ISE 302 –Veri Madenciliği

DR. ÖĞR. ÜYESİ ESİN AYŞE ZAİMOĞLU

esinzaimoglu@sakarya.edu.tr

Model Oluşturma

Veri Önişleme adımından sonra model oluşturma işlemi yapılır.

Modelin amacı :

- Sınıflandırma ?
- Tahmin ?

Model oluşturma aşamasında ilk yapılması gereken verinin uygun şekilde bölümlenmesidir.

Genellikle veriler eğitim, doğrulama ve test veri seti olarak 3'e ayrılır.

Bu işlem, modelin geliştirilmesi, doğrulanması ve sonuçlarının değerlendirilmesi için kullanılır.

Veri Madenciliği Teknikleri

- Veri Madenciliği yöntemlerini denetimli ve denetimsiz olmak üzere iki ana kategoriye ayırmak mümkündür.
- Veri Madenciliğinde iyi tanımlanmış veya kesin bir hedef olduğunda denetimli-gözetimli (supervised) ifadesi kullanılır.
- Elde edilmesi istenen sonuç için özel bir tanımlama yapılmamışsa veya belirsizlik söz konusu ise denetimsiz-gözetimsiz (unsupervised) ifadesi kullanılır
- Denetimli ve denetimsiz ifadeleri birbirinin tersine karşılık gelmektedir.

Denetimsiz Öğrenme

Veri setinde veri etiketlerinin (bağımlı değişken, yanıt değişken, hedef değişkenin)bulunmadığı durumlarda kısacası algoritmanın verileri öğreneceği bir eğitmeninin olmadığı ve algoritmanın farklı yöntemler kullanarak etiketleme yaptığı durumlara Denetimsiz Öğrenme (unsupervised learning) adı verilir.

Denetimsiz yöntemler daha çok veriyi anlamaya, tanımaya, keşfetmeye yönelik olarak kullanılan ve sonraki uygulanacak yöntemler için fikir vermeyi amaçlamaktadır.

Sınıf sayısı belirsizdir.

Denetimsiz Öğrenme Algoritmaları

Denetimsiz Öğrenme Algoritmaları

K-Means Kümeleme

Hiyerarşik Kümeleme

Doğrusal Olmayan Boyut Azaltma(Temel bileşen analizi -Principal component analysis) **(PCA, T-SNE)**

Doğrusal Olmayan Kümeleme (DBSCAN)

Denetimli Öğrenme

Denetimli yöntemler ise veriden bilgi ve sonuç çıkarmaya yönelik kullanılmaktadır

Bu nedenle denetimsiz bir yöntemle elde edilen bir bilgi veya sonucu, eğer mümkünse denetimli bir yöntemle teyit etmek, elde edilen bulguların doğruluğu ve geçerliliği açısından önem taşımaktadır.

Sınıf sayısı belirlidir.

Denetimli Öğrenme Yöntemleri

- En yakın k komşuluk (k-Nearest-Neighbor)
- Regresyon modelleri (Regression models)
- Kural çıkarımı (Rule induction)
- Karar ağaçları (Decision trees)
- Sinir ağları (Neural networks) Denetimsiz (Unsupervised) Veri Madenciliği yöntemleri:
- Aşamalı kümeleme (Hierarchical clustering)
- Kendi kendini düzenleyen haritalar (Self organized maps) olarak sınıflandırılabilir.

Denetimli vs. Denetimsiz Öğrenme

Özellikler	Denetimli Öğrenme	Denetimsiz Öğrenme
Etiket Gereksinimi	Evet, Etiketli Veri	Hayır, Etiketsiz Veri
Örnek Algoritma	Lineer Regresyon, Karar Ağaçları, Destek Vektör Makineleri (SVM), K-En Yakın Komşu (KNN)	K-Means Kümeleme, Hiyerarşik Kümeleme, DBSCAN
Amaç	Girdi ile Etiket Arasındaki İlişkiyi Öğrenmek	Veri Yapılarını ve Desenlerini Keşfetmek
Veri Ayrımı	Eğitim ve Test Verisi Olarak Ayrılır	Genellikle Tüm Veri Kullanılır
Geri Bildirim	Doğru Yanıtlarla Model Eğitilir	Etiketsiz Veri Yapısını Keşfeder
Performans Ölçütleri	Doğruluk, Hassasiyet, Geri Çağırma, F1 Skoru vb.	Kümeleme Kalitesi, Siluet Skoru, Elbow Yöntemi vb.
Örnek Uygulamalar	Hastalık Teşhisi, Müşteri Segmentasyonu, Ev Fiyat Tahmini vb.	Pazar Segmentasyonu, Anomali Tespiti, Veri Görselleştirme
Denetleyicinin Rolü	Veriye Dayalı Kararlar Alır ve Yönlendirir	Denetleyici yok, model veriyi doğal olarak keşfeder
Veri Seti	Etiketlenmiş Giriş ve Çıkış Verileri	Etiketlenmemiş Giriş Verileri

Veri Madenciliği Teknikleri

Sınıflandırma Yöntemleri

- Karar Ağaçları
- Lojistik Regresyon
- Naive Bayes Yöntemi
- Destek Vektör Makineleri
- K-NN

İlişkilendirme Yöntemleri

- Apriori- Birliktelik Analizi
- FP-Growth Yöntemi

Kümeleme Yöntemleri

- K-Means
- DBSCAN Yöntemi
- EM – GMM Yöntem

Sınıflandırma Yöntemleri

- Sınıflandırma, belli değişkenler bakımından benzer özelliklere sahip verilerin önceden belirlenmiş sınıflara ayrılması sürecidir.
- Sınıflandırma işlemi bir sınıflandırma modeline uygun olarak yerine getirilir.
- Bir sınıflandırma modeli; tahmin değişkenleri yardımıyla hedef değişkenin hesap edildiği bir fonksiyondur.
- Sınıflandırma modeli eğitim seti ile eğitilip, test veri seti ile değerlendirilir/doğrulanır.
- Sınıflandırma Uygulama Adımları
  - ✓Modelin oluşturulması
  - ✓Modelin değerlendirilmesi
  - ✓Modelin tahmin amaçlı kullanılması

Model Oluşturma-Veri Bölme

Eğitim Veri Seti: Bu veri seti, modelin öğrenmek için kullanacağı veri kümesidir. Model, eğitim veri seti üzerinde eğitilir ve içindeki örüntüleri öğrenir.

Doğrulama Veri Seti: Eğitim sırasında modelin performansını değerlendirmek için kullanılır.

Model, doğrulama veri seti üzerinde test edilir ve performansı gözlemlenir.

Bu adım, modelin aşırı uydurma (overfitting) veya aşırı genelleme (underfitting) gibi problemlerini tespit etmek için önemlidir.

Modelin doğrulama veri seti üzerinde iyi bir performans göstermesi, eğitim veri setine iyi uyumlandığını gösterebilir.

Model Oluşturma-Veri Bölme

Test Veri Seti: Son modelin genel performansını değerlendirmek için ayrılan veri kümesidir. Bu veri seti, modelin gerçek dünya verileri üzerinde nasıl performans göstereceğini belirlemek için kullanılır. Modelin test veri seti üzerinde iyi bir performans göstermesi, genel olarak iyi bir model olduğunu gösterebilir.

Veri bölme işlemi, genellikle veri setinin belirli bir oranını her bir sete ayırmak için rastgele bir şekilde yapılır.

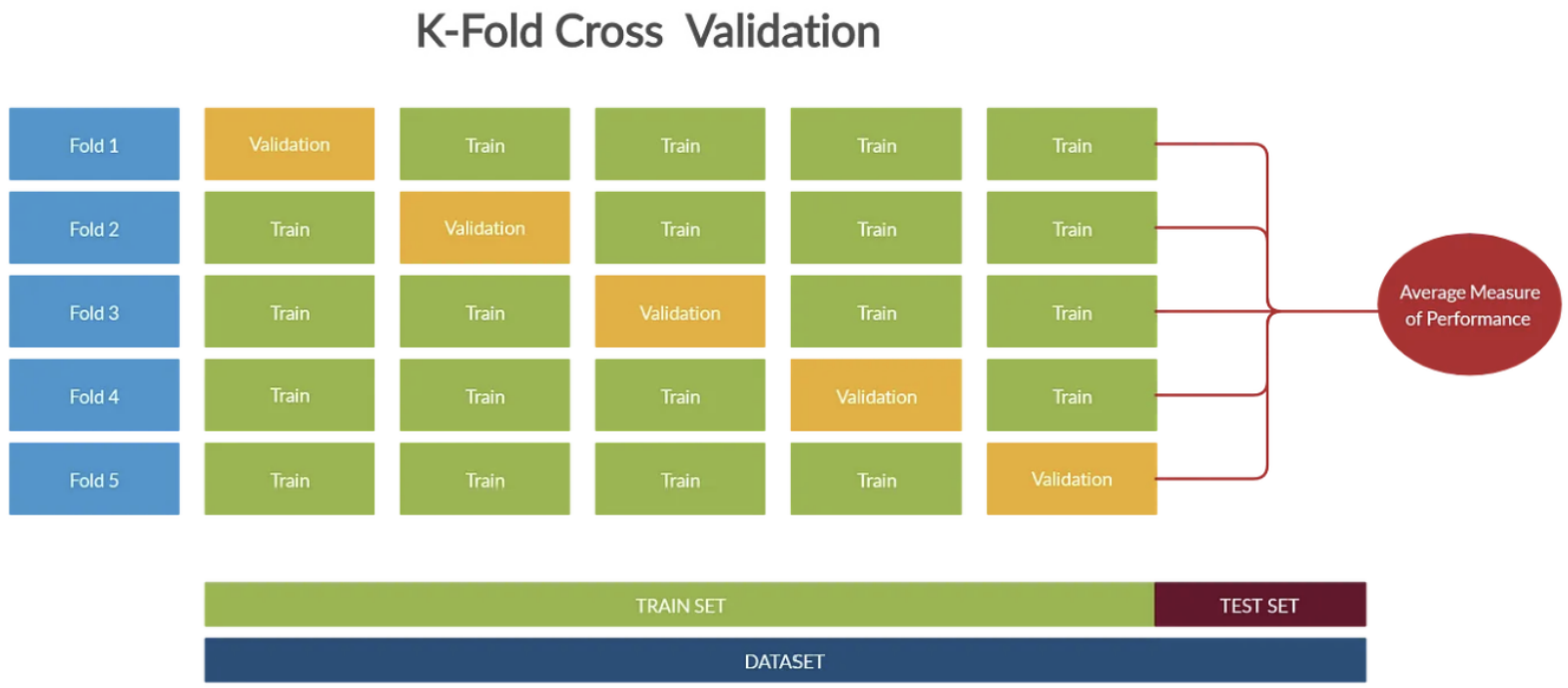
Örneğin, veri setinin %70'i eğitim, %15'i doğrulama ve %15'i test için ayrılabilir.

Ancak bu oranlar, projenin gereksinimlerine ve veri miktarına bağlı olarak değişebilir.

Önemli olan, modelin eğitim, doğrulama ve test veri setlerinde güvenilir bir şekilde değerlendirilmesidir.

K-fold Cross Validation(k-katlı çapraz doğrulama)

Literatürde bölmeleme işlemi k-fold cross validation(k-kadar çapraz doğrulama) olarak isimlendirilir.



Sınıflandırma Model Özellikleri

Doğruluk (Accuracy)	• Modelin, problemi gerçek sonuca en yakın şekilde çözebilmesi
Hız (Speed)	• Modeli oluşturmak için gerekli sürenin makul olması • Sınıflandırma için gerekli sürenin kabul edilebilir olması
Kararlılık (Robustness)	• Gürültülü & Eksik veri için de iyi sonuç vermesi • Farklı zamanlarda benzer sonuçlar verebilmesi
Ölçeklenebilirlik (Scalability)	• Büyük miktarda veri ile çalışabilmesi
Anlaşılabilirlik (Understandability)	• Kullanıcı tarafından yorumlanabilir olması

Sınıflandırma – Kullanım Alanları

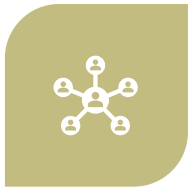
- İstisna Sapması
- Kredi başvurusu değerlendirme
- Hastalık teşhisi
- Ses tanıma
- Karakter tanıma
- Metinleri konularına göre ayırma
- Kullanıcı davranışları belirleme
- Resim tanıma

Sınıflandırma Yöntemleri

BAYES TABANLI SNIFLANDIRICILAR (BAYES CLASSİFİERS)

YAPAY SİNİR AĞLARI (ARTİFİCİAL NEURAL NETWORKS)

LOJİSTİK REGRESYON (LOGİSTİCS REGRESSİON)



İLİŞKİ TABANLI SINIFLANDIRICILAR (ASSOCİATİON-BASED CLASSİFİERS)

K-EN YAKIN KOMŞU YÖNTEMİ (K- NEAREST NEİGHBOOR METHOD)

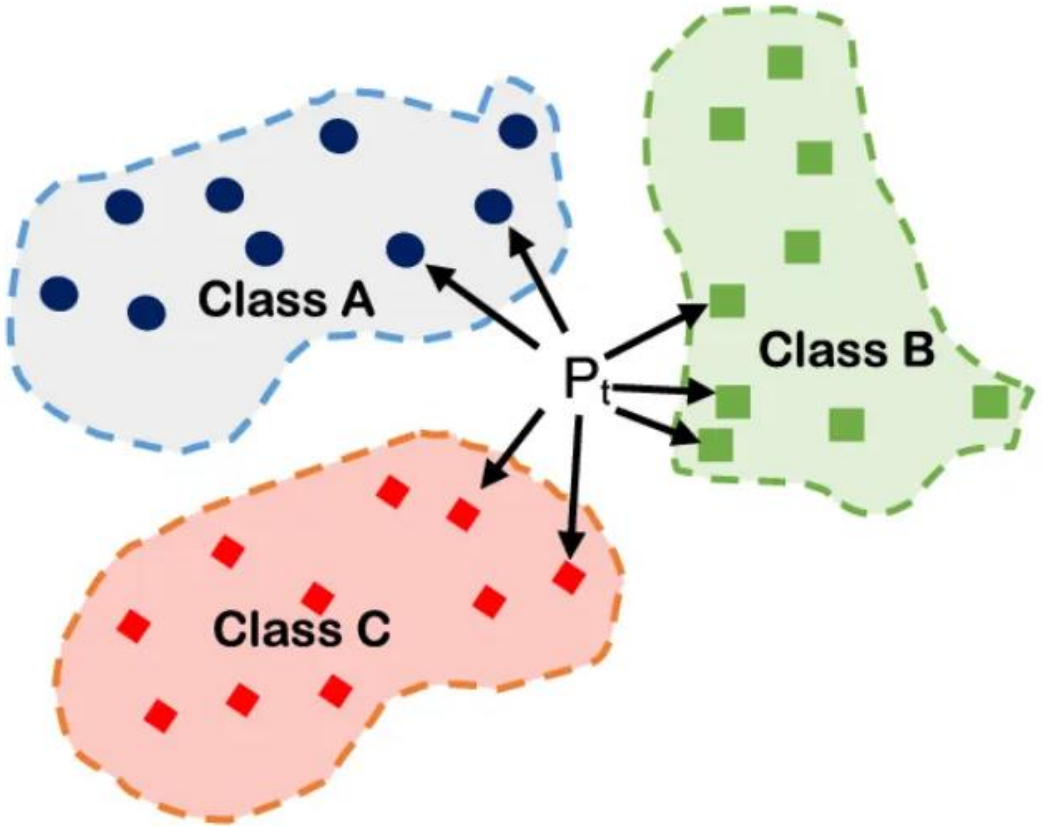
DESTEK VEKTÖR MAKİNELER (SUPPORT VECTOR MACHİNES)

GENETİK ALGORİTMALAR (GENETİC ALGORİTHMS)

KARAR AĞAÇLARI (DECİSİON TREES

k-En Yakın Komşu Yöntemi(k-nearest neighbors, KNN)

- Temel çalışma mantığı, verilerin birbirleriyle olan uzaklıkları ve benzerliklerini kullanarak sınıflandırma işlemi gerçekleştirmektedir.
- Karar sınıfı bilinmeyen bir veri geldiğinde verinin hangi sınıfa ait olduğunun belirlenmesi için sınıfı bilinmeyen veriye en yakın k adet veri belirlenir.
- Daha sonra ise, veri kendisine yakın olan bu k adet veriden hangisine daha çok benziyorsa onun sınıfında etiketlenir.



k-En Yakın Komşu Yöntemi Uygulama Aşaması

- En yakın komşu sayısı k değerini belirle
- Yeni bir değer ele al ve bütün değerler için bu değere uzaklıklarını hesapla
- Hesaplanan değerleri küçükten büyüğe sırala
- k. değere kadar olan değerler için sınıf değişkenini incele
- Maksimum frekanslı sınıfa yeni değeri ata

k-En Yakın Komşu Yöntemi

- Bu teknikte yeni bir durum daha önce sınıflandırılmış, benzer en yakın komşuluktaki k tane olaya bakılarak sınıflandırılır.
- Uzaklık ölçütü olarak genellikle Öklid uzaklıkları alınır.
- kk en yakın komşuluğundaki olayların ait olduğu sınıflar sayılır ve yeni durum sayısı fazla olan sınıfa dahil edilir.
- Bu yöntemin tercih edilme sebebi, sayısı bilinen veri kümeleri için hızlı ve verimli olmasıdır.

Veriler Arası Uzaklık Nasıl Hesaplanır?

Euclidean

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan

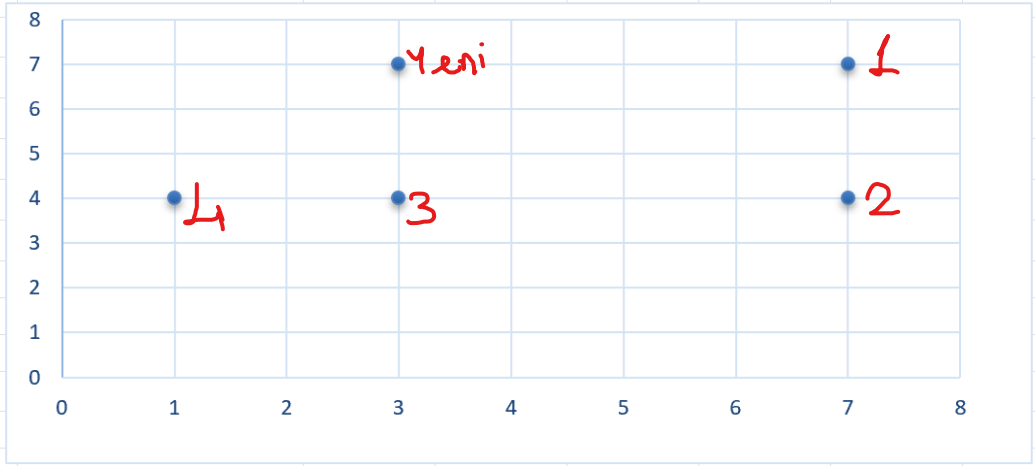
$$\sum_{i=1}^{k}|x_i - y_i|$$

Minkowski

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

k-En Yakın Komşu Yöntemi Örnek

	X	*Y*	*Sınıflama*
Nesne 1	7	7	Sınıf_1
Nesne 2	7	4	Sınıf_1
Nesne 3	3	4	Sınıf_2
Nesne 4	1	4	Sınıf_2
Yeni Nesne	3	7	???



$$Fark_e = \sqrt{(X_i - X_{yeni})^2 + (Y_i - Y_{yeni})^2}$$

	Yeni Nesne ile Uzalıklık
Nesne 1	4
Nesne 2	5
Nesne 3	3
Nesne 4	3,606

Sıralama	Sınıfları
Nesne 4	Sınıf_2
Nesne 3	Sınıf_2
Nesne 1	Sınıf_1
Nesne 2	Sınıf_1

K=3 için
Nesne 4, 3, 1 seçilir
Kazanan Sınıf_2'dir

k-En Yakın Komşu Yöntemi Örnek

- İki sütunumuz var: Parlaklık ve Doygunluk.
- Tablodaki her satırın Kırmızı veya Mavi sınıfı vardır.
- Yeni bir veri girişi yapmadan önce K değerinin 5 olduğunu varsayalım.

BRIGHTNESS	SATURATION	CLASS
40	20	Red
50	50	Blue
60	90	Blue
10	25	Red
70	70	Blue
60	10	Red
25	80	Blue

https://www.veribilimiokulu.com/k-en-yakin-komsu-k-nearest-neighbor-siniflandirma-python-ornek-uygulama/

k-En Yakın Komşu Yöntemi Örnek

- Yeni bir girişimiz var ancak henüz bir sınıfı yok.
- Sınıfını bilmek için, Öklid mesafe formülünü kullanarak yeni girişten veri setindeki diğer girişlere olan mesafeyi hesaplamamız gerekir.

BRIGHTNESS	SATURATION	CLASS
20	35	?

BRIGHTNESS	SATURATION	CLASS
40	20	Red
50	50	Blue
60	90	Blue
10	25	Red
70	70	Blue
60	10	Red
25	80	Blue

$$\text{Öklid Uzaklığı: } \left\{ d_{i,j} = \sqrt{\sum_{k=1}^{p} \left| x_{i,k} - x_{j,k} \right|^2} \right\}$$

k-En Yakın Komşu Yöntemi Örnek

X_2 = Yeni girişin parlaklığı (20).

X_1= Mevcut girişin parlaklığı.

Y_2 = Yeni girişin doygunluğu (35).

Y_1 = Mevcut girişin doygunluğu.

$d_1 = \sqrt{(20 - 40)^2 + (35 - 20)^2} = 25$

$d_2 = \sqrt{(20 - 50)^2 + (35 - 50)^2} = 33.54$

$d_3 = \sqrt{(20 - 60)^2 + (35 - 90)^2} = 68.01$

BRIGHTNESS	SATURATION	CLASS
20	35	?

BRIGHTNESS	SATURATION	CLASS	DISTANCE
40	20	Red	25
50	50	Blue	33.54
60	90	Blue	68.01
10	25	Red	?
70	70	Blue	?
60	10	Red	?
25	80	Blue	?

k-En Yakın Komşu Yöntemi Örnek

BRIGHTNESS	SATURATION	CLASS	DISTANCE
40	20	Red	25
50	50	Blue	33.54
60	90	Blue	68.01
10	25	Red	10
70	70	Blue	61.03
60	10	Red	47.17
25	80	Blue	45

BRIGHTNESS	SATURATION	CLASS	DISTANCE
10	25	Red	10
40	20	Red	25
50	50	Blue	33.54
25	80	Blue	45
60	10	Red	47.17
70	70	Blue	61.03
60	90	Blue	68.01

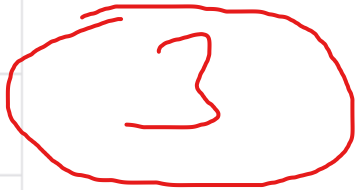
BRIGHTNESS	SATURATION	CLASS	DISTANCE
10	25	Red	10
40	20	Red	25
50	50	Blue	33.54
25	80	Blue	45
60	10	Red	47.17

Küçükten büyüğe sıralarsak

K değerini 5 seçtiğimiz için yalnızca ilk beş satırı dikkate alacağız.

k-En Yakın Komşu Yöntemi Örnek

BRIGHTNESS	SATURATION	CLASS	DISTANCE
10	25	Red	10
40	20	Red	25
50	50	Blue	33.54
25	80	Blue	45
60	10	Red	47.17



BRIGHTNESS	SATURATION	CLASS
40	20	Red
50	50	Blue
60	90	Blue
10	25	Red
70	70	Blue
60	10	Red
25	80	Blue
20	35	Red