Makine Öğrenmesine Giriş

Kursa Başlamadan Önce

- Bu kurs tamamlandığı takdirde giriş düzeyi yapay zeka algoritmaları ve veri analizine temel oluşturabilecek genel bilgileri edinmiş olacaksınız
- Spesifik konular anlaşılması zor ve kişide ders esnasında mantığın oturması kolay olmayacağından bol bol bireysel pratik gerekmektedir
- Slaytlar yazılara boğulmadan görsellerle anlatılacaktır. Bu yüzden ders esnasında not tutulması **son derece** önemlidir
- Konu başlıkları temel düzey algoritmalar için yeterli olduğundan başlıklar araştırılmalı, bol bol uygulama ve teorik bilgiler içeren sitelerde araştırma yapılmalıdır

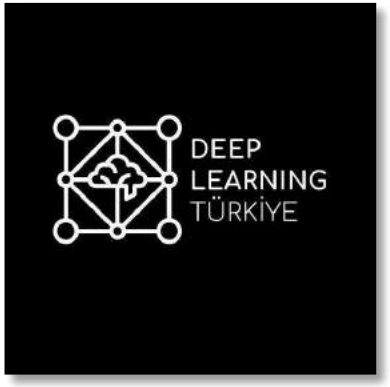
Bu Eğitimde Neler Öğreneceksiniz?

1. Makine Öğrenmesi konusuna hızlı bir giriş yapacak ve farklı alanlar ile ayrım anlamını inceleyeceğiz.
2. Kurs boyunca detaylı şekilde inceleyeceğimiz Makine Öğrenmesi tiplerini öğrenecek ve farklarını tartışarak tanışacağız.
3. Matematik ve İstatistik konusunun Makine Öğrenmesi ile ilişkisine değinecek ve ardından Veri Bilimi alanına giriş yapacağız.
4. Bir Makine Öğrenmesi modeli çalıştırma sırasında veriyi toplama, veriyi görselleştirme, veriyi modele uygun hale getirme, model seçimi yapma gibi model eğitiminin adımlarını görecek ve Makine Öğrenmesi terminolojileri ile tanışacağız.
5. Makine Öğrenmesi modeli geliştirirken kullanabileceğimiz çeşitli araçları görecek ve örnek veri setleri ile tanışacağız.

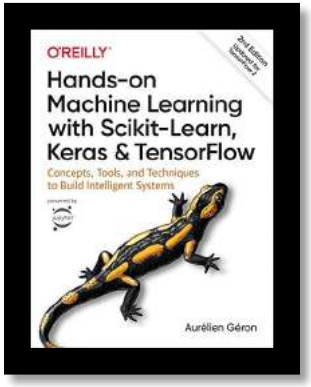
İrem Kömürcü

Github: irem-komurcu
Linkedin: iremkomurcu
Twitter: iremkomurcu

Kaynaklar



Deep Learning Türkiye



Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
Aurelien Geron

Stanford
University

Stanford University
stanford.edu/~shervine/
Shervine Amidi

Makine Öğrenmesi Nedir?

Makine Öğrenimi, bilgisayarları verilerden öğrenebilmeleri için programlama bilimidir (ve sanatıdır).

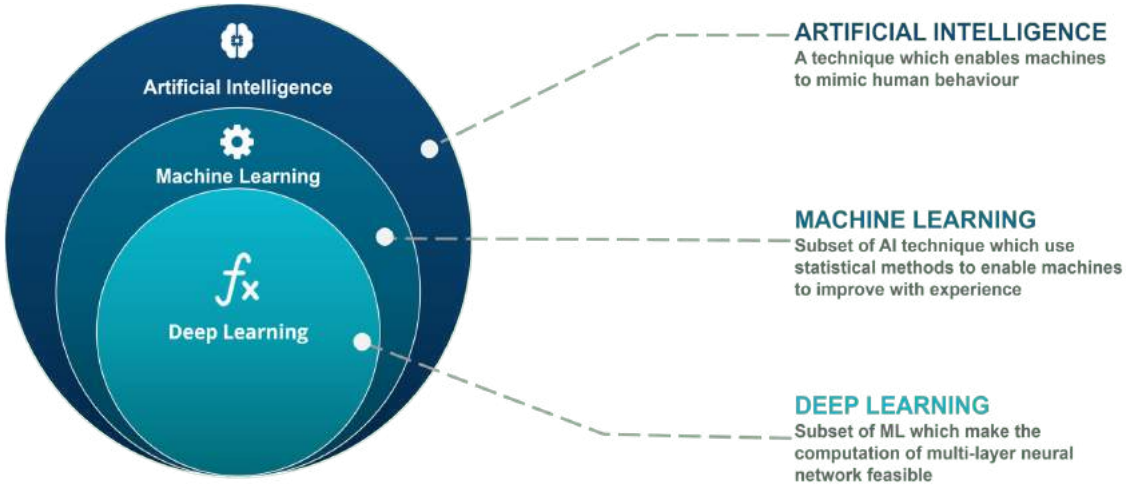
Makine Öğrenimi bilgisayarlara açıkça programlanmadan öğrenme yeteneği veren çalışma alanıdır.

—Arthur Samuel, 1959

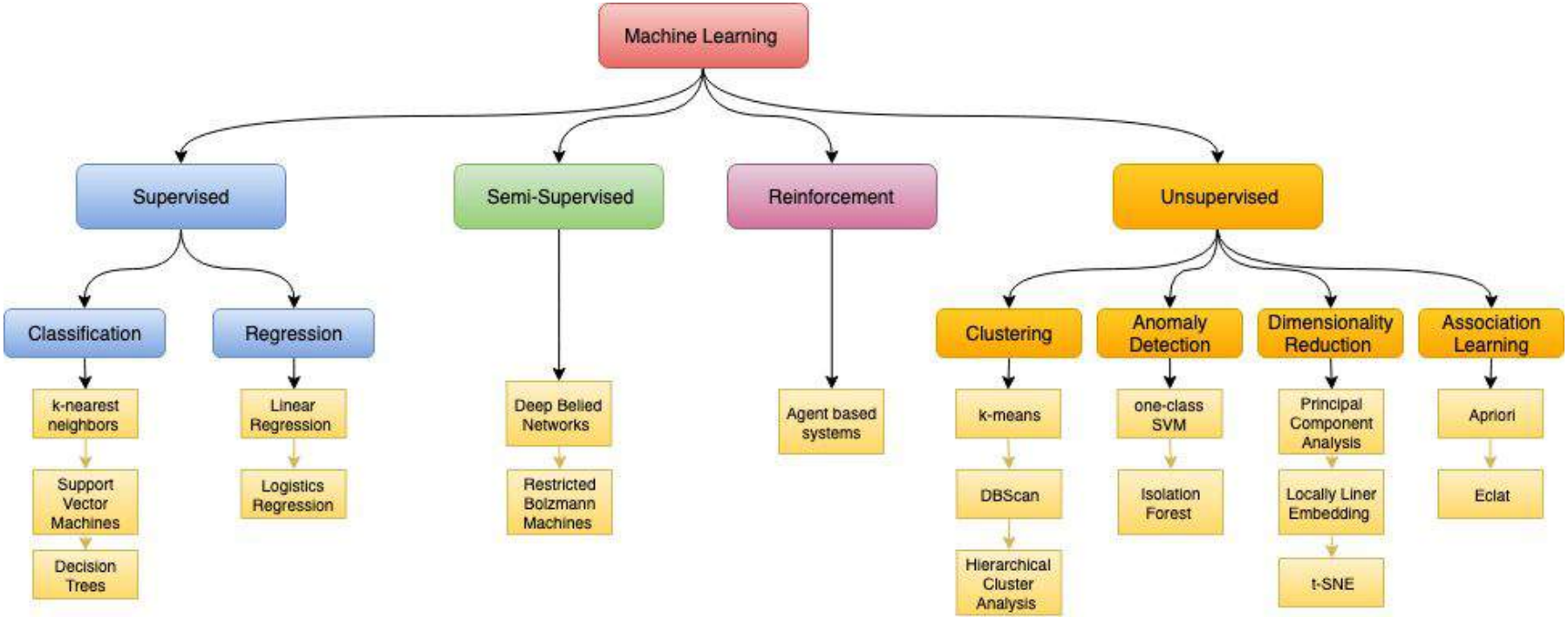
Makine Öğrenmesi

Makine Öğrenmesi, insanların öğrenme şeklini taklit etmek için veri ve algoritmaların kullanımına odaklanan ve doğruluğunu kümülatif olarak arttıran bir Yapay Zeka ve Bilgisayar Bilimi dalıdır.

Açıkça programlanmadan, deneyimlerden otomatik olarak öğrenir ve iyileştirme yeteneğine sahip sistemler oluşturmak için istatistiksel öğrenme algoritmalarını kullanır.



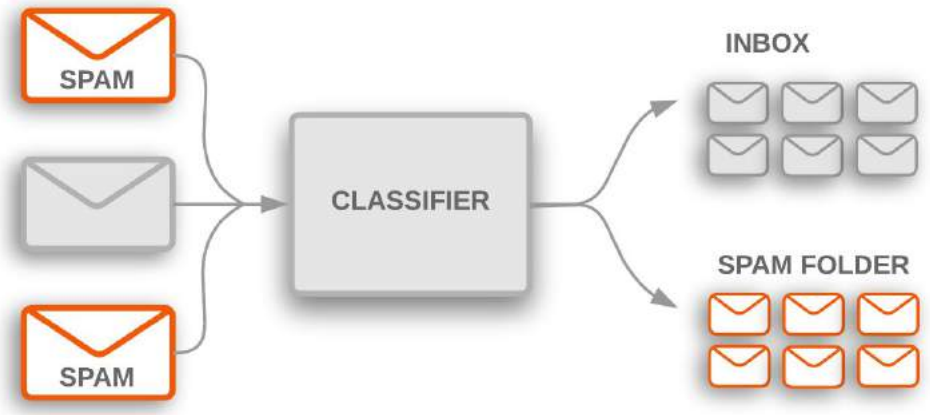
Makine Öğrenmesi Tipleri



Denetimli Öğrenme (Supervised Learning)

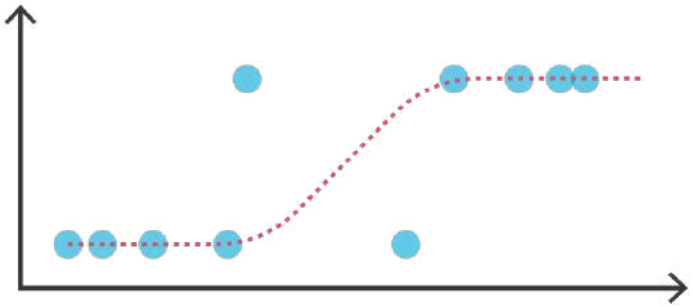
Denetimli Öğrenme; etiketlenmiş veri kümelerini bağımlı ve bağımsız değişkenler kullanarak işleme sokan makine öğrenmesi algoritmalarını içerir.

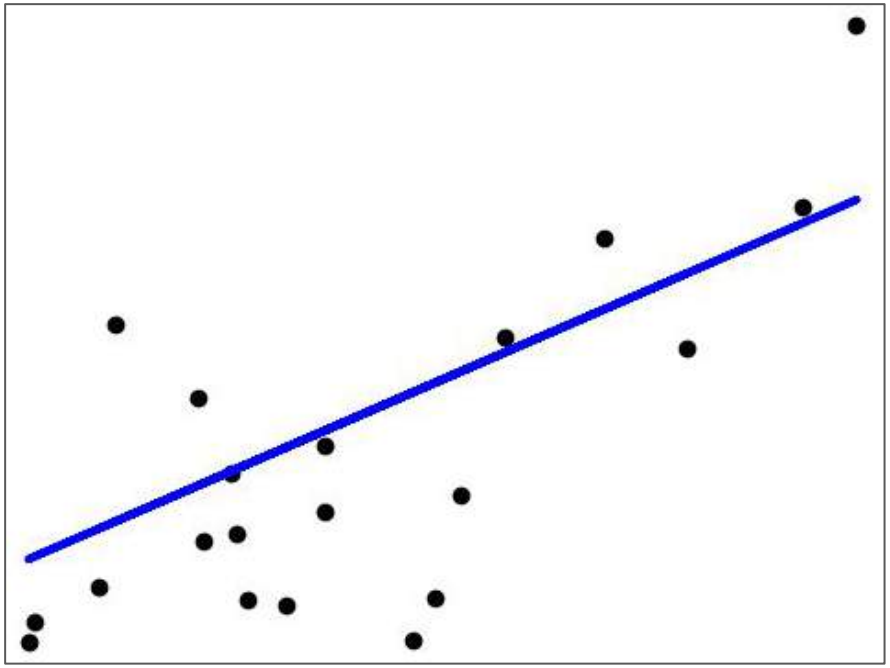
İstenmeyen e-posta filtresi buna iyi bir örnektir: Denetimli Öğrenme modeli, sınıflarıyla birlikte birçok örnek e-posta verisi ile eğitilmiştir ve yeni e-postaları nasıl sınıflandıracağını öğrenmelidir.



Denetimli Öğrenme Algoritma ve Mimarileri

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





Problemlere Gerçek Hayat Örnekleri

Regresyon;

- Ev fiyatı tahmini
- Kişinin yaş tahmini
- Bir çiçeğin çapının tahmini
- Muhtarlığa aday bir kişinin alacağı oy tahmini
- Arabanın motor hacminin L cinsinden tahmini

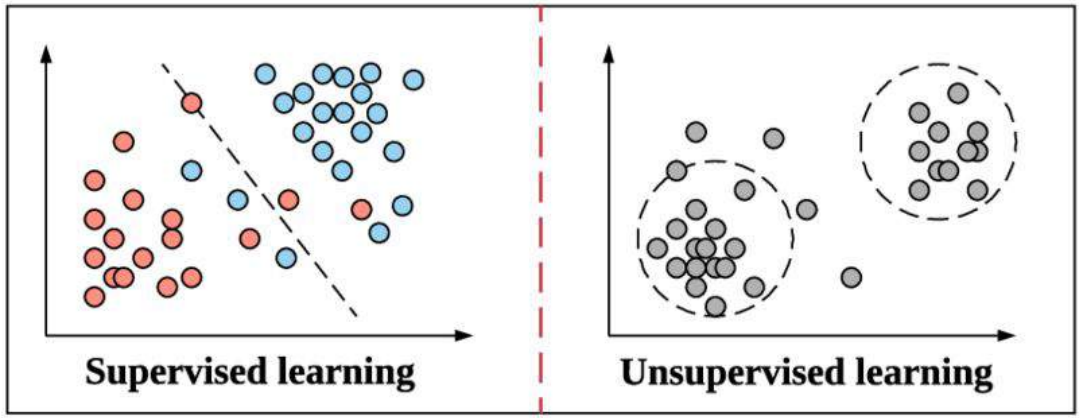
Sınıflandırma;

- Bir tümörün iyi huylu olup olmadığı
- Öğrencinin sınıfı geçip geçmeyeceği
- Bir e-mailin spam olup olmadığı
- Bir videodaki nesnenin [Köpek, kedi, kuş, at, otomobil, insan, araba] seçeneklerinden hangisinin olduğunun bulunması

Denetimsiz Öğrenme (Unsupervised Learning)

Unsupervised Learning, etiketlenmemiş veri kümelerini analiz etmek ve kümelemek için makine öğrenimi algoritmalarını kullanır.

Bu algoritmalar, insan müdahalesine ihtiyaç duymadan gizli kalıpları veya veri gruplamalarını keşfeder.



Neden Unsupervised Learning Kullanıyoruz?

Unsupervised Learning için en yaygın görevler;

- Kümeleme
- Yoğunluk tahmini
- Temsili öğrenme

Unsupervised Learning Algoritmaları;

- Küme sayımızın bilinmediği
- Etiketli eğitim verimizin bulunmadığı

durumlarda kullanılabilir.

Örnekleri

Güvenlik:

Veri kümelerindeki olağandışı veri noktalarını tanımlandığı kümeleme anormalliği algılaması

Pazarlama:

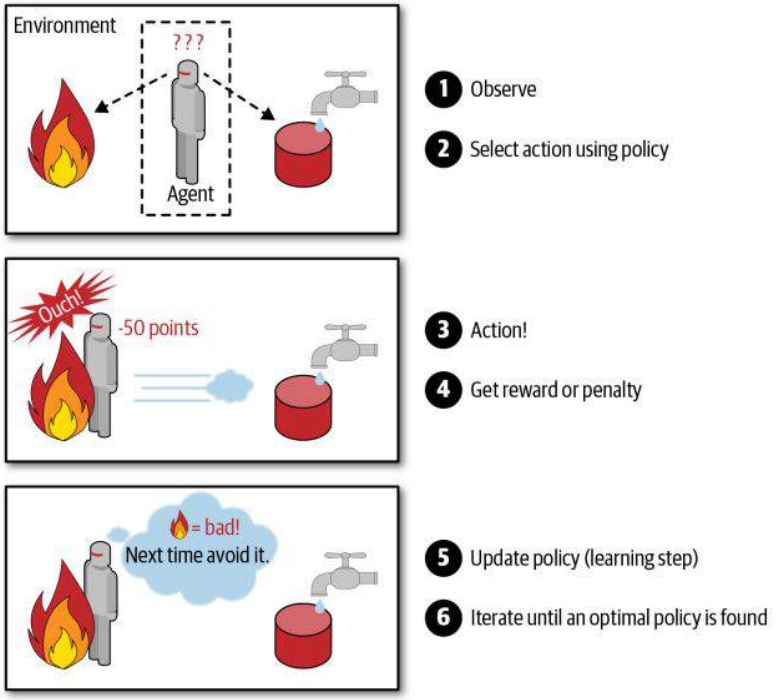
Veri noktaları arasındaki ilişkileri bulduğu ilişki madenciliği

Reinforcement Learning

Reinforcement Learning (Pekiştirmeli Öğrenme); amaca yönelik olarak, ajanların açık hedefler belirleyerek öğrenme gerçekleştirdiği Makine Öğrenmesi çeşididir.

Ajan, verilen modelimizi environment (çevre) ile etkileşiminden dönen geri bildirimleri kullanarak deneme yanılma yoluyla etkileşimli bir ortamda öğrenir.

Reinforcement Learning çevreyi (environment) gözlemleyebilir, eylemleri seçip gerçekleştirebilir ve karşılığında ödül (veya olumsuz ödül şeklinde cezalar) alabilir.



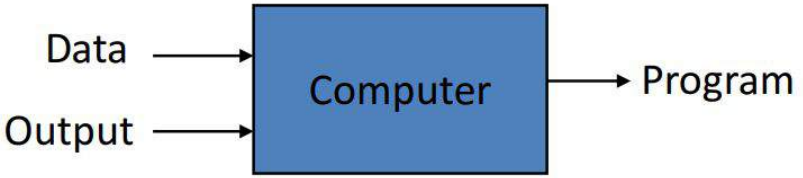
Geleneksel Programlama vs Makine Öğrenmesi

Geleneksel Programlama:

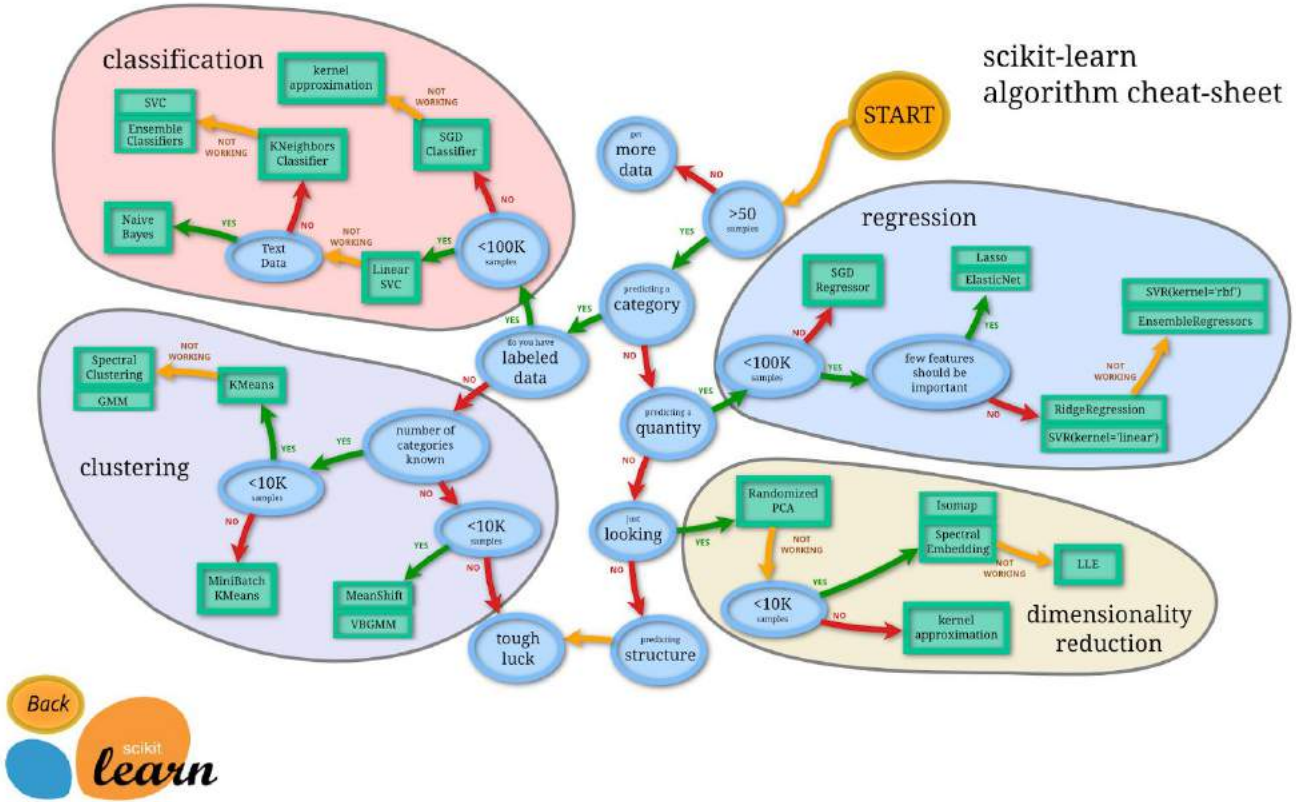
Bilgisayarınıza input verinizi ve kurallarınızı verirsiniz. Input değerleri bu kurallardan geçip işlenerek bir output değeri üretir.

Makine Öğrenmesi:

Bilgisayarınıza input ve output verinizi verirsiniz. Bilgisayar sizin için bu iki veri arasındaki kuralı belirler ve sonraki kullanımlarda input verileriniz bu kurallardan geçer.



Makine Öğrenmesi Algoritmaları

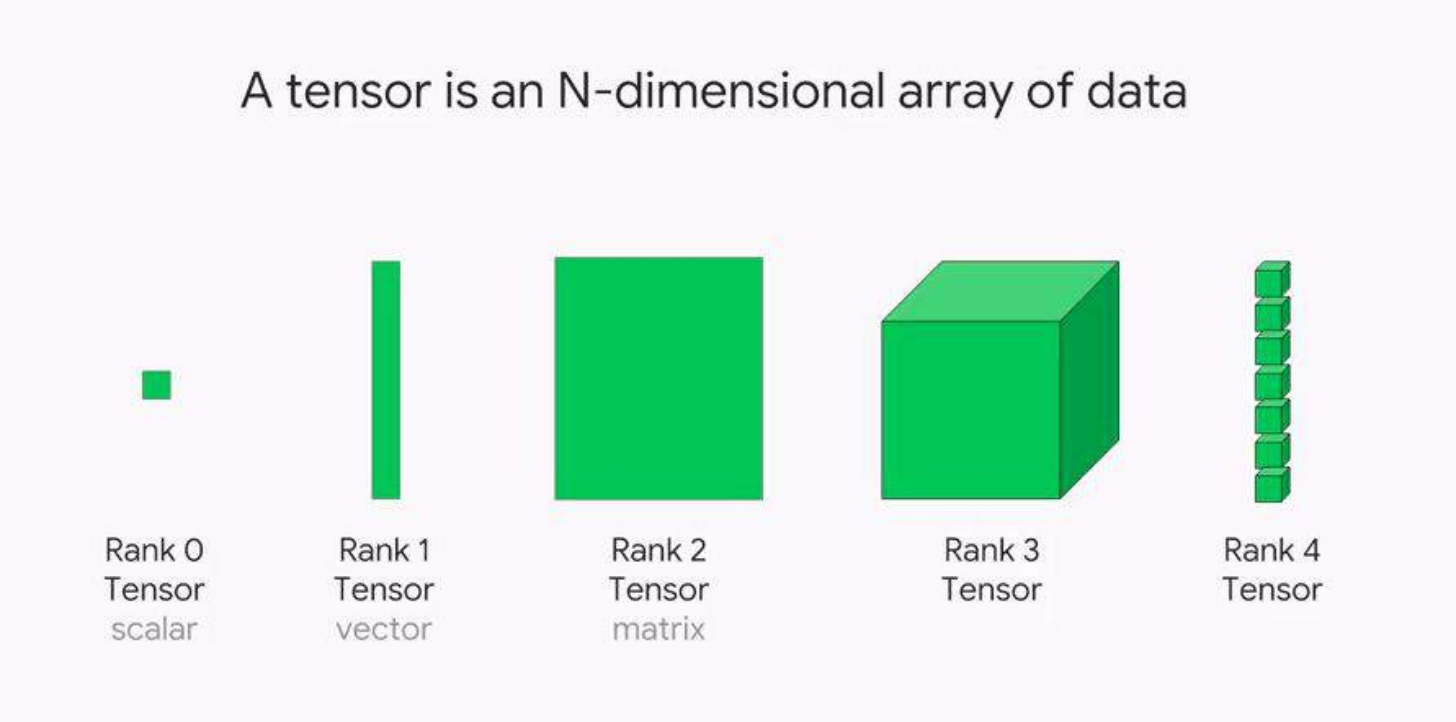


Makine Öğrenmesi Uygulamaları

- Her segmente farklı bir pazarlama stratejisi tasarlayabilmek için müşterileri satın alımlarına göre segmentlere ayırma
- Bir müşterinin ilgilenebileceği bir ürünü, geçmiş satın alımlara dayanarak önerilmesi
- Bir üretim hattındaki ürünlerin görüntülerini otomatik olarak sınıflandırmak için analiz edilmesi
- Haber makalelerini otomatik olarak sınıflandırma
- Tartışma forumlarında rahatsız edici yorumları otomatik olarak işaretleme
- Chatbot veya kişisel asistan oluşturma
- Birçok performans ölçümüne dayalı olarak şirketin gelecek yılki gelirini tahmin etme
- Uygulamanızın sesli komutlara tepki vermesini sağlama
- Bir oyun için akıllı bir bot oluşturma

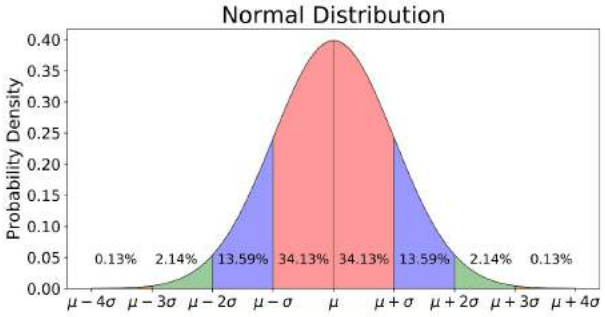
Makine Öğrenmesinde Matematik

Lineer Cebir ve Makine Öğrenmesi

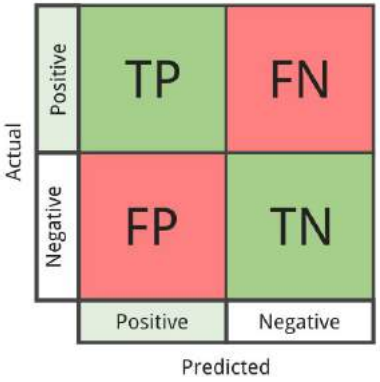


Olasılık ve İstatistik

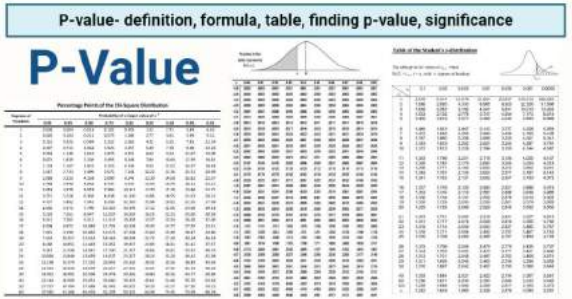
Dağılımlar



Model Performans Ölçümleri



Betimsel İstatistik ve İstatistik Testleri

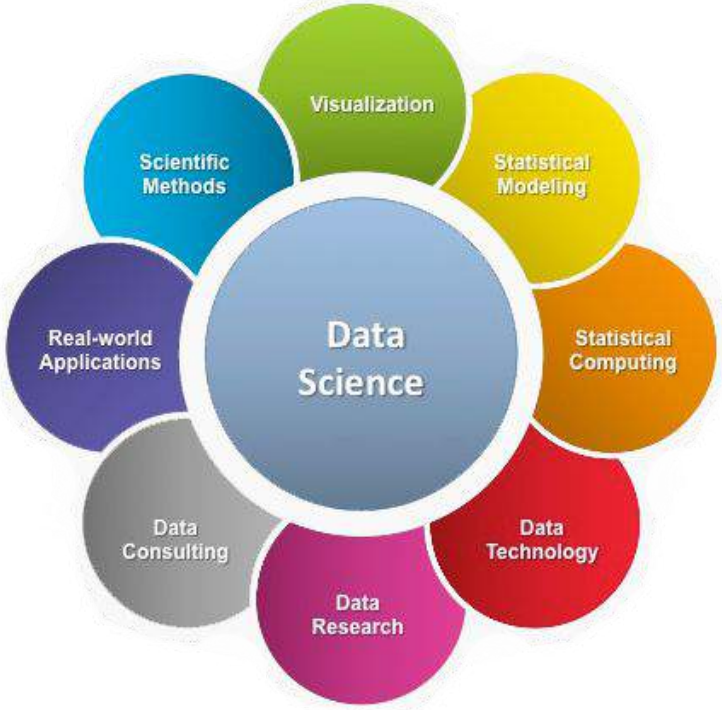


Veri Bilimi

Veri ile Çalışmak

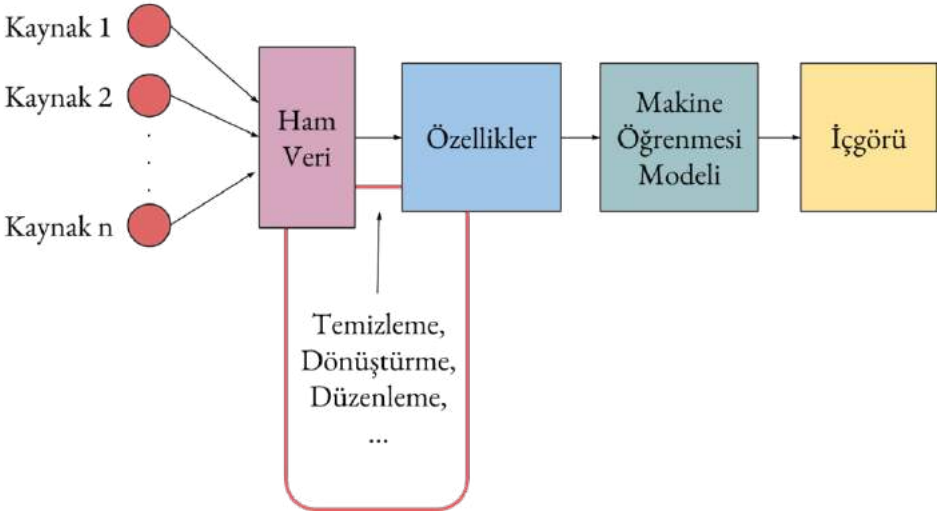
Veri bilimi, yapılandırılmış ve yapılandırılmamış verilerden bilgi ve öngörü elde etmek için bilimsel yöntemleri, süreçleri, algoritmaları ve sistemleri kullanan çok disiplinli bir alandır.

Veri bilimi, gerçek olayları verilerle anlamak ve analiz etmek için istatistikleri, veri analizini, makine öğrenimini ve ilgili yöntemlerini birleştirmek için kullanılan bir kavramdır. Matematik, istatistik, bilgisayar bilimi gibi alanlardan birçok teknik ve teori kullanır.



Feature Engineering

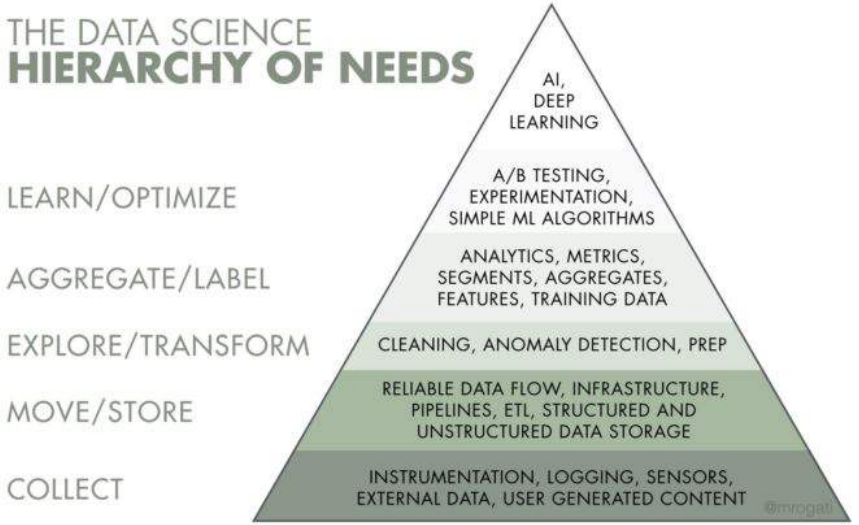
Feature Engineering (Özellik mühendisliği); ham verilerden özellikleri çıkarmak için alan bilgisini kullanma sürecidir. Özellikler, tahmine dayalı modeller tarafından kullanılır ve sonuçları etkiler.



Model Eğitimi

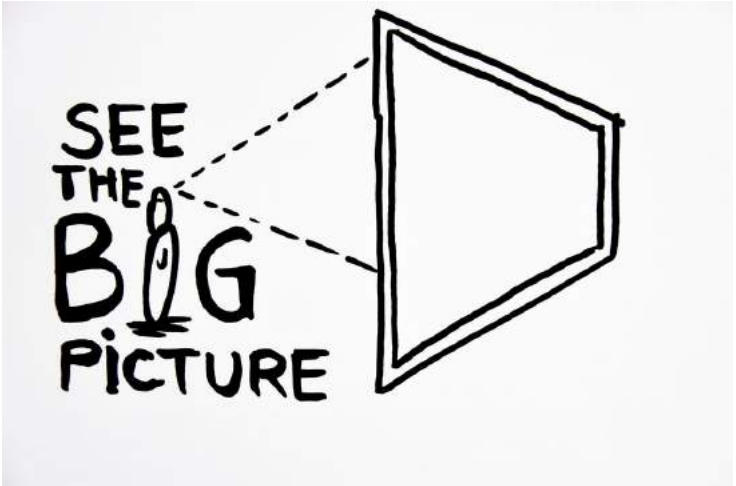
Makine Öğrenmesi Projesi Adımları

1. Büyük resme bakış ve projeyi anlamak
2. Veri toplamak
3. Veriyi incelemek ve görselleştirmek
4. Veriyi Makine Öğrenmesi Modellerine Uygun Hale getirmek
5. Model seçimi ve modelin eğitimi
6. Modelin optimize edilmesi
7. Modelin canlıya alınması



Büyük resme bakış ve projeyi anlamak

1. Hedefi iş açısından tanımlayın
2. Mevcut çözümler/geçici çözümler nelerdir (varsa)?
3. Bu sorunu nasıl çerçevelemelisiniz (denetimli/denetimsiz, çevrimiçi/çevrimdışı vb.)?
4. Performans nasıl ölçülmelidir?
5. İş hedefine ulaşmak için gereken minimum performans ne olurdu?
6. İnsan uzmanlığı mevcut mu?
7. Problemi manuel olarak nasıl çözersiniz?
8. Sizin (veya başkalarının) şimdiye kadar yaptığınız varsayımları listeleyin



Veri toplamak

Çalışma Ortamı Hazırlayın

Veriler makine öğrenmesi projelerinde düzenli ve uygun bir şekilde saklanmalıdır. Özellikle ham verinin zarar görmemesi, yapılarının bozulmaması gerekmektedir. Modelde kullanılacak ve önişleme tabi tutulan verilerin depolanması, modele gönderilen verilerinin saklanması gerekmektedir. Bu yüzden uygun veri tabanları kurulmalı veya klasör hiyerarşileri sağlanmalıdır.

Web Scraping

Get the Data

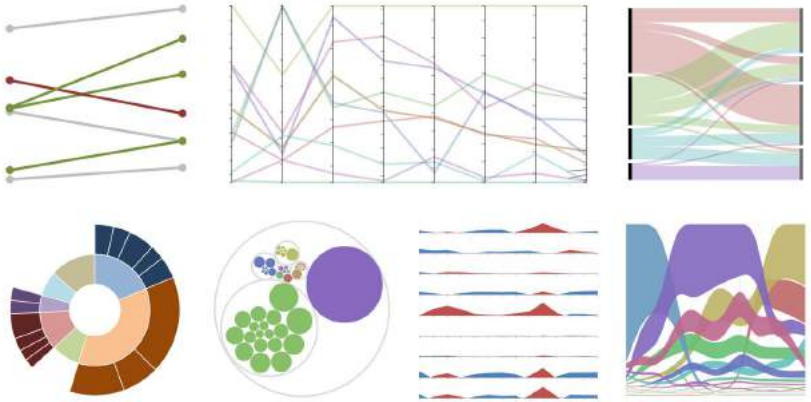
Veriyi elde edebilecek birçok kaynak bulunmaktadır, öğrenme aşamalarında internetten elde edilen veri setleri kullanılabilir, gerçek dünya veri setleriyle çalışmak biraz daha zorludur, bazen bu verilerin elde etmek kolay olmamakla birlikte, yazılımlar ile birlikte bu veriler internet üzerinden toplanabilir.

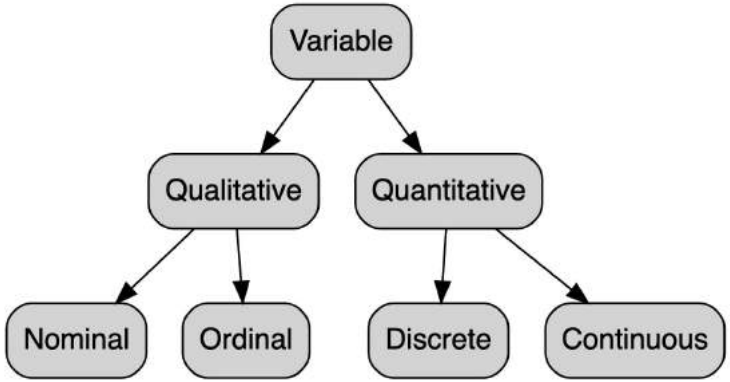
Structured Data

Veriyi incelemek ve görselleştirmek

Exploratory Data Analysis (EDA)

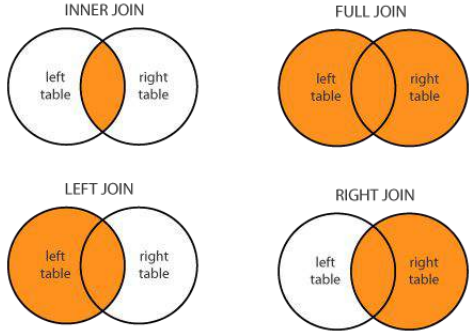
Özet istatistiklerle (ortalama, medyan, nicelikler, vb.) birlikte bir veri kümesinin resmini çizmeye yardımcı olan basit grafikler (örn. kutu grafikleri, dağılım grafikleri) oluşturmaya verilen, veriyi hızlı bir şekilde tanımamıza olanak sağlayan çalışmalara denir.

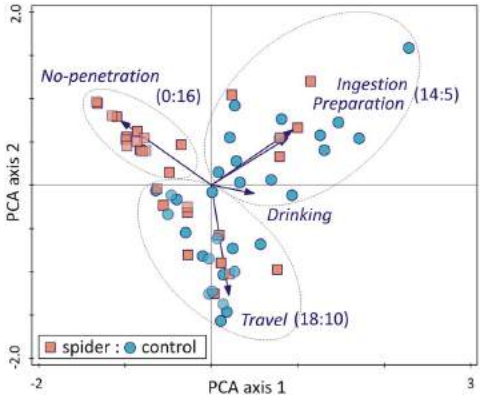




Veriyi Modele Uygun Hale getirmek

1. **Verilerin Temizlenmesi ve Düzenlenmesi**
   - Gereksiz veya Bilgi İçermeyen Verilerin Temizlenmesi
   - Hatalı verilerin düzeltilmesi
   - Farklı kaynaklardan gelen verilerin birleştirilmesi
2. **Veri Tiplerinin Belirlenmesi**
   - Tarih, Sayısal, Yazı vb. formatlardaki verilerin kontrol edilmesi
   - Uygun veri tipi dönüşümlerinin yapılması
3. **Veri Boyutunun İndirgenmesi**
   - PCA
   - Gereksiz kolonların atılması ve korelasyon analizi
4. **Veri dağılımlarının incelenmesi ve regularizasyon**
   - Min-max Ölçekleme
   - Standardizasyon

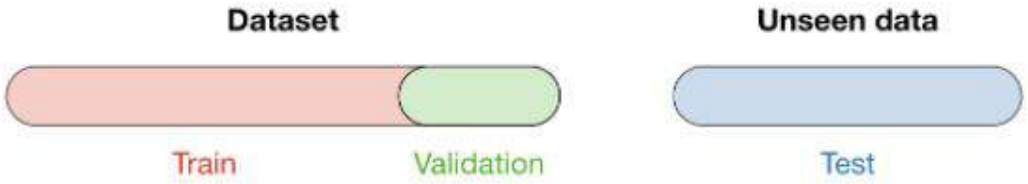




Model seçimi ve modelin eğitimi

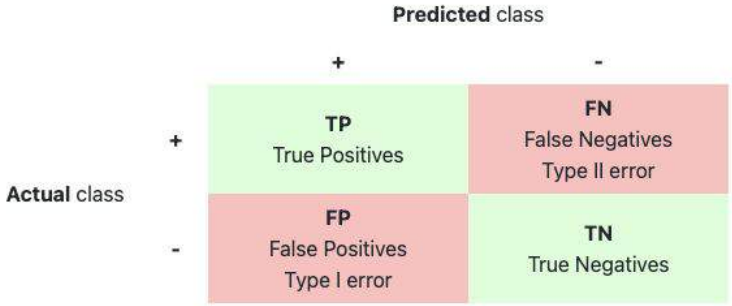
Training set	Validation set	Testing set
• Model is trained • Usually 80% of the dataset	• Model is assessed • Usually 20% of the dataset • Also called hold-out or development set	• Model gives predictions • Unseen data

Once the model has been chosen, it is trained on the entire dataset and tested on the unseen test set. These are represented in the figure below:



Başarı Performansları: Sınıflandırma Metrikleri

Karmaşıklık Matrisi



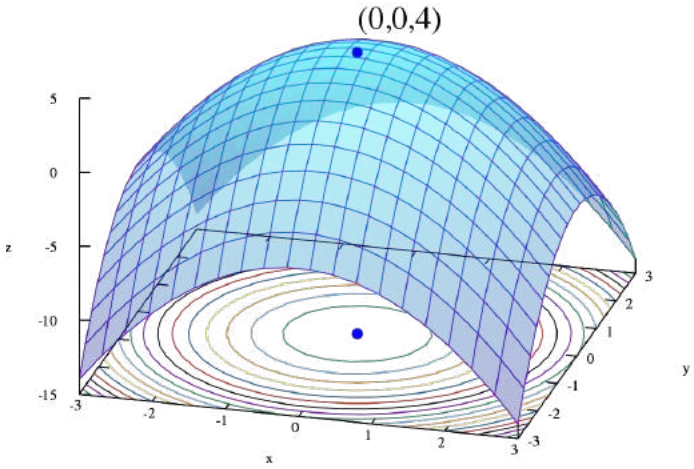
Metric	Formula	Interpretation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Overall performance of model
Precision	$\frac{TP}{TP + FP}$	How accurate the positive predictions are
Recall Sensitivity	$\frac{TP}{TP + FN}$	Coverage of actual positive sample
Specificity	$\frac{TN}{TN + FP}$	Coverage of actual negative sample
F1 score	$\frac{2TP}{2TP + FP + FN}$	Hybrid metric useful for unbalanced classes

Başarı Performansları: Regresyon Metrikleri

Mean squared error	$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$
Mean absolute percentage error	$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n} \left\lvert \frac{e_t}{y_t} \right\rvert$

Modelin optimize edilmesi

- Cross validation kullanarak hiperparametrelerde ince ayar yapın
- Hiperparametre aramaları yapın, gridsearch vs.
- Ensemble yöntemlerini deneyin. En iyi modellerinizi birleştirmek, genellikle tek tek çalıştırmaktan daha iyi performans sağlar
- Özellikle ince ayarın sonuna doğru ilerlerken, bu adım için mümkün olduğu kadar çok veri kullanmak isteyeceksiniz
- Nihai modelinizden emin olduğunuzda, genelleme hatasını tahmin etmek için test setindeki performansını ölçün



Modelin canlıya alınması (Deployment)

ML Deployment, verilere dayalı makine öğrenimi modelini mevcut bir üretim ortamına entegre edilmesidir.

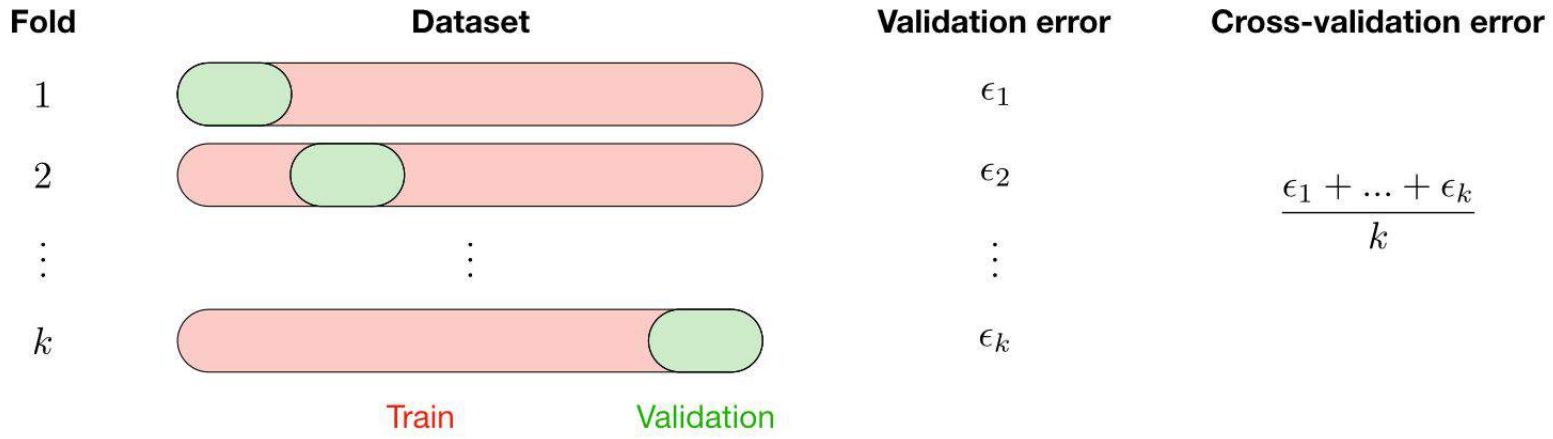
Test ortamında geliştirilen makine öğrenimi projesi son kullanıcıya hizmet vermek adına web servisleri gibi (SaaS, PaaS) platformlarında çalışmak için hazır hale getirilir.

- **Streamlit**, HTML veya CSS vb. bilgisi olmadan etkileşimli web uygulamaları oluşturmanıza yardımcı olan açık kaynakli br python kütüphanesidir
- **Docker**, geliştiricilerin kodlarını konteynerize hale getirmenin açık ara en popüler yollarından biridir
- **Heroku**, bir bulut bilişim uygulama altyapısı servis sağlayıcısıdır (PaaS)

Makine Öğrenmesi Terminolojileri

Cross Validation

CV olarak da adlandırılan çapraz doğrulama, ilk eğitim setine çok fazla dayanmayan bir model seçmek için kullanılan bir yöntemdir.



Bias - Variance Tradeoff

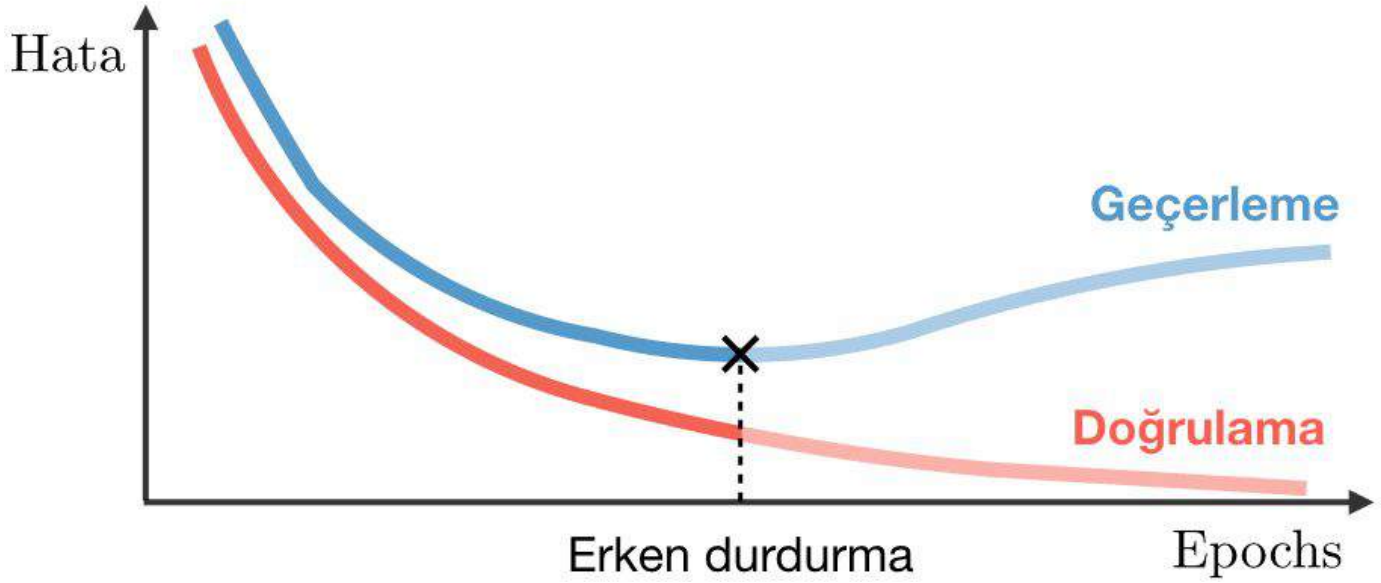
Bias: Bir modelin yanlılığı, beklenen tahmin ile verilen veri noktaları için tahmin etmeye çalıştığımız doğru model arasındaki farktır.

Varyans: Bir modelin varyansı, verilen veri noktaları için model tahmininin değişkenliğidir.

Sapma/varyans Tradeoff: Model ne kadar basitse yanlılık o kadar yüksek ve model ne kadar karmaşıksa varyans o kadar yüksek olur.

	Underfitting	Ideal	Overfitting
Semptomlar	• Yüksek eğitim hatası • Test hatasına yakın eğitim hatası • Yüksek bias	• Eğitim hatası, test hatasından biraz daha düşük	• Çok düşük eğitim hatası • Eğitim hatası, test hatasından oldukça düşüktür • Yüksek varyans

Early Stopping



Makine Öğrenmesinde Kullanılan Araçlar

Makine Öğrenmesinde Kullanılan Araçlar

Python, genel amaçlı bir programlama dilidir.

Yorumlanan ve dinamik bir dil olan Python, esas olarak nesne tabanlı programlama yaklaşımlarını ve fonksiyonel programlamayı desteklemektedir.

- Hızlı prototipleme
- Basit syntax
- Kolay kullanım
- Geniş topluluk

Linear Regression

```
from sklearn.tree import DecisionTreeClassifier

DTC = DecisionTreeClassifier(criterion='gini',
max_features=10, max_depth=5)

DTC = DTC.fit(X_train, y_train)

y_predict = DTC.predict(X_test)
```

Makine Öğrenmesinde Kullanılan Araçlar

NumPy, Python'da bilimsel hesaplamalarda kullanılan temel pakettir.

- Dizi oluşturma
- Vektörleştirme ve dilimleme
- Matrisler ve basit lineer cebir
- Veri dosyaları

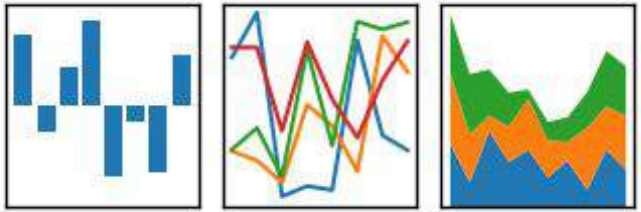
Makine Öğrenmesinde Kullanılan Araçlar

Pandas, veri analizi ve veri ön işlemeyi kolaylaştıran açık kaynak kodlu bir Python kütüphanesidir.

- Veri manipülasyonu için kullanışlı fonksiyonlar
- Farklı biçimler arasında veri okuma ve yazma araçları: CSV ve metin dosyaları, Microsoft Excel, SQL veritabanları
- Basit seviyede hızlı veri görselleştirme

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Makine Öğrenmesinde Kullanılan Araçlar

Matplotlib, Python programlama dili için bir veri görselleştirme ve çizim kütüphanesidir.

- Matplotlib grafik çizim paketi Python'la bilimsel programlamanın en önemli araçlarından birisidir
- Çok kuvvetli bir paket olan Matplotlib ile verileri etkileşimli olarak görselleştirebilir
- Basıma ve yayınlanmaya uygun yüksek kalitede çıktılar hazırlayabiliriz
- Hem iki boyutlu hem de üç boyutlu grafikler üretilebilir

Makine Öğrenmesinde Kullanılan Araçlar

Scikit-learn, Python programlama dili için ücretsiz bir yazılım makinesi öğrenme kütüphanesidir.

Doğrusal regresyon, lojistik regresyon, karar ağaçları, rastgele orman gibi birçok temel yöntemi bünyesinde bulundurur.

https://scikit-learn.org/stable/

Machine Learning with Scikit-Learn

Linear Regression

```
from sklearn.tree import DecisionTreeClassifier

DTC = DecisionTreeClassifier(criterion='gini',
max_features=10, max_depth=5)

DTC = DTC.fit(X_train, y_train)

y_predict = DTC.predict(X_test)
```

Veri Setleri

Kaggle

Kaggle, veri bilimcileri ve makine öğrenimi uygulayıcıları için çevrimiçi bir topluluktur.

Büyük veya küçük problem sahiplerinin ilgili problemi çözme amacıyla verilerini ve problemlerini dile getirdiği, katılımcıların ise verilen bilgiler dahilinde sorunu çözmek için yarışmalara katıldığı bir platformdur.

- Yüzlerce veri seti
- Ödüllü yarışmalar
- Eğitim ve rehberler

UCI

UCI, University of California, Irvine bünyesindeki machine learning and intelligent systems araştırma merkezi tarafından sunulan veri seti deposudur.

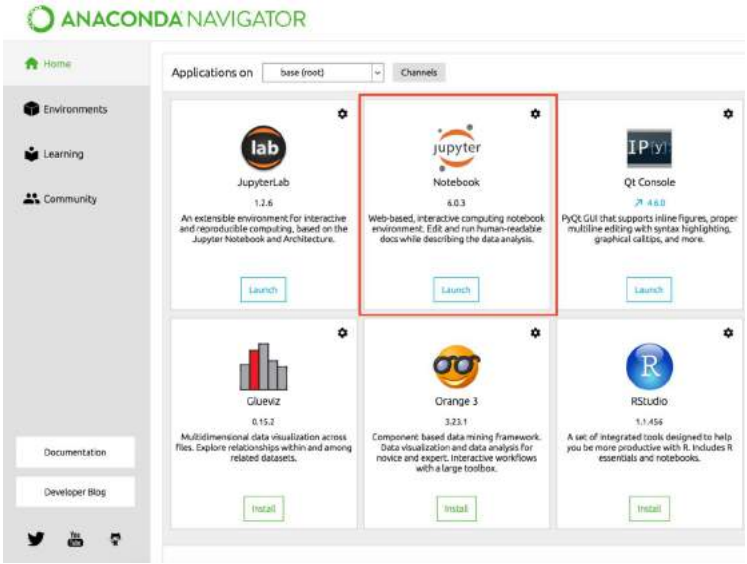
Şu anda makine öğrenimi topluluğuna hizmet olarak **588 veri setine** ev sahipliği yapmaktadır.

https://archive.ics.uci.edu/ml/index.php

Teknik Çalışma Ortamı

Working Environment

Derin öğrenme modelleri ile çalışırken kullanılabilecek bazı ortam ve araçlar;



Local Environment:

- Anaconda
- Spyder
- Jupyter Notebook

Cloud Environment:

- Google Colab
- AWS
- Azure

Component Needed:

- GPU, Cuda, cuDNN

Container Engine

- Docker

Google Colab

Colaboratory (ya da kısaca "Colab"), tarayıcınızda Python'u yazmanızı ve çalıştırmanızı sağlar.

- Hiç yapılandırma gerektirmez
- GPU'lara ücretsiz erişim imkanı sunar
- Kolay paylaşım imkanı sunar