Non-Alcoholic Fatty Liver Disease Veri Seti Analizi (Fibrozis)

Zehra Atmaca
23052605
zehra.atmaca@std.yildiz.edu.tr
Yıldız Teknik Üniversitesi
Matematik Mühendisliği

I. GİRİŞ

Karaciğer fibrozisi, karaciğer dokusunun yara dokusuyla yer değiştirmesi sonucu gelişen ve zamanla siroz ya da karaciğer yetmezliği gibi ciddi sağlık sorunlarına yol açabilen bir hastalıktır. Erken teşhis, fibrozis ilerlemesinin önlenmesi ve uygun tedavi stratejilerinin belirlenmesi açısından kritik öneme sahiptir. Ancak, fibrozis teşhisi genellikle invaziv yöntemlerle yapılır ve bu durum hem hasta için rahatsızlık yaratır hem de sağlık hizmetleri için yüksek maliyet anlamına gelir. Bu nedenle, biyokimyasal ve klinik verileri kullanarak, daha az invaziv ve hızlı bir şekilde teşhis koyabilecek makine öğrenimi modellerinin geliştirilmesi, sağlık hizmetlerinde önemli bir ihtiyaca yanıt verebilir. Bu çalışmada, NAFLD (Non-Alcoholic Fatty Liver Disease) hastalarından toplanmış biyokimyasal ve klinik verileri kullanarak fibrozis teşhisi için makine öğrenimi modelleri geliştirilmiştir.

A. Amaç

Çalışma, Logistic Regression ve Random Forest algoritmalarını kullanarak, fibrozis teşhisinde bu iki farklı modelin etkinliğini karşılaştırmayı amaçlamaktadır. Projenin temel hedefi, iki modelin performansını karşılaştırarak fibrozis teşhisi için en uygun makine öğrenimi yöntemini belirlemektir. Bu süreçte, veri setindeki eksik veriler ve sınıf dengesizliği gibi problemler özenle ele alınmıştır. Elde edilen sonuçlar, makine öğreniminin klinik karar destek sistemlerine entegre edilerek hastalık teşhisinde nasıl katkı sağlayabileceğini göstermeyi hedeflemektedir. Bu bağlamda, çalışma sadece fibrozis teşhisi için değil, genel olarak sağlık verilerinin analitik yöntemlerle değerlendirilmesi için bir çerçeve sunmayı amaçlamaktadır.

B. Veri Setlerinin Tanıtımı

Non-Alcoholic Fatty Liver Disease

Veri seti, NAFLD (Non-Alcoholic Fatty Liver Disease) hastalığına sahip bireylerden toplanmıştır. Toplam 605 gözlem ve 62 özellik içeren veri seti, biyokimyasal ölçümler, demografik bilgiler ve klinik değerlendirmeler gibi çeşitli verileri barındırmaktadır. Hedef değişken, bireyin fibrozis durumunu belirtmektedir (0: Fibrozis Yok, 1: Fibrozis Var). Ancak veri setinde sınıf dengesizliği (%67.6 Fibrozis Var, %32.4 Fibrozis Yok) ve eksik veri gibi zorluklar bulunmaktadır.

Bu veri setinde Fibrozis tanısında önemli rol oynayan özellikler şunlardır;

- Yaş (Age): Hastanın yaşı.
- Vücut Kitle İndeksi (BMI): Hastanın kilo ve boyuna göre hesaplanan sağlık göstergesi.
- Karaciğer Fonksiyon Testleri:

AST (Aspartat Aminotransferaz): Karaciğer hasarını belirleyen biyokimyasal ölçüm.

ALT (Alanin Aminotransferaz): Karaciğer enzim düzeyini gösteren bir ölçüm.

GGT (Gamma-Glutamil Transferaz): Karaciğer sağlığı ile ilgili önemli bir enzim.

- Diyabet Durumu: Hastanın diyabet varlığı (0: Yok, 1: Var).
- NAS Skoru (Kleiner Yöntemi): Karaciğer dokusunun histolojik değerlendirmesini sağlayan bir ölçüt.

II. GELİŞME

Veri setleri analiz edildikten sonra boş verileri çoğunlukta olan sütunlar belirlenerek veri ön işleme yapılır. Veri ön işlem yapıldıktan sonra sınıflandırma algoritmaları yapılarak sonuçlar değerlendirilir.

A. Veri Ön İşleme

Eksik Değerlerin İncelenmesi:

- Her bir özelliğin eksik veri oranı hesaplanmıştır.
- %50'den fazla eksik veri içeren sütunlar analiz dışı bırakılmıştır.

```
#Sütunlardaki eksik veri yüzdesinin kontrolü
missing_percentage = data.isnull().mean() * 100

#Çok eksik sütunları çıkarma (sınırı %50 belirledik )
data.drop(columns=missing_percentage[missing_percentage > 50].index, inplace=True)
```

Doldurma Yöntemleri:

- Sayısal özellikler Medyan değer ile doldurulmuştur. Örneğin, Bel Çevresi ve Hemoglobin-A1C gibi değişkenler için bu yöntem kullanılmıştır.
- Kategorik Özellikler Mod kullanılarak doldurulmuştur. Örneğin, Diyabet Durumu gibi kategorik değişkenler için bu yöntem uygulanmıştır.
- Forward Fill zaman sırasına dayalı eksik veriler için kullanılmıştır (ör. AST ve ALT). Veri setimiz klinik veriler içerdiğinden bu tekniği kullanmayı tercih edebiliri

Bu doldurma yöntemlerini veri setimizde klinik veriler bulunduğunu unutmadan ve aşağıda verilen data tipi ve eksik veri oranı gibi parametreleri dikkate alarak uygulamalıyız.

	Özellik Adı	Veri Tipi	Eksik Veri Oranı (%)
0	Yaş (Age)	Sayısal (int)	0.0
1	Vücut Kitle İndeksi (BMI)	Sayısal (float)	0.0
2	Bel Çevresi (Waist Circumference)	Sayısal (float)	4.8
3	Diyabet Durumu (Diyabetes Mellitus)	Kategorik (int)	0.0
4	AST	Sayısal (float)	0.0
5	ALT	Sayısal (int)	0.0
6	GGT	Sayısal (float)	0.5
7	Kleiner NAS Skoru	Kategorik (int)	0.0
8	Hemoglobin-A1C	Sayısal (float)	6.8

Aşağıda bu veri doldurma yöntemlerinin Python programı üzerinden üzerinden nasıl yapılacağı gösterilmiştir;

```
#Eksik verileri uygun metodlarla doldurma
fill_strategies = {
    'numeric_median': ['Age', 'Body Mass Index', 'Waist Circumference', 'Hemoglobin - A1C'],
    'forward_fill': ['AST', 'ALT', 'GGT'],
    'mode': ['Diyabetes Mellitus (No=0, Yes=1)', 'NAS score according to Kleiner']
}

for col in fill_strategies['numeric_median']:
    if col in data.columns:
        data[col] = data[col].fillna(data[col].median())
for col in fill_strategies['forward_fill']:
    if col in data.columns:
        data[col] = data[col].ffill()
for col in fill_strategies['mode']:
    if col in data.columns:
        data[col] = data[col].fillna(data[col].mode()[0])
```

Temizleme ve doldurma işlemlerinin ardından artık hedef ve özelliklerimizi tanımlayabiliriz.

```
#Hedef ve özelliklerin belirlenmesi
selected_features = [
    'Age', 'Body Mass Index', 'Waist Circumference', 'Diyabetes Mellitus (No=0, Yes=1)',
    'AST', 'ALT', 'GGT', 'NAS score according to Kleiner', 'Hemoglobin - A1C'
]
selected_features = [col for col in selected_features if col in data.columns]
X = data[selected_features]
y = data['Fibrosis status (No=0, Yes=1) (Fibrosis 1 and above, there is Fibrosis)']
```

Test ve Eğitim Süreci:

Bu çalışmada kullanılan veri seti, eğitim ve test olarak iki farklı bölüme ayrılmıştır. Bu bölme işlemi, modellerin performansını nesnel bir şekilde değerlendirmek için gereklidir. Eğitim seti, modelin öğrenme süreci için kullanılırken, test seti, modelin yeni verilere karşı nasıl genelleme yaptığını ölçmek için ayrılmıştır.

Veri setinin %80'i, modellerin eğitimi için kullanılmıştır. Bu, modelin hedef değişken (fibrozis durumu) ile bağımsız değişkenler (Yaş, BMI, AST, ALT vb.) arasındaki ilişkileri öğrenmesini sağlar.

Veri setinin %20'si, eğitilen modelin performansını değerlendirmek amacıyla ayrılmıştır. Test seti, modelin daha önce görmediği verilerden oluşur ve genelleme kapasitesini ölçer.

Veri bölme işlemi, scikit-learn kütüphanesindeki train_test_split fonksiyonu ile gerçekleştirilmiştir. Rastgelelik kontrolü için random_state=42 kullanılmış ve böylece bölme işlemi tekrarlanabilir hale getirilmiştir.

Aşağıda eğitim ve test setlerine ayırma işlemi Python'da gösterilmektedir;

```
#Eğitim ve Test verilerine bölme
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
#Modellerin Tanımlanması
models = {
    'Logistic Regression': LogisticRegression(random_state=42, max_iter=1000),
    'Random Forest': RandomForestClassifier(random_state=42, n_estimators=100)
}

# Model Eğitimi ve Değerlendirme
results = {}
conf_matrices = {}
for name, model in models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    y_proba = model.predict_proba(X_test)[:, 1] if hasattr(model, 'predict_proba') else None
    conf_matrices[name] = confusion_matrix(y_test, y_pred)
    results[name] = {
        'Accuracy': accuracy_score(y_test, y_pred),
        'Precision': precision_score(y_test, y_pred),
        'Recall': recall_score(y_test, y_pred),
        'F1 Score': f1_score(y_test, y_pred),
        'AUC': roc_auc_score(y_test, y_proba) if y_proba is not None else None,
    }
```

B. *Sınıflandırma Algoritmaları*

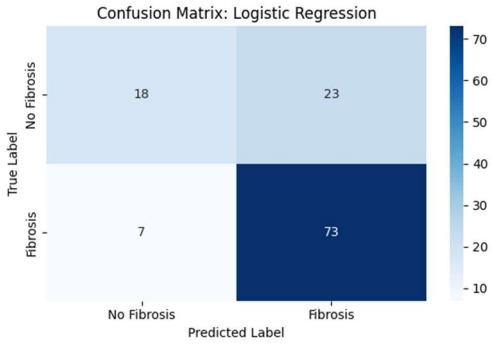
Logistic Regresyon

Bağımsız değişkenler ile hedef değişken arasında doğrusal bir ilişki olduğunu varsayan temel bir sınıflandırma algoritmasıdır. Modelin seçilmesindeki başlıca nedenler şunlardır:

- Logistic Regression, hızlı bir şekilde eğitilebilen ve uygulaması kolay bir modeldir. Özellikle, sınırlı sayıda gözlem ve özellik içeren veri setlerinde etkili bir şekilde çalışır.
- Logistic Regression, model çıktılarının klinik ortamlarda kolayca anlaşılabilir olması nedeniyle tercih edilmiştir.

Logistic Regression Algoritmasının Confusion Matrix Grafiği

```
# Confusion Matrix Görselleştirme
for name, matrix in conf_matrices.items():
    plt.figure(figsize=(6, 4))
    sns.heatmap(matrix, annot=True, fmt='d', cmap='Blues', xticklabels=['No Fibrosis', 'Fibrosis'],
                yticklabels=['No Fibrosis', 'Fibrosis'])
    plt.title(f'Confusion Matrix: {name}')
    plt.xlabel('Predicted Label')
    plt.ylabel('True Label')
    plt.tight_layout()
    plt.show()
```



Verilen Python koduyla elde ettiğimiz bu grafikten şu çıkarımları yapabiliriz;

Gerçek Pozitif (TP): 73
Gerçek Negatif (TN): 18
Yanlış Pozitif (FP): 23
Yanlış Negatif (FN)**:** 7

Yani model, toplam vakaların 91'ini başarıyla sınıflandırdı.
Model 30 yanlış sınıflandırma yaptı.

Duyarlılık: TP/TP+FN=73/73+7=0.91

Bu, modelin gerçek "Fibrosis" durumlarını doğru bir şekilde tespit etme oranının %91 olduğunu gösterir.

Kesinlik: TP/TP+FP=73/73+23=0.76

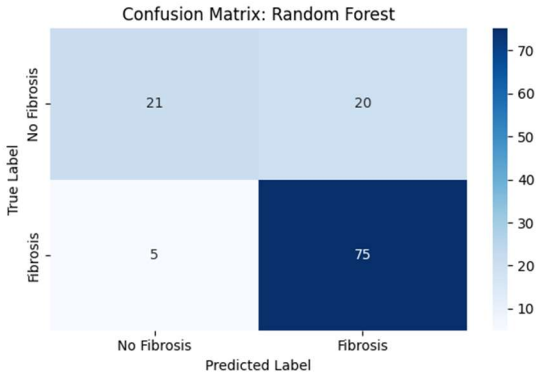
Model Fibrosis sınıfında yüksek bir duyarlılık sergilemektedir. Bu, modelin gerçek pozitifleri iyi bir şekilde tespit ettiğini gösterir.Ancak, "No Fibrosis" sınıfında yanlış pozitif oranı görece yüksektir. Bu durum, modelin bazı durumlarda "No Fibrosis" sınıfını yanlış bir şekilde "Fibrosis" olarak tahmin ettiğini göstermektedir.

Random Forest

Random Forest, birden fazla karar ağacının oluşturduğu bir topluluk modeli olup, daha karmaşık ilişkileri öğrenebilme kapasitesine sahiptir. Modelin seçilmesindeki başlıca nedenler şunlardır:

- Fibrozis teşhisi, bağımsız değişkenler ile hedef değişken arasında doğrusal olmayan ilişkiler içerebilir. Random Forest, bu karmaşık ilişkileri etkili bir şekilde öğrenebilir. Bu, yeni verilere karşı daha iyi performans gösterebilmesi anlamına gelir.
- Veri setindeki sınıf dengesizliği (%67.6 Fibrozis Var, %32.4 Fibrozis Yok), Random Forest ile daha iyi yönetilebilir. Model, azınlık sınıfı olan "Fibrozis Yok" durumunda daha iyi performans gösterme potansiyeline sahiptir.
- Random Forest, her bir özelliğin hedef değişken üzerindeki önemini belirleyebilir. Bu, hangi biyobelirteçlerin fibrozis teşhisi için kritik olduğunu anlamamıza yardımcı olur.
-

Random Forest Confusion Matrix Grafiği



Verilen grafikten şu çıkarımları yapabilriiz;

Gerçek Pozitif (TP): 75
Gerçek Negatif (TN): 21
Yanlış Pozitif (FP): 20
Yanlış Negatif (FN):5

Yani model, toplam vakaların 96'sını başarıyla sınıflandırdı.
Model 25 yanlış sınıflandırma yaptı.

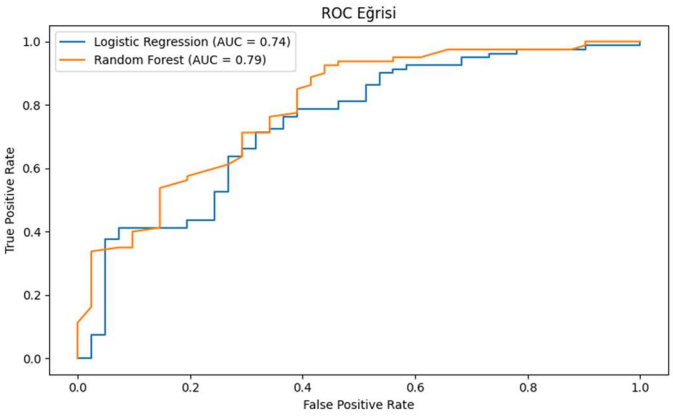
Duyarlılık: TP/TP+FN=75/75+5= 0.94

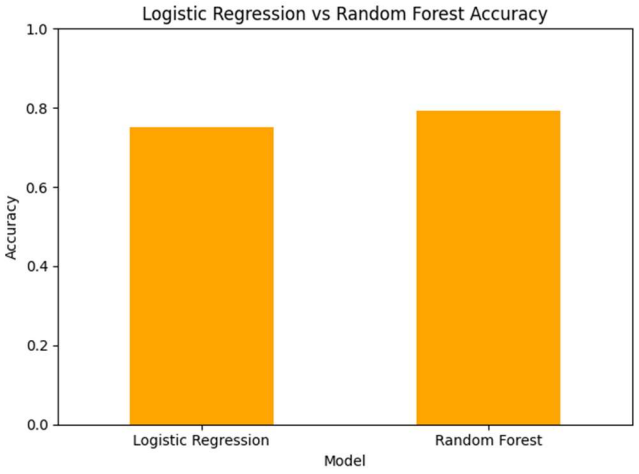
Bu, modelin gerçek "Fibrosis" sınıfını tespit etme oranının %94 olduğunu gösterir.

Kesinlik: TP/TP+FP = 75/75+20= 0.79

Random Forest modeli, "Fibrosis" sınıfındaki gerçek pozitifleri oldukça yüksek bir oranla tespit etmiştir. Bu, modelin "Fibrosis" durumlarını gözden kaçırma olasılığının düşük olduğunu gösterir.
Modelin "Fibrosis" tahminlerinin doğruluğu %79'dur. Bu, Logistic Regression modeline kıyasla bir miktar daha yüksektir.

ROC Eğrisi Grafik Analizi





Yukarıdaki grafik, Logistic Regression ve Random Forest modellerinin ROC (Receiver Operating Characteristic) eğrilerini göstermektedir.

Random Forest, daha yüksek AUC değeri (%79) ile, fibrozis teşhisi için daha etkili bir model olarak öne çıkmaktadır. Bu model, hem pozitif sınıfları doğru bir şekilde tespit etme hem de yanlış pozitif oranını düşük tutma konusunda daha başarılıdır. Logistic Regression, daha düşük AUC değeri (%74) ile, daha hızlı ve basit bir model olarak avantaj sağlar, ancak sınıflandırma performansı Random Forest'a kıyasla daha düşüktür.

III. SONUÇ

Random Forest, tüm metriklerde Logistic Regression'a göre daha iyi performans sergilemiştir. Özellikle sınıf dengesizliğindnen kaynaklanan zorlukları daha iyi yönetmiş ve yüksek bir AUC değeri elde etmiştir.

Logistic Regression ise daha hızlı çalışması ve modelin yorumlanabilirliği açısından avantaj sağlamıştır.

Bu sonuçlara dayanarak, fibrozis teşhisi için Random Forest modeli daha etkili bir seçimdir. Daha yüksek recall değeri, modelin hastaları gözden kaçırma olasılığını azalttığı için klinik uygulamalarda daha kritik bir rol oynayabilir. Ancak, Logistic Regression modeli de düşük hesaplama maliyeti ve kolay yorumlanabilirliği ile alternatif bir seçenek sunar.

Bunların yanında model şu geliştirmelere de açıktır;

Hiperparametre optimizasyonu yapılarak Random Forest modelinin performansı daha da artırılabilir.

Daha dengeli ve büyük veri setleriyle çalışmak, modellerin genelleme kapasitesini güçlendirebilir.

Logistic Regression'ın yorumlanabilirliğini ve Random Forest'ın doğruluğunu birleştiren hibrit modeller geliştirilebilir.

Kaynakça:

https://www.kaggle.com/datasets/sunilsah905/non-alcoholic-fatty-liver-disease/data