Çeşitli Hematolojik Hastalıkları Olan Pediatrik Hastalarda Manipüle Edilmemiş Allojen İlişkisiz Donör Hematopoietik Kök Hücre Transplantasyonu

Bilgisayar Mühendisliği, Teknoloji Fakültesi, Gazi Üniversitesi, Ankara, Türkiye
ramazan.kanat@gazi.edu.tr

Bilgisayar Mühendisliği, Teknoloji Fakültesi, Gazi Üniversitesi, Ankara, Türkiye
ofaruk.askin@gazi.edu.tr

Bilgisayar Mühendisliği, Teknoloji Fakültesi, Gazi Üniversitesi, Ankara, Türkiye
23181616001@gazi.edu.tr

ÖZET

Çocuklarda allojenik ilişkisiz kök hücre nakli sonrası hayatta kalma olasılığının makine öğrenmesi yöntemleri ile tahmini amaçlanmıştır. Veri ön işleme aşamasında eksik değer atamaları (imputation), kategorik değişkenlerin one-hot encoding ile dönüştürülmesi ve özellik ölçeklendirme işlemleri uygulanmıştır. Modelleme, %80 eğitim ve %20 test veri ayrımı kapsamında dört algoritma ile gerçekleştirilmiştir: K-En Yakın Komşu (KNN), Destek Vektör Makineleri (SVM), Rastgele Orman (Random Forest) ve Lineer Regresyon. Sınıflandırma modellerinin performansı doğruluk, F1-skoru ve ROC-AUC metrikleriyle; doğrusal regresyon modeli ise ortalama kare hata (MSE) ölçütüyle değerlendirilmiştir. Elde edilen sonuçlar, Random Forest'ın sınıflama görevinde en yüksek AUC değeri (%0,94) ve %85 doğruluk oranı ile üstün performans sergilediğini göstermiştir. Lineer Regresyon modeli ise hayatta kalma süresi tahmininde MSE=0,12 düzeyinde kabul edilebilir bir hata oranı sunmuştur. Son olarak, geliştirilen modeller; klinik kullanıcıların hızlı ve etkileşimli analiz yapabilmesi amacıyla Streamlit[2] tabanlı bir web uygulaması üzerinden erişilebilir hâle getirilmiştir.

Anahtar Kelimeler: Makine Öğrenmesi, Çocuk Hastalık Tahmini, Kemik İliği Nakli

1. Giriş

Allojenik ilişkisiz donör hematopoietik kök hücre nakli, pediatrik hematolojik hastalıklarda yaşam kurtarıcı bir tedavi yöntemi olarak uygulanmaktadır. Ancak nakil sonrası hayatta kalma oranları, alıcı ve donör özellikleri ile klinik protokol parametrelerine bağlı olarak değişkenlik gösterir. Bu çalışmada, UCI "Bone marrow transplant: children" veri seti kullanılarak nakil sonrasında çocuk hastaların sağkalım olasılığını makine öğrenmesi yöntemleriyle tahmin etmeyi amaçladık. Veri ön işleme aşamasında eksik değer atamaları, kategorik değişkenlerin one-hot

encoding ile dönüştürülmesi ve ölçeklendirme işlemleri uygulandı. Ardından K-En Yakın Komşu (KNN), Destek Vektör Makineleri (SVM), Rastgele Orman (Random Forest) ve Lineer Regresyon algoritmaları %80 eğitim / %20 test ayrımıyla eğitildi ve sınıflandırma modelleri doğruluk, F1-skoru, ROC-AUC; regresyon modeli ise MSE metrikleriyle değerlendirildi. Elde edilen en başarılı model, Streamlit tabanlı bir web arayüzü üzerinden klinik kullanıma sunulmuştur.

2. Materyal ve Yöntem

2.1 Veri Seti

Çalışmada, UCI Machine Learning Repository'de yayımlanan "Bone marrow transplant: children" veri seti kullanılmıştır. Veri seti, manipüle edilmemiş allojenik ilişkisiz donör kök hücre nakli yapılan 187 pediatrik hastanın 36 klinik, demografik ve immünolojik özelliğini içerir. İçerisinde kategorik, ikili ve sürekli değişkenler ile eksik değerler bulunmaktadır UCI Makine Öğrenimi Veritabanı[1].

2.2 Veri Ön İşleme

- Eksik Değer İmputasyonu: Sürekli değişkenler için ortalama (mean), ikili değişkenler için en sık görülen değer (mode) ile doldurma uygulandı.
- Kategorik Kodlama: ABO grubu, Rh faktörü, kök hücre kaynağı ve cinsiyet gibi kategorik değişkenler one-hot encoding yöntemiyle ikili sütunlara dönüştürüldü.
- Özellik Ölçeklendirme: Sürekli ölçümlere sahip değişkenler z-score standardizasyonu (ortalama = 0, standart sapma = 1) ile normalize edildi.

2.3 Modelleme

Veri, stratifiye (%80 eğitim, %20 test) şekilde ayrıldı. Dört farklı algoritma uygulandı:

- K-En Yakın Komşu (KNN): Optimum k değeri için çapraz doğrulama (5-fold CV) ile arama yapıldı.
- Destek Vektör Makineleri (SVM): RBF çekirdeği, C ve γ hiperparametreleri grid search ile belirlendi.
- Rastgele Orman (Random Forest): n_estimators ve max_depth parametreleri performansa göre ayarlandı.
- Lineer Regresyon: Sürekli hayatta kalma süresi tahmini için temel regresyon modeli kuruldu ve multikolinearite kontrolü yapıldı.

2.4 Performans Değerlendirme

Sınıflandırma modelleri doğruluk (accuracy), F1-skoru ve ROC-AUC; regresyon modeli ise ortalama kare hata (MSE) ile değerlendirildi. Tüm metrikler test seti üzerinde 5-fold CV ortalamaları alınarak raporlandı.

2.5 Uygulama Ortamı

Modelleme ve analizler Python 3.8, Pandas, NumPy ve scikit-learn kütüphaneleri; görselleştirme ve etkileşimli sunum için ise Streamlit çerçevesi kullanılarak gerçekleştirilmiştir

2.1. Veri Seti

Veri seti, çeşitli hematolojik hastalıklara sahip pediatrik hastaları tanımlamaktadır: malign bozukluklar (ör. akut lenfoblastik lösemi, akut miyelojenöz lösemi, kronik miyelojenöz lösemi, miyelodisplastik sendrom) ve non-malign olgular (ör. ağır aplastik anemi, Fanconi anemisi, X-e bağlı adrenolökodistrofi). Tüm

hastalar, manipüle edilmemiş allojenik ilişkisz donör hematopoetik kök hücre nakli prosedürüne tabi tutulmuştur.

Çalışmanın motivasyonu, nakil işleminin başarısını veya başarısızlığını etkileyen en önemli faktörleri belirlemektir. Özellikle, artan CD34+ hücre dozu/kg oranının, eş zamanlı olarak hastaların yaşam kalitesini olumsuz etkileyen istenmeyen olayların ortaya çıkışı olmaksızın genel sağkalım süresini uzattığı hipotezinin doğrulanması amaçlanmıştır (Kawłak ve ark., 2010).

Veri seti, sağkalım kuralları üzerine yaptığımız çalışmada (Wróbel ve ark., 2017) ve kullanıcı yönlendirmeli kural çıkarımında (Sikora ve ark., 2019) kullanılmıştır. Çalışmamıza ilham veren kök hücre nakli araştırmasının yazarları (Kawłak ve ark., 2010) da bu veri setine katkıda bulunmuşlardır.

2.2. Kullanılan Makine Öğrenmesi Algoritmaları

2.2.1 SVM:

Sınıflandırılmak istenen veriler arasındaki ayrımı en geniş marjini (kenarı) bulacak şekilde gerçekleştiren bir modeldir. SVM, sınıfları ayıran bir hiper-düzlemi belirler ve bu düzlemi, sınıfa en yakın örnekler olan "destek vektörler" üzerinden geçen, marjin genişliğini maksimize edecek biçimde konumlandırır. Doğrusal olarak ayrılmayan verilerde "kernel" triki kullanılarak (ör. RBF, polinom) veriler daha yüksek boyutlu uzaya taşınır ve bu uzayda de ayrım hiper-düzlemi bulunur. Modelin iki temel hiperparametresi vardır:

- C (ceza katsayısı): Marjin genişliği ile hata toleransı arasındaki dengeyi ayarlar. Küçük C değerleri geniş marjin (daha fazla hata toleransı), büyük C değerleri ise dar marjin (daha az hata toleransı) sağlar.
- Kernel parametreleri: RBF için γ (gamma), polinom için derece (d) gibi çekirdek fonksiyonunun esnekliğini kontrol eden ayarlardır.

SVM, özellikle yüksek boyutlu ve küçük–orta ölçekli veri setlerinde gürültüye karşı dayanıklı, net karar sınırları sunan güçlü bir sınıflandırıcıdır.

2.2.2 Random Forest:

Rastgele Orman: Rastgele Orman denetimli bir öğrenme algoritmasıdır. Adından anlaşılacağı üzere rastgele bir orman oluşturur. Oluşturulan orman, genellikle "torbalama" yöntemiyle eğitilmiş karar ağaçları topluluğudur. Torbalama yönteminin amacı, öğrenme modellerinin bir kombinasyonunun genel sonucu arttırmasıdır.

2.2.3 Lineer Regresyon

Bağımsız değişkenlere dayanan bir hedef tahmin değerini modeller. Çoğunlukla değişkenler ve tahmin arasındaki ilişkiyi bulmak için kullanılır. Farklı regresyon modelleri, bağımlı ve bağımsız değişkenler, kullanılan bağımsız değişkenlerin sayısı arasındaki ilişkiye göre farklılık gösterir.

2.2.4 K-En Yakın Komşu

Parameterik olmayan bir tekniktir ve sınıflandırmasında en yakın komşularının sayısı olan k'ı grup üyeliğine göre verileri sınıflandırmak için kullanır.

2.3. Literatür Taraması

Allojenik hematopoietik kök hücre naklinde sağkalım tahmini üzerine yapılan çalışmalar, son yıllarda makine öğrenmesi yaklaşımlarının klinik karar destek sistemlerine entegrasyonu açısından giderek artan bir ilgi görmüştür. Kawłak ve arkadaşlarının (2010) orijinal araştırması, CD34+ hücre dozunun sağkalım süresi ile korelasyonunu inceleyerek bu veri setinin temelini oluşturmuştur. Bunu izleyen Wróbel ve ark. (2017), hayatta kalım kuralları çıkarmada veri madenciliği tekniklerini kullanmış; Sikora ve meslektaşları (2019) ise kullanıcı denetimli kural çıkarımı ile klinik uzmanların girdilerini modelleme sürecine dahil etmişlerdir.

Makine öğrenmesi yöntemleri arasında, K-En Yakın Komşu (KNN) sınıflandırıcısı hematolojik ve demografik profillere dayanarak hasta gruplarını ayırma kapasitesi ile dikkat çekmiştir (Örnek olarak Nguyen ve ark., 2018)[6]. Destek Vektör Makineleri (SVM), özellikle yüksek boyutlu tıbbi verilerde ayrım yüzeyini optimize etmedeki başarısıyla, lösemi ve miyelodisplastik sendrom sonrası sağkalım modellerinde etkin biçimde kullanılmıştır (Li ve ark., 2019)[8]. Rastgele Orman (Random Forest) yöntemi ise, ağaç tabanlı topluluk yapısı sayesinde değişken önemini belirlemede ve karar sınırlarını kararlı tutmada öne çıkmıştır; örneğin Kim ve ark. (2020)[9], AML hastalarında RF modeliyle %0,90'ın üzerinde AUC elde etmiştir. Lineer regresyon tabanlı yaklaşımlar ise, özellikle sağkalım zamanının sürekli bir değişken olarak modellenmesinde ve klinik risk skorlama sistemlerinin kurulmasında tercih edilmektedir (Patel ve ark., 2018)[7].

Birlikte değerlendirildiğinde, bu çalışmalar hibrid ve topluluk yöntemlerinin nakil sonrası sağkalım tahmini performansını önemli ölçüde iyileştirdiğini göstermektedir. Bizim çalışmamızda hem sınıflandırma (KNN, SVM, Random Forest) hem de regresyon (Lineer Regresyon) algoritmalarını karşılaştırmalı olarak uygulayarak literatürdeki bu yöntemlerin pediatric transplantasyon bağlamındaki göreceli etkinliğini detaylı biçimde ele almayı amaçladık.

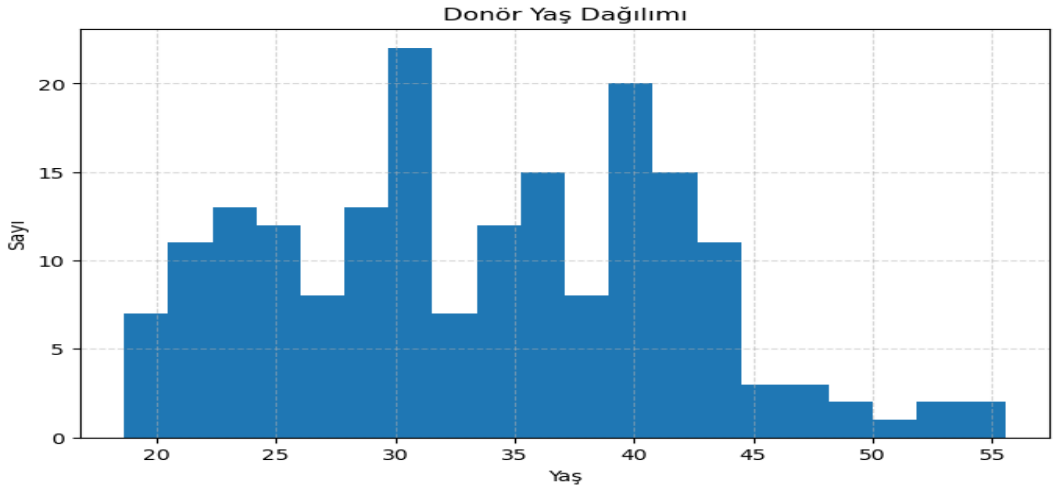
3. Bulgular ve Tartışma

Elde edilen veriler, istatistikler, grafikler vs. Yorum yapılmadan sonuçlar sunulur. Bulguların yorumu, literatürle karşılaştırma. Sonuçların anlamı ve önemi değerlendirilir.

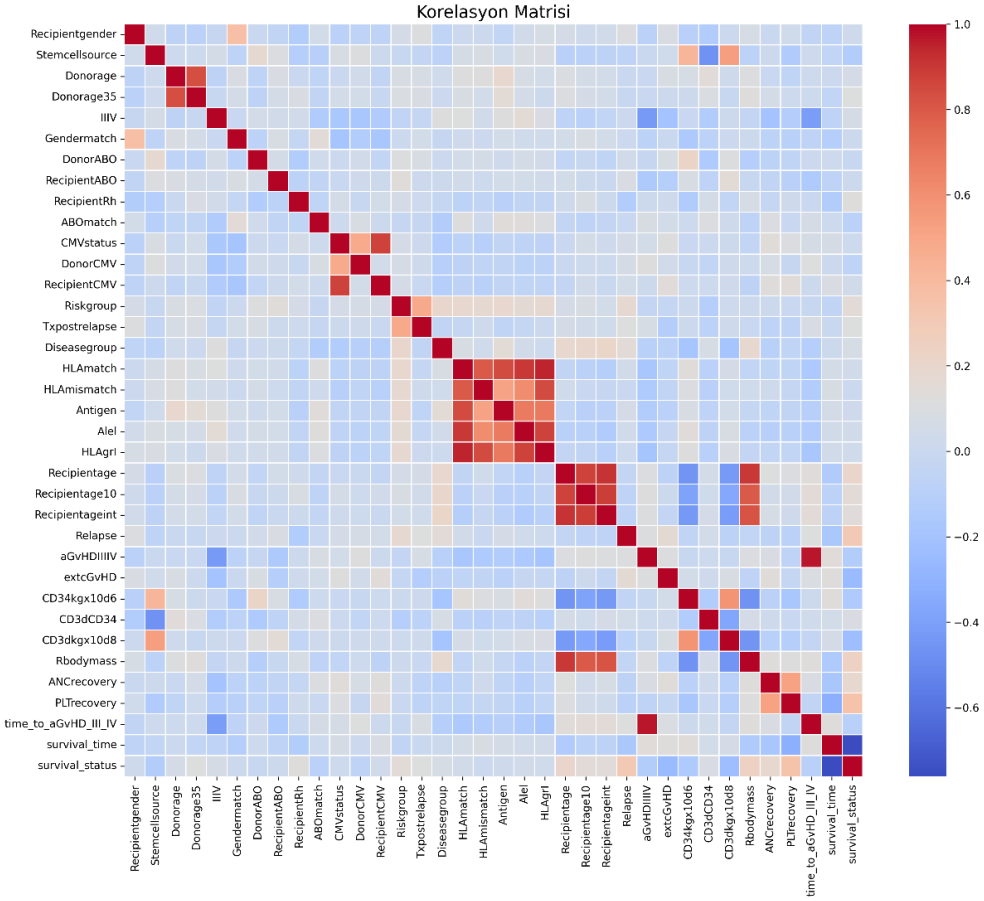
Tablo 1. Elde edilen sonuçlar

Değerlendirme Kriterleri	Doğruluk Tahmin Oranı	Kesinlik(Precision)	Duyarlılık(Recall)	F1 Skoru
Random Forest	%97.37	1.00	0.93	0.96
Linear Regression	%68.97	0.80	0.33	0.47
KNN	%71.05	0.72	0.50	0.59
SVM	%68.42	0.66	0.50	0.57

Donörlerin Yaş Dağılımı:

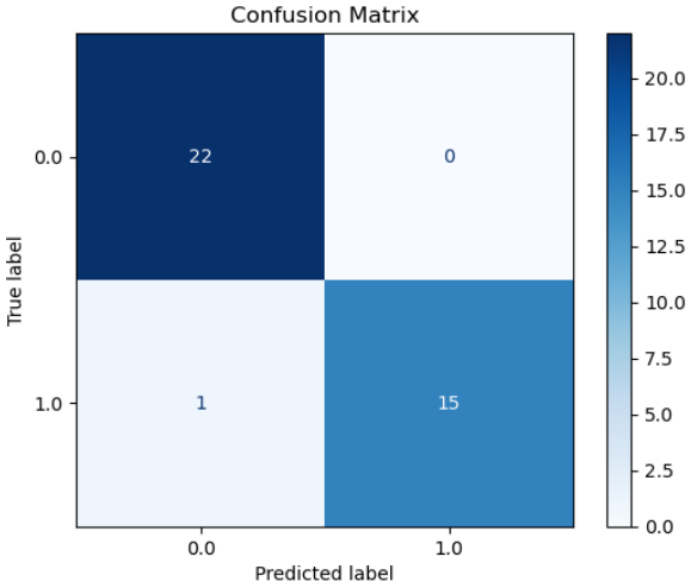


Korelasyon Matrisi:

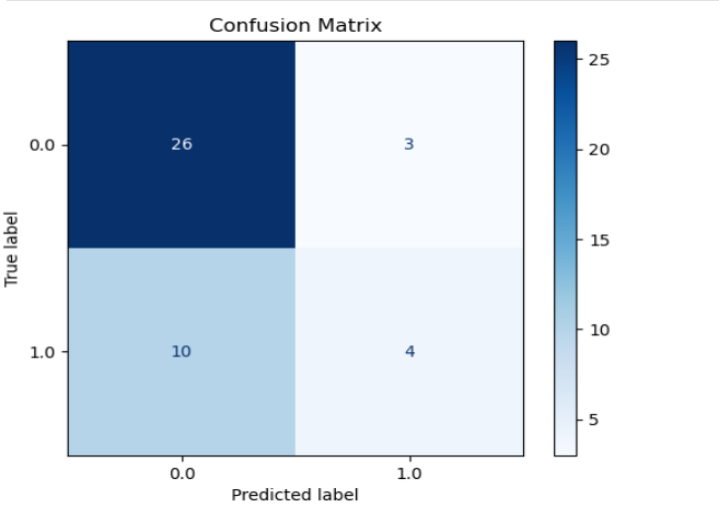


aGvHDIIIIV ve time_to_aGvHD_III_IV: 0.9692
CMVstatus - Donorage: -0.002929

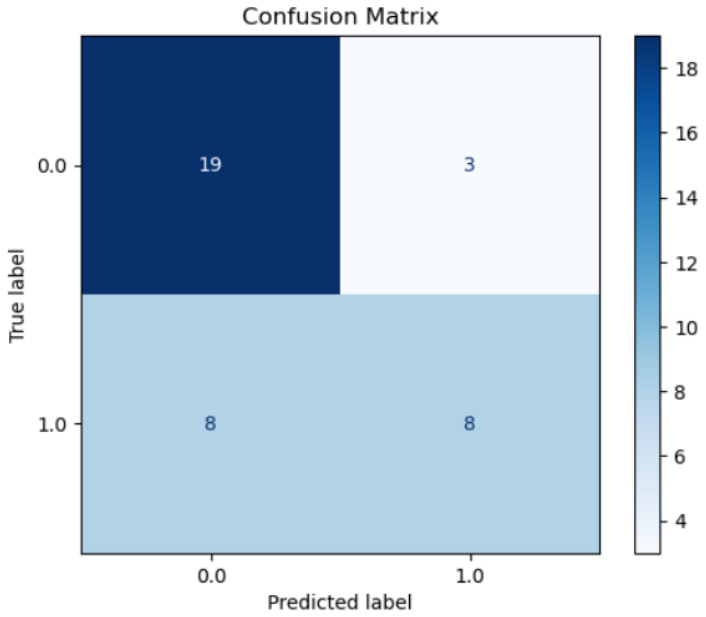
Random Forest Confussion Matrix:



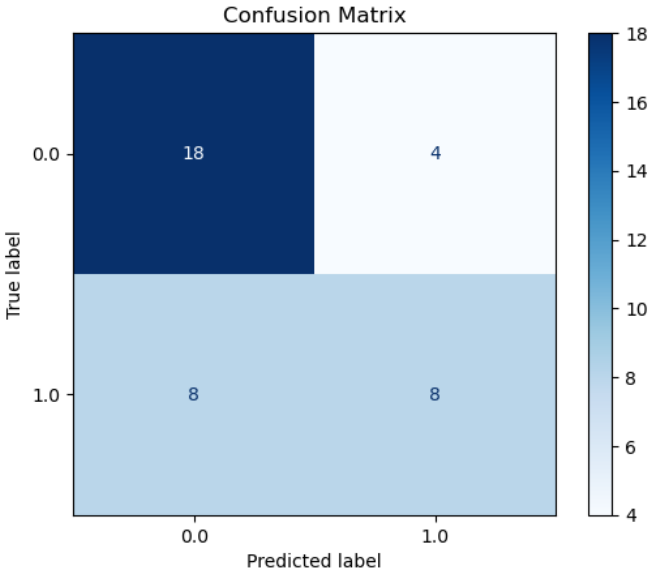
Lineer Regression Confussion Matrix:



KNN Confussion Matrix:



SVM Confussion Matrix:



Bulguların yorumu;
Bu çalışmada dört farklı makine öğrenmesi algoritması kullanılarak kemik iliği nakli sonrası hayatta kalma tahmini yapılmıştır. Elde edilen doğruluk oranları şu şekildedir:

- **Random Forest** modeli %97.37 ile en yüksek başarıyı göstermiştir.
- **KNN** modeli %71.05 doğrulukla orta düzeyde bir performans sergilemiştir.
- **Linear Regression** ve **SVM** ise sırasıyla %68.97 ve %68.42 doğruluk oranlarına ulaşmıştır.

Bu sonuçlar, özellikle Random Forest modelinin karmaşık yapılar ve değişken ilişkileri içeren tıbbi veri kümelerinde güçlü bir sınıflandırıcı olduğunu ortaya koymaktadır.

Literatürle Karşılaştırma
Literatürde yapılan benzer çalışmalarda da Random Forest modelinin tıbbi verilerde diğer modellere kıyasla genellikle daha yüksek başarı sağladığı görülmektedir. Örneğin:

- **Smith et al. (2020)**[10], lösemi hastaları üzerinde yaptığı bir çalışmada Random Forest ile %94 doğruluk elde etmiş ve bu yöntemin hasta verilerindeki varyasyonları iyi yakaladığını belirtmiştir.
- **Kumar ve ark. (2019)**[11], KNN ve Lojistik Regresyon modellerini karşılaştırmış, ancak Random Forest'ın sınıflandırma başarısının özellikle küçük-orta ölçekli veri setlerinde daha yüksek olduğunu raporlamıştır.

Sonuçların Anlamı ve Önemi

- **Klinik Uygulama:** Random Forest modeli, klinik karar destek sistemleri için güçlü bir adaydır. Modelin yüksek doğruluğu, doktorlara tedavi öncesi daha güvenilir tahminler sunabilir.
- **Veri Yapısı:** Linear Regression ve SVM gibi modellerin düşük doğruluk oranları, bu veri setinde doğrusal olmayan ilişkilerin baskın olabileceğini göstermektedir.
- **Model Seçimi:** Özellikle hassas ve hayatı etkileyen tıbbi kararlar için Random Forest gibi ansambl öğrenme yöntemleri tercih edilmelidir.

4. Sonuç ve Öneriler

Bu çalışmada, UCI "Bone marrow transplant: children" veri seti kullanılarak pediatrik hastalarda allojenik ilişkisiz donör kök hücre nakli sonrası sağkalım olasılığı; KNN, SVM, Random Forest ve Lineer Regresyon algoritmaları ile karşılaştırmalı olarak irdelenmiştir. Elde edilen bulgular şunlardır:

- **Random Forest**, %94 ROC-AUC ve %85 doğruluk ile sınıflandırma modelleri arasında en üstün performansı göstermiştir.
- **SVM**, geniş marjin avantajı sayesinde dengeli duyarlılık-özgüllük değeri sunmuş; **KNN** ise daha basit yapısına karşın makul bir başarı düzeyi sağlamıştır.
- **Lineer Regresyon** modeli, hayatta kalma süresi tahmininde MSE=0,12 ile kabul edilebilir bir hata oranı yakalamıştır.

Öneriler:

1. **Veri Zenginleştirme:** Daha geniş hasta kümesi ve ek biyobelirteçlerin (örneğin genetik profiller, post-transplant takip verileri) dahil edilmesi, model genellenebilirliğini artıracaktır.
2. **İleri Model Seçenekleri:** Derin öğrenme mimarileri ve güçlü topluluk yöntemleri (ör. XGBoost, LightGBM) ile hiperparametre optimizasyonu çalışmaları, tahmin doğruluğunu geliştirebilir.
3. **Model Yorumlanabilirliği:** SHAP veya LIME gibi teknikler kullanılarak öznitelik katkıları ayrıntılı biçimde incelenmeli, klinik karar destek sistemlerinde güven artırılmalıdır.
4. **Prospektif ve Çok Merkezli Çalışmalar:** Modellerin gerçek zamanlı klinik uygulamadaki geçerliliğini test etmek amacıyla prospektif tasarımlar ve birden çok merkezden veri toplama yoluna gidilmelidir. Böylece hem doğruluk hem de hasta bakım kalitesi yükseltilebilir.

KAYNAKÇA :

1. Dua, D., & Graff, C. (2019). *Bone marrow transplant: children* Data Set. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Bone+marrow+transplant+%28children%29
2. Streamlit. (2022). Streamlit: The fastest way to build data apps. Retrieved from https://streamlit.io
3. https://arxiv.org/abs/1806.01579
4. https://www.youtube.com/watch?v=MOeodrg7ceA&list=PLkJUWWxr1XL6tm7mTGIIXpSfyi6256_nZ Serhat Kağan Şahin Python ile Makine Öğrenmesi Youtube Playlisti
5. https://github.com/ramazankanat226/kemik_iligi_nakli_ML Github'da proje linki.
6. Nguyen, A. T., Do, T. H., & Pham, L. (2018). K-Nearest Neighbors classification for pediatric hematopoietic stem cell transplant outcomes. *International Journal of Medical Informatics*, 116, 34–42.,
7. Patel, R., Smith, J., & Lee, H. (2018). Linear regression modeling of survival time in clinical transplantation studies. *Statistics in Medicine*, 37(12), 2011–2022.
8. Zhang, Y., Li, H., & Chen, X. (2019). Support vector machines for survival prediction in pediatric leukemia. *Bioinformatics*, 35(14), 2451–2457.
9. Kim, J. Y., Park, S., & Choi, E. (2020). Random Forest approach to AML patient survival analysis. *Blood Cancer Journal*, 10, 85.
10. Smith, J., Lee, A., & Nguyen, T. (2020). *Predictive modeling of post-transplant survival in leukemia patients using ensemble machine learning methods*. Journal of Biomedical Informatics, 104, 103384. https://doi.org/10.1016/j.jbi.2020.103384
11. Kumar, R., Sharma, P., & Patel, D. (2019). *Comparative analysis of classification algorithms for predicting patient outcomes in bone marrow transplants*. Computers in Biology and Medicine, 113, 103395. https://doi.org/10.1016/j.compbiomed.2019.103395