İSTANBUL TİCARET
ÜNİVERSİTESİ

Yapay Zeka Proje Ödevi

Öğrenci Ad Soyad : Türker Ataç
Öğrenci Numara : 200030467

Bölüm : Bilgisayar Programcılığı
Ders : Yapay Zeka Uygulamaları

Projenin Amacı :

Veri setimiz, sağlık istatistikleri ve AIDS teşhisi konmuş hastaların kategorik bilgilerini içermektedir. Bu projenin amacı, veri setinde bulunan hastaların yaş, kilo, tedavi türü gibi çeşitli özelliklerine dayanarak AIDS hastalığına yakalanma olasılıklarını tahmin etmektir. İlk olarak yapay zeka modeline veri sağlanarak eğitilir, daha sonra test verileri ile sonuçlar kontrol edilir.

Veri Seti Hakkında :

Bu veri seti, 1996 yılında yayınlanmıştır ve AIDS hastalığına sahip olan hastaların sağlık istatistiklerini ve kategorik bilgilerini içermektedir.
Veri setinde bulunan sütunlar ve anlamları şu şekildedir:

1. **time**: Hastanın hastalığa yakalanma veya takibin sonlanma süresi
2. **trt**: Tedavi göstergeci (0 = Sadece ZDV; 1 = ZDV + ddI, 2 = ZDV + Zal, 3 = Sadece ddI)
3. **age**: Başlangıç yaşları (yıl cinsinden)
4. **wtkg**: Başlangıçta ağırlık (kg)
5. **hemo**: Hemofili durumu (0= hayır, 1= evet)
6. **homo**: Homoseksüel aktivite durumu (0= hayır, 1= evet)
7. **drugs**: IV ilaç kullanımı geçmişi (0= hayır, 1= evet)
8. **Karnof**: Karnofsky skoru (0-100 arasında bir ölçekte)
9. **oprior**: 175 öncesi ZDV dışı antiretroviral tedavi (0= hayır, 1= evet)
10. **z30**: 175 öncesinde 30 gün boyunca ZDV kullanımı (0= hayır, 1= evet)
11. **preanti**: 175 öncesindeki anti-retroviral tedavi günleri
12. **race**: Irk (0= Beyaz, 1= Beyaz olmayan)
13. **gender**: Cinsiyet (0= Kadın, 1= Erkek)
14. **str2**: Antiretroviral geçmişi (0= tecrübesiz, 1= tecrübeli)
15. **strat**: Antiretroviral geçmiş stratifikasyonu (1='Antiretroviral Tecrübesiz',2='> 1 ancak <= 52 haftalık önceki antiretroviral tedavi',3='> 52 hafta')
16. **symptom**: Semptomatik gösterge (0= asemptomatik, 1= semptomatik)
17. **treat**: Tedavi göstergeci (0= Sadece ZDV, 1= Diğerleri)
18. **offtrt**: 96+/-5 hafta öncesinde tedaviden çıkma göstergesi (0= hayır,1= evet)
19. **cd40**: Başlangıçta CD4 seviyesi
20. **cd420**: 20+/-5 hafta sonundaki CD4 seviyesi
21. **cd80**: Başlangıçta CD8 seviyesi
22. **cd820**: 20+/-5 hafta sonundaki CD8 seviyesi
23. **infected**: AIDS hastalığına sahip olup olmadığı (0= Hayır, 1= Evet)

AIDS_Classification

time	trt	age	wtkg	hemo	homo	drugs	karnof	oprior	z30	preanti	race	gender	str2	strat	symptom	treat	offtrt	cd40	cd420	cd80	cd820	infected
948	2	48	89.8128	0	0	0	100	0	0	0	0	0	0	1	0	1	0	422	477	566	324	0
1002	3	61	49.4424	0	0	0	90	0	1	895	0	0	1	3	0	1	0	162	218	392	564	1
961	3	45	88.452	0	1	1	90	0	1	707	0	1	1	3	0	1	1	326	274	2063	1893	0
1166	3	47	85.2768	0	1	0	100	0	1	1399	0	1	1	3	0	1	0	287	394	1590	966	0
1090	0	43	66.6792	0	1	0	100	0	1	1352	0	1	1	3	0	0	0	504	353	870	782	0
1181	1	46	88.9056	0	1	1	100	0	1	1181	0	1	1	3	0	1	0	235	339	860	1060	0
794	0	31	73.0296	0	1	0	100	0	1	930	0	1	1	3	0	0	0	244	225	708	699	1
957	0	41	66.2256	0	1	1	100	0	1	1329	0	1	1	3	0	0	0	401	366	889	720	0
198	3	40	82.5552	0	1	0	90	0	1	1074	0	1	1	3	1	1	1	214	107	652	131	1

Sınıflandırma Algoritmaları Hakkında :

1. Gradient Boosting Classifier:
Gradient Boosting, zayıf öğrenicileri (genellikle karar ağaçları) bir araya getirerek güçlü bir model oluşturur. Her bir öğrenici, önceki öğrenicinin hatalarını düzeltmeye odaklanır. Gradient Boosting, karmaşık ilişkileri modellemek için etkili bir algoritmadır.

2. Logistic Regression:
Lojistik Regresyon, bir olayın olasılığını tahmin etmek için kullanılan istatistiksel bir yöntemdir. İkili sınıflandırma problemlerinde kullanılır ve sonuç olasılık değeri olarak verilir.

3. Naive Bayes:
Naive Bayes, Bayes teoremi temelinde çalışan bir olasılık tabanlı sınıflandırma algoritmasıdır. Özellikler arasındaki bağımsızlık varsayımı yapar ve bu nedenle "naive" (saf) olarak adlandırılır.

4. Random Forest:
Random Forest, birçok karar ağacını bir araya getirerek sınıflandırma veya regresyon problemleri için kullanılan bir ensemble öğrenme algoritmasıdır. Her bir karar ağacı, rastgele özelliklerle oluşturulur ve ardından en yaygın sınıf veya ortalama değer alınarak sonuçlar birleştirilir.

5. K-Nearest Neighbors:
K-Nearest Neighbors, bir örneğin sınıfını belirlemek için çevresindeki en yakın k komşuların çoğunluk sınıfını kullanır. Basit ancak etkili bir sınıflandırma algoritmasıdır.

6. Decision Tree:
Decision Tree, veri kümesindeki özelliklerin belirli koşullar altında nasıl ayrılacağını öğrenen bir ağaç yapısıdır. Karar ağaçları, veriye dayalı bir karar alma sürecini temsil eder.

7. Support Vector Machine (SVM):
Support Vector Machine, iki sınıf arasındaki ayrımı belirleyen optimal hiper düzlemi bulmaya çalışan bir sınıflandırma algoritmasıdır. Karmaşık veri kümeleri için etkili bir şekilde çalışabilir.

8. Linear Discriminant Analysis (LDA):
Linear Discriminant Analysis, sınıflar arasındaki lineer ayrımı bulmaya çalışan bir sınıflandırma algoritmasıdır. Verileri boyut azaltma ve sınıflandırma için kullanılır.

9. AdaBoost Classifier:
AdaBoost, zayıf öğrenicileri bir araya getirerek güçlü bir sınıflandırıcı oluşturan bir ensemble öğrenme algoritmasıdır. Her bir öğrenici, önceki öğrenicinin hatalarını düzeltmeye odaklanır.

10. MLP Classifier:
•MLP (Multi-Layer Perceptron) Classifier, yapay sinir ağlarının bir türüdür. Derin öğrenme alanında kullanılan çok katmanlı bir yapay sinir ağı modelidir.

* **Büyük ve karmaşık veri setlerinde** Gradient Boosting, Random Forest, Support Vector Machine gibi karmaşık ve güçlü modellerin **daha yüksek doğruluk oranlarına sahiptir.**

Performans Metrikleri Hakkında :

Metrikler, bir sınıflandırma modelinin performansını değerlendirmek için kullanılır ve her biri modelin farklı yönlerini ölçer. Örneğin, doğruluk genel performansı ölçerken, hassasiyet ve duyarlılık dengesiz sınıflandırma problemlerinde önemlidir. F1 skoru, hassasiyet ve duyarlılık arasında bir denge sağlar, ROC AUC skoru ise sınıflandırıcının genel yeteneğini ölçer.

1. Doğruluk (Accuracy):

Doğruluk, doğru olarak sınıflandırılan örneklerin oranını ölçer. Genellikle bir modelin performansını değerlendirmek için en yaygın kullanılan metriktir. Ancak, dengesiz sınıflandırma problemlerinde doğruluk tek başına yeterli olmayabilir.

2. Hassasiyet (Precision):

Hassasiyet, pozitif olarak tahmin edilen örneklerin gerçekten pozitif olma oranını ölçer. Yani, yanlış pozitiflerin sayısını azaltmayı hedefler. Hassasiyet, yanlış alarm oranını belirler.

3. Duyarlılık (Recall):

Duyarlılık, gerçekten pozitif olan örneklerin ne kadarının doğru bir şekilde tanındığını ölçer. Yani, yanlış negatiflerin sayısını azaltmayı hedefler. Duyarlılık, kaçırılan pozitiflerin oranını belirler.

4. F1 Skoru:

F1 Skoru, hassasiyet ve duyarlılığın harmonik ortalamasıdır. Dengesiz veri setlerinde kullanışlıdır, çünkü hem yanlış pozitiflerin hem de yanlış negatiflerin etkisini dikkate alır.

5. ROC AUC Skoru (Receiver Operating Characteristic - Area Under Curve):

1. ROC Eğrisi, duyarlılık ve özgüllük arasındaki ilişkiyi gösteren bir grafiktir. ROC Eğrisi altında kalan alan (AUC), sınıflandırıcının performansını ölçer. AUC ne kadar yüksek olursa, modelin performansı o kadar iyidir. 0 ile 1 arasında bir değere sahip olur, 1 mükemmel bir sınıflandırıcıyı, 0.5 ise rastgele bir tahminciyi temsil eder.

Confusion Matrix Hakkında :

Sınıflandırma modelinin performansını değerlendirmek için kullanılan bir tablodur. Bu matris, tahmin edilen sınıflar ile gerçek sınıflar arasındaki ilişkiyi gösterir. Temel bileşenleri True Positive (TP), True Negative (TN), False Positive (FP) ve False Negative (FN) olarak adlandırılır.

1. True Positive (TP):

Gerçek pozitiflerin sayısını temsil eder. Bu, modelin doğru bir şekilde pozitif olarak sınıflandırdığı örneklerin sayısıdır. Örneğin, hastaların AIDS hastalığına sahip olduğunu doğru bir şekilde tahmin ettiğimiz örnekler TP'yi oluşturur.

2. True Negative (TN):

Gerçek negatiflerin sayısını temsil eder. Bu, modelin doğru bir şekilde negatif olarak sınıflandırdığı örneklerin sayısıdır. Örneğin, hastaların AIDS hastalığına sahip olmadığını doğru bir şekilde tahmin ettiğimiz örnekler TN'yi oluşturur.

3. False Positive (FP):

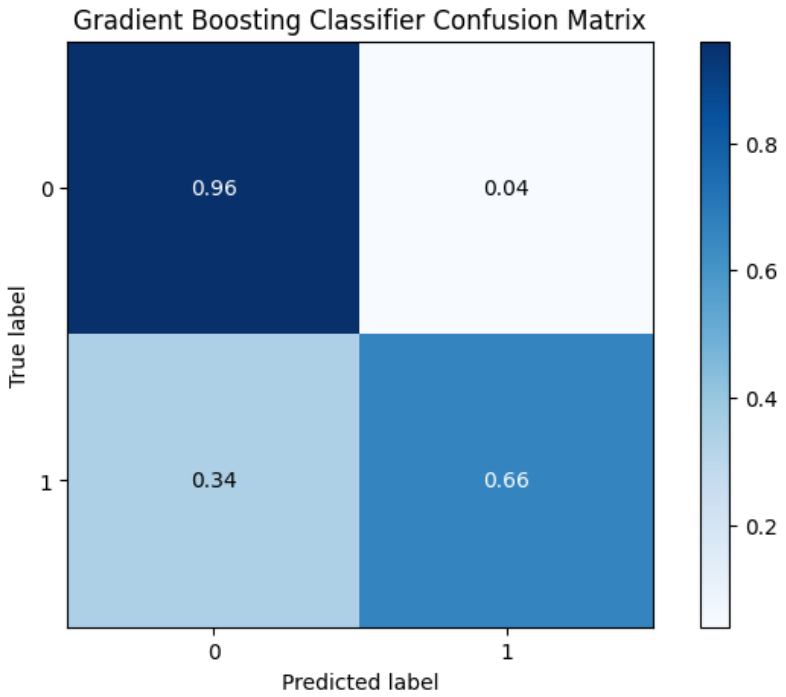
Yanlış pozitiflerin sayısını temsil eder. Modelin negatif olanları pozitif olarak yanlış bir şekilde sınıflandırdığı örneklerin sayısıdır. Örneğin, hastaların AIDS hastalığına sahip olmadığını düşündüğümüz ancak gerçekte hastalığa sahip olan örnekler FP'yi oluşturur. Bu, Tip I hata olarak da adlandırılır.

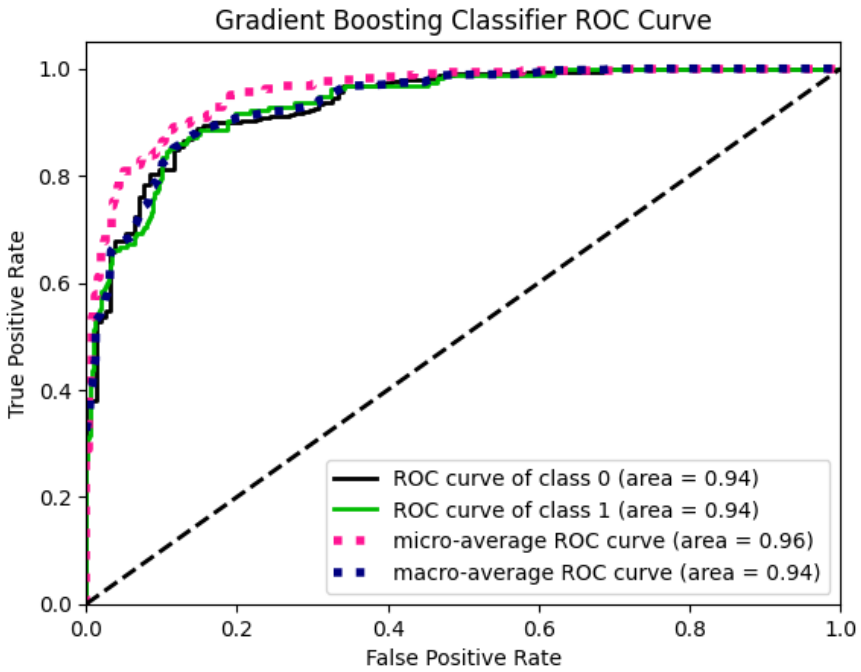
4. False Negative (FN):

Yanlış negatiflerin sayısını temsil eder. Modelin pozitif olanları negatif olarak yanlış bir şekilde sınıflandırdığı örneklerin sayısıdır. Örneğin, hastaların AIDS hastalığına sahip olduğunu düşündüğümüz ancak gerçekte hastalığa sahip olmayan örnekler FN'yi oluşturur. Bu, Tip II hata olarak da adlandırılır.

En İyi Sonucu Veren Algoritma ➔ Gradient Boosting Classifier

En iyi sonuç, Gradient Boosting Classifier algoritmasıyla elde edilmiştir. **Bu modelin Accuracy (Doğruluk) değeri 0.8863'dir.** Accuracy (Doğruluk), Modelin doğru tahmin ettiği toplam örnek sayısının, tüm örneklerin sayısına oranıdır. Yani, doğru sınıflandırılmış örneklerin oranını gösterir.





Feature Importance Hakkında :

Feature Importance değeri veya grafiği, bir modelin belirli özellikleri sınıflandırmadaki etkisini ölçer. Bu değerler, modele katkıda bulunan özelliklerin önem sırasını gösterir. Özelliklerin önemi, modelin tahmin yapma sürecinde hangi özelliklerin daha belirleyici olduğunu ve çıktıyı ne şekilde etkilediğini gösterir. Grafiğin yüksekliği, bir özelliğin önemini temsil eder. Yüksekliği ne kadar büyükse, o kadar önemli olduğunu gösterir.