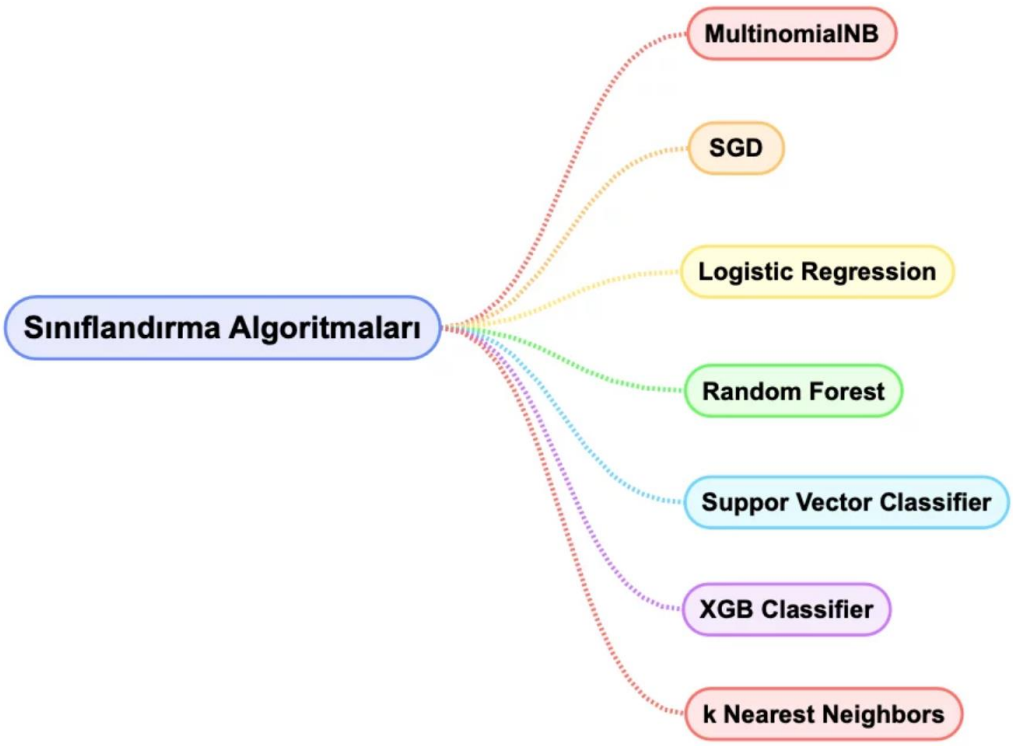


Sınıflandırma Nedir?

Makine öğreniminde sınıflandırma, öğeleri önceden kategorize edilmiş bir eğitim veri kümesine göre kategorilere ayırma sürecidir. Sınıflandırma, denetimli bir öğrenme algoritması olarak kabul edilir.

Sınıflandırma algoritmaları, yeni bir öğenin tanımlanan kategorilerden birine girme olasılığını hesaplamak için eğitim verilerinin kategorizasyonunu kullanır.

Örnek olarak gelen epostaların spam veya spam değil olarak işaretlenmesini verebiliriz.

Farklı sınıflandırma algoritmalarından bazıları;

- KNN ((**K**-Nearest Neighbors)
- Karar Ağaçları (Decision Trees)
- Naive Bayes
- SVM (Support Vector Machine)
- Random Forest - Rastgele Orman
- Logistic Regression - Lojistik Regresyon
- Neural Networks - Nöral Ağlar

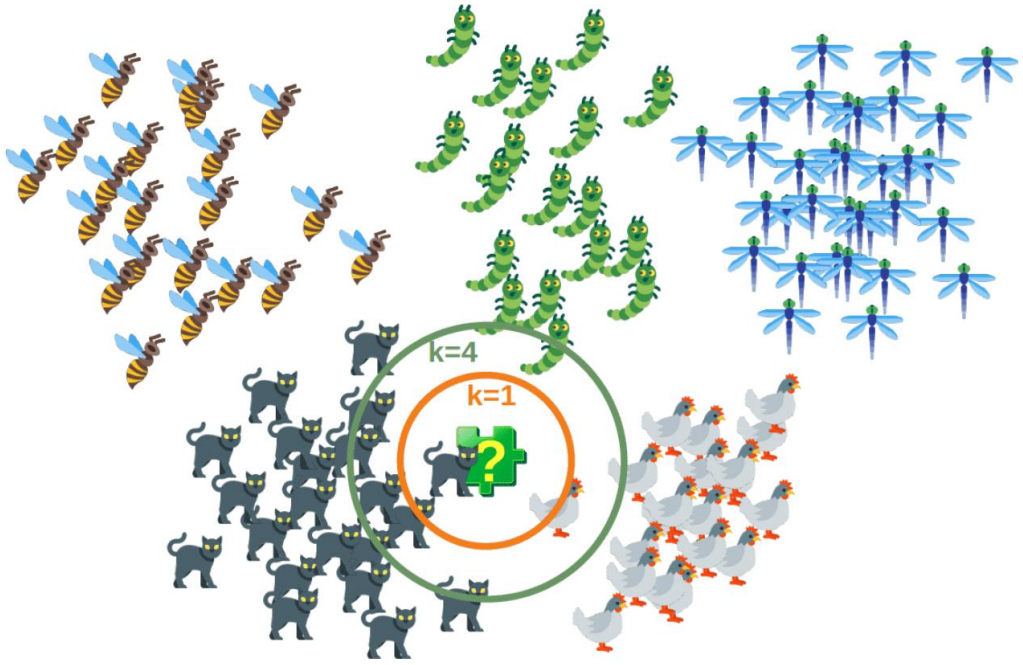
KNN (K-en yakın komşular)

KNN genellikle sınıflandırma problemleri için kullanılan bir gözetimli öğrenme algoritmasıdır. Benzerlik fonksiyonlarını kullanır.

KNN, bazı veri kümelerinde en yakın k veri noktasını bulmak için eğitim veri kümelerini kullanan bir algoritmadır. KNN, hem regresyon hem de sınıflandırma problemleri için kullanılan denetimli bir makine öğrenmesi algoritmasıdır. Genellikle örüntü tanıma için uygulanır.

Bu algoritma öncelikle verilerdeki tüm girdiler arasındaki mesafeyi depolar ve tanımlar, sorguya ve çıktılara en yakın girdiyi seçer.

KNN algoritmaları; gerçek hayatta parmak izi algılama, **kredi notu, borsa tahmini, kara para aklama analizi, iflas** ve **döviz kuru** alanlarında kullanılır.



Adım adım KNN:

1. **Veri setimizi algoritmaya veririz.**
2. **K, komşu değerini belirleriz.**
3. **Eğitim veri setinde kullanılan gözlemler ile sınıflandırma yapmak istediğimiz gözlem arasındaki mesafeleri hesaplarız.**
4. **Mesafeleri bir kümeye ekleriz.**
5. **Mesafeleri küçükten büyüğe sıralarız.**

6. **Gözlem ile arasındaki mesafe minimum olandan maximum olana doğru K kadar gözlem seçeriz. Bunlar en yakın komşulardır.**
7. **En uygun komşu kategorisi seçeriz.**
8. **KNN der ki, sınıflandırma yapmak istediğin gözlemin kategorisi, önceki adımda seçtiğimiz kategori ile aynı.**

Karar Ağaçları

Karar Ağacı algoritması, denetimli makine öğrenmesi türüdür. Regresyon ve sınıflandırma problemlerini çözmek için kullanılır. Amaç, gözlemlerden sonuçları işlemeye geçmek için bir karar ağacından yararlanmaktır.

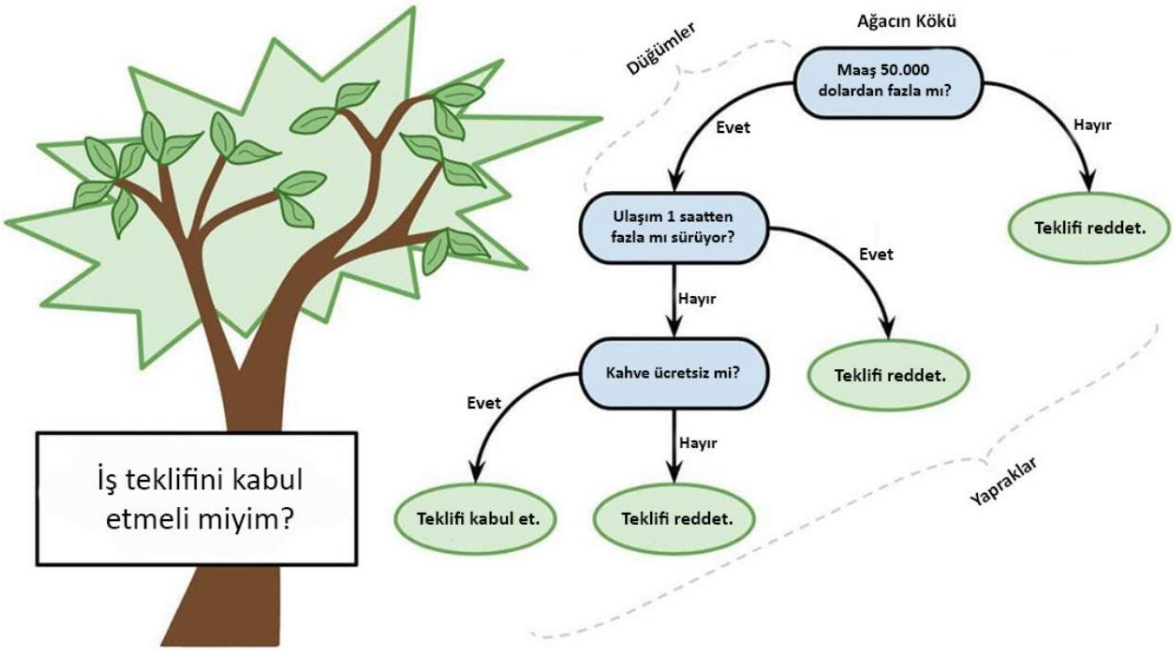
Karar ağaçlarının işlenmesi, eğitim verilerinden en uygun özniteliğin kök olarak seçildiği ve işlemin her dal için tekrarlandığı yukarıdan aşağıya bir yaklaşım benimser. Karar ağaçları genellikle şu alanlarda kullanılır:

- Bilgi yönetimi platformu oluşturma
- Uçuş seçme
- Oteller için yüksek doluluk tarihlerinin tahmin edilmesi
- Müşterilere ürün önerme
- Tahminleri tahmin etme
- Çeşitli alanlardaki olasılıkları belirleme

Adım adım Karar Ağaçları:

1. Veriseti içerisinden bir özellik seçilir.
2. Bu özelliğin, verinin ayrılmasındaki önemi hesaplanır.
3. Yukarıdaki işlem, bölme işini en iyi yapan özelliği bulana kadar tekrar edilir.
4. En iyi bölen özellik bulununca bir kural oluşturulur.
5. Yukarıdaki adımlar önem sırasına göre tüm özellikler için tekrarlanır.
6. Oluşturulan tüm kurallar verinin ayrılmasındaki önem sırasına göre sıralanır.

7. En önemli ayrıştırıcı özellik kök olur, önemi az olana doğru giderken yapraklara tek tek ulaşılır.



Naive Bayes

Koşullu olasılık kuralını kullanarak bir öğenin belirli bir kategoriye girme olasılığını hesaplayan bu algoritma, oldukça etkili bir denetimli makine öğrenimi algoritması olarak bilinir.

Sınıf değişkeninin değeri göz önüne alındığında, Bayes'in teoremini verilere uygulayarak, her özellik çifti arasında saf bir koşullu bağımsızlık varsayımı ile çalışır. Daha basit bir ifadeyle, B olayının meydana geldiği göz önüne alındığında, bir A olayının olma olasılığını bulmaya yardımcı olur. Kullanıldığı durumlar:

- İstenmeyen mesajları filtreleme
- Netflix gibi öneri sistemleri
- Teknoloji, politika veya sporla ilgili haber makalelerini sınıflandırma
- Sosyal medyada duygu analizi
- Yüz tanıma yazılımları

SVM

Bu algoritmada veriler, X / Y tahmininin ötesine geçebilen polarite derecesine göre sınıflandırılır. SVM, denetimli makine öğrenme algoritmaları arasında kategorize edilir ve öncelikle sınıflandırma ve regresyon analizi için kullanılır. Algoritma, bir kategoriye yeni örnekler ve veriler atayan modeller oluşturarak çalışır.

SVM, boyut sayısının örnek sayısından daha ağır bastığı durumlarda oldukça etkilidir ve bellek açısından son derece verimlidir.

SVM algoritmalarının bulunduğu uygulamalar:

- Biyoinformatik
- Metin ve Köprü Metni Kategorizasyonu
- El yazısı tanıma
- Tedavi amaçlı ilaç keşfi
- Görüntü sınıflandırma
- Yüz tanıma

Topluluk (Ensembling) Algoritması

Topluluk algoritmaları, daha doğru sonuçlar elde etmek için iki veya daha fazla makine öğrenmesi algoritmasının tahminini birleştirir. Sonuçları birleştirmek, oylama veya sonuçların ortalaması alınarak yapılabilir. Oylama genellikle regresyon sırasında sınıflandırma ve ortalama alma sürecinde kullanılır. Topluluk algoritmalarının 3 temel türü vardır: Bagging, Boosting ve Stacking.

Bagging algoritmaları, hepsi eşit büyüklükte farklı eğitim setlerinde paralel olarak çalıştırılır. Tüm algoritmalar daha sonra aynı veri kümesi kullanılarak test edilir ve genel sonuçları belirlemek için oylama kullanılır.

Boosting algoritmaları ise sıralı olarak çalıştırılır. Daha sonra genel sonuçlar ağırlıklı oylama kullanılarak seçilir.

Stacking algoritmalarının, üst üste yığılmış iki düzeyi bulunur: Temel düzey; algoritmaların bir kombinasyonudur ve üst düzey; temel düzey sonuçlarına dayalı bir meta algoritmadır.

Kümeleme (Clustering)

Kümeleme algoritmaları, veri noktalarını gruplamak için kullanılan denetimsiz algoritmalar grubudur. Aynı küme içindeki noktalar, farklı kümelerdeki noktalardan daha çok birbirine benzer.

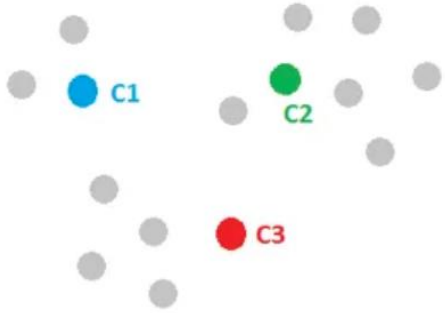
Uygulamaları, Python, SciPy, Sci-Kit Learn ve veri madenciliği gibi programlama dillerinde ve kitaplıklarında benzer ve ilgili web arama sonuçlarını kümelemeye kadar uzanır.

Kümeleme algoritmaları sahte haberleri belirleme, spam algılama ve filtreleme, kitapları veya filmleri türe göre sınıflandırma ve şehir planlaması sırasında popüler ulaşım yollarını belirleme gibi durumlar için kullanılır.

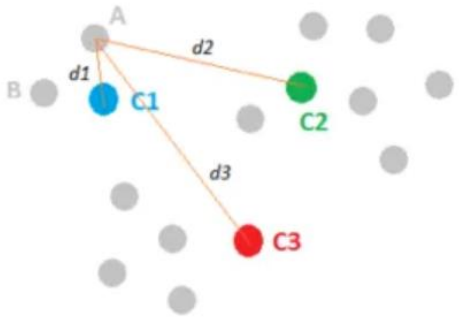
4 tür kümeleme algoritması bulunur:

- **Centroid Tabanlı Kümeleme**
  - Bu kümeleme algoritması (Centroid-based Clustering), verileri başlangıç koşullarına ve aykırı değerlere göre kümeler. K-means, en çok kullanılan centroid tabanlı kümeleme algoritmasıdır.
- **Dağıtım Tabanlı Kümeleme**
  - Bu kümeleme algoritması (Distribution-based Clustering), verilerin olasılık dağılımlarından oluştuğunu varsayar ve ardından verileri bu dağılımın çeşitli sürümlerinde kümeler.
- **Yoğunluğa Dayalı Kümeleme**
  - Bu kümeleme türünde (Density-based Clustering), algoritma yüksek yoğunluklu alanları rastgele şekilli dağılımlar oluşturan kümelere bağlar.
- **Hiyerarşik Kümeleme**
  - Bu algoritma (Hierarchical Clustering), hiyerarşik veri kümelerinden oluşan bir ağaç oluşturur. Küme sayısı, ağaç doğru seviyede kesilerek değiştirilebilir

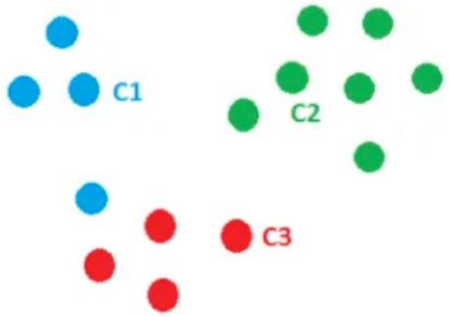
K-Means Kümeleme Algoritması
Adım adım çalışma mantığına bakalım. İlk adımda kümeler için bir merkez noktası rastgele olarak belirlenir.



Merkez eleman dışındaki elemanlar en yakın merkez noktasına göre kümelenir. Bunun için her bir elemanın merkez noktalara olan uzaklıkları ölçülür.

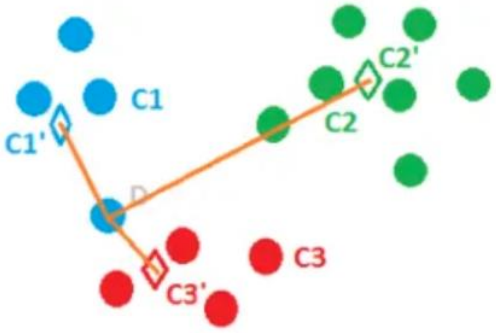


Kümelenen elemanlarımız böyle bir

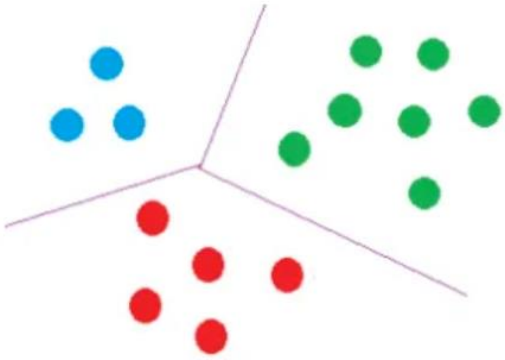


hal alır.

Ardından aynı kümeler içerisinde tekrar bir merkez noktası hesaplaması yapılır.



Ardından merkez noktası stabil hale gelinceye dek ilk iki adım tekrarlanır.



K-means algoritması çoğu büyük — küçük veri kümesi için ideal bir algoritma olsa da her algoritmada olduğu gibi onun da bazı problemleri vardır.

İlişkilendirme (Association)

İlişkilendirme algoritmaları, bazı öğelerin belirli bir veri kümesinde birlikte oluşma olasılığını keşfetmek için kullanılan denetimsiz algoritmalardır. Çoğunlukla alışveriş sepeti analizinde kullanılır. En çok kullanılan ilişkilendirme algoritması Apriori'dir:

Apriori Algoritması: İşlemsel veri tabanlarında yaygın olarak kullanılan bir madencilik algoritmasıdır. Apriori, sık kullanılan eşya setlerini çıkarmaya ve bu setlerden bazı ilişki kuralları oluşturmaya yarar.

Veri kümelerinde ortak öğe kümelerini arayarak çalışan ve daha sonra bunlar üzerinde ilişkiler kuran apriori; genellikle ilişkisel veri tabanlarında öğe kümesi madenciliği ve ilişkilendirme kuralı öğrenimi için kullanılır.

Bu algoritmanın arkasındaki fikir, daha kullanışlı bir ilişkilendirme oluşturmak için ilgili öğeleri mümkün olduğunca daha büyük bir kümeye genişletmektir. Büyük veri kümeleriyle kullanılabilir.