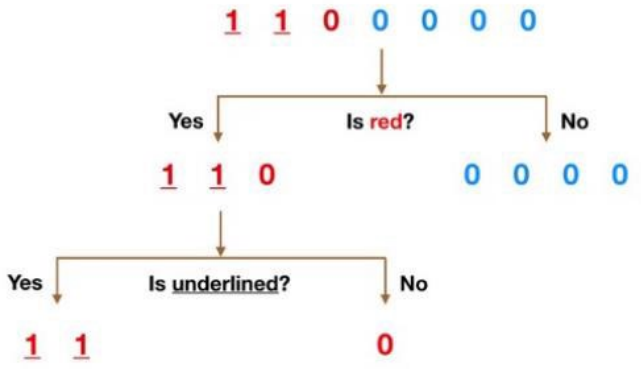
4.3.6. Rassal Orman Modeli Algoritması

Rastgele ormanlar, denetimli bir öğrenme algoritmasıdır. Hem sınıflandırma hem de regresyon için kullanılabilir.

Karar ağaçları:

Rastgele orman modelinin yapı taşları oldukları için karar ağaçlarının bilinmesi gerekmektedir. Oldukça sezgisel yaklaşımlar içerir. Çoğu insanın hayatlarının bir noktasında bilerek ya da bilmeyerek bir karar ağacı kullandığına bahse girerim.

Bir karar ağacının nasıl çalıştığını bir örnek üzerinden anlamak muhtemelen çok daha kolaydır.



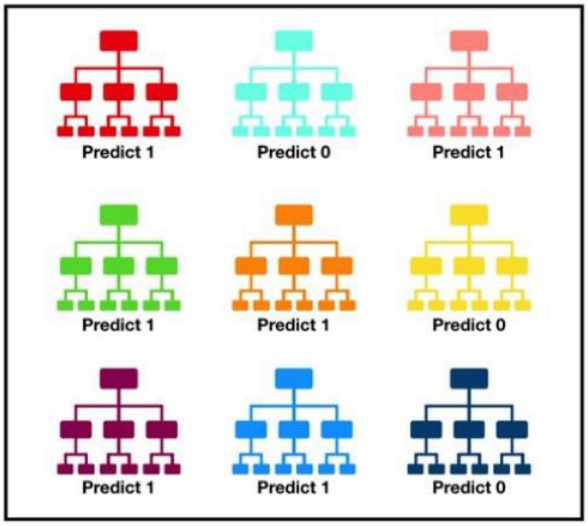
Veri setimizin soldaki şeklin üstündeki sayılardan oluştuğunu hayal edin. İki 1 ve beş 0'ımız var (1'ler ve 0'lar sınıflarımızdır) ve özelliklerini kullanarak sınıfları ayırmak istiyoruz. Özellikler renklidir (kırmızıya karşı mavi) ve gözlemin altı çizili olup olmadığıdır. Peki bunu nasıl yapabiliriz?

Renk, 0'lardan biri hariç tümü mavi olduğu için, ayrılması oldukça bariz bir özellik gibi görünüyor. Böylece "Kırmızı mı?" Sorusunu kullanabiliriz. ilk düğümümüzü ayırmak için. Bir ağaçtaki bir düğümü, yolun ikiye ayrıldığı nokta olarak düşünebilirsiniz - kriterleri karşılayan gözlemler Evet dalına ve Hayır dalına inmeyenler.

Hayır dalı (blues) artık 0'lardır, bu yüzden orada işimiz bitti, ancak Evet şubemiz yine de bölünebilir. Şimdi ikinci özelliği kullanıp "Altı çizili mi?" Diye sorabiliriz. ikinci bir bölme yapmak için.

Altı çizili iki 1, Evet alt dalına gider ve altı çizilmemiş 0, sağ alt daldan aşağı gider ve hepimiz işimiz biter. Karar ağacımız, verileri mükemmel bir şekilde bölmek için iki özelliği kullanabildi.

Rastgele Orman Sınıflandırıcısı:



Tally: Six 1s and Three 0s

Rastgele orman, adından da anlaşılacağı gibi, bir topluluk olarak çalışan çok sayıda bireysel karar ağacından oluşur. Rastgele ormandaki her bir ağaç bir sınıf tahmini verir ve en çok oyu alan sınıf, modelimizin öngörüsü haline gelir (yandaki şekle bakın).

Rastgele ormanın ardındaki temel kavram basit ama güçlü bir kavramdır - kalabalıkların bilgeliği. Veri biliminde konuşursak, rastgele orman modelinin bu kadar iyi çalışmasının nedeni şudur: Bir komite olarak faaliyet gösteren çok sayıda görece ilişkisiz model (ağaç), münferit kurucu modellerin herhangi birinden daha iyi performans gösterecektir. Modeller arasındaki düşük korelasyon anahtardır. Tıpkı düşük korelasyonlu yatırımların (hisse senetleri ve tahviller gibi) bir araya gelerek parçalarının toplamından daha büyük bir portföy oluşturması gibi, ilişkisiz modeller, bireysel tahminlerin herhangi birinden daha doğru olan topluluk tahminleri üretebilir. Bu harika etkinin nedeni, ağaçların birbirlerini kendi hatalarından korumalarıdır (sürekli aynı yönde hata yapmadıkları sürece). Bazı ağaçlar yanlış olabilirken, diğer birçok ağaç haklı olacaktır, bu nedenle bir grup olarak ağaçlar doğru yönde hareket edebilecektir. Bu nedenle, rastgele ormanın iyi performans göstermesi için ön koşullar şunlardır: Özelliklerimizde bazı gerçek sinyaller olması gerekir, böylece bu özellikler kullanılarak oluşturulan modeller rastgele tahmin etmekten daha iyi sonuç verir. Tek tek ağaçların yaptığı tahminlerin (ve dolayısıyla hataların) birbirleriyle düşük korelasyonlara sahip olması gerekir.

Özellik Rastgeleliği:

Normal bir karar ağacında, bir düğümü bölme zamanı geldiğinde, mümkün olan her özelliği göz önünde bulundururuz ve sağ düğümdekilerle sol düğümdeki gözlemler arasında en fazla ayrımı yaratanı seçeriz. Bunun aksine, rastgele bir ormandaki her ağaç yalnızca rastgele bir özellik alt kümesinden seçim yapabilir. Bu, modeldeki ağaçlar arasında daha fazla çeşitliliği zorlar ve sonuçta ağaçlarda daha düşük korelasyon ve daha fazla çeşitlilik ile sonuçlanır.