**IST4171-Veri Sınıflandırma Yöntemleri-Sorular-Grup 7**

**1.Soru : Decision Tree algoritmasıyla bir model oluşturuyorsunuz. Hangi durumda dallanma yapılacağını belirlemek için kullanılan bir ölçüt var mıdır? Eğer varsa, bu ölçüt nedir ve neye dayanarak karar verilir? Ayrıca, bu ölçütün modelin performansı üzerindeki etkisini nasıl değerlendirirsiniz?**

Decision Tree algoritması, bir özellikte dallanma yapmak için belirli bir ölçüt kullanır. Bu ölçüt, her bir özelliğin dallanma noktalarını belirlemede kullanılır ve genellikle veri setinin homojenliğini artırmak amacıyla seçilir. En yaygın olarak kullanılan ölçütler şunlardır:

Gini Impurity (Safiyet): Bu ölçüt, bir düğümün homojenliğini ölçer. Düğümdeki farklı sınıflardan örneklerin dağılımının ne kadar homojen veya heterojen olduğunu belirler. Gini impurity, bir düğümdeki sınıfların homojenliğinin azalması için hangi özelliklerin en iyi ayrımı sağladığını belirlemede kullanılır.

Entropy (Entropi): Entropi, bir düğümdeki belirsizliği ölçer. Bir özelliğin dallanma yapmasıyla, bir düğümdeki sınıfların ne kadar homojen hale geldiğini ve belirsizliğin azaldığını gösterir. Entropi, özelliklerin bir düğümü nasıl en iyi böleceğini belirlemek için kullanılır.

Information Gain (Bilgi Kazancı): Information gain, bir özelliğin dallanma yapmasının ardından belirli bir düğümdeki homojenlik veya entropideki azalma miktarını ölçer. Özellikler arasındaki en iyi ayrımı belirlemek için kullanılır.

Bu ölçütler, her bir özellik için dallanma noktalarını belirlemede kullanılır ve hangi özelliğin veri setini daha iyi böleceğini belirlemek için algoritma tarafından değerlendirilir.

**2.Soru: K-Nearest Neighbors (KNN) algoritmasında "K" değerini seçerken, bu değerin küçük veya büyük seçilmesinin model performansına etkileri nelerdir?**

K-Nearest Neighbors (KNN) algoritması, bir veri noktasını sınıflandırmak veya regresyon yapmak için çevresindeki "K" komşusunun etiketlerini veya değerlerini kullanır. "K" değeri, modelin performansını etkileyen önemli bir hiperparametredir. K değerini küçük veya büyük seçmenin model performansına olan etkileri şu şekilde özetlenebilir

Küçük K Değeri:

Daha Hassas Ama Gürültülü: Küçük bir K değeri, bir veri noktasını sınıflandırmak veya tahmin etmek için yalnızca birkaç komşuyu kullanır. Bu durum, modelin eğitim verilerine daha fazla uymasını sağlar, ancak gürültülü veri veya aykırı noktalara daha duyarlı hale gelir. Bu durumda, model eğitim verilerindeki küçük dalgalanmalara karşı aşırı uyum (overfitting) yapabilir.

Daha Az Hesaplama Gücü: Küçük K değerleri, tahmin yapmak için daha az komşu kullanması nedeniyle hesaplama açısından daha hızlıdır. Ancak, bu beraberinde daha fazla gürültüye ve veri setindeki dalgalanmalara duyarlılık getirir.

Büyük K Değeri:

Daha Pürüzsüz Tahminler: Büyük bir K değeri, daha fazla komşuyu dahil eder, bu da genellikle daha pürüzsüz tahminlere ve genelleme yeteneğine sahip bir model oluşturur. Bu durum, modelin genellikle daha iyi bir performans göstermesini sağlar, ancak eğitim verilerine aşırı uyma riskini azaltır.

Daha Fazla Hesaplama Gücü: Büyük K değerleri, daha fazla komşuyu dahil etmesi nedeniyle hesaplama açısından daha maliyetli olabilir. Ancak, bu genellikle modern bilgisayarların hesaplama gücüyle başa çıkılabilir.K değerini seçerken, problemin doğası, veri setinin büyüklüğü ve özellikleri göz önüne alınmalıdır. Genellikle, çeşitli K değerleriyle deneme yaparak doğru K değerini belirlemek için çapraz doğrulama gibi teknikler kullanılır.

**3.Soru: Bagging ve Boosting Yöntemleri arasındaki fark nedir?**

Bagging yönteminde tüm modeller paralel olarak inşa edilir, ve birbirlerinden bağımsızlardır. Modellerin çalışması tamamlandığında çıkan sonuçlar arasında en çok çıktısı verilen sınıf tahmin olarak alınır. Bu bir sınıflandırma algoritması ise en çok tahmin sonucuna ulaşılan sınıf çıktı olarak alınır, regresyon algoritması ise çıkan sonuçların ortalaması alınır. Boosting yönteminde ise öğrenme sonuçları bir sonraki öğrenici için kullanılır. Eğitim için ayrılan veri setinden bir temel öğrenici için rastgele seçim yapılır. Öğrenme gerçekleşir, model test edilir. Sonuçlardan yanlış sınıflandırılan örnekler belirlenir. Bunlar bir sonraki öğrenici için örnek seçiminde önceliklendirilir (seçilme olasılıkları arttırılır). Her seferinde bu bilgi güncellenir. Bagging yönteminde her bir örneğin seçilme şansı eşitken burada ağırlıklandırma vardır.