YENİDOĞANLARDA TOTAL BİLİRUBİNİN MAKİNE ÖĞRENMESİYLE TAHMİN EDİLMESİ

PREDICTION OF TOTAL BILIRUBIN LEVELS IN NEWBORNS USING MACHINE LEARNING

Kaan Kefal[1], Deniz İlhan Topcu[1], Taha Şahin[1], Aslıhan Abbasoğlu[2]

[1]İzmir Şehir Hastanesi, Tıbbi Biyokimya Bölümü, İzmir, Türkiye

[2]İzmir Şehir Hastanesi, Neonataloji Bölümü, İzmir, Türkiye

ORCID ID: KK: 0000-0001-9313-8735 **DİT:** 0000-0002-1219-6368 **TŞ:** 0009-0002-1960-0287 **AA:** 0000-0001-5435-646X

DOI: 10.52309/jaihs.1805577

MAKALE BİLGİLERİ

Anahtar Kelimeler:
Makine öğrenmesi, yenidoğan sarılığı, total bilirubin, gradient boosting, karar destek sistemi

ÖZET

Amaç: Yenidoğan döneminde sık görülen hiperbilirubinemi, zamanında tanı ve tedavi edilmediğinde ciddi nörolojik hasarlara neden olabilir. Bu çalışmanın amacı, makine öğrenmesi (ML) algoritmaları ile yenidoğanlarda total bilirubin düzeylerini tahmin eden modeller geliştirmek ve bu modellerin performansını değerlendirmektir.

Gereç ve Yöntem: İzmir Şehir Hastanesi'nde 318 yenidoğana ait 698 örnek retrospektif olarak analiz edilmiştir. Total bilirubin, hematokrit, doğum ağırlığı, gestasyonel yaş, yaş (gün) ve APGAR skoru gibi klinik-demografik veriler kullanılarak sekiz farklı ML algoritması (Gradient Boosting, Random Forest, Naive Bayes, Lojistik Regresyon, Yapay Sinir Ağı vb.) ile sınıflandırma modelleri oluşturulmuştur. Total bilirubin düzeyleri <12,5 ve ≥12,5 mg/dL olarak iki sınıfa ayrılmış, modellerin başarımı 10 kat çapraz doğrulama ile AUC, doğruluk ve F1 skoru gibi metrikler üzerinden değerlendirilmiştir. Model yorumlanabilirliği Decrease in AUC yöntemiyle analiz edilmiştir.

Bulgular: Gradient Boosting modeli test veri setinde %92 doğruluk, %0,90 F1 skoru ve 0,89 AUC değeri ile en başarılı model olarak belirlenmiştir. Düşük riskli (<12,5 mg/dL) olgular doğru tahmin edilirken, yüksek riskli (≥12,5 mg/dL) gruplarda hata oranları %90'ın üzerindedir. Değişken önem analizi, yaş (gün), doğum ağırlığı ve gestasyonel yaşın model üzerinde en belirleyici etkiye sahip olduğunu göstermiştir.

Sonuç: Makine öğrenmesi algoritmaları, özellikle düşük riskli yenidoğanlarda total bilirubin düzeylerini başarılı şekilde tahmin edebilmektedir. Ancak yüksek riskli grupların doğru tespiti için sınıf dengesizliğini azaltacak yöntemlerin (SMOTE, cost-sensitive learning vb.) kullanılması gerekmektedir.

ARTICLE INFO

Keywords:
Machine learning, neonatal jaundice, total bilirubin, gradient boosting, decision support system

ABSTRACT

Aim: Neonatal hyperbilirubinemia is a common condition that may lead to severe neurological damage if not diagnosed and treated promptly. This study aimed to develop machine learning (ML) models to predict total bilirubin levels in newborns and evaluate their performance.

Material and Method: A total of 698 samples from 318 newborns at İzmir City Hospital were retrospectively analyzed. Clinical and demographic variables, including total bilirubin, hematocrit, birth weight, gestational age, postnatal age (days), and Apgar scores, were used to develop classification models using eight ML algorithms (e.g., Gradient Boosting, Random Forest, Naive Bayes, Logistic Regression, Neural Networks). Total bilirubin levels were categorized as <12,5 mg/dL (low-risk) and ≥12,5 mg/dL (high-risk). Models were evaluated using 10-fold cross-validation and performance metrics such as AUC, accuracy, and F1 score. Model interpretability was assessed using the Decrease in AUC method.

Results: The Gradient Boosting model demonstrated the best performance on the test dataset with 92% accuracy, 0,90 F1 score, and an AUC of 0,89. While the models accurately predicted low-risk cases, their performance for high-risk (≥12,5 mg/dL) cases was limited, with error rates exceeding 90%. Feature importance analysis indicated that postnatal age (days), birth weight, and gestational age had the highest influence on predictions.

Conclusion: ML models, especially Gradient Boosting, can effectively predict low-risk total bilirubin levels in neonates. However, to improve the identification of high-risk cases, approaches addressing class imbalance (e.g., SMOTE, cost-sensitive learning) should be considered.

Makale Bilgisi | Article Information

Makale Türü | Article Type: Araştırma Makalesi | Research Article
Geliş Tarihi | Received: 17.10.2025
Kabul Tarihi | Accepted: 05.12.2025
Yayım Tarihi | Published: 30.12.2025

Çıkar çatışması
Yazarların çıkar çatışması bulunmamaktadır.
Finansman
Bu çalışma herhangi bir finansal destek almamıştır.
Teşekkürler
-
Etik Onay
-

Sorumlu Yazar | Correspondence Author Kaan Kefal

Address for Correspondence: Refik Şevket İnce Mah., 2148/11 Sok. No:1/11, 35540 Bayraklı, İzmir, Türkiye
Mail: kaankefal126@gmail.com

Yazar Katkıları
Motivasyon / Konsept: KK, TŞ
Çalışma Tasarımı: DİT, AA
Kontrol / Gözetim: DİT, AA
Veri Toplanması ve / veya İşlemesi: DİT, TŞ
Analiz ve / veya Yorum: AA
Literatür incelemesi:
Makalenin Yazılması:
Eleştirel İnceleme:

GİRİŞ

Gelişen perinatal bakım ve sağlık hizmetleri sayesinde, günümüzde daha küçük gestasyonel yaşa sahip yenidoğanların uygun şartlar altında yaşatılması mümkün hâle gelmiştir. Bu ilerlemeyle birlikte, prematüre ve düşük doğum ağırlıklı bebeklerin yaşam sürelerinde artış gözlenmekte; buna paralel olarak yenidoğan döneminde karşılaşılan klinik durumların sıklığında da artış yaşanmaktadır. Bu durumların başında ise, erken neonatal dönemde oldukça yaygın görülen sarılık gelmektedir. Yenidoğan sarılığı, total bilirubin düzeylerindeki artışla karakterize olup, zamanında tanı ve müdahale edilmediğinde ciddi nörolojik komplikasyonlara yol açabilmektedir. Bu nedenle, total bilirubin düzeylerinin doğru, güvenilir ve zamanında tespiti, tedavi sürecinin başarısı açısından kritik öneme sahiptir (1–5).

Total bilirubin düzeylerinin izlenmesinde yaygın olarak kullanılan yöntemler arasında transkütan bilirubin ölçümleri ile flebotomi yoluyla elde edilen serum örneklerinin analizi yer almaktadır. Transkütan cihazlar non-invaziv özellikleriyle avantaj sağlarken, ölçüm yöntemi ve cilt altı pigment dağılımına duyarlılıkları nedeniyle bazı klinik durumlarda sınırlı doğruluk gösterebilmektedir. Diğer yandan, serum bilirubin düzeylerinin ölçümü altın standart olarak kabul edilse de invaziv bir işlem oluşu enfeksiyon, hematom ve ağrı gibi komplikasyon risklerini beraberinde getirmektedir. Bu nedenle, klinik uygulamalarda her iki yöntemin avantaj ve dezavantajları göz önünde bulundurularak en uygun yaklaşımın seçilmesi gerekmektedir (6–10).

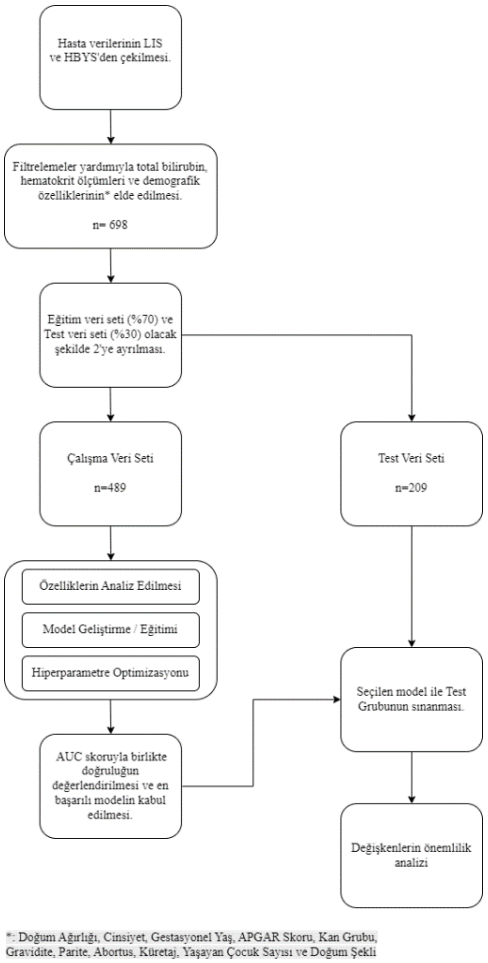
Yapay zeka algoritmaları ve makine öğrenmesi (ML), tıpta birçok farklı alanda yaygın olarak araştırılmaktadır (11). Özellikle ML tekniklerinin klinik karar süreçlerindeki kullanımı son yıllarda dikkat çekici biçimde artış göstermiştir. Klinik laboratuvar süreçlerinin optimize edilmesi, hasta sonuçlarının daha doğru değerlendirilmesi ve karar destek sistemlerinin geliştirilmesi gibi alanlarda ML uygulamaları önemli katkılar sunmaktadır. Ancak bu uygulamaların başarısı, model eğitiminde kullanılan verilerin doğruluğu, kalitesi ve klinik bağlamla uyumlu biçimde seçilmesine doğrudan bağlıdır (12–15).

Bu çalışmada, yenidoğanlarda total bilirubin düzeylerinin tahmin edilmesine yönelik çeşitli makine öğrenmesi algoritmaları uygulanmış; geliştirilen modeller çapraz doğrulama yöntemiyle değerlendirilmiştir. Amaç, klinik pratikte kullanılabilecek, güvenilir ve hassas öngörü sistemlerinin geliştirilmesine katkı sağlamak ve potansiyel olarak invaziv işlem gereksinimini azaltabilecek modeller oluşturmaktır. Böylece, hem sağlık hizmeti sunumunun etkinliği artırılmakta hem de hasta güvenliği ve konforu ön planda tutulmaktadır.

GEREÇ VE YÖNTEM

Çalışma planı

Çalışmanın genel iş akışı ve örnekleme süreci Şekil 1'de şematik olarak sunulmuştur.



Çalışma Popülasyonu

Bu çalışma, İzmir Şehir Hastanesi Kurumsal İnceleme Kurulu tarafından onaylandıktan sonra (Karar No: 2025 / 275), Temmuz 2025 ile Eylül 2025 tarihleri arasında İzmir Şehir Hastanesi'nde yürütülmüştür. Çalışmaya, aynı anda "Total Bilirubin" ve "Hematokrit" parametrelerinin ölçüldüğü ve ek olarak "Doğum Ağırlığı", "Cinsiyet", "Gestasyonel Yaş", "APGAR Skoru", "Kan Grubu", "Gravidite", "Parite", "Abortus", "Küretaj", "Yaşayan Çocuk Sayısı" ve "Doğum Şekli" gibi demografik ve klinik verilere de ulaşılan 318 farklı hastaya ait toplam 698 örnek dahil edilmiştir.

Örnekler, "Total Bilirubin" ve "Hematokrit" değerleri dikkate alınarak seçilmiştir. Makine öğrenmesi (ML) modeli geliştirme sürecinde, geniş bir total bilirubin aralığını kapsayacak biçimde ve herhangi bir altta yatan hastalık ayrımı yapılmaksızın hasta seçimi yapılmıştır. Tüm laboratuvar ölçümleri, numunelerin laboratuvara ulaşmasından itibaren en geç üç saat içinde tamamlanmıştır.

Analitik metotlar ve performans

Total bilirubin düzeyleri, laboratuvarımızda bulunan Roche Cobas c702 (Roche Diagnostics, Basel, Switzerland) otoanalizörü kullanılarak diazotize edilmiş sülfanilik asit ile gerçekleştirilen diazonium iyonu yöntemi ile ölçülmüştür. Serum total bilirubin için cut-off değeri 12,5 mg/dL olarak belirlenmiştir; bu eşik, literatürde fototerapi uygulama kriterleri ve klinik hiperbilirubinemi risk değerlendirmeleri temel alınarak klinik anlamlılık açısından seçilmiştir.

Hematokrit değerleri ise laboratuvarımızda bulunan Sysmex XN-2000 (Sysmex, Kobe, Japan) hematoloji analizörü ile eritrositlerin toplam atım yüksekliği esas alınarak belirlenmiştir.

Çalışma döneminde (01 Ocak 2025 – 31 Mart 2025) Cobas c702 cihazı için total bilirubin parametresinde Seviye 1 kontrolün ortalama değeri 0,95 mg/dL ve CV değeri %3,6; Seviye 2 kontrolün ortalama değeri 3,61 mg/dL ve CV değeri %4,26 olarak saptanmıştır.

Sysmex XN-2000 cihazı için hematokrit parametresinde; 43311101 LOT numaralı Seviye 1 kontrolün ortalama değeri %18,48 ve CV değeri %3,56, 43311102 LOT numaralı Seviye 2 kontrolün ortalama değeri %33,56 ve CV değeri %3,82, 43311103 LOT numaralı Seviye 3 kontrolün ortalama değeri ise %43,62 ve CV değeri %3,44 olarak belirlenmiştir.

EKK (eksternal kalite kontrol) değerlendirmeleri, Cobas c702 cihazı için RIQAS Monthly Clinical Chemistry, XN-2000 cihazı için ise RIQAS Monthly Haematology programları kullanılarak gerçekleştirilmiştir. İlgili üç aylık dönemde hematokrit için SDI (Z-skoru) değerleri Ocak -0,66, Şubat -0,65 ve Mart 0,01; total bilirubin için ise Ocak 0,23, Şubat 0,57 ve Mart 0,32 olarak saptanmış olup, tüm değerler ±2 sınırları içinde bulunmuş ve herhangi bir uygunsuzluk saptanmamıştır.

İlave olarak doğum ağırlığı, cinsiyet, gestasyonel yaş, APGAR skoru, kan grubu, gravidite, parite, abortus, küretaj, yaşayan çocuk sayısı ve doğum şekli gibi klinik ve demografik veriler Hastane Bilgi Sistemi (HIS) üzerinden retrospektif olarak elde edilmiştir.

APGAR'ın değerlendirilmesi

APGAR değerlendirmesi doğumdan sonraki 1. ve 5. dakikalarda olmak üzere iki kez yapılmaktadır. APGAR skoru; kalp atım hızı, solunum eforu, kas tonusu, refleks irritabilitesi ve cilt rengini içeren beş parametrenin her birine 0 ile 2 arasında puan verilerek hesaplanmasıdır. Elde edilen toplam skor 0 ile 10 arasında değişmektedir. Değerlendirmeler ilgili sağlık personeli tarafından standart prosedürlere uygun olarak gerçekleştirilmektedir. Çalışmada yalnızca kayıt altına alınmış ve eksiksiz veriye sahip olan olgular analize dahil edilmiştir (16,17).

Makine model öğreniminin geliştirilmesi

Bu çalışmadaki ML ile ilişkili tüm basamaklar Orange Data Mining yazılımı üzerinden gerçekleştirilmiştir (18). Total bilirubin düzeylerinin tahmini için tablo 1 de verilen parametreler kullanılarak klasifikasyon model geliştirilmesi

gerçekleştirilmiştir. Model geliştirme aşağıdaki adımlar izlenmiştir (19).

Özellik		n(%)	Minimum	Maksimum	Ortanca	Ortalama
Cinsiyet	Kadın	322 (46,1)				
	Erkek	376 (53,9)				
Yaş, gün	Kadın		0,00	103,00	10,00	18,00
	Erkek		0,00	153,00	7,00	19,00
Gestasyonel Yaş, hafta	Kadın		22,00	40,86	34,29	33,66
	Erkek		23,00	42,00	36,00	35,00
Total Bilirubin, mg/dL			0,15	23,60	3,74	4,94
Hematokrit, %			17,20	63,40	39,10	39,15
APGAR 1. Dakika			1,00	9	7,00	6,29
APGAR 5. Dakika			3,00	10,00	8,00	7,76
Doğum Ağırlığı, gr			300,00	5390,00	2500,00	2266,35
Gravidite			1,00	9,00		
Parite			1,00	6,00		
Abortus			0,00	5,00		
Küretaj			0,00	2,00		
Yaşayan			0,00	6,00		
Doğum Şekli			C/S (558)	NVY (140)		
Kan Grubu	Pozitif	A (219/42,0)	B (72/13,8)	AB (42/8,1)	0 (148/28,4)	
	Negatif	A (25/4,8)	B (3/0,6)	AB (0/0)	0 (12/2,3)	

Veri toplama

Bu çalışmaya, aynı anda istenmiş ve onaylanmış hematokrit ile total bilirubin testi sonuçlarına sahip olan ve bu testlerle ilişkili diğer değişkenleri de bulunan tüm yenidoğanlar dâhil edilmiştir. Dışlama ölçütü, bu parametrelerden herhangi birinin eksik olması olarak belirlenmiştir. Ancak veri kaybını en aza indirmek amacıyla, APGAR skoru verisi eksik olan örnekler çalışmadan çıkarılmamış; eksik veriler imputasyon yöntemiyle doldurulmuştur.

*Veri ön işleme ve özellik m*ühendisliği

Toplanan veriler total bilirubin düzeyleri benzer olacak şekilde öğrenme ve test veri seti olacak şekilde iki gruba ayrılmıştır. Bu ayrım öğrenme veri seti 419 (%60), test veri seti 279 (%40) olacak şekilde gerçekleştirilmiştir. Analizde süreklilik gösteren total bilirubin verileri tedavi sürecine yön verebilmesi ve hastanın klinik gidişatını etkileyebilmesi nedeniyle yüksek risk (≥12,5 mg/dL) ve düşük risk (<12,5 mg/dL) olmak üzere iki sınıfa ayrılmıştır. APGAR skorlarındaki verilerden 117 (%17)'sindeki eksik model bazlı imputasyon (simple tree) yardımıyla ortalama/en sık değer ile impute edilmiştir (20,21).

Model geliştirme

2 grup olacak şekilde yapılan sınıflandırmaya göre parametreleri kullanılarak Stochastic Gradient Descent, Random Forest, Yapay Sinir Ağları (Neural Network), Naive Bayes, Lojistik Regresyon, Gradient Boosting, CN2 Rule Induction ve AdaBoost algoritmaları kullanılarak modeller oluşturulmuştur. Her bir algoritma, model başarımı açısından çapraz doğrulama (k=10) ile değerlendirilmiştir. Benzer çalışma ve test veri setleri elde etmek amacıyla gruplar total bilirubin sınıfına göre tabakalı örnekleme yapılarak bölünmüş ve en iyi performans gösteren yöntemler belirlenmiştir.

Stochastic Gradient Descent algoritması, büyük veri setlerinde model parametrelerini hızlı ve etkili biçimde optimize etmek için kullanılmıştır. Random Forest, birden fazla karar ağacının çıktısını birleştirerek aşırı öğrenmenin önüne geçen güçlü bir topluluk yöntemi olarak uygulanmıştır. Yapay Sinir Ağları (Neural Network), verideki karmaşık ilişkileri öğrenebilmek amacıyla çok katmanlı yapısıyla modellenmiştir. Naive Bayes algoritması, değişkenler arasındaki koşullu bağımsızlık varsayımıyla hızlı ve basit sınıflandırmalar sağlamıştır. Lojistik Regresyon, ikili sınıflandırma için temel doğrusal modelleme yaklaşımı olarak kullanılmıştır. Gradient Boosting, zayıf sınıflayıcıların art arda eğitilmesiyle hataları minimize eden güçlü bir modelleme yöntemi sunmuştur. CN2 Rule Induction, veri üzerinden açık kurallar türeterek yorumlanabilir sınıflandırma sunmuştur. AdaBoost ise, yanlış sınıflandırılan örneklere daha fazla ağırlık vererek zayıf öğrenicilerin birleşimiyle güçlü bir sınıflayıcı oluşturmuştur (22–25). K = 10 olacak

şekilde çalışma veri setinde K-katlı çapraz doğrulama yöntemi uygulanmıştır (19).

Performans değerlendirmesi

Geliştirilen modellerin performansı çeşitli temel metrikler kullanılarak değerlendirilmiştir. Doğruluk (accuracy), modelin tüm tahminleri içerisindeki doğru sınıflandırma oranını göstererek genel başarıyı ölçmüştür. Hassasiyet (sensitivity / recall), pozitif vakaların doğru şekilde tanımlanma oranını yansıtarak özellikle klinik bağlamda önemli bir ölçüt olmuştur. F1 skoru, kesinlik (precision) ile hassasiyetin harmonik ortalaması olup sınıf dengesizliği durumlarında dengeli bir performans değerlendirmesi sunmuştur. İkili sınıflandırma görevleri için ROC eğrisi, farklı eşik değerlerinde doğru pozitif oranı ile yanlış pozitif oranı arasındaki ilişkiyi grafiksel olarak göstermiş; eğri altındaki alan (AUC) ise modelin sınıflar arasında ayırt etme yeteneğini nicel olarak ifade etmiştir (26). Ayrıca pozitif Prediktif Değer (PPV), yani pozitif olarak tahmin edilen vakaların gerçekten pozitif olma oranı ve Negatif Prediktif Değer (NPV), negatif olarak tahmin edilen vakaların gerçekten negatif olma oranı hesaplanmıştır. En iyi modelin seçilmesinde ise ilk önce ROC eğrisine bakılıp AUC hesaplanmış olup daha sonrasında doğrulukla beraber değerlendirilmiştir.

Seçilen modelin tahmin performansı, yeni ve daha önce görülmemiş (test veri seti) veriler üzerinde tekrar değerlendirilmiştir. Modelin yorumlanabilirliğini artırmak amacıyla, değişkenlerin tahmin sürecine katkısı AUC'de azalma (Decrease in AUC) yöntemiyle analiz edilmiştir. Bu yöntemde her bir değişkenin modelden çıkarılması veya rastgeleleştirilmesi sonrasında AUC performansındaki düşüş değerlendirilerek, ilgili değişkenin model için taşıdığı önem nicel olarak belirlenmiştir. Böylece modelde yer alan değişkenlerin göreli önemi ortaya konmuş ve karar sürecine en fazla etki eden değişkenler belirlenmiştir. (27–29).

BULGULAR

Tablo 1'de yenidoğan hastalara ait demografik özellikler ve makine öğrenimi modellerinin geliştirilmesinde kullanılan parametrelere ait tanımlayıcı istatistikler verilmiştir. Çalışmaya dahil edilen <12,5 mg/dL ve ≥12,5 mg/dL gruplarındaki katılımcı sayıları sırasıyla 658 ve 40 olarak belirlenmiştir. Bu istatistiklere bakıldığında doğum şekli açısından sezaryen doğumların (C/S) %80 olduğu gözlemlenmiştir. Ayrıca, çalışmaya dahil edilen verilerde AB Rh negatif kan grubuna sahip herhangi bir bireye rastlanmamıştır.

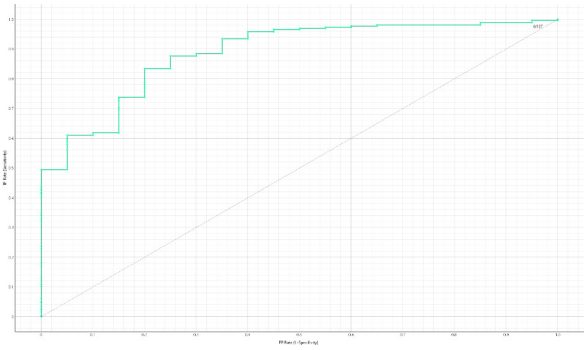
Tablo 2, geliştirilen modellerin hem çalışma hem de test veri setlerinde gerçek sonuçlara ne kadar yakın tahminler yaptığını ve bu tahminlerdeki oransal hata payını göstermektedir. Genel olarak çalışma veri setinde birçok modelin oldukça düşük hata oranlarıyla başarılı sonuçlar verdiği görülmektedir. Özellikle Gradient Boosting, Logistic Regression ve Stochastic Gradient Descent modelleri, çalışma veri setinde her iki bilirubin sınıfı için de %0 hata oranına ulaşmıştır. Test veri setinde ise modeller arasında performans farklılıkları belirginleşmektedir. Random Forest ve Logistic Regression modelleri <12,5 mg/dL sınıfında sıfıra yakın hata ile başarılı tahminlerde bulunurken, ≥12,5 mg/dL sınıfında bu modellerin başarısı ciddi şekilde azalmış, Logistic Regression ve Stochastic Gradient Descent modelleri bu sınıfı tamamen hatalı tahmin etmiştir. CN2 Rule Induction, Neural Network ve AdaBoost gibi modeller ise her iki sınıfta da daha dengeli ve kabul edilebilir hata oranları sergilemiştir. Bu durum, bazı modellerin özellikle ≥12,5 mg/dL gibi azınlık sınıfında yer alan total bilirubin değerlerini tahmin etmede yetersiz kaldığını, dolayısıyla dengesiz sınıflarla başa çıkma becerisinin model seçiminde kritik bir unsur olduğunu göstermektedir.

		Çalışma Veri Seti			Test Veri Seti		
Model	Sınıf	<12,5	≥12,5	Hata Oranı	<12,5	≥12,5	Hata Oranı
Gradient Boosting	<12,5	399	0	0%	256	3	1,16%
	≥12,5	0	20	0%	19	1	96%
Naive Bayes	<12,5	333	66	16,6%	226	33	12,7%
	≥12,5	5	15	25,0%	11	9	55%
Random Forest	<12,5	398	1	0,25%	259	0	0,00%
	≥12,5	12	8	60%	20	0	100%
CN2 Rule Induction	<12,5	399	0	0%	247	12	4,63%
	≥12,5	0	20	0%	17	3	85%
Logistic Regression	<12,5	399	0	0%	258	1	0,39%
	≥12,5	20	0	100%	20	0	100%
Neural Network	<12,5	397	2	0,50%	256	3	1,16%
	≥12,5	8	12	40%	17	3	85%
AdaBoost	<12,5	399	0	0%	248	11	4,24%
	≥12,5	0	20	0%	17	3	85%
Stochastic Gradient Descent	<12,5	399	0	0%	259	00	0%
	≥12,5	20	0	100%	20	0	100%

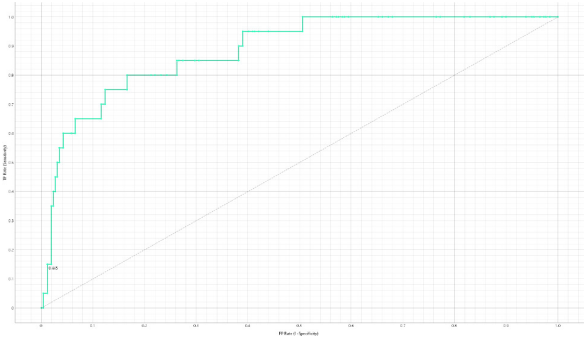
Tablo 3'te sunulan performans metrikleri incelendiğinde, test verisi üzerinde en yüksek AUC (0,89) değerine ulaşan Gradient Boosting modeli, aynı zamanda %92 doğruluk (CA) ve %90 F1 skoru ile genel sınıflama başarısı açısından en güçlü model olarak öne çıkmaktadır. Naive Bayes modeli, %87 F1 skoru ve %90 precision değeri ile güçlü sonuçlar üretmesine rağmen, AUC değerinin (0,80) daha düşük olması ayırt edicilik açısından sınırlı kaldığını göstermektedir. Random Forest, Neural Network ve Stochastic Gradient Descent modelleri %93 doğruluk oranına ulaşırken; Random Forest ve SGD modelleri ayrıca yüksek recall (%93) değerleriyle dikkat çekmektedir. Ancak Neural Network ve AdaBoost modellerinde AUC (<0,75) ve özellikle recall (<0,25) değerlerinin düşük olması, bu modellerin pozitif sınıfı ayırt etme başarısının zayıf olduğunu ortaya koymaktadır. CN2 Rule Induction ve Logistic Regression modelleri ise %90'ın üzerinde doğruluk ve %0,70'in üzerinde AUC değeri ile dengeli ve tutarlı performans sergilemiştir. Genel olarak AUC, modellerin sınıflar arasındaki ayırt edici gücünü yansıttığı için kritik bir ölçüt olarak değerlendirilmekte olup, bu bağlamda Gradient Boosting modelinin diğer modellere kıyasla üstün performans gösterdiği sonucuna ulaşılmaktadır.

Model	AUC	CA	F1	Prec	Recall
Gradient Boosting	0,89	0,92	0,90	0,89	0,92
Naive Bayes	0,80	0,84	0,87	0,90	0,84
Random Forest	0,75	0,93	0,89	0,86	0,93
CN2 Rule Induction	0,71	0,90	0,89	0,88	0,90
Logistic Regression	0,79	0,92	0,89	0,86	0,92
Neural Network	0,74	0,93	0,91	0,91	0,25
AdaBoost	0,55	0,90	0,89	0,88	0,13
Stochastic Gradient Descent	0,50	0,93	0,89	0,86	0,93
AUC: Area under ROC curve		CA: Classification accuracy		Prec: Precision	

Çalışmada kullanılan Gradient Boosting modelinin sınıflandırma başarımı, ROC analizi ile değerlendirilmiş ve sonuçlar grafik 1 ve grafik 2'de sunulmuştur. ≥12,5 değeri tahmin edilmesi gereken veri seti üzerinde elde edilen ROC eğrisi altında kalan alan 0,945 olarak hesaplanmış, bu da modelin yüksek ayırt edici güce sahip olduğunu göstermektedir. <12,5 değeri içeren veri seti üzerinde yapılan analizde ise AUC değeri 0.837 olarak bulunmuş olup, bu sonuç modelin bu grup üzerinde de oldukça iyi bir performans sergilediğine işaret etmektedir. Her iki durumda da elde edilen yüksek AUC değerleri, Gradient Boosting algoritmasının farklı veri dağılımları üzerinde etkili bir sınıflandırma yeteneğine sahip olduğunu ortaya koymaktadır.

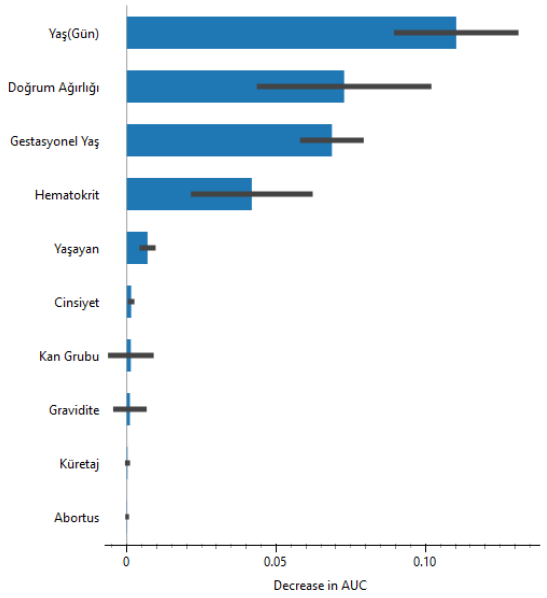


Gradient Boosting <12,5 ROC Grafik 1



Gradient Boosting ≥12,5 ROC Grafik 2

Gradient Boosting modeli ile yapılan analizde değişken önem düzeyleri grafik 3'te gösterilmiştir. Modele en fazla katkı sağlayan değişkenin Yaş (Gün) olduğu görülmekte olup, bunu sırasıyla Doğum Ağırlığı, Gestasyonel Yaş ve Hematokrit değişkenleri takip etmektedir. Bu değişkenlerin model performansı üzerinde anlamlı bir etkisi olduğu ve sınıflandırma başarısını önemli ölçüde artırdığı anlaşılmaktadır. Buna karşılık, Abortus, Küretaj, Gravidite, Cinsiyet ve Yaşayan gibi değişkenlerin modele katkısının oldukça sınırlı olduğu görülmüştür. Bu sonuçlar, modelin karar mekanizmasında özellikle yenidoğanın yaşı ve doğumla ilgili fizyolojik parametrelerin belirleyici olduğunu ortaya koymaktadır.



TARTIŞMA

Bu çalışma, yenidoğan bireylerde total bilirubin düzeylerinin tahminine yönelik çeşitli makine öğrenmesi (ML) algoritmalarının klinik karar destek süreçlerindeki potansiyel katkılarını değerlendirmeyi amaçlamıştır. Elde edilen bulgular, geliştirilen modellerin özellikle düşük bilirubin düzeylerini (<12,5 mg/dL) yüksek doğrulukla tahmin edebildiğini ve bu doğruluk sayesinde stabil seyir gösterebilecek hastaların önceden belirlenmesinde klinik ekipler için anlamlı bir destek aracı sunabileceğini göstermektedir. Nitekim test veri setinde Gradient Boosting modeli %92 doğruluk ve %90 F1 skoru ile düşük riskli sınıflarda başarılı bir performans ortaya koymuştur.

Ancak, modellerin ≥12,5 mg/dL düzeyindeki yüksek riskli vakaları sınıflandırmada belirgin yetersizlik sergilediği gözlemlenmiştir. Gradient Boosting algoritması bu grupta %96 oranında, Random Forest, Lojistik Regresyon ve Stochastic Gradient Descent modelleri ise %100 oranında hatalı sınıflandırma yapmıştır. Bu durum, yüksek riskli sınıfta pozitif prediktif değerin anlamlı ölçüde düştüğünü ve modellerin bu hasta grubunda güvenilir biçimde kullanılamayacağını göstermektedir. Bu bulgu, yalnızca doğruluk gibi tek bir performans metriğiyle model başarısının değerlendirilmesinin yetersiz kalabileceğini vurgulamaktadır. Özellikle sınıf dağılımının dengesiz olduğu klinik veri setlerinde, hassasiyet, özgüllük, F1 skoru, pozitif ve negatif prediktif değer gibi tamamlayıcı metriklerin dikkate alınması gerekmektedir.

Ek olarak, AUC değeri, sınıflar arasındaki ayırt ediciliği ölçen önemli bir gösterge olarak kabul edilmektedir. Çalışmamızda Gradient Boosting algoritmasının ≥12,5 mg/dL sınıfı için elde ettiği AUC değeri 0,828 olup, genel olarak kabul edilebilir bir ayırt edicilik düzeyini göstermektedir. Ancak bu sınıftaki %96'lık hata oranı, AUC'nin dengesiz sınıf dağılımına sahip alt gruplarda veya gözlem sayısı düşük sınıflarda tek başına yeterli bir ölçüt olmayabileceğini göstermektedir. Bu nedenle, modellerin yüksek riskli hastaları daha etkin şekilde tanıyabilmesi

için eğitim verisindeki örneklem dağılımının iyileştirilmesi gerekmektedir. Bu bağlamda, SMOTE (Synthetic Minority Over-sampling Technique), oversampling ve cost-sensitive learning gibi stratejilerin kullanımı sınıf dengesizliklerinin giderilmesi açısından önemli bir yaklaşım olarak öne çıkmaktadır. Tüm bu değerlendirmeler ışığında, çalışmanın klinik karar destek sistemlerine katkısı orta düzeyde olup, model performansının artırılmasına yönelik ileri optimizasyonlara gereksinim duyulmaktadır (30–32).

Neural Network ve AdaBoost modellerinde AUC (<0,75) ve özellikle hassasiyet (<0,25) değerlerinin düşük olması, bu modellerin pozitif sınıfı ayırt etme başarısının zayıf olduğunu ortaya koymuştur. Buna karşılık, Random Forest, Neural Network ve Stochastic Gradient Descent modelleri %93 gibi yüksek doğruluk oranlarına ulaşırken, Random Forest ve SGD yüksek recall (%93) değerleriyle dikkat çekmiştir. Bu modellerin yüksek doğruluklarına rağmen, özellikle yüksek riskli bilirubin seviyelerini tahmin etmedeki başarısızlıkları, sınıf dengesizliğinin performans üzerindeki belirleyici etkisini gözler önüne sermektedir.

Önemli Değişkenlerin Belirlenmesi ve Klinik Bağlantıları Modelde yer alan değişkenlerin önem dereceleri, AUC'de azalma yöntemiyle belirlenmiştir. Bu yöntem, her bir değişkenin modelden çıkarılması veya rastgeleleştirilmesi sonrasında AUC performansındaki düşüşü değerlendirerek, ilgili değişkenin model için taşıdığı önemi nicel olarak belirler. Analizler sonucunda, Gestasyonel Yaşın modeldeki en belirleyici faktör olduğu gösterilmiştir. Bunu sırasıyla Doğum Ağırlığı, Hematokrit ve Yaş (Gün) değişkenleri takip etmiştir. Bu bulgular, prematürelik ve düşük doğum ağırlığının yenidoğan hiperbilirubinemi gelişimindeki temel biyofizyolojik etkenlerden biri olduğu yönündeki literatürü desteklemektedir (33). Ayrıca, hematokrit düzeyinin, bilirubin üretiminin temel belirleyicisi olan hemoglobin metabolizmasıyla olan ilişkisi nedeniyle model performansına katkı sağlaması beklenen bir sonuçtur. Doğum şekli ve APGAR skorları gibi diğer klinik parametrelerin de tahmin sürecine katkı sağladığı, ancak Abortus, Küretaj, Gravidite, Cinsiyet ve Yaşayan Çocuk Sayısı gibi değişkenlerin modelde sınırlı etkiye sahip olduğu görülmüştür. Açıklanabilir yapay zeka tekniklerinden biri olan AUC'de azalma yöntemi, modelin yorumlanabilirliğini artırarak klinik karar destek sistemlerinde modele duyulan güveni arttırmaktadır (26).

Çalışmamız, makine öğrenmesinin sağlık alanındaki geniş potansiyelini bir kez daha ortaya koymaktadır. Literatürde belirttiği gibi, makine öğrenimi, büyük ve karmaşık veri kümelerinden anlamlı öngörüler çıkarma kapasitesiyle, istatistik ve bilgisayar biliminin kesişim noktasında yer almaktadır (12). Yenidoğan sarılığının yönetiminde, total serum bilirubin ölçümünün altın standart olmasına rağmen invaziv doğası, transkütan bilirubin ölçümlerini değerli bir non-invaziv alternatif haline getirmiştir. Ancak transkütan ölçümlerinin cilt rengi ve yüksek bilirubin konsantrasyonlarındaki sınırlılıkları literatürde mevcuttur. Bu çalışma, bu ölçüm kısıtlılıklarına ML modelleri ile çözüm arayışında önemli bir adım sunmaktadır.

Literatürde belirtildiği üzere, bilirubin seviyelerinin saatlik yaşa göre yorumlanması kritik öneme sahiptir. Bhutani ve ark. tarafından geliştirilen saat-spesifik nomogram, önemli hiperbilirubinemi riskini tahmin etmek için yaygın olarak kullanılan bir araçtır (4). Çalışmamızda yaş (gün) ve gestasyonel yaşın en önemli faktörler olarak belirlenmesi, bu nomogramların arkasındaki fizyolojik temelle uyumludur. Ayrıca, genetik varyantlar (örneğin UGT1A1 ve HMOX1 polimorfizmleri) ve bağlanmamış bilirubin seviyeleri gibi biyolojik belirteçlerin neonatal hiperbilirubinemi riskini etkilediği bilinmektedir (36,37). Serbest bilirubin seviyeleri, nörotoksisite için total serum bilirubin daha iyi bir gösterge olarak öne çıkmaktadır, ancak geniş çapta ulaşılabilir bir ölçüm yöntemi bulunmamaktadır (38). Gelecekteki ML modellerine bu tür ileri biyobelirteçlerin entegrasyonu, tahmin gücünü önemli ölçüde artırabilir.

Çalışmamızda Orange Data Mining yazılımının kullanılması, özellikle görsel programlama desteği ile ML modellerinin uygulanma-

sında erişilebilir ve etkin bir çözüm sunduğunu göstermektedir. Bu, veri madenciliği ve makine öğrenimi alanında eğitim ve uygulamalar için Orange'ın potansiyelini pekiştirmektedir.

Bu çalışmanın en önemli sınırlılıklarından biri, veri setindeki belirgin sınıf dengesizliğinin, özellikle ≥12,5 mg/dL düzeyindeki yüksek riskli hasta grubunun tahmininde ciddi performans kaybına yol açmasıdır. Bu sınıfta modellerin hata oranlarının %85–100 gibi oldukça yüksek değerlere ulaşması, geliştirilen sistemin klinikte yüksek riskli yenidoğanların güvenilir şekilde tanımlanmasında yetersiz kaldığını göstermektedir. Özellikle ≥12,5 mg/dL grubunun örneklem sayısının <12,5 mg/dL grubuna kıyasla sınırlı olması, modelin bu gruptaki tahmin başarısını doğrudan etkilemiş ve yüksek riskli olguların doğru sınıflandırılmasında güçlük yaratmıştır. Bu durum, mevcut haliyle modelin daha çok düşük riskli hastaların ayıklanmasında destekleyici bir araç olarak kullanılabileceğini, ancak tedavi kararlarını doğrudan yönlendirecek düzeyde klinik güven sunmadığını ortaya koymaktadır.

Bunun yanı sıra, geliştirilen modellerin harici (bağımsız) bir veri seti ile doğrulanmamış olması, elde edilen bulguların genellenebilirliği konusunda önemli bir kısıt oluşturmaktadır. Farklı merkezlerden ve farklı popülasyonlardan elde edilecek verilerle yapılacak harici validasyon çalışmaları, modelin gerçek klinik koşullardaki performansının güvenilir biçimde ortaya konulabilmesi açısından gereklidir.

Gelecekte yapılacak çalışmalarda, sınıf dengesizliğinin SMOTE, oversampling ve cost-sensitive learning gibi yöntemlerle giderilmesi ve çok merkezli harici validasyonların eklenmesi, modelin özellikle yüksek riskli hiperbilirubinemi olgularındaki tahmin başarısını artırarak klinik uygulanabilirliğini önemli ölçüde güçlendirecektir.

SONUÇ

Sonuç olarak çalışmanın makine öğrenmesi algoritmalarının yenidoğan sarılığı yönetiminde potansiyel klinik faydalarını ortaya koyarken, bu tür modellerin pratik kullanıma geçmeden önce çok yönlü ve dikkatli bir şekilde değerlendirilmesi gerektiğini göstermektedir. Özellikle yüksek riskli vakaların doğru bir şekilde tespit edilmesi için modellerin hassasiyetinin ve pozitif prediktif değerlerinin artırılmasına yönelik daha fazla araştırma ve geliştirme çabası gerekmektedir.

REFERANSLAR

1. Stevenson DK, Vreman HJ, Wong RJ. Bilirubin Production and the Risk of Bilirubin Neurotoxicity. *Semin Perinatol*. 2011;35(3):121-6. doi:10.1053/j.semperi.2011.02.005
2. Kirk JM. Neonatal jaundice: a critical review of the role and practice of bilirubin analysis. *Ann Clin Biochem*. 2008;45(Pt 5):452-62. doi:10.1258/acb.2008.008076
3. Hansen TWR, Wong RJ, Stevenson DK. Molecular Physiology and Pathophysiology of Bilirubin Handling by the Blood, Liver, Intestine, and Brain in the Newborn. *Physiol Rev*. 2020;100(3):1291-346. doi:10.1152/physrev.00004.2019
4. Bhutani VK, Vilms RJ, Hamerman-Johnson L. Universal bilirubin screening for severe neonatal hyperbilirubinaemia. *J Perinatol*. 2010;30(Suppl):S6-15. doi:10.1038/jp.2010.98
5. Carbonell X, Botet F, Figueras J, Riu-Godó A. Prediction of hyperbilirubinaemia in the healthy term newborn. *Acta Paediatr*. 2001;90(2):166-70. doi:10.1080/080352501300049343
6. Amin SB, Lamola AA. Newborn Jaundice Technologies: Unbound Bilirubin and Bilirubin Binding Capacity in Neonates. *Semin Perinatol*. 2011;35(3):134-40. doi:10.1053/j.semperi.2011.02.007
7. Moncrieff G. Bilirubin in the newborn: Physiology and pathophysiology. *Br J Midwifery*. 2018;26(6):362-70. doi:10.12968/bjom.2018.26.6.362
8. Hsia DYY, Allen FH, Diamond LK, Gellis SS. Serum bilirubin levels in the newborn infant. *J Pediatr*. 1953;42(3):277-85. doi:10.1016/S0022-3476(53)80182-4
9. Bhardwaj K, Locke T, Biringer A, Booth A, Darling EK, Dougan S, et al. Newborn Bilirubin Screening for Preventing Severe Hyperbilirubinemia and Bilirubin Encephalopathy: A Rapid Review. *Curr Pe-*