1-How would you define Machine Learning?

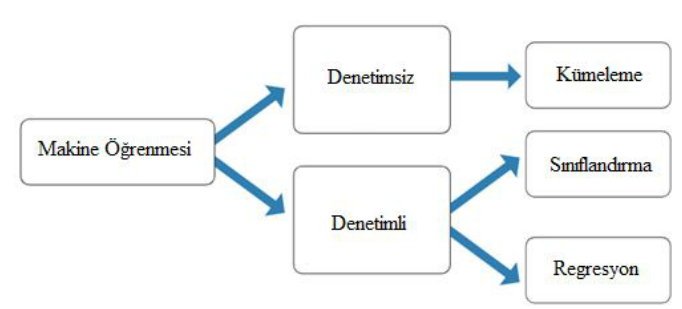
Makine öğrenmesi yapay zekânın bir alt bilim dalıdır. Bir tanım vermek gerekirse;

Matematiksel ve istatistiksel işlemler ile veriler üzerinden çıkarımlar yaparak tahminlerde bulunan sistemlerin bilgisayarlar ile modellenmesidir.

2) What are the differences between Supervised and Unsupervised Learning? Specify example 3 algorithms for each of these.

Makine Öğrenimi, analitik modellerin yetersiz kaldığı durumlarda sorunları çözmek için oluşturulmuştur. Denklemler ve yasalar ümit vaat etmediğinde eğitim verilerini kullanarak bir model elde etmek için Makine Öğrenimi tekniklerinden yararlanırız.

Makine öğrenimi algoritmaları, [insan](https://www.elektrikport.com/teknik-kutuphane/elektrik-teknolojisi-insan-hayatini-nasil-degistirdi/22420#ad-image-0) müdahalesi olmadan verilerden öğrenebilen ve deneyimler ile geliştirebilen programlardır.   Makine öğrenimi [algoritma](https://www.elektrikport.com/haber-roportaj/algoritma-ile-robotlara-yakinindaki-insan-yonlendirilmeleri-ogretiliyor/22289)ları denetimli öğrenme ve denetimsiz öğrenme olarak kategorize edilir.



**1-Denetimli Öğrenme**

Denetimli makine öğrenimi, belirsizlik dahilinde kanıta dayalı tahminler yapan bir model oluşturur. Denetimli bir öğrenme algoritması bilinen bir girdi verisi seti ve verilere bilinen yanıtları alır, ardından yeni verilere yanıt için makul tahminler oluşturmak üzere bir modeli eğitir. Tahmin etmeye çalıştığınız çıktı için bilinen verileriniz varsa denetimli öğrenmeyi kullanabilirsiniz. Bir örnek verecek olursak:  
Örneğin, iş yerinizden eve gitmenizin ne kadar süreceğini tahmin etmenize yardımcı olacak bir makine eğitmek istiyorsunuz. Burada, bir dizi etiketli veri oluşturarak başlıyorsunuz. Bu veriler şunları içerir.

* Hava koşulları
* Günün saati
* Bayram

Tüm bu detaylar sizin girdilerinizdir. Çıktı, o gün eve geri dönmek için geçen süredir. İçgüdüsel olarak dışarıda yağmur yağarsa, eve gitmenizin daha uzun süreceğini bilirsiniz. Ancak makinenin veri ve istatistiklere ihtiyacı var. Şimdi bu örneğin, işe gidip gelme zamanını belirlemesine yardımcı olan denetimli bir öğrenme modelini nasıl geliştirebileceğinizi görelim. Oluşturmanız gereken ilk şey bir eğitim seti. Bu eğitim seti toplam işe gidip gelme süresini ve hava durumu, zaman vb. gibi ilgili faktörleri içerecektir. Bu eğitim setine dayanarak, makineniz yağmur miktarı ile eve gitmek için alacağınız zaman arasında doğrudan bir ilişki olduğunu görebilir.

Böylece, ne kadar çok yağmur yağarsa, evinize geri dönmek için o kadar uzun süre araba kullanacağınızı tespit eder. Ayrıca işten ayrıldığınız zaman ile yolda olacağınız zaman arasındaki bağlantıyı da görebilir. 18.00' e ne kadar yakın olursanız eve dönmeniz o kadar uzun sürer. Makineniz etiketli verilerinizle bazı ilişkileri bulabilir.  Yağmurun insanların sürüş şeklini nasıl etkilediğini, günün belirli saatlerinde daha fazla insanın seyahat ettiğini görmeye başlar.

# Denetimli  Makine Öğrenmesi Algoritmaları

Denetimli öğrenme, tahmin modelleri geliştirmek için sınıflandırma ve regresyon tekniklerini kullanır.

* k-Nearest Neighbors
* Linear Regression
* Logistic Regression
* Support Vector Machines (SVMs)
* Decision Trees and Random Forests
* Neural networks

2-Denetimsiz Öğrenme

Gözetimsiz (denetimsiz) öğrenme, modeli denetlemenize gerek olmayan bir makine öğrenme tekniğidir. Bunun yerine, modelin bilgileri keşfetmek için kendi başına çalışmasına izin vermeniz gerekir. Denetimsiz öğrenme algoritmaları, denetimli öğrenmeye kıyasla daha karmaşık işleme görevleri gerçekleştirmenizi sağlar. Dentimsiz öğrenmede sistem öğretilmiyor, verilerden öğreniyor. Denetimsiz makine öğrenimi, verilerdeki bilinmeyen her türlü paterni bulur. Denetimsiz yöntemler, kategorizasyon için yararlı olabilecek özellikleri bulmanıza yardımcı olur.

Denetimsiz  Makine Öğrenmesi Türleri

## Kümeleme

Association

Aradaki fark, denetimli öğrenmede "kategoriler", "sınıflar" veya "etiketlerin" bilinmesidir. Denetimsiz öğrenmede, onlar değildir ve öğrenme süreci uygun "kategoriler" bulmaya çalışır. Her iki öğrenme türünde de, sınıflandırmada hangisinin en uygun olduğunu belirlemek için tüm parametreler göz önünde bulundurulur.

Denetimli veya denetimsiz seçmiş olup olmamanız, verilerinizin "kategorileri" nin ne olup olmadığını bilmenize bağlı olmalıdır. Biliyorsanız, denetimli öğrenmeyi kullanın. Eğer bilmiyorsanız, o zaman denetimsiz kullanın.

Çok sayıda parametreniz olduğundan ve hangilerinin alakalı olduğunu bilmediğinizden, ilgili olanları belirlemeye yardımcı olmak için [temel bileşen analizi](http://en.wikipedia.org/wiki/Principal_component_analysis) gibi bir şey kullanabilirsiniz.

3) What are the test and validation set, and why would you want to use them?

Makine öğrenmesi metotları kullanılarak büyük veri setleri uygun modeller elde edilir. Değerlendirme, hangi modelin daha iyi olduğunu bulmak ve gelecekte öğrenme modelinin ne kadar iyi çalışacağını anlamak için yardımcı olur. Sadece eğitim (training) datayı kullanmak kabul edilebilir durum değildir çünkü metot  overfitting (aşırı uyum) durumda olabilir. Bu durumdan kurtulmak için iki farklı yaklaşım önerilir: hold-out ve cross-validation. Overfitting’ten kaçınmak ve model performansını ölçmek için her iki yaklaşım bir test kümesi (metot eğitimde hiç kullanılmamış) kullanır.

Hold-out:

Elimizde büyük veri seti varsa biri eğitim, biri doğrulama ve biri de test olmak üzerine üç farklı veri seti elde edebiliriz.

Eğitim Verisi (Eğitim Kümesi – Örnekleme Kümesi)(Training Set): girdi özelliklerini karşı ve cevap (çıktı) özellik/lerini içerir.

Doğrulama veri seti (Validation dataset) eğitim aşamasında elde edilen modelin performansını değerlendirmek için kullanılan alt bir veri setidir. Ayrıca, bu veri seti hangi modelin iyi olduğunu belirlemek ve modeller için en uygun parametreleri ayarlamak için bir test platformu sağlar. Tüm modeller doğrulama veri setine gerek duymaz.

Test veri seti (Testing Dataset) (veri setinin alt kümelerinde olmayan hiç görülmemiş durumlar) modelin gelecekteki performansını değerlendirmek için kullanılır. Eğer test verisinde sonuçlar eğitim aşamasına göre kötü çıkıyorsa, overfitting ile karşı karşıya olduğumuz durumu çıkar.

Cross Validation (Çapraz Doğrulama)

Eğitim verisi alt kümelere ayrılır. Tek alt kümeyi eğitim için kullanıp diğer kalan kümeleri doğrulama işlemi için kullanılır. Bu işlem çapraz bir şekilde tüm alt kümeler için tekrarlanır. Bu işleme çapraz doğrulama denir. Bu işlem daha önceden belirlenen belli bir k sayısında yapılır. (Literatürde ten-cross validation ifadesine çok rastlarsınız.) Veri eşit boydaki k parçaya ayrılır ve k kez değerlendirilir.

Model değerlendirmeyi iki alt bölüme ayırabiliriz.

* Sınıflama Değerlendirme
* Regresyon Değerlendirme

4) What are the main preprocessing steps? Explain them in detail. Why we need to prepare our data?

Veri hazırlamak ya da veri ön analizleri, ham halde ve dağınık şekilde bulunan verileri analize hazır hale getirmek için yürütülen çalışmalardır.

Burada bahsettiğim hamlık kuşkusuz ki, eldeki verilerin bir makine öğrenmesi ya da diğer bir istatistiki çözüm modeline uygun olmaması durumudur. Elimizde ki ERP, CRM, B2B, B2C, Sosyal medya vs verileri çeşitli veri tabanlarında kendi içlerinde düzenli ve bütünleşik bir biçimde biriktirile biliyor olabilir. Fakat çözüm modellemesine girecek verilerin bu modellemeye uygun hale getirilmesi gerekmektedir.

Örneğin raporlama ve standart analiz işleri için şirketler verilerini veri ambarlarında raporlamaya ve standart analize hazır halde tutarlar. Veri bilim takımları da kullanacakları veri setleri analize hazır bir şeklide hazırlarlar. Analiz süreçleri çeşitli frekanslarla tekrar edilen süreçler olduğundan bu veri setleri sürekli güncellenir ve geliştirilir.

Bu çalışmalar hangi verinin nasıl toplanacağının karar verildiği genel veri modelleme sürecinde başlar. Veri modeli yapılacak analizlere göre şekillendirilir. Şirket hangi veriyi nereden nasıl toplayacak, nereye kayıt edecek gibi kararlar veri bilim süreçleri dahilinde değişebilir, geliştirilebilirdir.

Dolayısı ile Veri Bilim takımları şirketlerin genel veri modellerinde de söz sahibi olmalıdırlar.

Eldeki veri kaynakları veri bilim ile uyumlu hale getirildikten sonraki adım kuşkusuz ki şirket için oluşturulacak veri bilim süreçlerinin üzerine bina edileceği problemler ve çözümlerini betimlemektir. Bu problemlerin çözümü için kullanılacak modeller ortaya konur ve bu modellerin en iyi çalışması için veri temizlenir, düzenlenir.

5) How you can explore and analyse countionus and discrete variables?

Ayrık ve Sürekli Değişkenler

İstatistikte değişken, bir varlık, bir kişi, yer veya bir şey olarak ve değişkenin aldığı değer bir varlıktan diğerine değişebilir. Örneğin, Y değişkeninin bir sınavdaki bir öğrencinin notu olmasına izin verirsek Y, A, B, C, S ve F değerlerini alabilir. X değişkeninin bir sınıftaki bir öğrencinin yüksekliği olmasına izin verirsek, o zaman bir aralıkta herhangi bir gerçek değer alabilir.

Bu iki örnekten, değişkenin alanının normal aritmetik işlemlerle sayısal olup olmadığına bağlı olarak nicel ve nitel olarak iki değişken türü olduğu görülür. Bu nicel değişkenler iki türdendir: ayrık değişkenler ve sürekli değişkenler.

Ayrık bir değişken nedir?

Nicel değişken yalnızca sayılabilir en fazla sayıda değeri alabilirse, bu tür verilere ayrık veriler denir. Başka bir deyişle, değişkenin alanı en fazla sayılabilir olmalıdır. En fazla sayılabilir sayı sonlu veya sayılabilirdir. Bunun bir örneği bunu daha iyi açıklayacaktır.

Bir sınıfa beş soru testi verilir. Bir öğrencinin aldığı doğru cevapların sayısı X olsun. X'in olası değerleri 0, 1, 2, 3, 4 ve 5'tir; Sadece 6 olasılık ve sonlu bir sayıdır. Bu nedenle, X ayrı bir değişkendir.

Bir oyunda bir insan bir hedef vurmalı. Eğer hedefi vurana kadar bir kere atış yaparsak, Y'nin olası değerleri 1, 2, 3, 4 olacaktır … vb. Teorik olarak bu değerlerin sonlu bir sınırı yoktur. Ancak bu değerler sayılabilir. Dolayısıyla, Y değişkeni, "hedefe ulaşıncaya kadar bir atış kaç kez" olarak tanımlanır; ayrı bir değişkentir.

Bu iki örnekten, ayrı değişkenlerin çoğunlukla sayılar olarak tanımlandığı görülebilir.

Sürekli değişken nedir?

Bir aralıktaki olası tüm değerleri alabilen nicel değişken, sürekli veri olarak adlandırılır. Bu nedenle, sürekli değişken alanının aralığı (0, 5) ise, değişken 0 ile 5 arasında herhangi bir gerçek sayı değeri alabilir.

6) Analyse the plot given below. (What is the plot and variable type, check the distribution and make comment about how you can preproccess it.)

Negative Bionimal bir grafiktir.