardır. Ancak sonuç olarak amaç bunların giderilmesidir.

Sonuç olarak veri setimizin şu şekilde eksiksiz ve doğru olması gerekir:

Veri Temizleme / Ön-İşleme

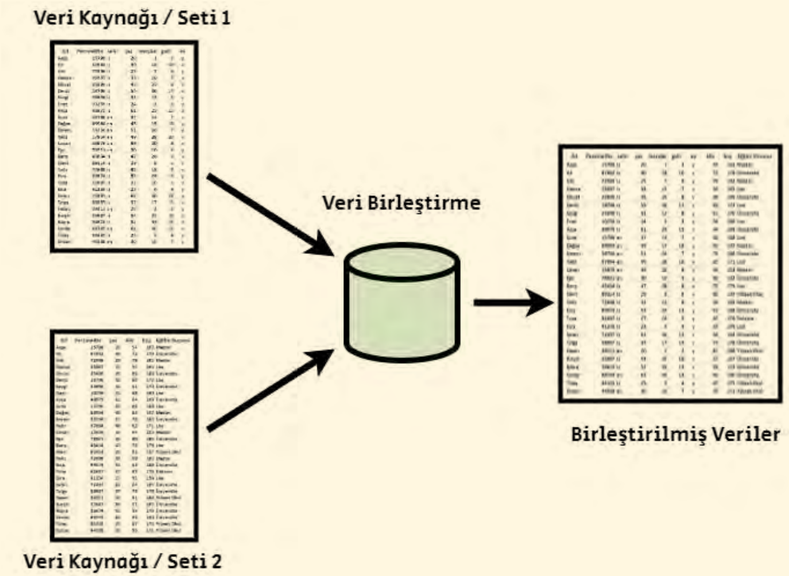
	sehir	yas	tecrube	gelir	ev
2	is	20	1	3	y
3	is	40	18	10	v
4	is		7	6	y
5	is	33	10	7	v
6	is	45	25	8	v
7	is	55	3 6	17	v
8	is	31	12	8	y
9	is	24	2	3	y
10	is	61	25	15	598
11	an	37	z	7	v
12	an	45	15	18	y
13	an	51		7	y
14	an	49	28	10	v
15	an	39	20	6	v
16	an	30	10	4	y
17	iz	47	28	9	v
18	iz	29	8	6	y
19	iz		15		v
20	iz	53	25	13	y
21	iz	27	10	5	12
22	iz	23	6	4	y
23	iz	62	40	12	v
24	is	37	17	15	v
25	an	20	2	2	
26	iz	54	35	10	v
27	is	52	33	15	v
28	an	62	40	13	v
29	iz	25	5	4	y
30	an	30	10	7	y

	sehir	yas	tecrube	gelir	ev
2	is	20	1	3	y
3	is	40	18	10	v
4	is	25	7	6	y
5	is	33	10	7	v
6	is	45	25	8	v
7	is	55	36	17	v
8	is	31	12	8	y
9	is	24	2	3	y
10	is	61	25	15	v
11	an	37	14	7	v
12	an	45	15	18	y
13	an	51	26	7	y
14	an	49	28	10	v
15	an	39	20	6	v
16	an	30	10	4	y
17	iz	47	28	9	v
18	iz	29	8	6	y
19	iz	32	15	9	v
20	iz	53	25	13	y
21	iz	27	10	5	y
22	iz	23	6	4	y
23	iz	62	40	12	v
24	is	37	17	15	v
25	an	20	2	2	y
26	iz	54	35	10	v
27	is	52	33	15	v
28	an	62	40	13	v
29	iz	25	5	4	y
30	an	30	10	7	y

Az veriden oluşan veri setlerinde çok fazla hata veya eksik olmayabilir. Ancak binlerce, on binlerce hatta yüzbinlerce satırdan oluşan veri setlerinde çoğunlukla bu tip hata veya eksiklikler olur. Veri işlemeden önce bunların giderilmesi gerekir.

VERİNİN BİRLEŞTİRİLMESİ

Farklı kaynaklardan gelen ilişkili verilerin birleştirilmesi işlemini ifade eder. Birleştirilmiş veriler daha zengin yeni bir kaynağı oluşturur. Bu yöntem için çeşitli teknikler uygulanır.



Örneğin elimizde şöyle iki adet veri seti olsun:

Ad	PersonelNo	sehir	yas	tecrube	gelir	ev
Ayşe	15798	is	20	1	3	y
Ali	87962	is	40	18	10	v
Veli	72596	is	25	7	6	y
Gamze	25987	is	33	10	7	v
Murat	25896	is	45	25	8	v
Deniz	18796	is	55	36	17	v
Sevgi	25896	is	31	12	8	y
Fırat	10259	is	24	2	3	y
Ayça	48975	is	61	25	15	v
Suna	15796	an	37	14	7	v
Doğan	88964	an	45	15	18	y
Kerem	58796	an	51	26	7	y
Yeliz	57964	an	49	28	10	v
Canan	15976	an	39	20	6	v
Ege	78921	an	30	10	4	y
Barış	45634	iz	47	28	9	v
Mert	89314	iz	29	8	6	y
Yeliz	72698	iz	32	15	9	v
Esra	89574	iz	53	25	13	y
Tuna	92687	iz	27	10	5	y
Esra	91236	iz	23	6	4	y
Selen	72397	iz	62	40	12	v
Tolga	56987	is	37	17	15	v
Hasan	39211	an	20	2	2	y
Burçin	55987	iz	54	35	10	v
Büşra	26674	is	52	33	15	v
Serdar	88745	an	62	40	13	v
Tülay	94335	iz	25	5	4	y
Selcan	44568	an	30	10	7	y

Ad	PersonelNo	yas	kilo	boy	Eğitim Durumu
Ayşe	15798	20	54	163	Master
Ali	87962	40	72	179	Üniversite
Veli	72596	25	78	192	Master
Gamze	25987	33	56	165	Lise
Murat	25896	45	89	189	Üniversite
Deniz	18796	55	65	173	Lise
Sevgi	25896	31	61	170	Üniversite
Nazlı	10259	24	58	160	Lise
Ayça	48975	61	64	169	Üniversite
Suna	15796	37	65	168	Lise
Doğan	88964	45	83	197	Master
Kerem	58796	51	79	186	Üniversite
Yeliz	57964	49	62	171	Lise
Canan	15976	39	64	153	Master
Ege	78921	30	80	182	Üniversite
Barış	45634	47	75	179	Lise
Mert	89314	29	92	197	Yüksek Okul
Yeliz	72698	32	59	165	Master
Esra	89574	53	63	168	Üniversite
Tuna	92687	27	87	176	Doktora
Esra	91236	23	55	159	Lise
Selen	72397	62	64	164	Üniversite
Tolga	56987	37	73	178	Üniversite
Hasan	39211	20	81	186	Yüksek Okul
Burçin	55987	54	57	167	Üniversite
Büşra	26674	52	59	173	Üniversite
Serdar	88745	62	93	190	Üniversite
Tülay	94335	25	67	175	Yüksek Okul
Selcan	44568	30	55	171	Yüksek Okul

Bu veri setlerindeki bazı alanlar ortak, ancak bir veri setinde olan alanlar diğerinde yok.

Bu iki veri setini birleştirip, yeni bir veri seti oluşturma işlemi veri birleştirmedir. Bazen aynı alanlara sahip veri setlerinin dikey olarak birleştirilmesi gerekebilir. Bu da veri hacmini arttırmak için yapılan birleştirmedir. Burada dikkat edilmesi gereken aynı alanlara (sütun başlıkları) sahip verilerin dikey olarak birleştirilebileceğidir.

Ad	PersonelNo	sehir	yas	tecrube	gelir	ev	kilo	boy	Eğitim Durumu
Ayşe	15798	is	20	1	3	y	54	163	Master
Ali	87962	is	40	18	10	v	72	179	Üniversite
Veli	72596	is	25	7	6	y	78	192	Master
Gamze	25987	is	33	10	7	v	56	165	Lise
Murat	25896	is	45	25	8	v	89	189	Üniversite
Deniz	18796	is	55	36	17	v	65	173	Lise
Sevgi	25896	is	31	12	8	y	61	170	Üniversite
Fırat	10259	is	24	2	3	y	58	160	Lise
Ayça	48975	is	61	25	15	v	64	169	Üniversite
Suna	15796	an	37	14	7	v	65	168	Lise
Doğan	88964	an	45	15	18	y	83	197	Master
Kerem	58796	an	51	26	7	y	79	186	Üniversite
Yeliz	57964	an	49	28	10	v	62	171	Lise
Canan	15976	an	39	20	6	v	64	153	Master
Ege	78921	an	30	10	4	y	80	182	Üniversite
Barış	45634	iz	47	28	9	v	75	179	Lise
Mert	89314	iz	29	8	6	y	92	197	Yüksek Okul
Yeliz	72698	iz	32	15	9	v	59	165	Master
Esra	89574	iz	53	25	13	y	63	168	Üniversite
Tuna	92687	iz	27	10	5	y	87	176	Doktora
Esra	91236	iz	23	6	4	y	55	159	Lise
Selen	72397	iz	62	40	12	v	64	164	Üniversite
Tolga	56987	is	37	17	15	v	73	178	Üniversite
Hasan	39211	an	20	2	2	y	81	186	Yüksek Okul
Burçin	55987	iz	54	35	10	v	57	167	Üniversite
Büşra	26674	is	52	33	15	v	59	173	Üniversite
Serdar	88745	an	62	40	13	v	93	190	Üniversite
Tülay	94335	iz	25	5	4	y	67	175	Yüksek Okul
Selcan	44568	an	30	10	7	y	55	171	Yüksek Okul

Sonuçta ortaya istediğimiz bütün alanları ve kayıtları kapsayan birleşik bir set çıkar. Birleştirilecek veriler değişik kaynak ve yapılardan gelebilir. 2 ve daha fazla sayıda veri kaynağı veya seti birleştirilebilir. Birleştirme esnasında olası tekrar ve karışıklıkların önlenmesi gerekir.

VERİ DÖNÜŞÜMÜ

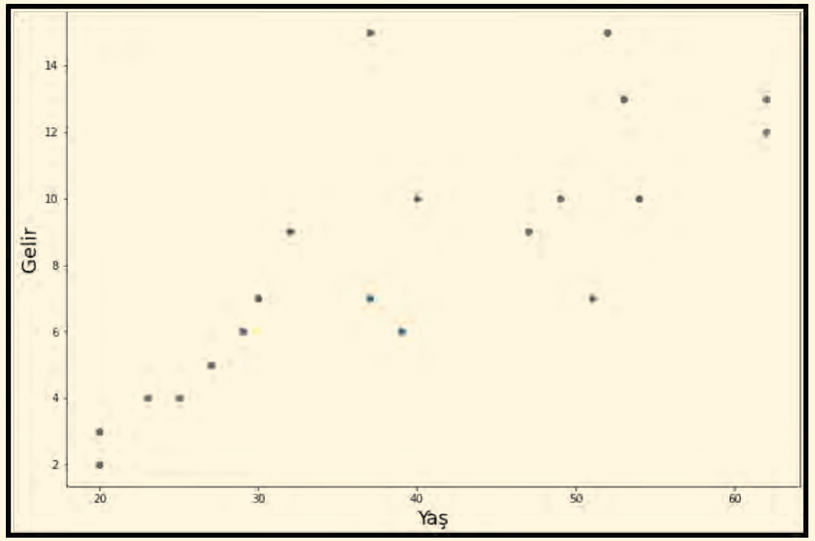
Verilerin özetlenmesi, normalleştirilmesi, genelleştirilmesini içeren birtakım dönüşümleri ifade eder. Örneğin farklı veri alanlarının ölçeklerini aynı yapıya getirmek için Normalizasyon teknikleri uygulanarak veri değerleri ortak bir zeminde karşılaştırılabilir hale getirilir. Böylece yüksek değerli verilerin, çok küçük değerli verilere baskınlığı engellenebilir. Özellikle mesafe hesaplamalı algoritmalarda normalleştirme son derece önemlidir. Min-max, Z-Score gibi pek çok normalleştirme teknikleri mevcuttur.

MIN-MAX NORMALLEŞTİRMESİ

Normalleştirmeye örnek olarak Maksimum – Minimum normalleştirmesi verilebilir. Özellikle alan değerleri arasında farklar olan veri setleri ile işlem yapılırken Normalleştirme yapmak son derece önemlidir. Aksi takdirde büyük alan değerlerine sahip olan veriler, daha küçük değerlere sahip olan alanlardaki değerleri etkisiz, önemsiz hale getirebilir. Özellikle mesafe algoritmalarına dayanan hesaplamalarda bu durum son derece kritiktir. Büyük değerlerin küçükleri "ezmemesi" için alanlar ölçeklemeyle aynı zemine getirilir. Ölçekleme her alanın kendi değerleri içinde yapılır. Böylece veri dağılımı değişmez, ancak her alan belirlenmiş yöntemle aynı ölçeğe getirilir.

Örneğin şu veri setine bakacak olursak:

Yaş	Gelir
20	3
40	10
37	7
51	7
49	10
39	6
47	9
29	6
32	9
53	13
27	5
23	4
62	12
37	15
20	2
54	10
52	15
62	13
25	4
30	7



Yaş alanına ait veriler "Gelir" alanınınkini bazen 6-7 katı kadar büyük olabiliyor.

Bu durum mesafeye dayalı bir algoritma hesaplamasında "Yaş" alanının daha baskın gelmesini sağlayacaktır. Çünkü bu iki alan herhangi bir "norm" a göre ölçekli değildir.

Bu verilerin dağılım grafiğine bakacak olursak Yaş-Gelir arasındaki grafik dağılımı bu şekilde. Ancak yaş alanı bariz bir şekilde gelir alanının değerlerinden fazla.

Bunun için her iki alanı da ortak zeminde buluşturmak için kendi içinde bir ölçeklemeye yani normalleştirmeye gitmek gerek.

Yukarıda da belirtiğim gibi pek çok normalleştirme metodu vardır. Bunlardan en yaygın kullanılanı maksimum minimum normalleştirmesidir.

Bu ölçeklemenin formülü:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Buna göre veri setindeki herhangi bir değer ile o veri setindeki en küçük değer arasındaki fark alınır ve o veri setinin maksimum ve minimum farkında bölünür.

Max-Min normalleştirmesinde bir alandaki değerler 0-1 arasında ölçeklenir.

Örneğin yaş alanına max-min normalleştirmesi uygulandığında şu değerler oluşur:

Max-Min Normalleştirmesi

Yaş	
20	0.000000
40	0.476190
37	0.404762
51	0.738095
49	0.690476
39	0.452381
47	0.642857
29	0.214286
32	0.285714
53	0.785714
27	0.166667
23	0.071429
62	1.000000
37	0.404762
20	0.000000
54	0.809524
52	0.761905
62	1.000000
25	0.119048
30	0.238095

Yaş alanındaki en küçük değer 20, en büyük değer 62. En küçük değer olan 20 sıfıra çevrilmiş, 62 değeri ise 1 olarak çevrilmiştir. Diğer bütün değerler bu aralıkta oluşmuştur.

Aynı şekilde "Gelir" alanını da max-min normalleştirmesine tabii tutarsak:

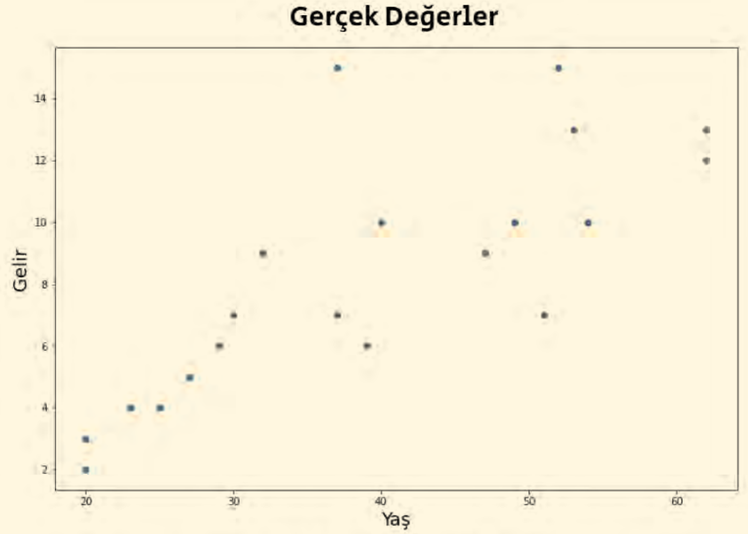
Max-Min Normalleştirmesi

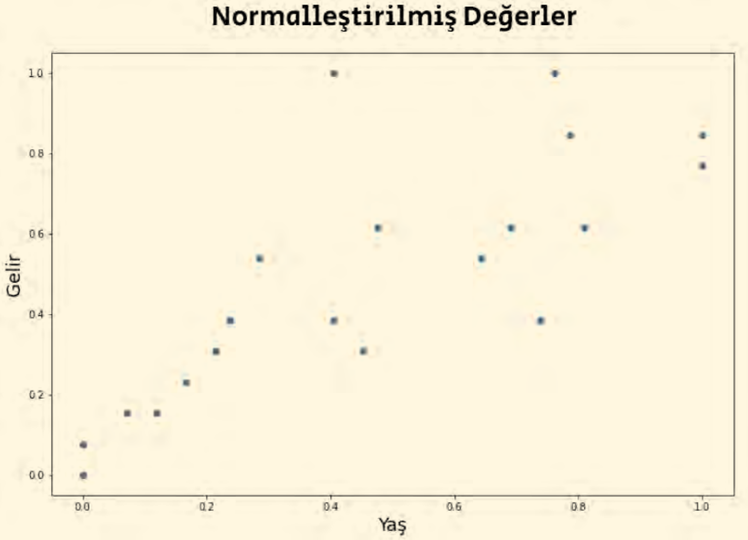
Gelir	
3	0.076923
10	0.615385
7	0.384615
7	0.384615
10	0.615385
6	0.307692
9	0.538462
6	0.307692
9	0.538462
13	0.846154
5	0.230769
4	0.153846
12	0.769231
15	1.000000
2	0.000000
10	0.615385
15	1.000000
13	0.846154
4	0.153846
7	0.384615

Burada da en küçük değer 2 ve yüksek değer 15. 2 değeri sıfıra, 15 değerleri de 1'e çevrilmiş. Veriler 1 ile 0 arasında ölçeklenmiş.

Bu normalleştirmelerle hem yaş hem de "gelir" alanları aynı zemine getirilmiştir. Böylece artık ikisi arasında özellikle mesafe ölçümlerine dayanan algoritmalarla işlemler yapılabilir.

Bu normalleştirmelerle veri dağılımında bir değişiklik olmamaktadır. Veriler aynen orijinal değerlerindeki gibi bir dağılım gösterirler. Sadece ölçekleri değişmiştir:



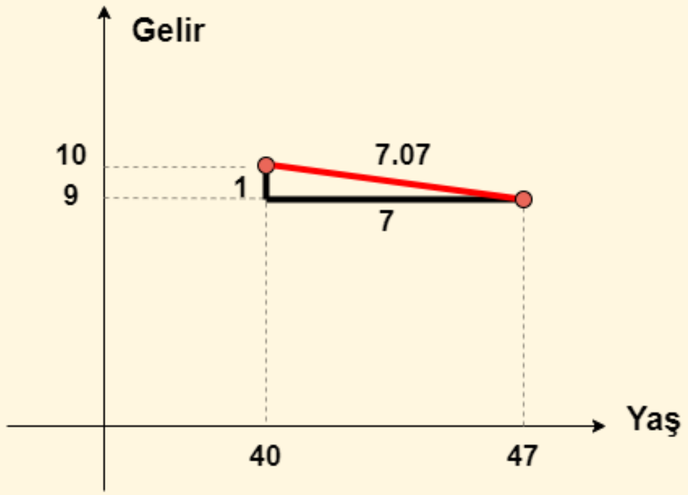


Gördüğünüz gibi dağılımda en ufak bir farklılık yok.

NORMALLEŞTİRMENİN FAYDALARI

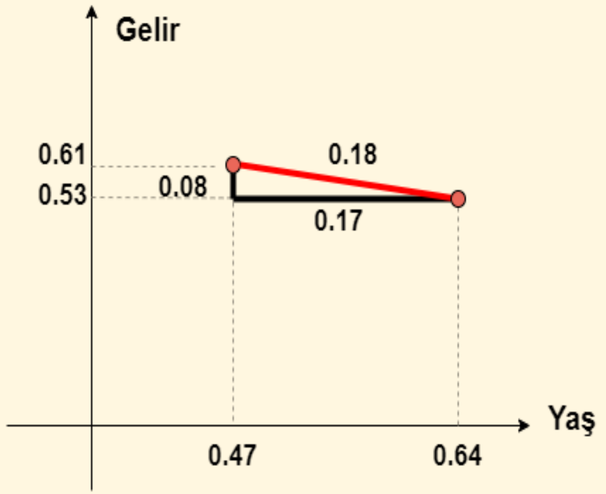
Yukarıda da belirttiğim gibi mesafeye dayalı algoritmalarda normalleştirme doğru uzunlukları, modellemeleri yapmak için gereklidir. Mesafeye dayalı algoritmalar, kümeleme, en yakın komşular, destek vektör makineleri gibi makine öğrenimi algoritmalarıdır. Bu algoritmalar herhangi iki veya daha fazla noktanın bir birlerine göre mesafesine bakarak belirli modeller oluştururlar. Buradaki alanların aynı zeminde normalleştirilmemiş olması, yukarıda bahsettiğim bir alanın diğerini "ezmesini" sağlayabilir. Yukarıdaki veri setimizden örnek verecek olursak. Yaş alanı, gelir alanı nicelikleri arasında 5-6 kata varan farklar var. Eğer bu alanları yukarıdaki gibi normalleştirmeden modellemeye çalışırsak birazdan örnekleyeceğim farklar çıkar.

Gerçek değeri 40'a 10 olan yani yaşı 40 ve geliri 10 olan nokta ile yaşı 47 ve geliri 9 olan iki veri noktası arasındaki mesafeyi hesaplarsak;



Öklid-Pisagor bağlantısına göre birin karesi ve 7 nin karesinin toplamının karakökü hesabına göre 7.07 gibi bir mesafe çıkar. Bu değer baskın olan yaş değeri farkı olan 7 değerine çok yakındır. Yani burada aslında "Gelir" alanından gelen değer, mesafe hesaplamasında nerdeyse hiç rol oynamamış. Mesafe bu değerin 7 katından fazla.

Şimdi yukarıdaki şekilde normalleştirilmiş verilerimizle mesafe hesaplaması yapalım:



Aynı verilerin normalleştirilmiş değerlerine göre yapılan hesaplamada bu sefer "Gelir" alanının mesafeye olan etkisi çok daha fazla olmuştur. "Yaş" alanı mesafeye yaklaşık yine aynı oranda etki ederken, "Gelir" alanı bu sefer mesafenin yarısından az bir etkide bulunmuş. Oysa normalleştirilmemiş değerlerde bu oran 7 de biriydi.

Alanlar arasındaki farkların daha da fazla olduğu durumlarda normalleştirme yapılmadığı takdirde hata payı da daha fazla olacaktır. Bazı veri setlerinin alanları arasında 10,100 katı hatta çok daha fazla nicelik farkı bulunmaktadır.

VERİ GÖRSELLEŞTİRME

Keşifsel veri analizinin en yaygın kullanılan metotlarından biri de verilerin Görselleştirilmesidir. Veri görselleştirme elbette sadece keşifsel veri analizinde değil genel anlamda yaygın kullanılan bir yöntemdir.

Görselleştirme verilerin anlaşılması için en etkin yöntemdir. Sadece rakamlardan, kategorilerden oluşan metinsel ifadeler veri hakkında çok fazla bir içgörü sunmazlar. Buna karşılık görselleştirilen veriler bize çok fazla bilgi verir.

Günümüzde veriler oldukça zengin alternatiflerle görselleştirilebiliyor. Bunun için pek çok yöntem ve görsel vardır. Devam eden kısımlarda veri görselleştirmede en yaygın kullanılan yöntemleri inceleyeceğiz.

Veri grafiklerini temel olarak 5 Grupta toplayabiliriz.

- Karşılaştırma Grafikleri
  - Çizgi
  - Bar
  - Radar
- İlişki Grafikleri
  - Scatter
  - Balon
  - Correlogram
  - Sıcaklık Haritaları
- Kompozisyon Grafikleri
  - Pasta Grafiği
  - Donut Grafiği
  - Yığın Bar Grafikler
  - Yığın Alan Grafikleri
  - Venn Diagram
- Dağılım Grafikleri
  - Histogram
  - Yoğunluk
  - Kutu
  - Keman
- Coğrafi Grafikler
  - Nokta Harita
  - Coropleth Harita
  - Bağlantı Harita

KARŞILAŞTIRMA GRAFİKLERİ

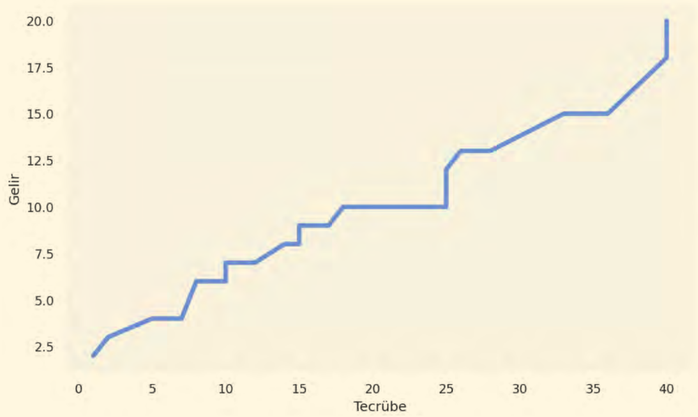
Karşılaştırma grafikleri değişkenlerin birbirine göre durumlarını karşılaştırmak veya zaman içindeki değişimlerini göstermek için kullanılır. Çizgi grafikler değişkenlerin zaman içindeki değişimini görselleştirmek için idealdir. Öğeler arasında karşılaştırma yapmak için bar grafikler iyi bir yöntemdir. Radar grafikleri, değişkenlerin durumlarını aynı eksen üzerinde daha net görünecek şekilde karşılaştırma sunarlar.

ÇİZGİ GRAFİKLER

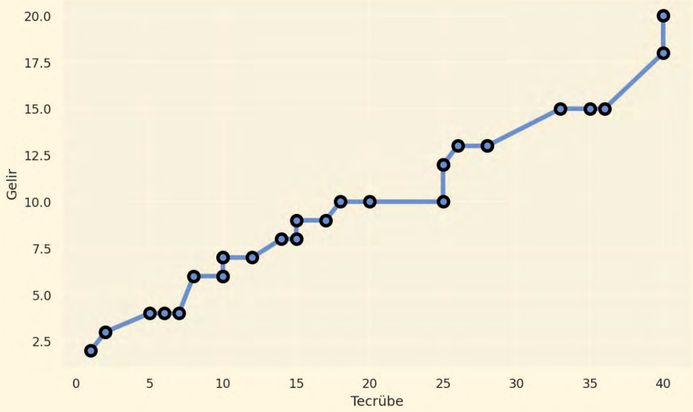
Çizgi Grafikler sürekli nicel verileri görüntülemek için kullanılır, veri noktalarını birbirine çizgilerle bağlarlar. En yaygın kullanılan grafik tiplerindendir.

Örneğin "Tecrübe" ile "Gelir" verileri arasındaki çizgi grafiği çizmek istersek:

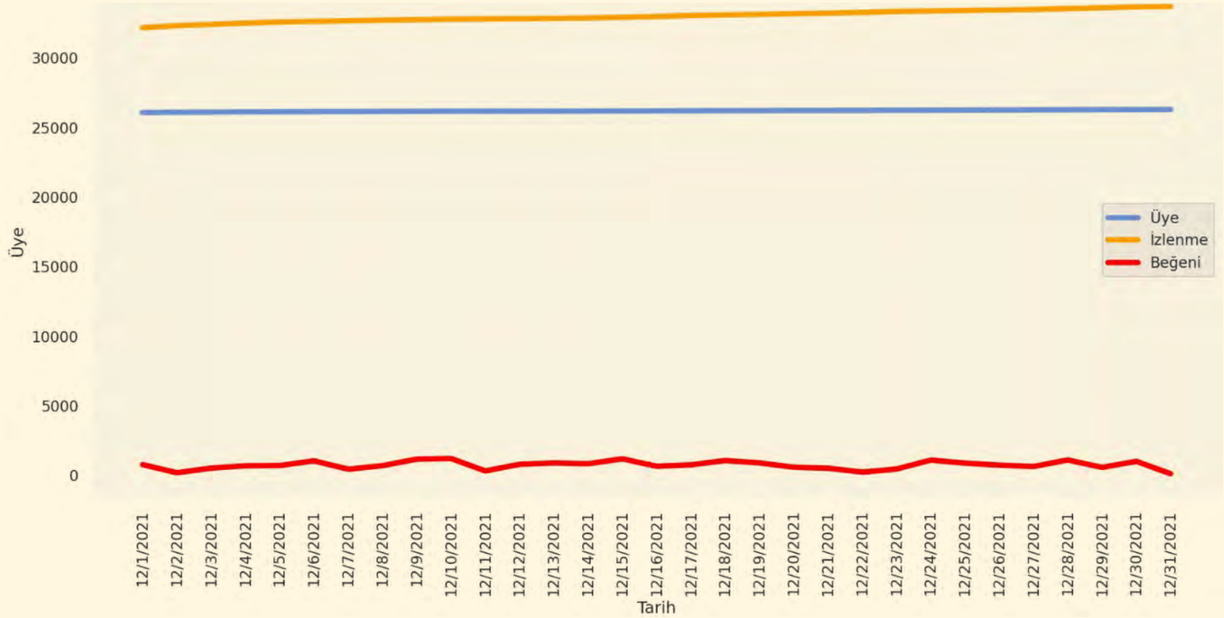
	tecrübe	gelir
0	1	3
1	18	10
2	7	6
3	10	7
4	25	8
5	36	12
6	12	8
7	2	3
8	25	15
9	14	7
10	15	18
11	26	7
12	28	10
13	20	6
14	10	4
15	28	9
16	8	6
17	15	9
18	25	13
19	10	20
20	6	4



Çizgi grafiklerde bazen veri noktaları da gösterilir:



Dilerseniz birden fazla değişkeni de grafik üzerinde gösterebilirsiniz. Böylece karşılaştırmalı veri analizi yapma imkanınız oluşur:



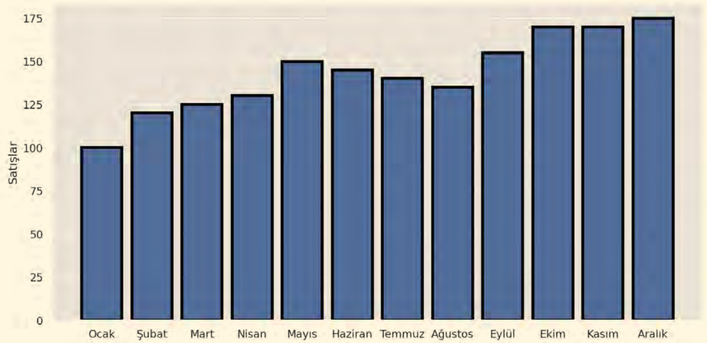
10 ve daha fazla sürekli nicel veriyi görüntülemek için çizgi grafikler iyi bir seçimdir. Daha düşük sayıdaki verileri görüntülemek için "bar" grafikler daha iyi bir çözüm olabilir.

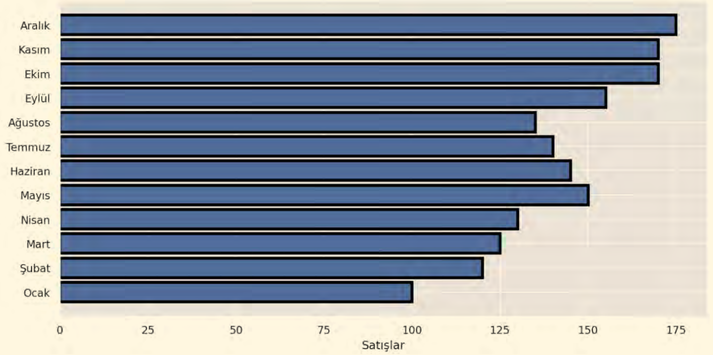
BAR GRAFİKLER

Değerlerin dikdörtgen çubuklarla gösterildiği grafik tipleridir. Yatay veya dikey olabilir. Kategorik verilerin değerlerini göstermek için kullanılır.

Aylar	Satış
Ocak	100
Şubat	120
Mart	125
Nisan	130
Mayıs	150
Haziran	145
Temmuz	140
Ağustos	135
Eylül	155
Ekim	170
Kasım	170
Aralık	175

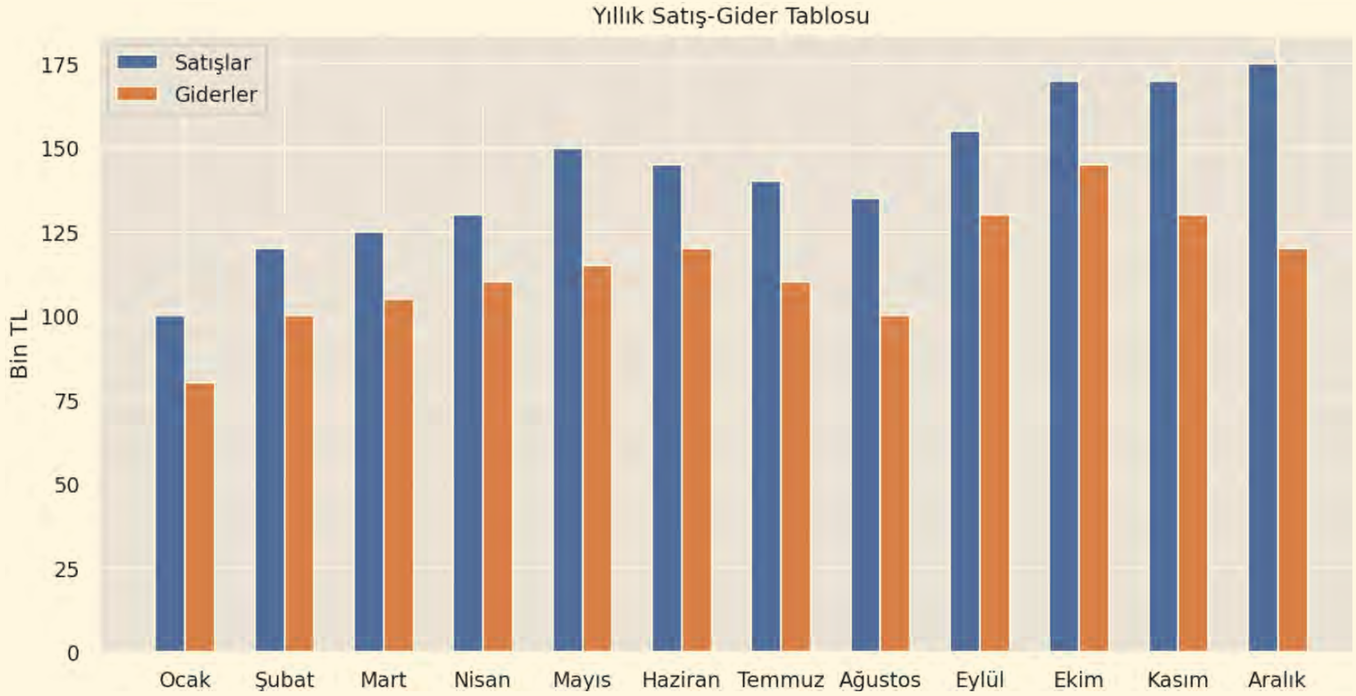
Elimizde yandaki gibi bir veri seti olsun. Bu veri setinde satış değerleri aylık bazda gösteriliyor. Bu veri setini hem dikey hem de yatay bar grafiğinde göstermek gerekirse aşağıdaki grafikleri elde ederiz.





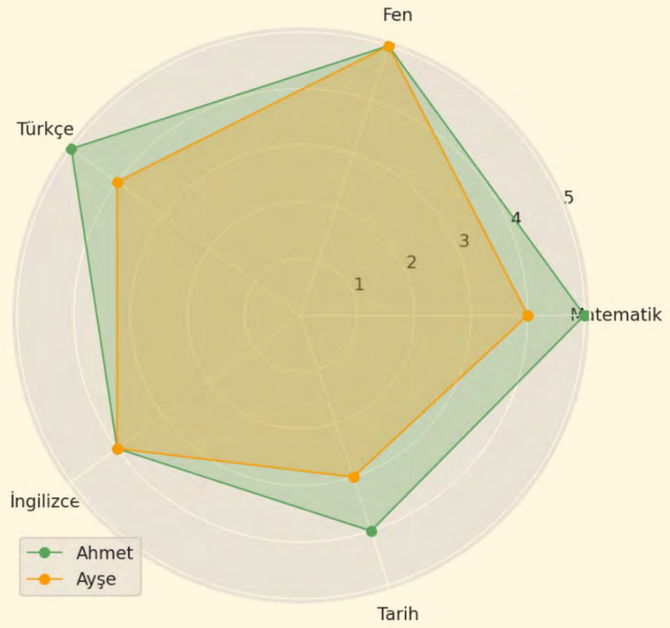
Bar grafiklerini ileride göreceğimiz histogram grafikleriyle karıştırmamak gerekir. Bar grafikler farklı kategorileri veya değişkenleri karşılaştırırken, histogramlar tek bir değişkenin dağılımını gösterir. Diğer yaygın yapılan bir hata gruplar ya da kategoriler arasındaki merkezi eğilimleri göstermek için bar grafikleri kullanmaktır. Bu istatistiksel ölçümleri veya dağılımları göstermek için kutu veya keman grafikler gibi başka görselleştirme teknikleri kullanılır.

Bar grafikleri karşılaştırma amaçlı da kullanabilirsiniz:



RADAR GRAFİKLERİ

Örümcek ya da Ağ grafikleri olarak da bilinir. Bu grafikte değişkenler(kategoriler) bir daire etrafında sıralanır ve değerler dairenin merkezinden değişkenlere doğru eşit biçimde ölçeklenir. Her bir değişkenden diğerine düz bir çizgi çizilerek çokgenler oluşturarak değerler karşılaştırılır.



Örneğin bu grafikte 5 farklı ders için Ahmet ile Ayşe'nin notları karşılaştırılıyor. Notlar 5 üzerinden. Daire merkezden itibaren 5 eşit parçaya bölünmüş. Bunlar not değerlerini temsil ediyor ve her ders için bir not mevcut. Buna göre Ahmet ile Ayşe'nin notları aynı eksen üzerinde karşılaştırmalı olarak görülebiliyor.

Radar grafikleri birden çok nicel değişkeni karşılaştırmak için çok idealdir.

İLİŞKİ GRAFİKLERİ

Veriler arasındaki ilişkileri göstermek için uygulanan görselleştirme metotlarıdır.

SCATTER PLOT

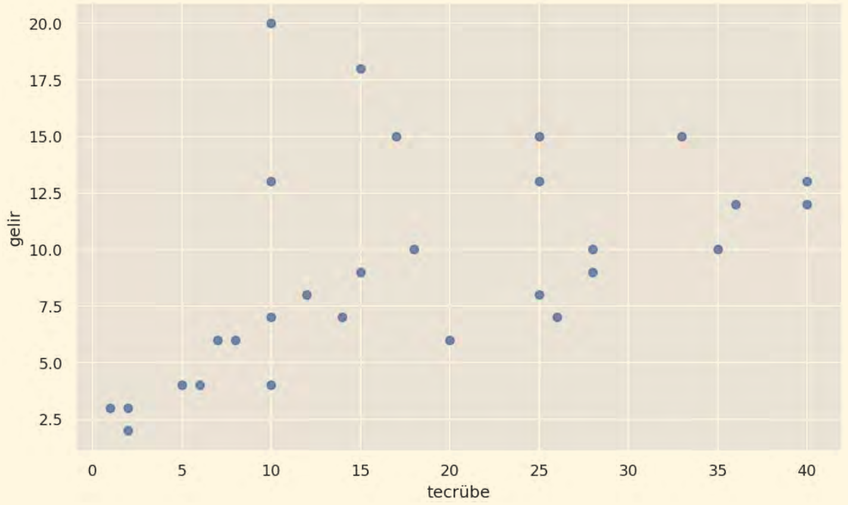
İki eksene dayalı sayısal veri noktalarını gösteren grafiklerdir. İki veri arasında bir ilişki olup olmadığını anlayabilirsiniz. Scatter grafiklerinde Farklı renkler kullanıp gruplar arasındaki ilişkileri de görselleştirebilirsiniz. Scatter grafiklerinin bir versiyonu olan kabarcık (balon) grafikleri de özellikle nicelikle ilgili bir boyutu da devreye sokar. Böylece baloncukların boyutuna göre, ilişkiler arasındaki niceliksel bağlantılar kolayca gözlemlenebilir.

Özetle iki değişken arasındaki verileri görselleştirebilen scatter grafikleri, renklerin ve boyutların devreye girmesiyle 4 boyutlu bir hale gelebilir. Önce iki boyutlu bir scatter grafiği çizelim, verilerimiz şunlar olsun:

tecrübe	gelir
1	3
2	2
2	3
5	4
6	4
7	6
8	6
10	20
10	4
10	13
10	7
12	8
14	7
15	9
15	18

tecrübe	gelir
17	15
18	10
20	6
25	15
25	8
25	13
26	7
28	9
28	10
33	15
35	10
36	12
40	12
40	13

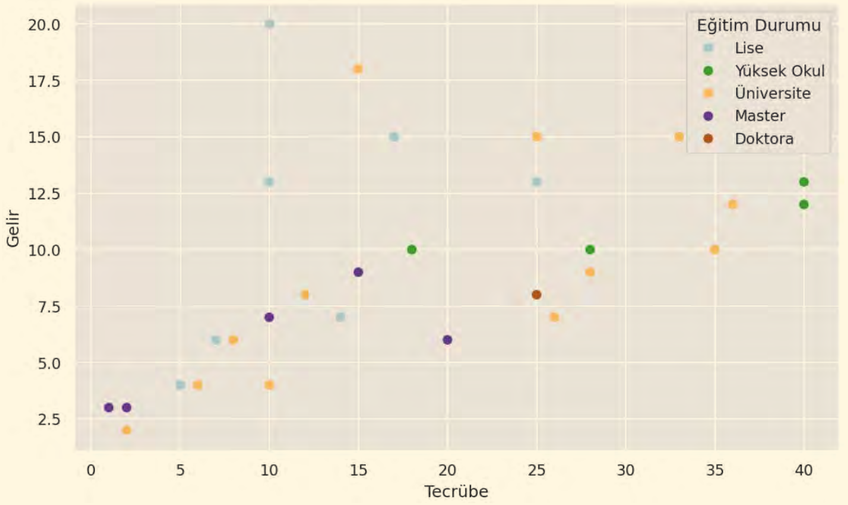
Yandaki tecrübe ve gelir verilerinin arasındaki ilişkiyi daha net görebilmek için scatter grafiğini çizebiliriz:



Bu grafikten tecrübe arttıkça gelirin de arttığın çok daha net gözlemleyebiliyoruz.

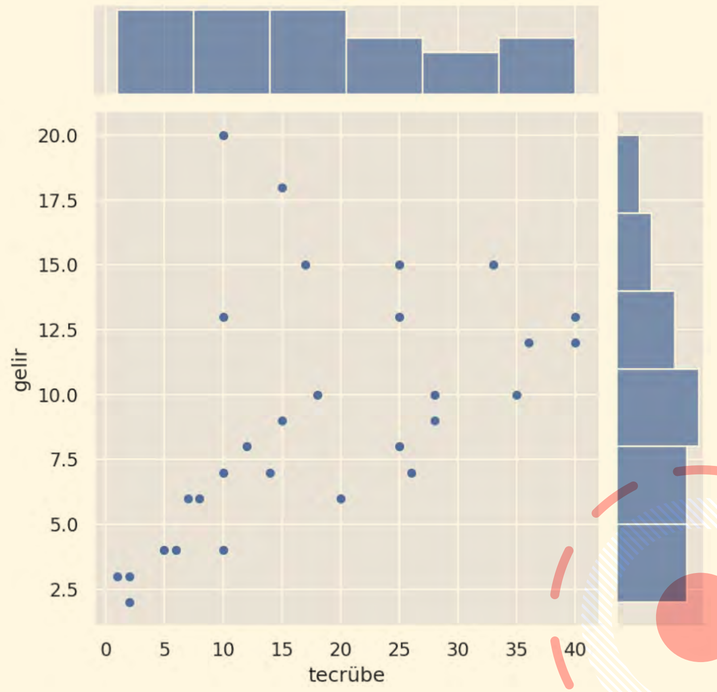
Eğer bu ilişkide başka bir değişkeni de kullanmak istersek, örneğin tecrübe ile gelir dağılımı arasındaki ilişkide "Eğitim durumu" dağılımı nasıldır diye merak edersek o zaman, eğitim durumlarını bu dağılımda farklı renklerde gösterebiliriz:

Böylece grafiğe 3. Bir boyut ekleyip, tecrübe-gelir arasındaki ilişkiyi gözlemleyebildiğimiz gibi, hangi eğitim grubunun hangi noktalarda olduğunu da görebiliriz.



BELİRLİ ARALIK DAĞILIMLARI (HISTOGRAM) İLE SCATTER GRAFİKLERİ

Scatter grafiklerindeki her bir değişkenin belirli aralıktaki dağılımlarınıda gösteren entegre grafiklerde vardır. Yine kendi verileri setimizden gidecek olursak bu grafiklerin görünümü aşağıdaki gibidir:



Örneğin bu grafikte 5 ile 8 gelir aralığında 6 nokta varken, 8 ile 10 arasında 7 nokta vardır. Buna göre sağ taraftaki histogram oluşmaktadır.

Diğer bir değişken olan tecrübede 20-26 arasında 4 nokta varken 28-33 arasında 3 nokta mevcut.

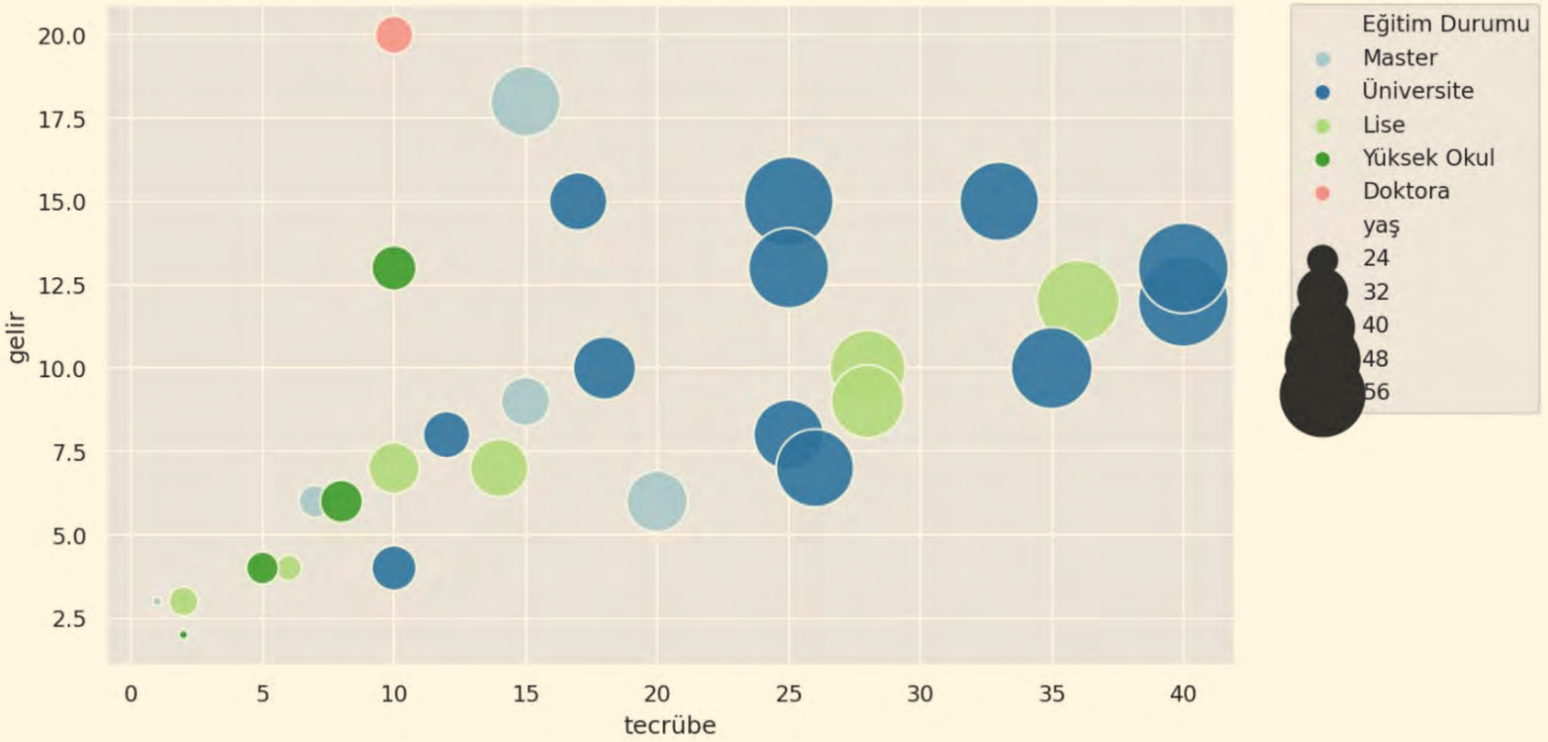
KABARCIK (BALONCUK) GRAFİKLERİ

Scatter grafiklerine çok benzeyen kabarcık grafiklerinde niceliksel değerleri gösteren 4. Bir boyutta ekleyebilirsiniz. Yukarıdaki örneğimize "yaş" alanın da eklendiğin düşünelim:

tecrübe	gelir	yaş	Eğitim Durumu
1	3	20	3
2	2	20	1
2	3	24	0
5	4	25	1
6	4	23	0
7	6	25	3
8	6	29	1
10	20	27	4
10	4	30	2
10	13	30	1
10	7	33	0
12	8	31	2
14	7	37	0
15	9	32	3
15	18	45	3

tecrübe	gelir	yaş	Eğitim Durumu
17	15	37	2
18	10	40	2
20	6	39	3
25	15	61	2
25	8	45	2
25	13	53	2
26	7	51	2
28	9	47	0
28	10	49	0
33	15	52	2
35	10	54	2
36	12	55	0
40	12	62	2
40	13	62	2

Buna ait kabarcık grafiği şu şekildedir:

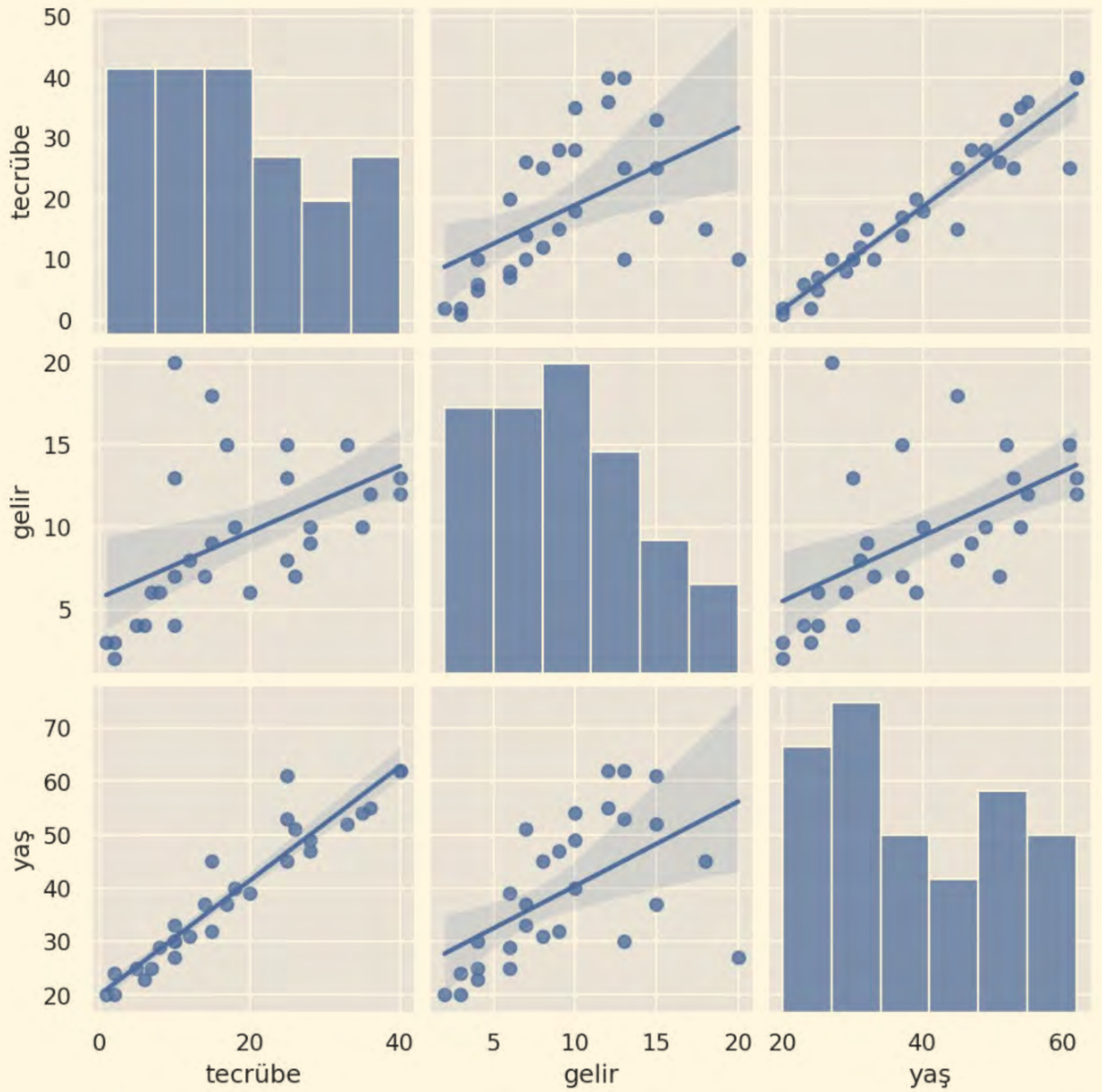


Gördüğünüz gibi nokta büyüklükleri yaş gruplarına göre farklılık gösteriyor. Böylece "tecrübe-gelir" ilişkisini görüntülerken aynı zamanda eğitim durumu ve "yaş" arasındaki ilişkilerle olan bağlantısını da sadece bir grafikte gözlemleyebiliyoruz.

CORRELOGRAM – KORELASYON MATRİSLERİ

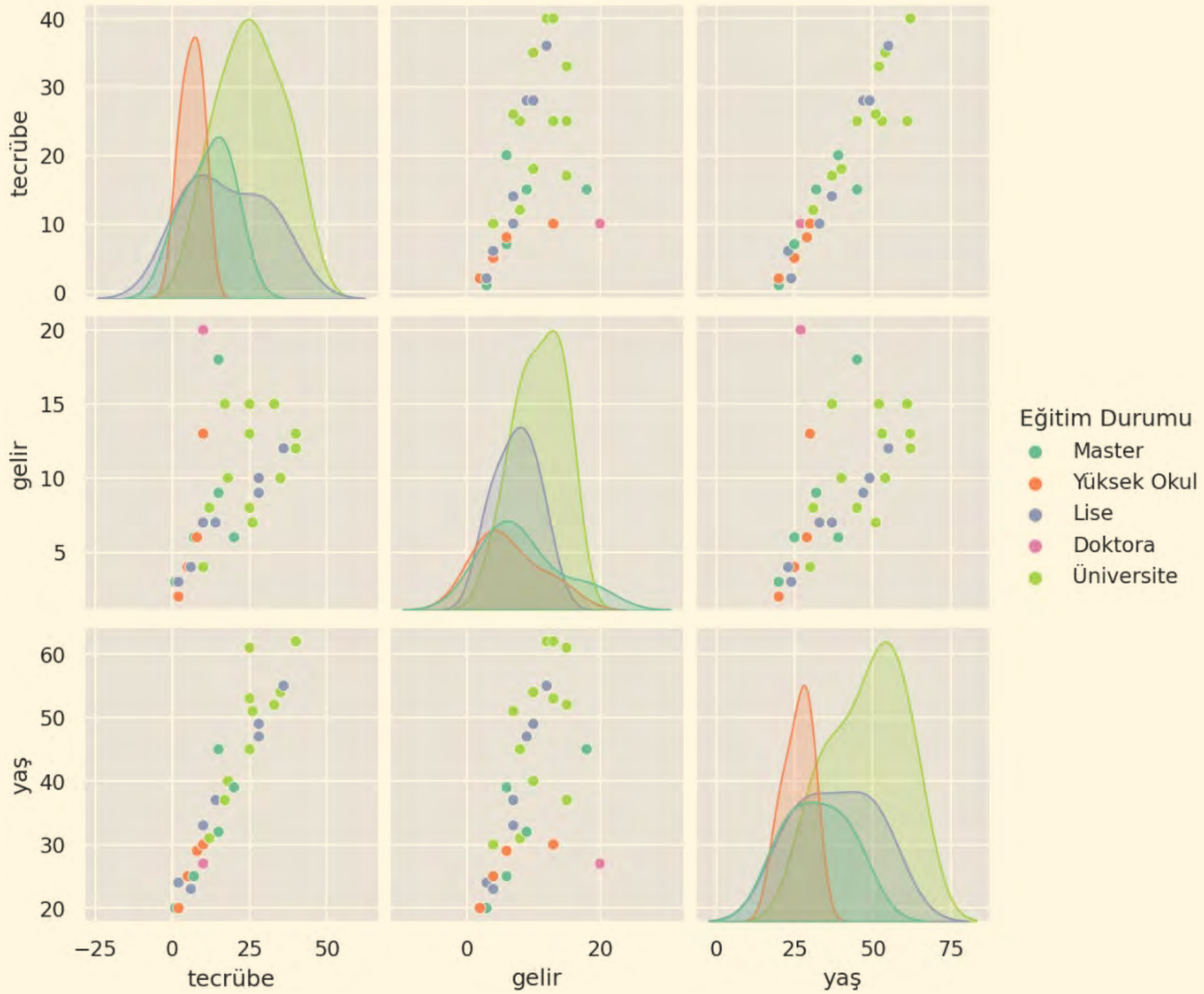
Bir korelasyon matrisi, İlişki grafikleri ile (Scatter) ileride göreceğimiz Histogram (dağılım grafikleri) grafiklerinin matris şeklinde sunulduğu grafik tipleridir.

Özellikle keşifsel veri analizinde oldukça yaygın kullanılırlar. Yukarıda yaptığımız scatter grafiklerinin her bir değişkene göre aynı anda gösterebilirler. Sayısal değerlerin olduğu alanları grafiklerler. Çaprazda kendi alanı ile eşleştiği yerlerde dağılım (histogram) grafiğini gösterirler:



Örneğin bu correlogram da x ekseni doğrultusunda “tecrübe”, “gelir”, “yaş” alanları alınmış, y ekseni boyunca da yin “yaş”, “gelir” ve “tecrübe” alanları alınmış. Böylece her bir alanın birbirine göre grafiği çizilmiş. Örneğin x eksenindeki tecrübe alanı ile y eksenindeki “yaş” alanının grafiği görüldüğü gibi yine y ekseninde “tecrübe” ye ait “gelir” grafiği de görülebilmekte. Böylece bütün alanlara ait ilişkisel dağılımlar izlenebilmektedir. X ve y ekseninde aynı alan olduğu zaman o alana ait değer dağılımları (histogramlar) gösterilmektedir. Örneğin “yaş” alanının histogramı x ve y eksenindeki “yaş” , “yaş” alanlarının kesişimde görülmektedir. Başka bir deyişle matrisin sol üst den sağ alt çaprazına kadar olan diagonel Histogramlardır.

Eğer dilersek renk faktörünü de katarak aynen Scatter grafiklerinde olduğu gibi kategorik bir alanın bu dağılımdaki durumunu gözlemleyebiliriz:



Bu tip matrisler makine öğrenimi öncesi keşifsel veri analizi konusunda çok hızlı bir bakış açısı ve içgörü sunarlar.

HEATMAP

Isı haritaları kategorik değişkenlerin satırlara ve sütunlara yerleştirildiği, sayısal değerlerin renk dolgunluğu olarak temsil edildiği grafik tipleridir. Bu yapısıyla matrissel dağılımların görselleştirilmesinde oldukça başarılıdır. Matrisin her bir değeri açıktan koyuya doğru renk doygunlukları ile temsil edilir.

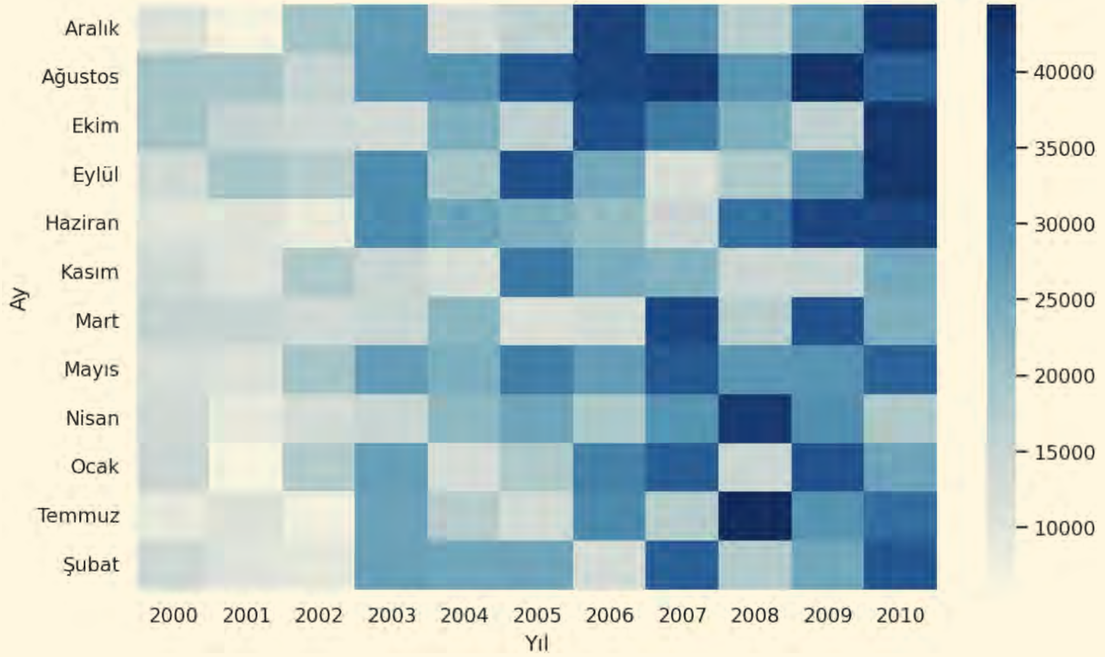
Örneğin yıllar ve aylara göre satış verilerinin tutulduğu şöyle bir veri setimiz olsun:

	Yıl	Ay	Satışlar
0	2000	Ocak	14999
1	2000	Şubat	14057
2	2000	Mart	15923
3	2000	Nisan	13852
4	2000	Mayıs	12458
...	...	...	...
127	2010	Ağustos	36563
128	2010	Eylül	42829
129	2010	Ekim	42574
130	2010	Kasım	24951
131	2010	Aralık	41984

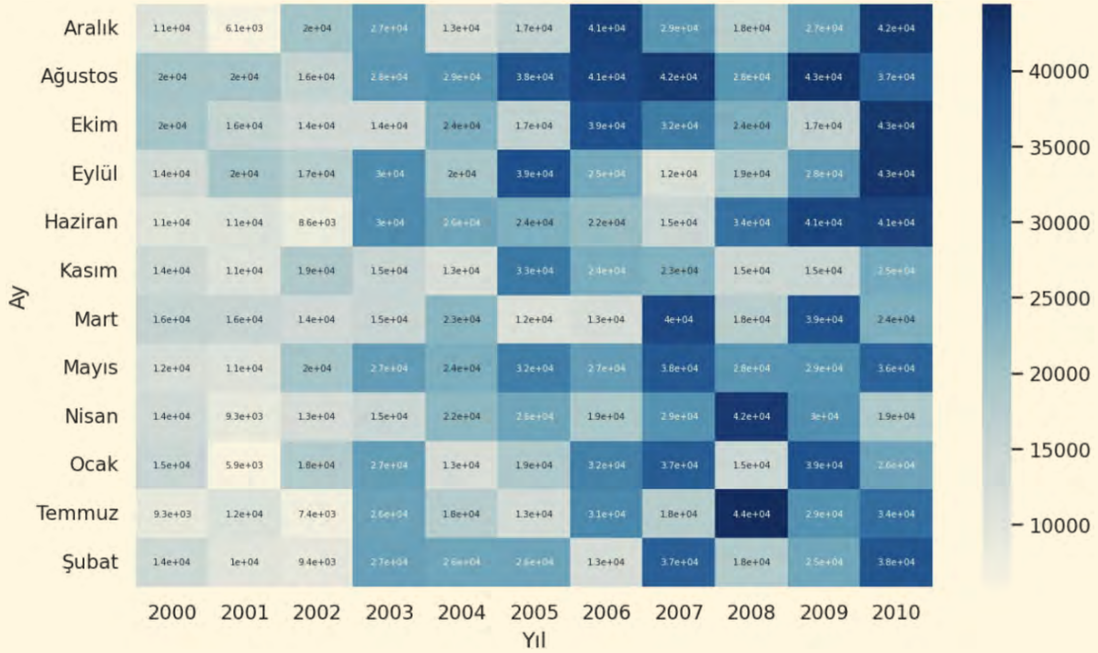
Bu verilerin özet tablosu(Pivot) şu şekildedir:

Yıl	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Ay											
Aralık	11105	6115	20058	27365	12973	16648	41113	28572	17545	26802	41984
Ağustos	19995	19723	15899	28288	28758	38236	40655	41876	28489	42898	36563
Ekim	19733	15771	13992	13957	23942	16608	38920	32357	24001	16701	42574
Eylül	13779	19710	17477	29808	19702	38888	25067	12277	18780	28143	42829
Haziran	11130	11407	8604	29997	25851	24049	21593	15128	33862	40610	40842
Kasım	14155	11376	18626	15273	12536	33083	24472	23393	14879	15104	24951
Mart	15923	15508	13971	15276	22760	11797	13057	40227	17613	38618	24072
Mayıs	12458	11351	19599	27103	24175	31983	27124	37596	28303	28877	35942
Nisan	13852	9326	13026	15140	22368	25628	19058	29040	42153	29545	18778
Ocak	14999	5887	18163	26929	13084	18722	31663	36793	14990	38560	26015
Temmuz	9327	11657	7350	26075	17630	13462	30542	17787	44381	29019	34381
Şubat	14057	10354	9440	26542	25644	25865	13098	36550	18239	25469	38043

Buradaki verilerin niceliklerini gözle takip etmek oldukça zor. Oysa bir ısı haritası bu işi çok daha kolay hale getirir:



Burada daha koyu renkler yüksek değerleri temsil etmektedir. Haritanın ölçeği sağ tarafta gözükmektedir. Bu şekilde bir görselleştirme satış değerlerinin yüksek olduğu alanları çok daha etkin bir şekilde gösterecektir. Eğer dilersek renklerle birlikte değerleri de alanlarda gösterebiliriz:

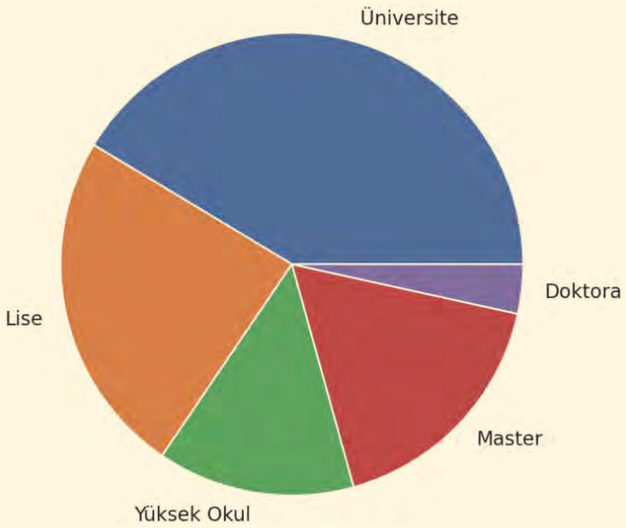


KOMPOZİSYON (BİLEŞİM) GRAFİKLERİ

Verileri bir bütünün parçası olarak düşünüyorsanız bileşim (kompozisyon) grafikleri ideal görselleştirme araçlarıdır. Devam dene başlıklarda bileşim grafiklerini inceleyeceğiz.

PASTA GRAFİKLERİ

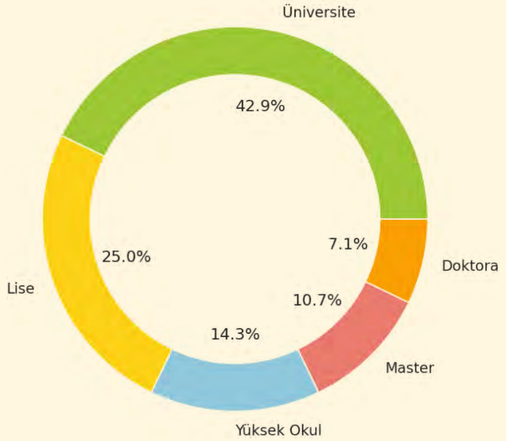
Bir daireyi dilimlere bölerek sayısal oranları gösterir. Her dilim bir kategorinin bir oranını temsil eder. Tam daire %100'e eşittir.



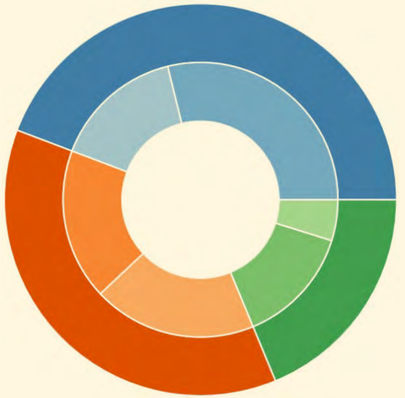
Pasta grafikleri bir bütünün parçalarını kıyaslamak için kullanılırlar.

DONUT GRAFİĞİ

Pasta Grafiğinin alternatifi olarak "Donut" grafiği de kullanılabilir. Bu grafik de yine bir bütünün parçalarını birbirine göre oransal olarak sunar. Bütün 100 üzerinden değerlendirilir:



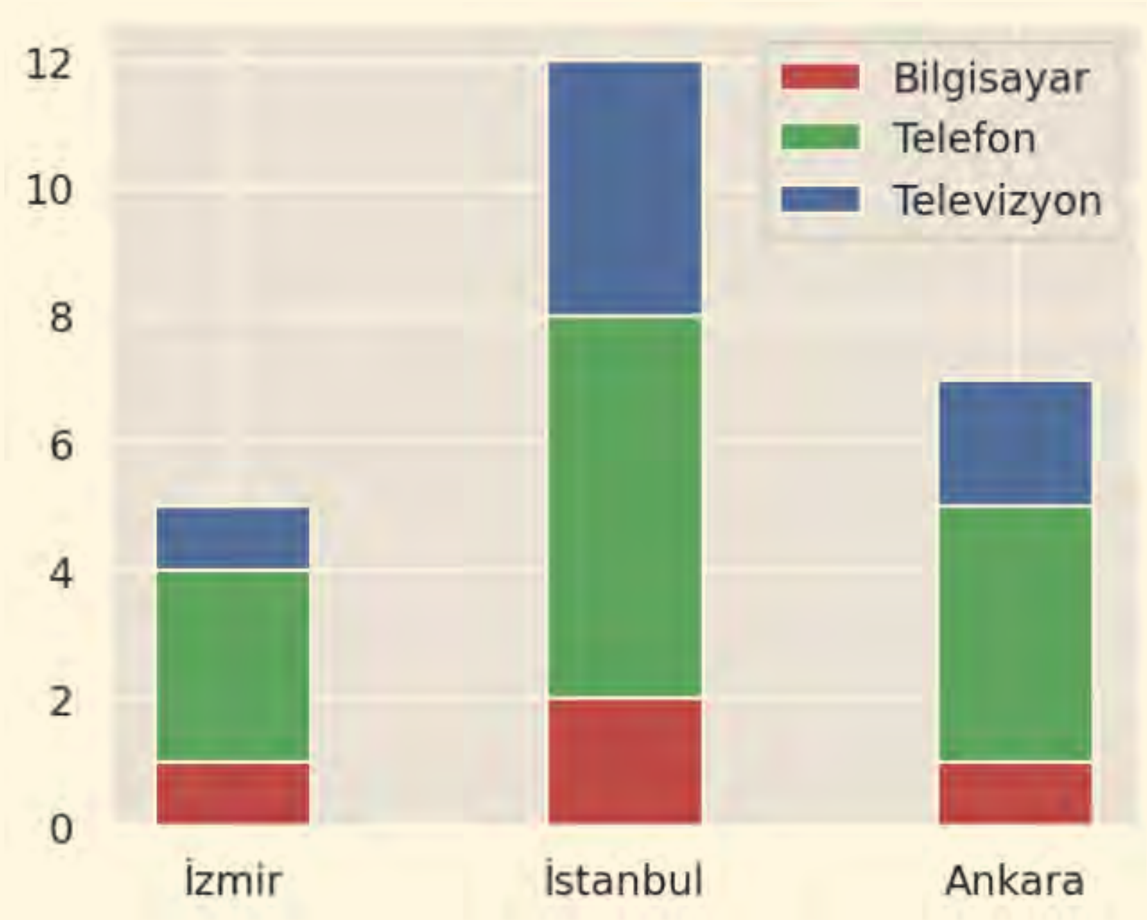
Alt Gruplarında gösterildiği donut grafikleri oluşturabiliriz:



YIĞIN BAR GRAFİĞİ (STACKED BAR CHART)

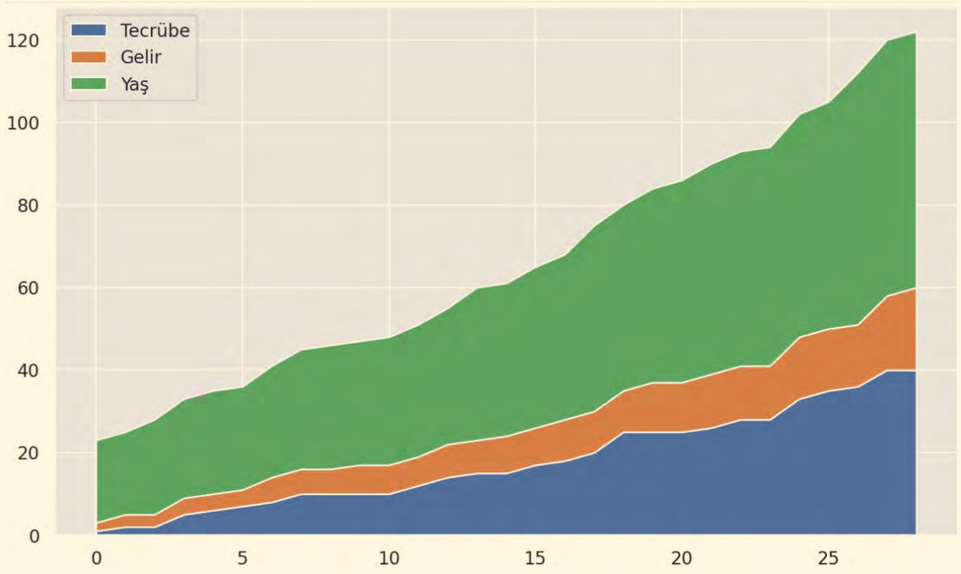
Bir kategori altındaki alt kategorileri de gösteren bar grafikleridir. Alt kategorilerin genel kategoriye oranını göstermek için kullanılır.

Yandaki yığın bar grafiğinde illere göre bilgisayar, telefon ve Televizyon satışları gösterilmiştir. Buna göre her bir kategori kendi değerlerini il bazında diğer bir kategorinin üzerine yığılarak göstermektedir. Örneğin İstanbul'da 2 birim bilgisayar satışına karşılık, 6 birim Telefon ve 4 birim Televizyon satışı yapılmış. İstanbul'daki toplam satış 12 birim olmuş.



YIĞIN ALAN GRAFİĞİ (STACKED AREA CHART)

Yığılmış alan grafiği, aynı grafik üzerinde birden fazla grubun değerinin değişimini gösteren bir grafik türüdür. Alan grafiğinin bir türüdür ancak bu tipte birden fazla grup aynı grafik üzerinde üst üstü bindirilerek sunulur.

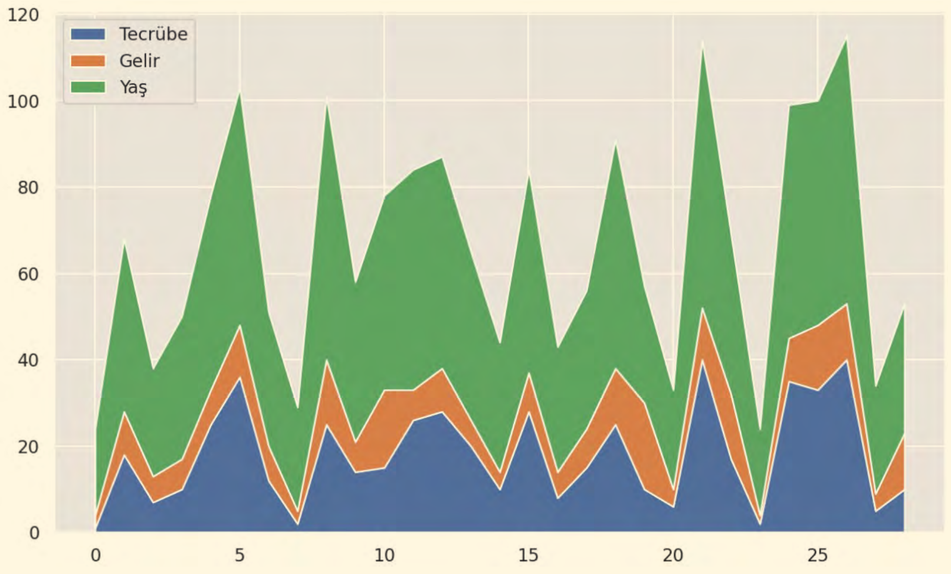


Örneğin bu grafikte Tecrübe, Gelir ve Yaş alanları kendi değerleri ile üst üste binerek (yığın) gösterilmiştir.

Tecrübenin maksimum değerine bakarsanız 40 olduğunu görürsünüz. Gelir alanının değeri ise 20 dir, bu ikisi üst üstü bindiğinde grafikte 60 skalasında gözüküyor. Aynı şekilde Yaş alanının maksimum değeri 62 ve bu alanda diğerlerinin üstüne binip 122 değerinde gösteriliyor.

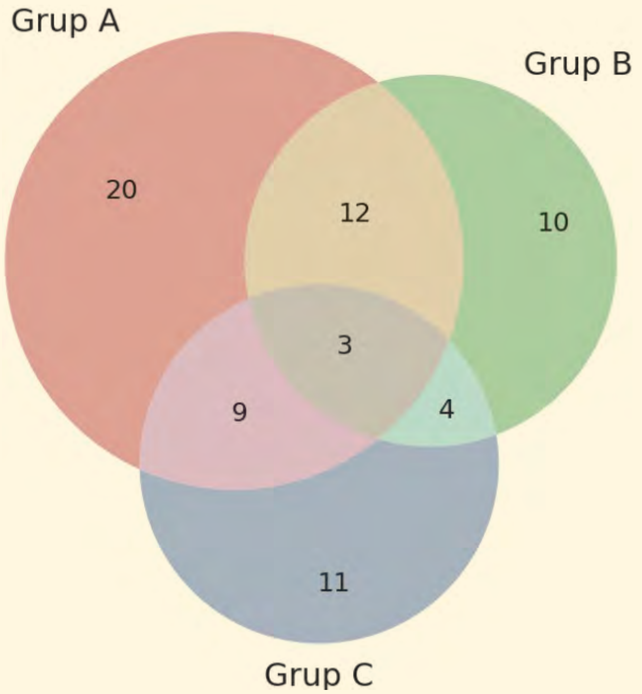
Böylece bütün alanların gidişatını kıyaslamalı olarak görme imkanımız oluyor.

Bu alanları sayısal olarak sıralamadan da gösterebiliriz:



VENN (KÜME) DIAGRAMLARI

Bir venn veya küme diagramı, kümeler arasındaki mantıksal tüm ilişkileri gösteren grafik türleridir:



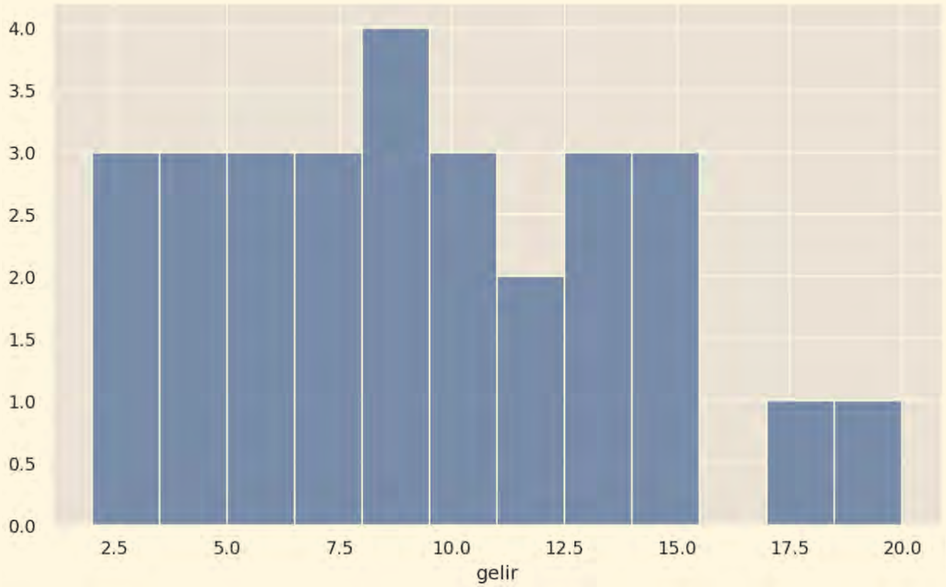
Her küme bir daire ile gösterilir, dairelerin boyutu, kümenin büyüklüğünü temsil eder. Örtüşmenin boyutu da her kesişmenin büyüklüğünü gösterir.

DAĞILIM GRAFİKLERİ

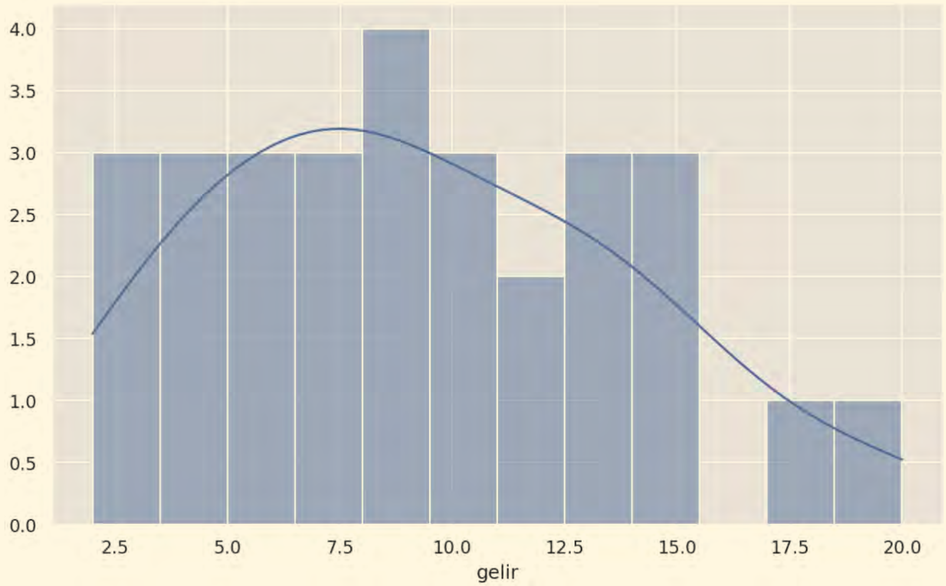
Dağılım grafikleri verilerin nasıl dağıldığı ile ilgili bir içgörü sağlarlar. Tek bir değişkene ait dağılımlarda histogram denilen grafikler kullanılır. Birden fazla dağılımlar için farklı grafik türleri kullanılır. Aşağıda bunlar incelenmiştir.

HISTOGRAM

Histogram tek bir sayısal değişkenin veri dağılımını görselleştirir. Her çubuk, belirli bir aralığın frekansını temsil eder. Değerlerin nerelerde yoğunlaştığı, aykırı değerlerin görülmesini sağlarlar.

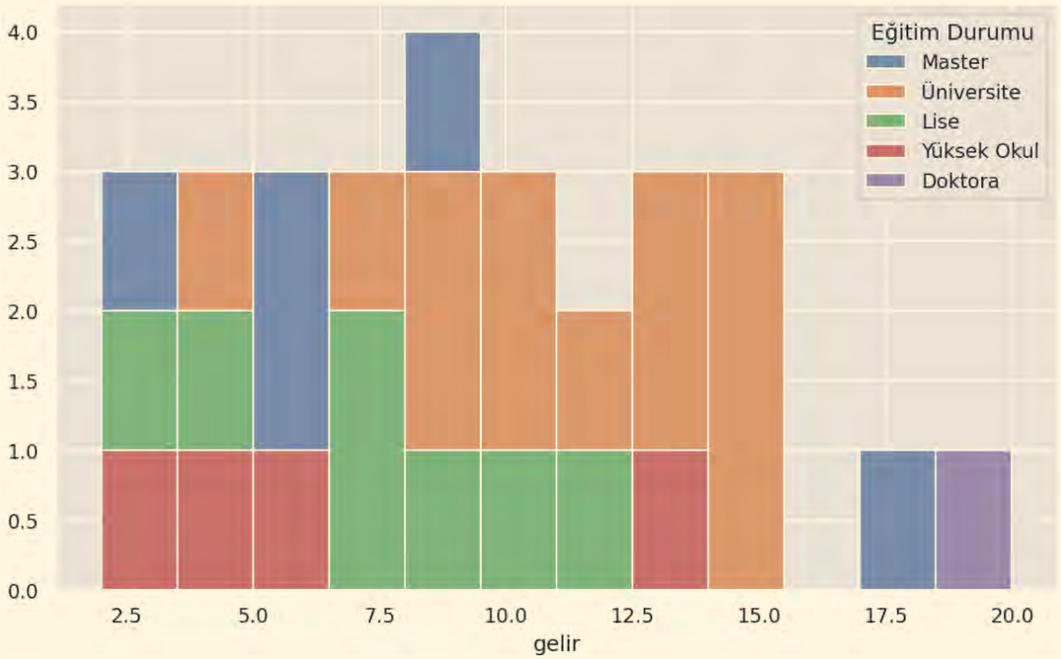


Buna göre örneğin 5-6 arasında 3 adet veri, 15'de 3 adet veri ve 17.5 ile 20 arsında 1 adet veri bulunmaktadır. Dağılımın yoğunluğunun nerelerde tepe noktası yaptığını nerelerde azaldığını gösteren bir çizgi de bulundurabilir bu grafikler:

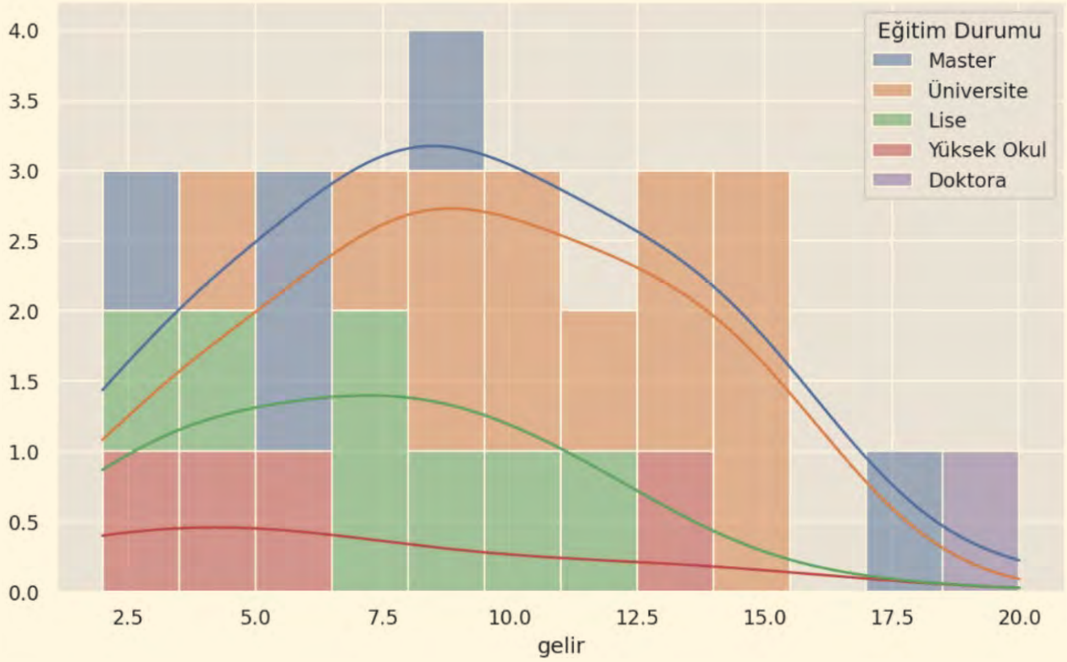


Buna göre gelir dağılımının en yoğun olduğu aralık 6 ila 10 TL olduğu bölge. Bu tip çizgiler Olasılık yoğunluk fonksiyonları tarafından oluşturulan (Kernel Density Estimation, KDE) , değişkenin herhangi bir değerine karşılık gelen olası dağılım değerini döndürürler. Bunlara kısaca KDE veya yoğunluk grafikleri de denir.

Eğer Dilersek dağılım grafiklerinde kategorik alanların gösterimini de yapabiliriz.



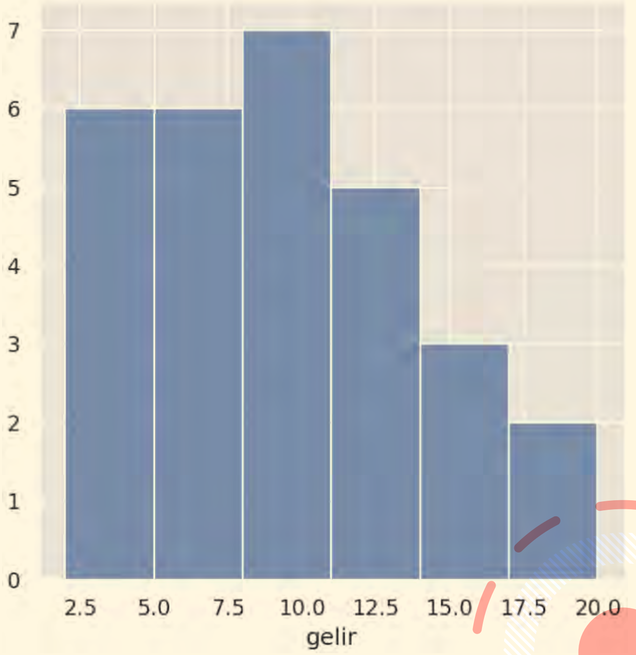
Bu grafik bir önceki dağılım grafiğinin aynısı farklı olarak dağılımların hangi kategorilere ait (Eğitim Durumu) olduğunu da gösteriyor. Eğer dilersek bu grafikler üzerinden KDE eğrilerini de her bir kategorinin yoğunluğu olarak gösterebiliriz:



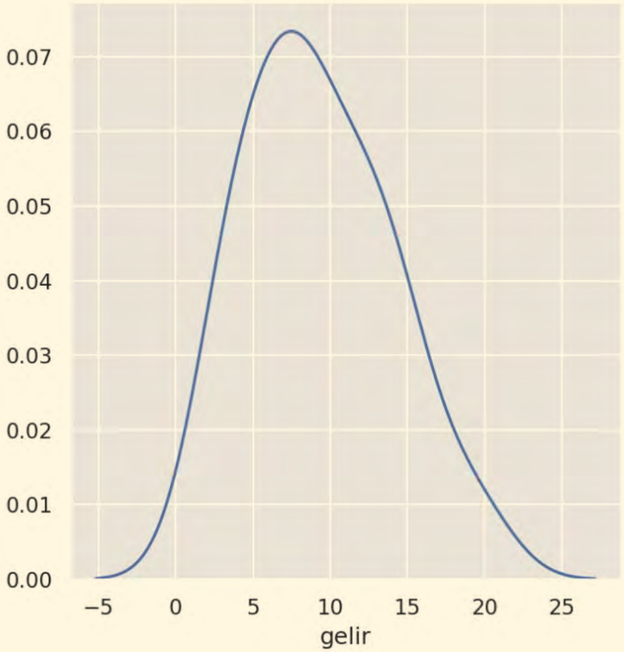
DENSITY (YOĞUNLUK) GRAFİKLERİ

Yukarıda histogramların üzerinden gördüğümüz yoğunluk grafikleri, tek başına da kullanılabilir. Daha düzgün dağılımları izlememize imkan tanıyan sürekli bir grafik yapısındadır. Histogramdaki kutu yapısında farklı olarak veri aralıkları yerine her bir noktaya ait farklı değerler sunarlar.

Örneğin Yukarıdaki verilerimizin farklı aralıklardaki histogram grafiği şu şekilde oluşmaktadır:

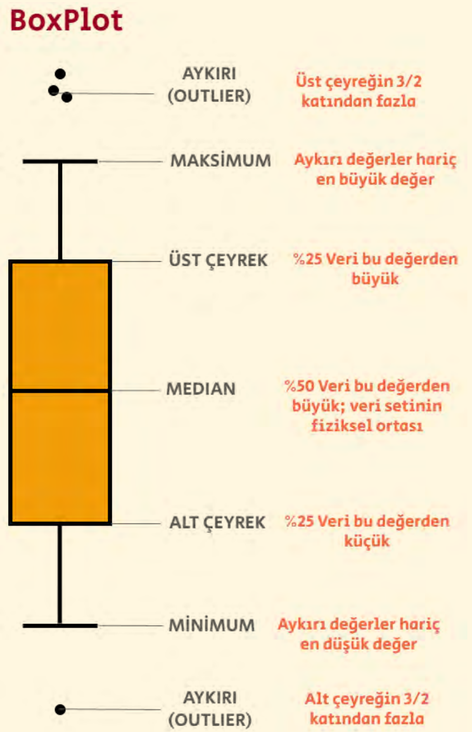


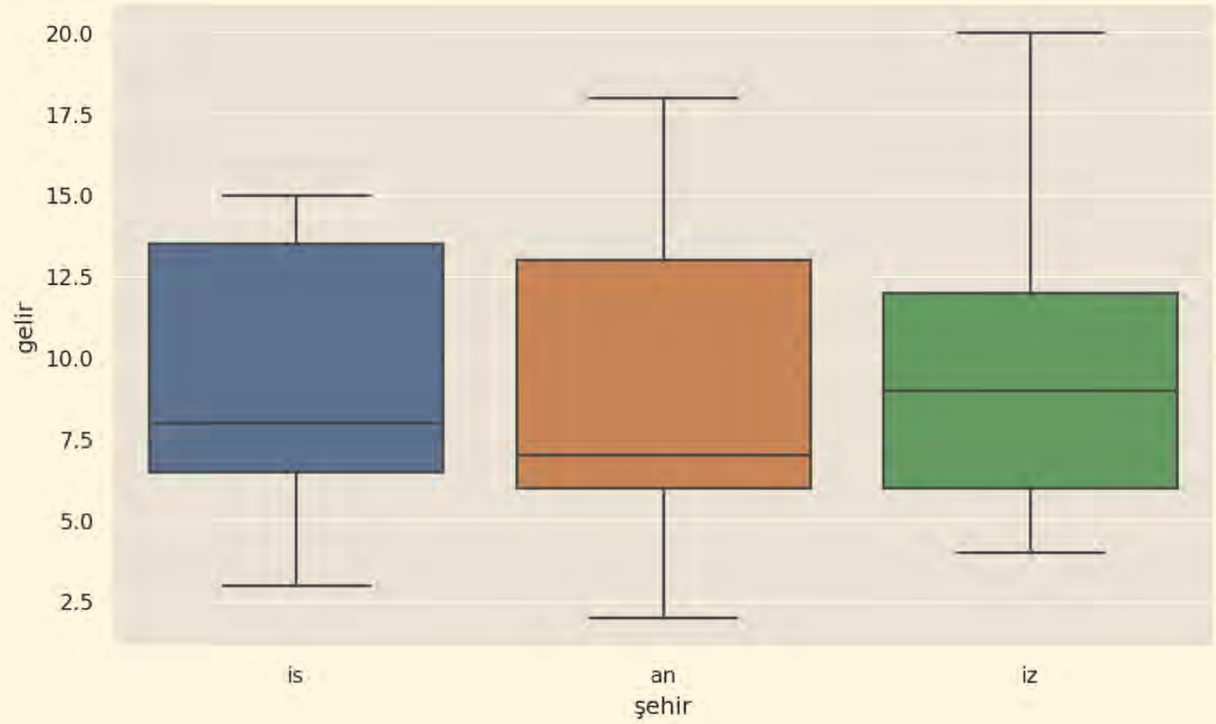
Bunun yoğunluk grafiği:



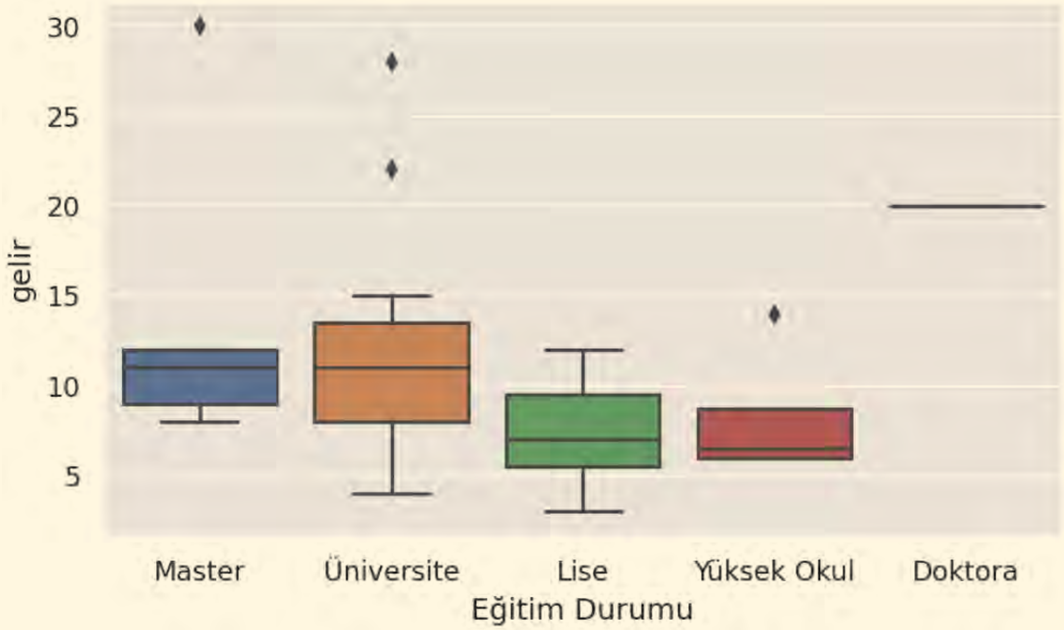
KUTU GRAFİKLERİ (BOX PLOT)

Veriler üzerinde istatistiksel dağılımları gözlemlememizi sağlayan grafik türüdür. Maksimum minimum, medyan gibi değerleri gösterir. Ayrıca üst çeyrek ile alt çeyrek arasındaki veri dağılımlarıyla aykırı verileri de görselleştirir.





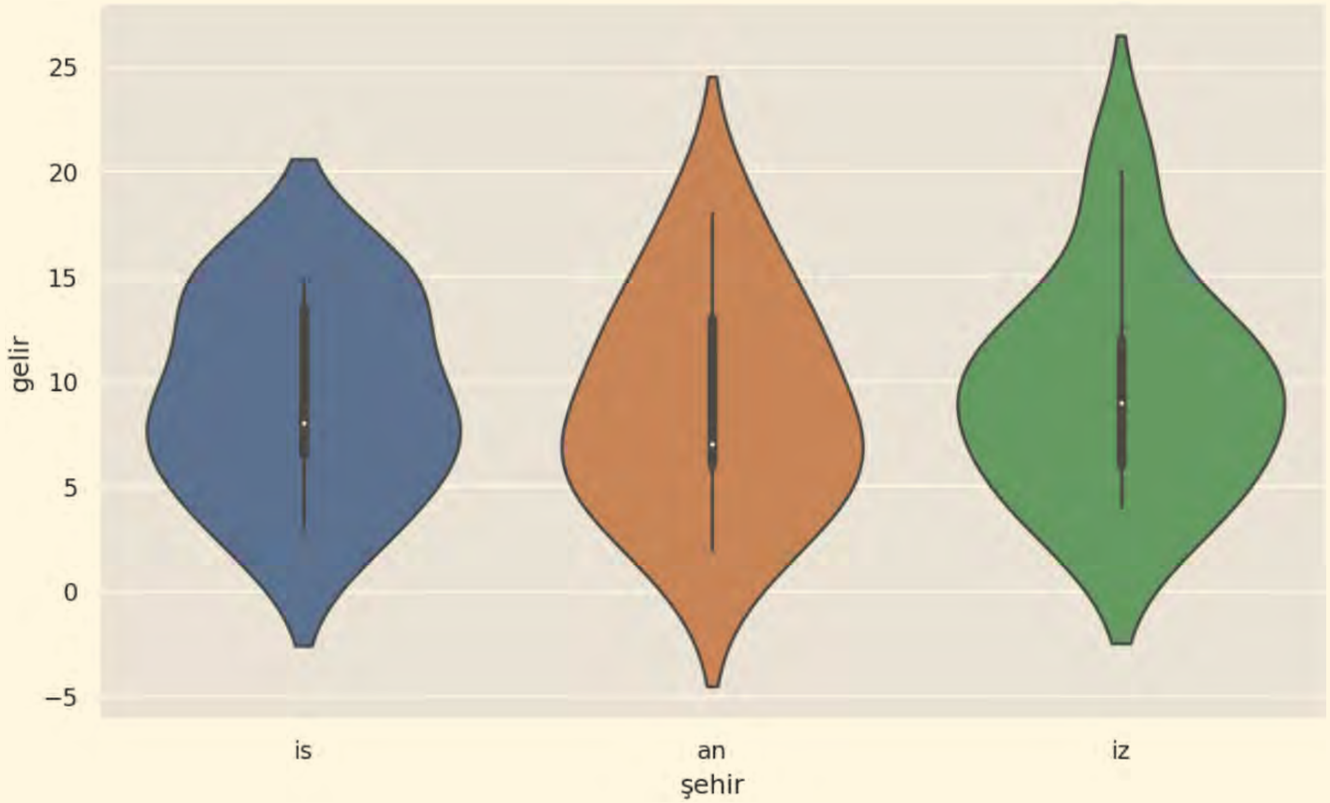
Örneğin bu grafikte İstanbul Ankara ve İzmir'e göre gelir dağılımları gösterilmiştir (veriler gerçek değerlere dayanmamaktadır, örnekleme amaçlıdır) Buna göre İstanbul için gelir dağılımının %50'si 7-14 arasındayken Ankara'da yaklaşık 6-13 arası, İzmir'deyse 6-12 arasında dağılım oluşuyor. Diğer değerler bu dağılımların ya üstünde ya da altındadır. En üstteki ve alttaki değer bir çizgi ile gösterilir, buna kedi bıyığı denir. Kedi bıyığını aşan değerler de olabilir, bunlarada aykırı değerler denir. Kutunun dışında kalan bu değerlerde kalan %50 veriyi temsil ederler. Örneğin:



Örneğin bu tablodaki eğitim durumuna göre gelir dağılımı kutu grafiklerinde üst dağılımın çok üstünde Aykırı değerler var. Master eğitimi gelirinde 30 değeri, diğer dağılımlara göre oldukça fazla ve aykırı bir değer oluşturuyor. Aynı şekilde Üniversite eğitimi gelirindeki 22 ve 28 değerleri dağılıma göre oldukça aykırı gelirleri gösteriyor.

KEMAN (VIOLIN) GRAFİKLERİ

Keman grafikleri, kutu grafikleri ile yoğunluk grafiklerinin bir birleşimidir. Hem dağılımı hem de bu dağılımdaki yoğunluğu görselleştirirler.



Ortadaki kalın siyah çubuk alt ve üst çeyrekler arasındaki aralığı temsil eder. İnce siyah çizgi kutu grafiğindeki kedi bıyıklarını temsil eder. Beyaz nokta ise medyanı temsil eder. Bu merkez çizginin her iki tarafında da veri yoğunluğu görselleştirilir.

TEMEL İSTATİSTİK

Türk Dil kurumuna göre istatistiğin tanımı şu şekildedir:

"Bir sonuç çıkarmak için verileri yöntemli bir biçimde toplayıp sayı olarak belirtme işi, sayımlama."

Oxford sözlüğünde de benzer bir tanımla karşılaşırız:

"Belli alanlardaki bilgileri, olguları, bir sonuç çıkarmak amacıyla, yöntemli bir biçimde toplayıp sayılar halinde gösterme işi."

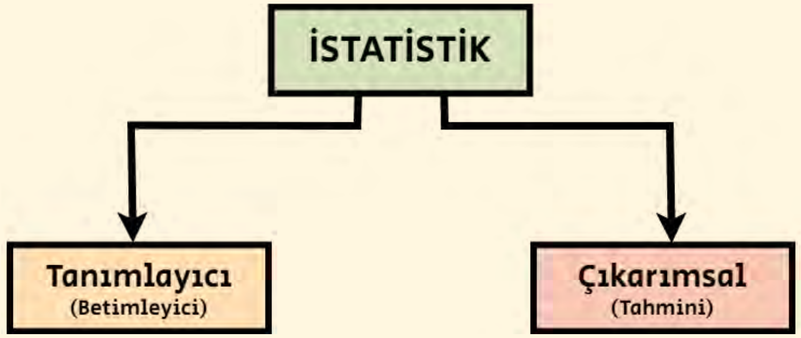
Bu tanımlara baktığımızda istatikte, belirli alanlarda (amaçla) verilerin "yöntemli" bir şekilde toplanması (veri toplama-data seti), gösterilmesi (raporlama-görselleştirme) ve sonuçların çıkarılması (veri analizi – Keşifsel veri analizi) işleri vardır.

Buraya kadar bu yöntemlerin tamamını inceledik.

İstatistikteki temel amaçlardan biri, veri setlerindeki çok fazla veriden, anlayışımızı kolaylaştıracak, belirli matematiksel işlemlerle o veri setini temsil edecek, örnekleyecek verileri bulmaktır. Böylece belki, onbinlerce, yüzbinlerce hatta milyonlarca kayıttan oluşan bir veri seti hakkında bir iç görü veya öngörüye sahip olabilir, o veri setinden bir takım anlamlı sonuçlar çıkartabiliriz.

Devam eden bölümde istatistiğin, bu dağılımları örnekleyen temel konularını ve temel istatistiğe ait, özellikle makine öğrenimi, derin öğrenme konusuna katkı sağlayan bölümlerini inceleyeceğiz.

İstatistiğin temel alanlarına baktığımızda, veri biliminde yukarıda anlattığım pek çok kavramın altyapısını oluşturduğu görürüz. İstatistiği temel olarak iki alana ayırabiliriz: Tanımlayıcı (Betimsel) ve Çıkarımsal (tahmini).



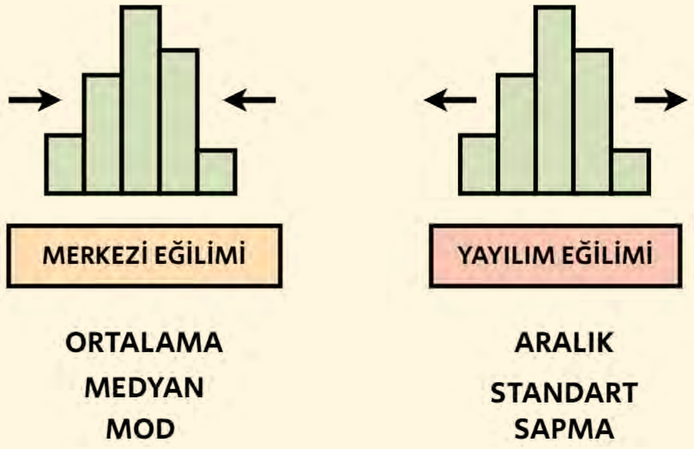
Bu alanlara baktığımızda tanımlayıcı istatistiğin, yukarıda gördüğümüz veri analizinin altyapısını oluşturduğunu, verileri tanımamıza, dağılımlarını anlamamıza yardım eden yöntemler olduğunu görürüz.

İstatistiki yöntemler veri bilimindeki veri görselleştirme veya raporlamayla yapılan ve diğer araçları kullanan yöntemlere matematiksel ve aritmetiksel altyapı oluştururlar. Daha basit söylemek gerekirse tanımlayıcı istatistik, verileri anlamamız için matematiksel-aritmetiksel yöntemler sunar.

TANIMLAYICI İSTATİSTİK

Günümüzde çok fazla veriyle çalışmaktayız. Tanımlayıcı istatistik işte bu verileri en iyi tanımlayabileceğimiz yöntemleri sunar. On binlerce, yüzbinlerce hatta milyonlarca kayıttan oluşan bir sayı seti ile çalıştığımızı düşünün. Bu sayı setindeki verileri anlayabilmemiz o sayı setine ait bir takım özel-özet verileri bulmamız gerekir. Bu özet örneklemeler veya yöntemler sayesinde o binlerce, yüzbinlerce veriden oluşan sayı seti hakkında bir içgörü sahibi olabiliri. Bu sayı setini en iyi tanımlayan bu matematiksel ve aritmetiksel yöntemler, toplanan sayı seti hakkında belirli yorumları, analizleri ve belki de çıkarımları yapmamıza olanak verir.

Tanımlayıcı istatistik temel olarak iki ana yöntemle yapılır: Merkezi eğilim ve Dağılım (yayılım) eğilimi.



Her iki yönteminde matematiksel karşılıkları vardır.

MERKEZİ EĞİLİM

Şemada da görüleceği gibi merkezi eğilim istatistiğinin kullandığı 3 temel matematiksel işlem vardır:

- Aritmetiksel Ortalama
- Medyan
- Mod

Bu yöntemlerin hepsi verilerin hangi merkezde toplandığına yönelik matematiksel çıkarımlardır. Her bir yöntem mevcut verilerin hangi ortalamada, ortada veya tekrarda temsil edildiğine ait sayısal sonuçlar üretir.

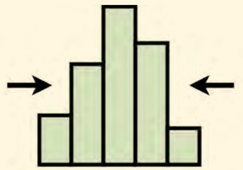
Her bir yöntem çok sayıda verinin temsilini oluşturan sayılar üretir. Merkez eğilimlerini temsil eden bu sayıların yorumlanması verinin hangi amaçla kullanıldığına bağlı olarak değişik faydalı içgörüler ve temsiller sağlarlar.

Şu şekilde de ifade edebiliriz, her bir yöntem, verilerin temsil ettiği amaca göre faydalı temsiller oluşturabilir. Örneğin araba fiyatlarını temsil etmek amacıyla aritmetiksel ortalama almak medyana göre daha faydalı bir temsil sunmaya bilir. Çünkü yaygın bir aralıktaki araba fiyatlarının aritmetiksel ortalamasını, en pahalı bir-iki araba bozabilir.

Bu durumda medyan ve/veya mod temsillerini almak daha doğru olabilir. Ama başka bir durumda, bir sınavdaki aritmetiksel ortalamayı almak, o sınavın başarısını göstermek adına daha doğru olabilir. Dolayısı ile bu yöntemlerin birinin diğerine göre bir avantajı yoktur. Temsil edecekleri veri setine göre farklılık gösterebilirler. Ancak sonuç olarak hepsi mevcut veriyle ilgili merkezi eğilim bilgilerini döndürürler.

ARİTMETİKSEL ORTALAMA

Geometrik, harmonik ortalama gibi başka ortalamalar da vardır ancak ortalama deyince aklımıza aksi belirtilmemişse aritmetik ortalama gelmelidir. Bir sayı setinin merkezi eğilimi hakkında o setteki bütün sayıların toplamının, veri sayısına bölünmesiyle elde edilen sayısal bir temsil değeri verir.



MERKEZİ EĞİLİM

ORTALAMA

$x = 5, 12, 12, 15, 20$

1 2 3 4 5

$$x_{ortalama} = \frac{5 + 12 + 12 + 15 + 20}{5}$$

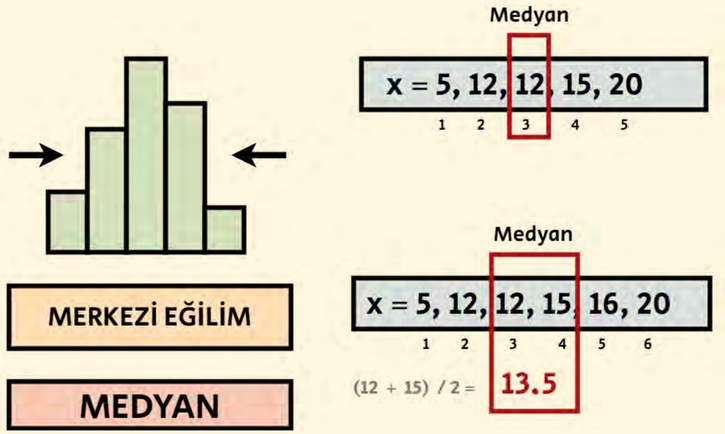
$$x_{ortalama} = 12.8$$

Bu değer o veri setindeki sayıların matematiksel orta değerini temsil eder. Veri setindeki diğer değerler bu temsilden daha büyük veya daha küçük olabilir hatta bu değere eşit olabilir ama bu sonuç temsili matematiksel bir orta değerdir. Yukarıdaki örnekte bir veri setindeki sayıların ortalaması 12.8 bulunmuştur. Bu değer veri seti içindeki sayılarda yok, ancak veri setinin merkezi ortalamasını temsil etmektedir. Eğer veri setimizdeki sayıları görmemiş olsanız ve birisi size 5 tane verisi olan setin ortalaması 12.8 deseydi, sayıların dağılımı konusunda aşağı yukarı bir fikriniz olurdu. Aynı mantığı büyük veri setleri için de düşünebilirsiniz. Çok büyük veri setlerinin ortalamasını söylemek o veri setindeki dağılım hakkında bir fikir verir. Elbette çok aykırı değerler olabilir, ancak sonuçta merkezi olarak o veri setindeki sayıların değerleri hakkında bir fikriniz olur. Çok küçük veri setlerinde veya çok aykırı durumlarda ortalama almak veri setindeki sayıların merkez dağılımı hakkında doğru bilgi vermeyebilir.

Örneğin x= 2, 5,7,8,42 şeklinde bir veri setimiz olsun. Bu veri setinin merkezi eğilimini aritmetik ortalmayla hesaplayacak olsak yine 12.8 çıkar. Ancak bu temsili değer veri setindeki genel dağılım ve merkezde olan dağılımı tam temsil etmez. Bunun için diğer merkez eğilim yöntemlerine de başvurmakta, ortalamayla birlikte değerlendirmekte fayda vardır.

MEDYAN

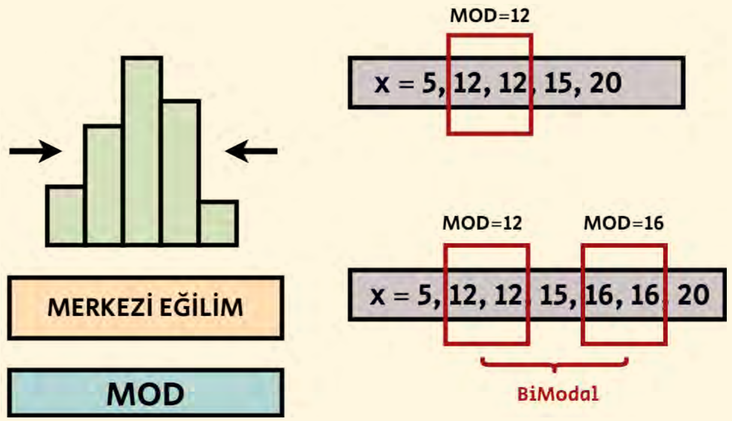
Medyan, sayı setindeki sayıların fiziksel olarak orta noktasını alır. Sayısal Veri setindeki sayılar sıraya dizilir ve bu sıranın tam ortası o veri setinin medyanı olur. Türkçeye "ortanca" olarak da çevrilir.



Yukarıda yaptığımız son örnekteki x= 2, 5,7,8,42 sayı setindeki medyan, setin tam ortasında olan 7 değeridir. Bu setin aritmetiksel ortalaması 12.8 fakat medyanı 7 dir. Bu sayılardan oluşan veri setini görmemiş birine bu değerleri verdiğimizde - veya sizin hiç görmediğinizi varsayalım- veri merkezi ve bu verilerin dağılımı konusunda ortalamaya göre çok daha net fikri olacaktır. Bu örneği yine çok ama çok büyük veri setlerini anlamak için uyarlayın. O veri setlerindeki her bir veriyi tek tek gözlemlemek yerine bu parametrelerle değerlendirdiğinizde o veri seti hakkında bir fikir sahibi olabilirsiniz.

MOD

Bir sayı setinde en sık rastlanan sayıdır. X=5,12,12,15,20 sayı setinde en sık rastlanan veya tekrar eden sayı 12 dir, dolayısı ile bu sayı setinin mod'u 12 dir.



Eğer bir sayı setinde birden fazla tekrar eden sayılar varsa o setin birden fazla "mod"u vardır demektir. Eğer iki modu varsa Bimodal daha fazla varsa "Multimodal" set olarak adlandırılır.

MERKEZİ EĞİLİMİN DEĞERLENDİRİLMESİ

Merkez eğilimin 3 temel yöntemini inceledik. Bölümünde başında belirttiğim gibi, bu metotların herhangi birisinin diğerine bir üstünlüğü yoktur. Sadece bir sayı setindeki merkez dağılımlarını anlamamıza yönelik yöntemlerdir ve çoğunlukla 3'ü bir arada kullanılmalıdır.

Yukarıdakine benzer bir örnek vermek gerekirse, şöyle bir sayı setimiz olsun:

X= 5, 5, 6, 8, 150

Eğer bu sayı setini örneğin araba fiyatları için değerlendiriyor olsaydık ve sadece aritmetiksel ortalamaya baksaydık (5+5+6+9+150)/5 = 35 değeri elde ederdik. Buna göre araba fiyatlarının merkeze göre durumunun 35 lira olduğunu düşünürüz. Ancak sayı setinin dağılımına baktığımızda 35 sayısının bizi yanıltabileceğini görüyorsunuz. Çünkü aykırı veya sıra dışı bir veri ortalamayı etkilemiş. Aykırı ve veya sıra dışı verilerin tam bir tanımı olmamasına rağmen sayı setindeki diğer verilere bakınca hemen anlaşılır.

Bu durumda aritmetiksel ortalamaya göre bir fikir sahibi olmak çok doğru yorumlar yapmamızı sağlamayabilir.

Bu sayı setinin bir de medyanına bakalım. Setin tam ortasında 6 değeri var. Bu değer araba fiyatlara hakkında daha iyi bir fikir vermektedir. Aynı şekilde "mod" a baktığımızda 5 değerini görürsünüz. Bu değer de oldukça iyi bir yaklaşım sunar. Diğer bir ifadeyle bu sayı setini bir araba fiyatı değerlemesine göre yorumlamak için medyan ve mod çok daha iyi bir iş çıkarıyor.

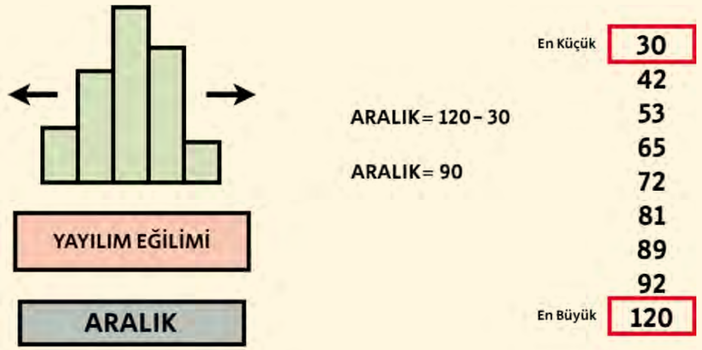
Ancak aynı sayı setini örneğin 5 aylık satış yorumlaması için değerlendirirsek aritmetiksel ortalama olan 35 değeri daha iyi fikir vermektedir.

YAYILIM EĞİLİMİ

Tanımlayıcı istatistiğin temel iki alanından bir merkezi eğilim ise diğeri de yayılım eğilimidir. Yayılım eğilimi merkezi eğilimin tam tersidir. Yayılım eğilimi sayılar arasındaki uzaklığı ifade eder. Sayılar arasındaki uzaklığı ifade etmenin değişik yöntemleri vardır. Bunlardan biri Açıklıktır (aralık).

AÇIKLIK (ARALIK)

Aralık değeri bir sayısal veri setindeki en düşük değer ile en yüksek değer arasındaki farkı belirtir.



ARALIK= 120 – 30

ARALIK= 90

En Küçük	30
	42
	53
	65
	72
	81
	89
	92
En Büyük	120

Fark büyüdükçe sayılar arasındaki mesafe artıyor demektir. En büyük ve en küçük sayı arasındaki mesafe (fark) büyükse, açıklıkta büyük oluyor demektir. Tabii ki tam tersi de geçerlidir, değer küçükse açıklıkta küçük demektir. Sayılar arasındaki bu mesafe, verinin dağılımının ne kadar birbirine yakın veya açık olduğunu belirtir.

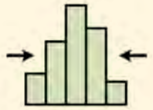
Aşağıdaki gibi veri dağılımları olan iki sayı seti olsun:

X= -5, 0, 5, 10, 15 y= 3, 4, 5, 6, 7

MERKEZ – YAYILIM EĞİLİMİ

x = –5, 0, 5, 10, 15
y = 3, 4, 5, 6, 7

MERKEZ EĞİLİMİ



$x_{ortalama} = 5$ $x_{medyan} = 5$

$y_{ortalama} = 5$ $y_{medyan} = 5$

YAYILIM EĞİLİMİ

$x_{Aralık} = 20$ (15 - (-5))

$y_{Aralık} = 4$ (7 - 3)

y Sayıları, x'e göre bir birine daha yakındır

X ve y nin ortalaması ve medyanı aynıdır. Bu sayı setlerindeki değerleri görmeyen biri bu iki setteki sayıların birbirine çok benzer olduğunu düşünebilir. Ancak aralık değerine baktığımızda aslında bunun böyle olmadığını iki setteki sayılar arasında oldukça fark olduğunu çıkartabilmeliyiz. Setlerdeki değerlere baktığımızda bunu kolayca gözlemleyebiliyoruz. Tabi burada sadece 5 sayıdan oluşan her bir set için bunu gözlemlemek kolay. Ancak binlerce, on binlerce hatta milyonlarca sayıdan oluşan bir sayı setinin dağılımını anlayabilmek ve sayı setlerini birbiriyle karşılaştırmak için bu yöntemler oldukça faydalıdır.

Örneğin iki sınıftaki sınav sonuçlarının ortalaması aynı olabilir. Ancak bu bilgi iki sınıftaki not dağılımları hakkında bir fikir vermez. Bunun için iki sınıftaki not aralıklarını bilmemiz gerekir. Böylece not dağılımlarının dengeli olup olmadığını da anlarız.

Bazı durumlarda aralık değeri de dağılım hakkında tam bilgi döndürmeyebilir bunun için, varyant ya da standart sapma hesaplamalarını kullanmak daha sağlıklıdır.

ORTALAMA MUTLAK SAPMA

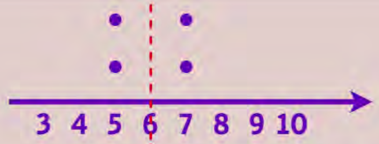
Bu hesaplama yöntemi, sayı dağılımlarının, ortalamadan farklarının ortalamasını belirtir. İfade biraz karışık gibi görünse de aslında mantık çok basittir. Bir sayı setindeki sayıların ortalaması alındıktan sonra her bir sayının bu ortalamayla olan farkının mutlak değeri alınıp, bu farkların ortalaması bulunur.

Böylece her hangi bir sayı setinin ortalamaya olan uzaklıkları hakkında bir fikir edinilmiş olur. Bu değer arttıkça dağılımın daha geniş olduğu tersi durumda ise daha dar alanda olduğunu varsayabiliriz.

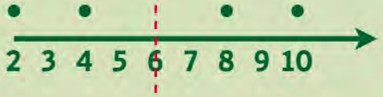
YAYILIM EĞİLİMİ

ORTALAMA MUTLAK SAPMA (OMS)

$x = 5, 5, 7, 7$ $\quad x_{ortalama} = 6$



$y = 2, 4, 8, 10$ $\quad y_{ortalama} = 6$



$$x_{OMS} = \frac{|5-6| + |5-6| + |7-6| + |7-6|}{4} \qquad x_{OMS} = 1$$

$$y_{OMS} = \frac{|2-6| + |4-6| + |8-6| + |10-6|}{4} \qquad y_{OMS} = 3$$

VARYANS – STANDART SAPMA

Sayı dağılımlarını hesaplamanın en yaygın yöntemlerinden iki tanesi varyans ve standart sapmadır. Aslında standart sapma varyansın karekökünü alarak hesaplanan bir adım sonraki değeridir.

Ortalama mutlak sapma mantığına benzeyen varyans hesaplamasında, sayı setindeki her bir değerin ortalama ile olan farkının karesinin ortalaması alınır. İfade biraz karışık ama örneklere baktığımızda hesaplamanın aslında ne kadar basit olduğunu göreceksiniz. Ortalama mutlak sapmadan farklı olarak bu sefer mutlak değer yerine değerlerin karesi alınır. Bu arada, istatistikte veri setlerine, "yığın", "ana kütle" veya popülasyon ismi de verilir. Aşağıdaki tabloda bu 2 farklı popülasyona ait varyans ve standart sapma hesaplaması görülmektedir.

Varyans Standart Sapma (Popülasyon)

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$$

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

$x = 5, 5, 7, 7$ (1 2 3 4) $N = 4$

$x_{ortalama} = 6$

3 4 5 6 7 8 9 10

$y = 2, 4, 8, 10$ (1 2 3 4) $N = 4$

$y_{ortalama} = 6$

2 3 4 5 6 7 8 9 10

$$x_{varyans} = \frac{(5-6)^2 + (5-6)^2 + (7-6)^2 + (7-6)^2}{4} \qquad x_{varyans} = 1$$

$$y_{varyans} = \frac{(2-6)^2 + (4-6)^2 + (8-6)^2 + (10-6)^2}{4} \qquad y_{varyans} = 10$$

$$x_{StandartSapma} = \sqrt{x_{varyans}} = \sqrt{1} = 1$$

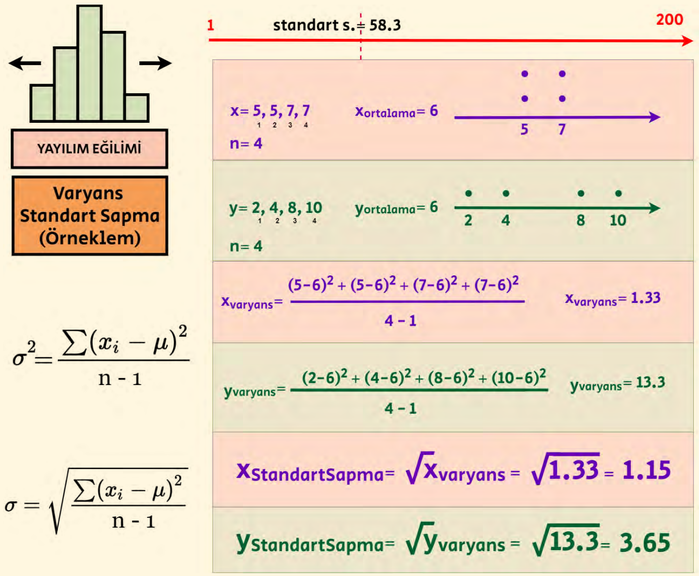
$$y_{StandartSapma} = \sqrt{y_{varyans}} = \sqrt{10} = 3.16$$

Gerek ortalama mutlak sapma olsun, gerek varyans olsun gerekse de standart sapma hesaplamaları olsun bunlar sayı setindeki sayıların birbirine olan mesafe-aralıkları hakkında bilgi vermektedir. Değeri yüksek olan çıktılar sayı aralıklarının fazla olduğunu, tersi ise aralıkların daha az olduğunu gösterir.

ÖRNEKLEM VARYANSI – STANDART SAPMASI

İstatistikte genelde bir sayı setinin (popülasyonun) tamamı kullanılmaz. Çünkü çoğunlukla popülasyon çok fazla olur. Örneğin Türkiye nüfusu, illerin nüfusu ile ilgili istatistiki işlemler yapıldığını düşünün. Bu illerdeki nüfusun tamamı yerine, buradan örneklemler alınır.

Örneklem varyans ve standart sapma hesaplamaları, popülasyon yani bütün veriye ait hesaplamalardan biraz farklıdır. Bu hesaplamada bölen olarak kullanılan değer, popülasyon sayısının bir eksiği olarak kullanılır.



$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{n - 1}$$

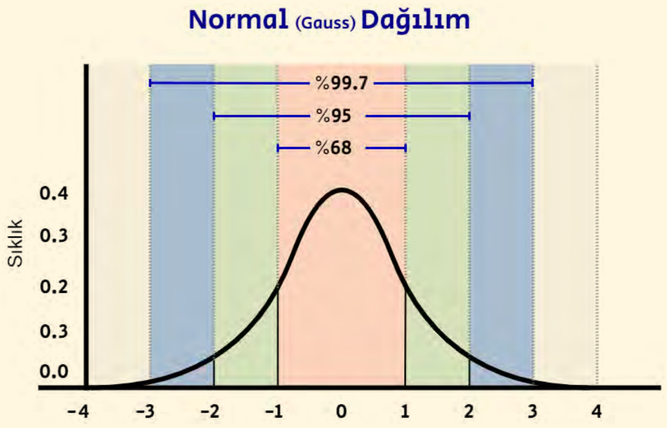
$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n - 1}}$$

Yanda popülasyon varyans ve standart sapmasındaki veri setini 200 verisi bulanan bir veri setinden (popülasyondan) örneklem olarak alındığını düşünün. 1 den 200 e kadar olan sayılardan oluşan bu veri setinin standart sapma değeri 58 dir (yuvarlamalardan dolayı bazı farklar olabilir). Bu 200 adet veriden yukarıda gördüğümüz 4 er adet örneklem alınmıştır. Popülasyon hesaplamalarındaki değerleri kullanmak istediğimden örneklemler bu şekilde oluştu. Normal koşullarda 50, 70,120, 180 veya 20,95,145,185 gibi örneklemler almak daha gerçekçi olacaktır. Ancak fark etmez, örnekler arasındaki uyumluluğu korumak adına aynı verilerle ilerleyelim. Bu seferki 4 değer bir popülasyon değeri değil bir popülasyondan alınan örneklerdir. O zaman hesaplama biraz farklılaşıyor. Popülasyon standart sapma değerine bakarsak 58 görürüz. Aldığımız örneklemler çok küçük değerlere sahip olduğu için popülasyon standart sapma değerinin çok uzağında olacaktır. Ancak ne kadar yakınsarsa o kadar gerçeğe yaklaşmış olacaktır. Popülasyon formülüne göre hesaplanırsa, bölen 4 olacaktır. Ancak bölen büyük olduğunda bu dağılımlar için, örneklem verilerinin standart sapması gerçek değerden uzaklaşacaktır. Bunun yerine 1 azaltılarak daha az bir bölen elde ediliyor. Bu sayede gerçek standart sapmaya daha yaklaşılmış oluyor. X veri setinde 1 standart sapması, 1.15 e, y veri setinde 3.16 standart sapması 3.65 değerine çıkıyor. Diğer bir ifadeyle 58 değerine daha da yaklaşıyor. Yukarıda da belirttiğim gibi daha dengeli örneklemle de gerçek değere daha yakın değerler elde edilir.

Örneklem değerlerinin aralıkları ve popülasyon standart sapmasına olan mesafeleri azaldıkça, böleni daha küçük olan hesaplamalar, popülasyon standart sapmasına daha yakın değerler döndürmektedir. Böleni daha küçük derken bu da örneklem sayısının bir eksiği noktasında optimizedir. Bundan daha farklı değerler popülasyon değerinden uzak sonuçlar döndürmektedir. Farklı Veri aralığı ve merkeze olan uzaklıklardan oluşan örneklemlerle yapılan hesaplamaların ortalamasında yaklaşık 2/3 oranında n-1 bölenin, popülasyon değerine en yakın sonuçları ürettiği hesaplamalarla ispatlanmıştır. Buna karşılık veri setinin en uzak 2 uç noktasından alınan örneklemler daha yüksek bir bölen gerektirir fakat bu örneklemlerin sayısı diğer örneklemlere göre daha azdır. Bundan ötürü, örneklem varyans veya standart sapma hesaplamalarında bölen sayı seti sayısının bir eksiği olarak alınır.

NORMAL DAĞILIMI

İstatistikte verinin dağılımıyla ilgili, isimlendirilmiş bir çok metot vardır. Bunlardan en bilineni normal dağılımıdır. Gauss dağılımı olarak da bilinir. Merkez değerinin etrafında ağırlıklı dağılıma sahip, simetrik, çan şeklinde bir eğridir. Bu dağılımda merkezden uçlara doğru gidildikçe sıklığın(yoğunluğun) azaldığını görürüz. Ortalama, median ve Mod normal dağılım için eşittir.

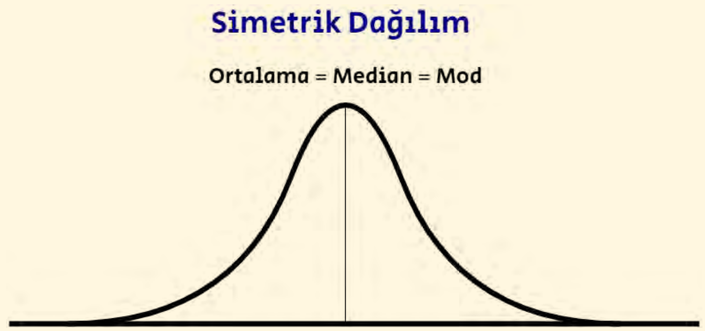


EĞRİLİK - ÇARPIKLIK

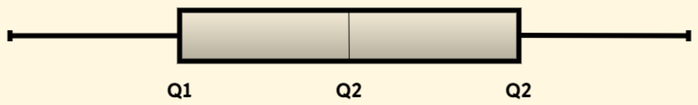
Eğrilik veya Çarpıklık, bir simetri göstergesidir. Bir dağılımın simetriden ne kadar saptığını gösterir. Simetri, Sol (negatif çarpık dağılım olarak da bilinir) ve sağ (pozitif çarpık dağılım olarak da bilinir) taraf çarpıklığı olarak 3 temel dağılım vardır.

SİMETRİK DAĞILIM

Genellikle çan eğrisi olarak görünür. Mükemmel bir simetriyi tanımlar ancak gerçek dünyada bu simetriyi yakalamak mümkün değildir. Diğer bir ifadeyle sıfır çarpıklık olması mümkün değildir. Bunun yerine sıfıra mümkün olduğunca yakın çarpıklıktan bahsedilir.

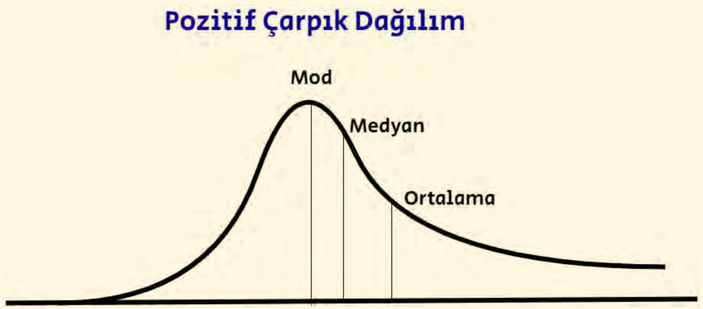


Kutu grafiğindeki gösterimi ise şu şekildedir.



POZİTİF ÇARPIKLIK

Sağ tarafa doğru giden uzun bir kuyruk varsa pozitif çarpıklıktan bahsedilir. Kuyruk merkez değerin pozitif tarafında olduğundan, dağılım pozitif olarak çarpıtılır. Simetrik dağılımın aksine veriler merkezin her iki tarafına eşit olarak dağılmaz.



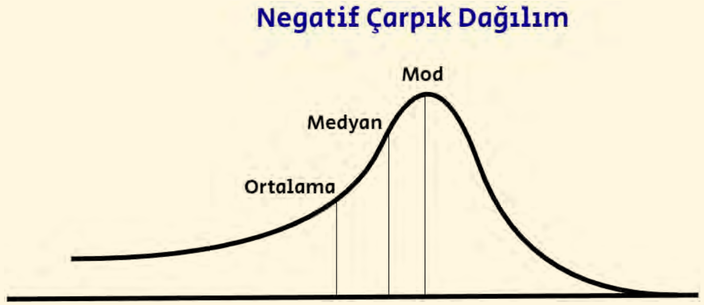
Grafikten ortalama değerinin (sıklık, frekans veya yoğunluk değil, x ekseni değeri) en yüksek değer olduğunu görebiliriz. Ortalama değeri medyan ve mod takip ediyor. Eğrinin kuyruğu sağ doğru yani yüksek değerlere doğru olduğundan, ortalama medyandan daha büyüktür. Mod ise en yüksek frekansta (sıklık, yoğunlukta) gerçekleşir. Bu eğrinin sıralaması Mod<Medyan<Ortalama şeklindedir.

Bu dağılımın kutu grafiği şu şekildedir:

Bıyıkların uzunluklarındaki farklılıklar (maks bıyığının, min bıyığında daha büyük olması) kutu grafiğinin Q2 noktasını merkeze alabilir.

NEGATİF ÇARPIKLIK

Sağ taraf çarpıklığının tam tersi olarak düşünebilirsiniz. Bu sefer kuyruk sola doğru uzar. Bu sefer ortalama daha az olan kısma doğru kayar, değer olarak en küçük değere sahiptir. Onu medyan ve mod takip eder:



Negatif çarpık dağılımın kutu grafiğide şu şekildedir:

ÇIKARIMSAL İSTATİSTİK

İstatistiğin temel alanlarından biri açıklayıcı istatistikse diğeri de çıkarımsal istatistiktir. Buna tahmini istatistik de denir. Popülasyonlardan (bütün veri seti) alınan örneklemlerle, büyük popülasyon hakkında çıkarımlarda, tahminlerde bulunur. Çıkarımsal istatistiğin değişik metotları vardır.

Tanımlayıcı istatistik mevcut veriyi anlamaya yönelikken, Çıkarımsal istatistik, verilerden tahminler, çıkarımlar yapmaya yönelik metotlardır.

YAPAY ZEKA - MAKİNE ÖĞRENİMİ

Son yıllarda veriyle çalışma yeni algoritma ve tekniklerin devreye girmesiyle ivme kazanmıştır. Belirli matematiksel algoritmalarla mevcut veriden, insan müdahalesi olmadan öğrenen yapılar geliştirilebilmektedir. Diğer bir ifadeyle makineler, veriden kendi başına öğrenebilmektedir. Özellikle 80'li yıllardan itibaren kullanılmaya başlayan bu algoritma ve teknikler, 2010'lu yıllarda derin öğrenme algoritmalarının devreye girmesiyle başka bir evreye geçmiş, günümüz Yapay Zekasının altyapıları oluşturulmuştur.

Günümüz Makine öğrenimi ve Yapay Zeka teknikleri pek çok disiplin ve teknolojiyi içinde barındırır. Pek çok kaynağa göre veri biliminin alanlarından biri olarak incelenir ancak, kendi başlarına da bağımsız birer alandırlar.

Makine Öğrenimi – Yapay Zeka Alt Disiplin ve teknolojilerine şöyle bakacak olursak

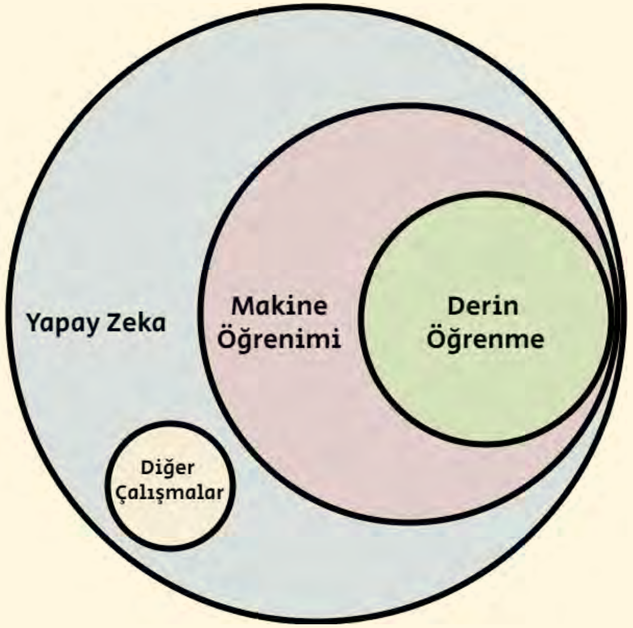
- Matematik
- İstatistik
- Veri – Veri Bilimi pek çok konusu
- Programlama

Makine Öğrenimi ve Yapay Zeka bütün bu teknoloji ve disiplinlerden oluşur, ancak hiç biri tek başına onu oluşturmaz.

Makine Öğrenimi ve Yapay Zekada, veri biliminin diğer alanlarından farklı bir şekilde yoğun olarak programlama ve kodlama teknikleri kullanılır. Özellikle Yapay Zeka konuları, derin öğrenme algoritmalarının ihtiyaç duyduğu çok fazla hesaplatma işlemi yaparlar. Hacimli hesaplatmalar yapabilmemizi sağlayan altyapılar da, programlama sayesinde kurgulanabilir.

Yapay Zeka – Derin Öğrenme konularının son yıllardaki yükselişinin en temel sebeplerinden biri de hesaplatma (Computing) gücümüzün her geçen gün daha da artmasıdır. Diğer bir sebebi de verinin her geçen gün daha da büyümesidir.

Özetle söylemek gerekirse, Yapay zekanın son yıllardaki yükselişi, büyük verinin oluşması ve bu veriyi işleyebilecek işlem gücüne (programlama ve programlama altyapıları) sahip olmamız ile mümkün olmuştur.



Günümüz Yapay Zekası, Makine Öğrenimi kategorisinde sınıflandırılır. Makine öğreniminin algoritmalarından biri olan Derin Öğrenme algoritması ve yapay sinir ağları ve onun türevleri, günümüz Yapay Zeka çalışmalarının konusunu oluşturur. Ancak bu demek değildir ki Yapay Zeka alanında başka çalışmalar yok. Yapay Zeka alanında Makine öğrenimi ve derin öğrenme teknolojileri dışında da, gerek akademik, gerek şirketler özelinde bir takım çalışmalar vardır ancak bunlar günlük hayatımıza veya endüstriyel kullanım pratiklerine ulaşmamıştır. En azından şimdilik.

MAKİNE ÖĞRENİMİ TİPLERİ VE ALGORİTMALARI

Makine Öğrenimindeki "Makine" ifadesi soyut olarak bir bilgisayarı temsil eder. Diğer bir ifadeyle işlem yapabilen donanımları. Bu paralel çalışan pek çok sunucu bilgisayardan oluşabileceği gibi, bir ufak gömülü işlemci de olabilir.

Yukarıda da belirttiğim gibi, makine öğrenimi, makinelerin verilerden kendi başlarına öğrenebilmelerini ifade eder. Aslında buradaki "Öğrenme", mevcut veriden belirli bir "örüntü" oluşturmasıdır. Mevcut verideki ilişkileri, etkileşimleri modelleyebilmesidir. Bu örüntü veya model sayesinde yeni gelecek veriyle tahminlerde bulunabilir. Yeni veriye ait tahminlerini bu modele göre yapabilir.

Makinelerin öğrenme tiplerine yakından bakacak olursak, temel olarak 3 başlık altında toplanırlar:

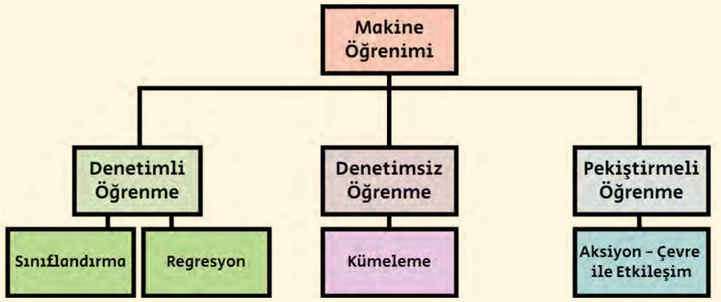
- Denetimli Öğrenme
- Denetimsiz Öğrenme
- Pekiştirmeli Öğrenme

Denetimli öğrenmede, elimizde eğitim için bir veri seti mevcuttur. Bu veri setinde girdiler (sorular) ve çıktılar(cevaplar) biliniyordur ve makine bu girdi ve çıktılara göre bir model oluşturur.

Denetimsiz öğrenmede, elimizde herhangi bir veri seti bulunmaz, girdiler-çıktılar veya sorular-cevapları hakkında bir fikrimiz yoktur, makine mevcut veriye bakarak bir sonuç üretir.

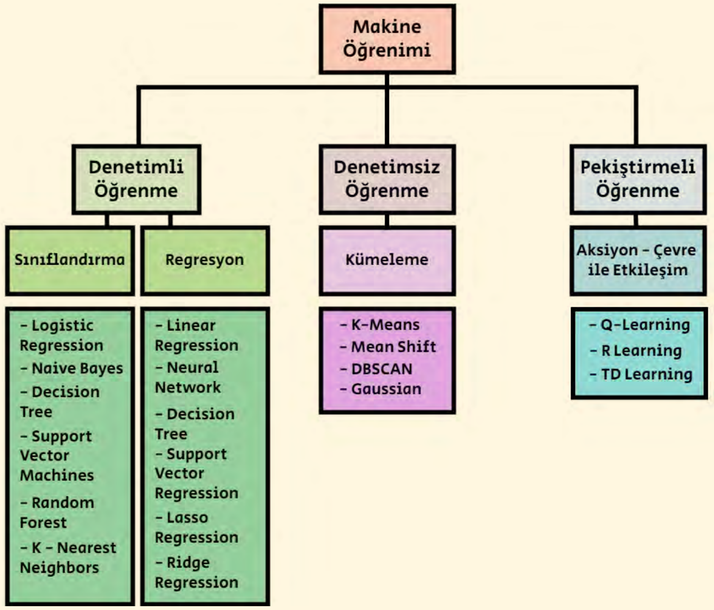
Denetimli ve Denetimsiz öğrenmenin bileşiminden oluşan bir de yarı denetimli öğrenme vardır (Semi-Supervised Learning).

Pekiştirmeli öğrenme ise, etki-tepki, ödül-ceza mantığına dayanır, belirli eylemlere verilen cevaplara dayanarak bir model oluşturur.



Denetimli öğrenme (supervised learning) temel iki soruna çözüm bulmaya çalışır; bunlardan biri sınıflandırma diğeri de Regresyon sorunlarıdır.

Sınıflandırma, herhangi bir verinin hangi sınıfa ait olduğu ile ilgilidir. Regresyonda, verinin değeri ile ilgilidir. Başka bir anlatımla, sınıflandırma problemleri kategori belirtirken, Regresyon problemleri değer belirtir. Her bir sorun için bir takım matematiksel algoritmalar kullanılır. Bu algoritmalar ve eldeki eğitim verisiyle, makine bir model oluşturur. Yeni gelecek soru(n)ların cevapları bu modele dayanarak "tahmin" edilir. Bazı algoritmalar hem regresyon hem de sınıflandırma problemleri için kullanılır.

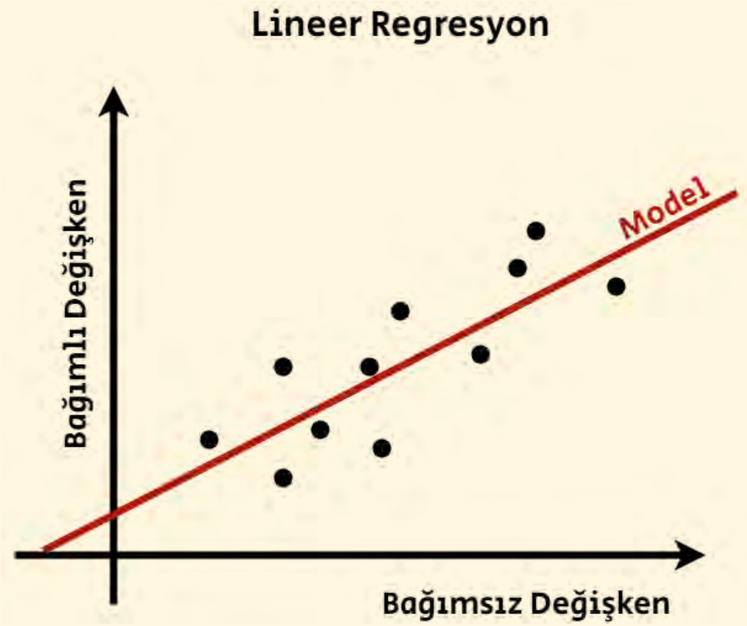


LINEER REGRESYON ALGORİTMASI

Makine öğrenimi regresyon problemlerinde en yaygın kullanılan algoritmadır.

Bağımsız değişkenle(r) (özellikler) ile bağımlı değişken arasındaki ilişkiyi tanımlar. Farklı regresyon algoritmaları vardır. Lineer regresyon bağımlı ve bağımsız değişkenler arasındaki doğrusal ilişkiyi modeller.

Veri dağılımının, bu şekilde bir eğilim göstermesi, lineer regresyon modelinin doğruluğunu arttırır.



Regresyon algoritması, diğer bütün algoritmalarda olduğu gibi, mevcut veriler ve çıktılarına bakarak, olabilecek en optimum genel (ilkel) çizgiyi belirler. Bu çizgi, aslında lineer regresyon modelinin kendisidir.

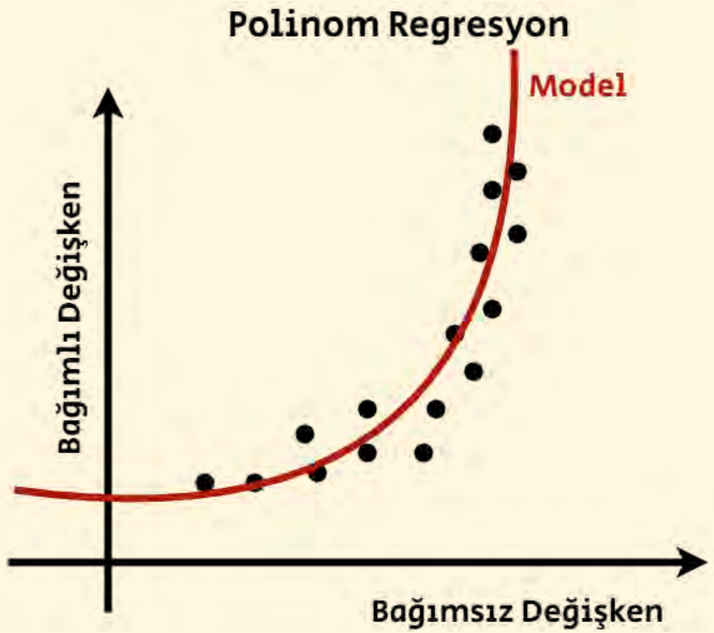
Bu modele dayanarak, yeni gelen veri veya verilerin değeri tahmin edilir. Regresyonun Çoklu doğrusal regresyon ve Polinom regresyon gibi tipleri vardır.

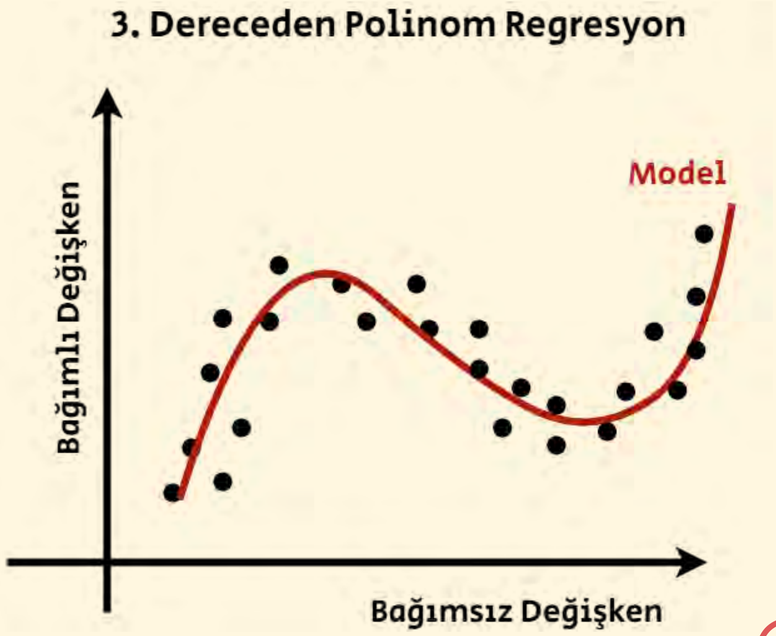
Çoklu doğrusal regresyonda bağımsız değişken sayısı veya özellik birden fazladır. Bunun çıktısı olan bağımlı değişken bir tanedir. Örneğin geçmiş ayların satışlarına göre önümüzdeki ayların satışını tahmin etmek bir bağımsız değişkenli (özellikli) lineer regresyondur, ancak fiyat, mevsim gibi özellikleri de ekleyerek tahmin yapmak istersek bu çoklu linear regresyon problemine girer.

POLİNOM REGRESYON

Polinom regresyonda bağımsız değişken ile bağımlı değişken arasında lineer değil eğrisel bir ilişki vardır. Veri dağılımına göre bazen düz bir çizgi yerine, eğrisel (polinom) çizgiler daha isabetli modellemeler oluştururlar.

Veri dağılımına göre polinomları oluşturan fonksiyonların güçleri artabilir. Örneğin yandaki eğri 2. Dereceden bir polinom iken aşağıdaki eğri 3. Dereceden bir polinomu temsil eder:

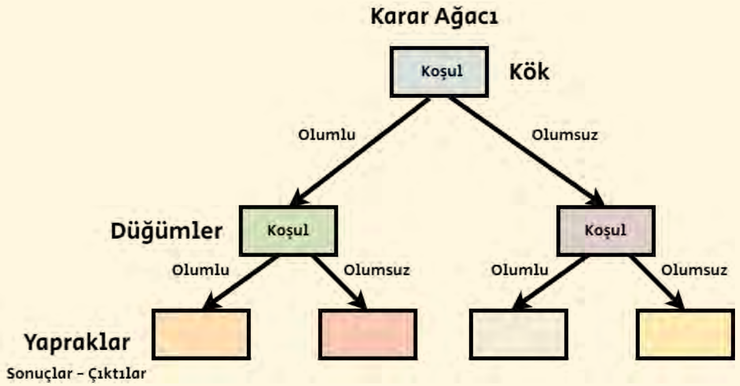




KARAR AĞAÇLARI

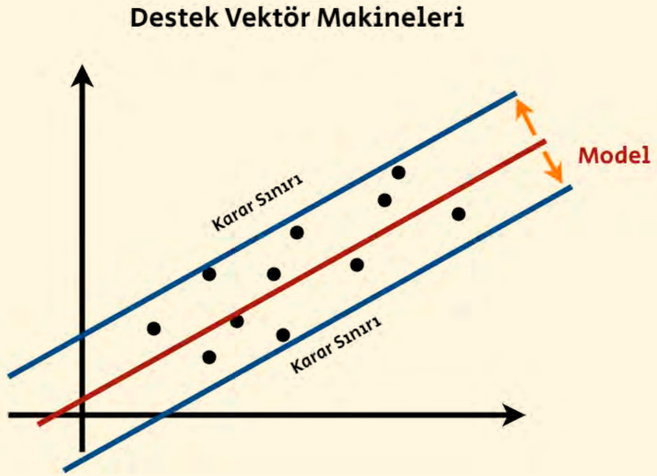
Hem regresyon hem de sınıflandırma için kullanılan bu algoritmada, değişik koşullara göre üretilen çıktılar üzerinden ilerleme sağlanır. Daha fazla bölünemeyen (bilgi döndürmeyen) en son çıktı cevabı oluşturur.

Bu algoritmada bir kök üzerinden ayrışan, ağaç yapısına benzeyen bir yapı vardır. Her bir düğüm farklı olumlu veya olumsuz bir çıktı oluşturur. En son kısımda nihai çıktıların olduğu yaprak denilen son çıktılar bulunur. Bu yapısıyla biraz kural tabanlı sistemlere benzerler:



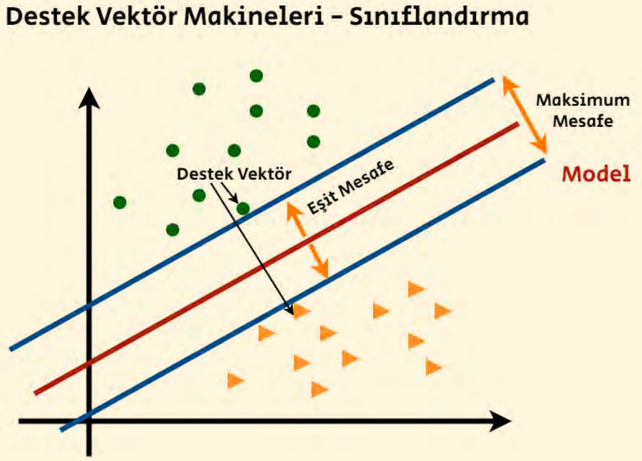
DESTEK VEKTÖR MAKİNELERİ (SUPPORT VECTOR MACHINES)

Destek Vektör Makineleri aynen karar ağaçları gibi hem regresyon hem de sınıflandırma için kullanılan bir algoritmadır. Verideki Destek noktalarına göre işlem yapılan bu algoritma regresyon için kullanıldığında karar sınırlarının aralığını minimum tutarak alabileceği maksimum veriyi kapsamayı hedefler.



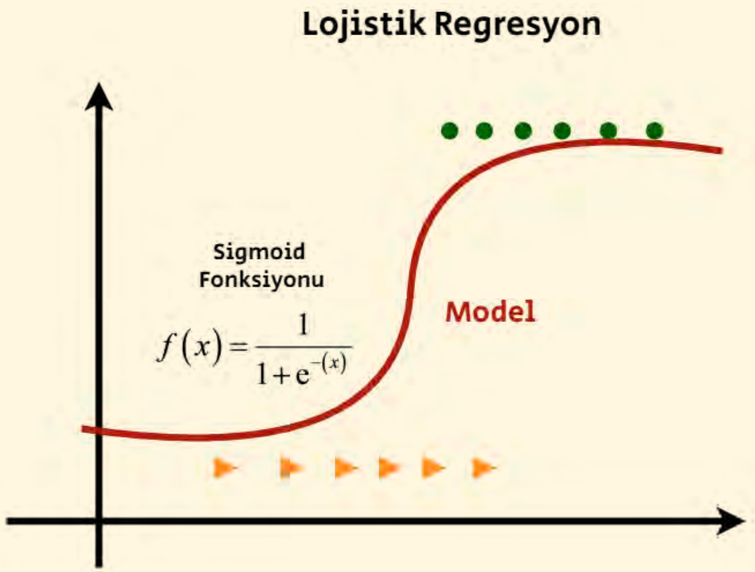
Karar sınırları ile belirlenen bu alanın ortası regresyon çizgisini belirler.

Destek vektör makineleri algoritması sınıflandırma için kullanıldığında, olabilecek en fazla aralığı temsil eden destek verileri (destek vektörleri) tespit edilir, bu aralığı her iki destek noktasına göre eşit bölen çizgi sınıflandırmayı oluşturur.



LOJİSTİK REGRESYON

Adında her ne kadar Regresyon ifadesi olsa da Lojistik regresyon sınıflandırma için kullanılır. İki farklı sınıfı ayıran bu algoritmadaki regresyon ifadesi, sürekliliği belirtmek için kullanılır. Lojistik regresyon algoritması matematikteki sigmoid fonksiyonuna dayanır. S şekline benzer, sürekli bir çizgi oluşturabilen bu fonksiyon, iki sınıfı tek bir formülle ayırabilme özelliğine sahiptir:

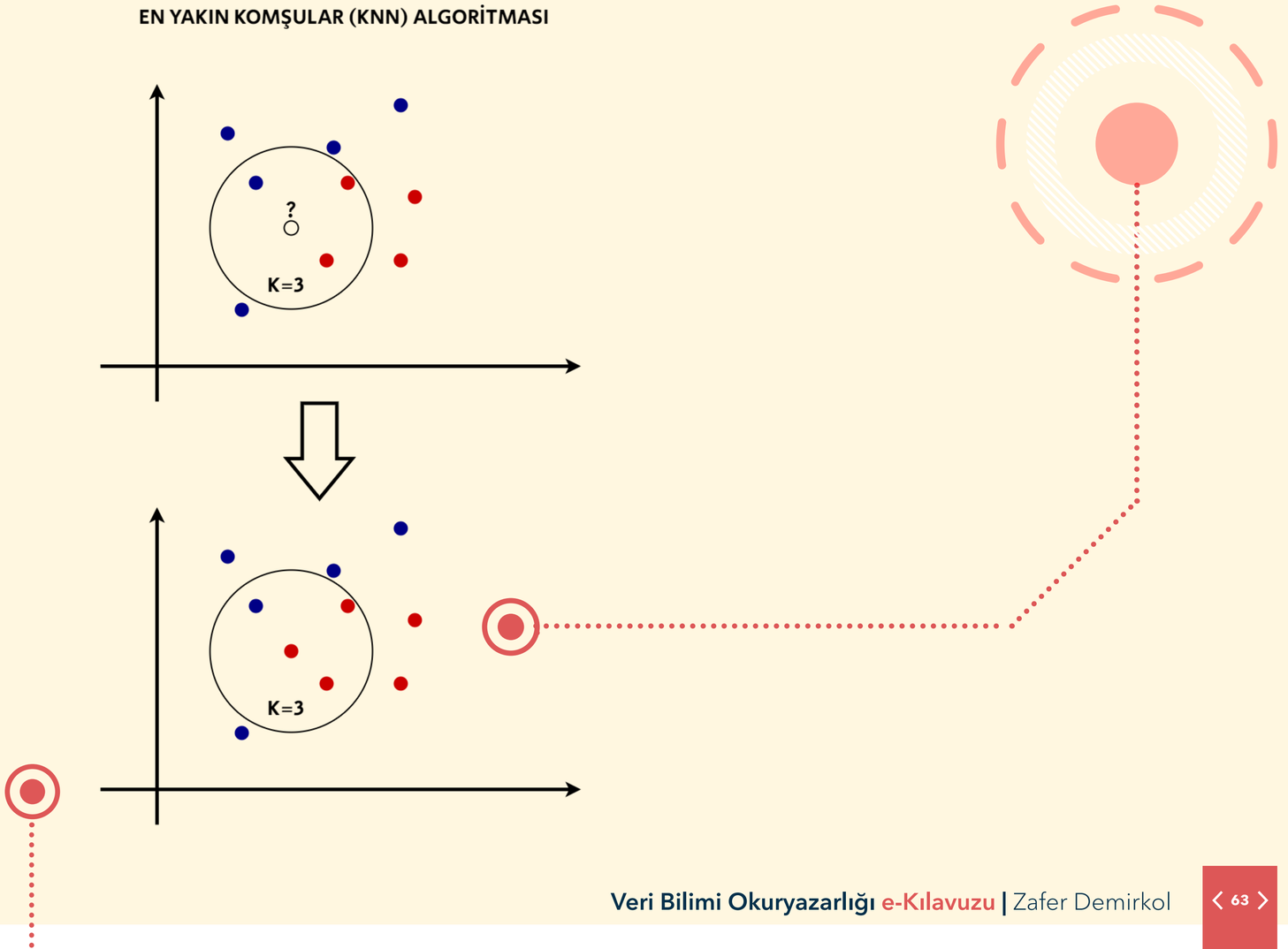


Lojistik regresyonun bir özelliği de olasılık yüzdelik değerleri de döndürebilmesidir.

EN YAKIN KOMŞULAR ALGORİTMASI

Başına bir de “K” harfi alan bu algoritmanın İngilizce karşılığı “k-Nearest Neighbors” dır ve kısaca KNN algoritması olarak adlandırılır. KNN sınıflandırma sorunlarını çözmek için kullanılan bir algoritmadır.

KNN algoritması oldukça basittir. Buna göre, sınıflandırmak istediğimiz verinin, elimizdeki gerçek verilere (eğitim verisi) olan mesafesine bakılır. En yakın olan belirli sayıdaki mevcut verinin sayısına göre hangi sınıfa ait olduğuna karar verilir. Kaç yakın verinin alınacağını K parametresi belirler. Algoritmadaki K parametresi bunu belirtir. Örneğin K parametresi 3 olarak belirlenirse, yeni verinin sınıfı, kendisine en yakın 3 mevcut veri noktasının sayıca fazla olan sınıfına dahil edilerek belirlenir. Buna göre en yakınındaki iki veya üç noktaya sahip olan sınıfa dahil olur.



NAIVE BAYES ALGORİTMASI

Sınıflandırma için kullanılan bu algoritma teoremi oluşturan matematikçinin adıyla anılmaktadır. Bu algoritma olasılık hesaplamasına ve teoremine dayanır. Buna göre A'nin B'ye göre olma olasılığı ile B'nin A'ya göre olma olasılığı arasında matematiksel bir bağlantı vardır ve bu formüle edilebilir. Bu formül şu şekildedir:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

S= Sınıf

Ö= Özellik (Veri)

$$P(S|Ö) = \frac{P(Ö|S) * P(S)}{P(Ö)}$$

Bunu sınıflandırma mantığı ile yorumlarsak, bir özelliğin (verinin) bir sınıfa ait olma özelliği ile, bir sınıfın o özelliğe ait olma olasılıkları arasında ve her birinin genel olarak olma olasılıkları arasında yukarıda belirtilen bir formül vardır. Böylece mevcut verilere (eğitim verisi – denetimli öğrenme) dayalı olasılıklarla yeni verinin hangi sınıfa ait olduğu bu formüle göre hesaplanır.

KÜMELEME ALGORİTMALARI – K-MEANS KÜMELEME ALGORİTMASI

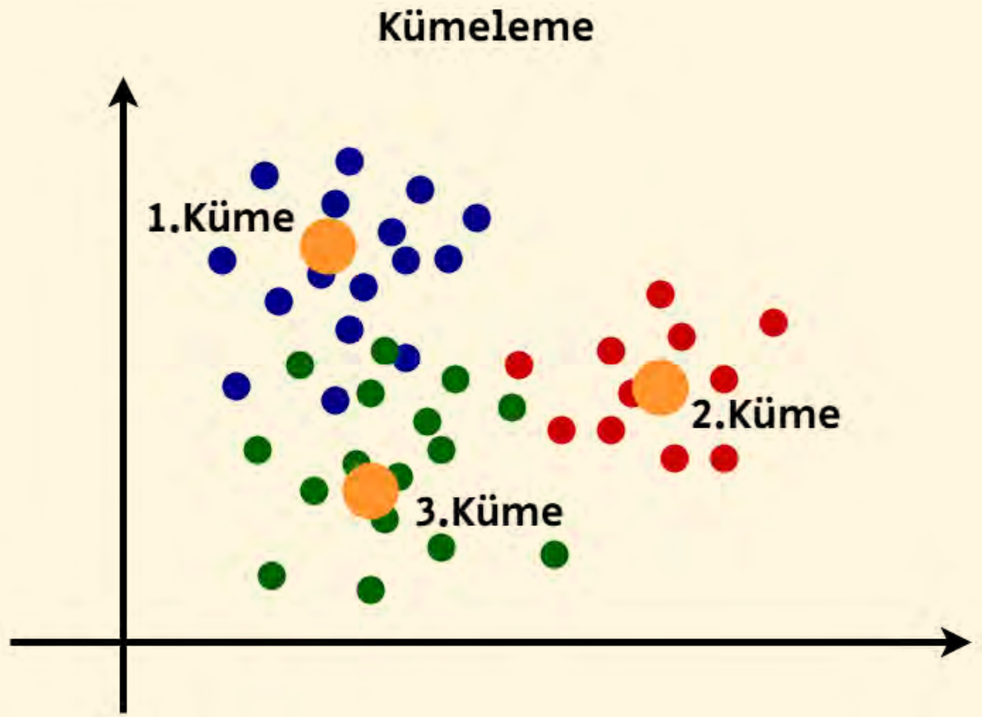
Kümeleme algoritmaları şu ana kadar gördüğümüz algoritmalardan kategorik olarak farklıdır. Bu algoritmalar denetimsiz öğrenme kategorisine girer. Yukarıdaki bölümlerde de bahsedildiği gibi denetimsiz öğrenmede, eğitim amaçlı herhangi bir veri seti bulunmaz. Verinin dağılımına göre, kümelenme alanları tespit edilir. Örneğin müşteri davranışları analizinde, herhangi bir konudaki eğilimlerin belirlenmesinde bu algoritmalar kullanılır. Konu hakkında önceden herhangi bir bilgimiz yoktur. Verinin kümelenmesine göre fikir sahibi olmaya çalışırız.

Kümelemede en yaygın kullanılan algoritma K-Means algoritmasıdır. Algoritmanın yapısı son derece basittir.

Buna göre öncelikle verilerimizi kaç kümeye ayıracağımıza karar vermeliyiz. Bu sayı algoritmanın başındaki "K" ifadesini temsil eder. Genelde 2-3 veya 4 kümeleme oluşturulur.

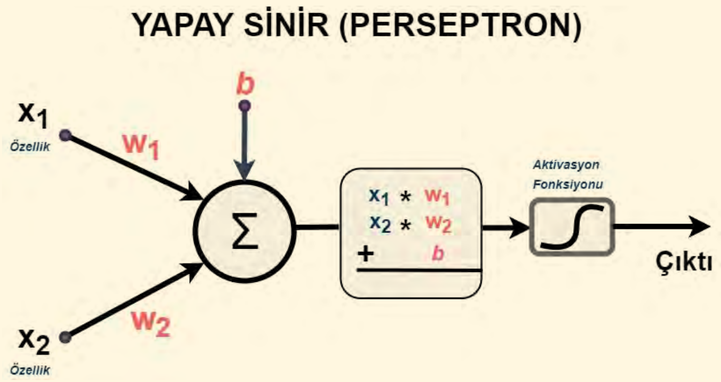
Daha sonra kaç küme oluşturacaksak o küme sayısı kadar noktayı verilerimiz içinde rastgele konumlandırırız. Bir sonraki adımda bu noktalara en yakın veriler tespit edilir ve ilk kümeleme oluşturulur. Mevcut kümelemeye göre noktalarımızın yeni yerleri, kümelerin ağırlık merkezleri olarak tespit edilir. Noktaların bu yeni konumuna göre tekrar en yakın veriler tespit edilir ve bu işlem artık noktalarımızın yerleri değişmeyecek veya çok az değişecek tekrara ulaşıncaya kadar devam eder.

Böylece belirlediğimiz kümeleme sayısına göre veriler en doğru şekilde bir kümeye dahil edilir ve kümeler oluşturulmuş olur. Küme noktaları ve veriler arasındaki hesaplamalar genelde Öklid-Pisagor teoremlerine dayalı mesafe hesaplamalarıyla pek çok tekrarla yapılır.



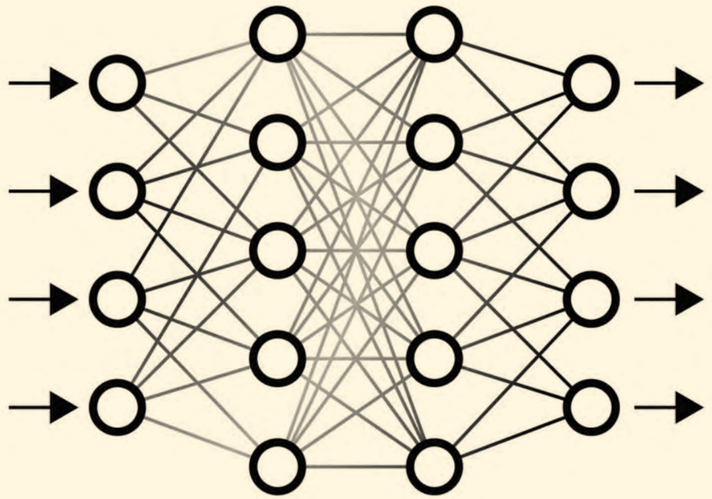
GÜNÜMÜZ YAPAY ZEKA TEKNOLOJİ VE DİSİPLİNLERİ

Günümüz Yapay Zeka teknoloji ve disiplinleri, "Derin Öğrenme" ve "Yapay Sinir" ağlarına dayalı, büyük veriden öğrenebilen bir yapıya sahiptir. Bu teknoloji kendi içinde de "Bilgisayar Görmesi" (Computer Vision), Doğal Dil işleme (NLP) vb. alt alanlara sahiptir. Derin öğrenme altyapısı, mevcut verilere dayalı bir eğitim sürecinden geçtiği için, denetimli öğrenme algoritmaları içinde yer alır (istisnai bazı durumlar dışında). Yapay Sinir ağlarının temelini "Yapay Sinir" oluşturur. Yapay Sinir "Perceptron" olarak da adlandırılır.

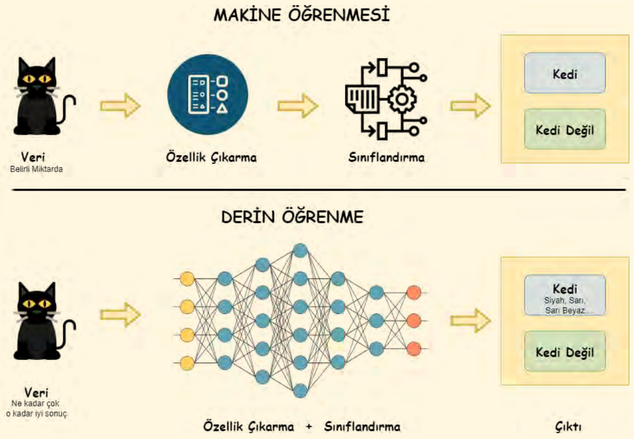


Yapay Sinirler bir araya gelerek "Yapay Sinir Ağları" nı oluştururlar.

Bu ağlar değişik mimarilerde olabilirler. Bu mimarileri belirleyen de problemin kategorisidir.



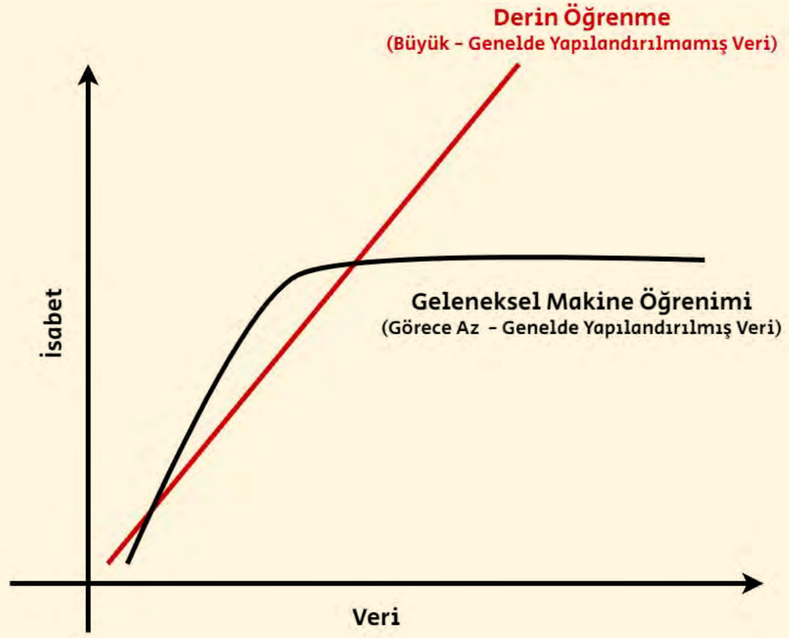
Yapay Sinir ağları hem Regresyon hem de Sınıflandırma sorunlarının çözümünde kullanılırlar. Geleneksel Makine Öğrenimi algoritmalarının tersine, öncesinde bir özellik çıkarımına ve uzmanlık bilgisine ihtiyaç duymazlar.



Geleneksel makine öğreniminin çözemediği, özellikle yapılandırılmamış veriye dayalı problemleri çözebilirler. Örneğin Yapay Sinir Ağları algoritmasına dayalı Bilgisayar Görmesi (Computer Visison) alanı, geleneksel Makine Öğrenimine göre açık ara farkla isabetli modeller oluşturabilmektedir. Bilgisayar görmesi alanında geleneksel Makine Öğrenimi algoritmaları yetersizdir, bir nesneyi tanımlayamazlar. Ancak Yapay Sinir Ağları günümüzde bir nesneyi insan görüşünden daha isabetli bir şekilde tanımlayabilmektedir (görebilmektedir). Bu tip örnekleri farklı alanlar için de çoğaltmak mümkündür. Örneğin Doğal Dil İşleme alanı (NLP - Natural Language Processing). Burada da özellikle Yapay Sinir Ağları algoritmalarının kullanılmasından sonra çok isabetli ve şaşırtıcı sonuçlar alınmıştır. Günümüzde daha pek çok alanda Derin Öğrenme ve Yapay Sinir Ağları kullanılmakta ve bu algoritma ve mimariler günümüz "Yapay Zeka" uygulamalarının altyapısını oluşturmaktadır.

GELENEKSEL MAKİNE ÖĞRENİMİ – YAPAY SİNİR AĞLARI / DERİN ÖĞRENME

Yapay Sinir Ağlarının geleneksel Makine Öğrenimi algoritmalarına göre en ayırt edici özelliklerinden biri, Yapay Sinir Ağlarının çok fazla veriye ihtiyaç duymasıdır. Yapay Sinir Ağları günümüzde "Büyük Veri" olarak adlandıran çoğunlukla yapılandırılmamış verilerle çok isabetli modeller oluşturabilmektedir. Diğer bir ifadeyle ne kadar çok veri olursa o kadar fazla isabetli modeller oluşturabilmektedir. Geleneksel Makine Öğrenimi modelleri ise görece daha az veriye ihtiyaç duyarlar ve büyük bir çoğunlukla da yapılandırılmış veri setleri üzerinde çalışabilirler. Geleneksel Makine Öğrenimi algoritmalarına dayalı model oluşturmalarda belirli bir veri miktarından sonra, algoritmaya daha fazla veri eklemek genelde performansı arttırmaz. Diğer bir ifadeyle doyum noktasından sonra veri arttıkça -çoğunlukla- isabet artmaz.



SAS HAKKINDA

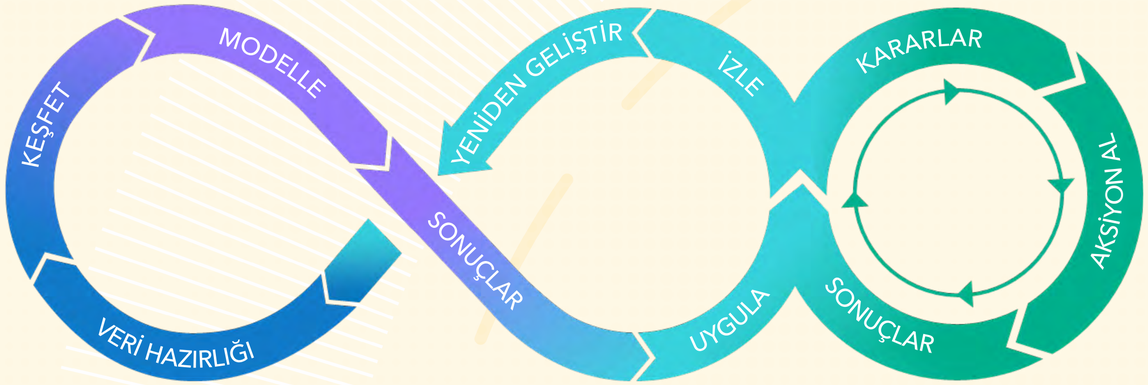
SAS analitik alanında liderdir.
SAS, yenilikçi yazılım ve hizmetlerle dünyanın her yerinden müşterilere veriyi zekaya dönüştürmeleri için destek ve ilham verir. SAS, dünyadaki müşterilerine "**THE POWER TO KNOW**" anlayışıyla hizmet vermektedir.

ANALİZ ET
Mevcut durumu iyi analiz edin.

İYİLEŞTİR
Vatandaşların ihtiyacı olan hizmetleri belirleyin.

YENİDEN KURGULA
Kurumların kamuya hizmet etme şekillerini yeniden tasarlayın.

Veriyi anlamlı hale getirmek için analitiğe ihtiyaç vardır > Anlamın değer bulabilmesi için kararlara ihtiyacı vardır.



ANALİTİK — BT — İŞ

- Karar odaklı
- Geliştirici-dostu
- Otomatikleştirilmiş
- Yönetilen
- Sürekli güncellenen
- Bulutta

İŞ VAKALARI

- Durumsal Farkındalık ve Kritik Görev Analizi
- Tıbbı Kaynak Optimizasyonu
- Temas İzleme Optimizasyonu
- Ekonomik İyileşme Analizleri
- Sosyal Yardımlar Acil Durum Programı
- Nüfus Hareketliliği

sas.com