Makine Öğrenmesi Sınıflandırma Algoritmaları

Sınıflandırma

Sınıflandırma Nedir?

•Sınıflandırma, verilen veri noktalarının sınıfını tahmin etme işlemidir.

- Sınıflandırma, nesnelerin ve fikirlerin önceden belirlenmiş kategoriler halinde tanınması, anlaşılması ve gruplandırılması süreci olarak tanımlanır, Buna “alt popülasyonlar” denir.

•En çok kullanılan 5 sınıflandırma algoritması,

- Random Forest
- Logistic Regression
- Naive Bayes
- K-Nearest Neighbors
- Decision Tree
- Support Vector Machines

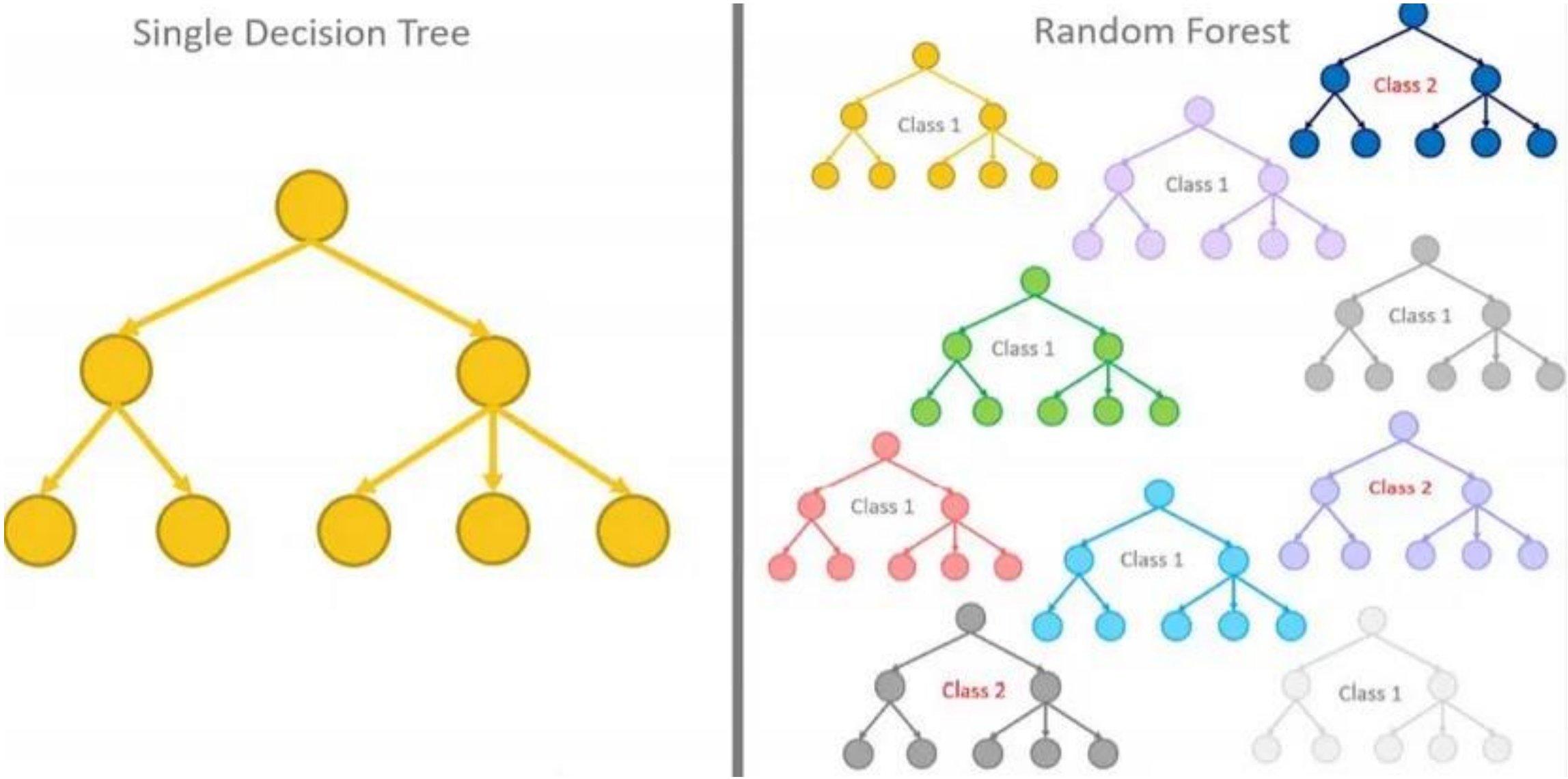
Random Forest

•Random forest, hem regresyon hem de sınıflandırma problemlerinde kullanılmaktadır. Algoritma, birden fazla karar ağacı üreterek sınıflandırma işlemi esnasında sınıflandırma değerini yükseltmeyi hedefler.

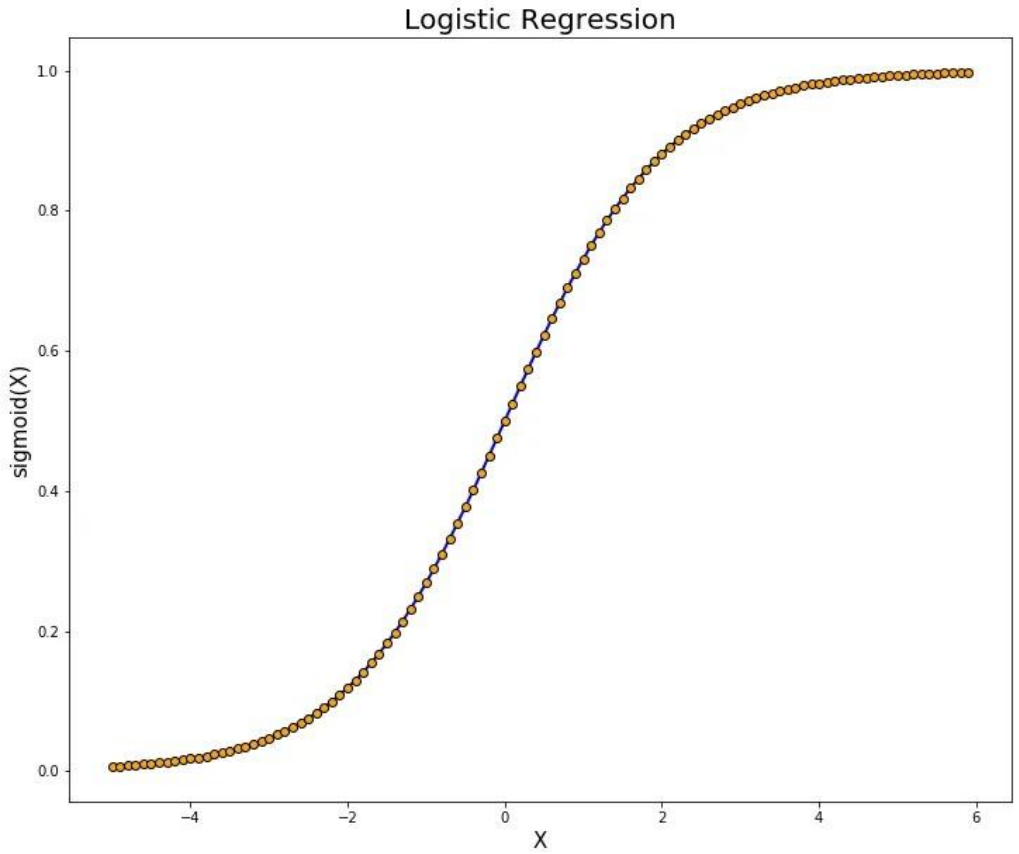
- Sınıflandırma ve Regresyon problemlerinde yaygın olarak kullanılan Denetimli Makine Öğrenimi Algoritmasıdır. Farklı örnekler üzerine karar ağaçları oluşturur ve regresyon durumunda sınıflandırma ve ortalama için çoğunluk oylarını alır.

•Random forest algoritması birbirinden bağımsız olarak çalışan birçok karar ağacının bir araya gelerek aralarından en yüksek puan alan değerin seçilmesi işlemidir.

- Sınıflandırma ve Regresyon problemlerinde yaygın olarak kullanılan Denetimli Makine Öğrenimi Algoritmasıdır.
- Random forest algoritması, elinde yeterli miktarda ağaç varsa aşırı öğrenme sorununu azaltır. Az oranda bir veri hazırlığına ihtiyaç duyar. Farklı örnekler üzerine karar ağaçları oluşturur ve regresyon durumunda sınıflandırma ve ortalama için çoğunluk oylarını alır


Single Decision Tree
Random Forest
Class 1
Class 2
Class 1
Class 1
Class 1
Class 1
Class 1
Class 2
Class 1
Class 2
Class 1

Logistic Regression



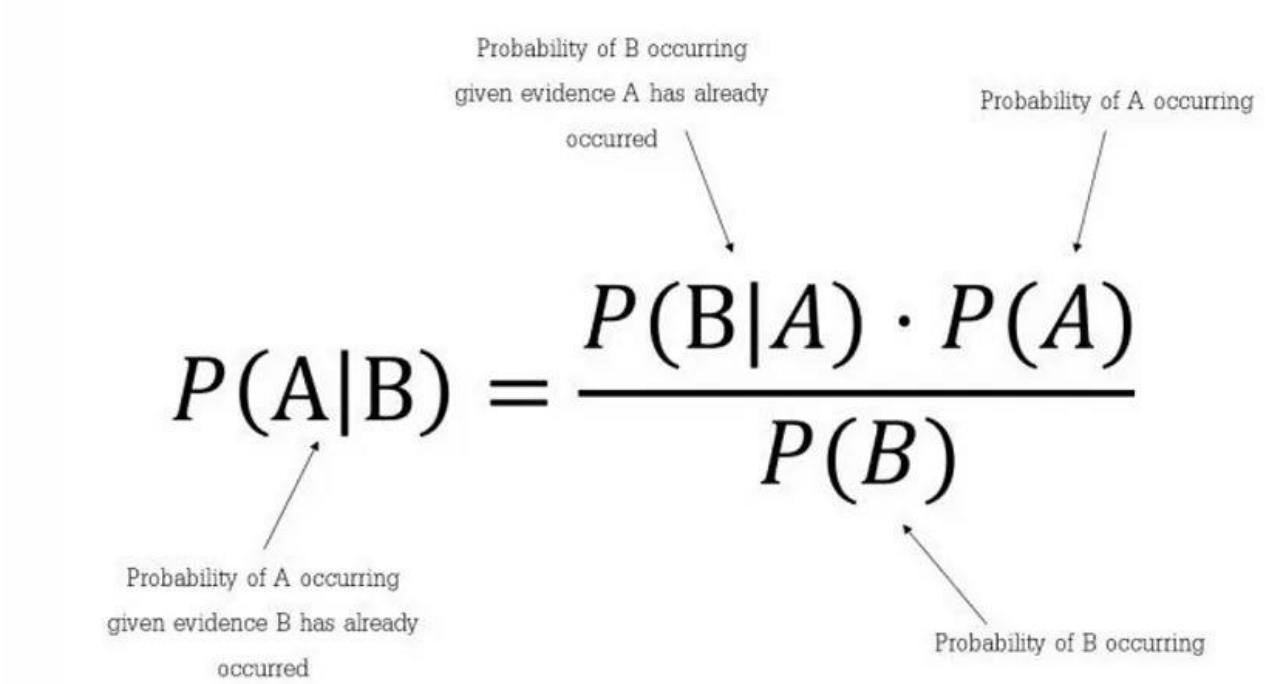
•Lojistik regresyon, gözlemleri ayrı bir sınıf kümesine atamak için ve sınıflandırma problemleri için kullanılan bir makine öğrenme algoritmasıdır.

- Öngörücü bir analiz algoritmasıdır ve olasılık kavramına dayanır.
- Bir olasılık değeri döndürmek için lojistik sigmoid fonksiyonunu kullanarak çıktısını dönüştürür.

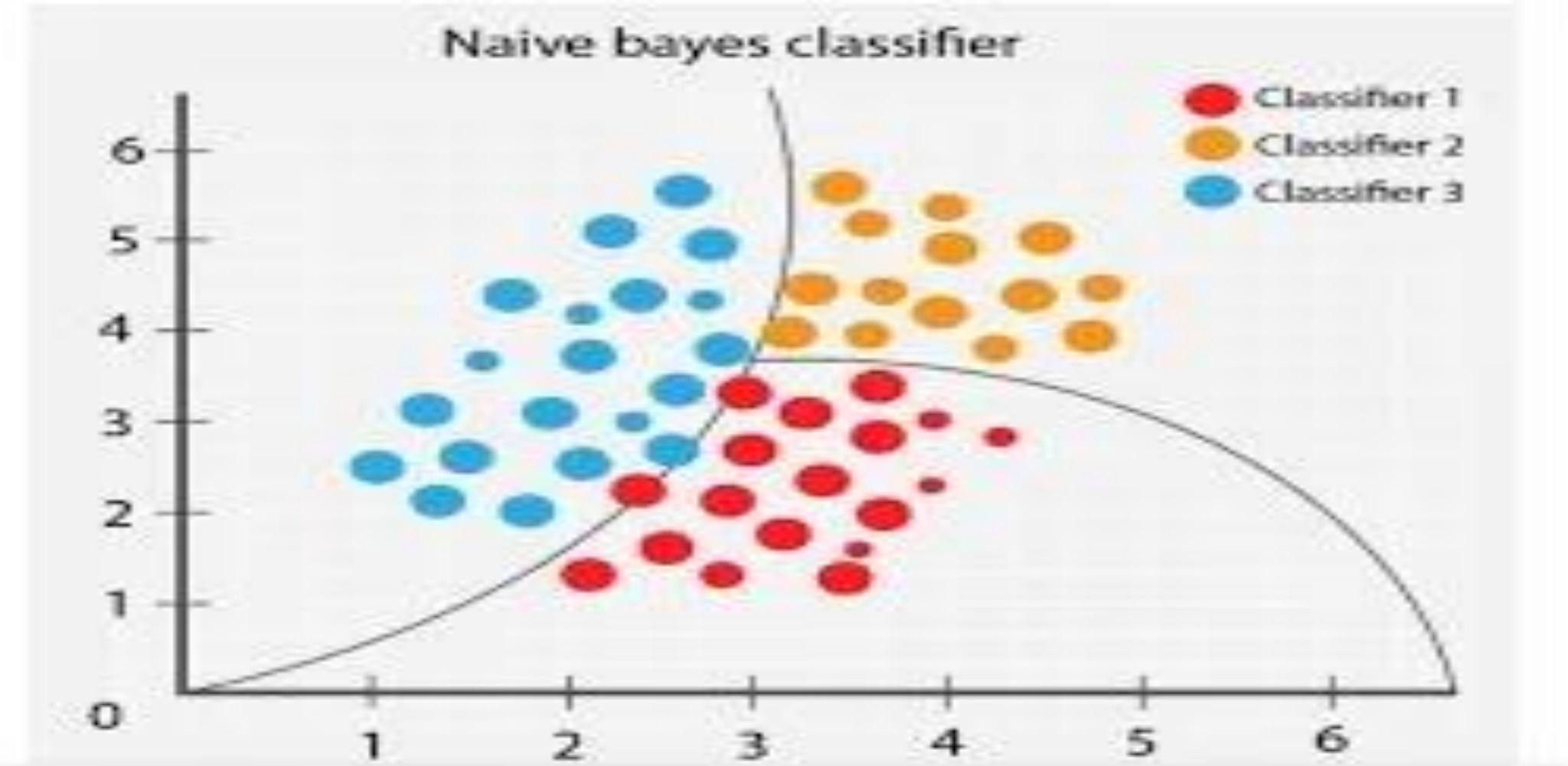
•Lojistik Regresyona Doğrusal Regresyon modeli diyebiliriz, ancak Lojistik Regresyon daha karmaşık bir maliyet fonksiyonu kullanır, bu maliyet fonksiyonu doğrusal bir fonksiyon yerine ‘Sigmoid fonksiyonu’ veya ‘lojistik fonksiyon’ olarak da bilinir.

- Lojistik regresyon hipotezi, maliyet fonksiyonunu 0 ile 1 arasında sınırlama eğilimindedir. Bu nedenle, doğrusal fonksiyonlar onu temsil edemez, çünkü lojistik regresyon hipotezine göre mümkün olmayan 1'den büyük veya 0'dan küçük bir değere sahip olabilir.

Formül,


Probability of B occurring given evidence A has already occurred
Probability of A occurring
$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{B}|A) \cdot P(A)}{P(B)}$
Probability of A occurring given evidence B has already occurred
Probability of B occurring

Naive Bayes

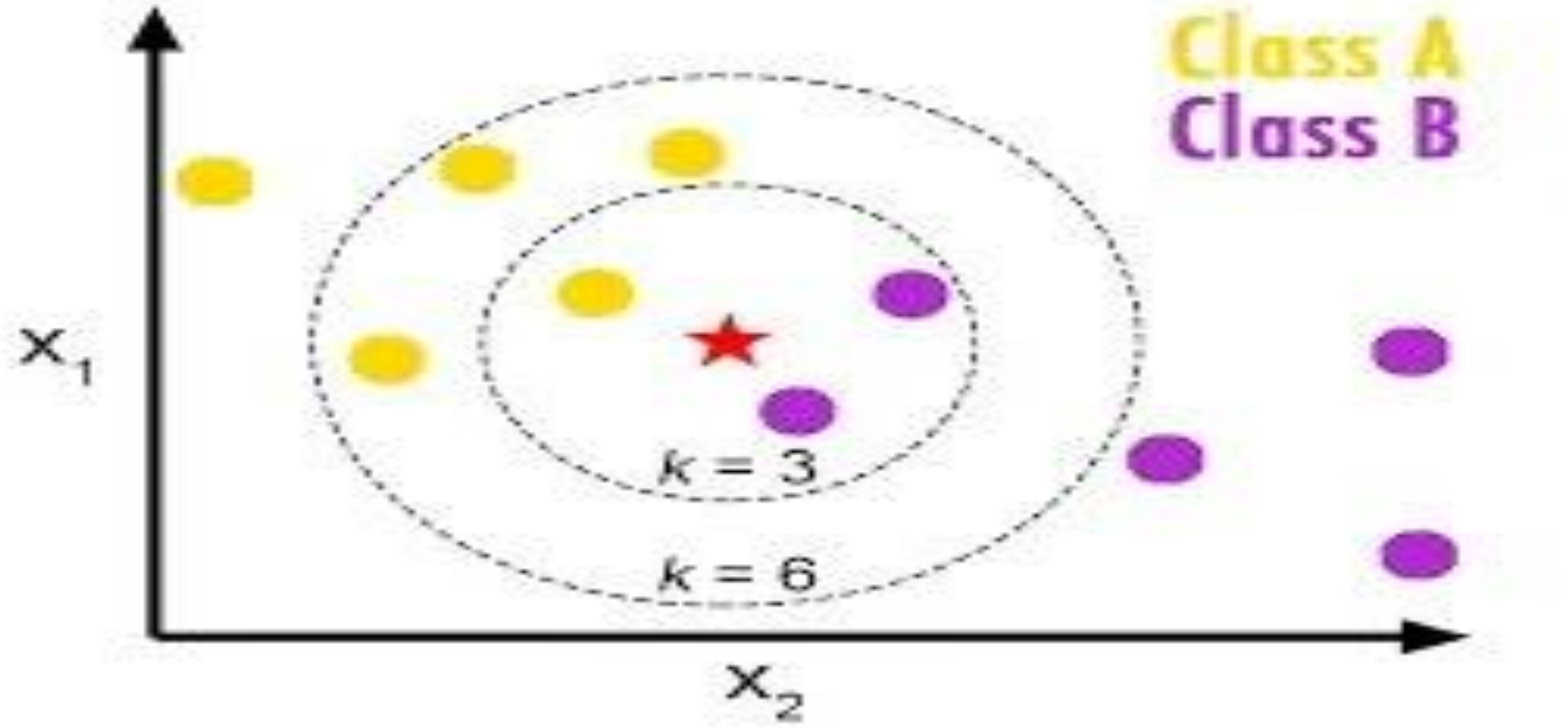

Naive bayes classifier
Classifier 1
Classifier 2
Classifier 3
0
1
2
3
4
5
6
1
2
3
4
5
6

Naïve Bayes, çok çeşitli sınıflandırma görevlerinde kullanılan Bayes Teoremine dayanan olasılıksal bir makine öğrenme algoritmasıdır
Formül

$$P(c \mid x) = \frac{P(c \mid x)\, P(c)}{P(x)}$$

- Naïve Bayes sınıflandırmasında sisteme belirli bir oranda öğretilmiş veri sunulur (Örn: 100 adet). Öğretim için sunulan verilerin mutlaka bir sınıfı/kategorisi bulunmalıdır. Öğretilmiş veriler üzerinde yapılan olasılık işlemleri ile sisteme sunulan yeni test verileri, daha önce elde edilmiş olasılık değerlerine göre işletilir ve verilen test verisinin hangi kategoride olduğu tespit edilmeye çalışılır. Öğretilmiş veri sayısı arttıkça test verisinin gerçek kategorisini tespit etme şansı o kadar yükselir.

K-Nearest Neighbors



•K-en yakın komşular (KNN) algoritması, hem sınıflandırma hem de regresyon problemlerini çözmek için kullanılabilecek basit, uygulaması kolay bir denetimli makine öğrenimi algoritmasıdır.

- Her ne kadar KNN algoritması k-means algoritmasındaki benzer özellikler taşısa da büyük farklılıklar da içermektedir. KNN algoritması bir eğitim verisi içerirken k-means algoritması bir eğitim verisi içermez. Yeni bir değer geldiğinde K değerine mesafeler hesaplanır ve yeni değer bir kümeye ilave edilir. Mesafe hesaplama işleminde ise k-means ve hiyerarşik kümeleme de kullanılan öklid uzaklığı, manhattan uzaklığı gibi mesafe hesaplama yöntemleri kullanılabilir

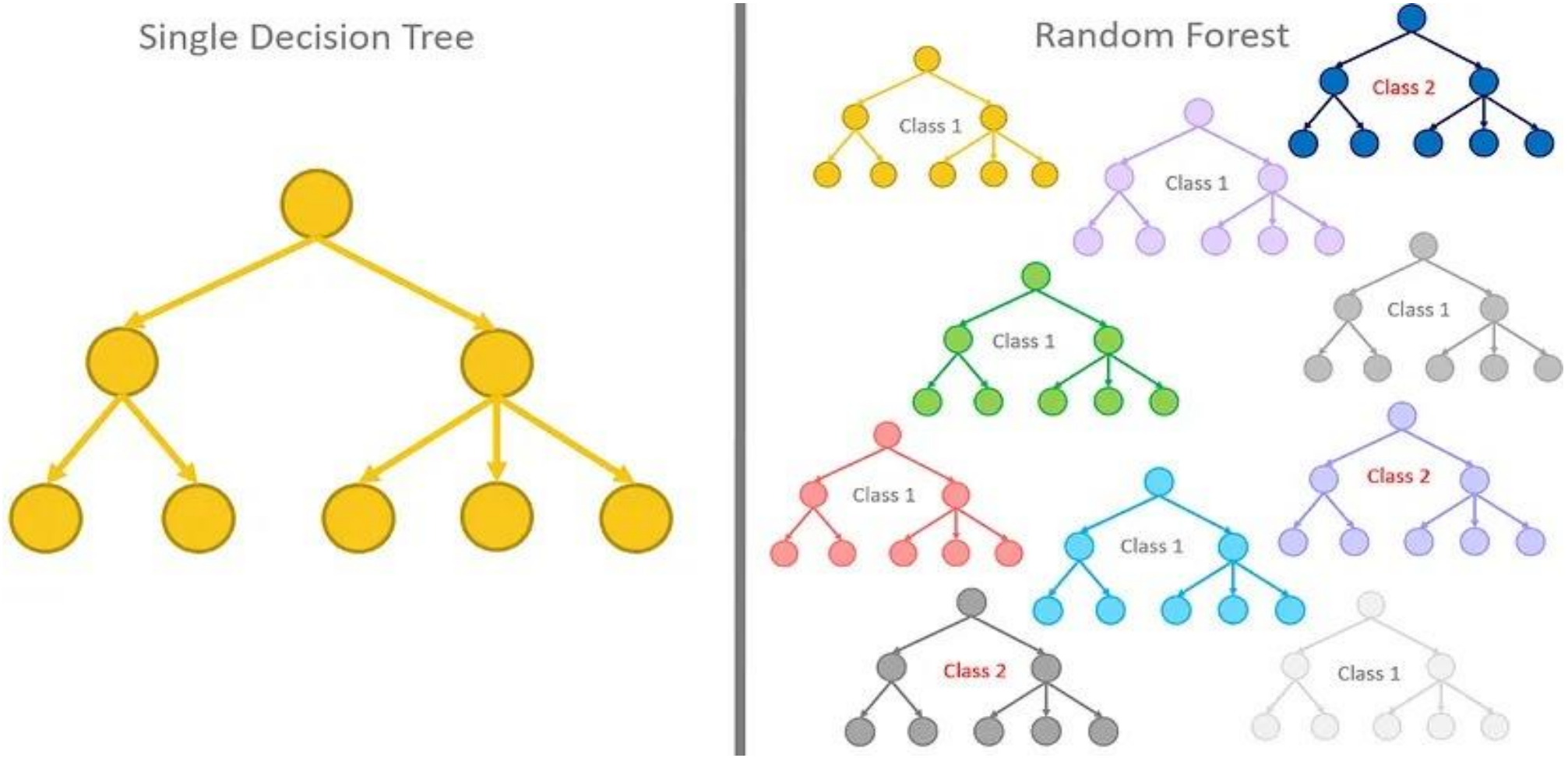
•KNNnin en temel haliyle uygulanması son derece kolaydır ve yine de oldukça karmaşık sınıflandırma görevlerini yerine getirir.

- Özel bir eğitim aşamasına sahip olmadığı için tembel bir öğrenme algoritmasıdır.

- Bunun yerine, yeni bir veri noktasını veya örneğini sınıflandırırken eğitim için tüm verileri kullanır.

- KNN, parametrik olmayan bir öğrenme algoritmasıdır, yani altta yatan veriler hakkında hiçbir şey varsaymaz. Bu son derece kullanışlı bir özelliktir, çünkü gerçek dünya verilerinin çoğu, doğrusal ayrılabilirlik, tekdüze dağılım vb. gibi herhangi bir teorik varsayımı gerçekten takip etmemektedir

Decision Tree



•Karar Ağacı algoritması, denetimli öğrenme algoritmaları ailesine aittir. Diğer denetimli öğrenme algoritmalarının aksine, karar ağacı algoritması regresyon ve sınıflandırma problemlerini çözmek için de kullanılabilir.

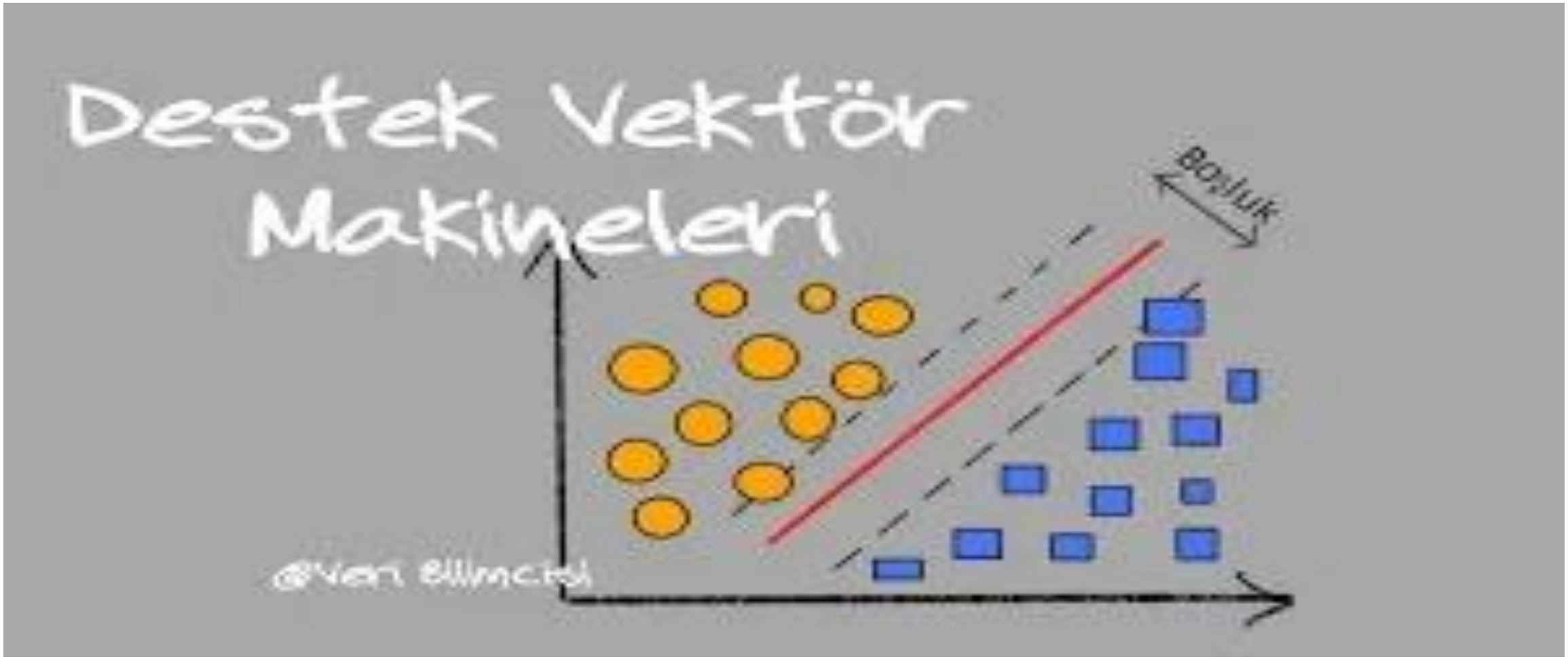
- Bir Karar Ağacı kullanmanın amacı, önceki verilerden (eğitim verileri) çıkarılan basit karar kurallarını öğrenerek hedef değişkenin sınıfını veya değerini tahmin etmek için kullanılabilecek bir eğitim modeli oluşturmaktır.
- Karar Ağaçlarında, bir kayıt için bir sınıf etiketini tahmin etmek amacıyla ağacın kökünden başlarız.
- Kök özniteliğin değerlerini kaydın özniteliğiyle karşılaştırırız. Karşılaştırma temelinde, bu değere karşılık gelen dalı takip eder ve bir sonraki düğüme atlarız

Karar Ağacı Türleri

•Karar ağacı türleri, sahip olduğumuz hedef değişkenin türüne dayanır.

- 2 tip olabilir:
- Kategorik Değişken Karar Ağacı:
- Kategorik bir hedef değişkene sahip olan Karar Ağacı, daha sonra Kategorik değişken karar ağacı olarak adlandırılır.
- Sürekli Değişken Karar Ağacı:
- Karar Ağacı sürekli bir hedef değişkene sahiptir, daha sonra Sürekli Değişken Karar Ağacı olarak adlandırılır.
- Kullanımı ve yorumlanması çok kolay olduğu için Makine Öğrenmesinde kullanılan en yaygın ve pratik yöntemlerden biridir.
- Karar ağaçları baş aşağıdır, bu da kökün en üstte olduğu anlamına gelir ve daha sonra bu kök çeşitli düğümlere bölünür.
- Karar ağaçları, bir grup if-else ifadesi olarak tanımlanabilir. Koşulun doğru olup olmadığını kontrol eder ve akışta, bu karara bağlı olarak bir sonraki düğüme geçer

Support Vector Machines



- Destek Vektör Makinesi (SVM) Sınıflandırma ve Regresyon problemleri için kullanılan en popüler Denetimli Öğrenme algoritmalarından biridir.

•Öncelik olarak Sınıflandırma problemleri için kullanılır.

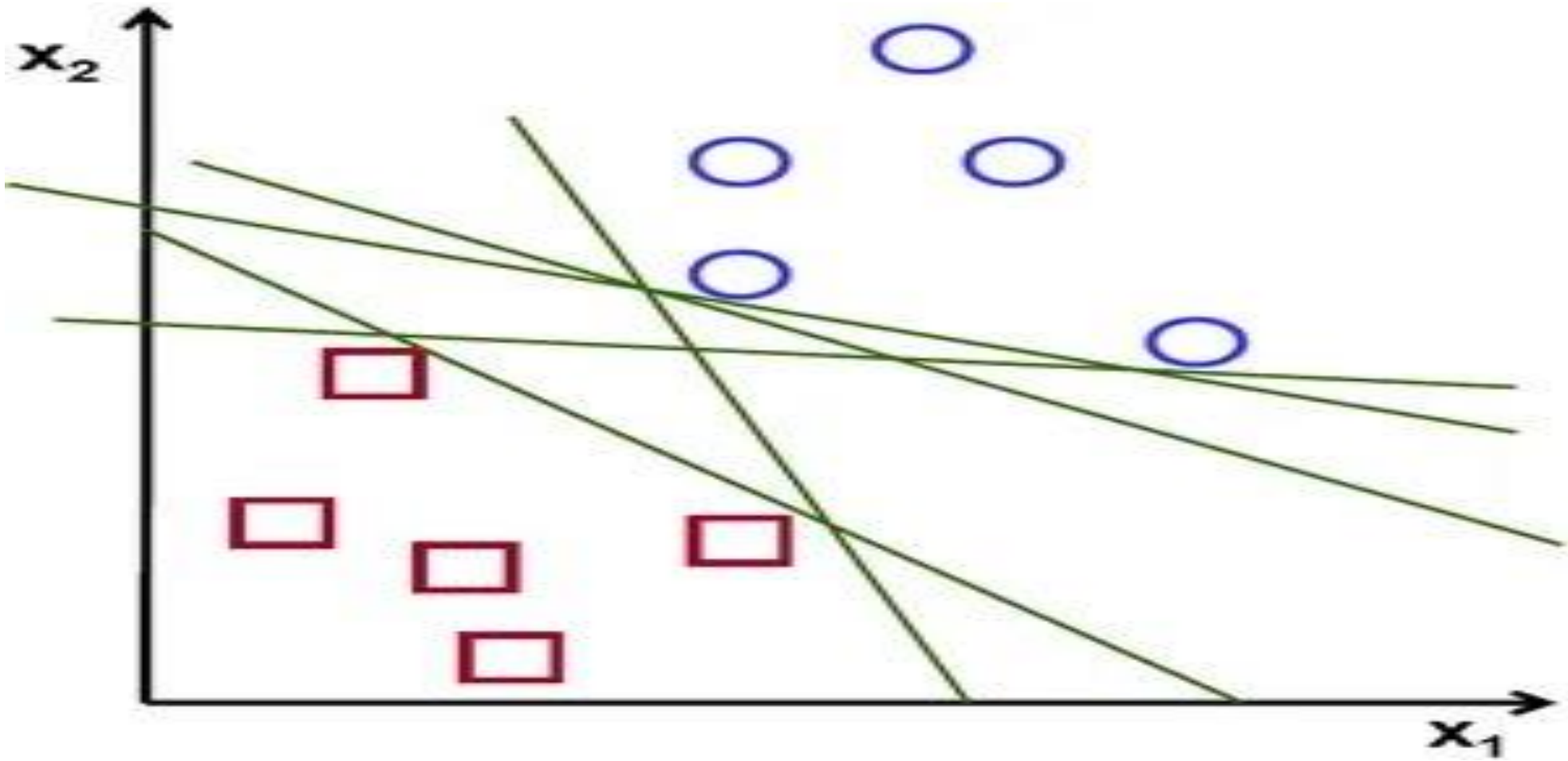
- SVM algoritmasının amacı, n boyutlu uzayı sınıflara ayırabilen en iyi çizgiyi veya karar sınırını oluşturmaktır, böylece gelecekte yeni veri noktasını kolayca doğru kategoriye koyabiliriz. Bu en iyi karar sınırına hiperdüzlem denir.

- SVM algoritmasındaki Hiperdüzlem ve Destek Vektörleri:

- Hiperdüzlem:

- Sınıfları n boyutlu uzayda ayırmak için birden fazla çizgi/karar sınırı olabilir, ancak veri noktalarını sınıflandırmaya yardımcı olan en iyi karar sınırını bulmamız gerekir. Bu en iyi sınır, SVMnin hiperdüzlemi olarak bilinir.

- Hiperdüzlemin boyutları, veri kümesinde bulunan özelliklere bağlıdır, 2 özellik varsa hiperdüzlem düz bir çizgi olacaktır. Ve eğer 3 özellik varsa, hiperdüzlem 2 boyutlu bir düzlem olacaktır.


x_2
x_1

•Öncelik olarak Sınıflandırma problemleri için kullanılır.

- SVM algoritmasının amacı, n boyutlu uzayı sınıflara ayırabilen en iyi çizgiyi veya karar sınırını oluşturmaktır, böylece gelecekte yeni veri noktasını kolayca doğru kategoriye koyabiliriz. Bu en iyi karar sınırına hiperdüzlem denir.
- SVM algoritmasındaki Hiperdüzlem ve Destek Vektörleri:
- Hiperdüzlem:
- Sınıfları n boyutlu uzayda ayırmak için birden fazla çizgi/karar sınırı olabilir, ancak veri noktalarını sınıflandırmaya yardımcı olan en iyi karar sınırını bulmamız gerekir. Bu en iyi sınır, SVMnin hiperdüzlemi olarak bilinir.
- Hiperdüzlemin boyutları, veri kümesinde bulunan özelliklere bağlıdır, 2 özellik varsa hiperdüzlem düz bir çizgi olacaktır. Ve eğer 3 özellik varsa, hiperdüzlem 2 boyutlu bir düzlem olacaktır.