



Feature Engineering Approach for Disease Severity Classification in Hemodialysis Diseases

Ahmet Nihat Velioğlu^{1*}, Adnan Özhan Pekmezci²

¹Department of Computer Engineering, Iskenderun Technical University, Hatay, Türkiye.

²Internal Medicine, İskenderun State Hospital, Turkey

ABSTRACT

Hemodialysis patients require continuous monitoring because clinical and laboratory findings may reflect different levels of disease severity. In this study, a multiclass machine learning pipeline was developed for predicting hemodialysis disease severity as Mild, Moderate, and Severe. In the proposed pipeline, domain-oriented features such as blood pressure reduction, ultrafiltration per kilogram, ultrafiltration per hour, rate-pressure product, potassium-ultrafiltration interaction, interaction terms, and train-only class centroid distances were generated. SMOTE was applied only to the training partition, while the test set was kept in its natural imbalanced distribution. The proposed feature engineering + train-only SMOTE + XGBoost model achieved 0.9950 accuracy, 0.9968 macro precision, 0.9728 macro recall, and 0.9844 macro F1-score on the locked test set. Additional validation experiments showed stable performance across random seeds and 10-fold cross-validation, while the label shuffle test reduced macro F1-score to 0.2855, supporting that the model did not learn a trivial target proxy. The findings indicate that high classification performance can be obtained without applying SMOTE before the train-test split, and that leakage-aware evaluation is essential when reporting clinical machine learning results.

ARTICLE INFO

Received 2026-05-21,

Accepted 2026-06-29,

Publication Date 2026-06-30

Keywords:

Hemodialysis, XGBoost, SMOTE, feature engineering, data leakage, disease severity classification

Distributed Under CC-BY 4.0



*Corresponding Author: Ahmet NİHAT, E-mail: ahmetnihat.lee21@iste.edu.tr

Hemodiyaliz Hastalarında Hastalık Şiddeti Sınıflandırması için Öznitelik Mühendisliği Yaklaşımı

ÖZET

Hemodiyaliz hastalarında klinik ve laboratuvar bulguları hastalık şiddetinin farklı düzeylerini yansıtabilmektedir. Bu çalışmada, hemodiyaliz hastalık şiddetini Mild, Moderate ve Severe sınıfları altında tahmin etmek amacıyla çok sınıflı bir makine öğrenmesi yaklaşımı geliştirilmiştir. Önerilen pipeline içinde kan basıncı düşüşü, kilo başına ultrafiltrasyon, saatlik ultrafiltrasyon, rate-pressure product, potasyum-ultrafiltrasyon etkileşimi, değişkenler arası çarpım öznitelikleri ve yalnızca eğitim verisi üzerinden hesaplanan sınıf merkezi uzaklıkları üretilmiştir. SMOTE işlemi yalnızca eğitim kümesine uygulanmış, test kümesi doğal dengesiz dağılımıyla korunmuştur. Önerilen öznitelik mühendisliği + yalnızca eğitimde SMOTE + XGBoost modeli kilitli test seti üzerinde 0.9950 doğruluk, 0.9968 macro precision, 0.9728 macro recall ve 0.9844 macro F1-skoru elde etmiştir. Ek doğrulama deneylerinde farklı random seed değerlerinde ve 10-katlı çapraz doğrulamada performansın kararlı olduğu; etiket karıştırma testinde macro F1-skorunun 0.2855 seviyesine düşmesiyle modelin basit bir hedef kopyası öğrenmediği görülmüştür. Bulgular, SMOTE işlemini train-test ayırımından önce uygulamadan da yüksek sınıflandırma başarısı elde edilebileceğini ve klinik makine öğrenmesi çalışmalarında sızıntı farkındalıklı değerlendirmenin zorunlu olduğunu göstermektedir.

MAKALE BİLGİSİ

Keywords:

Hemodiyaliz, XGBoost, SMOTE, öznitelik mühendisliği, veri sızıntısı, hastalık şiddeti sınıflandırması

Distributed Under CC-BY 4.0



GİRİŞ

Kronik böbrek hastalığı, ilerleyici yapısı ve çoklu komplikasyonları nedeniyle sağlık sistemleri üzerinde önemli bir yük oluşturmaktadır. Hastalığın ileri evrelerinde hemodiyaliz, böbreğin süzme işlevini kısmen yerine getirmek için düzenli olarak uygulanan temel tedavi yöntemlerinden biridir. Hemodiyaliz hastalarında yaş, diyabet, hipertansiyon, üre, kreatinin, albümin, hemoglobin, Kt/V, URR, kan basıncı, ultrafiltrasyon ve damar yolu tipi gibi parametreler hastanın klinik durumunu değerlendirmede birlikte ele alınmaktadır. KDIGO 2024 kılavuzu kronik böbrek hastalığında risk değerlendirmesi ve komplikasyon yönetiminin hasta merkezli bakım için temel olduğunu vurgularken, KDOQI hemodiyaliz yeterliliği kılavuzu Kt/V, tedavi süresi, ultrafiltrasyon ve kan basıncı kontrolü gibi parametrelerin klinik takipteki önemini ortaya koymaktadır (KDIGO CKD Work Group, 2024; National Kidney Foundation, 2015; Chowdhury ve ark. 2024).

Tıbbi veri setlerinde sınıf dengesizliği, eksik değerler, heterojen ölçüm aralıkları ve değişkenler arası doğrusal olmayan ilişkiler model performansını doğrudan etkileyen önemli sorunlardır. Hemodiyaliz hastalarının Mild, Moderate ve Severe gibi şiddet sınıflarına ayrılması, sadece genel doğruluk oranı ile değil, özellikle azınlık sınıfların doğru tanınmasıyla değerlendirilmelidir. Çünkü çoğunluk sınıfına yönelen bir model yüksek accuracy değeri üretebilir; buna karşın klinik açıdan önemli olan hafif veya orta şiddetli hasta gruplarını kaçırabilir. Bu nedenle macro precision, macro recall, macro F1-score, balanced accuracy ve karışıklık matrisi gibi metrikler tıbbi sınıflandırma problemlerinde daha açıklayıcıdır.

Hemodiyaliz alanında makine öğrenmesi uygulamaları son yıllarda hızla artmıştır. Garcia-Montemayor ve ark. (2021), hemodiyaliz hastalarında mortalite tahmini için random forest modellerinin lojistik regresyona göre daha yüksek ayırt edici performans sağlayabildiğini göstermiştir. Sheng ve ark. (2020), hemodiyalize yeni başlayan hastalarda ilk yıl mortalitesini tahmin etmek için makine öğrenmesi modelleri geliştirmiş ve bu tür modellerin yüksek riskli hasta gruplarını erken dönemde belirlemede kullanılabileceğini belirtmiştir. Rankin ve ark. (2022) ise XGBoost algoritmasını kullanarak diyaliz başlangıcından sonraki ilk 90 gün içindeki mortaliteyi tahmin etmeye yönelik geniş ölçekli bir model geliştirmiştir. Bu çalışmalar, hemodiyaliz verilerinin makine öğrenmesi modelleriyle işlenebileceğini göstermektedir; ancak aynı zamanda model validasyonu, veri sızıntısı kontrolü ve klinik genellenebilirlik konularının dikkatle ele alınması gerektiğini de ortaya koymaktadır.

Bu çalışmanın doğrudan referans aldığı çalışma, Wafa ve ark. (2025) tarafından yayımlanan “Analisis Kinerja XGBoost pada Data Seimbang dan tidak Seimbang dalam Klasifikasi Tingkat Keparahan Pasien Hemodialisis” başlıklı makaledir. Söz konusu çalışmada Kaggle kaynaklı 5.000 hemodiyaliz kaydı kullanılmış, XGBoost algoritması ile Mild, Moderate ve Severe sınıfları tahmin edilmiş ve sınıf dengesizliğini gidermek için SMOTE uygulanmıştır. Makalede SMOTE öncesi 60:40 bölünmede accuracy 0.7055, 70:30 bölünmede 0.7000 olarak; SMOTE sonrası ise 60:40 bölünmede 0.8566 ve 70:30 bölünmede 0.8705 olarak raporlanmıştır. Bununla birlikte çalışmada

SMOTE işleminin veri bölme işleminden önce uygulanmasının aşırı iyimser performans tahminleri oluşturabileceği sınırlılık olarak belirtilmiştir (Wafa et al., 2025).

SMOTE, azınlık sınıfları temsil eden sentetik örnekler üretmek için geliştirilen yaygın bir yeniden örnekleme yöntemidir (Chawla et al., 2002). Ancak SMOTE gibi resampling tekniklerinin tüm veri setine uygulanıp daha sonra train-test ayrımı yapılması, test örneklerine ait bilgilerin eğitim sürecine dolaylı olarak taşınmasına neden olabilir. imbalanced-learn dokümantasyonu bu durumu veri sızıntısına yol açan yaygın bir hata olarak tanımlar ve resampling işleminin yalnızca eğitim verisi üzerinde yapılması gerektiğini vurgular. Benzer şekilde scikit-learn dokümantasyonu, preprocessing veya fit işlemlerinde test setinin kullanılmasının model skorlarını aşırı iyimser hale getirebileceğini belirtir. Bu nedenle tıbbi sınıflandırma çalışmalarında modelin yüksek başarısı kadar, bu başarının nasıl elde edildiği de kritik öneme sahiptir.

Bu çalışmanın amacı, hemodiyaliz şiddet sınıflandırması probleminde XGBoost tabanlı önceki yöntemi yeniden değerlendirmek, train-test ayrımından önce uygulanan SMOTE yaklaşımının metodolojik riskini tartışmak ve veri sızıntısı riskini azaltan bir öznitelik mühendisliği pipeline'ı ile elde edilen sonuçları sunmaktır. Çalışmanın temel katkıları şunlardır: (i) referans makaledeki XGBoost tabanlı yaklaşım yeniden üretilmiş ve elde edilen sonuçlar karşılaştırılmıştır; (ii) SMOTE, encoding, scaling ve öznitelik mühendisliği işlemleri yalnızca eğitim verisi üzerinde uygulanacak şekilde düzenlenmiştir; (iii) hemodiyaliz klinik mantığına uygun yeni öznitelikler üretilmiştir; (iv) model başarısı yalnızca tek test sonucu ile değil, seed stability, 10-fold cross-validation, label shuffle ve feature importance sanity analizleri ile desteklenmiştir.

MATERYAL VE YÖNTEM

Veri Seti

Çalışmada kullanılan veri seti Kaggle üzerinde yayımlanan “Hemodialysis Realtime Hospital Dataset” veri setidir (GSK RD Solutions, 2023). Veri setinde 5.000 hasta kaydı, 26 giriş değişkeni ve “Disease Severity” hedef değişkeni bulunmaktadır. Hedef değişken Mild, Moderate ve Severe olmak üzere üç sınıftan oluşmaktadır. Veri setindeki sınıf dağılımı dengesizdir: Mild sınıfında 145, Moderate sınıfında 1.184 ve Severe sınıfında 3.671 örnek bulunmaktadır. Bu dağılım, modelin çoğunluk sınıfına yönelme riskini artırdığı için değerlendirmede macro metriklere özel önem verilmiştir.

Tablo 1. Çalışmada kullanılan veri seti

Sınıf	Örnek Sayısı	Oran
Mild	145	2.90%
Moderate	1184	23.68%
Severe	3671	73.42%

Giriş değişkenleri demografik bilgiler, komorbiditeler, kan basıncı ölçümleri, laboratuvar değerleri ve diyaliz sürecine ilişkin parametrelerden oluşmaktadır. Yaş, kilo, kreatinin, üre, potasyum, hemoglobin, hematokrit, albümin, Kt/V, URR, idrar çıkışı, kuru ağırlık ve sıvı çekim hızı gibi sayısal değişkenlerin yanında cinsiyet, diyabet, hipertansiyon, böbrek yetmezliği nedeni, diyalizat içeriği, damar yolu tipi ve diyalizör tipi gibi kategorik değişkenler de kullanılmıştır.

Öznitelik Mühendisliği

Öznitelik mühendisliği aşamasında amaç, modelin yalnızca ham değişkenleri değil, klinik olarak anlamlı etkileşimleri ve hemodiyaliz sürecini temsil eden türetilmiş göstergeleri de öğrenebilmesini sağlamaktır. Üre, Kt/V, kan basıncı, kalp atış hızı, ultrafiltrasyon ve potasyum gibi değişkenler hemodiyaliz takibinde ayrı ayrı önemli olmakla birlikte, hastanın şiddet sınıfı çoğu zaman bu değişkenlerin birlikte oluşturduğu fizyolojik tabloyla ilişkilidir. Bu nedenle aşağıdaki yeni öznitelikler oluşturulmuştur.

Sınıf merkezi uzaklığı öznitelikleri veri sızıntısı açısından özellikle dikkatli tasarlanmıştır. Bu özniteliklerde her sınıfın merkezi yalnızca eğitim verisi kullanılarak hesaplanmış, test setindeki örnekler bu eğitim merkezlerine olan uzaklıklarına göre dönüştürülmüştür. Böylece test setinin hedef sınıf bilgisi öznitelik üretiminde kullanılmamıştır. Bu yaklaşım, modelin sınıf yapısına ait temsil gücünden faydalanmasını sağlarken test verisinin eğitim sürecine sızmasını engellemeyi amaçlamaktadır.

Modelleme Yaklaşımı

Ana sınıflandırıcı olarak XGBoost kullanılmıştır. XGBoost, gradyan artırılmalı karar ağaçlarını ölçeklenebilir ve düzenleştirilmiş bir yapıda uygulayan güçlü bir ensemble algoritmasıdır (Chen & Guestrin, 2016). Tabular klinik veri setlerinde doğrusal olmayan ilişkileri ve değişken etkileşimlerini yakalayabilmesi nedeniyle bu çalışma için uygun görülmüştür. Önerilen temel modelde `learning_rate=0.1`, `max_depth=6`, `n_estimators=200` ve `random_state=42` ayarları kullanılmıştır. Model çok sınıflı sınıflandırma yapacak şekilde eğitilmiş, performans test setinde accuracy, precision, recall, F1-score, balanced accuracy ve confusion matrix ile değerlendirilmiştir.

Bu çalışmada hiperparametre seçimi, en yüksek skoru aramaya yönelik kapsamlı bir GridSearch veya Optuna optimizasyonu olarak değil, referans çalışmadaki XGBoost-SMOTE yaklaşımının sızıntısız deneysel kurgu içinde yeniden değerlendirilmesini sağlayacak kontrollü bir modelleme tercihi olarak ele alınmıştır. Bu nedenle XGBoost parametreleri, referans çalışmayla karşılaştırılabilirliği koruyacak ve XGBoost'un tabular verilerdeki güçlü temel performansını temsil edecek sabit değerler üzerinden yürütülmüştür (Chen & Guestrin, 2016; Wafa et al., 2025). Bu tercih, çalışmanın ana odağını hiperparametre optimizasyonundan ziyade veri sızıntısını engelleyen pipeline tasarımına, eğitim ve test işlemlerinin ayrıştırılmasına ve klinik olarak anlamlandırılabilir öznitelik mühendisliğine yöneltmektedir.

Tablo 2. Üretilen yeni öznitelikler ve klinik/matematikselanlamları

Yeni Öznitelik	Hesaplama Mantığı	Açıklama
bp_reduction	Diyaliz öncesi kan basıncı - diyaliz sonrası kan basıncı	Diyaliz seansı sonrası basınç değişimini temsil eder.
bp_reduction_pct	$\text{bp_reduction} / \text{diyaliz öncesi kan basıncı}$	Kan basıncı düşüşünü görece oranla gösterir.
rate_pressure_product	Heart Rate x Pre-Dialysis Blood Pressure	Kardiyovasküler yük / stres göstergesi olarak kullanılır.
uf_per_kg	$\text{Fluid Removal Rate} / \text{Weight}$	Kilo başına sıvı çekim yoğunluğunu gösterir.
uf_per_hour	$\text{Fluid Removal Rate} / \text{Dialysis Duration}$	Saatlik sıvı çekim yükünü gösterir.
k_uf_interaction	Potassium x Fluid Removal Rate	Elektrolit durumu ile ultrafiltrasyon etkileşimini temsil eder.
Weight_mul_Hypertension	Weight x Hypertension	Kilo ve hipertansiyon birlikteliğini temsil eder.
Gender_mul_Kidney_Failure_Cause	Gender x Kidney Failure Cause	Cinsiyet ve böbrek yetmezliği nedeni etkileşimini temsil eder.
Gender_mul_Diabetes	Gender x Diabetes	Cinsiyet ve diyabet birlikteliğini temsil eder.
Gender_mul_Hypertension	Gender x Hypertension	Cinsiyet ve hipertansiyon birlikteliğini temsil eder.
Age_mul_Weight	Age x Weight	Yaş ve kilo bilgisinin birleşik etkisini temsil eder.
dist_to_class_*	Eğitim verisiyle hesaplanan sınıf merkezlerine uzaklık	Örneklerin Mild, Moderate ve Severe sınıf merkezlerine benzerliğini gösterir.

Sınıf dengesizliği için SMOTE yöntemi kullanılmıştır; ancak referans makaledeki riskli sıradan farklı olarak SMOTE yalnızca eğitim verisine uygulanmıştır. Test seti dengelenmemiş ve doğal sınıf dağılımı ile korunmuştur. Bu tercih, modelin gerçek kullanım senaryosuna daha yakın bir dağılım üzerinde test edilmesini sağlamaktadır. Ayrıca modelin azınlık sınıfları öğrenip öğrenmediğini daha doğru gözlemlemek için macro average metrikleri raporlanmıştır (Chawla et al., 2002; imbalanced-learn developers, n.d.; scikit-learn developers, n.d.).

Doğrulama ve Sızıntı Kontrol Deneyleri

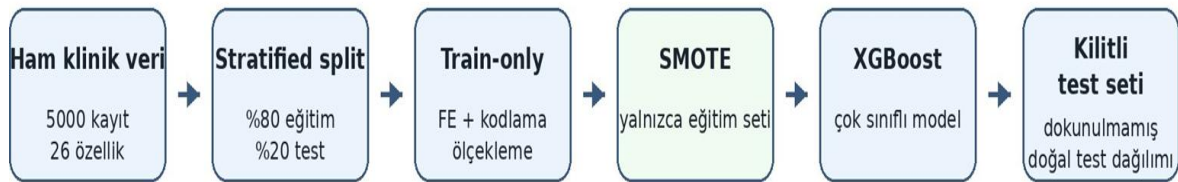
Modelin tek bir train-test bölünmesine bağımlı kalmadığını göstermek amacıyla ek doğrulama deneyleri yapılmıştır. İlk olarak beş farklı random seed ile train-test ayrımı tekrarlanmış ve her çalıştırmada model yeniden eğitilmiştir. İkinci olarak 10-katlı stratified cross-validation uygulanmıştır. Üçüncü olarak etiket karıştırma testi (label shuffle test) yapılmış; bu testte eğitim etiketleri karıştırılarak modelin gerçek klinik örüntüleri mi yoksa hedefe yakın sızıntılı bir değişkeni mi öğrendiği kontrol edilmiştir. Etiketler karıştırıldığında performansın rastgele tahmine yakın seviyeye düşmesi, modelin doğrudan hedef kopyası öğrenmediğine yönelik önemli bir sanity check olarak değerlendirilmiştir. Son olarak feature importance sanity analizi ile sınıf merkezi uzaklığı değişkenlerinin modele aşırı baskın olup olmadığı incelenmiştir.

BULGULAR

Deneyel Tasarım

Deneyel tasarım üç ana basamaktan oluşmaktadır. İlk basamakta referans makaledeki raporlanan sonuçlar literatür sonucu olarak ele alınmış ve makaledeki XGBoost hiperparametrelerine yakın ayarlarla yeniden üretim deneyi gerçekleştirilmiştir. İkinci basamakta, veri sızıntısı riskini azaltan doğru temel model kurulmuş; bu modelde train-test ayrımı başta yapılmış, veri ön işleme işlemleri yalnızca eğitim verisi üzerinde fit edilmiştir. Üçüncü basamakta ise domain-aware öznitelik mühendisliği ile desteklenmiş ve SMOTE işlemi yalnızca eğitim setine uygulanmış XGBoost modeli test edilmiştir.

Sızıntısız hemodiyaliz şiddet sınıflandırma iş akışı



Tüm fit / fit_transform işlemleri yalnızca eğitim verisi üzerinde yürütülmüş; test seti model seçimi ve öznitelik üretiminde kullanılmamıştır.

Şekil 1. Sızıntısız hemodiyaliz şiddet sınıflandırma pipeline akışı

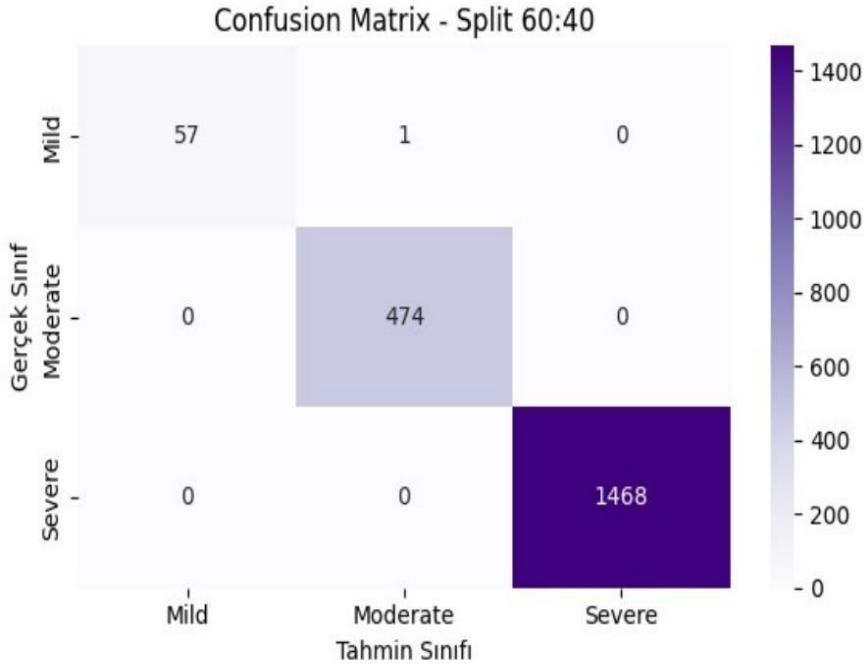
Bu tasarımda test seti başlangıçta ayrılmış ve model geliştirme sürecinde kullanılmamıştır. Böylece test seti, modelin daha önce görmediği doğal ve dengesiz dağılımı temsil eden kilitli değerlendirme kümesi olarak korunmuştur. SMOTE, scaler, encoder ve sınıf merkezi tabanlı öznitelikler yalnızca eğitim verisi üzerinde fit edilmiş; test verisine yalnızca eğitimde öğrenilen dönüşümler uygulanmıştır.

Referans Çalışma ve Yeniden Üretim Sonuçları

Referans çalışmada XGBoost modeli dengesiz veri üzerinde 60:40 ve 70:30 bölünmeler için yaklaşık 0.70 doğruluk elde etmiş; SMOTE uygulaması sonrasında accuracy değerleri sırasıyla 0.8566 ve 0.8705 seviyelerine yükselmiştir. Ancak aynı çalışmada SMOTE işleminin split öncesinde uygulanmasının performansı aşırı iyimser gösterebileceği sınırlılık olarak belirtilmiştir. Bu çalışmada yapılan yeniden üretim denemesinde, makaledeki XGBoost ayarlarına yakın hiperparametrelerle yalnızca orijinal öznitelikler kullanıldığında 60:40 bölünmede 0.9995 accuracy ve 0.9968 macro F1; 70:30 bölünmede 0.9980 accuracy ve 0.9945 macro F1 elde edilmiştir. Bu sonuç, veri setindeki hedef sınıfların ham değişkenler üzerinden oldukça güçlü biçimde ayrışabildiğini göstermektedir (Wafa et al., 2025).

Tablo 3. Referans çalışma ve yeniden üretim sonuçlarının karşılaştırılması

Deney	Bölünme	Accuracy	Macro F1
Wafa et al. (2025) - Dengesiz veri	60:40	0.7055	-
Wafa et al. (2025) - SMOTE sonrası	60:40	0.8566	0.85
Wafa et al. (2025) - Dengesiz veri	70:30	0.7000	-
Wafa et al. (2025) - SMOTE sonrası	70:30	0.8705	-
Bu çalışma - referans yönteme yakın yeniden üretim	60:40	0.9995	0.9968
Bu çalışma - referans yönteme yakın yeniden üretim	70:30	0.9980	0.9945



Şekil 2. Makale yöntemi yeniden üretimi 60:40 split test karışıklık matrisi

Sızıntısız Ham Veri Temel Modeli

Feature engineering ve SMOTE kullanılmadan kurulan ham veri temel modelinde train-test ayrımı başlangıçta yapılmış ve test seti eğitim sürecinden izole tutulmuştur. Bu model test setinde 0.9980 accuracy, 0.9981 macro precision, 0.9871 macro recall ve 0.9925 macro F1-score elde etmiştir. Ham veri modelinin yüksek başarısı, veri setinde Urea, Diabetes, Hypertension, Age ve Kt/V gibi değişkenlerin hedef sınıflar ile güçlü ilişki taşıdığını göstermektedir. Bununla birlikte ham veri modelinde klinik anlamlı türetilmiş göstergeler bulunmadığı için önerilen sistemde öznitelik mühendisliği ayrıca değerlendirilmiştir.

Ham modelin yüksek performansı, bu veri seti özelinde ham klinik/laboratuvar değişkenlerinin sınıflar arasında güçlü ayrıştırıcı bilgi taşıdığını göstermektedir. Ancak tıbbi makine öğrenmesi bağlamında yalnızca daha yüksek accuracy veya macro F1-score elde etmek, modelin klinik karar destek sistemine daha uygun olduğu anlamına gelmemektedir. Ham modelin güçlü sonucu, karşılaştırma için önemli bir referans çizgisi oluştururken; öznitelik mühendisliği ile desteklenen pipeline, kan basıncı değişimi, ultrafiltrasyon yükü ve kardiyovasküler stres gibi klinik olarak yorumlanabilir ara göstergeleri modele dahil ederek karar mantığının açıklanabilirliğini artırmayı hedeflemektedir (KDIGO CKD Work Group, 2024; National Kidney Foundation, 2015).

Önerilen Öznitelik Mühendisliği + Sızıntısız SMOTE + XGBoost Modeli

Önerilen modelde train-test ayrımı %80 eğitim ve %20 test olacak şekilde stratified olarak yapılmıştır. Eğitim setinde 4.000, test setinde 1.000 örnek yer almıştır. Öznitelik mühendisliği, encoding, scaling ve SMOTE işlemleri yalnızca eğitim seti üzerinde gerçekleştirilmiştir. Modelin eğitim setinde tüm sınıfları doğru ayırdığı, test setinde ise sınıflar arasında oldukça düşük hata yaptığı görülmüştür. Test setindeki 29 Mild örneğin 27'si doğru sınıflandırılmış, 237 Moderate örneğin 234'ü doğru sınıflandırılmış ve 734 Severe örneğin tamamı doğru sınıflandırılmıştır.

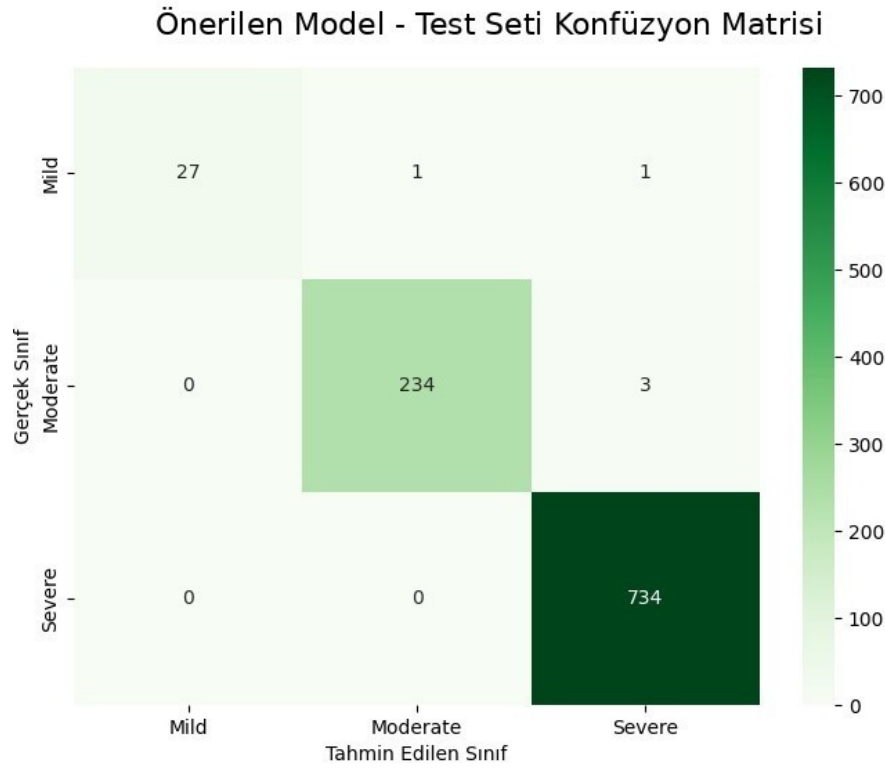
Bu çalışma kapsamında FE + train-only SMOTE + XGBoost yapısı, yalnızca ham performans metriklerini yükseltmeye çalışan bir model olarak değil, klinik bilgiyle zenginleştirilmiş ve veri sızıntısı riski azaltılmış bir sistem yaklaşımı olarak değerlendirilmiştir. Özellikle Mild sınıfının veri setinde düşük oranda temsil edilmesine rağmen önerilen modelin bu sınıfta 0.9310 recall ve 0.9643 F1-score üretmesi, pipeline'ın yalnızca çoğunluk sınıfına yönelmediğini ve azınlık sınıf duyarlılığını koruyabildiğini göstermektedir. Bu yönüyle önerilen sistem, ham modelden daha yüksek skor üretme iddiasından ziyade, klinik anlamlı öznitelikler, train-only yeniden örnekleme ve doğal dağılımda test prensiplerini birlikte uygulayan daha bütünlüklü bir modelleme çerçevesi sunmaktadır (Chawla et al., 2002; imbalanced-learn developers, n.d.).

Tablo 4. Önerilen modelin kilitli test seti sınıflandırma raporu

Sınıf/Metrik	Precision	Recall	F1-score
Mild	1.0000	0.9310	0.9643
Moderate	0.9957	0.9873	0.9915
Severe	0.9946	1.0000	0.9973

Model Karşılaştırması

Tablo 5'te referans çalışma değerleri Wafa ve ark. (2025) tarafından raporlanan sonuçlardan alınmış; bu çalışma sonuçları ise aynı veri seti üzerinde gerçekleştirilen yeniden üretim ve önerilen sızıntısız pipeline deneylerini göstermektedir.



Şekil 3. Önerilen modelin kilitli test seti karışıklık matrisi

Tablo 5. Referans çalışma ve önerilen pipeline sonuçlarının karşılaştırılması

Karşılaştırılan Senaryo	Değerlendirme Düzeni	Accuracy	Macro F1
Referans çalışma: XGBoost + SMOTE (Wafa ve ark., 2025)	70:30	0.8705	-
Yapılan referans çalışma: yakın yeniden üretim	60:40	0.9995	0.9968
Yapılan önerilen FE + train-only SMOTE + XGBoost pipeline	80:20	0.9950	0.9844

Karşılaştırma sonuçları, önerilen pipeline'ın değerlendirilmesinde tek ölçütün en yüksek test skoru olmadığını göstermektedir. Referans yönteme yakın yeniden üretim sonucunun yüksek performans üretmesi, veri setindeki ham değişkenlerin güçlü ayrışma taşıdığına işaret etmektedir. Buna karşın

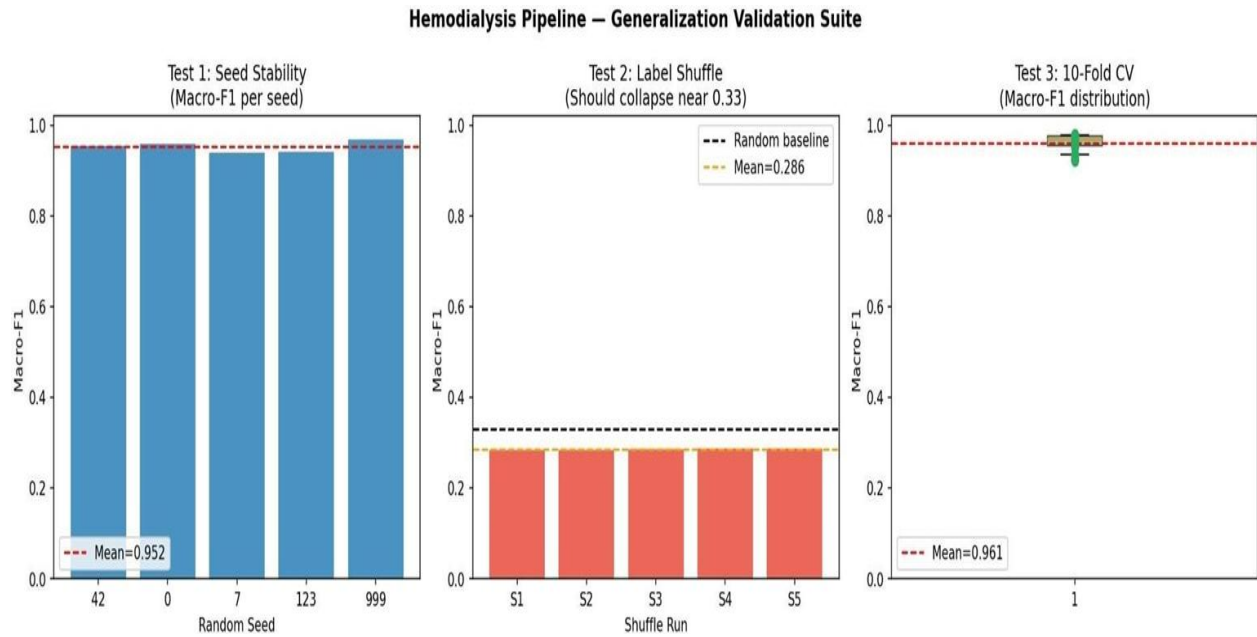
FE + train-only SMOTE + XGBoost pipeline, veri sızıntısı kontrolünü, azınlık sınıf duyarlılığını ve klinik öznitelik yorumlanabilirliğini aynı deneysel çatı altında birleştirdiği için bu çalışmada daha güvenli ve açıklanabilir sistem yaklaşımı olarak ele alınmıştır. Bu yorum, tıbbi yapay zeka çalışmalarında yalnızca accuracy odaklı değerlendirme yerine sınıf dengesizliği, macro metrikler, test setinin korunması ve klinik anlamlılık eksenlerinin birlikte değerlendirilmesi gerektiği görüşüyle uyumludur (Chawla et al., 2002; scikit-learn developers, n.d.).

Doğrulama Deneyleri

Seed stability testinde beş farklı random seed ile model yeniden eğitilmiş ve ortalama macro F1-score 0.9519 ± 0.0128 olarak hesaplanmıştır. 10-katlı çapraz doğrulamada ortalama macro F1-score 0.9611 ± 0.0188 seviyesinde kalmıştır. Label shuffle testinde ise etiketler karıştırıldığında macro F1-score 0.2855 seviyesine düşmüştür. Çok sınıflı bir problemde rastgele taban çizgisinin yaklaşık 0.33 civarında olması beklendiğinden, bu düşüş modelin hedef değişkenin yapısını gerçek klinik örüntüler üzerinden öğrendiğini ve basit bir target proxy'ye dayanmadığını desteklemektedir.

Tablo 6. Genellenebilirlik ve sanity check sonuçları

Doğrulama Testi	Tekrar/ Kat	Accuracy	Balanced Accuracy	Macro F1	ROC-AUC
Seed stability	5	0.9908	0.9888	0.9519	0.9995
10-fold cross-validation	10	0.9922	0.9905	0.9611	0.9992
Label shuffle	5	0.7276	0.3321	0.2855	0.5227



Şekil 4. Seed stability, label shuffle ve 10-fold cross-validation doğrulama görseli

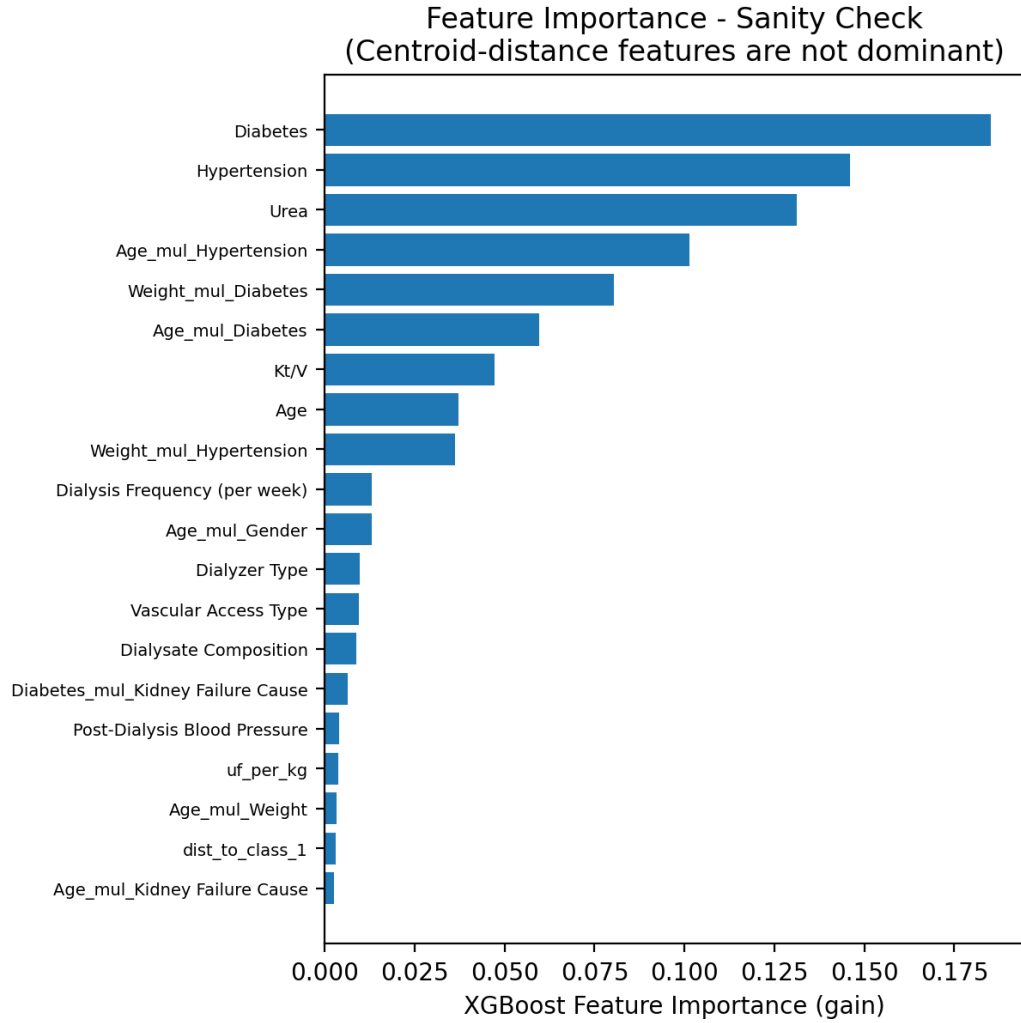
Feature Importance Analizi

Feature importance analizi, modelin en çok Diabetes, Hypertension, Urea, Age-Hypertension etkileşimi, Weight-Diabetes etkileşimi, Age-Diabetes etkileşimi, Kt/V ve Age değişkenlerinden faydalandığını göstermektedir. Bu dağılım klinik açıdan makul bir yapı sunmaktadır; çünkü diyabet ve hipertansiyon kronik böbrek yetmezliği ve hemodiyaliz sürecinde önemli komorbiditeler olup, üre ve Kt/V diyaliz yeterliliği ve toksin yükünü temsil eden temel göstergelerdir. Ayrıca sınıf merkezi uzaklığı değişkenlerinden yalnızca dist_to_class_1'in ilk 20 değişken arasına düşük önemle girmesi, bu özneliliklerin model üzerinde aşırı baskın olmadığına yönelik ek bir sanity check olarak yorumlanmıştır.

Tablo 7. Doğrulama pipeline'ında ilk 15 öznelilik önem değeri

Öznelilik	Importance
Diabetes	0.1852
Hypertension	0.1459
Urea	0.1311
Age_mul_Hypertension	0.1014
Weight_mul_Diabetes	0.0805
Age_mul_Diabetes	0.0596
Kt/V	0.0472
Age	0.0373
Weight_mul_Hypertension	0.0363
Dialysis Frequency (per week)	0.0130
Age_mul_Gender	0.0130
Dialyzer Type	0.0097
Vascular Access Type	0.0094
Dialysate Composition	0.0089
Diabetes_mul_Kidney Failure Cause	0.0063

Şekil 5. Feature importance sanity check grafiği



Sınıf Bazlı Değişken Aralıkları

Sınıf bazlı min-max değerleri, bazı değişkenlerin hedef sınıflarla güçlü ayrışma gösterdiğini ortaya koymaktadır. Özellikle Urea değişkeninde Mild sınıfı 15.20-29.90 aralığında kalırken, Moderate sınıfı 15.00-50.00 ve Severe sınıfı 15.00-100.00 aralığına ulaşmaktadır. Kt/V değişkeninde Mild ve Moderate sınıfları 1.20-2.00 aralığında iken Severe sınıfında alt sınır 1.00 seviyesine inmektedir. Bu dağılımlar, modelin yüksek başarısının yalnızca algoritmik karmaşıklıktan değil, veri setindeki sınıflar arası belirgin klinik/laboratuvar ayrışmalardan da kaynaklandığını göstermektedir.

Tablo 8. Seçilmiş ham ve türetilmiş özniteliklerin sınıf bazlı min-max aralıkları

Öznitelik	Mild Min/Max	Moderate Min/Max	Severe Min/Max
Age	18.00 / 75.00	18.00 / 75.00	18.00 / 89.00
Weight	51.40 / 119.60	50.00 / 120.00	50.00 / 120.00
Diabetes	0.00 / 0.00	0.00 / 1.00	0.00 / 1.00
Hypertension	0.00 / 0.00	0.00 / 1.00	0.00 / 1.00
Pre-Dialysis	100.00 / 178.00	100.00 / 179.00	100.00 / 179.00
Blood Pressure			
Post-Dialysis	80.00 / 139.00	80.00 / 139.00	80.00 / 139.00
Blood Pressure			
Heart Rate	60.00 / 99.00	60.00 / 99.00	60.00 / 99.00
Creatinine	1.04 / 9.86	1.02 / 9.99	1.00 / 9.99
Urea	15.20 / 29.90	15.00 / 50.00	15.00 / 100.00
Potassium	3.50 / 5.50	3.50 / 5.50	3.50 / 5.50
Hemoglobin	8.00 / 14.90	8.00 / 15.00	8.00 / 15.00
Albumin	3.00 / 5.00	3.00 / 5.00	3.00 / 5.00
Dialysis Duration (hours)	3.00 / 5.00	3.00 / 5.00	3.00 / 5.00
Dialysis Frequency (per week)	1.00 / 3.00	1.00 / 3.00	1.00 / 3.00
Kt/V	1.20 / 2.00	1.20 / 2.00	1.00 / 2.00
URR	65.00 / 75.00	65.00 / 75.00	65.00 / 75.00
Dry Weight (kg)	45.60 / 112.90	45.10 / 114.90	45.00 / 115.00
Fluid Removal Rate (ml/hour)	201.20 / 499.10	200.10 / 499.60	200.00 / 499.90
bp_reduction	-31.00 / 95.00	-37.00 / 98.00	-39.00 / 99.00
bp_reduction_pct	-0.30 / 0.54	-0.37 / 0.55	-0.39 / 0.55
rate_pressure_pr oduct	6000.00 / 17325.00	6060.00 / 17622.00	6060.00 / 17721.00
uf_per_kg	1.93 / 8.56	1.75 / 9.52	1.70 / 9.98
uf_per_hour	41.10 / 151.24	40.52 / 160.52	40.32 / 166.53
k_uf_interaction	866.45 / 2639.52	704.90 / 2695.14	703.15 / 2746.15
dist_to_class_0	638.98 / 6641.61	397.73 / 7195.21	316.99 / 8395.42
dist_to_class_1	452.85 / 6759.85	450.29 / 7105.54	309.11 / 8200.51
dist_to_class_2	447.08 / 7080.74	348.39 / 7066.18	243.43 / 7625.40

TARTIŞMA

Bu çalışmada elde edilen bulgular iki açıdan önemlidir. İlk olarak, referans çalışmada SMOTE sonrası görülen performans artışı, sınıf dengesizliği probleminin model başarısını doğrudan etkilediğini göstermektedir. Ancak SMOTE'un veri bölme işleminden önce uygulanması, test setinin doğal dağılımını bozabilir ve eğitim verisine test örneklerinden türeyen sentetik bilgi taşıyabilir. Bu durum literatürde veri sızıntısı riski olarak değerlendirilmektedir. Bu nedenle bu

çalışmada SMOTE yalnızca eğitim kümesine uygulanmış ve test kümesi doğal dengesiz haliyle korunmuştur. Böylece model başarısı daha gerçekçi bir değerlendirme ortamında ölçülmüştür (Chawla et al., 2002; imbalanced-learn developers, n.d.; scikit-learn developers, n.d.).

İkinci olarak, ham veriyle kurulan sızıntısız XGBoost modelinin bile çok yüksek başarı göstermesi, veri setinin hedef sınıflar bakımından güçlü biçimde ayrıştığını düşündürmektedir. Bu durum iki şekilde yorumlanmalıdır. Olumlu açıdan bakıldığında, üre, diyabet, hipertansiyon, Kt/V ve yaş gibi değişkenler hastalık şiddetiyle açık biçimde ilişkili görünmektedir. Daha temkinli açıdan bakıldığında ise veri setinin sentetik veya kural tabanlı biçimde üretilmiş olma olasılığı, dış doğrulama yapılmadan klinik genellenebilirlik iddiasında bulunmayı sınırlar. Bu nedenle çalışmada yüksek accuracy değeri tek başına yeterli kanıt olarak sunulmamış; label shuffle, seed stability, cross-validation ve feature importance sanity testleriyle model davranışı ayrıca incelenmiştir.

Ham veri XGBoost modelinin sayısal olarak önerilen FE + train-only SMOTE modelinden daha yüksek accuracy ve macro F1 değeri üretmesi ayrıca dikkatle yorumlanmalıdır. Bu sonuç, önerilen modelin değerini azaltmaktan ziyade, veri setinin ham değişkenler üzerinden güçlü ayrışma taşıdığını ortaya koymaktadır. Buna karşın FE + train-only SMOTE modeli, klinik olarak yorumlanabilir türetilmiş değişkenleri, azınlık sınıf odaklı öğrenme yaklaşımını ve sızıntısız modelleme sürecini birlikte değerlendiren optimize edilmiş bir sistem yaklaşımı olarak sunulmuştur. Dolayısıyla bu çalışmada “daha iyi sistem” ifadesi yalnızca en yüksek test skoru anlamında değil; test setini koruyan, klinik öznitelik mantığını modele dahil eden, sınıf dengesizliğini eğitim aşamasında yöneten ve karar destek bağlamında daha açıklanabilir bir pipeline ortaya koyan bütüncül yapı anlamında kullanılmaktadır.

Önerilen feature engineering yaklaşımı, ham değişkenleri klinik olarak daha anlamlı yapılara dönüştürmeyi amaçlamıştır. Kan basıncı düşüşü ve kan basıncı düşüş yüzdesi, diyaliz öncesi ve sonrası hemodinamik değişimi temsil etmektedir. Kilo başına ultrafiltrasyon ve saatlik ultrafiltrasyon, sıvı çekiminin hasta ağırlığı ve seans süresi ile ilişkili yükünü göstermektedir. Rate-pressure product, kalp hızı ile kan basıncını birleştirerek kardiyovasküler yükü temsil etmektedir. Potasyum-ultrafiltrasyon etkileşimi ise elektrolit durumu ile sıvı çekimi arasındaki ortak etkiyi modelin öğrenebilmesini sağlamaktadır. Bu tür özellikler, modelin yalnızca ham değerleri değil, değişkenler arası klinik ilişkileri de değerlendirmesine yardımcı olur (KDIGO CKD Work Group, 2024; National Kidney Foundation, 2015).

Modelin azınlık sınıf olan Mild üzerinde 0.9310 recall ve 0.9643 F1-score elde etmesi önemlidir. Çünkü sınıf dağılımında Mild örnekleri yalnızca %2.90 oranındadır. Sınıf dengesizliği olan tıbbi veri setlerinde çoğunluk sınıfı olan Severe sınıfında yüksek başarı elde etmek görece kolay olabilir; fakat azınlık sınıflarında dengeli performans göstermek modelin klinik karar destek potansiyeli açısından daha anlamlıdır. Bu nedenle macro F1-score değerinin 0.9844 olması, modelin yalnızca çoğunluk sınıfını ezberlemediğini göstermektedir.

SONUÇ

Bu çalışmada hemodiyaliz hastalarında hastalık şiddeti sınıflandırması için XGBoost tabanlı, veri sızıntısı riskini azaltan ve öznitelik mühendisliği ile desteklenen bir pipeline geliştirilmiştir. Referans makalede kullanılan XGBoost-SMOTE yaklaşımı yeniden değerlendirilmiş ve SMOTE'un train-test ayrımından önce uygulanmasının metodolojik risk taşıdığı vurgulanmıştır.

Önerilen pipeline'da train-test ayrımı başlangıçta yapılmış; feature engineering, encoding, scaling ve SMOTE yalnızca eğitim seti üzerinde uygulanmıştır.

Önerilen model kilitli test setinde 0.9950 accuracy ve 0.9844 macro F1-score elde etmiştir. Mild, Moderate ve Severe sınıflarında dengeli performans sağlanmış; özellikle azınlık sınıf olan Mild için yüksek F1-score elde edilmiştir. Ham veri XGBoost modeli sayısal olarak daha yüksek skor üretmiş olsa da, FE + train-only SMOTE + XGBoost pipeline klinik yorumlanabilirlik, sızıntısız yeniden örnekleme ve karar destek sistemine dönüştürülebilirlik açısından daha bütünlüklü bir sistem yaklaşımı sunmaktadır. Ek doğrulama deneyleri, model başarısının tek bir bölünmeye bağlı olmadığını ve etiketler karıştırıldığında performansın rastgele seviyeye düştüğünü göstermiştir. Sonuç olarak, hemodiyaliz şiddet sınıflandırmasında yüksek başarı elde etmek kadar, bu başarının sızıntısız, yeniden üretilebilir ve klinik olarak yorumlanabilir bir pipeline ile gösterilmesi de önemlidir.

KAYNAKÇA

- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357. <https://doi.org/10.1613/jair.953>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. <https://doi.org/10.1145/2939672.2939785>
- Chowdhury, S., Kutlu, Y., & Pekmezci, A. (2024). Diabetic Retinopathy Diagnosis Using Deep Learning. *Journal of Artificial Intelligence with Applications*.
- Garcia-Montemayor, V., et al. (2021). Predicting mortality in hemodialysis patients using machine learning analysis. *Clinical Kidney Journal*, 14(5), 1388-1395. GSK RD Solutions. (2023). Hemodialysis Realtime Hospital Dataset. Kaggle. <https://www.kaggle.com/datasets/gskrdsolutions/hemodialysis-realtime-hospital-dataset>
- imbalanced-learn developers. (n.d.). Common pitfalls and recommended practices. *imbalanced-learn documentation*. https://imbalanced-learn.org/stable/common_pitfalls.html
- KDIGO CKD Work Group. (2024). KDIGO 2024 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney International*, 105(4S), S117-S314. <https://doi.org/10.1016/j.kint.2023.10.018>
- National Kidney Foundation. (2015). KDOQI clinical practice guideline for hemodialysis adequacy: 2015 update. *American Journal of Kidney Diseases*, 66(5), 884-930. <https://doi.org/10.1053/j.ajkd.2015.07.015>
- Rankin, S., Han, L., Scherzer, R., Tenney, S., Keating, M., Genberg, K., Rahn, M., Wilkins, K., Shlipak, M., & Estrella, M. (2022). A machine learning model for predicting mortality

- within 90 days of dialysis initiation. *Kidney360*, 3(9), 1556-1565. <https://doi.org/10.34067/KID.0007012021>
- scikit-learn developers. (n.d.). Common pitfalls and recommended practices. scikit-learn documentation. https://scikit-learn.org/stable/common_pitfalls.html
- Sheng, K., Zhang, P., Yao, X., Li, J., He, Y., & Chen, J. (2020). Prognostic machine learning models for first-year mortality in incident hemodialysis patients: Development and validation study. *JMIR Medical Informatics*, 8(10), e20578. <https://doi.org/10.2196/20578>
- Wafa, M. T., Puspaningrum, E. Y., & Rahajoe, A. D. (2025). Analisis kinerja XGBoost pada data seimbang dan tidak seimbang dalam klasifikasi tingkat keparahan pasien hemodialisis. *JiIP - Jurnal Ilmiah Ilmu Pendidikan*, 8(12), 13928-13936. <https://doi.org/10.54371/jiip.v8i12.9998>