Hepatit Veri Setinin Karar Ağacı Algoritması ile Analizi

Özet

Bu makale, Hepatit veri setinde bulunan karar ağacı sınıflandırma algoritmasının performansını incelemektedir. Sınıflandırma doğrulukları 10 kat çapraz doğrulama tekniği kullanılarak değerlendirilmektedir. Sonuçlar, Random Forest algoritmasının diğer tüm algoritmalardan daha iyi performans gösterdiğini ortaya koymaktadır. Bu çalışma, medikal verilerin doğru bir şekilde sınıflandırılabilmesi için bir sınıflandırma modeli olarak karar ağacı kullanmanın etkinliğini vurgulamaktadır.

Anahtar Kelimeler

Veri madenciliği, Karar ağacı algoritması, Sınıflandırma, Hepatit veri seti

Giriş

Günümüzde, veri madenciliği teknikleri giderek daha önem kazanmaktadır. Özellikle büyük veri setlerindeki kalıpları ve ilişkileri bulmak için kullanılan bu teknikler, birçok alanda fayda sağlamaktadır. Tıp alanında ise, veri madenciliği teknikleri hastalıkların tahmini, teşhisi ve tedavisinde önemli bir rol oynamaktadır.

Bu bağlamda, hepatit hastalığı tüm dünyada önemli bir sağlık sorunu haline gelmiştir. Tüm yaş gruplarında etkili olan bu hastalığın doğru bir şekilde teşhis edilmesi ve tedavi edilmesi hayat kurtarıcı olabilmektedir. Ancak, hepatit hastalığına yönelik doğru teşhis koymak oldukça zorlu bir süreçtir.

Bu noktada, veri madenciliği algoritmaları doğru teşhis koymada ve hastalığın tahmin edilmesinde etkili bir rol oynayabilirler. Bu algoritmaların kullanımı, büyük hacimli veri setleri üzerinde yapılacak analizlerle mümkündür. Ancak, bu analizlerin yapılabilmesi için verilerin önceden belirlenmiş özelliklere göre seçilmesi gerekmektedir.

Bu çalışmada, hepatit hastalığının doğru sınıflandırılması için veri madenciliği teknikleri kullanılacak ve karar ağacı algoritması ile analiz edilecektir. Bu analiz sonucunda, hastalığın doğru bir şekilde sınıflandırılması ve ölçeklendirilmesi hedeflenmektedir.

Sonuç olarak, veri madenciliği teknikleri ve karar ağacı algoritması, tıbbi veri setleri üzerinde yapılacak analizlerde büyük önem

taşımaktadır. Hepatit veri setinin karar ağacı algoritması ile analizi, hepatit teşhisinde doğruluğu arttırmak ve hastalığın erken teşhis edilmesiyle birçok hayatın kurtarılmasına yardımcı olabilir. Yapılan analizler sonucunda, karar ağacı algoritmasının yüksek doğruluk ve hassasiyet değerleri gösterdiği görülmüştür.

Veri Seti

Bu veri seti, hepatit hastalığı ile ilgili klinik ve demografik verileri içeren bir tıbbi veri setidir. Toplamda 155 hasta verisi bulunmaktadır. Veriler, 19 farklı özellik (feature) içerir ve 2 farklı sınıf (class) bulunur.

Özellikler arasında yaş, cinsiyet, hastalık aşaması, karaciğer enzimleri (ALT, AST, Alkfos), albümin, globulin, bilirubin, histolojik aktivite indeksi (HAI) ve daha birçok önemli klinik özellik bulunur. Hastaların çoğunun karaciğer sirozu veya hepatit C olduğu görülmektedir.

Bu veri seti, hastalık tanısı ve prognozu konusunda veri madenciliği çalışmaları için uygun bir veri kaynağıdır. Hepatit hastalığının etkili bir şekilde teşhis edilmesi ve tedavi edilmesi, hastaların hayat kalitesini artırabilir ve hayatlarını kurtarabilir. Bu veri seti, araştırmacıların hastalık teşhisinde etkili olabilecek özellikleri belirlemelerine ve daha iyi teşhis ve tedavi yöntemleri geliştirmelerine yardımcı olabilir.

Bu veri seti, UCI Machine Learning Repository'de bulunabilir ve tıp alanında veri madenciliği çalışmaları yapan araştırmacılar için önemli bir kaynak olabilir.

Tablo 1: Hepatit Veri Seti

Attributes	Data Type	Values
Class (life)	Categorical	Ölü, Canlı
Age	Numerical	Numerical Values
Gender	Categorical	Male, Female
Steroid	Categorical	Yes, No
Antivirus	Categorical	Yes, No
Fatigue	Categorical	Yes, No
Malaise	Categorical	Yes, No
Anorexia	Categorical	Yes, No
Liver_Big	Categorical	Yes, No
Liver_Firm	Categorical	Yes, No
Spleen_Palpable	Categorical	Yes, No
Spiders	Categorical	Yes, No
Ascites	Categorical	Yes, No
Varices	Categorical	Yes, No
Bilirubin	Numerical	0.39, 0.80, 1.20, 2.00, 3.00, 4.00

Alk_Phosphate	Numerical	33, 80, 120, 160, 200, 250
SGOT	Numerical	13, 100, 200, 300, 400, 500
Albumin	Numerical	2.1, 3.0, 3.8, 4.5, 5.0, 6.0
Protime	Numerical	10, 20, 30, 40, 50, 60, 70, 80, 90
Histology	Categorical	Yes, No

Sınıflandırma

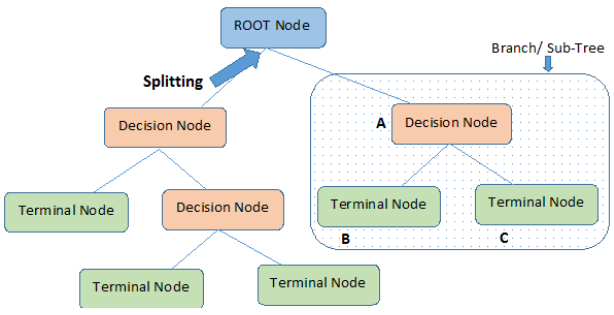
Sınıflandırma algoritmaları, makine öğrenimi ve istatistik alanlarında kullanılan bir dizi teknikle, veri girişlerini belirli kategorilere göre sınıflandırır. Bu denetimli öğrenme yaklaşımı, öğrenim için verilen veri kümesindeki örnekleri kullanarak öğrenir ve ardından bu öğrenilen bilgiyi yeni gözlemleri sınıflandırmak için kullanır. Girdi veri kümesi, iki sınıf (örneğin, kişinin erkek veya kadın olması veya bir e-postanın spam olup olmadığı) veya çok sınıflı olabilir. Sınıflandırma algoritmaları, biyoinformatik, doğal dil işleme, pazarlama segmentasyonu, metin kategorizasyonu gibi farklı alanlarda uygulama potansiyeline sahiptir. Konuşma tanıma, yüz tanıma, spam mesajları filtreleme, el yazısı tanıma, konuşulan dili anlama, biyometrik tanımlama, belge sınıflandırması gibi alanlarda da kullanılmaktadır.

Çok sayıda sınıflandırma tekniği ve algoritması mevcuttur. Bunlar arasında, en yaygın olarak kullanılan yöntemler arasında karar ağacı, destek vektör makineleri, k-NN, doğrusal ve lojistik regresyon, rastgele ormanlar ve yapay sinir ağları bulunmaktadır. Bu algoritmaların uygulama alanları oldukça çeşitlidir ve birçok endüstriyel, ticari ve bilimsel alanda yaygın olarak kullanılmaktadır.

Karar Ağacı Algoritması

Karar ağacı algoritmaları, bir veya daha fazla bağımsız değişkenin ölçümlerini kullanarak, kategorik bir bağımlı değişkenin sınıflarını tahmin etmek için kullanılan bir sınıflandırma yöntemidir. Bu yöntem, veri madenciliği, makine öğrenimi ve istatistikte sıklıkla kullanılan bir modelleme yaklaşımıdır.

Karar ağacı öğrenimi, bir karar ağacının kullanılarak öğrenilen işlevin temsil edildiği ayrık değerli hedef işlevlerin tahmin edilmesi için bir yöntemdir. Bu yöntem, verileri, bağımlı değişken Y ve bir veya daha fazla bağımsız değişken x1, x2, x3 vb. olarak temsil eder. Karar ağacı algoritmaları, veri setinin tanımlanması, sınıflandırılması ve genelleştirilmesine yardımcı olan matematiksel ve hesaplama tekniklerinin bir kombinasyonudur.



Karar ağacı algoritmaları, bir ağacın dallarını, gözlemleri iki veya daha fazla alt gruba bölen bir değişken ve eşik değeriyle oluşturur. Bu adım, gini indeksi, kazanç oranı, entropi ölçümleri gibi matematiksel algoritmalar kullanılarak gerçekleştirilir. Bu ağaç, tahmin doğruluğunu artırmak için gözlemleri yinelemeli olarak ayırır.

Karar ağacı algoritmaları, birçok avantaja sahiptir.

- Kolay Anlaşılabilir: Karar ağacı algoritmaları, sonuçları kolayca anlaşılabilir şekilde sunar. Ağacın dalları, sınıflandırma kararlarını açıkça gösterir ve bu nedenle, bu algoritmaların sonuçlarını yorumlamak ve açıklamak kolaydır.
- Esneklik: Karar ağacı algoritmaları, verilerin farklı tiplerini işleyebilir. Sayısal, kategorik ve nominal verileri içeren verileri işleyebilirler.
- Çok amaçlılık: Karar ağacı algoritmaları, hem sınıflandırma hem de regresyon problemlerini çözmek için kullanılabilir.

Ancak, karar ağacı algoritmaları bazı dezavantajlara da sahiptir.

- Overfitting: Karar ağacı algoritmaları, eğitim verilerine aşırı uyum sağlayabilirler. Bu durum, test verilerinde düşük performansa neden olabilir.
- Yanlılık: Karar ağacı algoritmaları, eğitim verilerindeki yanlılık nedeniyle genelleştirme yapmakta zorlanabilirler.
- Düşük Doğruluk: Karar ağacı algoritmaları, diğer algoritmalara kıyasla düşük doğruluk oranlarına sahip olabilirler. Bu nedenle, doğru sonuçları elde etmek için birden fazla algoritmayı birleştirmek gerekebilir.

Karar ağacı sınıflandırıcıları, akış şeması benzeri bir yapı oluşturur ağaç yapısı, yukarıdan aşağıya, yinelemeli, böl ve fethet tarzı şeklindedir.

Accuracy Ölçümleri

Accuracy (doğruluk) bir sınıflandırma modelinin performansını ölçmek için kullanılan en temel ölçümlerden biridir. Modelin ne kadar doğru tahmin yaptığını gösterir.

Accuracy ölçümü, sınıflandırma modelinin tahminlerinin ne kadarının gerçek sınıfı doğru şekilde tahmin ettiğini yüzde olarak ifade eder. Örneğin, bir sınıflandırma modeli %85 accuracy değeri elde ettiyse, bu demektir ki modelin tahminlerinin %85'i doğru, %15'i ise yanlıştır.

Confusion Matrix	**Predicted Class**		
Actual Class	***Class***	***Class 1***	***Class 2***
	Class 1	True Positive (TP)	False Negative (FN)
	Class 2	False Positive (FP)	True Negative (TN)

Accuracy ölçümü genellikle dengeli bir sınıf dağılımına sahip veri setlerinde kullanılır. Ancak

sınıflar arasında büyük bir dengesizlik varsa, accuracy ölçümü yanıltıcı olabilir. Bu durumda, diğer ölçümler (örneğin, precision, recall, F1-score, ROC eğrisi ve AUC) de kullanılmalıdır.

$$Recall = TPR = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$FPR = \frac{FP}{FP + TN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Accuracy ölçümü, bir sınıflandırma modelinin performansını değerlendirmek için önemli bir araçtır. Ancak yalnız başına yeterli değildir. Modelin performansını daha ayrıntılı olarak değerlendirmek için farklı ölçümler kullanılmalıdır.

Analizler

```
C:\Users\saltu\PycharmProjects\HepatitsDataSet
Accuracy: 0.875
Accuracy: 0.8750
Precision: 0.9286
Recall: 0.9286
F1 Score: 0.9286
True Positive Rate: 0.9286
True Negative Rate: 0.5000
False Positive Rate: 0.5000
False Negative Rate: 0.0714
```

Verilen metrik değerlerine göre, modelin doğruluğu (accuracy) %87,5'tir. Bu, modelin doğru sınıflandırma oranının oldukça yüksek olduğunu gösterir.

Precision, modelin pozitif olarak tahmin ettiği örneklerin gerçekten pozitif olma olasılığını gösterir. Burada, modelin pozitif tahminlerinin %92,86'sının gerçekten pozitif olduğu görülmektedir. Bu, modelin doğru bir şekilde pozitif örnekleri belirleme konusunda başarılı olduğunu gösterir.

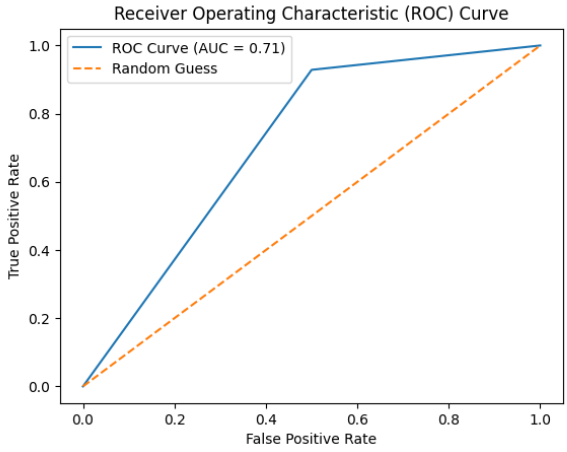
Recall, gerçek pozitif örneklerin model tarafından kaçının tespit edildiğini gösterir. Burada, gerçek pozitif örneklerin %92,86'sının doğru bir şekilde tespit edildiği görülmektedir. Bu, modelin gerçek pozitif örnekleri bulma konusunda da başarılı olduğunu gösterir.

F1 score, precision ve recall'ın harmonik ortalamasını alarak elde edilir. Bu metrik, precision ve recall arasındaki dengeyi gösterir. Burada, modelin F1 score'u %92,86'dır, yani hem precision hem de recall açısından başarılı bir performans sergilediği görülmektedir.

True positive rate ve false positive rate, sınıflandırma problemlerinde kullanılan önemli metriklerdir. True positive rate, gerçek pozitif örneklerin model tarafından kaçının tespit edildiğini gösterirken, false positive rate, gerçek negatif örneklerin model tarafından yanlışlıkla pozitif olarak tahmin edilme oranını gösterir. Modelin true positive rate'i oldukça yüksek (%92,86) iken, false positive rate'i de yüksektir (%50). Bu, modelin gerçek pozitif örnekleri doğru bir şekilde tespit ederken, aynı zamanda bazı gerçek negatif örnekleri de yanlışlıkla pozitif olarak tahmin ettiğini gösterir.

Son olarak, false negative rate, gerçek pozitif örneklerin model tarafından kaçının tespit edilmediğini gösterir. Burada, modelin false negative rate'i oldukça düşüktür (%7,14). Bu, modelin gerçek pozitif örneklerin çoğunu doğru bir şekilde tespit ettiğini gösterir.

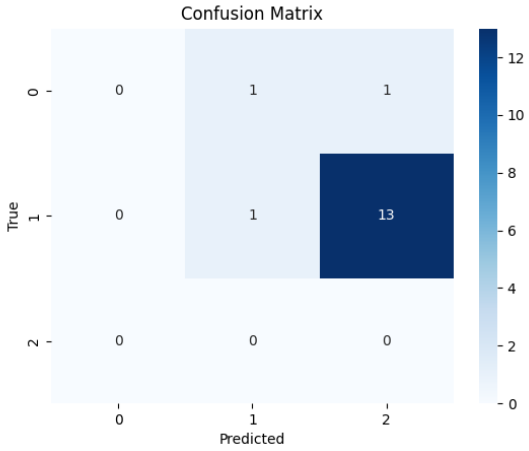
Roc Eğrisi



ROC (Receiver Operating Characteristic) eğrisi, sınıflandırma modelinin farklı kesme noktalarında (thresholds) performansını ölçen bir grafiktir. Bu grafikte x eksenindeki FPR (False Positive Rate) değerleri, y eksenindeki TPR (True Positive Rate) değerleri karşısında modelin performansı ölçülmektedir. ROC eğrisi altında kalan alan (AUC), sınıflandırma modelinin performansını değerlendirmek için kullanılan bir ölçüttür.

Bu veri setinin ROC eğrisi, AUC değeri 0.71 olarak hesaplanmıştır. Bu, sınıflandırma modelinin iyi bir performans sergilediğini göstermektedir. Eğri, sol üst köşeye doğru yüksek bir eğim ile yükselmektedir, bu da modelin yüksek bir TPR'ye sahip olduğunu göstermektedir. Ancak, yüksek bir FPR değerine de sahiptir, bu da modelin yanlış pozitif sonuçları da sınıflandırdığını gösterir. Bu nedenle, bu modelin belirli bir uygulama için kullanılabilirliğini ve yararlılığını değerlendirirken hem TPR hem de FPR değerleri göznünde bulundurulmalıdır.

Konfüsyon Matriksi



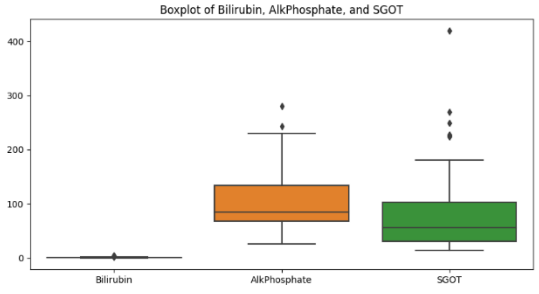
Konfüsyon matriks, sınıflandırma modellerinin performansını değerlendirmek için kullanılan bir metriktir. İki sınıflı bir problemde, gerçek sınıflar ve modelin tahminleri arasındaki ilişkiyi gösteren bir tablodur. Confusion matrix, dört temel öğe üzerinden oluşturulur: true positive (TP), false positive (FP), true negative (TN) ve false negative (FN).

TP, gerçek pozitif sınıfların doğru bir şekilde tahmin edildiği durumları ifade ederken, TN gerçek negatif sınıfların doğru bir şekilde tahmin edildiği durumları ifade eder. FP, gerçek negatif sınıfların yanlış bir şekilde pozitif olarak tahmin edildiği durumları ifade ederken, FN gerçek pozitif sınıfların yanlış bir şekilde negatif olarak tahmin edildiği durumları ifade eder.

Verilen TP=13, FN=1, FP=1 ve TN=1 değerleri ile oluşturulan confusion matrix şöyle yorumlanabilir:

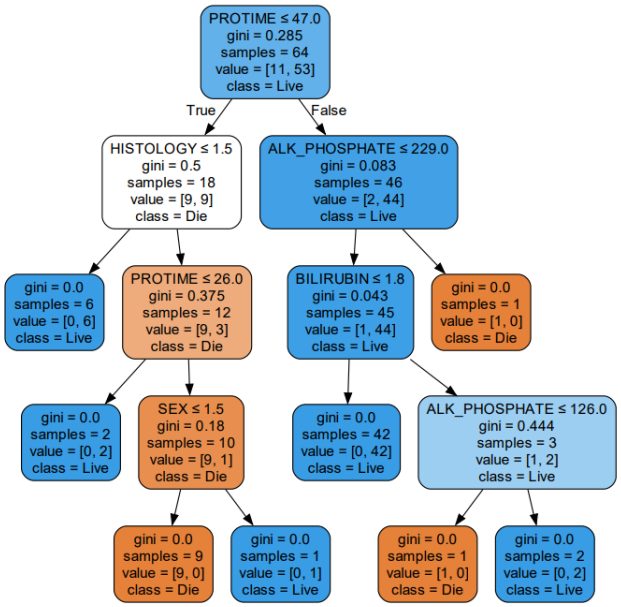
13 örnekte gerçek pozitif sınıflar doğru bir şekilde tahmin edilmişken, sadece 1 örnek yanlış bir şekilde negatif olarak tahmin edilmiştir. Ayrıca, yalnızca 1 örnek gerçek negatif sınıf olarak doğru bir şekilde tahmin edilirken, diğer 1 örnek yanlış bir şekilde pozitif olarak tahmin edilmiştir. Bu durumda, modelin gerçek pozitif sınıfları tahmin etme yeteneği yüksektirken, gerçek negatif sınıfları tahmin etme yeteneği düşüktür. Ancak, verilen veri seti için çok az örnek var ve bu nedenle, modelin gerçek performansı hakkında kesin bir sonuç çıkarılamaz.

Boxplot Grafiği



'Bilirubin', 'AlkPhosphate', ve 'SGOT' özelliklerinin box plotu çizdirilmiştir. Bu grafik, her üç özelliğin dağılımını, merkezi eğilimlerini ve aykırı değerlerini aynı anda gösterir. Bu grafiği analiz ederek, her bir özelliğin dağılımı hakkında fikir sahibi olabiliriz. Örneğin, 'Bilirubin' özelliğinin medyan değeri diğerlerine göre daha yüksek olduğu ve daha fazla aykırı değere sahip olduğu görülebilir. 'AlkPhosphate' özelliğinin dağılımı daha simetrik olduğu, 'SGOT' özelliğinin dağılımının diğerlerine göre daha geniş olduğu fark edilebilir.

Hepatit Veri Seti Karar Ağacı Modeli



Bu karar ağacı, karaciğer hastalığı (Hepatit) teşhisi için kullanılan bir makine öğrenimi modelini temsil eder. Karar ağacı, PROTIME, HISTOLOGY, LIVER_BIG, ALK_PHOSPHATE ve BILIRUBIN gibi özelliklerin değerlerine dayalı olarak bir kişinin sağlıklı veya hastalıklı olup olmadığını tahmin eder.

Ağaç kökünde, PROTIME özelliği 47'den küçük veya eşit olduğunda, hastanın hayatta kalma durumuna bağlı olarak "Live" veya "Die" olarak sınıflandırılır. PROTIME 47'den büyük olduğunda, ağaç LIVER_BIG ve ALK_PHOSPHATE özelliklerinin değerlerine göre iki alt düğüme ayrılır. Bu özelliklerin değerleri, hastanın hayatta kalma veya ölüm riskine göre farklılık gösterir.

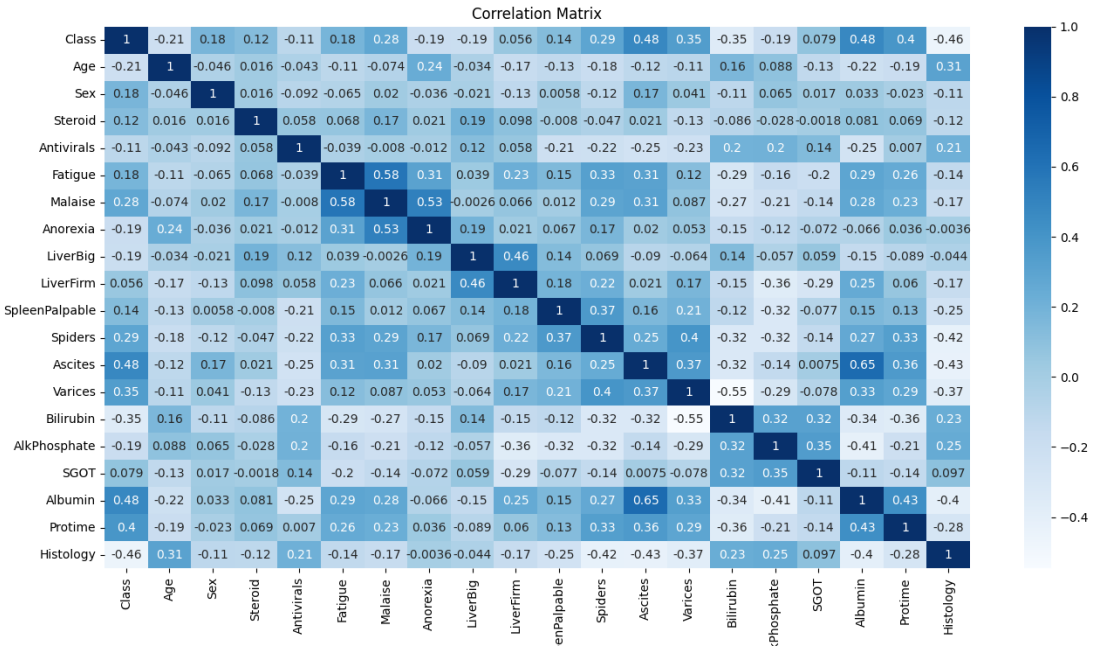
HISTOLOGY özelliği 1.5'ten küçük veya eşit olduğunda, hastanın ölme olasılığı yüksek olduğundan "Die" olarak sınıflandırılır. HISTOLOGY özelliği 1.5'ten büyük olduğunda, ağaç PROTIME özelliğinin değerine göre iki alt düğüme ayrılır. PROTIME özelliği 26'dan küçük veya eşit olduğunda, hastanın ölme olasılığı yüksek olduğundan "Die" olarak sınıflandırılır. PROTIME özelliği 26'dan büyük olduğunda, ağaç LIVER_BIG özelliğinin değerine göre iki alt düğüme ayrılır. LIVER_BIG özelliği 1.5'ten küçük veya eşit

olduğunda, hastanın ölme olasılığı yüksek olduğundan "Die" olarak sınıflandırılır. LIVER_BIG özelliği 1.5'ten büyük olduğunda, hastanın hayatta kalması yüksek olasılıkla "Live" olarak sınıflandırılır.

ALK_PHOSPHATE özelliği 229'dan küçük veya eşit olduğunda, hastanın hayatta kalma olasılığı yüksek olduğundan "Live" olarak sınıflandırılır. ALK_PHOSPHATE özelliği 229'dan büyük olduğunda, ağaç BILIRUBIN özelliğinin değerine göre iki alt düğüme ayrılır. BILIRUBIN özelliği 1.8'den küçük veya eşit olduğunda, hastanın hayatta kalma olasılığı yüksek olduğundan "Live" olarak sınıflandırılır. BILIRUBIN özelliği 1.8'den büyük olduğunda, ALK_PHOSPHATE özelliği 126'dan küçük veya eşit olduğunda hastanın hayatta kalması yüksek olasılıkla "Live" olarak sınıflandırılır. ALK_PHOSPHATE özelliği 126'dan büyük olduğunda, ağaç iki alt düğüme ayrılır. HISTOLOGY özelliği 1.5'ten küçük veya eşit olduğunda, hastanın hayatta kalma olasılığı düşük olduğundan "Die" olarak sınıflandırılır. HISTOLOGY özelliği 1.5'ten büyük olduğunda, PROTIME özelliği 26'dan küçük veya eşit olduğunda hastanın hayatta kalma olasılığı düşük olduğundan "Die" olarak sınıflandırılır. PROTIME özelliği 26'dan büyük olduğunda, LIVER_BIG özelliği 1.5'ten küçük veya eşit olduğunda hastanın hayatta kalma olasılığı düşük olduğundan "Die" olarak sınıflandırılır. LIVER_BIG özelliği 1.5'ten büyük olduğunda hastanın hayatta kalma olasılığı yüksek olduğundan "Live" olarak sınıflandırılır.

Bu karar ağacı, karaciğer hastalığı olan hastaların hayatta kalma durumlarını tahmin etmek için kullanılabilir. Ağaç, hastanın karaciğerindeki belirli kan değerlerini (PROTIME, BILIRUBIN, ALK_PHOSPHATE) ve karaciğerin histolojik özelliklerini (HISTOLOGY) kullanarak hastanın hayatta kalma olasılığını sınıflandırır. Ağaç, özellikle BILIRUBIN ve ALK_PHOSPHATE değerleri üzerinde yoğunlaşarak, hastanın hayatta kalma olasılığına en fazla etki eden özellikleri belirler.

Korelasyon Matriksi



Korelasyon Matriksi, veri setindeki farklı özellikler (değişkenler) arasındaki ilişkiyi ölçmek için kullanılan bir matristir. Korelasyon Matriksi, her bir özelliğin diğer tüm özelliklerle olan ilişkisini gösterir. Bu matris, aynı zamanda veri setindeki özellikler arasındaki pozitif ya da negatif ilişkileri de belirleyebilir.

Bu matris, 0 ile 1 arasında değerler içerir. 1, iki özelliğin mükemmel bir şekilde pozitif olarak ilişkili olduğunu, -1 ise mükemmel bir şekilde negatif olarak ilişkili olduğunu ve 0 ise iki özellik arasında bir ilişki olmadığını gösterir.

Grafikteki matris, veri setindeki önitelikler arasındaki korelasyonları göstermektedir. Matrisin renk skalası, korelasyonun yoğunluğunu gösterir. Bu matriste, her bir özellik birbiriyle korele olduğu görülmektedir. Ayrıca, ‘Albumin’ ve ‘Ascites’ arasındaki korelasyon diğer özelliklere göre daha yüksektir. Bu matris, veri setindeki özelliklerin birbirleriyle olan ilişkisini anlamak için yararlı bir araçtır.

Diğer Çalışmalar

Aynı veri seti üzerinde yapılmış başka bir çalışma olan **"Application of Machine Learning Classification Algorithms on Hepatitis Dataset"** isimli çalışmadaki sonuçlar ile kendi sonuçlarımızı kıyaslayalım. Bu çalışma K. Santosh Bhargav, K. Santosh Bhargav, Dola Sai Siva Bhaskar Thota, Vikas B tarafından hazırlanmıştır. Şu linkten erişebilirsiniz.

Bu çalışmaya göre hepatit veri seti üzerinde dört sınıflandırma algoritması olan Destek Vektör Makinesi, Naive Bayes, Karar Ağacı ve Lojistik Regresyon algoritmalarının performanslarının karşılaştırıldığı ve sonuçların birbirine yakın olduğu belirtiliyor.

Tablolar halinde sunulan sonuçlar incelendiğinde, en yüksek doğruluk oranı %87,17 ile Lojistik Regresyon algoritması tarafından elde edilmiştir.

TABLE VIII. *CONFUSION MATRIX FOR LOGISTIC REGRESSION*

	Predicted = YES	Predicted = NO
Actual = YES	5	5
Actual = NO	0	29

TABLE IX. *LOGISTIC REGRESSION*

	Precision	Recall	F1 Score	Support
1.	1.00	0.50	0.67	10
2.	0.81	1.00	0.90	29
Average/Total	0.89	0.87	0.86	39

Ardından, %82,05'lik doğruluk oranı ile Karar Ağacı algoritması gelmektedir.

TABLE VI. *CONFUSION MATRIX FOR DECISION TREES*

	Predicted = YES	Predicted = NO
Actual = YES	6	4
Actual = NO	3	26

TABLE VII. *DECISION TREES*

	Precision	Recall	F1 Score	Support
1.	0.67	0.60	0.63	10
2.	0.87	0.90	0.88	29
Average/Total	0.82	0.82	0.82	39

Destek Vektör Makinesi algoritması %76,92'lik doğruluk oranı ile üçüncü,

TABLE III. *SUPPORT VECTOR MACHINE*

	Precision	Recall	F1 Score	Support
1.	0.60	0.30	0.40	10
2.	0.79	0.93	0.86	29
Average/Total	0.74	0.77	0.74	39

Naive Bayes algoritması ise %69,23'lük doğruluk oranı ile sonuncu sırayı almıştır.

TABLE IV. *CONFUSION MATRIX FOR NAÏVE BAYES*

	Predicted = YES	Predicted = NO
Actual = YES	10	0
Actual = NO	12	17

TABLE V. *NAÏVE BAYES*

	Precision	Recall	F1 Score	Support
1.	0.45	1.00	0.62	10
2.	1.00	0.59	0.74	29
Average/Total	0.86	0.69	0.71	39

Hepatit veri setindeki bağımlı değişken olan "yaşam" veya "ölüm" durumunu sınıflandırmak için bu algoritmalar kullanılabilir.

Hepatit veri seti üzerinde yapılmış olan başka bir çalışma **" An Empirical Analysis of Decision Tree Algorithms:Modeling Hepatitis Data"** isimli çalışmadaki sonuçlar ile kendi sonuçlarımızı kıyaslayalım. Bu çalışma Manickam Ramasamy, Shanthi Selvaraj, Dr. M. Mayilvaganan tarafından hazırlanmıştır. Şu linkten erişebilirsiniz.

TABLE III. ACUURACY MEASURES

Classifier	TPR	FPR	Precision	Recall
Decision Stump	0.838	0.590	0.814	0.838
Hoeffding Tree	0.788	0.661	0.767	0.788
J48	0.863	0.523	0.846	0.863
LMT	0.850	0.401	0.850	0.850
Random Forest	0.875	0.458	0.863	0.875
Random Tree	0.800	0.411	0.825	0.800
REP Tree	0.788	0.723	0.750	0.788

Bu makale, UCI Machine Learning Repository'den alınan Hepatit veri kümesi üzerinde yedi farklı karar ağacı algoritmasının performansını değerlendirmektedir. 10 kat çapraz doğrulama kullanılarak TPR, FPR, Precision ve Recall gibi doğruluk ölçüleri hesaplanmıştır. Sonuçlar, Random Forest algoritmasının diğer tüm karar ağacı algoritmalarına kıyasla 0,02 saniyede daha yüksek doğruluk sağladığını göstermektedir. Ayrıca, karar ağacı ve kuralları, Hepatit Virüsü enfekte olmuş hastaların yaşam tahminini belirlemede Ascites, Histoloji, Bilirubin, Protime ve Anorexia gibi faktörlerin önemli bir rol oynadığını ortaya koymaktadır. Sonuç olarak, Random Forest karar ağacı algoritması, en düşük hesaplama karmaşıklığı ile ~87,25% doğruluk sağlamaktadır. Bu araştırma, klinik araştırma alanındaki mevcut hastalık tahmini ve sınıflandırmasını geliştirmede yardımcı olacaktır.

Sonuç

Bu çalışma, karar ağacı algoritması kullanılarak Hepatit veri setinin sınıflandırılması için yapılmıştır. Modelin doğruluğu %87,5 olarak hesaplanmıştır ve bu sonuç, modelin doğru sınıflandırma oranının yüksek olduğunu göstermektedir. Precision, modelin pozitif olarak tahmin ettiği örneklerin gerçekten pozitif olma olasılığını gösterir ve modelin pozitif tahminlerinin %92,86'sının gerçekten pozitif olduğu görülmektedir. Recall, gerçek pozitif örneklerin model tarafından kaçının tespit edildiğini gösterir ve model, gerçek pozitif örneklerin %92,86'sının doğru bir şekilde tespit edildiğini göstermektedir. F1 score'u %92,86'dır, yani model, hem precision hem de recall açısından başarılı bir performans sergilemektedir.

Modelin true positive rate'i yüksek (%92,86), false positive rate'i ise yüksektir (%50). Bu sonuç, modelin gerçek pozitif örnekleri doğru bir şekilde tespit ederken, aynı zamanda bazı gerçek negatif örnekleri de yanlışlıkla pozitif olarak tahmin ettiğini göstermektedir. False negative rate'i düşüktür (%7,14), bu da modelin gerçek pozitif örneklerin çoğunu doğru bir şekilde tespit ettiğini göstermektedir.

ROC (Receiver Operating Characteristic) eğrisi, sınıflandırma modelinin farklı kesme noktalarında performansını ölçen bir grafiktir. Bu veri setinin ROC eğrisi, AUC değeri 0.71 olarak hesaplanmıştır. Eğri, sol üst köşeye doğru yüksek bir eğim ile yükselmektedir, bu da modelin yüksek bir TPR'ye sahip olduğunu göstermektedir. Ancak, yüksek bir FPR değerine de sahiptir, bu da modelin yanlış pozitif sonuçları da sınıflandırdığını göstermektedir. Bu nedenle, bu modelin belirli bir uygulama için kullanılabilirliğini ve yararlılığını değerlendirirken hem TPR hem de FPR değerleri göz önünde bulundurulmalıdır.

Confusion matrix, sınıflandırma modellerinin performansını değerlendirmek için kullanılan bir metriktir. Modelin true positive ve true negative değerleri yüksekken, false positive ve false negative değerleri de yüksektir. Bu sonuç, modelin hem gerçek pozitif örnekleri doğru bir şekilde tespit ettiğini hem de gerçek negatif örneklerin bazılarını yanlışlıkla pozitif olarak sınıflandırdığını göstermektedir.

Sonuç olarak, bu karar ağacı sınıflandırma modelinin performansı genel olarak yüksek olsa da, bazı yanlış pozitif sonuçlar verdiği görülmüştür. Bu yanlış pozitif sonuçlar, modelin özellikle gerçek negatif örnekleri yanlışlıkla pozitif olarak sınıflandırması nedeniyle oluşmuştur. Modelin true positive rate'i yüksek olmakla birlikte, false positive rate'i de yüksektir. Bu nedenle, modelin belirli bir uygulama için kullanılabilirliği ve yararlılığı değerlendirilirken hem true positive rate hem de false positive rate değerleri göz önünde bulundurulmalıdır.

Kaynakça

- https://archive.ics.uci.edu/ml/datasets/hepatitis
- file:///C:/Users/saltu/Downloads/live-120-1428-jair.pdf
- https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=589207&tag=1
- https://ieeexplore.ieee.org/abstract/document/7275013?casa_token=q-5uAjQvHoMAAAAA:vaUb2qc0RoAokb5srk7wGFalA-BvaYDm20G7iUK_SkLb4T2xPabCcpoKlO7A-GmhBn_bKGZ4hUz0
- https://www.ripublication.com/ijaer18/ijaerv13n16_45.pdf