[Şekil 1 Veri Bilimi Bileşenleri 3](file:///C:\Users\IlknurKaraduman\Desktop\bitirme_ilknur.docx#_Toc39332317)

[Şekil 2 Gartner Analytic Ascendancy Model (Gartner, 2012, kaynak) 4](file:///C:\Users\IlknurKaraduman\Desktop\bitirme_ilknur.docx#_Toc39332318)

[Şekil 3 CRISP-DM Adımları ve Akışı (21, kaynak) 5](file:///C:\Users\IlknurKaraduman\Desktop\bitirme_ilknur.docx#_Toc39332319)

[Şekil 4 Çalışan İşten Ayrılma/ Ayrılmama Yüzdesi 10](file:///C:\Users\IlknurKaraduman\Desktop\bitirme_ilknur.docx#_Toc39332320)

[Şekil 5 Kutu Grafiği ile Veri Görselleştime 10](file:///C:\Users\IlknurKaraduman\Desktop\bitirme_ilknur.docx#_Toc39332321)

[Şekil 6 Plotly Parallel Coordinates Plot ile Veri Görselleştirme 11](file:///C:\Users\IlknurKaraduman\Desktop\bitirme_ilknur.docx#_Toc39332322)

# GİRİŞ

## Amaç ve Hedef

# VERİ BİLİMİ

Veri bilimi veriden faydalı bilgi çıkarma sürecidir. Elimizde bulunan veri kaynakları üzerinden analitik düşünce becerilerimiz ve veri analitiğini kullanarak faydalı bilgiler elde ederiz. Bu bilgileri aksiyona çevirdiğimiz uçtan uca sürece veri bilimi denir. Veriden faydalı bilgi çıkarma sürecini yöneten kişi ise veri bilimcidir.

Matematik & İstatistik, Bilgisayar Bilimleri ve İş-Sektör Bilgisi, veri bilimini meydana getiren bileşenlerdir.

## Veri Analitiği

Şekil 1 Veri Bilimi Bileşenleri

Veriden faydalı bilgi çıkarmak için analitik türleri kullanılır.

### Betimleyici Analitik (Descriptive Analytics)

Elimizdeki veriyi betimlediğimizde (ortalama, standart sapma, mod, medyan ve veri görselleştirme teknikleri) oluşturulan raporlar mevcut durumu göz önüne sunar. Problemi ve veriyi anlamayı sağlar. Veriyi tanımladığımız analizdir. Ne oldu sorusuna yanıt aranır.

### Teşhis / Tanı Analitiği (Diagnostic Analytics)

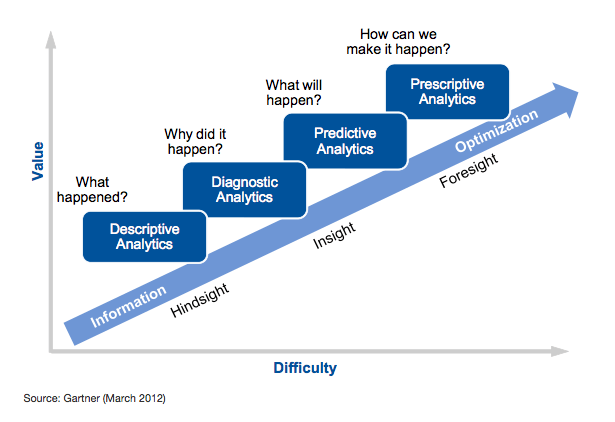
Veriyi betimlediğimiz zaman tespit ettiğimiz gözlemlerin yani veriden çıkarmış olduğumuz anlamlara neden-sonuç bağlamında sorgulamalar yaptığımızda teşhis/tanı durumunun konulmasını içeren analizdir. Neden oldu sorusuna yanıt aranır.

### Tahminsel Analitik (Predictive Analytics)

Gelecek ile ilgili tahminler yapmak için kullanılır. Ne olacak sorusuna yanıt aranır. Burada istatistiksel modellemeler, ekonometrik modellemeler ve makine öğrenmesi modelleri kullanılır.

### Yönergeli Analitik (Prescriptive Analytics)

Geleceğe yönelik tahmin değerleri göz önüne alınarak ne yapılması gerektiği sonucuna varan analizdir. Ne, nasıl olmalı sorularına yanıt aranır. Bu doğrultuda alınan iş/aksiyon kararları uygulanır.



Şekil 2 Gartner Analytic Ascendancy Model (Gartner, 2012, [kaynak](http://www.ithappens.nu/levels-of-data-analytics/))

### CRISP-DM

Veri madenciliği için sektörler arası standart süreci ele alan modeldir. Veri bilimi süreçlerinin iş dünyasındaki ele alınışını ifade eder. Şirketler genellikle bu döngüyü takip etmektedir. CRISP-DM metodolojisi, veriden doğru sonuçları elde etmeyi amaçlayan altı adımdan oluşan döngüsel bir süreçten ibarettir.

İşletme anlayışı (Business understanding), gerçekleştirilen proje hedeflerini ve gereksinimlerini anlamak ve verileri kullanmak için gereken bir sorun tanımlamaktır. Yani bu aşamada iş süreçlerini anlayarak problemin ne olduğu ve problem çözümü neticesinde ne gibi çıktıların beklendiği tanımlanır. Öncelikle problemi anlamak daha sonra veriyi anlamak gereklidir. Veriyi anlama (Data understanding), veri toplama işlemi sonrası veri setini bütünsel olarak gözden geçirerek çözüme ulaşmayı kolaylaştırır. Veri setindeki gözlem, öznitelik, veri tipleri ve eksik veya gürültülü veri sayısı gibi bilgiler bu aşamada tespit edilir. Gerekli olduğu taktirde ilave veri toplama veya mevcut veriyi zenginleştirme işlemleri yapılır. Veri hazırlama (Data preparation), ham veriden nihai veri setini oluşturmaktır. Veri bilimi sürecinin büyük bir kısmını kapsar. Veri üzerinde yapılacak işlemlere ve bu işlemlerin hangi yöntemlerle yapılacağına karar verilir. Modelleme (Modeling) aşamasında, tanımlanan problemin çözümüne yönelik bir makine öğrenmesi veya istatistiksel model geliştirilir. Değerlendirme (Evaluation) aşamasında, tüm adımların genel bir değerlendirmesi yapılır. İlk aşamada problem çözümü neticesinde belirlenen çıktıları ne ölçüde sağladığı test edilir. Modelleme ve değerlendirme, uygun algoritmayı seçerek oluşturulan modeli eğitim seti üzerinden eğitip sonrasında doğruluğunu doğrulama seti üzerinden test etmektir. Son aşama olan gerçekleme ise ortaya çıkan sonuçta bir başarım elde ediliyor ise kullanıma sokmaktır. Aşağıdaki şekilde görüleceği üzere bazı adımlar arasında döngüler mümkündür ve bunların proje yönetimi açısından önemi yüksektir.

### Çevik (Agile)

Şekil 3 CRISP-DM Adımları ve Akışı (21, [kaynak](http://veribilimlerienstitusu.com/veri-madenciligi-proje-dongusu-crisp-dm/))

Zaman içinde projelerin karmaşıklaşması, teknolojinin gelişimi gibi pek çok problemler proje sürecinin yönetilmesini zorlaştırır. Agile, çözümlerin ekipler arasındaki iş birliği ile geliştiği yinelemeli gelişim fikrini merkez alan yazılım geliştirme metodolojilerini ifade eder. Tekrarlanan kısa döngüler ile proje yönetimini sağlar.

(Scrum/sprint….)

## Görsel Analitik

Görsel analitik temel olarak değerlendirme, planlama ve karar vermeyi doğrudan destekleyen daha derin kavrayışlar elde etmeyi sağlayan analitik akıl yürütme tekniklerine odaklanır. Basit bir grafik veri analistinin zihnine diğer herhangi bir cihazdan daha fazla bilgi getirir (John Tukey).

Verinin belli bir kısmı için değerlendirmeler yaptığımızda doğru sonuca ulaşamayabiliriz. Veriyi doğru temsil ettiğimizde büyük resmi çok net görebiliriz. Çalışmaya başlamadan önce verinin ana hatlarının görsel teknikler kullanılarak ortaya koyulması gerekir. Elimizde çok karmaşık bir veri bulunsa bile görselleştirme sonucunda verinin anlamına çok basit bir şekilde ulaşılabiliriz.

### Veri Görselleştirmenin Önemi

(Veriyi anlama / aykırı değerleri daha kolay farketmek vb bahset)

# PROBLEME YAKLAŞIM

## Projenin Yapısı

Gerçekleştirmiş olduğum proje üç farklı veri bilimi çalışmasından oluşmaktadır.

Birinci sprint içerisinde veriyi anlama, veriye bakınca anlaşılamayan bilgileri çeşitli veri görselleştirme teknikleri kullanarak ortaya çıkarma, analiz etme ve yorumlama süreçlerini global bir konu olan COVID -19 ile ilgili veri seti üzerinde gerçekleştirdim.

İkinci sprint kapsamında bir insan kaynakları veri setini kullanarak şirketteki çalışanların işten ayrılıp ayrılmama durumunu tahmin eden makine öğrenmesi modelini oluşturdum. Birinci sprintte olduğu gibi burada da problemi anladıktan sonra veriyi anlama ve veri görselleştirme ile farklı anlamlar çıkarma adımlarını uyguladım. Veri ön işleme yaparak makine öğrenmesi modeli için veriyi uygun hale getirdim. Daha sonra oluşturduğum modeli eğitim ve test aşaması sonucunda modelimin başarı durumunu değerlendirdim.

Üçüncü sprint içerisinde ise farklı derin öğrenme algoritma ve yöntemleri ile zaman serisi analizi gerçekleştirdim. Bu aşamada TCMB verilerini kullandım.

## Projenin Programlama Dili

Basit bir söz dizimi ile yorumlanan, nesneye dayalı, öğrenmesi kolay, açık kaynak kodlu bir programlama dilidir. Ustalaşması en kolay dillerden biridir. Veri analizi sürecini çok hızlı ve kolay hale getiren pek çok kütüphane barındırmaktadır. R ve Matlab analitik amaçlar için iyi olsalar da üretim ortamlarına kolayca entegre olmazlar. Veri bilimciler için çeşitli veri bilimi projeleri / uygulamaları için kullanılan en iyi dillerden biri olması nedeniyle Python’ı tercih ettim.

## Geliştirilen Ortam

Jupyter Notebooks çok yönlü, paylaşılabilir, dönüştürülmesi kolay ve aynı ortamda kod, görsel, metin, grafik oluşturmayı tek bir yerde sağlayan web tabanlı bir interaktif ortamdır. Veri bilimcilerinin düzenli bir şekilde hem kodlama hem de raporlama yaparak belge oluşturmalarına ve paylaşmalarına olanak tanır. Ayrı ayrı bağımsız bloklar içinde kod gerçekleştirilebilir. Böylece, kod en baştan yürütülmeden belirli bir kod bloğu üzerinde işlem tamamlanır. Parametrelere ince ayar yapmak gibi değişiklikler için kod blokları kolaylık sağlar. Veri bilimi iş süreçlerini gerçekleştirmek için kullanışlı bir araç olduğundan bu ortamı tercih ettim. Hem Python hem de Jupyter Notebooks için Anaconda dağıtımını kulandım. Anaconda her ikisinin de kurulumunu gerçekleştirmektedir.

## Kütüphaneler

Bazı özelleştirilmiş kütüphaneler bulunmaktadır. Eğer Python ile veri bilimi projesi yapmak istiyorsanız bu kütüphaneler ihtiyaçlarınızı gidermenizi sağlar.

(Dask bahsedilebilir)

### Numpy

NumPy (Numerical Python), Python ile bilimsel hesaplamaları hızlı bir şekilde yapabileceğimiz bir matematiksel kütüphanedir. Python’ın sayısal işlemlerinin yetersiz kaldığı anlarda kullanılır. Python listelerinden farkı örneğin n elemanlı bir liste için n adet veri tipi bilgisi tutulurken n elemanlı bir dizi için sabitlenmiş tip olarak bir adet veri tipi bilgisi tutulmaktadır yani numpy dizileri (array) veriyi daha verimli biçimde saklamaktır. Diziler ve matrisler üzerinde yüksek performans sağlar. Çok karmaşık matematiksel işlemleri çok daha kolay gerçekleştirmemizi sağlar.

### Pandas

Veri bilimi sürecinin büyük bir kısmının veriyi hazırlamak olduğu öngörülmektedir. Çünkü eğer kullanacağınız veri kötü ise veri bilimi projenizin sonucunda başarısız olmanızın olasılığı yüksektir. Pandas veri manipülasyonu için kullanılan bir kütüphanedir. Veri keşfi ve veri temizleme aşamalarını kolaylaştırır. Python ile verilerinizi anlamak ve üzerinde birtakım işlemler yapmak için oldukça yararlıdır. Pandas ile veri analizi için temel veri yapıları Series ve DataFrame’lerdir. Pandas ile veri çerçeveleri oluşturabilir, verileri birleştirebilir, yeniden şekillendirebilir, veri tiplerini düzenleyebilir, aykırı değerleri silebilir, boş değerleri doldurabilir yani verinizi istediğiniz gibi analiz edebilirsiniz. Pandas özellikle zaman indeksli veriler açısından çok uygun araçlara sahiptir.

### Matplotlib

Numpy dizileri üzerine kurulmuş, John D. Hunter tarafından yazılmış, bir veri görselleştirme kütüphanesidir. Verilerinizin çizgi, histogram, dağılım gibi çeşitli grafiklerini oluşturabilirsiniz. Ayrıca pandas içerisine entegre edilmiş plot özelliğini kullanabilirsiniz. Yazı tipleri, çizgi stilleri, grafik boyutu, kullanılacak renkler, başlık ve eksen isimleri gibi tanımlamalar birçok ek kod satırı ile yapılsa da kullanıcı tarafından tam kişiselleştirmeye yöneliktir. Üçüncü taraf paketlerin üzerine eklenmesiyle kapsamlı, hızlı bir şekilde statik görselleştirmeler üretilebilir.

### Seaborn

Matplotlib üzerine inşa edilmiş, Python veri görselleştirme kütüphanesidir. Daha az kod ile daha spesifik grafikler oluşturabilirsiniz. Seaborn bir Pandas veri çerçevesi (dataframe) ile çalışmak için daha entegredir.

### Plotly

Bir başka veri görselleştirme kütüphanesi ise Plotly. Çizgi, çubuk, pasta grafiği gibi temel grafiklerin yanında, histogram, kutu grafiği gibi istatistiksel grafikler, ısı haritaları, kontur grafikleri gibi bilimsel tablolar, finansal tablolar, haritalar, 3D grafikler, animasyonlar ve daha fazlasını oluşturabilirsiniz. Diğer veri görselleştirme kütüphanelerinden en büyük farkı etkileşimli grafiklerin oluşturulmasıdır. Yani biz bir grafiğin üzerinde gezdiğimizde o noktaya ait bilgilerin bize sunulması, grafiği yakınlaştırma, belirli aralıklarını görme, grafiği dinamik olarak istenilen etiketler için oluşturma gibi imkanlar tanır. Ayrıntılı grafikler oluşturmak için iyi bir tercihtir.

(arkasında javascript çalışıyor bunu belirt)

### Folium

Uzamsal, coğrafi verileri görselleştirmek için kullanılan bir Python kütüphanesidir.

### Scikit-Learn

Makine öğrenmesi kütüphanesidir. Sınıflandırma, regresyon, kümeleme gibi problemler için doğrusal regresyon (linear regression), lojistik regresyon (logistic regression), destek vektör matrisleri (support vector machine), rastgele ormanlar (random forests), k-en yakın komşu (k-neighbours) ve benzeri algoritmalar içermektedir.

### Scipy

SciPy, matematik, bilim ve mühendislik için çok sayıda farklı problemlerin çözümünü içeren bir kütüphanedir. İstatistik, optimizasyon, lineer cebir, enterpolasyon gibi çok yaygın kullanılan işlemler içerir.

### Statsmodels

İstatistiksel model tahmini, istatistiksel testler, istatistiksel veri araştırması için sınıflar ve işlevler sağlayan kısacası istatistiksel hesaplamalar için bir Python modülüdür. Örneğin elinizde zaman serisi verisi var ise ARIMA ve benzeri yöntemleri kullanarak

### Facebook Prophet

Facebook araştırmacıları tarafından oluşturulan hem Python hem de R'de bulunan Prophet, doğrusal olmayan trendlerin günlük, haftalık, yıllık mevsimsellik ve tatillerin etkilerine uygun bir modele dayanan zaman serisi verileri tahmini için kullanılır. Verilerdeki değişim noktalarını seçerek trendlerdeki değişiklikleri kavrar. Mevsimsel etkileri güçlü olan ve birkaç geçmiş mevsim verisi bulunan zaman serileriyle daha iyi sonucu verir. Prophet, eksik veri ve trenddeki kaymalara karşı dayanıklıdır ve aykırı değerleri iyi işler.

### Keras

Google mühendisi olan François Chollet tarafından geliştirilmiştir. TensorFlow, CNTK ve Theano üstünde çalışan, ücretsiz, açık kaynaklı, Python ile yazılmış bir sinir ağı kütüphanesidir. Favori derin öğrenme kütüphanesi olan Keras, geniş kapsamlı ve genişletilmesi kolaydır. Basit ve modüler olduğundan yeni başlayanlar için en iyisidir. CPU ve GPU'da sorunsuz çalışır. Amaç yapay sinir ağlarının kolay ve hızlı prototiplenmesidir. Keras esnektir, tekrarlayan sinir ağları, evrişimli sinir ağları, derin sinir ağları ve benzeri mimarileri inşa etmenizi sağlar. Model oluşturup, eğitmeyi hızlı ve kolay hale getirir.

## Modeller

# UYGULAMALAR

## Sprint 1: Veri Görselleştirmesi

### Veri Seti

### Kullanılan Kütüphaneler

### Sonuçlar

## Sprint 2: İnsan Kaynakları Analizi

### Veri Seti

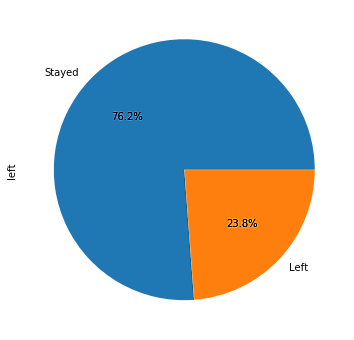
Bir çalışanın bir kuruluş içinde tutulmasını öngören şirketler, ayrılma potansiyelindeki çalışanlarını geri kazanmak isterler. Şirketlerin büyüme gerçekleştirmesi için büyük oranda personelin elde tutulması gerekir. Çalışanların kaybedilmesi hem kuruluşun moralini, inancını etkiler hem de yeni çalışana harcanılan zaman ve diğer etkenler neticesinde mevcut çalışanı elde tutmak yeni bir çalışanın işe alınmasından daha makuldür. Uyguladıkları veri bilimi çalışması sonucunda işten ayrılacağı tahmininin yapıldığı çalışanlar öngörülerek insan kaynakları yönetiminin daha iyi bir elde tutma planları oluşturması, çalışan memnuniyetinin artması sağlanır. Şirketler için çalışan memnuniyeti oldukça önemlidir. Örneğin işten ayrılacağı belirtilen çalışanların bilgilerine bakarak çok fazla saat mi çalışıyor, çok fazla proje sorumluluğu mu var gibi sorunları görebilir, iyileştirmeler yapılabilir.

İkinci sprint kapsamında çalışanın şirketten ayrılıp ayrılmayacağını sınıflandırabilen makine öğrenmesi modelinin oluşturulması için ele aldığım CSV formatındaki insan kaynakları verisi, 11 adet özellik ve 14999 kayıt içermektedir. Yani veri setinde her satır bir çalışanı, her sütun ise çalışan özelliklerini ifade etmektedir. Çalışan özellikleri aşağıdaki gibidir.

|  |  |  |
| --- | --- | --- |
| Özellik | Açıklaması | Veri Tipi |
| empid | Çalışan ID'si | int64 |
| satisfaction\_level | Çalışan memnuniyet seviyesi (0-1 arasında değer) | float64 |
| last\_evaluation | Çalışanın işi hakkındaki son değerlendirmesi (0-1 arasında değer) | float64 |
| number\_project | Çalışanın şirkette hazırladığı projeler | float64 |
| average\_montly\_hours | Çalışanın aylık ortalama çalıştığı saat | float64 |
| time\_spend\_company | Çalışanın şirkette çalıştığı süre | float64 |
| Work\_accident | Çalışan kaza geçirdi mi? (1: Evet, 0: Hayır) | float64 |
| left | Çalışan şirketten ayrıldı mı? (1: Evet, 0: Hayır) | float64 |
| promotion\_last\_5years | Çalışan son 5 yılda terfi aldı mı? (1: Evet, 0: Hayır) | int64 |
| dept | Çalışanın çalıştığı departman | object |
| salary | Çalışanın maaşı (High/Medium/Low) | object |

Tablo 1 İnsan Kaynakları Veri Seti Özellikleri

Bağımlı yani hedef değişken çalışanın şirketten ayrılma durumunu bildiren left özniteliğidir. Veride 3571 adet pozitif, 11427 adet negatif örnek vardır. Pozitif örnekler işten ayrılma durumunu, negatif örnekler ise işte kalma durumunu kastetmektedir.

Veride toplam 21 adet boş değer bulunmaktadır. Bu eksik verileri istersek direkt olarak veri setinden atabilir ya da ilgili yerin ortalama, medyan gibi değerleri ile doldurabiliriz. Örneğin bir sütun yani özellik %70’ten fazla boş değer bulunduruyor ise en mantıklısı o sütunu tamamen silmektir. Departman, işten ayrılma durumu ve iş kazası geçirme durumu NaN olan kayıtları veri setinden sildim. Sayısal değişkenlerin NaN değerlerini, işten ayrılma durumu, iş kazası geçirme durumu ve bulunduğu departmana göre gruplayarak ortalama değer ile doldurdum. Böylece veri setinde boş değer kalmamış oldu. Ayrıca empid sütununun değerlerinin hepsi benzersiz (unique) olduğundan daha doğru bir model oluşturmak için bu özelliği kaldırdım.

Şekil 4 Çalışan İşten Ayrılma/ Ayrılmama Yüzdesi

Veri setinde boş değerler kalmasa da aykırı değerler söz konusu olduğundan veri en iyi durumda değildir. Çünkü eğitilen makine öğrenmesi modelinde yanlılığa neden olur, genelleme kaygısı oluşturur. Aykırı gözlem, verideki genelin dışına çıkan, diğerlerinden çok farklı olan gözlemlere yani aykırı değeri barından gözleme denir. Aykırı değer ise aykırılığı belirten değerdir. Bu aykırı değerleri de anlamanın en kolay yolu veri görselleştirme teknikleridir. Mesela her özellik için kutu grafiği çizdirdiğimiz zaman aykırı değerler göze çarpar, maksimum ve minimum değerlerini çok daha rahat görürüz.

Şekil 5 Kutu Grafiği ile Veri Görselleştime

A picture containing map, umbrella, rain

Description automatically generatedDescribe işlevi ile de verinin özet istatistiklerini gözden geçirerek aykırı değerler bulunabilir. Ele aldığım veri seti için describe tablosuna bakılarak değer aralığı 0-1 olması gereken satisfaction\_level ve last\_evaluation özniteliklerinin maksimum değerlerinin 1’den büyük olduğu ve proje sayısının 55 gibi çok büyük bir değer olduğu fark edilir. Bu özniteliklerin 1’den büyük olan değerlerini içeren kayıtları veri setinden silebilir veya ortalama gibi doldurma yöntemleri kullanılarak doldurulabilir. Aşağıda bulanan bir başka görselleştirme ile de aykırı gözlemler dikkat çekmektedir.

Şekil 6 Plotly Parallel Coordinates Plot ile Veri Görselleştirme

### Kullanılan Kütüphaneler

### Sonuçlar

## Sprint 3: Zaman Serisi Analizi

### Veri Seti

### Kullanılan Kütüphaneler

### Sonuçlar

# SONUÇ

Sonuçlar ve öneriler