Alkol Dışı Yağlı Karaciğer Hastalığı Tanısında SVM ve Random Forest Tahmin Modellerinin Karşılaştırmalı Analizi

Ayça Durgut
Yıldız Teknik Üniversitesi, Matematik Mühendisliği Bölümü
İstanbul, Türkiye
ayca.durgut@outlook.com

I. GİRİŞ

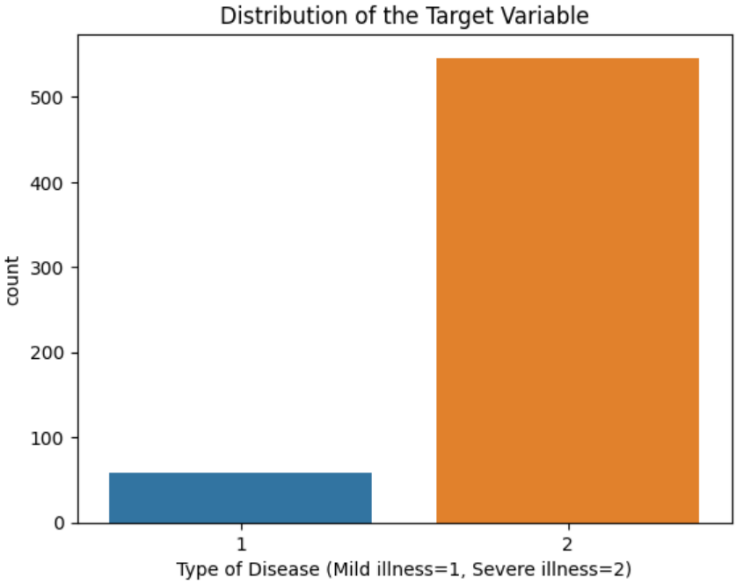
Bu çalışmada Alkol Dışı Yağlı Karaciğer Hastalığı (Non-Alcoholic Fatty Liver Disease/NAFLD) dünya genelinde yaygın bir karaciğer hastalığıdır. Bu çalışmada, ilgili veri seti ile makine öğrenimi algoritmaları kullanarak sınıflandırma modelleri geliştirilmesine ve tanı ve hastalık şiddetinin tahmin edilmesinde odaklanılmaktadır. Bu amaçla, Non-Alcoholic Fatty Liver Disease veri seti kullanılmıştır. Veri seti 605 gözlemden ve 62 değişkenden oluşmaktadır. İlgili veri seti, hastalara ait demografik bilgiler (yaş, cinsiyet, boy, kilo), biyokimyasal ölçümler (AST, ALT, kreatinin gibi) ve klinik tanılar (diyabet, hipertansiyon, fibrozis durumu) gibi özellikleri içermektedir. Çalışmanın temel amacı hastalığın şiddeti ve tipi gibi hedef değişkenlerin tahmin edilmesi için Support Vector Machine (SVM) ve Random Forest modellerini geliştirmek ve sonuçlarını karşılaştırmaktır.

Veri setindeki değişkenlerin bir kısmında eksik değerler bulunmaktadır. Bu değerler uygun yöntemlerle doldurulmuştur. Veri setinde sınıf dengesizliği sorunu vardır; "hafif hastalık" sınıfı 59 gözlem içerirken, "şiddetli hastalık" sınıfı 546 gözlem içermektedir. Bu durum, modellerin performansını etkilediğinden modellerde iyileştirmeler yapılmıştır.

Bu çalışmada, veri setini temizleme, özellik seçimi ve ön işleme adımları uygulanmıştır. Daha sonra, sınıflandırma amacıyla SVM ve Random Forest algoritmaları kullanılmıştır. Modellerin performanslarını karşılaştırmak için doğruluk (accuracy), F1 skoru ve ROC-AUC gibi metrikler değerlendirilmiştir.

II. GELİŞME

İlk olarak, hedef değişkenin ("Type of Disease: Mild illness = 1, Severe illness = 2") dağılımını inceledik. Elde edilen sonuçta sınıf dengesizliği ile karşılaştık. "Hafif hastalık" sınıfı yalnızca 59 gözlem içerirken, "şiddetli hastalık" sınıfı 546 gözlem içeriyordu. Bu durum, sınıflandırma algoritmalarının "hafif hastalık" sınıfını öğrenmede zorlanabileceğinin işaretini verdi ve modelleme aşamasında gerekli değişiklikler ve önlemler göz önünde bulunduruldu.



Veri setindeki "Patient No" sütunu, model için anlamlı bilgi taşımadığı ve yalnızca gözlem numarasını ifade etmektedir. Sınıflandırma algoritmasına bir katkısı olmadığı ve modelleme sürecini etkileyebileceği göz önünde bulundurulmuştur.

Veri setinde yüzde 25'ten fazla eksik değere sahip sütunlar analiz edilmiştir. Bu sütunlar arasında karaciğer hastalığı tanısında etkili olabilecek değişkenler olduğundan ("Vitamin D" ve Microalbumin/Creatinine Ratio") bu sütunlar veri setinden çıkarılmamıştır. Bunun yerine eksik değerlerin doldurulması ile devam edilmiştir.

Eksik verilerin doldurulması aşamasında veri setindeki tüm değişkenlerin nümerik değişkenler olduğu göz önüne alınmıştır. Değişkenlerden bazıları birbirleri ile bağlantılı

olduğu ya da özel hesaplamalar gerektirdiği için ayrı ayrı incelenmiş, geri kalan değişkenler için ise eksik değerler medyan ile doldurulmuştur. Medyan değeri kullanılmasının nedeni, bu yöntemin uç değerlerin etkisini azaltmasıdır.

Özel hesaplamalar gerektiren sütunlar ise ayrı ayrı doldurulmuştur. Bu sütunlar ve hesaplama yöntemleri şunlardır:

BUN (Blood Urea Nitrogen): Medyan değeri ile doldurulmuştur.

Ceruloplasmin: Medyan değeri ile doldurulmuştur.

Vitamin D: Yaş gruplarına göre medyan kullanılarak doldurulmuştur.

Glucose: Medyan değeri ile doldurulmuştur.

Insulin: Medyan değeri ile doldurulmuştur.

HOMA (Homeostatic Model Assessment): Bu sütun, "Glucose" ve "Insulin" değerlerinin formül yardımıyla hesaplanmasıyla doldurulmuştur:

HOMA = (Glucose x Insulin) / 405

Insulin Resistance (HOMA'ya Göre): "HOMA" değerleri ≥ 2.5 ise direnç var (1), değilse yok (2) olarak etiketlenmiştir.

NAS Score ve Fibrosis: NAS skoruna ve fibrozis değerine dayalı hesaplamalar yapılarak özel bir sütun oluşturulmuştur:

Eğer NAS Score >= 4 ve Fibrosis >= 2 ise 1, aksi takdirde 0 değeri verilmiştir.

Diagnosis According to SAF: NAS skoruna göre bir sütun oluşturulmuştur:

Eğer NAS Score ≥ 5 ise 1, aksi takdirde 2 değeri verilmiştir.

Eksik değerler doldurulduktan sonra korelasyon analizi ile devam edilmiştir. Hedef değişken ile korelasyonu 0.05'in altında olan sütunlar (ALP, Total Bilirubin, LDL, Microalbumin Spot Urine, Mean Platelet Volume), hedef değişkenle anlamlı bir ilişki göstermediği için çıkarılmıştır. Hedef değişkenle düşük korelasyona sahip sütunların, modelin performansına katkı sağlamayabileceği ve hesaplama yükünü artırabileceği göz önüne alınmıştır. Kendi aralarında ve 0.8'den yüksek korelasyon gösteren değişkenler de incelenmiş ve bu sütunlardan bazıları modelden çıkarılmıştır. Bu sütunlar şunlardır:

Height (Boy): Waist Circumference (Bel Çevresi) ile yüksek korelasyon gösterdiği için çıkarılmıştır. Bel çevresi, karaciğer hastalıklarında daha doğrudan bir gösterge olduğundan bu değişken modelde tutulmuştur.

AST ve ALT: Her iki değişken karaciğer fonksiyonlarıyla ilişkili olsa da birinin tutulması modelde bilgi kaybını önlemek için yeterli görülmüştür.

Veri setindeki nümerik değişkenler, Standard Scaler kullanılarak ölçeklendirilmiştir. Bu işlem özellikle SVC (Support Vector Classifier) gibi mesafe tabanlı modeller için gereklidir çünkü bu tür algoritmalar farklı ölçeklerdeki değişkenlerden eşit şekilde etkilenemez. Aynı zamanda bu işlem Random Forest için herhangi bir etki yaratmayacağından iki model için de aynı şekilde ilerlenmiştir.

Veri seti üzerindeki hazırlık işlemleri tamamlandıktan sonra son kontroller yapılmıştır ve sınıflandırma modellerinin performansını değerlendirmek için Support Vector Classifier (SVC) ve Random Forest algoritmaları kullanılarak ilk modeller oluşturulmuştur. Bu modeller oluşturulurken SVC için RBF Kernel, ve ceza parametresi C değeri 1 ve gamma değeri 0.1 kullanılmıştır. Sınıf dengesizliğini dengelemek adına, bu modelde "class_weight=balanced" parametresi kullanılmıştır. Random Forest modeli için ise ağaç sayısı 100 olarak belirlenmiş, sabit bir rastgele durum için random_state=42 parametresi eklenmiştir.

İlk değerlendirmeler sonucunda, test veri setinde her iki model de "hafif hastalık" sınıfını yeterince iyi tahmin edememiştir. SVC modeli, "hafif hastalık" sınıfında düşük precision ve recall değerlerine sahiptir ve Random Forest modeli genel doğrulukta daha iyi bir performans sergilemesine rağmen aynı sınıfta benzer düşük performans göstermiştir. Test veri setinde alınan düşük skorların üzerine eğitim veri setinde her iki modelin de çok yüksek doğruluk sağladığı tespit edilmiş, bu modellerin overfitting yaptığını ortaya koyulmuştur. Bunun üzerine, modellerin genelleme kabiliyetini artırmak amacıyla bazı düzenlemeler yapılmıştır.

SVC modeli için, ceza parametresi C değerinin 0.1'e düşürerek bir deneme yapıldı. Fakat bu durum model performansını daha da kötüleştirdi. Bunun üzerine "RBF kernel" yerine "linear kernel" tercih edilerek model yeniden eğitilmiştir. Linear kernel, genelleme performansını artırma potansiyeline sahiptir ve linear kernel ile yapılan deneme, hafif hastalık sınıfında daha iyi sonuçlar vermiştir. Ceza parametresi C bu aşamada yeniden 1 olarak belirlenmiştir.

Random Forest modelinde ise her bir dalın minimum örnek sayısını 5'e çıkarılarak genelleme kabiliyeti artırılmıştır ve maksimum derinlik 10 ile sınırlandırılarak ağaçların gereksiz karmaşıklığı engellenmiştir. Ayrıca, sınıf ağırlıkları hedef değişkenin dengesiz dağılımını hesaba katacak şekilde yeniden optimize edilmiştir.

Bu düzenlemelerden sonra modeller hem test hem de eğitim veri setleri üzerinde yeniden değerlendirilmiş ve özellikle "hafif hastalık" sınıfında belirgin bir gelişim

olduğunu ortaya koyulmuştur. Ayrıca eğitim seti ve test setindeki sonuçların birbirine yakın olması, overfitting sorunlarının büyük ölçüde giderildiğini göstermiştir. Precision, recall, F1-score ve ROC-AUC gibi metrikler üzerinden yapılan karşılaştırmalar, iyileştirilmiş modellerin genelleme performansını artırdığını doğrulamıştır. Bu sürecin sonunda performansları artırılmış ve sınıf dengesizliğine rağmen anlamlı sonuçlar elde edilmiştir.

III. SONUÇ

Bu çalışma kapsamında, karaciğer hastalığı tanısında kullanılması hedeflenen sınıflandırma modellerinin performansı Random Forest ve SVM algoritmaları kullanılara karşılaştırılmıştır. İlk değerlendirmelerde, Random Forest modeli genellikle daha yüksek doğruluk oranlarına sahip olmasına rağmen, modelin sınıf dengesizliği nedeniyle "hafif hastalık" sınıfında düşük precision ve recall değerleri sergilediği gözlemlenmiştir. Benzer şekilde, SVC modeli RBF kernel ile çalıştırıldığında hedef değişkenin sınıf dengesizliği sebebiyle "hafif hastalık" sınıfında düşük performans göstermiştir.

Parametre optimizasyonunda, SVC için C değerinin 0.1 olarak seçilmesi model performansını düşürmüştür. Bunun üzerine, linear kernel kullanılarak model yeniden eğitilmiş ve özellikle sınıf dengesizliğine daha dayanıklı sonuçlar elde edilmiştir. SVC'nin linear kernel ile ROC-AUC değeri 0.93 olarak hesaplanmış ve bu da modelin sınıf ayrımında daha iyi bir performans sergilediğini göstermiştir. Random Forest modeli, ROC-AUC değeri 0.99 ile genel olarak daha yüksek doğruluk sağlamış, fakat sınıf dengesizliği sorununa tam çözüm üretememiştir. Random Forest modeline uygulanan parametre optimizasyonları modelin genelleme kabiliyetini artırarak overfitting durumunu önlemiştir. Sonuç olarak, ROC-AUC ve F1-score gibi metrikler göz önünde bulundurulduğunda, Random Forest modeli genel olarak daha iyi bir performans sunmuş, ancak linear kernel kullanılan SVC modeli dengeli sınıflandırma açısından daha iyi sonuç vermiştir. Bu durum, sınıf dengesizliği ile başa çıkmak için kernel seçiminin ve parametre optimizasyonunun önemini ortaya koymuştur. Elde edilen sonuçlar, daha dengeli ve daha çok veri içeren bir veri seti ile bu modellerin karaciğer hastalığı tanısında kullanılabileceğini göstermiştir.

IV. KAYNAKÇA

Kaggle Notebook: Steffi P. (n.d.). *EDA - Non Alcohol Fatty Liver Disease*. Retrieved from https://www.kaggle.com/code/steffip/eda-non-alcohol-fatty-liver-disease

Kaggle Notebook: Chanchal M. (n.d.). *Liver Disease Prediction Using 7 Models*. Retrieved from https://www.kaggle.com/code/chanchal24/liver-disease-prediction-using-7-models/notebook

Kaggle Notebook: Aymen Mouffok. (n.d.). *Non Alcoholic Fatty Liver Disease - 92% Accuracy*. Retrieved from https://www.kaggle.com/code/aymenmouffok/non-alcoholic-fatty-liver-disease-92-accuracy

OpenAI ChatGPT. *Generative AI Assistant*. Accessed in December 2024 for domain expertise, clarifications, and analytical discussions.