



**TALASEMİ HASTALIĞININ ÖZNİTELİK SEÇİM
YÖNTEMLERİ KULLANARAK MAKİNE
ÖĞRENMESİ ALGORİTMALARI İLE TEŞHİS
EDİLMESİ**

Ece Gülşah ABBASOĞULLARI

Danışman: Dr. Öğr. Üyesi Faruk Baturalp GÜNAY
Yüksek Lisans Tezi
Bilgisayar Mühendisliği Anabilim Dalı
2024
(Her hakkı saklıdır.)

T.C.
ATATÜRK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

**TALASEMİ HASTALIĞININ ÖZNİTELİK SEÇİM YÖNTEMLERİ KULLANARAK
MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE TEŞHİS EDİLMESİ**

(Diagnosis of Thalassemia Disease With Machine Learning Algorithms Using Feature
Selection Methods)

YÜKSEK LİSANS TEZİ

Ece Gülşah ABBASOĞULLARI

Danışman: Dr. Öğr. Üyesi Faruk Baturalp GÜNAY

Erzurum
Eylül, 2024

KABUL VE ONAY TUTANAĞI

Ece Gülşah ABBASOĞULLARI tarafından hazırlanan “**Talasemi Hastalığının Öznelik Seçim Yöntemleri Kullanarak Makine Öğrenmesi Algoritmaları ile Teşhis Edilmesi**” başlıklı çalışması 17/09/2024 tarihinde yapılan tez savunma sınavı sonucunda başarılı bulunarak jürimiz tarafından Bilgisayar Mühendisliği Ana Bilim Dalı, Bilgisayar Mühendisliği Bilim Dalında yüksek lisans tezi olarak kabul edilmiştir.

Jüri Başkanı:	Doç. Dr. Ferhat BOZKURT <i>Atatürk Üniversitesi</i>	
Danışman:	Dr. Öğr. Üyesi Faruk Baturalp GÜNAY <i>Atatürk Üniversitesi</i>	
Jüri Üyesi:	Dr. Öğr. Üyesi Erkan GÜLER <i>Giresun Üniversitesi</i>	

Bu tezin Atatürk Üniversitesi Lisansüstü Eğitim ve Öğretim Yönetmeliği'nin ilgili maddelerinde belirtilen şartları yerine getirdiğini onaylarım.

Prof. Dr. Saltuk Buğrahan CEYHUN
Enstitü Müdürü

Not: Bu tezde kullanılan özgün ve başka kaynaklardan yapılan bildiriş, çizelge, şekil ve fotoğrafların kaynak olarak kullanımı, 5846 sayılı Fikir ve Sanat Eserleri Kanunundaki hükümlere tabidir.

ETİK BİLDİRİM VE İNTİHAL BEYAN FORMU

Yüksek Lisans Tezi olarak Dr. Öğr. Üyesi Faruk Baturalp GÜNAY danışmanlığında sunulan **“Talasemi Hastalığının Öz nitelik Seçim Yöntemleri Kullanarak Makine Öğrenmesi Algoritmaları ile Teşhis Edilmesi”** başlıklı çalışmanın tarafımızdan bilimsel etik ilkelere uyularak yazıldığını, yararlanılan eserlerin kaynakçada gösterildiğini, Fen Bilimleri Enstitüsü tarafından belirlenmiş olan Turnitin Programı benzerlik oranlarının aşılmadığını ve aşağıdaki oranlarda olduğunu beyan ederiz.

Tez Bölümleri	Tezin Benzerlik Oranı (%)	Maksimum Oran (%)
Giriş	6	30
Kuramsal Temeller	9	30
Materyal ve Metot	10	35
Araştırma Bulguları ve Tartışma	1	20
Sonuçlar ve Öneriler	17	20
Tezin Geneli	4	25

Not: Yedi kelimeye kadar benzerlikler ile Başlık, Kaynakça, İçindekiler, Teşekkür, Dizin ve Ekler kısımları tarama dışı bırakılabilir. Yukarıdaki azami benzerlik oranları yanında tek bir kaynaktan olan benzerlik oranlarının %5’den büyük olmaması gerekir.

Beyan edilen bilgilerin doğru olduğunu, aksi halde doğacak hukuki sorumlulukları kabul ve beyan ederiz.

Tez Yazarı (Öğrenci)	Tez Danışmanı
Ece Gülşah ABBASOĞULLARI	Dr. Öğr. Üyesi Faruk Baturalp GÜNAY
17.9.2024	17.9.2024
İmza:	İmza:

* Tez ile ilgili YÖKTEZ’de yayınlamasına ilişkin bir engelleme var ise aşağıdaki alanı doldurunuz.

☐ Tezle ilgili patent başvurusu yapılması / patent alma sürecinin devam etmesi sebebiyle Enstitü Yönetim Kurulunun/....../.... tarih ve sayılı kararı ile teze erişim 2 (iki) yıl süreyle engellenmiştir.

☐ Enstitü Yönetim Kurulunun/....../.... tarih ve sayılı kararı ile teze erişim 6 (altı) ay süreyle engellenmiştir.

TEŞEKKÜR

Tez çalışmam süresince motivasyon, tecrübe ve yardımları ile desteğini esirgemeyen danışman hocam Sayın Dr. Öğr. Üyesi Faruk Baturalp GÜNEY'a çok teşekkür ederim.

Bilgisayar Mühendisliğine dair bilgilerin temellerini oluşturan, emek veren ve eğiten Atatürk Üniversitesi Bilgisayar Mühendisliği Bölümü Öğretim Üyelerine çok teşekkür ederim.

Tez çalışmamda, bana destek, özveri ve yardımlarından dolayı değerli Atatürk Üniversitesi Araştırma Hastanesi Bilgi İşlem Merkezi çalışma ekibime çok teşekkür ederim.

Hayatımın her anında beni destekleyen, her türlü maddi ve manevi anlamda yanımda olan, bana inanan canım aileme çok teşekkür ederim.

Ece Gülşah ABBASOĞULLARI

ÖZET

YÜKSEK LİSANS TEZİ

TALASEMİ HASTALIĞININ ÖZNİTELİK SEÇİM YÖNTEMLERİ KULLANARAK MAKİNE ÖĞRENMESİ ALGORİTMALARI İLE TEŞHİS EDİLMESİ

Ece Gülşah ABBASOĞULLARI

Danışman: Dr. Öğr. Üyesi Faruk Baturalp GÜNAY

Amaç: Bu çalışmada, talasemi hastalarının teşhis edilmesi amaçlanmıştır. Erzurum Atatürk Üniversitesi Araştırma Hastanesi hematoloji bölümüne gelen 803 hastanın bilgilerinden oluşan veri setinin, öznitelik seçim yöntemleri kullanılarak makine öğrenmesi algoritmaları ile oluşturulan modeller aracılığı ile sınıflandırılması amaçlanmıştır. Oluşturulan modellerin öznitelik seçimi yapıldığında veya yapılmadığında elde edilen performansları ile değerlendirmeleri yapıp karşılaştırma yapılmıştır.

Yöntem: Tez çalışması için kullanılan veri seti için makine öğrenmesi algoritmalarından modeller oluşturulmuştur. Veri seti, oluşturulan modellerin eğitilmesi ve test edilmesi için kullanılmak üzere %70'i eğitim, %30'u test olarak ayrılmıştır. Veri ön işleme adımlarından sonra oluşturulan modellere k katlı çapraz doğrulama yöntemi, SFS, SBS, SFFS ve SBFS öznitelik seçim yöntemleri uygulanmıştır. Bir yöntem uygulanmadan ve uygulandıktan sonra model eğitimi yapıp sonuçların performans metrikleriyle karşılaştırma yapılarak sonuçta en iyi performans sonucunun hangi modelden elde edildiğine bakılmıştır. Bu çalışmada, veri ön işleme, model oluşturma, k katlı çapraz doğrulama, öznitelik seçimi, verilerin eğitilmesi ve test edilme işlemleri gerçekleştirilmiştir. Bu işlemler, Google Colaboratory bulut ortamında, Python programlama dili ile gerçekleştirilmiştir.

Bulgular: Bu çalışmada, ML modelleri herhangi bir yöntem kullanılmadan önce sınıflandırılmış daha sonra k katlı çapraz doğrulama yöntemi kullanılıp sonuçları karşılaştırılmıştır. Daha sonra modeller üzerinde SFS, SBS, SFFS ve SBFS öznitelik seçim yöntemleri uygulanmıştır. Herhangi bir yöntem uygulamadan önce RF modeli %89,63 doğruluk değeri ile en yüksek başarı performansı elde edilirken, k katlı çapraz doğrulama yöntemi ile RF modeli %87.00 doğruluk değeri ile en yüksek başarı performansı elde edilmiştir. Öznitelik seçim yöntemleri ile RF modelinde en yüksek başarı performanslarını göstermiştir. Sırasıyla, %89,86 SBS, %89,68 SFFS, %89,50 SFS, %89,32 SBFS doğruluk değerleri elde edilmiştir.

Sonuç: Çalışmada kullanılan yöntemlerden elde edilen sonuçlar birbirleriyle yakın değerler elde edilmiştir. Tüm bu çalışmalardan yola çıkarak, ML sınıflandırma yöntemlerinin hastalık tespitinde yüksek başarı sonuçları elde edildiği görülmüştür.

Anahtar Kelimeler: makine öğrenmesi, öznitelik seçimi, talasemi, akdeniz anemisi, çapraz doğrulama.

Eylül 2024, 83 sayfa

ABSTRACT

MASTER'S THESIS

DIAGNOSIS OF THALASSEMIA DISEASE WITH MACHINE LEARNING ALGORITHMS FEATURE SELECTION METHODS

Ece Gülşah ABBASOĞULLARI

Supervisor: Assist. Prof. Dr. Faruk Baturalp GÜNAY

Purpose: In this study, aimed to diagnose thalassemia patients. The data set, consisting of the information of 803 patients who came to the hematology department of Erzurum Atatürk University Research Hospital, was aimed to be classified through models created with machine learning algorithms using feature selection methods. The performances of the created models were evaluated and compared with and without feature selection.

Method: Models were created from machine learning algorithms for the data set used for the thesis study. The data set is divided into 70% training and 30% testing to be used for training and testing the created models. After the data preprocessing steps, k-fold cross-validation method, SFS, SBS, SFFS and SBFS feature selection methods were applied to the models created. Before and after applying a method, model training was performed and the results were compared with performance metrics to see which model achieved the best performance result. In this study, data preprocessing, model creation, k-fold cross-validation, feature selection, data training and testing processes were performed. These processes were performed in the Google Colaboratory cloud environment with the Python programming language.

Findings: In this study, the ML models were classified before using any method and then the results were compared using the k-fold cross-validation method. Then, SFS, SBS, SFFS and SBFS feature selection methods were applied on the models. Before applying any method, the highest success performance was obtained with 89.63% accuracy value in the RF model, while the highest success performance was obtained with 87.00% accuracy value in the RF model with the k-fold cross-validation method. The highest success performances were obtained in the RF model with the feature selection methods. Respectively, 89.86% SBS, 89.68% SFFS, 89.50% SFS, 89.32% SBFS accuracy values were obtained.

Results: The results obtained from the methods used in the study were close to each other. Based on all these studies, it was seen that BC classification methods achieved high success results in disease detection.

Keywords: machine learning, feature selection, thalassemia, mediterranean, cross validation.

September 2024, 83 pages

İÇİNDEKİLER

KABUL VE ONAY TUTANAĞI.....	i
ETİK BİLDİRİM VE İNTİHAL BEYAN FORMU	ii
TEŞEKKÜR	iii
ÖZET	iv
ABSTRACT	v
İÇİNDEKİLER.....	vi
TABLolar DİZİNİ.....	viii
ŞEKİLLER DİZİNİ.....	xi
KISALTMALAR VE SİMGELER DİZİNİ.....	xiii
GİRİŞ.....	1
KURAMSAL TEMELLER.....	3
MATERYAL VE METOT	7
Veri Seti	8
Veri Ön İşleme	10
Öznitelik Seçimi (Feature Selection)	12
Makine Öğrenmesi (Machine Learning-ML).....	14
Naive Bayes – (NB)	15
K En Yakın Komşu – (K Nearest Neighbour – KNN)	16
Destek Vektör Makinesi (Support Vector Machine – SVM).....	17
Lojistik Regresyon (Logistic Regresyon – LR)	17
Karar Ağacı (Decision Tree – DT).....	17
Rastgele Orman (Random Forest – RF).....	18
Sınıflandırma Algoritmalarının Performansını Test Etme Ölçütleri.....	19
Kesinlik (Precision).....	20
Duyarlılık (Recall)	20
F1 Skoru (F1 Score).....	20
Doğruluk (Accuracy-ACC).....	21
ROC – AUC Eğrisi (ROC – AUC Curve-ROC-AUC)	21
K Katlı Çapraz Doğrulama (K Fold Cross Validation-CV).....	21
ARAŞTIRMA BULGULARI VE TARTIŞMA.....	23
Öznitelik Seçim Yöntemleri Uygulanmadan Önce Sınıflandırma.....	23
Öznitelik Seçim Yöntemleri Uygulandıktan Sonra Sınıflandırma.....	30
SFS öznitelik seçim yöntemi ile ML algoritmalarının modellenmesi	30

SBS öznitelik seçim yöntemi ile ML algoritmalarının modellenmesi	37
SFFS öznitelik seçim yöntemi ile ML algoritmalarının modellenmesi	44
SBFS öznitelik seçim yöntemi ile ML algoritmalarının modellenmesi.....	51
SONUÇ VE ÖNERİLER	63
KAYNAKLAR.....	65
ÖZGEÇMİŞ.....	68



TABLÖLAR DİZİNİ

Tablo 1. Talasemi Hastalığı Tahmini için Veri Seti Öznitelikleri	8
Tablo 2. Karmaşıklık Matrisi (Confusion Matrix).....	20
Tablo 3. Veri Setinde Yöntem Uygulanmadan Önce NB Performans Metrikleri	23
Tablo 4. Veri Setinde Yöntem Uygulanmadan Önce KNN Performans Metrikleri	25
Tablo 5. Veri Setinde Yöntem Uygulanmadan Önce SVM Performans Metrikleri	25
Tablo 6. Veri Setinde Yöntem Uygulanmadan Önce LR Performans Metrikleri.....	26
Tablo 7. Veri Setinde Yöntem Uygulanmadan Önce RF Performans Metrikleri.....	27
Tablo 8. Veri Setinde Yöntem Uygulanmadan Önce DT Performans Metrikleri	28
Tablo 9. K Katlı Çapraz Doğrulama Uygulamadan Önce ve Sonrası Modellerin Doğruluk Değerleri	30
Tablo 10. SFS Yöntemi İçin Kullanılan Parametreler	30
Tablo 11. NB-SFS Seçilen En İyi Öznitelikler	31
Tablo 12. SFS Öznitelik Seçim Öncesi ve Sonrası NB Doğruluk Oran Karşılaştırması.....	31
Tablo 13. KNN-SFS Seçilen En İyi Öznitelikler	32
Tablo 14. SFS Öznitelik Seçim Öncesi ve Sonrası KNN Doğruluk Oran Karşılaştırması	32
Tablo 15. SVM-SFS Seçilen En İyi Öznitelikler	33
Tablo 16. SFS Öznitelik Seçim Öncesi ve Sonrası SVM Doğruluk Oran Karşılaştırması	34
Tablo 17. LR-SFS Seçilen En İyi Öznitelikler	34
Tablo 18. SFS Öznitelik Seçim Öncesi ve Sonrası LR Doğruluk Oran Karşılaştırması	35
Tablo 19. RF-SFS Seçilen En İyi Öznitelikler	35
Tablo 20. SFS Öznitelik Seçim Öncesi ve Sonrası RF Doğruluk Oran Karşılaştırması	36
Tablo 21. DT-SFS Seçilen En İyi Öznitelikler	36
Tablo 22. SFS Öznitelik Seçim Öncesi ve Sonrası DT Doğruluk Oran Karşılaştırması	37
Tablo 23. SFS Öznitelik Seçim Öncesi ve Sonrası ML Algoritmalarının Doğruluk Oran Karşılaştırması	37
Tablo 24. SBS Yöntemi İçin Kullanılan Parametreler	38
Tablo 25. NB-SBS Seçilen En İyi Öznitelikler	38
Tablo 26. SBS Öznitelik Seçim Öncesi ve Sonrası NB Doğruluk Oran Karşılaştırması	39
Tablo 27. KNN-SBS Seçilen En İyi Öznitelikler	39
Tablo 28. SBS Öznitelik Seçim Öncesi ve Sonrası KNN Doğruluk Oran Karşılaştırması	40
Tablo 29. SVM -SBS Seçilen En İyi Öznitelikler	40
Tablo 30. SBS Öznitelik Seçim Öncesi ve Sonrası SVM Doğruluk Oran Karşılaştırması	41
Tablo 31. LR-SBS Seçilen En İyi Öznitelikler	41

Tablo 32. SBS Öznitelik Seçim Öncesi ve Sonrası LR Doğruluk Oran Karşılaştırması.....	42
Tablo 33. RF-SBS Seçilen En İyi Öznitelikler	42
Tablo 34. SBS Öznitelik Seçim Öncesi ve Sonrası RF Doğruluk Oran Karşılaştırması	43
Tablo 35. DT-SBS Seçilen En İyi Öznitelikler.....	43
Tablo 36. SBS Öznitelik Seçim Öncesi ve Sonrası DT Doğruluk Oran Karşılaştırması	44
Tablo 37. SBS Öznitelik Seçim Öncesi ve Sonrası ML Algoritmalarının Doğruluk Oran Karşılaştırması	44
Tablo 38. SFFS Yöntemi İçin Kullanılan Parametreler	45
Tablo 39. NB-SFFS Seçilen En İyi Öznitelikler.....	45
Tablo 40. SFFS Öznitelik Seçim Öncesi ve Sonrası NB Doğruluk Oran Karşılaştırması	46
Tablo 41. KNN-SFFS Seçilen En İyi Öznitelikler	46
Tablo 42. SFFS Öznitelik Seçim Öncesi Ve Sonrası KNN Doğruluk Oran Karşılaştırması ..	47
Tablo 43. SVM-SFFS Seçilen En İyi Öznitelikler	47
Tablo 44. SFFS Öznitelik Seçim Öncesi ve Sonrası SVM Doğruluk Oran Karşılaştırması ...	48
Tablo 45. LR-SFFS Seçilen En İyi Öznitelikler	48
Tablo 46. SFFS Öznitelik Seçim Öncesi ve Sonrası LR Doğruluk Oran Karşılaştırması	49
Tablo 47. RF-SFFS Seçilen En İyi Öznitelikler	49
Tablo 48. SFFS Öznitelik Seçim Öncesi ve Sonrası RF Doğruluk Oran Karşılaştırması	50
Tablo 49. DT-SFFS Seçilen En İyi Öznitelikler	50
Tablo 50. SFFS Öznitelik Seçim Öncesi ve Sonrası DT Doğruluk Oran Karşılaştırması.....	51
Tablo 51. SFFS öznitelik seçim öncesi ve sonrası ML algoritmalarının doğruluk oran karşılaştırması.....	51
Tablo 52. SBFS Yöntemi için Kullanılan Parametreler	52
Tablo 53. NB-SBFS Seçilen En İyi Öznitelikler	52
Tablo 54. SBFS Öznitelik Seçim Öncesi ve Sonrası NB Doğruluk Oran Karşılaştırması	53
Tablo 55. KNN-SBFS Seçilen En İyi Öznitelikler	53
Tablo 56. SBFS Öznitelik Seçim Öncesi ve Sonrası KNN Doğruluk Oran Karşılaştırması...	54
Tablo 57. SVM-SBFS Seçilen En İyi Öznitelikler	54
Tablo 58. SBFS Öznitelik Seçim Öncesi ve Sonrası SVM Doğruluk Oran Karşılaştırması...	55
Tablo 59. LR-SBFS Seçilen En İyi Öznitelikler.....	55
Tablo 60. SBFS Öznitelik Seçim Öncesi Ve Sonrası LR Doğruluk Oran Karşılaştırması.....	56
Tablo 61. RF-SBFS Seçilen En İyi Öznitelikler	56
Tablo 62. SBFS Öznitelik Seçim Öncesi ve Sonrası RF Doğruluk Oran Karşılaştırması.....	57
Tablo 63. DT-SBFS Seçilen En İyi Öznitelikler	57
Tablo 64. SBFS Öznitelik Seçim Öncesi ve Sonrası DT Doğruluk Oran Karşılaştırması	58

Tablo 65. SBFS Öznitelik Seçim Öncesi ve Sonrası ML Algoritmalarının Doğruluk Oran Karşılaştırması	58
Tablo 66. Öznitelik Seçim Yöntemleri Uygulamadan Önce ve Sonra ML Algoritmalarının Performans Karşılaştırmaları	59



ŞEKİLLER DİZİNİ

Şekil 1. Makine öğrenmesi algoritmalarının akış şeması.....	7
Şekil 2. Talasemi hastalığı olan (1) ve sağlıklı kişi (0) dağılımı.....	8
Şekil 3. Veri seti istatistik bilgisi.	9
Şekil 4. Veri seti özniteliklerin dağılımı.	9
Şekil 5. Değişkenler arasındaki ısı haritası.	10
Şekil 6. Orijinal veri seti mcv kutu grafik gösterimi.	11
Şekil 7. Veri ön işleme sonrasında mcv kutu grafik gösterimi.	11
Şekil 8. SFS yöntemi çalışma akışı.	13
Şekil 9. SBS yöntemi çalışma akışı.....	13
Şekil 10. Karar Ağacı algoritma yapısı.	18
Şekil 11. Rastgele orman algoritması çalışma şeması.....	19
Şekil 12. ROC – AUC eğri grafiği.	21
Şekil 13. k katlı çapraz doğrulama.	22
Şekil 14. NB modeli karmaşıklık matrisi.	24
Şekil 15. KNN algoritması için optimum k değerini belirleme grafiği.....	24
Şekil 16. KNN modeli karmaşıklık matrisi.	25
Şekil 17. SVM modeli karmaşıklık matrisi.	26
Şekil 18. LR modeli karmaşıklık matrisi.....	27
Şekil 19. RF modeli karmaşıklık matrisi.....	28
Şekil 20. DT modeli karmaşıklık matrisi.	29
Şekil 21. Sınıflandırma sonucunda oluşan DT yapısı.	29
Şekil 22. SFS yöntemi sonucu NB modeli performans gösterimleri.	31
Şekil 23. SFS yöntemi sonucu KNN modeli performans gösterimleri.	32
Şekil 24. SFS yöntemi sonucu SVM modeli performans gösterimleri.	33
Şekil 25. SFS yöntemi sonucu LR modeli performans gösterimleri.....	34
Şekil 26. SFS yöntemi sonucu RF modeli performans gösterimleri.	35
Şekil 27. SFS yöntemi sonucu DT modeli performans gösterimleri.....	36
Şekil 28. SBS yöntemi sonucu NB modeli performans gösterimleri.	38
Şekil 29. SBS yöntemi sonucu KNN modeli performans gösterimleri.....	39
Şekil 30. SBS yöntemi sonucu SVM modeli performans gösterimleri.....	40
Şekil 31. SBS yöntemi sonucu LR modeli performans gösterimleri.	41
Şekil 32. SBS yöntemi sonucu RF modeli performans gösterimleri.....	42
Şekil 33. SBS yöntemi sonucu DT modeli performans gösterimleri.	43

Şekil 34. SFFS yöntemi sonucu NB modeli performans gösterimleri.	45
Şekil 35. SFFS yöntemi sonucu KNN modeli performans gösterimleri.	46
Şekil 36. SFFS yöntemi sonucu SVM modeli performans gösterimleri.	47
Şekil 37. SFFS yöntemi sonucu LR modeli performans gösterimleri.	48
Şekil 38. SFFS yöntemi sonucu RF modeli performans gösterimleri.	49
Şekil 39. SFFS yöntemi sonucu DT modeli performans gösterimleri.	50
Şekil 40. SBFS yöntemi sonucu NB modeli performans gösterimleri.	52
Şekil 41. SBFS yöntemi sonucu KNN modeli performans gösterimleri.	53
Şekil 42. SBFS yöntemi sonucu SVM modeli performans gösterimleri.	54
Şekil 43. SBFS yöntemi sonucu LR modeli performans gösterimleri.	55
Şekil 44. SBFS yöntemi sonucu RF modeli performans gösterimleri.	56
Şekil 45. SBFS yöntemi sonucu DT modeli performans gösterimleri.	57
Şekil 46. ROC-AUC eğrisi.	59
Şekil 47. SBS ile ROC-AUC eğrisi.	60
Şekil 48. SFS Yöntemi ile Modellerin Doğruluk Oranlar Grafiği	60
Şekil 49. SBS Yöntemi ile Modellerin Doğruluk Oranlar Grafiği.	61
Şekil 50. SFFS Yöntemi ile Modellerin Doğruluk Oranlar Grafiği	61
Şekil 51. SBFS Yöntemi ile Modellerin Doğruluk Oranlar Grafiği	62

KISALTMALAR VE SİMGELER DİZİNİ

AUC	: Eğri Altında Kalan Alan (Area Under the Curve)
CV	: K Katlı Çapraz Doğrulama (K Fold Cross Validation)
DT	: Karar Ağacı (Decision Tree)
EM	: Gömülü Yöntemler (Embedded Methods)
FN	: Yanlış Negatif (False Negative)
FM	: Filtreleme Yöntemler (Filter Methods)
FP	: Yanlış Pozitif (False Positive)
KNN	: K En Yakın Komşu (K Nearest Neighbour)
LR	: Lojistik Regresyon (Logistic Regresyon)
ML	: Makine Öğrenmesi (Machine Learning)
NB	: Naive Bayes
RF	: Rastgele Orman (Random Forest)
ROC	: Alıcı İşlem Karakteristikleri (Receiver Operating Characteristic)
SBS	: Ardışık Geri Yönde Seçim (Sequentinal Backward Selection)
SBFS	: Ardışık Geri Yönde Kayan Seçim (Sequentinal Backward Floating Selection)
SFS	: Ardışık İleri Yönde Seçim (Sequentinal Forward Selection)
SBFS	: Ardışık Geri Yönde Kayan Seçim (Sequentinal Backward Floating Selection)
SVM	: Destek Vektör Makineleri (Support Vector Machine)
TN	: Gerçek Negatif (True Negative)
TP	: Gerçek Pozitif (True Positive)
YZ	: Yapay Zeka (Artificial Intelligence)
WM	: Sarmal Yöntemler (Wrapper Methods)

GİRİŞ

Kalıtsal hastalıklar her geçen gün artış göstermektedir. Bireylerde hayatları boyunca hastalıkları geç fark etme, geç tedaviye başlama ve düzenli tedavi olmadığı durumda ölüm riski kaçınılmaz olmaktadır. Bu hastalıklar kadın, erkek ve aynı zamanda çocuklarda sıkça görülmektedir. Talasemi (Akdeniz Anemisi), bireylerde sıkça rastlanan kalıtsal hastalıklardan birisidir. Talasemi, hemoglobın sentezinde kalıtsal bozuklukların neden olduğu dünya genelinde çok yaygın bir gen hastalığıdır (Mohammed & Al-Tuwaijari, 2021). Talasemi hastalığı ilk olarak 1925 yılında Thomas Cooley tarafından “anemi” ve “splenomegali” olan bebeklerde görülüp tanımlanmıştır. İleri tarihlerde bu tip vakaların artması ile bu hastalığa “Van Jaksch anemisi”, “splenik anemisi”, “eritroblastozis”, “Akdeniz anemisi” isimleri verilmiştir. George Whipple ve Lesley Bradford tarafından 1936 yılında araştırdıkları vakaların Akdeniz taraflarındaki ülkelerde sıkça rastlandığını gördükleri için bu hastalığa “thalassemia” adını vermişlerdir. Talasemi hastalığının sadece Akdeniz etrafındaki ülkelerde değil diğer ülkelerde de olduğu saptanmıştır (Demir, 2019).

Dünya Sağlık Örgütü raporlarına bakıldığında, dünya üzerinde %5,1 talasemi hastalığı ve 266 milyon talasemi taşıyıcısı bulunduğu görülmektedir. Talasemi hastalığı Türkiye’de yaklaşık olarak 1.300.000 talasemi taşıyıcısı var olduğu ve 4500 bireyin de talasemi hastası olduğu görülmektedir. Akraba evliliğinin çok sık görüldüğü ülkemizde doğan β -talasemi tanısı konan çok fazla çocuk olduğu görülmektedir. Türkiye ‘de talasemi hastalığının görülme sıklığı %2,1 olarak belirlenip talasemi taşıyıcılığı dünyada en fazla %18’lik bir oranla Kıbrıs’ta görülmektedir. Talasemi hastalığını önlemek için hastalara çeşitli programlar ve geliştirilmiş tedaviler uygulanmıştır. Bu uygulamalar sonucunda hastalık oranında düşüş olmuştur. Bu hastalığın diğer yerlerden daha sık görülmesinden dolayı bu programlar ve tedaviler ilk olarak Kıbrıs’ta uygulanmıştır (Karaaziz & Okyayuz, 2020).

Talasemi hastalığı tanısı için genetik testler yaygın olarak kullanılmaktadır. Genetik testlerin yanı sıra çeşitli tarama yöntemleri hastalığın teşhis edilmesinde ilk adım olarak tercih edilmektedir. Talasemi hastalığında geleneksel tarama yöntemlerinden olan kan yayması, tam kan sayımı, ozmotik kırılabilirlik testi, hemoglobın elektrofarezi gibi yöntemleri uzun yıllar boyunca hastalığın teşhis edilmesinde klinik uygulamalarda kullanılmaktadır (Gao & Liu, 2022). Talasemi hastaları için erken ve doğru tarama yapmak oldukça önemlidir. Son yıllarda Yapay Zekâ (YZ) teknolojileri birçok alanda özellikle tıp alanında çeşitli hastalıkların teşhis,

tedavi ve tanı aşamalarında büyük bir başarıyla kullanılmaktadır. YZ uygulamaları denilince ilk olarak Makine Öğrenmesi (ML) yöntemleri gelmektedir. ML, verileri daha verimli bir şekilde kullanmaları gerektiğini makinelere öğretmek için kullanılmaktadır. Veriler görüntülendikten sonra elde edilen bilgiler yorumlanamayabilir. Bu bilgileri yorumlayabilmek için ML yöntemleri kullanılır. ML yöntemleri ile örnek veri setleri üzerinde sınıflandırma ya da tahmin uygulamaları yapılmaktadır. Bir sorunu çözmek için en iyi, tek tip algoritma yoktur. Kullanılan ML algoritmasının türü, problemin türüne, veri seti değişkenlerinin sayısına ve probleme uygun modelin türüne bağlıdır (Mahesh, 2020).

ML sınıflandırma algoritmalarının kullanımı, hastalıkların erken teşhisi ve tedavi süreçlerinin iyileştirilmesi açısından büyük potansiyele sahiptir. Özellikle büyük veri setleriyle çalışıldığında, ML sınıflandırma algoritmaları sayesinde daha doğru ve hızlı teşhisler konulabilir. ML teknikleri, sağlık hizmetlerinin kalitesini artırarak, hastaların daha iyi tedavi edilmesine olanak tanımaktadır (Veranyurt et al., 2020).

Bu çalışmada, talasemi hastalığının öznelik seçim yöntemleri ile ML algoritmaları kullanarak erken teşhis edilmesine katkı sağlamaktır. Bu sayede talasemi hastalığıyla meydana gelen risklerin ML algoritmaları ile erken teşhis edilmesi ile beraber hastalar için oluşan bu risklerin en aza indirilip var olan uygun tedavi yöntemleri ile daha kaliteli yaşam sürmesine katkıda bulunması amaçlanmıştır.

KURAMSAL TEMELLER

Ferih ve arkadaşları (Ferih et al., 2023) yaptıkları çalışmada, ML tekniklerini kullanarak talasemi hastalığı tahmini için modelleme yapmışlardır. Tam kan sayımı testi ile talasemi tanısına yönelik parametreler açıklanmıştır. Çalışmada model olarak, K-En Yakın Komşu (K-Nearest Neighbour-KNN), Naive Bayes (NB), Karar Ağacı (Decision Tree-DT) ve Çok Katmanlı Algılayıcı (MLP) kullanılmıştır. Çalışma sonucunda %92 doğrulukla MLP ile en başarılı sonuca ulaşmıştır. MCH, MCHC ve RDW gibi parametreler kullanılarak talasemi ve diğer kan hastalıklarının teşhisinde en iyi yöntem olduğunu göstermiştir.

Liu ve arkadaşları (Liu, 2024) yaptıkları bir çalışmada, hastaların Akdeniz anemisine sahip olup olmadığını tahmin etmek için iki model önermektedir. Çalışmada kullanılan veriler Çin' in Guangxi Zhuang Özerk Bölgesi'ndeki bir hastanede bulunan klinik kayıtlardan elde edilmiştir. Veri seti, 110 farklı gen içeren her bir örnek ve toplam 110 tane gözlemlenen gösterge ya da değişken ile sonuçlanmış olan toplam 60 bireyin genetik örneklerinden oluşmaktadır. Çalışmada kullanılan modeller, Temel Bileşen Analizi ile Lojistik regresyon (PCA- Logistic Regresyon-LR) ve kısmi en küçük kareler yöntemi (PLS) modelleridir. Çalışma sonucunda PCA-LR modelinin tahmin doğruluğu %87,5 ve PLS modelinin tahmin doğruluğu ise %92,5 olarak elde edilmiştir.

Devanath ve arkadaşları (Devanath et al., 2022) yaptıkları bir çalışmada, YZ'nin önemli bir parçası olan ML yöntemi ile talasemi varlığının öngörülmesi için modeller oluşturulmuştur. Modelin geliştirilmesi için 297 talasemi pozitif ve 297 talasemi negatif kişilerden elde edilen veriler kullanılmıştır. Uygulanan modeller, KNN, LR, SVM, NB, RF, Uyarlanabilir Güçlendirme (ADA Boosting), Xgboost, DT, MLP ve Gradient Boosting sınıflandırıcısıdır. Çalışma sonucunda uygulanan 10 model arasında ADABOOST modeli %100 ile en iyi sonucu vermiştir.

Phirom ve arkadaşları (Phirom et al., 2022) yaptıkları bir çalışmada, α + talasemi tahmini için kırmızı kan hücresi (RBC) parametrelerini kullanan ML tabanlı model geliştirmek ve teşhis performansını tek bir RBC parametresi kullanan geleneksel bir yöntem ile karşılaştırmak amaçlanmıştır. Hemoglobin H hastalığı bulunan ve fetüs açısından risk altında olan çiftler üzerinde çalışma yapılmıştır. Doğum öncesi talasemi hastalarından elde edilen yaş, cinsiyet, RBC (Hb, Hct, MCV, MCH, MCHC, RDW ve RBC sayısı) parametreleri ile makine öğrenmesi tabanlı yöntem kullanılmıştır. Çalışmada, CNN, SVM, MLP, RF, PLS, LR, ET,

LGBM, XGB, DT, KNN yöntemleri kullanılmıştır. 10 kat çapraz doğrulama yöntemi kullanılmıştır. Çalışma sonucunda modelde en iyi sonuç %100 MLP ile elde edilmiştir.

Laengsri ve arkadaşları (Laengsri et al., 2019) yaptıkları bir çalışmada, talasemi ile demir eksikliği anemisini ayırt etmek için web tabanlı bir tahminleme aracı oluşturmak amaçlanmıştır. Çalışmada Temmuz 2014'ten Eylül 2016'ya kadar Madidol Üniversitesi Tıbbi Teknoloji Fakültesi Tıbbi Laboratuvar Hizmetleri Merkezi'nden elde edilen klinik laboratuvar verilerinden oluşan veri seti üzerinden çalışma gerçekleştirilmiştir. Çalışmada KNN, Yapay Sinir Ağı, MLP, SVM ve DT modelleri uygulanmıştır. Uygulanan modeller içerisinde %99,12 doğruluk oranı ile MLP en iyi sonucu vermiştir.

Yağmur ve arkadaşları (Yağmur et al., 2023) yaptıkları bir çalışmada, Yapay Sinir Ağları (YSA) yöntemi kullanılarak bir sınıflandırma yapılması amaçlanmıştır. Çalışma, Kaggle veri tabanından alınan anemi veri seti üzerinde yapılmıştır. Yapılan çalışmada Öğrenmeli Vektör Kuantalama (LVQ), Rekabetçi Katman Sinir Ağı (CLNN), Örüntü Tanıma Yapay Sinir Ağı (PRNN), Kendiliğinden Organize Olan Harita (Paokanta et al.) NN modelleri kullanılmıştır. Kullanılan modeller ile elde edilen doğruluk oranları ve karışıklık matrisleri verilerek yöntemler arasında karşılaştırma yapılmıştır. Çalışma sonucunda %99,88 doğruluk, %99,86 özgüllük, %99,43 kesinlik oranı ile PRNN ile en iyi başarı performansı elde edilmiştir.

Farzaliyev ve arkadaşları (Farzaliyev et al., 2023) yaptıkları bir çalışmada, çocuklar üzerinde anemi hastalığının teşhis edilmesinde Topluluk Öğrenme Yöntemlerinin (Esemble Learning) kullanılması amaçlanmıştır. Çalışma Irak'ın Haditha Genel Hastanesi ve kliniklerden elde edilen 600 hastadan oluşan veri seti üzerinde yapılmıştır. Çalışmada birçok ML yöntemi kullanılmıştır. DT, SVM, RF, LR, KNN kullanılan yöntemlerdir. Bu yöntemlere ek olarak Topluluk öğrenme yöntemleri olan Torbalama (Bagging), Artırma (Boosting) ve İstifleme (Staking) yöntemleri kullanılmıştır. Çalışma sonucunda ML yöntemlerinden %98 doğruluk oranı ile DT en yüksek başarı performansı elde edilmiştir. Topluluk öğrenme yöntemlerinden %91 doğruluk oranı ile Boosting en yüksek başarı performansı elde edilmiştir.

Fu ve arkadaşları (Fu et al., 2021) yaptıkları bir çalışmada yetişkin Tayvanlı 350 hasta anemi tanısı, tam kan hücresi sayımı ve hemoglobin gen profili içeren bir veri seti üzerinde retrospektif olarak inceleme yapılmıştır. Çalışma SVM algoritması kullanılarak gerçekleştirilmiştir. Duyarlılık, özgüllük, pozitif ve negatif tahmin değerleri elde edilip incelenmiştir. Çalışma sonucunda model performansı, AUC değeri 0,76 olarak elde edilmiştir. Bu sonuç doğrultusunda Tayvanlı talasemi hastaları için uygun bir araç geliştirilmiştir.

Das ve arkadaşları (Das et al., 2022) yaptıkları bir çalışmada β -Talasemi taraması için kırmızı kan hücresi (RBC) parametrelerine ve 13 makine öğrenme algoritması ile 26 ayırım

formülü önerilmiştir. Önerilen bu formüller ile doğum öncesi kadın verileri üzerindeki performansları ile SCSBTT formülü ile yetmişten fazla ülkede SUSOKA adı verilen bir Android uygulaması arasında karşılaştırma yapmak ve değerlendirmek amaçlanmıştır. Çalışma Hindistan'ın Chandigarh kentindeki PGIMER'den doğum öncesi 2942 kadından oluşan bir veri seti üzerinde gerçekleştirilmiştir. Çalışmada, RF, SVM, LR, KNN, DTC, ETC, BC, MLP, ADA, LDA, GBC, ELM gibi makine öğrenmesi algoritmaları kullanılmıştır. Çalışma sonucunda ELM ve GBC modellerinin tüm ayırt edici formüllerle karşılaştırılması durumunda AUC-ROC ölçümlerinin daha başarılı performans elde edildiği görülmüştür.

Paokanta ve arkadaşları (Paokanta et al., 2010) yaptıkları bir çalışmada makine öğrenmesi tekniklerine uygun veri türü bulmak için veri türü değişerek farklı sınıflandırma yöntemlerinin karşılaştırmalı performans sonuçlarını sağlamayı amaçlamaktadır. Çalışmada, KNN, MLP, NB, BNs (Bayesian Networks), MLR (Multinomial Logistic Regression) teknikleri kullanılmıştır. Çalışma sonucunda sırasıyla KNN %88,98, MLP %87,40 ve NB %84,25 doğruluk değerleri elde edilmiştir.

Mo ve arkadaşları (Mo et al., 2023) yaptıkları çalışmada kırmızı kan hücresi (RBC) değerlerine dayalı geleneksel istatistiksel yöntemlerin yerini makine öğrenmesi yöntemleri almaktadır. Bu yöntemlerin yanı sıra daha iyi performans gösteren derin sinir ağları (DNNs) ile çalışma yapılmıştır. Çalışmada genetik testler ve 11 özellik bulunan 8693 veri bulunan bir veri seti üzerinde testler gerçekleştirilmiştir. Veri seti üzerinde 11 DNN modeli ve 4 geleneksel istatistiksel model kullanılmıştır. Çalışma sonucunda DNN modeli mevcut geleneksel tarama modelinden daha iyi performans gösterdiği görülmüştür. Kullanılan özellikler arasında en faydalı olanlar RDW ve yaş olarak değerlendirilmiştir.

Mirghaed ve arkadaşları (Mirghaed et al., 2020) yaptıkları bir çalışmada talasemi hastaları üzerinde transfüzyon sonrası bir komplikasyon riskinin olup olmadığı makine öğrenmesi algoritmaları kullanılarak tahmin edilmesi amaçlanmıştır. Çalışma için, 2018 yılında Tahran şehrindeki 12 ve Mazandaran şehrindeki 14 talasemi merkezinde bulunan 3489 hastadan alınan verilerle oluşturulan veri seti kullanılmıştır. Bu veri seti üzerinde makine ve derin öğrenme tekniklerinden oluşan farklı sınıflandırma modeli eğitilip üzerinde çalışma yapılmıştır. Transfüzyon sonrası komplikasyon riskini tahmin etmede makine öğrenimi yöntemlerinin çalışma sonucunda iyi bir doğruluk elde edilmiştir. Derin öğrenme yöntemleri kullanarak doğruluk değeri %76 ile en iyi tahmin sonucunu üretmiştir. Bu yöntem ile transfüzyon sonrası komplikasyon yaşayacak olan hastaların %77'si nakil öncesinde durum tespit edilmektedir.

Sadiq ve arkadaşları (Sadiq et al., 2021) Pakistan’ da yapılan bir çalışmada, Punjab Talasemi önleme programından alınan 5066 hastanın tam kan sayımı verileri üzerinde eğitilen bir yapay zeka modeli, talasemi taşıyıcılarını tespit etmede kullanılmıştır. Çalışmada bu önerilen SGR-VC modeli; Destek Vektör Makinesi (SVM), Gradyan Artırma Makinesi (GBM) ve Rastgele Orman (RF) makine öğrenme algoritmalarının birleşiminden oluşmaktadır. Bu kullanılan model, tam kan sayımı (CBC) testlerine ait özellikleri, hastaların demografik bilgilerini ve hedef etiketleri kullanarak, %93 doğruluk elde ederek talasemi taşıyıcılarını ve olmayan bireyleri başarıyla ayırt edebildiğini kanıtlamıştır.

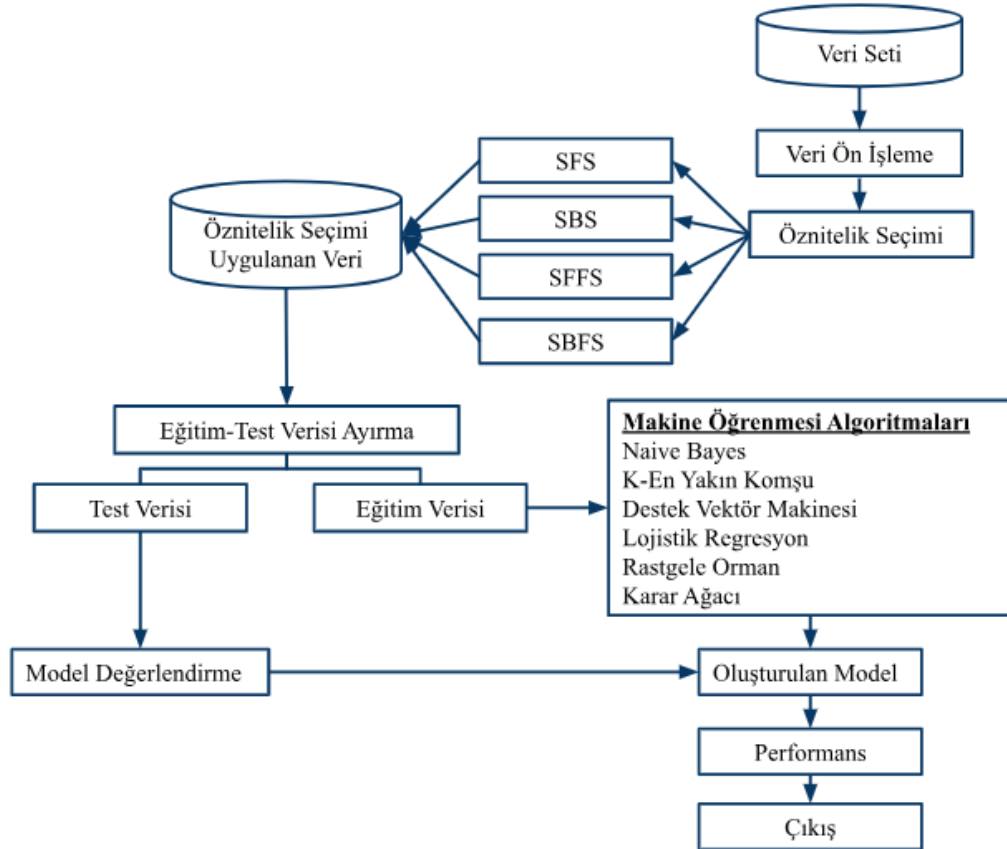
Masala ve arkadaşları (Masala et al., 2013) İtalya’ da yaptıkları bir çalışmada, Radyal Temel Fonksiyon (RBF) makine öğrenimi algoritmasına dayanan bir model geliştirilmiş ve veri seti olarak 304 hastanın tam kan sayımı kayıtları kullanılmıştır. Veri setindeki tüm hastaları modeller doğru sınıflandırmıştır. Çalışmada, Olasılıksal Sinir Ağı (PNN) ve K-En Yakın Komşu (KNN) algoritmaları da denenmiş fakat RBF’ deki yüksek doğruluğu elde edememişlerdir. RBF modelinin hız ve verimlilik açısından en iyi olduğu görülmüştür.

Noferesti ve arkadaşları (Noferesti et al., 2019) yaptıkları bir çalışmada, Dr. Haidari’ nin İran’ ın Zahedan şehrinde bulunan laboratuvarından tam kan sayımı veri seti toplanmıştır. Çalışmada, farklı makine öğrenimi modellerini kendi topladıkları veri seti ile eğiterek karşılaştırma yapılmıştır. Veri setini eğitmek için Karar Ağaçları (DT), Naive Bayes (NB), Bagging, Destek Vektör Sınıflandırması (SVC) ve AdaBoost gibi çeşitli makine öğrenimi modelleri kullanılmıştır. Modellerin performans karşılaştırması %96 DT, %76,6 NB, %80,2 AdaBoost, %95,5 SVC ve %96,6 Bagging doğruluk oranları elde edilmiştir. Bagging modeli ile en iyi başarı performansı elde edilmiştir.

MATERYAL VE METOT

Tez çalışmasında veri seti 2008-2023 yılları arasında Erzurum Atatürk Üniversitesi Araştırma Hastanesi Hematoloji bölümüne gelen 0-81 yaş aralığına sahip hasta verilerinden oluşmaktadır. Retrospektif bir çalışma olduğu için hastalara ait veriler hasta kayıtlarının tutulduğu veri tabanından çekilmiştir. Bu çalışma için etik kurul izni Atatürk Üniversitesi Girişimsel Olmayan Klinik Araştırmalar Etik Kurulu'nun 07/06/2024 tarihli ve B.30.2.ATA.0.01.00/378 numaralı kararı ile alınmıştır. Talasemi hastalığı tahmini için NB, KNN, SVM, LR, RF ve DT gibi ML algoritmaları kullanılmıştır. Bu algoritmalar öznelik seçme yöntemleri uygulanarak talasemi hastası ve sağlıklı kişileri başarılı bir şekilde sınıflandırılmasını sağlayacak en yüksek doğruluk değerinin elde edildiği yöntem belirlenmeye çalışılmıştır. Çalışma Google Colaboratory bulut ortamında ve Python programlama dili ile gerçekleştirilmiştir.

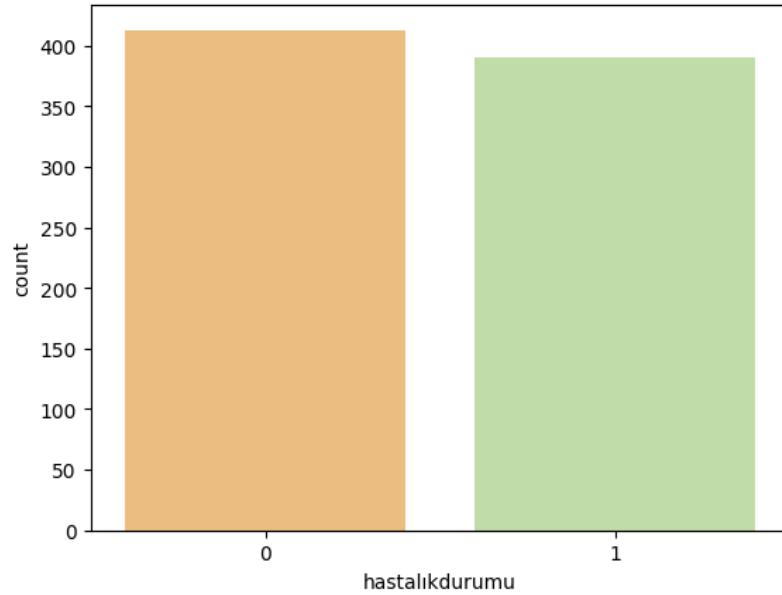
Talasemi hastalığının tahmini için ML algoritmaları ile gerçekleştirilen çalışmanın akış şeması Şekil 1'de gösterilmiştir. Akış şemasındaki her aşama aşağıdaki alt bölümlerde detaylandırılarak açıklanmıştır.



Şekil 1. Makine öğrenmesi algoritmalarının akış şeması.

Veri Seti

Bu çalışmada, veri seti Erzurum Atatürk Üniversitesi Araştırma ve Eğitim Hastanesi Hematoloji Bölümüne gelen hasta bilgilerinden oluşmaktadır. Veri seti Şekil 2’de görüldüğü gibi hastaneye gelen 803 hasta içerisinde 390 talasemi hastası olan ve 413 sağlıklı bireyden oluşmaktadır. Örneklem yaş aralığı 0-81 arasındadır.



Şekil 2. Talasemi hastalığı olan (1) ve sağlıklı kişi (0) dağılımı.

Veri seti 9 bağımsız değişken (hemoglobin, hemoglobin a, hemoglobin a2 vs.) ve bir bağımlı değişkenden (hastalık durumu) oluşmaktadır. Tablo 1’de veri seti öznitelikler, açıklamaları ve birer örnekle ayrıntılı bir şekilde gösterilmiştir.

Tablo 1. Talasemi Hastalığı Tahmini için Veri Seti Öznitelikleri

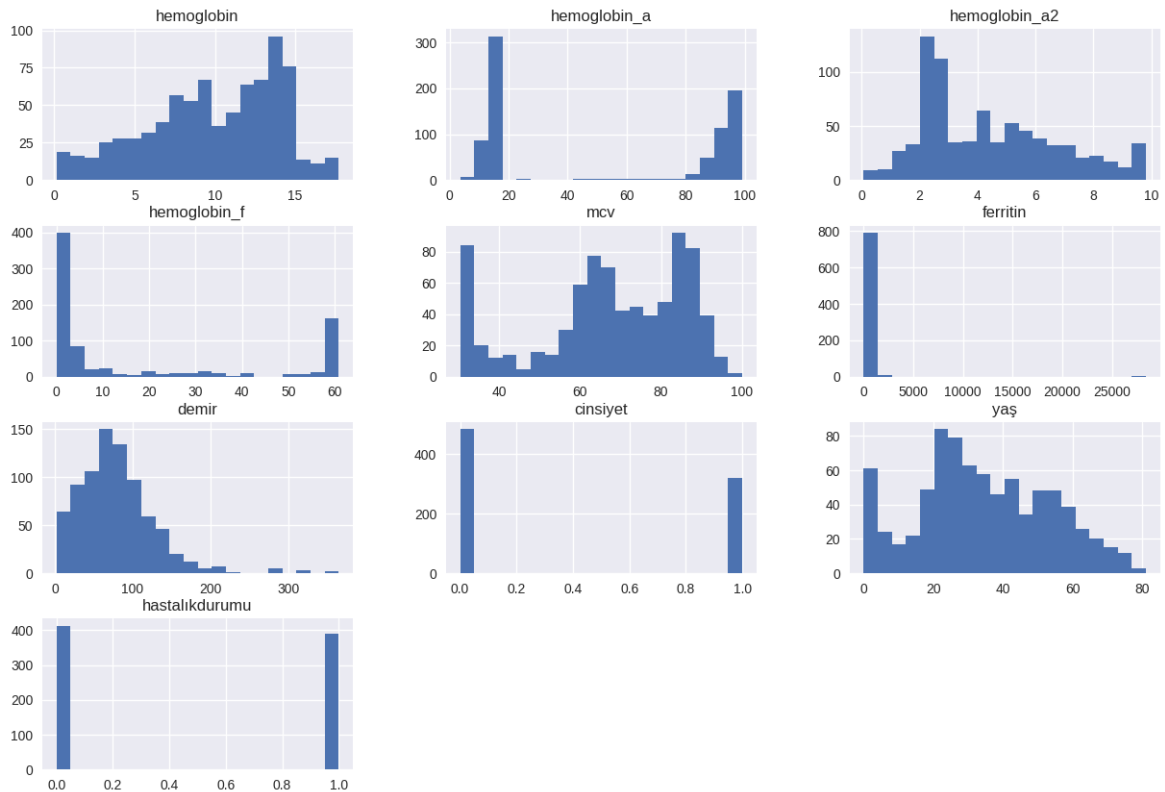
Öznitelikler	Açıklama	Örnek
hemoglobin	kandaki hemoglobin değeri	56
hemoglobin a	kandaki hemoglobin a değeri	94,26
hemoglobin a2	kandaki hemoglobin a2 değeri	7,2
hemoglobin f	kandaki hemoglobin f değeri	18,49
mcv	kandaki ortalama hücre hacmi	69,7
ferritin	kandaki ferritin değeri	8,05
demir	kandaki demir değeri	13
cinsiyet	kadın, erkek	(0,1)
yaş	yaş	26
hastalık durumu	sağlıklı, hasta	(0,1)

Veri setindeki hastalara ait verilerde kaç kayıt olduğu, özniteliklerin ortalama değeri, standart sapma değeri, minimum ve maksimum değeri gibi istatistik bilgileri Şekil 3'te gösterilmiştir.

	count	mean	std	min	25%	50%	75%	max
hemoglobin	803.0	9.896639	4.105754	0.150000	7.090000	10.300000	13.400000	17.700000
hemoglobin_a	803.0	52.230951	39.041988	3.600000	14.586340	16.977135	94.050000	99.100000
hemoglobin_a2	803.0	4.412949	2.336446	0.069137	2.487156	4.040000	6.049358	9.805597
hemoglobin_f	803.0	18.355261	24.485436	0.002473	0.650695	3.200000	36.017642	60.914394
mcv	803.0	67.352668	18.709565	30.362902	58.950000	69.000000	83.800000	100.100000
ferritin	803.0	246.318245	1742.217683	0.800000	23.607940	70.200000	155.148391	28392.000000
demir	803.0	79.439448	48.827132	2.000000	46.410770	74.000000	100.985000	365.000000
cinsiyet	803.0	0.398506	0.489896	0.000000	0.000000	0.000000	1.000000	1.000000
yaş	803.0	34.530511	18.721079	0.000000	22.000000	33.000000	49.000000	81.000000
hastalıkl durumu	803.0	0.485679	0.500106	0.000000	0.000000	0.000000	1.000000	1.000000

Şekil 3. Veri seti istatistik bilgisi.

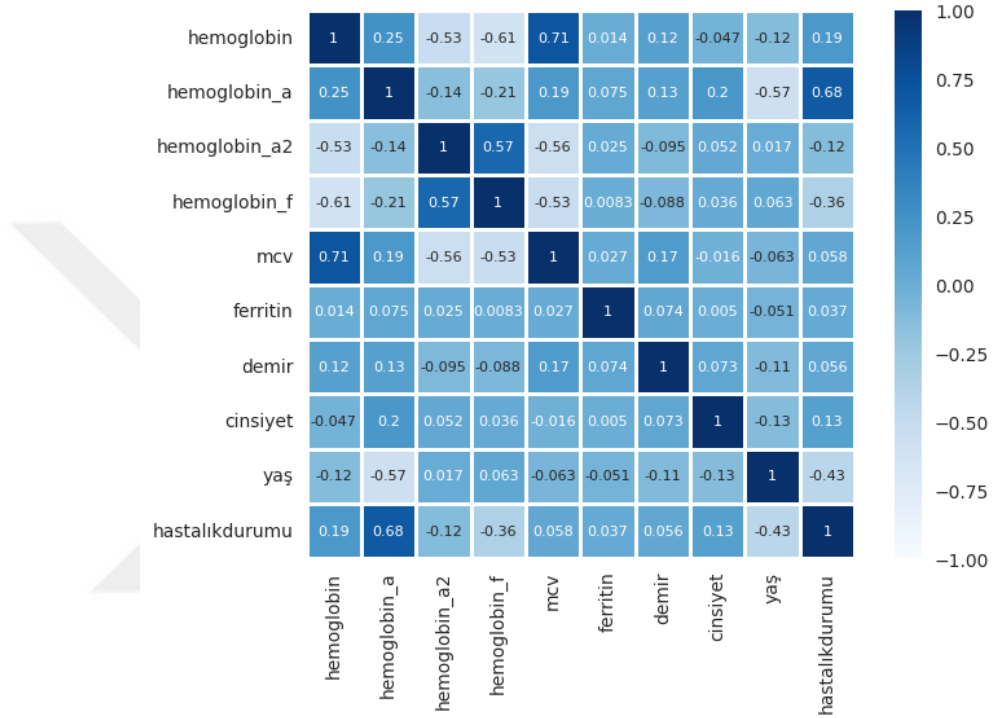
Veri setinde bulunan özniteliklerin dağılım grafikleri Şekil 4'de gösterilmiştir.



Şekil 4. Veri seti özniteliklerin dağılımı.

Veri setinde bulunan değişkenlerin birbirleriyle olan ilişkileri Şekil 5'te ısı haritası ile görülmektedir. Isı haritasında 1.0'a olan yakınlık değişkenler arasında pozitif ilişkiyi ve -1.0'e yakınlık ise değişkenler arasındaki negatif ilişkiyi göstermektedir. Şekil 5 incelendiğinde veri

setinde bulunan öznitelikler arasındaki korelasyon ısı haritası görülmektedir. Isı haritasına bakıldığında, öznitelikler arasında pozitif yönde 0.71 korelasyon değeri ile en yüksek ilişkiye sahip hemoglobin, mcv özniteliklerinin olduğu görülmektedir. Buna göre hemoglobin değerindeki artış ile mcv değerinin de arttığı söylenebilir. Şekilde öznitelikler arasında negatif yönde -0.61 korelasyon değeri ile en yüksek ilişkiye sahip hemoglobin, hemoglobin f özniteliklerinin olduğu görülmektedir. Bu sonuca göre hemoglobin değerindeki artışın hemoglobin f değerinde azalmasına neden olduğu yorumlanabilir.

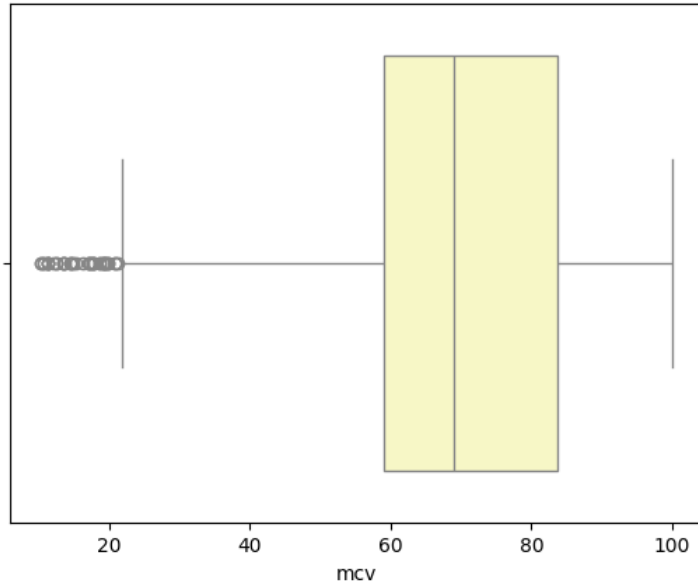


Şekil 5. Değişkenler arasındaki ısı haritası.

Veri Ön İşleme

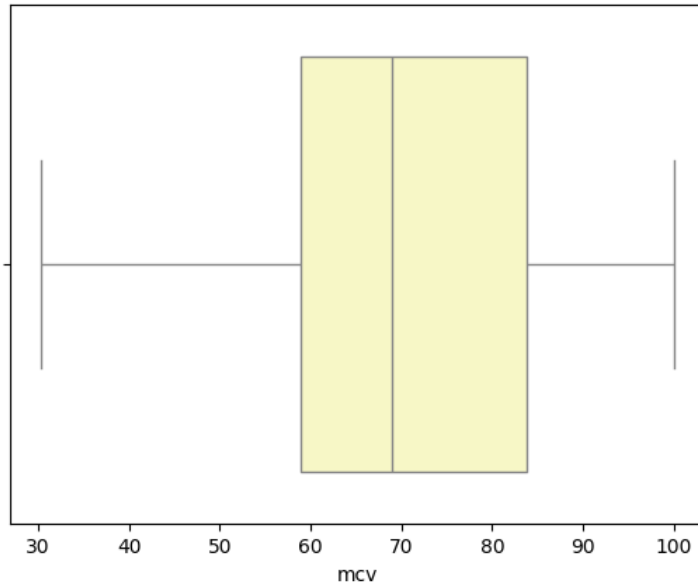
Veri ön işleme, kullanılan ML yöntemleri için önemli bir adımdır. Makine öğrenmesi sınıflandırma algoritmalarının doğru ve yüksek performanslı çalışmalarına neden olan birçok etken bulunmaktadır. Eksik gözlem, sınıflar arası dengesizlik, aykırı gözlem, gereksiz bilgiler bu etkenler arasındadır. Veri setinde bu etkenlerin olması eğitim aşamasında öğrenmeyi zor hale getirir. Bu çalışmada sınıflandırma sonuçlarından daha yüksek performans elde etmek amacı ile eksik veri tamamlama, tekrarlayan veri temizleme, aykırı değer yakalama, kategorik verileri sayısal verilere dönüştürme, veri ölçeklendirme gibi birçok veri ön işleme adımları uygulanmıştır. ML yöntemlerini uygulamadan önce veri setinde eksik verilerin var olup olmadığına bakılmış ve eksik veri olmadığı görülmüştür. Veri setinde aykırı değerler tespit edilmiş ve tek değişkenli yöntemlerden biri olan çeyrekler açıklığı yöntemi ile aykırı değerler

iyileştirilmiştir. Şekil 6’da veri setinde bulunan ‘mcv’ özneliği örnek olarak alınarak kutu grafiği (boxplot) gösterilmiştir.



Şekil 6. Orijinal veri seti mcv kutu grafik gösterimi.

Orijinal veri setinde aykırı değerler tespit edilip veri ön işleme yapıldıktan sonra ‘mcv’ özneliğindeki verilerin dağılımı Şekil 7’de gösterilmiştir.



Şekil 7. Veri ön işleme sonrasında mcv kutu grafik gösterimi.

Bu işlemler sonrasında veri setine normalizasyon işlemi uygulanmıştır. Verileri standart normal dağılıma dönüştürmek için Python programlama “sklearn.preprocessing” kütüphanesinde olan “StandartScaler” ölçeklendirme yöntemi kullanılmıştır. Bu ölçeklendirmede, veri setinde var olan her öznelik için ortalama değer çıkarılır ve standart sapmaya bölünür. Bu işlem sonucunda ise, ortalama değeri 0 ve standart sapma 1 olan normal

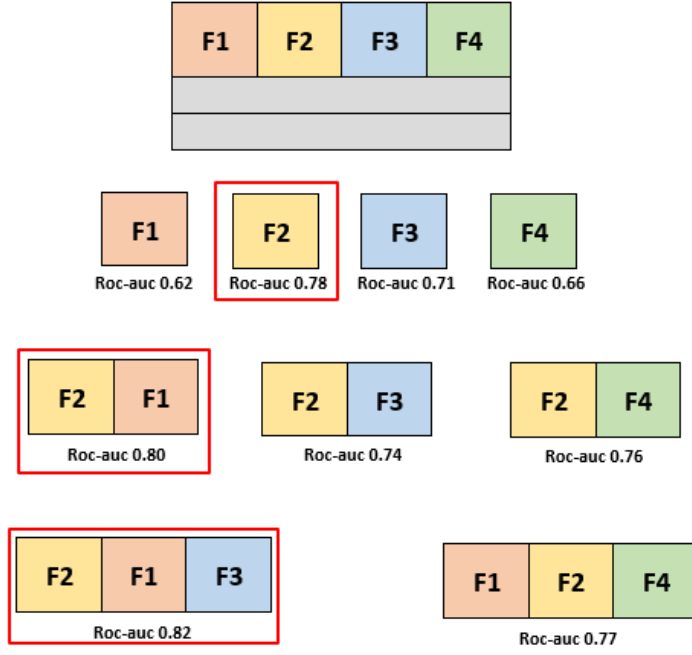
dağılım elde edilir. Bu sayede ölçek farklılıklarının ML yöntemleri üzerinde olumsuz etki oluşturmasının önüne geçilmesi sağlanır.

Öznitelik Seçimi (Feature Selection)

Öznitelik seçimi, sınıflandırma problemlerinde ve öğrenme algoritmalarında önemli bir yere sahiptir. ML algoritmalarının doğruluğunu ve verimliliğini artırmak amacı ile gereksiz veya önemsiz özniteliklerin çıkarılması sürecidir. Bu süreç, modelin performansını artırmak ve hesaplama maliyetlerini düşürmek için kullanılmaktadır. Bu yöntemlerde ML algoritmaları kullanılmaktadır. Öznitelik seçimi çeşitli yöntemlerle yapılmaktadır. Bu yöntemler, istatistiksel bilgi içeren Filtreleme Yöntemler (Filter Methods – FM), veri setinde bulunan öznitelikler üzerinde uygulanan Sarmal Yöntemler (Wrapper Methods – WM) ve Gömülü Yöntemler (Embedded Methods – EM) olmak üzere üç gruba ayrılmıştır (Saeys et al., 2007). Bu tez çalışmasında ele alınacak olan yöntem WM yöntemidir.

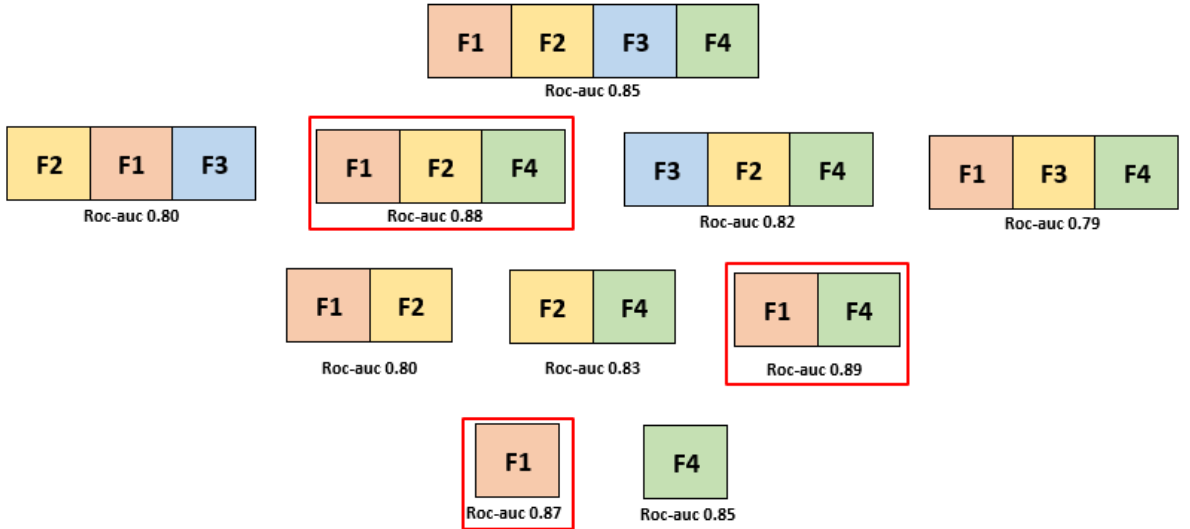
WM, öznitelik seçme sürecinde ML algoritmalarını doğrudan kullanarak özniteliklerin alt kümesini belirlemektedir. Bu yöntemler, oluşturulan modelin tahmin doğruluğunu artırmayı hedefler ve genellikle daha yüksek doğruluk sunar. WM yöntemleri üç ana adımda işlenmektedir. İlk olarak, farklı öznitelik alt kümeleri oluşturulur. İkinci olarak, oluşturulan bu alt kümeler seçilen bir ML algoritması kullanılarak değerlendirilir. Son olarak, en iyi performansı gösteren alt küme seçilir (Agraz, 2023). Bu çalışmada, Ardışık İleri Yönde Seçim (Sequential Forward Selection – SFS), Ardışık Geri Yönde Seçim (Sequential Backward Selection – SBS), Ardışık İleri Yönde Kayan Seçim (Sequential Forward Selection – SFFS), Ardışık Geri Yönde Kayan Seçim (Sequential Backward Floating Selection – SBFS) olarak dört adımda öznitelik seçme yöntemi gerçekleştirilecektir.

SFS yöntemi, öznitelik seçim sürecinde sıklıkla kullanılan tekniklerden birisidir. Bu yöntem, boş bir öznitelik kümesi ile başlayarak adım adım en iyi öznitelikleri seçer ve alt kümeye ekler. Her eklenen yeni öznitelik, mevcut öznitelik kümesi ile birlikte performans açısından değerlendirildikten sonra eklenir ve bu süreç durdurma kriteri sağlanana kadar devam eder (Cengiz & Demir, 2020). Şekil 8’de SFS yönteminin çalışma adımları gösterilmiştir.



Şekil 8. SFS yöntemi çalışma akışı.

SBS yöntemi, öznelilik seçimi sürecinde kullanılan önemli tekniklerinden birisidir. SBS, tüm özneliliklerden başlayarak her adımda en az önemli özelliği çıkarır ve bu işlem, en iyi alt küme bulunana kadar devam eder. Bu yöntem, özellikle karmaşık veri setlerinde gereksiz özneliliklerin çıkarılması ve model performansının artırılması yönünden kullanılması faydalıdır (Dursun et al., 2017). Şekil 9’da SBS yönteminin çalışma adımları gösterilmiştir.



Şekil 9. SBS yöntemi çalışma akışı.

SFFS yöntemi modelin doğruluğunu artırmak amacı ile en iyi öznelilik alt kümelerini belirlemeyi hedeflemektedir. SFS yönteminin bir çeşididir. SFS’de başlangıçta boş bir öznelilik kümesi ile başlanır ve her adımda mevcut kümeye en çok katkı sağlayan öznelilik eklenir.

Ancak SFFS yönteminde, eklenen her öznitelik sonrasında, kümenin performansını optimize etmek amacıyla geri adımlar atarak önceki adımlarda eklenmiş bazı öznitelikleri çıkarma esnekliğine sahiptir. Bu sayede, sadece ekleme değil aynı zamanda çıkarma işlemleri de yapılır. Bu durum, yöntemin kayan yapısını oluşturur ve daha optimal bir öznitelik alt kümesi elde edilmesini sağlar (Budak, 2018).

SBFS yöntemi, öznitelik seçim sürecinde kullanılan bir yöntemdir ve SFFS yöntemine benzer bir mantık ile çalışır. Ancak bu yöntemde başlangıç olarak tüm öznitelikler dahil edilir ve her adımda en az bilgi taşıyan öznitelik çıkarılır. Bu işlem sırasında, önceki adımlarda çıkarılan öznitelikler yeniden değerlendirilir ve gerekirse ekleme yapılır. Bu yöntem, özniteliklerin eklenip çıkarılarak en uygun öznitelik alt kümesinin dinamik bir şekilde belirlenmesini sağlar (Budak, 2018).

Bu çalışmada, var olan dört yöntem sınıflandırma algoritmalarında uygulanmıştır.

Makine Öğrenmesi (Machine Learning-ML)

ML yöntemi, verilerden öğrenme yeteneğine sahip algoritmaların geliştirilmesini amaçlayan bir YZ dalıdır. Temel olarak, bilgisayarların programlanmadan belirli görevleri nasıl yerine getireceğini öğrenme sürecidir. Bu teknoloji, çeşitli veri türlerini analiz ederek desenleri ve ilişkileri tanımlar ve bu bilgiler doğrultusunda gelecekteki veri üzerinde tahminler yapabilir veya kararlar alabilir (Özdemir & Örsülü, 2019; Şapcı & Pektaş, 2021).

ML, çeşitli istatistiksel, olasılıksal optimizasyon tekniklerini kullanarak, makinelerin geçmişteki örneklerden öğrenmesine ve büyük, gürültülü ya da karışık veri kümelerinden ayırt edilmesi zor verileri tespit etmeye olanak tanır (Cruz & Wishart, 2006). Veri analizi ve modelleme yetenekleri sayesinde, geleneksel yöntemlere kıyasla daha hızlı, daha doğru ve maliyet etkin çözümler sunmaktadır. ML temel olarak üç ana gruba ayrılmaktadır. Bunlar denetimli öğrenme (supervised learning), denetimsiz öğrenme (unsupervised learning) ve pekiştirmeli öğrenme (reinforcement learning).

Denetimli öğrenme, ML alanında yaygın olarak kullanılan bir tekniktir. Bu yöntem, belirli bir veri kümesi üzerinde eğitim yaparak yeni gelen veri sınıflandırma ya da tahmin etme yeteneğine sahip modeller oluşturmayı amaçlamaktadır. Denetimli öğrenme, etiketli bir veri kümesi kullanarak çalışmaktadır yani her giriş verisi bir doğru sonuç ile ilişkilidir (Calayır & Kabak, 2021). Bu süreçte, algoritma giriş verileri ve bu verilere karşılık gelen etiketler arasındaki ilişkiyi öğrenir. Denetimli öğrenme, genel olarak sınıflandırma ve regresyon problemlerinde kullanılmaktadır. Sınıflandırma problemlerinde algoritma, verileri önceden

tanımlanmış sınıflara ayırırken, regresyon problemlerinde ise sürekli bir çıktı tahmin edilmektedir. Bu öğrenme farklı alanlarda etkili bir şekilde kullanılabilir.

Denetimsiz öğrenme, ML yöntemlerinin önemli bir dalıdır ve veriler üzerinde herhangi bir etiketleme veya sınıflandırma yapmadan çalışır. Bu öğrenme yöntemi, veriler arasındaki gizli yapıları ve desenleri keşfetmek için kullanılır. Denetimsiz öğrenme, denetimli öğrenmeden farklı olarak belirli bir çıktıyı tahmin etmek yerine, veri setindeki ilişkileri ve yapıları keşfetmeye odaklanır. Veri madenciliği, anomali tespiti ve müşteri segmentasyonu gibi birçok uygulama alanında kullanılmaktadır (Sözer, 2022). Bu tür uygulamalarda, anormal davranışları tespit etmek için verilerdeki kalıpları belirlemek büyük önem taşımaktadır.

Pekiştirmeli öğrenme, bir ajanın belirli bir ortamda ödül veya ceza alarak en iyi stratejiyi öğrenmesini sağlayan bir ML tekniğidir. Bu öğrenme türü, ajanın çevresi ile etkileşime girerek deneyimlerinden öğrenmesini ve zaman içinde performansını iyileştirmesini hedefler. Pekiştirmeli öğrenme, genellikle oyunlar ve robotik gibi dinamik ve etkileşimli alanlarda kullanılmaktadır (Ospanov & Demirci, 2023).

Bu tez çalışmasında, ML sınıflandırma algoritmalarından NB, KNN, SVM, LR, RF ve DT kullanılarak gerekli çalışmalar yapılacaktır.

Naive Bayes – (NB)

NB algoritması, ML 'nde sıkça kullanılan bir denetimli öğrenme algoritmasıdır. Bu algoritma, Bayes teoremi temelinde çalışır ve sınıflandırma problemlerinde özellikle metin analizinde etkili bir şekilde kullanılmaktadır. NB algoritmasının 'naive' yani 'saf' olarak adlandırılmasının nedeni, sınıf koşullu bağımsızlık varsayımı yapması olarak nitelendirilir. Bu varsayımda, kullanılan her özelliğin diğerleri ile bağımsız olduğunu varsayar ve bu durum hesaplamaların basit olmasını ve uygulanmasını kolaylaştırır. Hesaplaması formül (1)'de gösterilmiştir.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \quad (1)$$

$P(A)$ = A olayının bağımsız olarak gerçekleşme olasılığı

$P(B)$ = B olayının bağımsız olarak gerçekleşme olasılığı

$P(A|B)$ = B olayının doğru olduğu durumda A olayının gerçekleşme olasılığı

$P(B|A)$ = A olayının doğru olduğu durumda B olayının gerçekleşme olasılığı

NB algoritması, seçilen bir elemanın her durumun olasılığını hesaplar ve elde edilen olasılık değeri en yüksek olan hangisi ise ona göre sınıflandırma yapılır. Küçük veya dengesiz

dağılılan veri setleri üzerinde de bu algoritma çalışmaktadır ve basit sınıflandırma yöntemleri arasında yer alır. Geniş bir uygulama alanına sahiptir ve özellikle metin ve belge sınıflandırması gibi problemlerde etkin bir şekilde kullanılmaktadır. Algoritma çalışırken değişkenler arasındaki ilişki gösterilmez ve büyük, dinamik veri setleri ile çalışılması durumunda veri setinin eğitilmesi aşamasında yavaş çalışmasına sebep olmaktadır.

K En Yakın Komşu – (K Nearest Neighbour – KNN)

KNN algoritması, ML yöntemlerinde sıkça kullanılan bir denetimli öğrenme algoritmasıdır. Bu algoritma, sınıflandırma ve regresyon problemlerinde kullanılan basit bir yapıdır. Bu algoritmanın temel fikri bir veri noktasını sınıflandırmak ya da tahmin etmek amacı ile en yakınındaki k adet komşunun etiketlerini ya da değerlerini kullanmaktır. Bu k değeri yapılan hesaplamalardan sonra oluşan verilerden en yakın kaç adet veriye bakılacağını belirleyen değerdir ve genellikle tek sayı olarak belirlenmektedir. Bu algortmada genel olarak Öklid, Manhattan veya Minkowski mesafesi gibi uzaklık ölçüm yöntemleri kullanılır. Çalışma performansı k komşu sayısına, uzaklık ölçüleri gibi parametrelere bağlıdır. Öklid mesafesi, veri noktaları arasındaki geometrik uzaklıkları hesaplamak için kullanılır. k boyutlu uzayda öklid uzaklık hesaplaması formül (2)'de gösterilmiştir.

$$\sqrt{\sum_{i=1}^k (x_i - y_i)^2} \quad (2)$$

Diğer uzunluk hesaplama yöntemi olan Manhattan ise noktalar arasındaki mutlak uzaklıkların toplamı olarak hesaplanır. Uzaklık formülü formül (3)' de gösterilmiştir.

$$\sum_{i=1}^k |x_i - y_i| \quad (3)$$

En son uzunluk yöntemi ise k sayıda değişkene bağlı olarak uzaklık hesaplanan Minkowski yöntemidir. Uzaklık formülü formül (4) 'te gösterilmiştir.

$$\left(\sum_{i=1}^k |x_i - y_i|^p \right)^{1/p} \quad (4)$$

KNN algoritması diğer algoritmalarından farklı olarak önce bir model oluşturmadan verileri doğrudan sınıflandırma yapmak için kullanır. Bu nedenle modelin hiçbir detayının dikkate alınmasına gerek duyulmaz ve ayarlanması gereken tek parametre en yakın komşu sayısı olan k değeridir. Bu algoritmanın çalışmalarda eğitimi, anlaşılması ve uygulanması

oldukça kolaydır. Büyük verilerde kullanımı zordur. Bu durumun sebebi kullanılan veriler arasındaki uzaklık hesaplanırken bilgilerin saklanarak yapılmasıdır.

Destek Vektör Makinesi (Support Vector Machine – SVM)

SVM algoritması, ML yöntemlerinde sıkça kullanılan bir denetimli öğrenme algoritmasıdır. Bu algoritma, sınıflandırma ve regresyon problemlerinde kullanılan güçlü bir ML algoritmasıdır. Veri noktalarını sınıflandırmak ya da bir regresyon çizgisini belirlemek için kullanılmaktadır.

SVM' nin temel amacı, veri noktalarını sınıflandırmak için optimum bir karar sınırı oluşturmaktır. Oluşturulan bu sınır, veri noktalarının en iyi şekilde ayrılmasını sağlayacak şekilde belirlenmektedir ve bu noktalar destek vektörleri olarak adlandırılır. Bu algoritma, doğrusal ve doğrusal olmayan veri setlerinde uygulanabilir, hızlı tahmin gerçekleştirirler ve hızlı tahminleme yapabilirler. Büyük verilerde kullanımı eğitim süresinin uzun olması nedeniyle tercih edilmemektedir.

SVM algoritmaları doğrusal destek vektör makineleri ve doğrusal olmayan destek vektör makineleri adı altında iki gruba ayrılmaktadır.

Lojistik Regresyon (Logistic Regresyon – LR)

LR algoritması, ML yöntemlerinde sıkça kullanılan bir denetimli öğrenme algoritmasıdır. Bu algoritma, sınıflandırma problemlerinde kullanılan temel istatistiksel bir modeldir. LR algoritmasında temel amaç bağımsız (giriş) değişkenlerinin bir lineer kombinasyonu ile bağımlı (sınıf etiketi) değişkenin olasılığını tahmin etmektir. Bu tahmin için sigmoid fonksiyonu kullanılmaktadır. Bu fonksiyon verileri 0 ve 1 arasında sıkıştırmak için kullanılan bir fonksiyondur (Sevli, 2022). Sınıflandırma yapılmasını sağlayan sigmoid fonksiyonu formül (5) 'te gösterilmiştir.

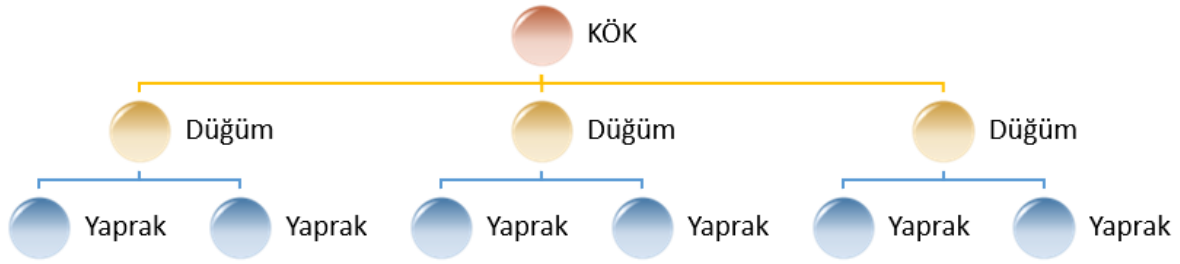
$$\sigma(x) = \frac{1}{e^{-(Qx)}} \quad (5)$$

Formülde, x değeri girdi verilerini temsil etmektedir. Q katsayı (ağırlık) değeridir. LR algoritması veri seti doğrusal ayrılabilirdiği durumlarda oldukça iyi performans göstermektedir ve uygulanması, yorumlanması oldukça kolaydır.

Karar Ağacı (Decision Tree – DT)

DT algoritması, ML yöntemlerinde sıkça kullanılan bir denetimli öğrenme algoritmasıdır. Veri madenciliğinde çok sık kullanılan sınıflandırma ve regresyon problemlerinde kullanılan bir ML algoritmasıdır. Bu yöntem, verileri ağaç yapısı üzerinden

analiz eder ve karar düğümleri, dallar ve yaprak düğümlerinden oluşmaktadır (Atasoy & Kara, 2023). Şekil 11’de DT algoritmasının yapısı gösterilmiştir.



Şekil 10. Karar Ağacı algoritma yapısı.

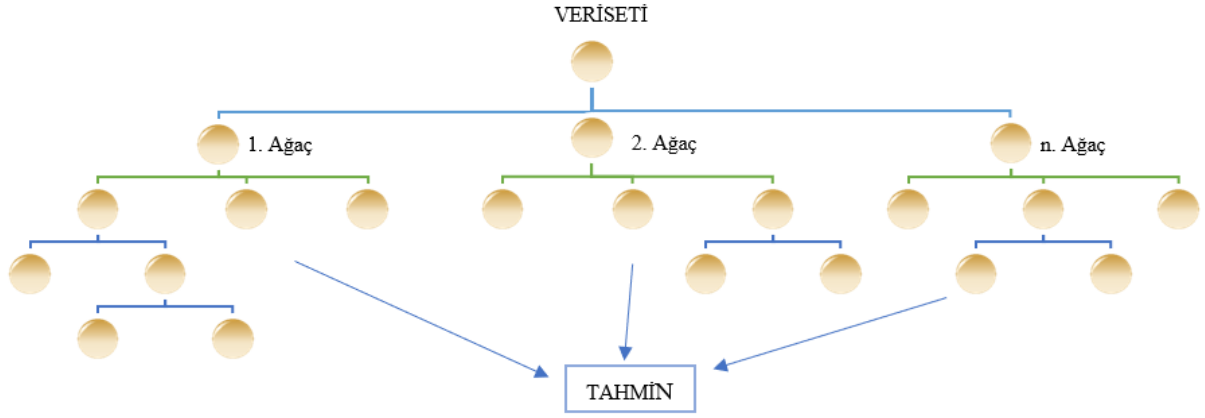
DT çalışma şekli, kök düğümünden başlayarak çalışılan veri seti özelliklerine göre dallara ayrılır. Her bir düğümde bir karar verilir ve verilen bu kararlar sonucunda son yaprak düğümlere ulaşır. Yaprak düğümler, sınıflandırma problemlerinde sınıf etiketlerini, regresyon problemlerinde ise tahmin edilen değerleri ifade etmektedir (Doğruel & Fırat, 2021).

DT algoritmalarının önemli avantajlarından birisi, verileri açıklayıcı bir şekilde görselleştirmesi ve karar verme sürecini kolaylaştırmasıdır. Büyük veri setlerinde etkili bir şekilde çalışabilir ve eksik veriler ile başa çıkabilir ancak karar ağaçlarının aşırı uyum sorununa eğilimli olabileceği söylenebilir ve bu sorunu azaltmak için budama (pruning) teknikleri kullanılabilir (Atasoy & Kara, 2023).

Rastgele Orman (Random Forest – RF)

RF algoritması, ML yöntemlerinde sıkça kullanılan bir denetimli öğrenme algoritmasıdır. Sınıflandırma algoritmaları arasında sıkça kullanılan regresyon problemlerinde ve sınıflandırmalarda kullanılan bir yöntemdir (Breiman, 2001). RF, birden fazla DT ’nin birleşiminden oluşur ve her bir ağacın elde edilen sonuçları üzerinden bir tahmin yapılmaktadır.

RF çalışma şekli, eğitim aşamasında veri setinden rastgele örnekler seçerek birçok DT oluşturur. Bu oluşturulan her bir DT, veri setindeki özelliklerin rastgele bir alt kümesi ile eğitilir. RF algoritmasının diğer algoritmalarından farklı kılan bir özelliği oluşturduğu farklı DT modelleri oluşturması ve değerlendirme yaparken farklı ağaçları bir araya getirip kullanabilmesidir. RF algoritmasının çalışma şeması Şekil 12’de gösterilmiştir.



Şekil 11. Rastgele orman algoritması çalışma şeması.

RF algoritması sayısal ve kategorik sınırlaması olmadan tüm etiketler için kullanılabilir. RF algoritmasında çeşitlendirme iki şekilde sağlanır. İlk olarak her ağaç için eğitim verisi alt kümelerin rastgele seçimi ile, ikinci olarak da her bir düğüm bölünmesi için özelliklerin rastgele alt kümelerinin seçilmesi olarak sağlanmaktadır. Bu çeşitlendirme, oluşan ağaçlar arasındaki korelasyonu azaltır ve modelin performansını artırır. Sınıflandırma problemlerinde, her bir ağaç sınıf tahmini yapar ve bu sınıflandırma çoğunluğa göre belirlenir (Sel, 2021).

Sınıflandırma Algoritmalarının Performansını Test Etme Ölçütleri

Çalışmada kullanılan sınıflandırma sonuçlarının ve başarı performansların belirlenmesinde kullanılan çeşitli yaklaşımlar vardır. Sınıflandırmada kullanılan algoritmaların performans değerlendirmesinde karmaşıklık matrisleri oluşturulmuş daha sonra Kesinlik (Precision), Duyarlılık (Recall), F1-Skoru (F1-Score), Doğruluk (Accuracy), İşlem Karakteristik Eğrisi (ROC-AUC) gibi kriterlere göre algoritma performanslarının başarı oranları arasında karşılaştırma yapılmıştır.

İkili sınıflandırmada, problem için elde edilen sonuçlar Pozitif (P) ve Negatif (N) olarak etiketlenmektedir. 2x2 karmaşıklık matrisi olası dört sonuç için her biri tarafından tahmin edilen örneklerin sayısını vermektedir. Bu tahminler; doğru pozitif (True Positive-TP), yanlış pozitif (False Positive-FP), doğru negatif (True Negative-TN) ve yanlış negatif (False Negatif-FN) kullanılarak değerlendirme yapılır (Ozcift & Gulten, 2011). Kullanılan sınıflandırıcıların performans ölçütlerini türetmek için Tablo 2’de gösterilen karmaşıklık matris tablosu kullanılmaktadır.

Tablo 2. Karmaşıklık Matrisi (Confusion Matrix)

	Gerçek Pozitif (1)	Gerçek Negatif (0)
Tahmin Pozitif (1)	True Positive (TP)	False Positive (FP)
Tahmin Pozitif (0)	False Negative (FN)	True Negative (TN)

TP: Talasemi hastası olarak etiketlenmiş örneğin talasemi hastası olması

FP: Talasemi hastası olmayan bir örneğin talasemi hastası olarak etiketlenmesi

TN: Talasemi hastası olmayan bir örneğin hasta olmaması

FN: Talasemi hastası olan bir örneğin talasemi hastası olmadığına etiketlenmesi diye açıklanabilir.

Kesinlik (Precision)

Kesinlik metriği, oluşturulan modelde pozitif denilen örneklerin ne kadarının gerçekten pozitif olduğunu göstermektedir.

$$Kesinlik = \frac{TP}{TP + FP} \quad (6)$$

Duyarlılık (Recall)

Duyarlılık metriği; oluşturulan modelde pozitif denilen örneklerin ne kadarının pozitif tahmin edildiğini göstermektedir.

$$Duyarlılık = \frac{TP}{TP + FN} \quad (7)$$

F1 Skoru (F1 Score)

F1 Skoru, kesinlik ve duyarlılık metriklerinin arasındaki ve dengeyi gösterir ve harmonik ortalama ile hesaplanmaktadır (Sevli, 2022).

$$F1 Skoru = 2x \frac{Duyarlılık \times Kesinlik}{Duyarlılık + Kesinlik} \quad (8)$$

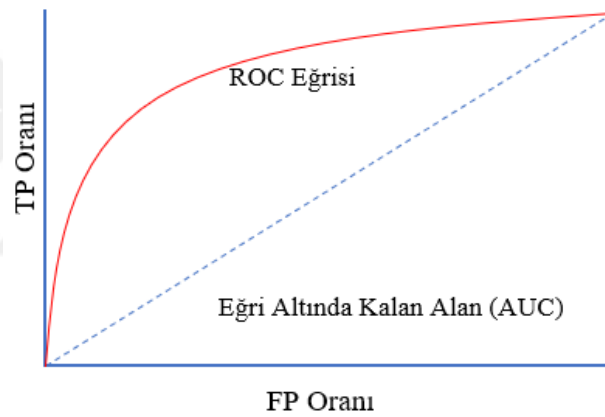
Doğruluk (Accuracy-ACC)

Doğruluk, oluşturulan modelin genel başarısını göstermektedir. Veri setinin dengeli dağılması durumunda etkili olduğu söylenebilir.

$$\text{Doğruluk} = \frac{TP + TN}{TP + FP + TN + FN} \quad (9)$$

ROC – AUC Eğrisi (ROC – AUC Curve-ROC-AUC)

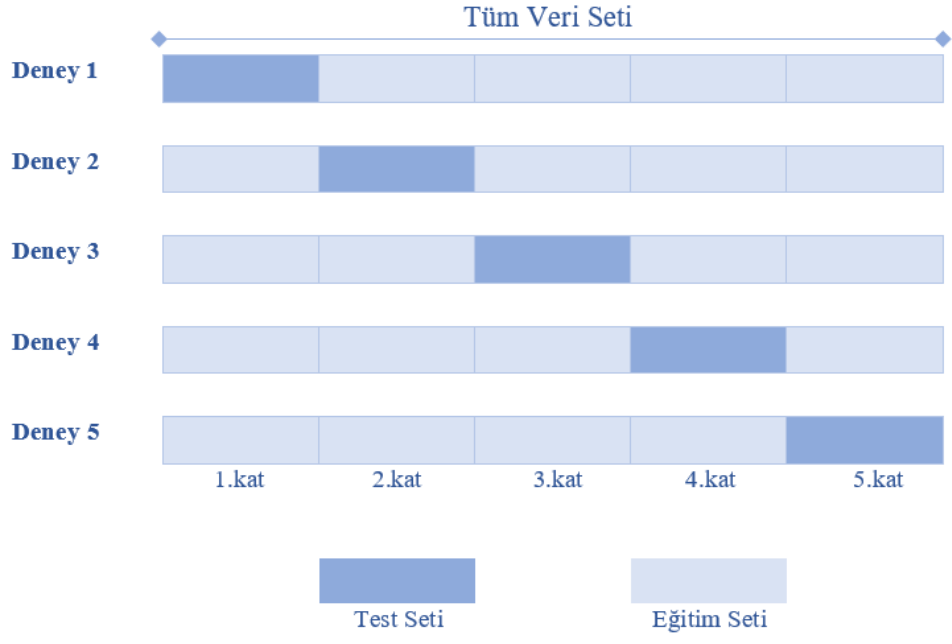
ROC eğrisi, duyarlılık ve yanlış pozitif (FP) ile oranından elde edilir. Eksende (0,0) ve (1,1) noktaları arasında yer almaktadır. AUC, ROC eğrisi altında kalan alanı ifade eder ve 0 ile 1 arasında değer almaktadır. Eğri altında kalan alan (AUC) değeri ne kadar 1'e yakın olursa model performans başarısı o kadar yüksek olmaktadır (Eröz, 2010). Şekil 13'te ROC-AUC eğri grafiği gösterilmiştir.



Şekil 12. ROC – AUC eğri grafiği.

K Katlı Çapraz Doğrulama (K Fold Cross Validation-CV)

CV, oluşturulan model performansını etkileyen önemli bir adımdır. k değerinin seçimi performans başarısı için oldukça önemlidir. Genel olarak çalışmalarda k değeri 5 veya 10 olarak verilmektedir (Nti et al., 2021). Bu yöntemde veri rastgele karıştırılır ve belirlenen k değerine göre parçalanır. Her parça test ve eğitim verisi olarak kullanılır. Bu yöntemle hata oluşması azaltılmış olur. Bu tez çalışmasında, k değeri 5 seçilmiştir. Şekil 10'da 5 tane k değeri için eğitim ve test verilerine göre çapraz doğrulama adımları gösterilmiştir.



Şekil 13. k katlı çapraz doğrulama.

ARAŞTIRMA BULGULARI VE TARTIŞMA

Bu çalışmada, talasemi hastası olan ve sağlıklı 803 adet örnekten oluşan veri seti üzerinde veri ön işleme, öznitelik seçme işlemleri uygulanıp NB, KNN, SVM, LR, RF, DT makine öğrenme algoritmaları ile ayrı ayrı sınıflandırılma işlemleri yapılmıştır. Çalışmada, 803 adet örnekten oluşan veri ön işleme adımları gerçekleştirilen veri seti, %70'i eğitim, %30'u test verisi olarak kullanılmak üzere ayrılmıştır.

Veri setine uygulanacak olan ML algoritmalarının daha iyi sonuçlar vermesi için, verileri normal dağılıma dönüştürmek gerekmektedir. Normal dağılıma dönüştürmek için Standart Scaler yöntemi kullanılmıştır. Bu yöntem, var olan her özellik için ortalama bir değer çıkarır. Bu ortalama değer standart sapmaya bölünür. Sonuçta ise ortalama değerinin 0, standart sapma değerinin 1 olduğu normal dağılım elde edilmiş olur. Yöntemin hesaplaması formül (10)'da gösterilmiştir.

$$\frac{X - \text{mean}(X)}{\text{std}(X)} \quad (10)$$

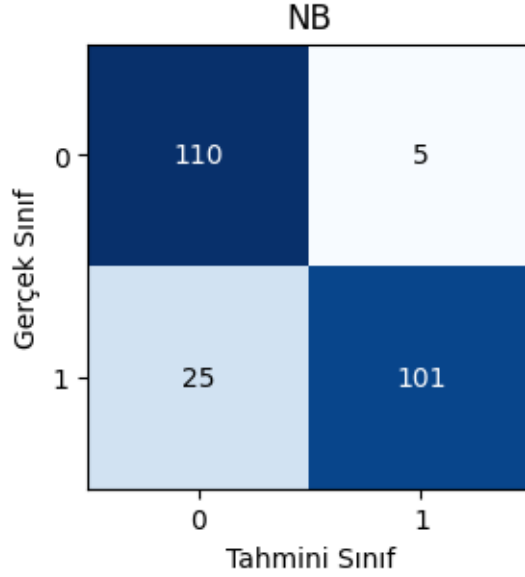
Öznitelik Seçim Yöntemleri Uygulanmadan Önce Sınıflandırma

Tez çalışmasında, yapılan veri ön işleme, veri ölçeklendirme adımlarından sonra herhangi bir öznitelik seçme ve CV yöntemi kullanılmadan önce oluşturulan ML modellerinden NB algoritmasından elde edilen sonuçlar Tablo 3'te gösterilmiştir. Elde edilen sonuç doğrultusunda NB algoritmasının en yüksek performansı %87.55 doğruluk değeri ve %87.91 AUC değeri ile göstermiştir.

Tablo 3. Veri Setinde Yöntem Uygulanmadan Önce NB Performans Metrikleri

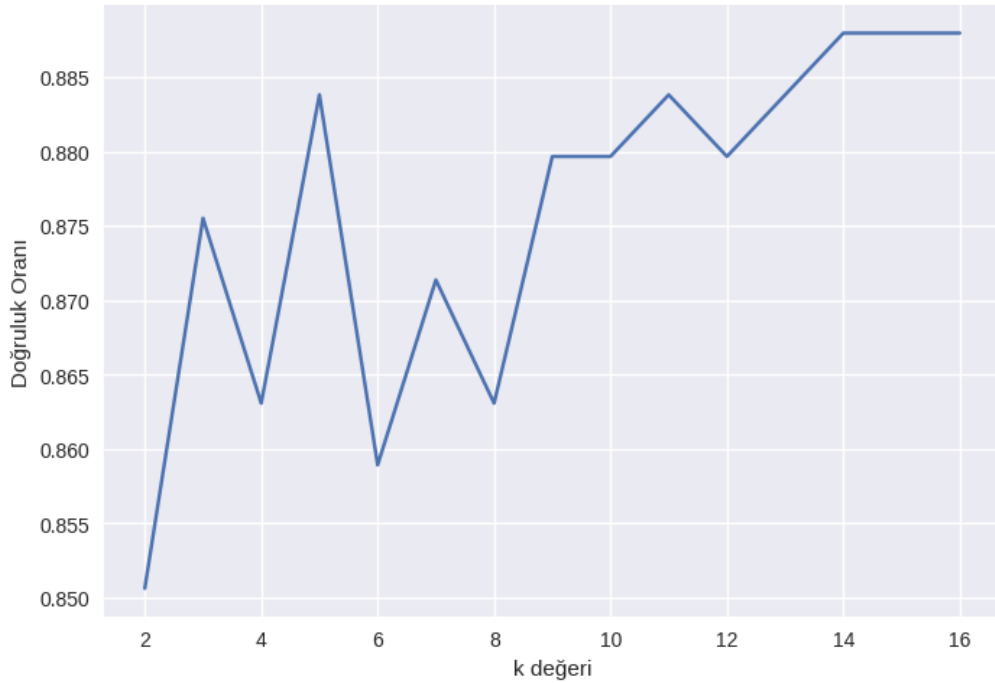
ML Algoritması	Doğruluk (%)	Duyarlılık (%)	Kesinlik (%)	F1 Skor (%)	AUC (%)
NB	87.55	87.55	88.70	87.51	87.91

Şekil 14'te gösterilen NB modeline ait karmaşıklık matrisi değerlendirildiğinde 5 talasemi hastası olan bireyin ve 25 talasemi hastası olmayan bireyin yanlış sınıflandırıldığı görülmektedir.



Şekil 14. NB modeli karmaşıklık matrisi.

Tez çalışmasında, veri setine yapılan veri ön işleme, veri ölçeklendirme adımlarından sonra KNN algoritması uygulanmıştır. Sınıflandırma yaparken “n_neighbors” parametresi yani k değeri için 2 ile 17 arasında tüm değerler denenmiş ve en optimum değer 14 olarak belirlenmiştir. Veri seti için elde edilen en uygun k değeri Şekil 15’te gösterilmiştir. Bununla birlikte uzaklık ölçütü olarak “minkowski” ölçüm yöntemi kullanılmıştır.



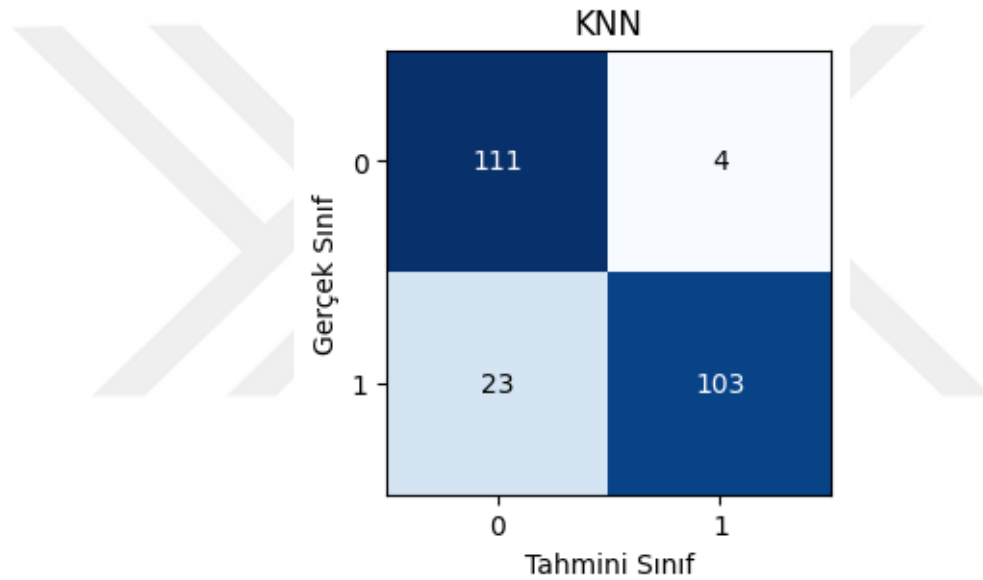
Şekil 15. KNN algoritması için optimum k değerini belirleme grafiği.

Yapılan işlemler sonrasında, KNN algoritmasından elde edilen sonuçlar Tablo 4'te gösterilmiştir. Elde edilen sonuç doğrultusunda KNN algoritmasının en yüksek performansı %88.80 doğruluk değeri ve %89.13 AUC değeri ile göstermiştir.

Tablo 4. Veri Setinde Yöntem Uygulanmadan Önce KNN Performans Metrikleri

ML Algoritması	Doğruluk (%)	Duyarlılık (%)	Kesinlik (%)	F1 Skor (%)	AUC (%)
KNN	88.80	88.80	89.86	88.77	89.13

Şekil 16'da gösterilen KNN modeline ait karmaşıklık matrisi değerlendirildiğinde 4 talasemi hastası olan bireyin ve 23 talasemi hastası olmayan bireyin yanlış sınıflandırıldığı görülmektedir.



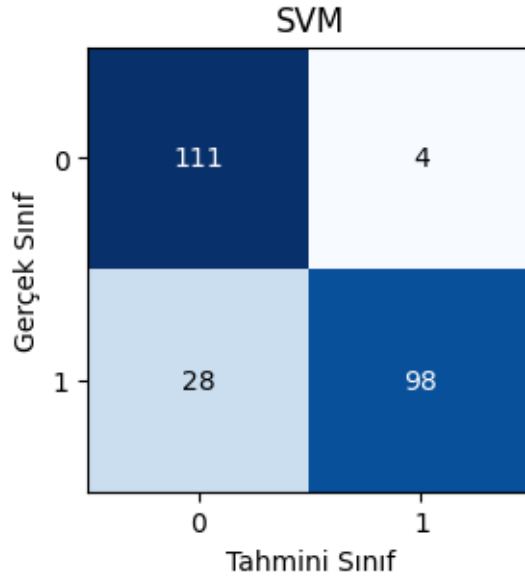
Şekil 16. KNN modeli karmaşıklık matrisi.

Tez çalışmasında, veri setine yapılan veri ön işleme, veri ölçeklendirme adımlarından sonra SVM algoritması uygulanmıştır. SVM algoritmalarında model performansını etkileyen çekirdek (kernel) fonksiyonları vardır. Sınıflandırma yaparken “linear” çekirdek fonksiyonu kullanılmıştır. Yapılan işlemler sonrasında, SVM algoritmasından elde edilen sonuçlar Tablo 5'te gösterilmiştir. Elde edilen sonuç doğrultusunda SVM algoritmasının en yüksek performansı %86.72 doğruluk değeri ve %87.15 AUC değeri ile göstermiştir.

Tablo 5. Veri Setinde Yöntem Uygulanmadan Önce SVM Performans Metrikleri

ML Algoritması	Doğruluk (%)	Duyarlılık (%)	Kesinlik (%)	F1 Skor (%)	AUC (%)
SVM	86.72	86.72	88.34	86.65	87.15

Şekil 17’de gösterilen SVM modeline ait karmaşıklık matrisi değerlendirildiğinde 4 talasemi hastası olan bireyin ve 28 talasemi hastası olmayan bireyin yanlış sınıflandırıldığı görülmektedir.



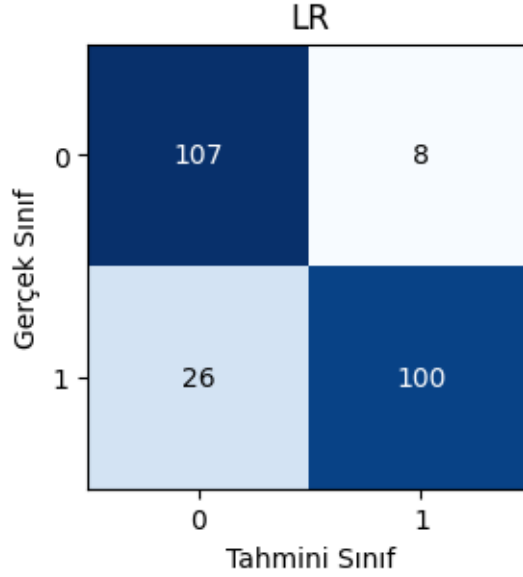
Şekil 17. SVM modeli karmaşıklık matrisi.

Tez çalışmasında, veri setine yapılan veri ön işleme, veri ölçeklendirme adımlarından sonra LR algoritması uygulanmıştır. Sınıflandırma yaparken “random_state” parametresi 42 olarak alındı. Bu parametre sınıflandırma aşamasında veri setinde farklı yerlerden bölme yaparak ezberlemenin önüne geçer. Yapılan işlemler sonrasında, LR algoritmasından elde edilen sonuçlar Tablo 6’da gösterilmiştir. Elde edilen sonuç doğrultusunda LR algoritmasının en yüksek performansı %85.89 doğruluk değeri ve %86.20 AUC değeri ile göstermiştir.

Tablo 6. Veri Setinde Yöntem Uygulanmadan Önce LR Performans Metrikleri

ML Algoritması	Doğruluk (%)	Duyarlılık (%)	Kesinlik (%)	F1 Skor (%)	AUC (%)
LR	85.89	85.89	86.80	85.86	86.20

Şekil 18’de gösterilen LR modeline ait karmaşıklık matrisi değerlendirildiğinde 8 talasemi hastası olan bireyin ve 26 talasemi hastası olmayan bireyin yanlış sınıflandırıldığı görülmektedir.



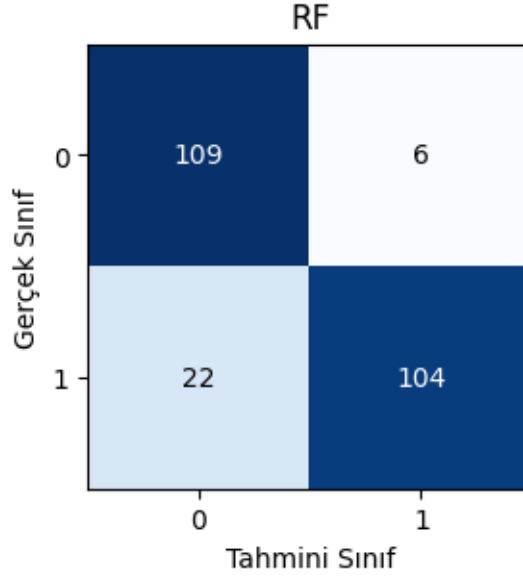
Şekil 18. LR modeli karmaşıklık matrisi.

Tez çalışmasında, veri setine yapılan veri ön işleme, veri ölçeklendirme adımlarından sonra RF algoritması uygulanmıştır. Sınıflandırma yaparken “n_estimators” parametresi yani ağaç sayısı belirlenmesi için farklı sayılarda ağaç sayıları ile model oluşturulmuştur. Yapılan işlemler sonrasında, RF algoritmasından elde edilen sonuçlar Tablo 7’de gösterilmiştir. Elde edilen sonuç doğrultusunda RF algoritmasının en yüksek performansı 50 ağaç sayısı kullanılarak, %89.63 doğruluk değeri ve %89.78 AUC değeri ile göstermiştir.

Tablo 7. Veri Setinde Yöntem Uygulanmadan Önce RF Performans Metrikleri

ML Algoritması	Ağaç Sayısı	Doğruluk (%)	Duyarlılık (%)	Kesinlik (%)	F1 Skor (%)	AUC (%)
RF	50	89.63	89.88	89.63	89.63	89.78
	100	88.80	88.80	89.47	88.79	89.06
	150	89.21	89.21	89.97	89.20	89.49

Şekil 19’da gösterilen RF modeline ait karmaşıklık matrisi değerlendirildiğinde 6 talasemi hastası olan bireyin ve 22 talasemi hastası olmayan bireyin yanlış sınıflandırıldığı görülmektedir.



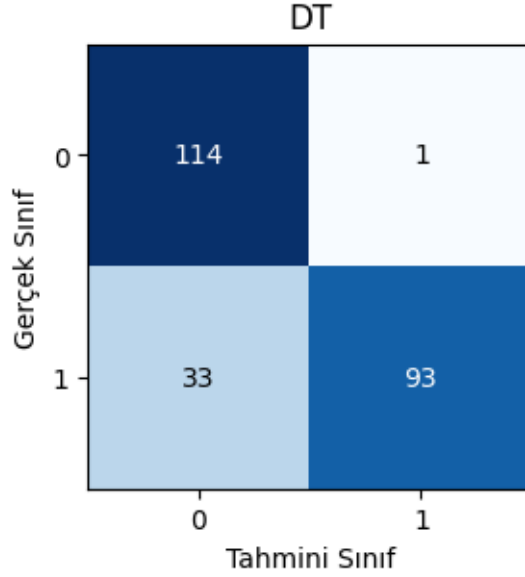
Şekil 19. RF modeli karmaşıklık matrisi.

Tez çalışmasında, veri setine yapılan veri ön işleme, veri ölçeklendirme adımlarından sonra DT algoritması uygulanmıştır. Sınıflandırma yaparken “criterion” parametresi yani her bir düğümün information gain değerini görmek için “entropy” olarak tanımlanmıştır. Oluşturulan modellerde “max_depth” yani ağaç derinliği 1, 2, 3, 4 ve 5 olarak seçilmiştir. Ağaç derinliği kök düğüm ile oluşan yaprak arasındaki yolu belirtir ve ağacın maksimum derinliği demektir. Sınıflandırma sonunda optimum ağaç derinliği ile sınıflandırma sonuçları doğruluk, hassasiyet, duyarlılık, f1 skor ve ROC-AUC ölçüm metrikleri kullanılarak değerlendirilmiştir. Tablo 8’de DT algoritması sonucunda oluşturulan modelin 1-5 aralığında ağaç derinliğine göre elde edilen performans metrik değerleri gösterilmiştir. Elde edilen sonuç doğrultusunda DT algoritmasının en yüksek doğruluk ve AUC değerine, 5 ağaç derinliğinde %85.89 doğruluk değeri ve %86.46 AUC değeri ile ulaşılmıştır.

Tablo 8. Veri Setinde Yöntem Uygulanmadan Önce DT Performans Metrikleri

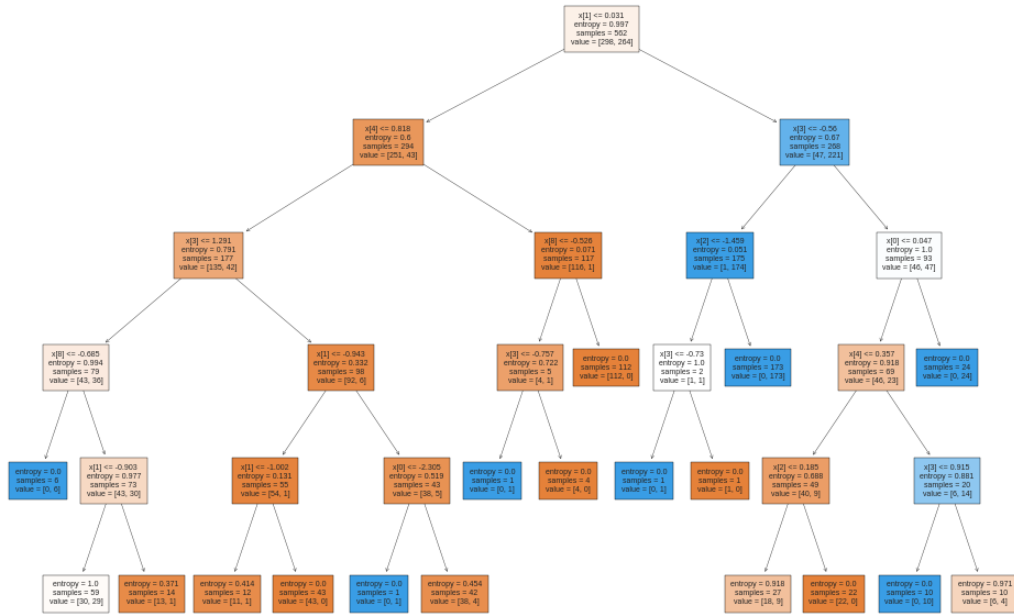
ML Algoritması	Ağaç Derinliği	Doğruluk (%)	Duyarlılık (%)	Kesinlik (%)	F1 Skor (%)	AUC (%)
DT	1	83.40	83.40	83.70	83.41	83.55
	2	83.40	83.40	83.70	83.41	83.55
	3	82.99	82.99	87.46	82.62	83.73
	4	85.06	85.06	87.51	84.91	85.60
	5	85.89	85.89	88.73	85.73	86.46

Şekil 20’de gösterilen DT modeline ait karmaşıklık matrisi değerlendirildiğinde 1 talasemi hastası olan bireyin ve 33 talasemi hastası olmayan bireyin yanlış sınıflandırıldığı görülmektedir.



Şekil 20. DT modeli karmaşıklık matrisi.

Seçilen parametreler doğrultusunda oluşan DT yapısı Şekil 21’de gösterilmiştir



Şekil 21. Sınıflandırma sonucunda oluşan DT yapısı.

Çalışılan modellere herhangi bir yöntem uygulamadan yapılan sınıflandırma sonucunda %89.63 doğruluk ile RF diğer modellere göre en yüksek performansı göstermiştir. Veri setinin %70’i eğitim, %30’u test verisi olarak kullanılmak üzere ayrılmıştır. Eğitim ve test verisi olarak ayrılan veri setinde değişkenlerin dağılımlarına bağlı olarak sapmalar olabilmektedir. Bu sapmaların önüne geçebilmek için kullanılan modelde k katlı çapraz doğrulama (k fold cross validation-CV) yöntemi uygulanmıştır. Tez çalışmasında, k sayısı 5 olarak alınmıştır. CV yöntemi uygulandıktan sonra sınıflandırma sonucunda %87.00 doğruluk ile RF diğer modellere

göre en yüksek performansı göstermiştir. CV uygulanma sonucunda model performanslarında çok azda olsa azalma olmuştur. Modellerin CV yöntemi uygulamadan öncesi ve sonrası doğruluk değerleri Tablo 9’da gösterilmiştir.

Tablo 9. K Katlı Çapraz Doğrulama Uygulamadan Önce ve Sonrası Modellerin Doğruluk Değerleri

ML Algoritmaları	CV Uygulanmayan ACC (%)	CV Uygulanan ACC (%)
NB	87.55	84.64
KNN	88.80	83.15
SVM	86.72	82.89
LR	85.89	83.77
RF	89.63	87.00
DT	85.89	86.01

Öznitelik Seçim Yöntemleri Uygulandıktan Sonra Sınıflandırma

ML sınıflandırma işlemlerinde en iyi değerlere ulaşabilmek için her sınıflandırma algoritması için öznitelik seçim yöntemleri kullanılmıştır. Sırasıyla SFS, SBS, SFFS, SBFS yöntemleri modellere uygulanmıştır.

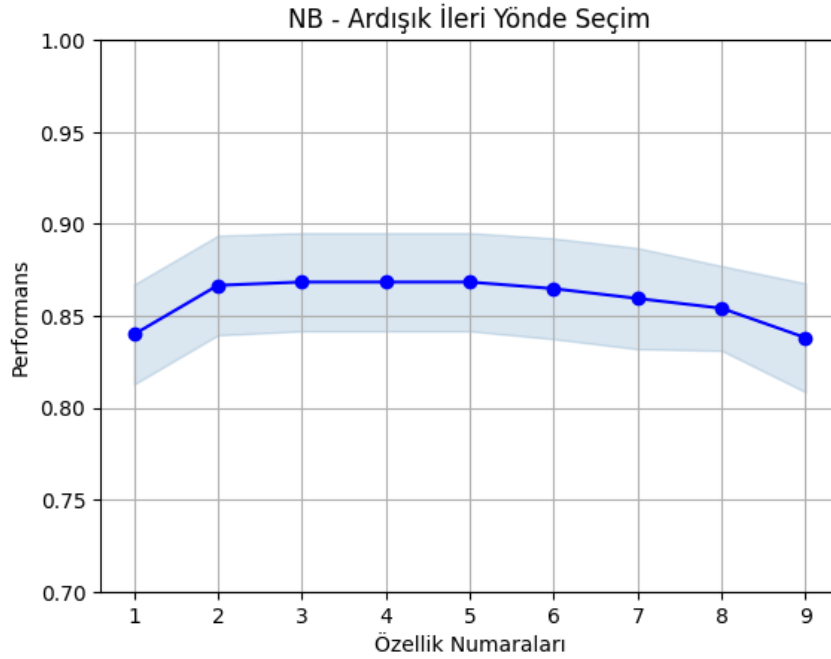
SFS öznitelik seçim yöntemi ile ML algoritmalarının modellenmesi

SFS yöntemi, her model için uygulanmıştır. Bu yöntemde bazı parametreler kullanılmıştır. İlk olarak kaç öznitelik kullanılacağına karar veren “k_features” parametresi (1,9) aralığı seçilmiştir. Veri setinde 10 öznitelik bulunmaktadır. Bu parametre ile özniteliklerin hepsine bakar ve en yüksek doğruluk değerine hangi öznitelikler ile ulaşırsa onları seçer. SFS yöntemi ardışık ileriye doğru olacağı için “forward” parametresi “True” ve “floating” parametresi de “False” olarak kullanılmıştır. İşlemden her tekrarda model özetini ayrıntılı görmek için “verbose” parametresi kullanılmıştır. Model sonucunu doğruluk değeri olarak görmek için “scoring” parametresi “accuracy” olarak seçilmiştir. Yöntemler uygulanırken k katlı çapraz doğrulama kullanılıp k değeri 5 seçilmiştir. SFS yöntemi için seçilen parametreler Tablo 10’da gösterilmiştir.

Tablo 10. SFS Yöntemi İçin Kullanılan Parametreler

Parametreler	Seçilen Değerler
k_features	(1,9)
forward	True
floating	False
verbose	2
scoring	accuracy
cv	5

SFS öznitelik seçim yöntemi NB algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 22’de SFS öznitelik seçim yöntemi sonucu NB modeli performansları gösterilmiştir.



Şekil 22. SFS yöntemi sonucu NB modeli performans gösterimleri.

Veri setinde bulunan 10 öznitelik, SFS özellik seçim yöntemi uygulandıktan sonra 3 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznelikler Tablo 11’de gösterilmiştir.

Tablo 11. NB-SFS Seçilen En İyi Öznelikler

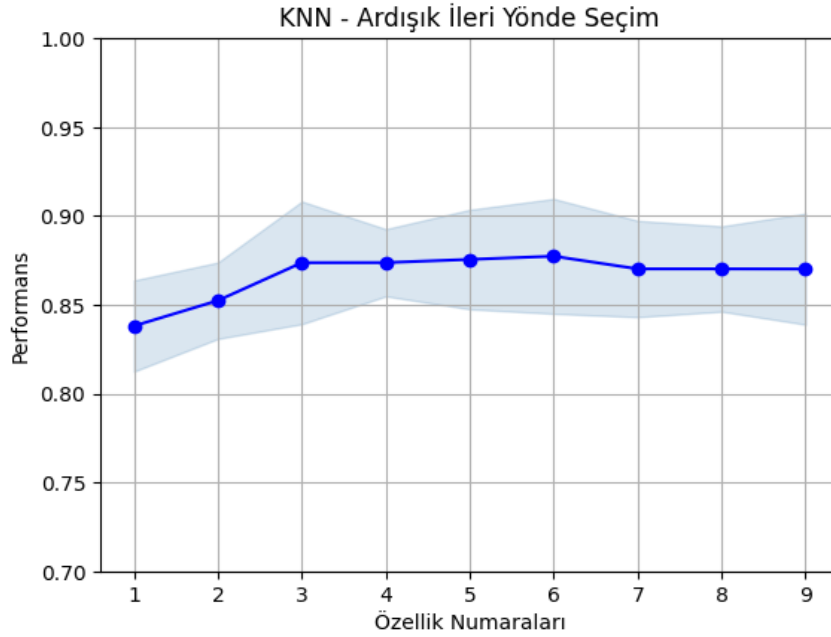
Öznitelikler	Öznitelik İndeksleri
hemoglobin a	1
hemoglobin f	3
mcv	4

Veri setine öznelik seçim yöntemleri kullanmadan öncesi ve SFS öznelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden NB algoritması sonuç değerleri Tablo 12’de gösterilmiştir. SFS öznelik seçim sonrasında NB algoritmasının %86.83 doğruluk oranı ile arttığı görülmektedir.

Tablo 12. SFS Öznelik Seçim Öncesi ve Sonrası NB Doğruluk Oran Karşılaştırması

ML Algoritmaları	NB Doğruluk Oranları (%)	
	SFS Öznelik Seçim Öncesi	SFS Öznelik Seçim Sonrası
NB	84.64	86.83

SFS öznitelik seçim yöntemi KNN algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 23’de SFS öznitelik seçim yöntemi sonucu KNN modeli performansları gösterilmiştir.



Şekil 23. SFS yöntemi sonucu KNN modeli performans gösterimleri.

Veri setinde bulunan 10 öznitelik, SFS özellik seçim yöntemi uygulandıktan sonra 6 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznelikler Tablo 13’te gösterilmiştir.

Tablo 13. KNN-SFS Seçilen En İyi Öznelikler

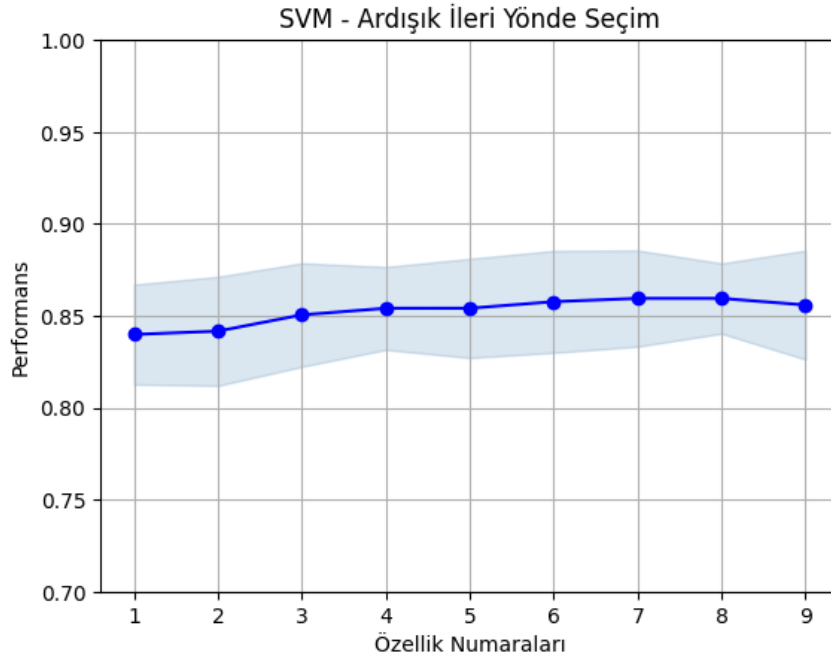
Öznitelikler	Öznitelik İndeksleri
hemoglobin	0
hemoglobin a	1
hemoglobin f	3
ferritin	5
cinsiyet	7
yaş	8

Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SFS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden KNN algoritması sonuç değerleri Tablo 14’te gösterilmiştir. SFS öznitelik seçim sonrasında KNN algoritmasının %87.72 doğruluk oranı ile arttığı görülmektedir.

Tablo 14. SFS Öznelik Seçim Öncesi ve Sonrası KNN Doğruluk Oran Karşılaştırması

ML Algoritmaları	KNN Doğruluk Oranları (%)	
	SFS Öznelik Seçim Öncesi	SFS Öznelik Seçim Sonrası
KNN	83.15	87.72

SFS öznitelik seçim yöntemi SVM algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 24’te SFS öznitelik seçim yöntemi sonucu SVM modeli performansları gösterilmiştir.



Şekil 24. SFS yöntemi sonucu SVM modeli performans gösterimleri.

Veri setinde bulunan 10 öznitelik, SFS özellik seçim yöntemi uygulandıktan sonra 7 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznitelikler Tablo 15’te gösterilmiştir.

Tablo 15. SVM-SFS Seçilen En İyi Öznitelikler

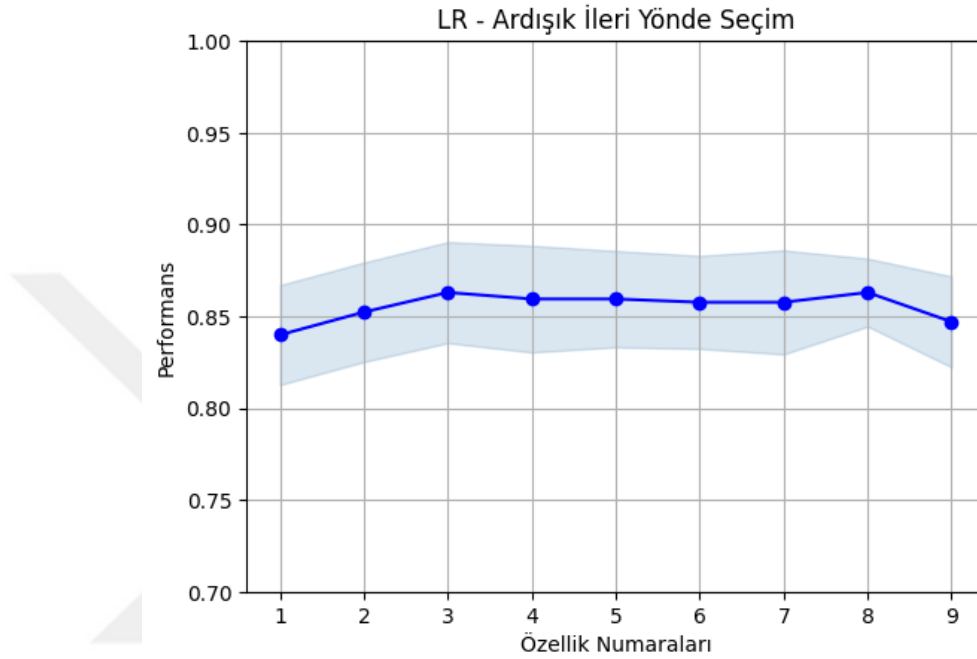
Öznitelikler	Öznitelik İndeksleri
hemoglobin	0
hemoglobin a	1
hemoglobin a2	2
hemoglobin f	3
mcv	4
demir	6
yaş	8

Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SFS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden SVM algoritması sonuç değerleri Tablo 16’da gösterilmiştir. SFS öznitelik seçim sonrasında SVM algoritmasının %85.94 doğruluk oranı ile arttığı görülmektedir.

Tablo 16. SFS Öznitelik Seçim Öncesi ve Sonrası SVM Doğruluk Oran Karşılaştırması

ML Algoritmaları	SVM Doğruluk Oranları (%)	
	SFS Öznitelik Seçim Öncesi	SFS Öznitelik Seçim Sonrası
SVM	82.89	85.94

SFS öznitelik seçim yöntemi LR algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 25’te SFS öznitelik seçim yöntemi sonucu LR modeli performansları gösterilmiştir.

**Şekil 25.** SFS yöntemi sonucu LR modeli performans gösterimleri.

Veri setinde bulunan 10 öznitelik, SFS özellik seçim yöntemi uygulandıktan sonra 3 özniteliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznitelikler Tablo 17’de gösterilmiştir.

Tablo 17. LR-SFS Seçilen En İyi Öznitelikler

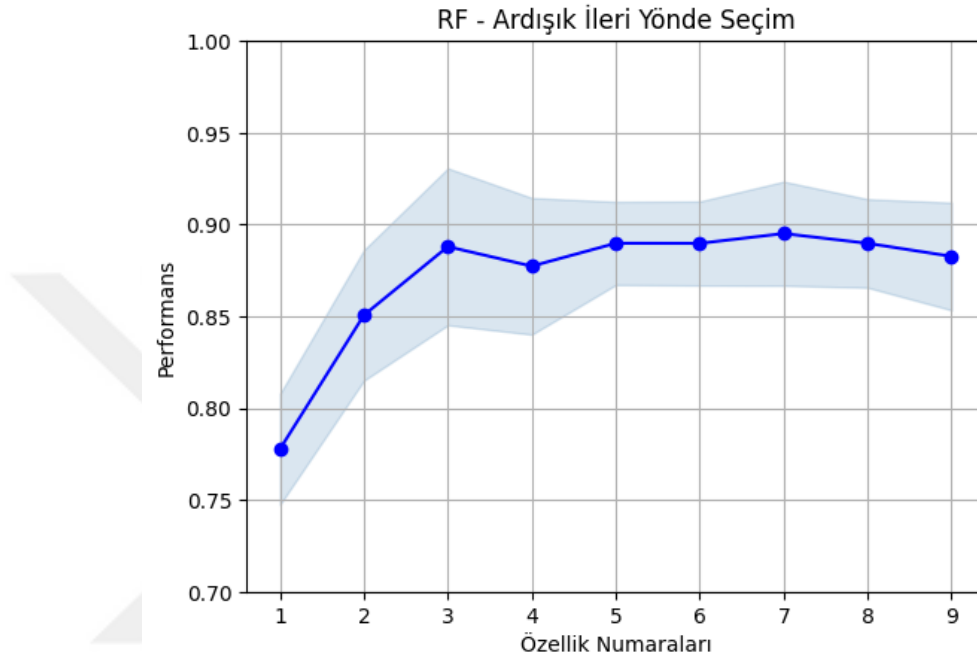
Öznitelikler	Öznitelik İndeksleri
hemoglobin a	1
hemoglobin a2	2
hemoglobin f	3

Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SFS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden LR algoritması sonuç değerleri Tablo 18’de gösterilmiştir. SFS öznitelik seçim sonrasında LR algoritmasının %86.29 doğruluk oranı ile arttığı görülmektedir.

Tablo 18. SFS Öznitelik Seçim Öncesi ve Sonrası LR Doğruluk Oran Karşılaştırması

ML Algoritmaları	LR Doğruluk Oranları (%)	
	SFS Öznitelik Seçim Öncesi	SFS Öznitelik Seçim Sonrası
LR	83.77	86.29

SFS öznitelik seçim yöntemi RF algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 26’da SFS öznitelik seçim yöntemi sonucu RF modeli performansları gösterilmiştir.

**Şekil 26.** SFS yöntemi sonucu RF modeli performans gösterimleri.

Veri setinde bulunan 10 öznitelik, SFS özellik seçim yöntemi uygulandıktan sonra 7 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznelikler Tablo 19’da gösterilmiştir.

Tablo 19. RF-SFS Seçilen En İyi Öznitelikler

Öznitelikler	Öznitelik İndeksleri
hemoglobin	0
hemoglobin a	1
hemoglobin a2	2
hemoglobin f	3
mcv	4
ferritin	5
cinsiyet	7

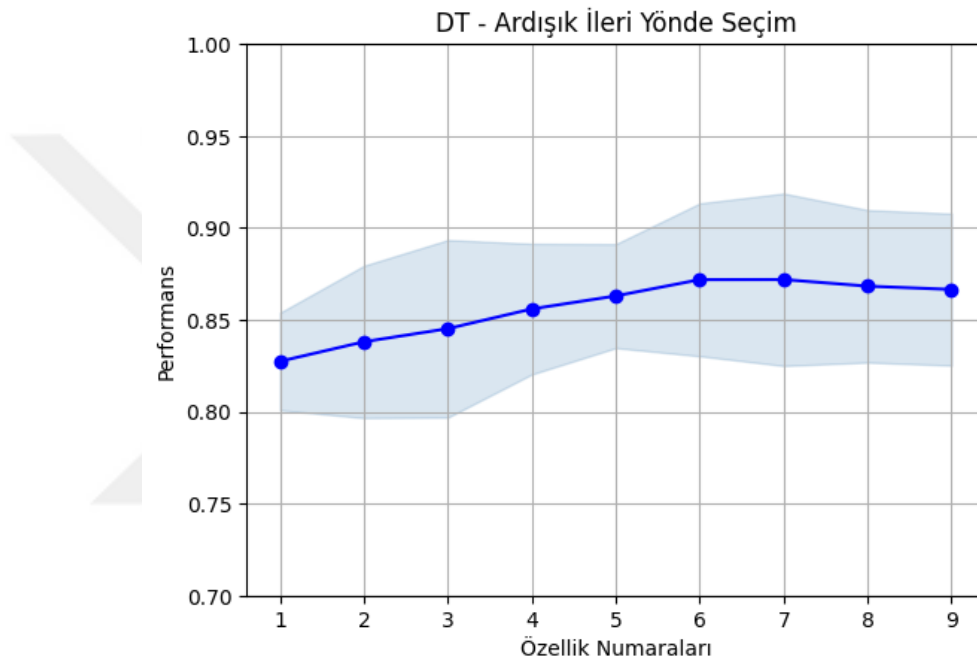
Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SFS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden RF algoritması sonuç değerleri Tablo

20’de gösterilmiştir. SFS öznitelik seçim sonrasında RF algoritmasının %89.50 doğruluk oranı ile arttığı görülmektedir.

Tablo 20. SFS Öznitelik Seçim Öncesi ve Sonrası RF Doğruluk Oran Karşılaştırması

ML Algoritmaları	RF Doğruluk Oranları (%)	
	SFS Öznitelik Seçim Öncesi	SFS Öznitelik Seçim Sonrası
RF	87.00	89.50

SFS öznitelik seçim yöntemi DT algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 27’de SFS öznitelik seçim yöntemi sonucu DT modeli performansları gösterilmiştir.



Şekil 27. SFS yöntemi sonucu DT modeli performans gösterimleri.

Veri setinde bulunan 10 öznitelik, SFS özellik seçim yöntemi uygulandıktan sonra 7 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznelikler Tablo 21’de gösterilmiştir.

Tablo 21. DT-SFS Seçilen En İyi Öznelikler

Öznitelikler	Öznitelik İndeksleri
hemoglobin	0
hemoglobin a	1
hemoglobin a2	2
hemoglobin f	3
mcv	4
ferritin	5
cinsiyet	7

Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SFS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden DT algoritması sonuç değerleri Tablo 22’de gösterilmiştir. SFS öznitelik seçim sonrasında DT algoritmasının %87.18 doğruluk oranı ile arttığı görülmektedir.

Tablo 22. SFS Öznitelik Seçim Öncesi ve Sonrası DT Doğruluk Oran Karşılaştırması

ML Algoritması	DT Doğruluk Oranları (%)	
	SFS Öznitelik Seçim Öncesi	SFS Öznitelik Seçim Sonrası
DT	86.01	87.18

SFS yöntemi sonucunda çalışmada kullanılan tüm ML algoritmalarında doğruluk oranlarının arttığı gözükmemektedir. Tablo 23’te SFS öznitelik seçim yöntemi kullanılmadan önce ML algoritmaları arasında DT %86.01 doğruluk oranı ile en yüksek performansı, SFS öznitelik seçim yöntemi kullanılarak ML algoritmaları arasında RF %89.50 doğruluk oranı ile en yüksek performans elde edildiği görülmektedir.

Tablo 23. SFS Öznitelik Seçim Öncesi ve Sonrası ML Algoritmalarının Doğruluk Oran Karşılaştırması

ML Algoritmaları	ML Doğruluk Oranları (%)	
	SFS Öznitelik Seçim Öncesi	SFS Öznitelik Seçim Sonrası
NB	84.64	86.83
KNN	83.15	87.72
SVM	82.89	85.94
LR	83.77	86.29
RF	87.00	89.50
DT	86.01	87.18

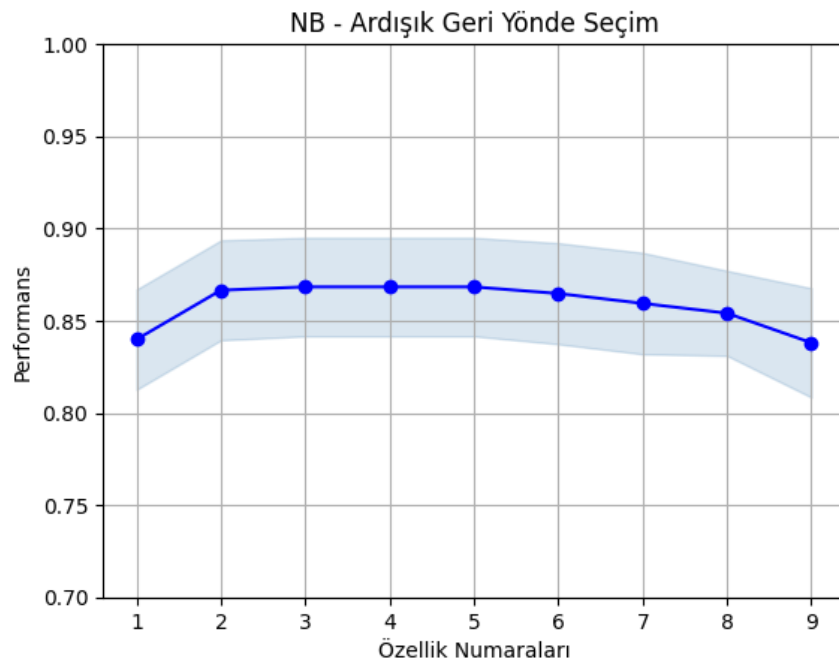
SBS öznitelik seçim yöntemi ile ML algoritmalarının modellenmesi

SBS yöntemi, her model için uygulanmıştır. Bu yöntemde bazı parametreler kullanılmıştır. İlk olarak kaç öznitelik kullanılacağına karar veren “k_features” parametresi (1,9) aralığı seçilmiştir. Veri setinde 10 öznitelik bulunmaktadır. Bu parametre ile özniteliklerin hepsine bakar ve en yüksek doğruluk değerine hangi öznitelikler ile ulaşırsa onları seçer. SBS yöntemi ardışık geriye doğru olacağı için “forward” parametresi “False” ve “floating” parametresi de “False” olarak kullanılmıştır. İşlemden her tekrarda model özeti ayrıntılı görmek için “verbose” parametresi kullanılmıştır. Model sonucunu doğruluk değeri olarak görmek için “scoring” parametresi “accuracy” olarak seçilmiştir. Yöntemler uygulanırken katlı çapraz doğrulama kullanılıp k değeri 5 seçilmiştir. SBS yöntemi için seçilen parametreler Tablo 24’te gösterilmiştir.

Tablo 24. SBS Yöntemi İçin Kullanılan Parametreler

Parametreler	Seçilen Değerler
k_features	(1,9)
forward	False
floating	False
verbose	2
scoring	accuracy
cv	5

SBS öznetelik seçim yöntemi NB algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 28’de SBS öznetelik seçim yöntemi sonucu NB modeli performansları gösterilmiştir.

**Şekil 28.** SBS yöntemi sonucu NB modeli performans gösterimleri.

Veri setinde bulunan 10 öznetelik, SBS özellik seçim yöntemi uygulandıktan sonra 5 özneteliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznetelikler Tablo 25’te gösterilmiştir.

Tablo 25. NB-SBS Seçilen En İyi Öznetelikler

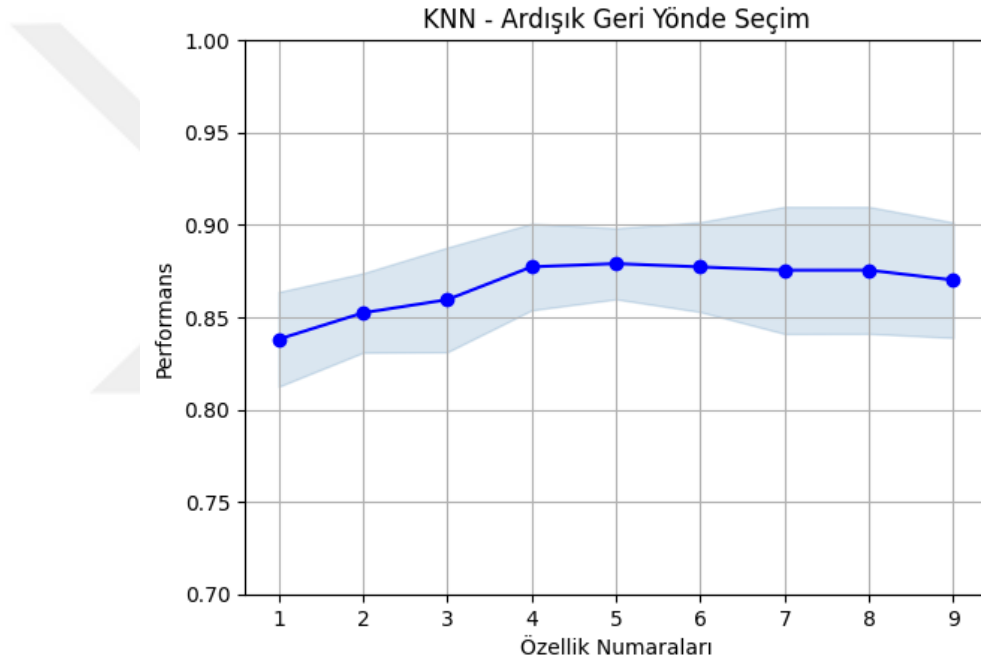
Öznetelikler	Öznetelik İndeksleri
hemoglobin	0
hemoglobin a	1
hemoglobin a2	2
hemoglobin f	3
mcv	4

Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SBS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden NB algoritması sonuç değerleri Tablo 26’da gösterilmiştir. SBS öznitelik seçim sonrasında NB algoritmasının %86.83 doğruluk oranı ile arttığı görülmektedir.

Tablo 26. SBS Öznitelik Seçim Öncesi ve Sonrası NB Doğruluk Oran Karşılaştırması

ML Algoritması	NB Doğruluk Oranları (%)	
	SBS Öznitelik Seçim Öncesi	SBS Öznitelik Seçim Sonrası
NB	84.64	86.83

SBS öznitelik seçim yöntemi KNN algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 29’da SBS öznitelik seçim yöntemi sonucu KNN modeli performansları gösterilmiştir.



Şekil 29. SBS yöntemi sonucu KNN modeli performans gösterimleri.

Veri setinde bulunan 10 öznitelik, SBS özellik seçim yöntemi uygulandıktan sonra 5 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznelikler Tablo 27’de gösterilmiştir.

Tablo 27. KNN-SBS Seçilen En İyi Öznitelikler

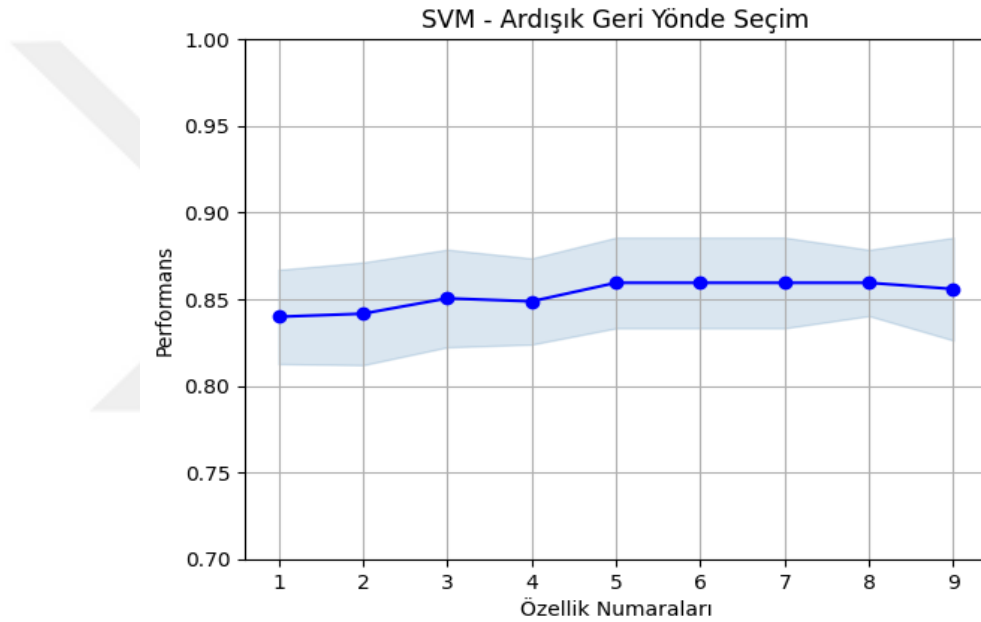
Öznitelikler	Öznitelik İndeksleri
hemoglobin	0
hemoglobin a	1
hemoglobin a2	2
hemoglobin f	3
mcv	4

Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SBS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden KNN algoritması sonuç değerleri Tablo 28’de gösterilmiştir. SBS öznitelik seçim sonrasında KNN algoritmasının %87.89 doğruluk oranı ile arttığı görülmektedir.

Tablo 28. SBS Öznitelik Seçim Öncesi ve Sonrası KNN Doğruluk Oran Karşılaştırması

ML Algoritması	KNN Doğruluk Oranları (%)	
	SBS Öznitelik Seçim Öncesi	SBS Öznitelik Seçim Sonrası
KNN	83.15	87.89

SBS öznitelik seçim yöntemi SVM algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 30’da SBS öznitelik seçim yöntemi sonucu SVM modeli performansları gösterilmiştir.



Şekil 30. SBS yöntemi sonucu SVM modeli performans gösterimleri.

Veri setinde bulunan 10 öznitelik, SBS özellik seçim yöntemi uygulandıktan sonra 7 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznelikler Tablo 29’da gösterilmiştir.

Tablo 29. SVM -SBS Seçilen En İyi Öznitelikler

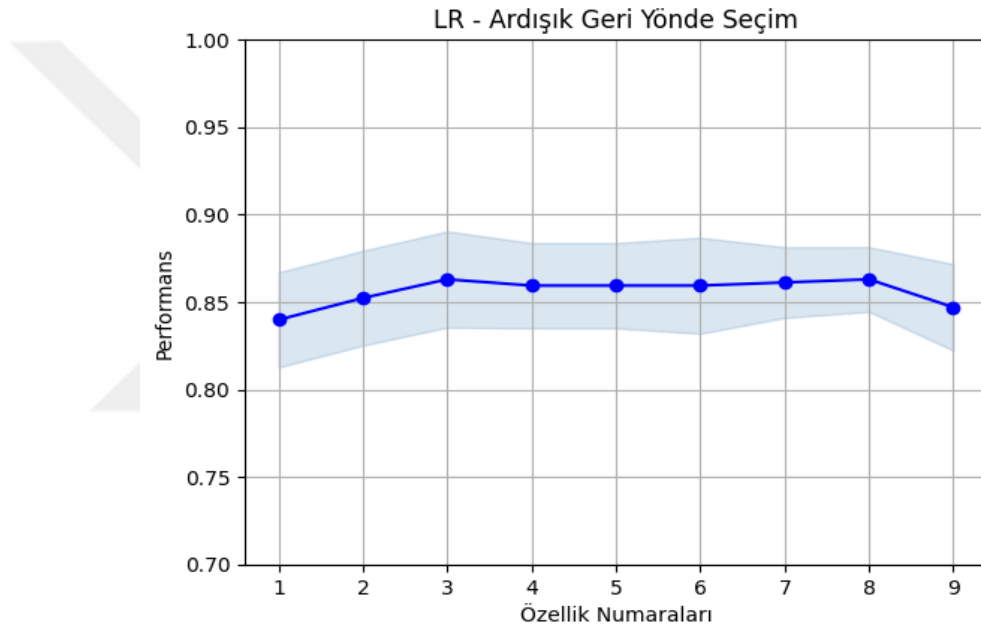
Öznitelikler	Öznitelik İndeksleri
hemoglobin	0
hemoglobin a	1
hemoglobin a2	2
hemoglobin f	3
mcv	4
demir	6
cinsiyet	7
yaş	8

Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SBS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden SVM algoritması sonuç değerleri Tablo 30’da gösterilmiştir. SBS öznitelik seçim sonrasında SVM algoritmasının %85.94 doğruluk oranı ile arttığı görülmektedir.

Tablo 30. SBS Öznitelik Seçim Öncesi ve Sonrası SVM Doğruluk Oran Karşılaştırması

ML Algoritması	SVM Doğruluk Oranları (%)	
	SBS Öznitelik Seçim Öncesi	SBS Öznitelik Seçim Sonrası
SVM	82.89	85.94

SBS öznitelik seçim yöntemi LR algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 31’de SBS öznitelik seçim yöntemi sonucu LR modeli performansları gösterilmiştir.



Şekil 31. SBS yöntemi sonucu LR modeli performans gösterimleri.

Veri setinde bulunan 10 öznitelik, SBS özellik seçim yöntemi uygulandıktan sonra 8 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznelikler Tablo 31’de gösterilmiştir.

Tablo 31. LR-SBS Seçilen En İyi Öznitelikler

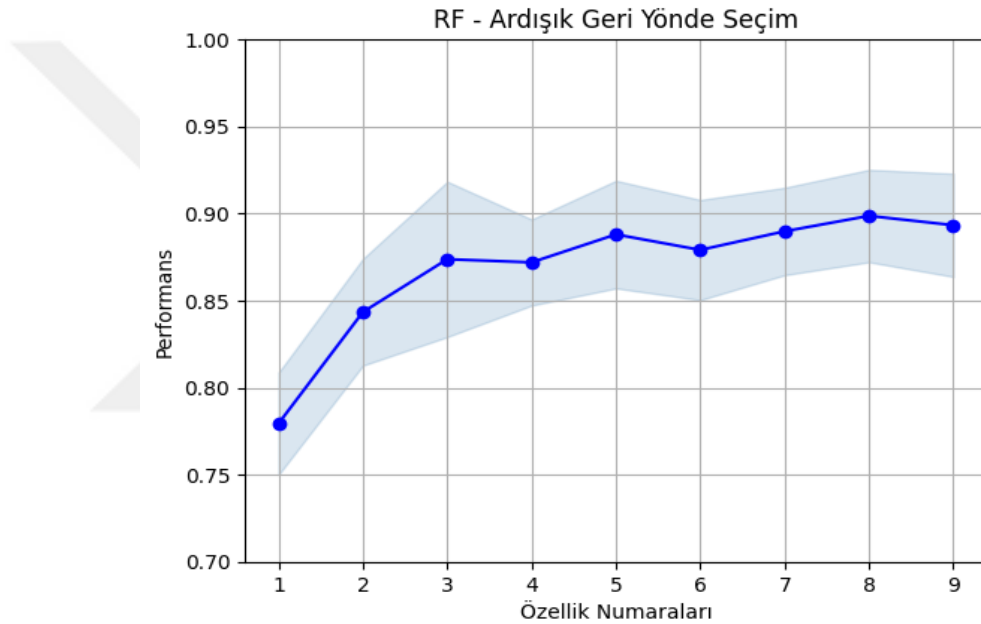
Öznitelikler	Öznitelik İndeksleri
hemoglobin	0
hemoglobin a	1
hemoglobin a2	2
hemoglobin f	3
ferritin	5
demir	6
cinsiyet	7
yaş	8

Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SBS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden LR algoritması sonuç değerleri Tablo 32’de gösterilmiştir. SBS öznitelik seçim sonrasında LR algoritmasının %86.29 doğruluk oranı ile arttığı görülmektedir.

Tablo 32. SBS Öznitelik Seçim Öncesi ve Sonrası LR Doğruluk Oran Karşılaştırması

ML Algoritması	LR Doğruluk Oranları (%)	
	SBS Öznitelik Seçim Öncesi	SBS Öznitelik Seçim Sonrası
LR	83.77	86.29

SBS öznitelik seçim yöntemi RF algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 32’de SBS öznitelik seçim yöntemi sonucu RF modeli performansları gösterilmiştir.



Şekil 32. SBS yöntemi sonucu RF modeli performans gösterimleri.

Veri setinde bulunan 10 öznitelik, SBS özellik seçim yöntemi uygulandıktan sonra 8 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznelikler Tablo 33’te gösterilmiştir.

Tablo 33. RF-SBS Seçilen En İyi Öznitelikler

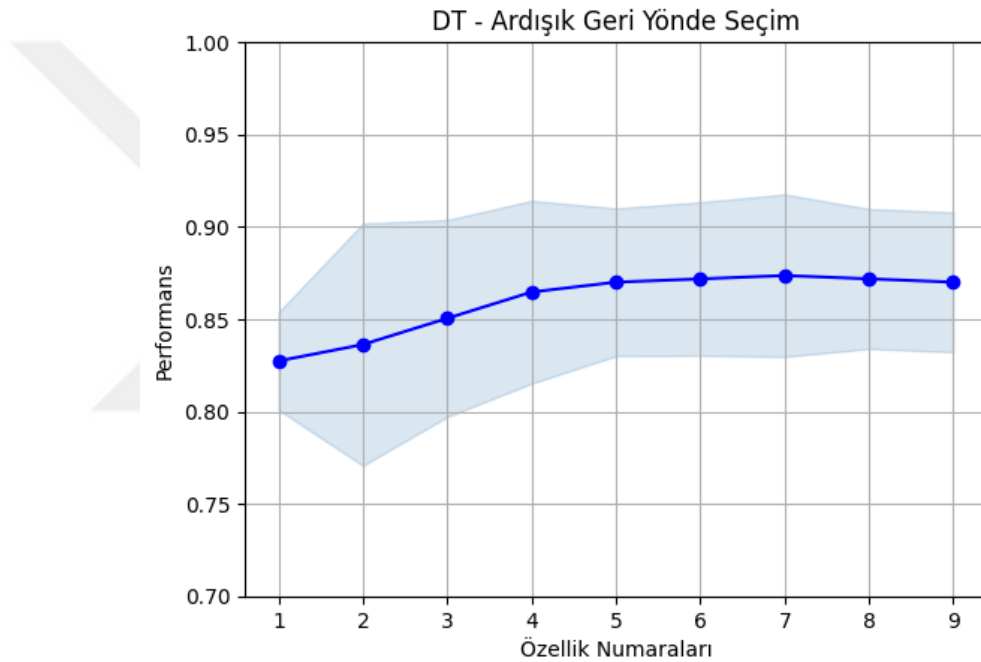
Öznitelikler	Öznitelik İndeksleri
hemoglobin	0
hemoglobin a	1
hemoglobin a2	2
hemoglobin f	3
mcv	4
demir	6
cinsiyet	7
yaş	8

Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SBS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden RF algoritması sonuç değerleri Tablo 34'te gösterilmiştir. SFS öznitelik seçim sonrasında RF algoritmasının %89.86 doğruluk oranı ile arttığı görülmektedir.

Tablo 34. SBS Öznitelik Seçim Öncesi ve Sonrası RF Doğruluk Oran Karşılaştırması

ML Algoritması	RF Doğruluk Oranları (%)	
	SBS Öznitelik Seçim Öncesi	SBS Öznitelik Seçim Sonrası
RF	87.00	89.86

SBS öznitelik seçim yöntemi DT algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 33'te SBS öznitelik seçim yöntemi sonucu DT modeli performansları gösterilmiştir.



Şekil 33. SBS yöntemi sonucu DT modeli performans gösterimleri.

Veri setinde bulunan 10 öznitelik, SBS özellik seçim yöntemi uygulandıktan sonra 7 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznelikler Tablo 35'te gösterilmiştir.

Tablo 35. DT-SBS Seçilen En İyi Öznitelikler

Öznitelikler	Öznitelik İndeksleri
hemoglobin	0
hemoglobin a	1
hemoglobin a2	2
hemoglobin f	3
mcv	4
ferritin	5
cinsiyet	7

Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SBS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden DT algoritması sonuç değerleri Tablo 36’da gösterilmiştir. SBS öznitelik seçim sonrasında DT algoritmasının %87.36 doğruluk oranı ile arttığı görülmektedir.

Tablo 36. SBS Öznitelik Seçim Öncesi ve Sonrası DT Doğruluk Oran Karşılaştırması

ML Algoritması	DT Doğruluk Oranları (%)	
	SBS Öznitelik Seçim Öncesi	SBS Öznitelik Seçim Sonrası
DT	86.01	87.36

SBS yöntemi sonucunda çalışmada kullanılan tüm ML algoritmalarında doğruluk oranlarının arttığı gözükmemektedir. Tablo 37’de SBS öznitelik seçim yöntemi kullanılmadan önce ML algoritmaları arasında RF %87.00 doğruluk oranı ile en yüksek performansı, SBS öznitelik seçim yöntemi kullanılarak ML algoritmaları arasında yine RF %89.86 doğruluk oranı ile en yüksek performans elde edildiği görülmektedir.

Tablo 37. SBS Öznitelik Seçim Öncesi ve Sonrası ML Algoritmalarının Doğruluk Oran Karşılaştırması

ML Algoritmaları	ML Doğruluk Oranları (%)	
	SBS Öznitelik Seçim Öncesi	SBS Öznitelik Seçim Sonrası
NB	84.64	86.83
KNN	83.15	87.89
SVM	82.89	85.94
LR	83.77	86.29
RF	87.00	89.86
DT	86.01	87.36

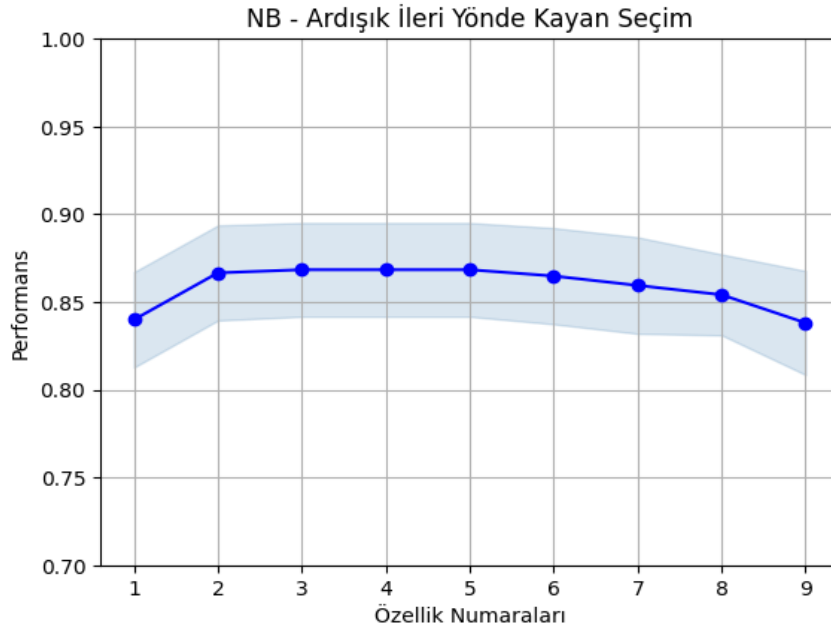
SFFS öznitelik seçim yöntemi ile ML algoritmalarının modellenmesi

SFFS yöntemi, her model için uygulanmıştır. Bu yöntemde bazı parametreler kullanılmıştır. İlk olarak kaç öznitelik kullanılacağına karar veren “k_features” parametresi (1,9) aralığı seçilmiştir. Veri setinde 10 öznitelik bulunmaktadır. Bu parametre ile özniteliklerin hepsine bakar ve en yüksek doğruluk değerine hangi öznitelikler ile ulaşırsa onları seçer. SFFS yöntemi ardışık ileri olacağı için “forward” parametresi “True” ve ileri kayan seçim olacağı için ise “floating” parametresi de “True” olarak kullanılmıştır. Kayan seçim olması demek daha önce baktığı özniteliklere tekrar bakabilmesidir. İşlemde her tekrarda model özetini ayrıntılı görmek için “verbose” parametresi kullanılmıştır. Model sonucunu doğruluk değeri olarak görmek için “scoring” parametresi “accuracy” olarak seçilmiştir. Yöntemler uygulanırken k katlı çapraz doğrulama kullanılıp k değeri 5 seçilmiştir. SFFS yöntemi için seçilen parametreler Tablo 38’de gösterilmiştir.

Tablo 38. SFFS Yöntemi İçin Kullanılan Parametreler

Parametreler	Seçilen Değerler
k_features	(1,9)
forward	True
floating	True
verbose	2
scoring	accuracy
cv	5

SFFS öznelik seçim yöntemi NB algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 34’te SFFS öznelik seçim yöntemi sonucu NB modeli performansları gösterilmiştir.

**Şekil 34.** SFFS yöntemi sonucu NB modeli performans gösterimleri.

Veri setinde bulunan 10 öznelik, SFFS özellik seçim yöntemi uygulandıktan sonra 3 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznelikler Tablo 39’da gösterilmiştir.

Tablo 39. NB-SFFS Seçilen En İyi Öznelikler

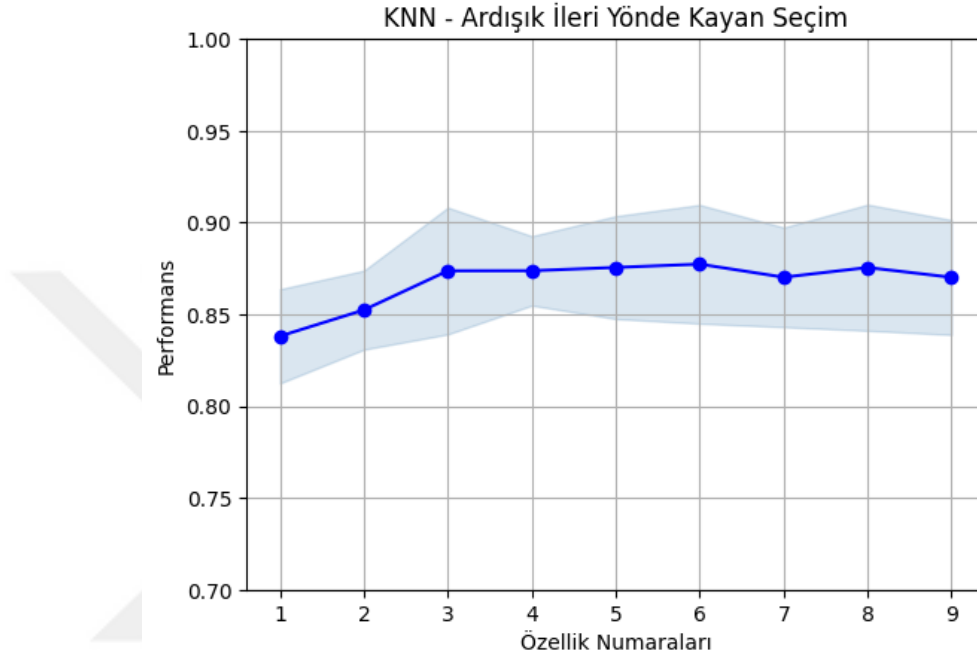
Öznelikler	Öznelik İndeksleri
hemoglobin a	1
hemoglobin f	3
mcv	4

Veri setine öznelik seçim yöntemleri kullanmadan öncesi ve SFFS öznelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden NB algoritması sonuç değerleri Tablo 39’da gösterilmiştir. SFFS öznelik seçim sonrasında NB algoritmasının %86.83 doğruluk oranı ile arttığı görülmektedir.

Tablo 40. SFFS Öznitelik Seçim Öncesi ve Sonrası NB Doğruluk Oran Karşılaştırması

ML Algoritması	NB Doğruluk Oranları (%)	
	SFFS Öznitelik Seçim Öncesi	SFFS Öznitelik Seçim Sonrası
NB	84.64	86.83

SFFS öznitelik seçim yöntemi KNN algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 35’te SFFS öznitelik seçim yöntemi sonucu KNN modeli performansları gösterilmiştir.

**Şekil 35.** SFFS yöntemi sonucu KNN modeli performans gösterimleri.

Veri setinde bulunan 10 öznitelik, SFFS özellik seçim yöntemi uygulandıktan sonra 6 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznitelikler Tablo 41’de gösterilmiştir.

Tablo 41. KNN-SFFS Seçilen En İyi Öznitelikler

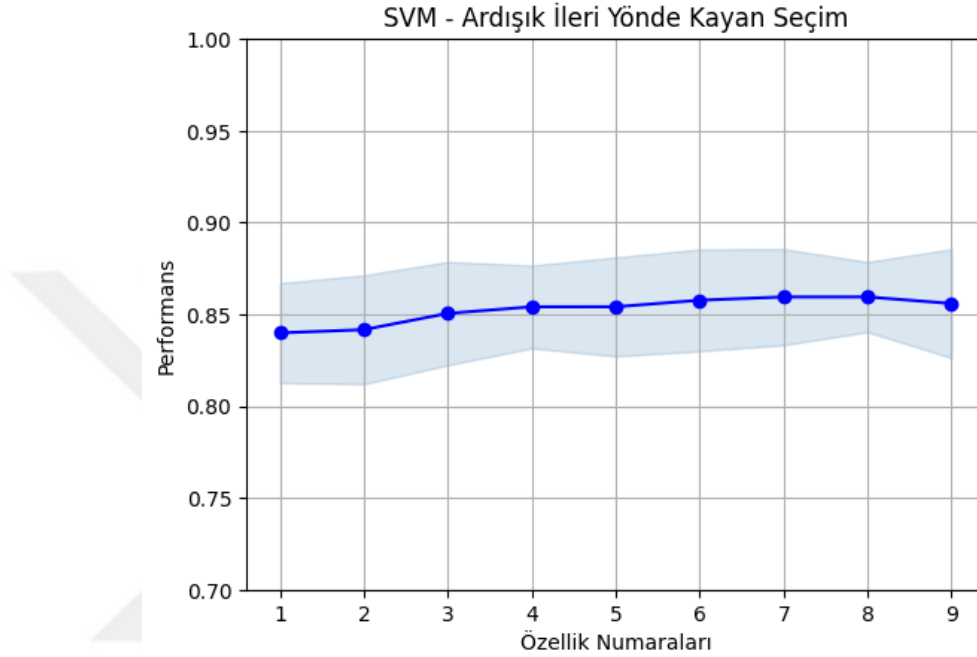
Öznitelikler	Öznitelik İndeksleri
hemoglobin	0
hemoglobin a	1
hemoglobin f	3
ferritin	5
cinsiyet	7
yaş	8

Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SFFS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden KNN algoritması sonuç değerleri Tablo 42’de gösterilmiştir. SFFS öznitelik seçim sonrasında KNN algoritmasının %87.72 doğruluk oranı ile arttığı görülmektedir.

Tablo 42. SFFS Öznitelik Seçim Öncesi Ve Sonrası KNN Doğruluk Oran Karşılaştırması

ML Algoritması	KNN Doğruluk Oranları (%)	
	SFFS Öznitelik Seçim Öncesi	SFFS Öznitelik Seçim Sonrası
KNN	83.15	87.72

SFFS öznitelik seçim yöntemi SVM algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 36’da SFFS öznitelik seçim yöntemi sonucu SVM modeli performansları gösterilmiştir.

**Şekil 36.** SFFS yöntemi sonucu SVM modeli performans gösterimleri.

Veri setinde bulunan 10 öznitelik, SFFS özellik seçim yöntemi uygulandıktan sonra 7 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznitelikler Tablo 43’te gösterilmiştir.

Tablo 43. SVM-SFFS Seçilen En İyi Öznitelikler

Öznitelikler	Öznitelik İndeksleri
hemoglobin	0
hemoglobin a	1
hemoglobin a2	2
hemoglobin f	3
mcv	4
demir	6
yaş	8

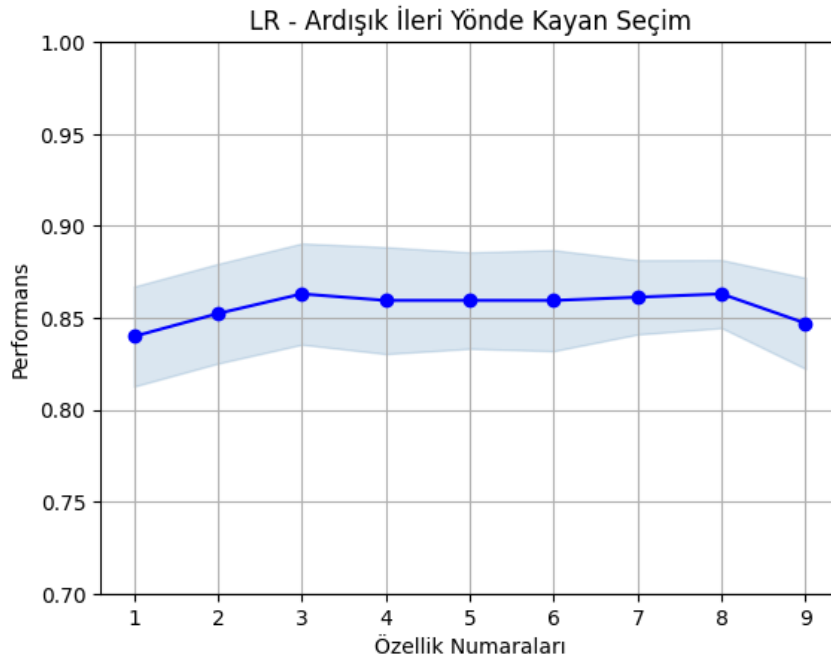
Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SFFS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden SVM algoritması sonuç değerleri

Tablo 44’te gösterilmiştir. SFFS öznelik seçim sonrasında SVM algoritmasının %85.94 doğruluk oranı ile arttığı görülmektedir.

Tablo 44. SFFS Öznelik Seçim Öncesi ve Sonrası SVM Doğruluk Oran Karşılaştırması

ML Algoritması	SVM Doğruluk Oranları (%)	
	SFFS Öznelik Seçim Öncesi	SFFS Öznelik Seçim Sonrası
SVM	82.89	85.94

SFFS öznelik seçim yöntemi LR algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 37’de SFFS öznelik seçim yöntemi sonucu LR modeli performansları gösterilmiştir.



Şekil 37. SFFS yöntemi sonucu LR modeli performans gösterimleri.

Veri setinde bulunan 10 öznelik, SFFS özellik seçim yöntemi uygulandıktan sonra 3 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznelikler Tablo 45’te gösterilmiştir.

Tablo 45. LR-SFFS Seçilen En İyi Öznelikler

Öznelikler	Öznelik İndeksleri
hemoglobin a	1
hemoglobin a2	2
hemoglobin f	3

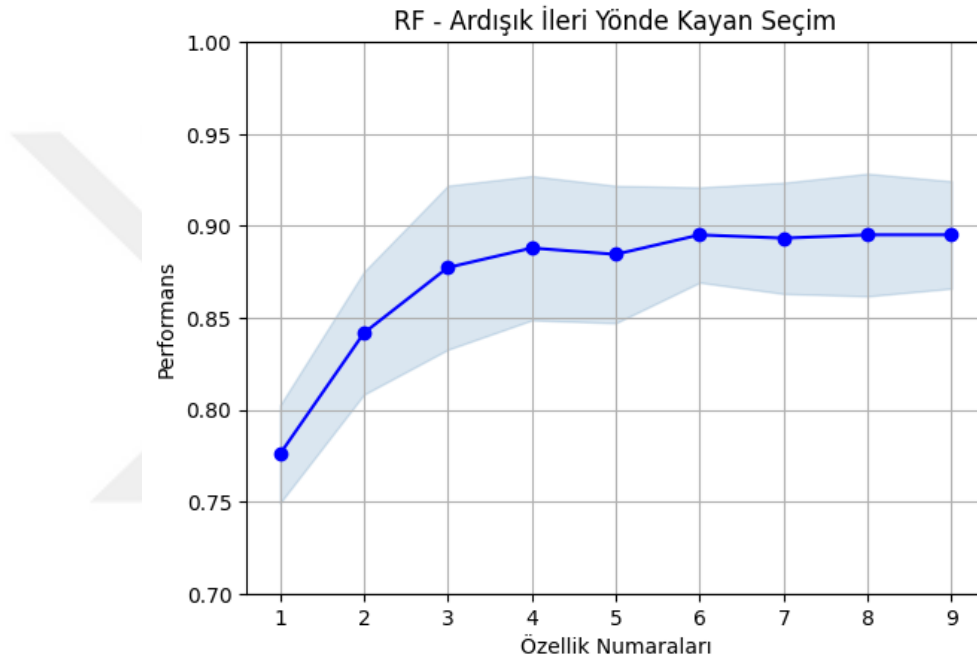
Veri setine öznelik seçim yöntemleri kullanmadan öncesi ve SFFS öznelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden LR algoritması sonuç değerleri Tablo

46’da gösterilmiştir. SFFS öznelik seçim sonrasında LR algoritmasının %86.29 doğruluk oranı ile arttığı görülmektedir.

Tablo 46. SFFS Öznelik Seçim Öncesi ve Sonrası LR Doğruluk Oran Karşılaştırması

ML Algoritması	LR Doğruluk Oranları (%)	
	SFFS Öznelik Seçim Öncesi	SFFS Öznelik Seçim Sonrası
LR	83.77	86.29

SFFS öznelik seçim yöntemi RF algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 38’de SFFS öznelik seçim yöntemi sonucu RF modeli performansları gösterilmiştir.



Şekil 38. SFFS yöntemi sonucu RF modeli performans gösterimleri.

Veri setinde bulunan 10 öznelik, SFFS özellik seçim yöntemi uygulandıktan sonra 7 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznelikler Tablo 47’de gösterilmiştir.

Tablo 47. RF-SFFS Seçilen En İyi Öznelikler

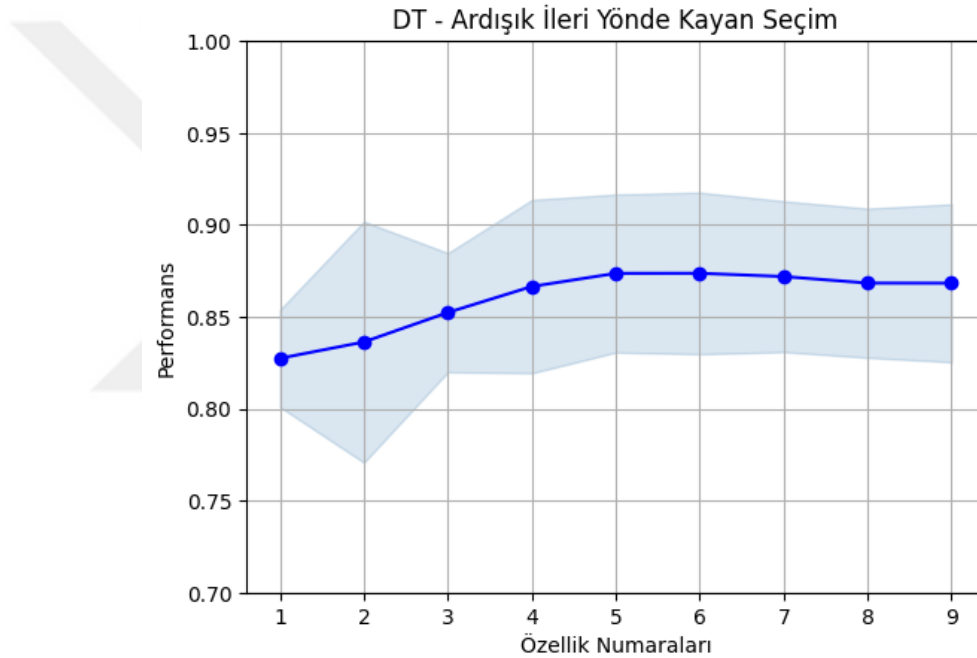
Öznelikler	Öznelik İndeksleri
hemoglobin	0
hemoglobin a	1
hemoglobin a2	2
hemoglobin f	3
mcv	4
cinsiyet	7
yaş	8

Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SFFS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden RF algoritması sonuç değerleri Tablo 48’de gösterilmiştir. SFFS öznitelik seçim sonrasında RF algoritmasının %89.68 doğruluk oranı ile arttığı görülmektedir.

Tablo 48. SFFS Öznitelik Seçim Öncesi ve Sonrası RF Doğruluk Oran Karşılaştırması

ML Algoritması	RF Doğruluk Oranları (%)	
	SFFS Öznitelik Seçim Öncesi	SFFS Öznitelik Seçim Sonrası
RF	87.00	89.68

SFFS öznitelik seçim yöntemi DT algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 39’da SFFS öznitelik seçim yöntemi sonucu DT modeli performansları gösterilmiştir.



Şekil 39. SFFS yöntemi sonucu DT modeli performans gösterimleri.

Veri setinde bulunan 10 öznitelik, SFFS özellik seçim yöntemi uygulandıktan sonra 6 özniteliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznitelikler Tablo 49’da gösterilmiştir.

Tablo 49. DT-SFFS Seçilen En İyi Öznitelikler

Öznitelikler	Öznitelik İndeksleri
hemoglobin	0
hemoglobin a	1
hemoglobin f	3
mcv	4
ferritin	5
cinsiyet	7

Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SFFS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden DT algoritması sonuç değerleri Tablo 50’de gösterilmiştir. SFFS öznitelik seçim sonrasında DT algoritmasının %87.36 doğruluk oranı ile arttığı görülmektedir.

Tablo 50. SFFS Öznitelik Seçim Öncesi ve Sonrası DT Doğruluk Oran Karşılaştırması

ML Algoritması	DT Doğruluk Oranları (%)	
	SFFS Öznitelik Seçim Öncesi	SFFS Öznitelik Seçim Sonrası
DT	86.01	87.36

SFFS yöntemi sonucunda çalışmada kullanılan tüm ML algoritmalarında doğruluk oranlarının arttığı gözükmemektedir. Tablo 51’de SFFS öznitelik seçim yöntemi kullanılmadan önce ML algoritmaları arasında RF %87.00 doğruluk oranı ile en yüksek performansı, SFFS öznitelik seçim yöntemi kullanılarak ML algoritmaları arasında RF %89.68 doğruluk oranı ile en yüksek performans elde edildiği görülmektedir.

Tablo 51. SFFS öznitelik seçim öncesi ve sonrası ML algoritmalarının doğruluk oran karşılaştırması

ML Algoritmaları	ML Doğruluk Oranları (%)	
	SFFS Öznitelik Seçim Öncesi	SFFS Öznitelik Seçim Sonrası
NB	84.64	86.83
KNN	83.15	87.72
SVM	82.89	85.94
LR	83.77	86.29
RF	87.00	89.68
DT	86.01	87.53

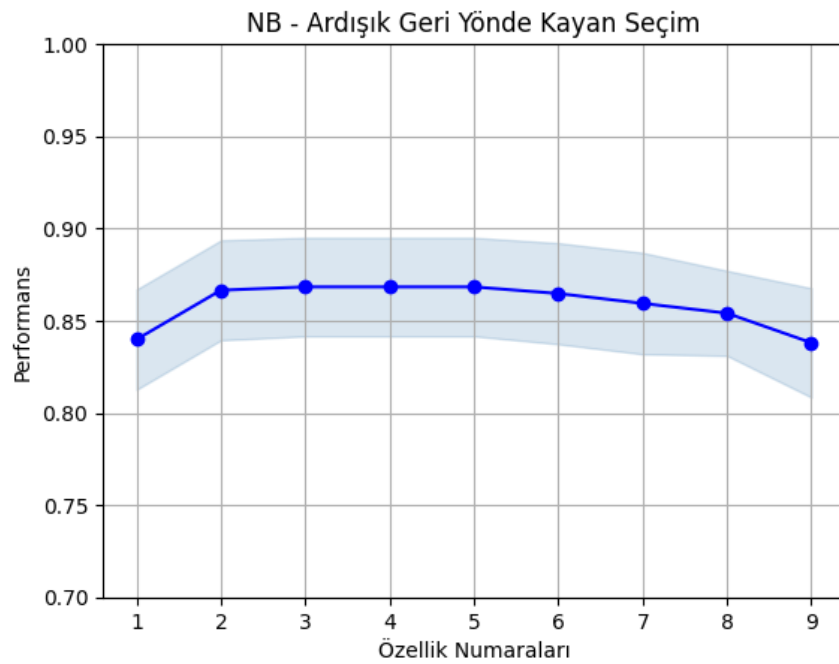
SBFS öznitelik seçim yöntemi ile ML algoritmalarının modellenmesi

SBFS yöntemi, her model için uygulanmıştır. Bu yöntemde bazı parametreler kullanılmıştır. İlk olarak kaç öznitelik kullanılacağına karar veren “k_features” parametresi (1,9) aralığı seçilmiştir. Veri setinde 10 öznitelik bulunmaktadır. Bu parametre ile özniteliklerin hepsine bakar ve en yüksek doğruluk değerine hangi öznitelikler ile ulaşırsa onları seçer. SBFS yöntemi ardışık geriye doğru olacağı için “forward” parametresi “False” ve geri kayan seçim olacağı için “floating” parametresi de “True” olarak kullanılmıştır. İşlemde her tekrarda model özetini ayrıntılı görmek için “verbose” parametresi kullanılmıştır. Model sonucunu doğruluk değeri olarak görmek için “scoring” parametresi “accuracy” olarak seçilmiştir. Yöntemler uygulanırken k katlı çapraz doğrulama kullanılıp k değeri 5 seçilmiştir. SBFS yöntemi için seçilen parametreler Tablo 52’de gösterilmiştir.

Tablo 52. SBFS Yöntemi için Kullanılan Parametreler

Parametreler	Seçilen Değerler
k_features	(1,9)
forward	False
floating	True
verbose	2
scoring	accuracy
cv	5

SBFS öznelik seçim yöntemi NB algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 40'ta SBFS öznelik seçim yöntemi sonucu NB modeli performansları gösterilmiştir.

**Şekil 40.** SBFS yöntemi sonucu NB modeli performans gösterimleri.

Veri setinde bulunan 10 öznelik, SBFS özellik seçim yöntemi uygulandıktan sonra 5 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznelikler Tablo 53'te gösterilmiştir.

Tablo 53. NB-SBFS Seçilen En İyi Öznelikler

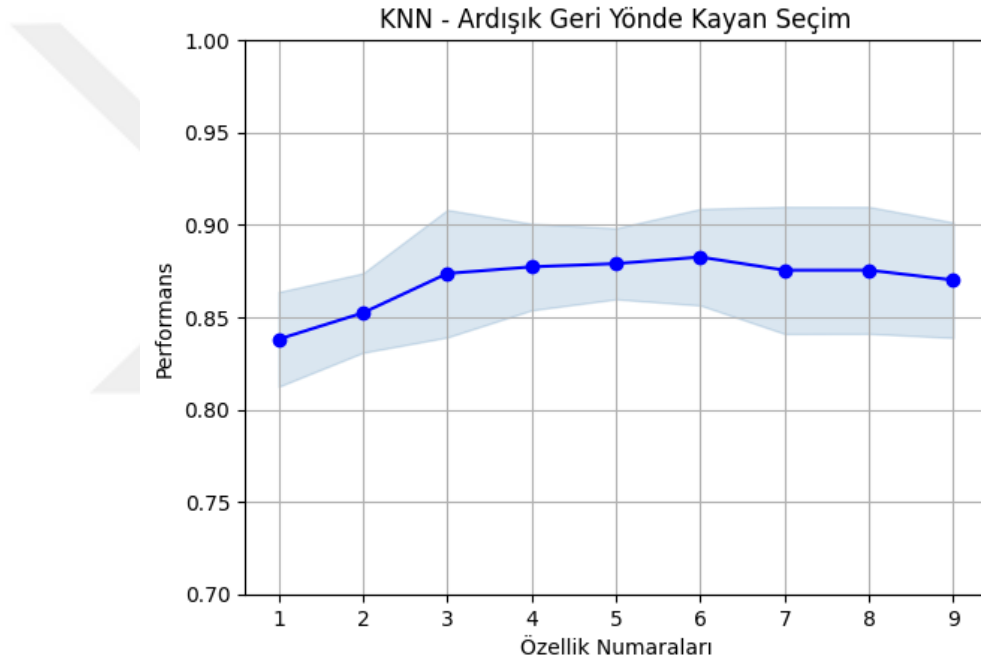
Öznelikler	Öznelik İndeksleri
hemoglobin	0
hemoglobin a	1
hemoglobin a2	2
hemoglobin f	3
mcv	4

Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SBFS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden NB algoritması sonuç değerleri Tablo 54'te gösterilmiştir. SBFS öznitelik seçim sonrasında NB algoritmasının %86.83 doğruluk oranı ile arttığı görülmektedir.

Tablo 54. SBFS Öznitelik Seçim Öncesi ve Sonrası NB Doğruluk Oran Karşılaştırması

ML Algoritması	NB Doğruluk Oranları (%)	
	SBFS Öznitelik Seçim Öncesi	SBFS Öznitelik Seçim Sonrası
NB	84.64	86.83

SBFS öznitelik seçim yöntemi KNN algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 41'de SBFS öznitelik seçim yöntemi sonucu KNN modeli performansları gösterilmiştir.



Şekil 41. SBFS yöntemi sonucu KNN modeli performans gösterimleri.

Veri setinde bulunan 10 öznitelik, SBFS özellik seçim yöntemi uygulandıktan sonra 6 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznelikler Tablo 55'te gösterilmiştir.

Tablo 55. KNN-SBFS Seçilen En İyi Öznitelikler

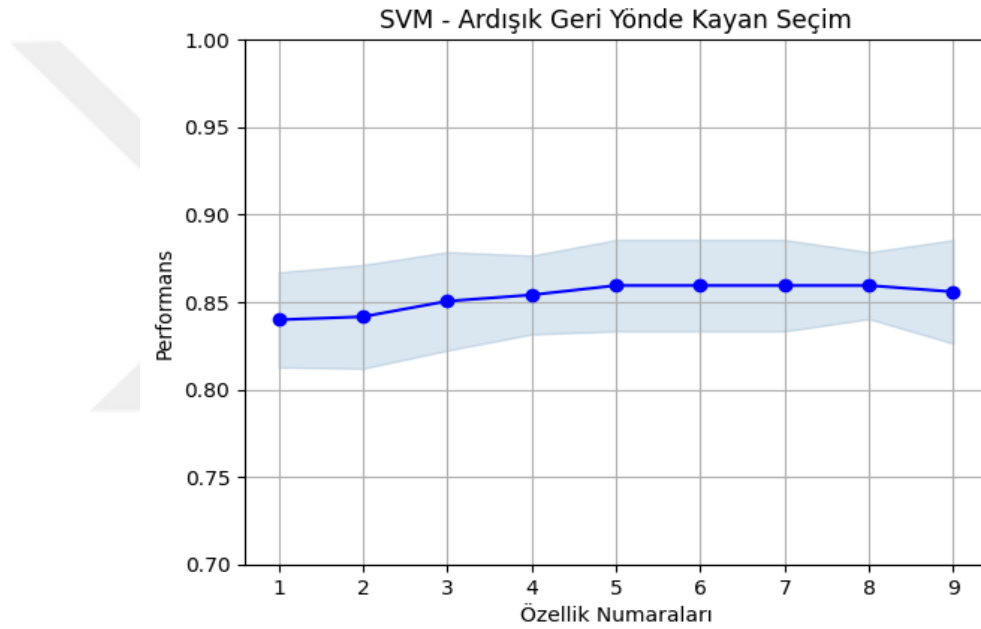
Öznitelikler	Öznitelik İndeksleri
hemoglobin	0
hemoglobin a	1
hemoglobin a2	2
hemoglobin f	3
mcv	4
yaş	8

Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SBFS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden KNN algoritması sonuç değerleri Tablo 56’da gösterilmiştir. SBFS öznitelik seçim sonrasında KNN algoritmasının %88.25 doğruluk oranı ile arttığı görülmektedir.

Tablo 56. SBFS Öznitelik Seçim Öncesi ve Sonrası KNN Doğruluk Oran Karşılaştırması

ML Algoritması	KNN Doğruluk Oranları (%)	
	SBFS Öznitelik Seçim Öncesi	SBFS Öznitelik Seçim Sonrası
KNN	83.15	88.25

SBFS öznitelik seçim yöntemi SVM algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 42’de SBFS öznitelik seçim yöntemi sonucu SVM modeli performansları gösterilmiştir.



Şekil 42. SBFS yöntemi sonucu SVM modeli performans gösterimleri.

Veri setinde bulunan 10 öznitelik, SBFS özellik seçim yöntemi uygulandıktan sonra 7 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznitelikler Tablo 57’de gösterilmiştir.

Tablo 57. SVM-SBFS Seçilen En İyi Öznitelikler

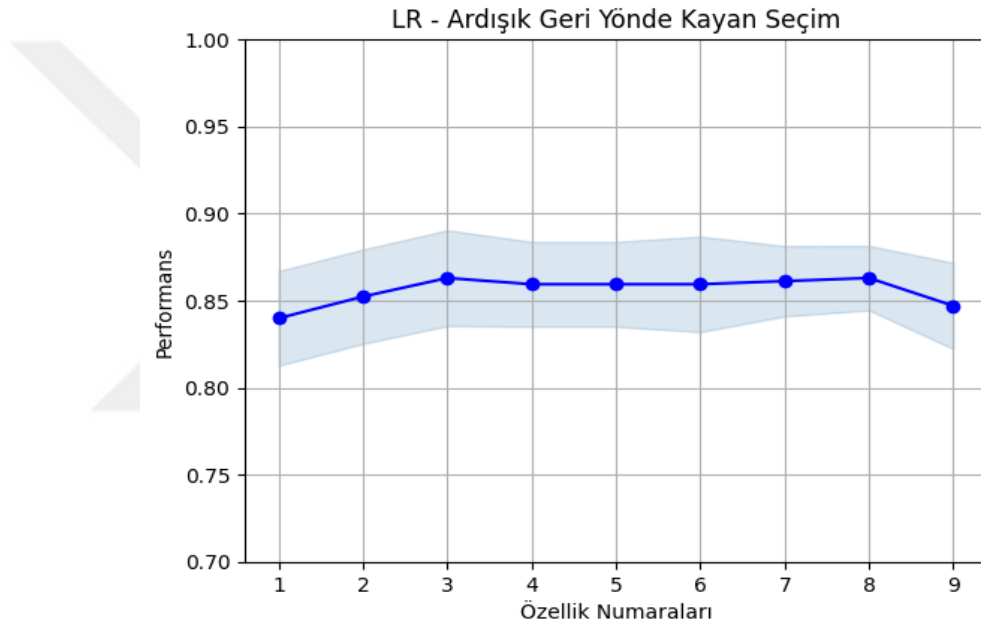
Öznitelikler	Öznitelik İndeksleri
hemoglobin	0
hemoglobin a	1
hemoglobin a2	2
hemoglobin f	3
mcv	4
demir	6
cinsiyet	7
yaş	8

Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SBFS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden SVM algoritması sonuç değerleri Tablo 58’de gösterilmiştir. SBFS öznitelik seçim sonrasında SVM algoritmasının %85.94 doğruluk oranı ile arttığı görülmektedir.

Tablo 58. SBFS Öznitelik Seçim Öncesi ve Sonrası SVM Doğruluk Oran Karşılaştırması

ML Algoritması	SVM Doğruluk Oranları (%)	
	SBFS Öznitelik Seçim Öncesi	SBFS Öznitelik Seçim Sonrası
SVM	82.89	85.94

SBFS öznitelik seçim yöntemi LR algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 43’te SBFS öznitelik seçim yöntemi sonucu LR modeli performansları gösterilmiştir.



Şekil 43. SBFS yöntemi sonucu LR modeli performans gösterimleri.

Veri setinde bulunan 10 öznitelik, SBFS özellik seçim yöntemi uygulandıktan sonra 8 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznelikler Tablo 59’da gösterilmiştir.

Tablo 59. LR-SBFS Seçilen En İyi Öznitelikler

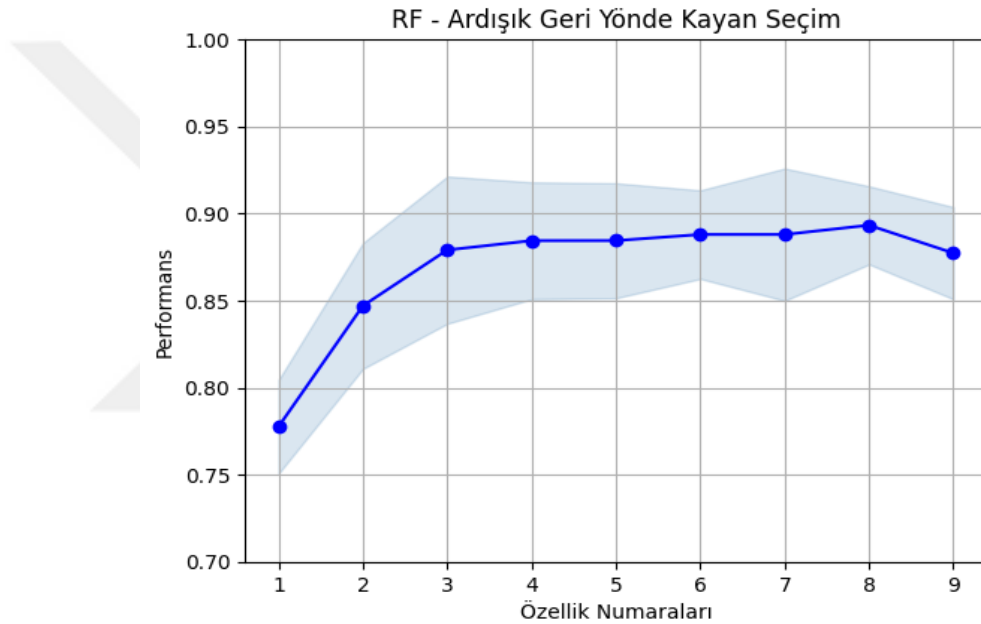
Öznitelikler	Öznitelik İndeksleri
hemoglobin	0
hemoglobin a	1
hemoglobin a2	2
hemoglobin f	3
ferritin	5
demir	6
cinsiyet	7
yaş	8

Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SBFS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden LR algoritması sonuç değerleri Tablo 60’da gösterilmiştir. SBFS öznitelik seçim sonrasında LR algoritmasının %86.29 doğruluk oranı ile arttığı görülmektedir.

Tablo 60. SBFS Öznitelik Seçim Öncesi Ve Sonrası LR Doğruluk Oran Karşılaştırması

ML Algoritması	LR Doğruluk Oranları (%)	
	SBFS Öznitelik Seçim Öncesi	SBFS Öznitelik Seçim Sonrası
LR	83.77	86.29

SBFS öznitelik seçim yöntemi RF algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 44’te SBFS öznitelik seçim yöntemi sonucu RF modeli performansları gösterilmiştir.



Şekil 44. SBFS yöntemi sonucu RF modeli performans gösterimleri.

Veri setinde bulunan 10 öznitelik, SBFS özellik seçim yöntemi uygulandıktan sonra 8 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznelikler Tablo 61’de gösterilmiştir.

Tablo 61. RF-SBFS Seçilen En İyi Öznitelikler

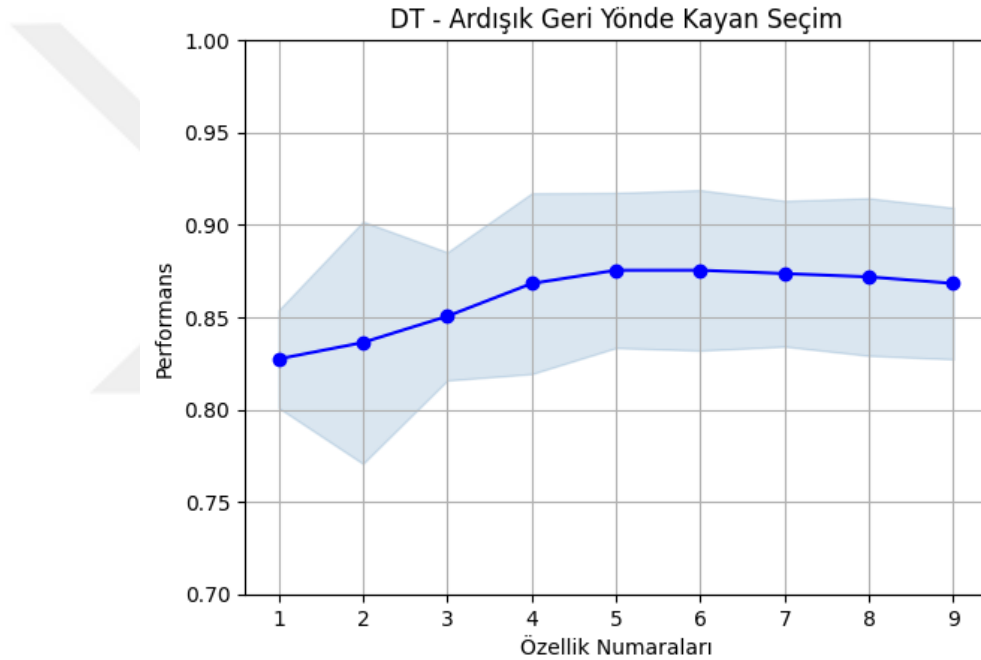
Öznitelikler	Öznitelik İndeksleri
hemoglobin	0
hemoglobin a	1
hemoglobin a2	2
hemoglobin f	3
mcv	4
demir	6
cinsiyet	7
yaş	8

Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SBFS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden RF algoritması sonuç değerleri Tablo 62’de gösterilmiştir. SBFS öznitelik seçim sonrasında RF algoritmasının %89.32 doğruluk oranı ile arttığı görülmektedir.

Tablo 62. SBFS Öznitelik Seçim Öncesi ve Sonrası RF Doğruluk Oran Karşılaştırması

ML Algoritması	RF Doğruluk Oranları (%)	
	SBFS Öznitelik Seçim Öncesi	SBFS Öznitelik Seçim Sonrası
RF	87.00	89.32

SBFS öznitelik seçim yöntemi DT algoritmasına uygulanmıştır. Uygulama sonrasında Şekil 45’te SBFS öznitelik seçim yöntemi sonucu DT modeli performansları gösterilmiştir.



Şekil 45. SBFS yöntemi sonucu DT modeli performans gösterimleri.

Veri setinde bulunan 10 öznitelik, SBFS özellik seçim yöntemi uygulandıktan sonra 5 özneliğe indirgenmiştir. Yapılan modelleme sonucunda en iyi performansı veren öznelikler Tablo 63’te gösterilmiştir.

Tablo 63. DT-SBFS Seçilen En İyi Öznitelikler

Öznitelikler	Öznitelik İndeksleri
hemoglobin	0
hemoglobin a	1
hemoglobin f	3
mcv	4
ferritin	5

Veri setine öznitelik seçim yöntemleri kullanmadan öncesi ve SBFS öznitelik seçme yöntemi kullanım sonrası uygulanan ML yöntemlerinden DT algoritması sonuç değerleri Tablo 64'te gösterilmiştir. SBFS öznitelik seçim sonrasında DT algoritmasının %87.54 doğruluk oranı ile arttığı görülmektedir.

Tablo 64. SBFS Öznitelik Seçim Öncesi ve Sonrası DT Doğruluk Oran Karşılaştırması

ML Algoritması	DT Doğruluk Oranları (%)	
	SBFS Öznitelik Seçim Öncesi	SBFS Öznitelik Seçim Sonrası
DT	86.01	87.54

SBFS yöntemi sonucunda çalışmada kullanılan tüm ML algoritmalarında doğruluk oranlarının arttığı gözükmemektedir. Tablo 65'te SBFS öznitelik seçim yöntemi kullanılmadan önce ML algoritmaları arasında RF %87.13 doğruluk oranı ile en yüksek performansı, SBFS öznitelik seçim yöntemi kullanılarak ML algoritmaları arasında yine RF %89.32 doğruluk oranı ile en yüksek performans elde edildiği görülmektedir.

Tablo 65. SBFS Öznitelik Seçim Öncesi ve Sonrası ML Algoritmalarının Doğruluk Oran Karşılaştırması

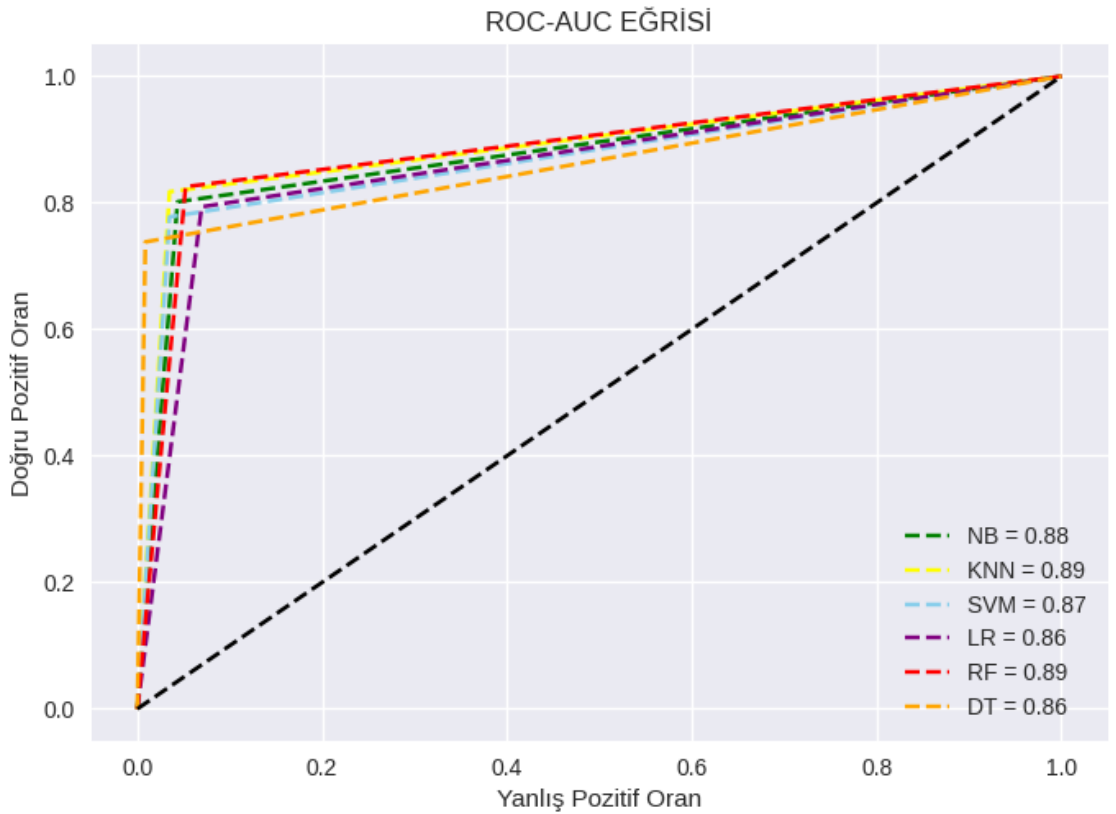
ML Algoritmaları	ML Doğruluk Oranları (%)	
	SBFS Öznitelik Seçim Öncesi	SBFS Öznitelik Seçim Sonrası
NB	84.64	86.83
KNN	83.15	88.25
SVM	82.89	85.94
LR	83.77	86.29
RF	87.13	89.32
DT	85.89	87.54

Bu tez çalışmasında, veri ön işleme ve CV yöntemi uygulanan veri setine sırasıyla SFS, SBS, SFFS ve SBFS öznitelik seçim yöntemleri uygulanmıştır ve performans sonuçları gösterilmiştir. Öznitelik seçim yöntemleri uygulamadan önce ve sonra ML algoritmalarının doğruluk performansları ve karşılaştırılması Tablo 66'da gösterilmiştir.

Tablo 66. Öznitelik Seçim Yöntemleri Uygulamadan Önce ve Sonra ML Algoritmalarının Performans Karşılaştırmaları

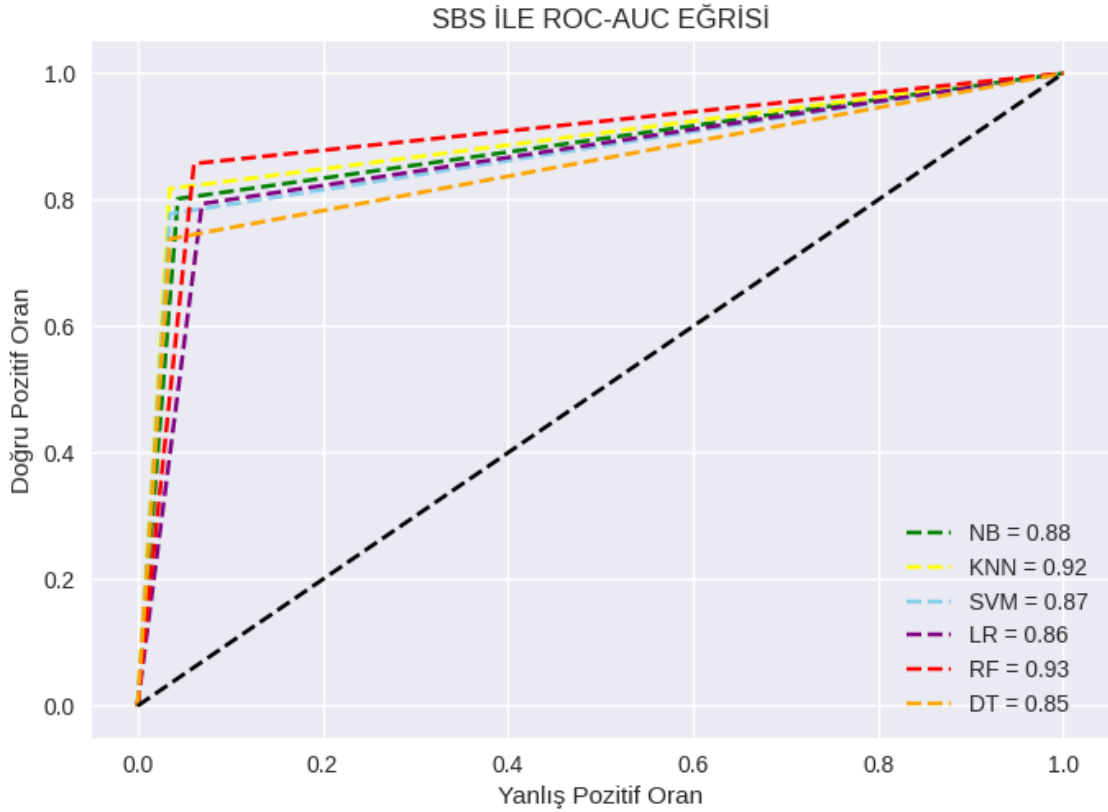
Oluşturulan Modeller	Öznitelik Seçim Yöntemleri Uygulamadan Önce ve Sonra ML algoritmalarının ACC ve AUC Oranları (%) (5 katlı çapraz doğrulama)					
	METRİKLER	NON-YÖNTEM	SFS	SBS	SFFS	SBFS
NB	ACC	84.64	86.83	86.83	86.83	86.83
	AUC	87.73	87.97	87.08	87.97	87.08
KNN	ACC	83.15	87.72	87.89	87.72	88.25
	AUC	89.03	94.21	95.05	94.21	96.01
SVM	ACC	82.89	85.94	85.94	85.94	85.94
	AUC	87.59	92.11	91.89	92.11	91.89
LR	ACC	83.77	86.29	86.29	86.29	86.29
	AUC	88.75	89.58	90.49	89.58	90.49
RF	ACC	87.00	89.50	89.86	89.68	89.32
	AUC	89.32	95.94	95.26	94.46	94.79
DT	ACC	86.01	87.18	87.36	87.53	87.54
	AUC	86.47	91.34	93.77	93.77	90.58

Tablo 66’da görüldüğü gibi öznitelik seçim yöntemleri uygulama sonrasında elde edilen %89.86 doğruluk oranı ile RF algoritması SBS öznitelik seçim yöntemiyle en iyi performansı göstermiştir.



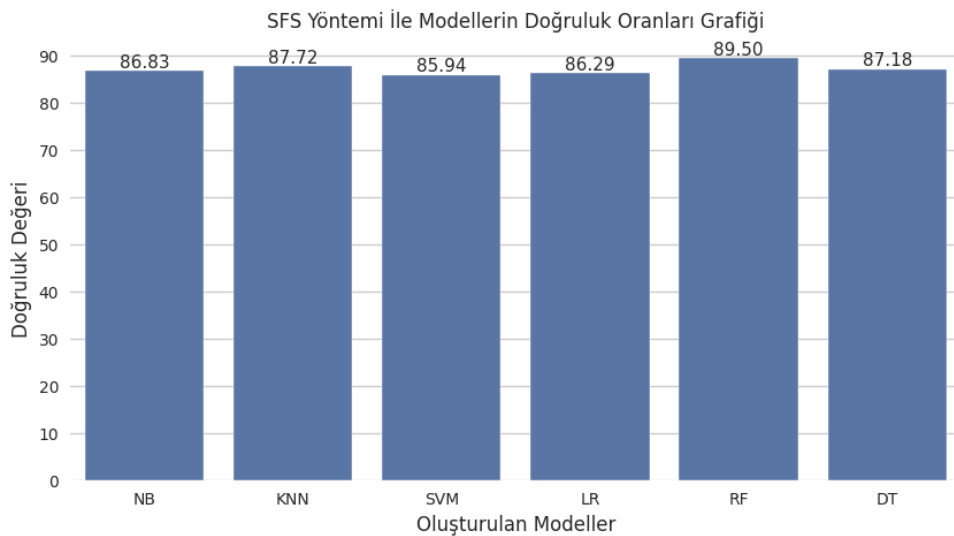
Şekil 46. ROC-AUC eğrisi.

Şekil 46’da yöntem kullanılmadan elde edilen modellere bağlı olarak ROC-AUC eğrisi elde edilmiştir. Şekilde görüldüğü gibi %89 doğruluk değerinde KNN ve RF modellerinin diğer modellere göre yüksek elde edildiği söylenebilir.



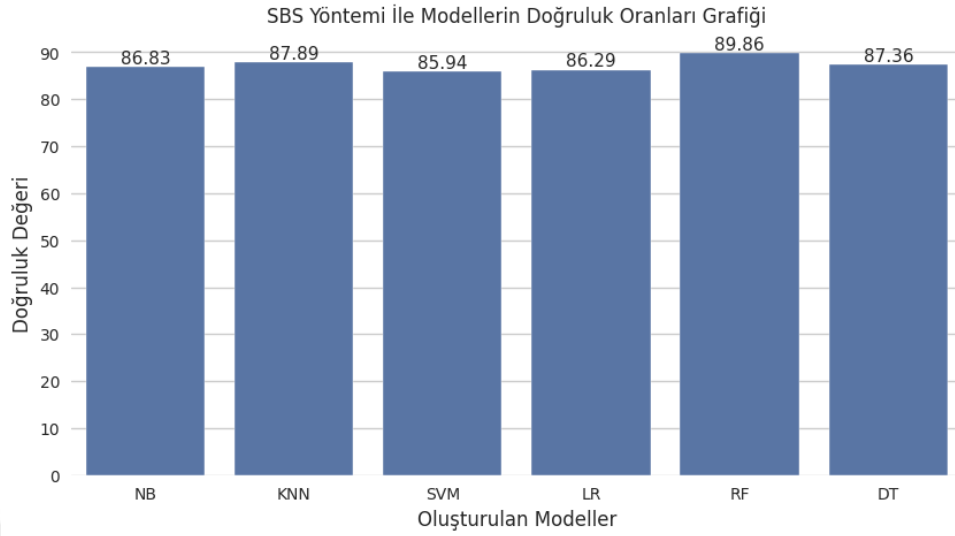
Şekil 47. SBS ile ROC-AUC eğrisi.

Şekil 46’da SFS yöntemi kullanılarak elde edilen modellere bağlı ROC-AUC eğrisi elde edilmiştir. Şekilde görüldüğü gibi %93 AUC değerinde RF modelinin diğer modellere göre yüksek değer elde edildiği söylenebilir.



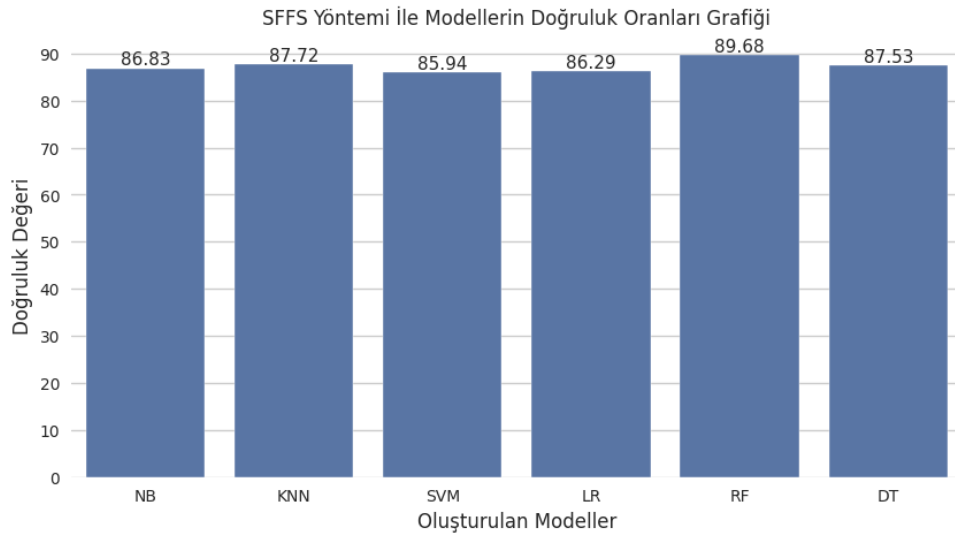
Şekil 48. SFS Yöntemi ile Modellerin Doğruluk Oranlar Grafiği

Şekil 48 incelendiğinde RF algoritmasının %89.50 ile yüksek başarılı performans elde edilmiştir.



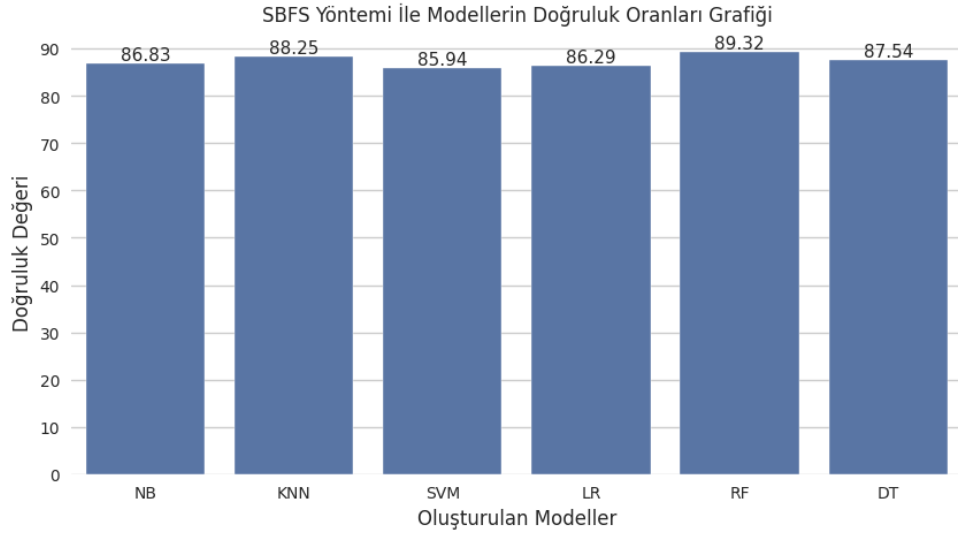
Şekil 49. SBS Yöntemi ile Modellerin Doğruluk Oranlar Grafiği

Şekil 49 incelendiğinde RF algoritmasının %89.86 ile yüksek başarılı performans elde edilmiştir.



Şekil 50. SFFS Yöntemi ile Modellerin Doğruluk Oranlar Grafiği

Şekil 50 incelendiğinde RF algoritmasının %89.68 ile yüksek başarılı performans elde edilmiştir.



Şekil 51. SBFS Yöntemi ile Modellerin Doğruluk Oranlar Grafiği

Şekil 51 incelendiğinde RF algoritmasının %89.32 ile yüksek başarılı performans elde edilmiştir.

SONUÇ VE ÖNERİLER

Bu tez çalışmasında, talasemi teşhisi için öznitelik seçme yöntemleri uygulanıp, ML algoritmaları ile sınıflandırma yapılmıştır. Bununla birlikte talasemi hastası ve sağlıklı bireylerin başarılı bir şekilde sınıflandırılmasını sağlayan yüksek doğruluk değerlerinin elde edildiği yöntemler değerlendirilip, karşılaştırmalar yapılmıştır. Veri seti Erzurum Atatürk Üniversitesi Araştırma ve Eğitim Hastanesi Hematoloji Bölümüne gelen hasta bilgilerinden oluşmaktadır. 803 örnekten oluşan veri seti 0-81 yaş aralığında örneklem ve 9 öznitelikten oluşmaktadır. Bu öznitelikler, hemoglobin, hemoglobin a, hemoglobin a2, hemoglobin f, mcv, ferritin, demir, cinsiyet ve yaş olmaktadır.

Veri seti üzerinde veri ön işleme, öznitelik seçme işlemleri uygulanıp NB, KNN, SVM, LR, RF, DT olan ML algoritmaları ile ayrı ayrı sınıflandırılma işlemleri yapıp modeller kurulmuştur. İlk olarak sınıflandırma sonuçlarından daha yüksek performans elde etmek ve doğru bir şekilde sınıflandırma yapabilmek amacı ile veri ön işleme adımları gerçekleştirilmiştir. Bunlar; eksik veri tamamlama, tekrarlayan veri temizleme, aykırı değer yakalama, kategorik verileri sayısal verilere dönüştürme, veri ölçeklendirme gibi birçok veri ön işleme adımları uygulanmıştır. Veri ön işleme adımları gerçekleştirilen veri seti, %70'i eğitim, %30'u test verisi olarak kullanılmak üzere ikiye ayrılmıştır. ML sınıflandırma işlemlerinde en iyi değerlere ulaşabilmek için her sınıflandırma algoritması için öznitelik seçim yöntemleri kullanılmıştır. Sırasıyla SFS, SBS, SFFS, SBFS yöntemleri modellere uygulanmıştır. Yapılan çalışmada sınıflandırma sonucunda, oluşturulan modellerin performanslarının değerlendirilmesinde farklı performans metrikleri kullanılmıştır. Performans değerlendirmesinde karmaşıklık matrisleri oluşturulmuş daha sonra kesinlik, duyarlılık, f1-skoru, doğruluk ve AUC gibi kriterlere göre model performanslarının başarı oranları arasında karşılaştırma yapılmıştır.

Bu tez çalışmasında, modeller kurulduktan sonra herhangi bir öznitelik seçme ve CV yöntemi kullanılmadan önce oluşturulan ML modellerinde sınıflandırma yapılmış ve başarı sonuçları değerlendirilmiştir. Bu değerlendirme sonucunda, RF modeli %89,63 doğruluk değeri ile en yüksek başarı performansı elde edilmiştir. Diğer modellerde sırası ile, %88,80 KNN, %87,55 NB, %86,72 SVM, %85,89 LR ve DT modellerinin doğruluk değerleri elde edilmiştir. Daha sonra modellere CV uygulanmış ve k sayısı 5 olarak alınmıştır. Bu yöntem ile modellerin başarı performansları değerlendirilmiştir. CV uygulanmadan önce olduğu gibi CV uygulandıktan sonra da RF modeli %87.00 doğruluk değeri ile en yüksek başarı performansı

elde edilmiştir. Diğer modellerde sırası ile %86,01 DT, %84,64 NB, %83,77 LR, %83,15 KNN ve %82,89 SVM modellerinin doğruluk değerleri elde edilmiştir. CV ile az da olsa doğruluk değerlerinde azalma görülmüştür.

Daha sonra oluşturulan modellere öznitelik çıkarma yöntemleri sırası ile uygulanmıştır. SFS yöntemi sonucunda, RF modeli %89,50 doğruluk, %95,94 AUC değeri ile en yüksek, SVM modeli %85,94 doğruluk, %92,11 AUC değeri ile en düşük başarı performansı elde edilmiştir. SBS yöntemi sonucunda, RF modeli %89,86 doğruluk, %95,26 AUC değeri ile en yüksek, SVM modeli %85,94 doğruluk, %91,89 AUC değeri ile en düşük başarı performansı elde edilmiştir. SFFS yöntemi sonucunda, RF modeli %89,68 doğruluk, %94,46 AUC değeri ile en yüksek, SVM modeli %85,94 doğruluk, %92,11 AUC değeri ile en düşük başarı performansı elde edilmiştir. SBFS yöntemi sonucunda, RF modeli %89,32 doğruluk, %94,79 AUC değeri ile en yüksek, SVM modeli %85,94 doğruluk, %91,89 AUC değeri ile en düşük başarı performansı elde edilmiştir. Bu sonuçlar doğrultusunda, öznitelik seçim yöntemleri uygulama sonrasında elde edilen %89.86 doğruluk oranı ile RF algoritması SBS öznitelik seçim yöntemiyle en iyi performansı göstermiştir.

Tez çalışması sonucunda, önerilen modeller doğrultusunda yeni bir veri seti kullanılarak talasemi hastalığı tahmini gerçekleştirilmiştir. Literatüre yeni bir veri seti ile katkı sağlanmıştır. Farklı ML sınıflandırma yöntemleri, literatürden farklı veri seti ve parametreler ile talasemi tahmini için başarılı performans sonuçları elde edildiği görülmektedir. Gerçekleştirilen bu çalışma ile bireylerin, talasemi hastası olup olmadığını doğru bir şekilde, erken teşhis edilmesi doktorların hastalık için erken ve doğru karar vermesine olanak sağlayacaktır. İlerde yapılacak olan çalışmalarda daha büyük veri seti üzerinde, farklı parametreler eklenerek talasemi hastalıklarının teşhisinde ML yöntemlerinin yanı sıra derin öğrenme yöntemleri ile hibrit modeller oluşturup, bu modeller ile etkili tahminleme ve analizler yapılması hedeflenmektedir.

KAYNAKLAR

- Agraz, M. (2023). Comparison of Feature Selection Methods in Breast Cancer Microarray Data. *Medical Records*, 5(2), 284-289.
- Atasoy, D., & Kara, H. (2023). Karar ağacı optimizasyon algoritması üzerine bir çalışma. *Journal of the Institute of Science and Technology*, 13(2), 1247-1255.
- Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.
- Budak, H. (2018). Özellik seçim yöntemleri ve yeni bir yaklaşım. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 22, 21-31.
- Calayır, G. N., & Kabak, M. (2021). Bakım için makine öğrenme tekniklerinin analizi ve bir uygulama. *Journal of Turkish Operations Management*, 5(1), 662-675.
- Cengiz, T., & Demir, M. C. (2020). Detection and classification of muscle activation in EMG data acquired by Myo armband. *Avrupa Bilim ve Teknoloji Dergisi*, 178-183.
- Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2, 117693510600200030.
- Das, R., Saleh, S., Nielsen, I., Kaviraj, A., Sharma, P., Dey, K., & Saha, S. (2022). Performance analysis of machine learning algorithms and screening formulae for β -thalassemia trait screening of Indian antenatal women. *International Journal of Medical Informatics*, 167, 104866.
- Demir, A. (2019). *Beta talasemi majör tanılı hastalarda transfüzyon öncesi ve sonrası amino asit profilinin incelenmesi/Investigation of amino acid profile before and after transfusion in patients with beta thalassemia major*
- Devanath, A., Akter, S., Karmaker, P., & Sattar, A. (2022). Thalassemia Prediction using Machine Learning Approaches. 2022 6th International Conference on Computing Methodologies and Communication (ICCMC),
- Doğruel, M., & Fırat, S. Ü. (2021). Veri Madenciliği Karar Ağaçları Kullanarak Ülkelerin İnovasyon Değerlerinin Tahmini ve Doğrusal Regresyon Modeli ile Karşılaştırmalı Bir Uygulama. *Istanbul Business Research*, 50(2), 465-493.
- Dursun, Ö. O., Toraman, S., & Türkoğlu, A. (2017). Comparison Of The Classification Performances Of Criminal Tendencies Of Schizophrenic Patients By Artificial Neural Networks And Support Vector Machine. *European Journal of Technique (EJT)*, 7(2), 177-185.
- Eröz, B. (2010). *Veri yapısına bağlı olarak Roc eğrisi altında kalan alana ilişkin istatistiksel yöntemlerin karşılaştırılması* Sağlık Bilimleri Enstitüsü].
- Farzaliyev, E., Saihood, Q., & Sonuç, E. (2023). Çocuklarda Anemi Hastalığının Teşhisinde Topluluk Öğrenme Yöntemlerinin Kullanılması. 1 st International Conference on Recent Academic Studies,
- Ferih, K., Elsayed, B., Elshoeibi, A. M., Elsabagh, A. A., Elhadary, M., Soliman, A., Abdalgayoom, M., & Yassin, M. (2023). Applications of artificial intelligence in thalassemia: A comprehensive review. *Diagnostics*, 13(9), 1551.
- Fu, Y.-K., Liu, H.-M., Lee, L.-H., Chen, Y.-J., Chien, S.-H., Lin, J.-S., Chen, W.-C., Cheng, M.-H., Lin, P.-H., & Lai, J.-Y. (2021). The tvgh-nycu thal-classifier: Development of a machine-learning classifier for differentiating thalassemia and non-thalassemia patients. *Diagnostics*, 11(9), 1725.

- Gao, J., & Liu, W. (2022). Advances in screening of thalassaemia. *Clinica Chimica Acta*, 534, 176-184.
- Karaaziz, M., & Okyayuz, Ü. H. (2020). Bir talasemi hastasının hastalık ile uyumluluğunun incelenmesi: olgu sunumu. *Cukurova Medical Journal*, 45(1), 362-369.
- Laengsri, V., Shoombuatong, W., Adirojananon, W., Nantasenamat, C., Prachayasittikul, V., & Nuchnoi, P. (2019). ThalPred: a web-based prediction tool for discriminating thalassemia trait and iron deficiency anemia. *BMC medical informatics and decision making*, 19, 1-14.
- Liu, S. (2024). An Application of Machine Learning to Thalassemia Diagnosis. *Journal of Computer and Communications*, 12(2), 211-230.
- Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)*. [Internet], 9(1), 381-386.
- Masala, G. L., Golosio, B., Cutzu, R., & Pola, R. (2013). A two-layered classifier based on the radial basis function for the screening of thalassaemia. *Computers in biology and medicine*, 43(11), 1724-1731.
- Mirghaed, M. T., Hedayatirad, S., Farhadi, Z., Sepehrian, R., Salemi, M., & Raeisi, P. (2020). Correlation of General Health Status of Patients with Major Thalassemia with Demographic Characteristics. *J. Health Promot. Manag*, 9(1), 18-24.
- Mo, D., Zheng, Q., Xiao, B., & Li, L. (2023). Predicting thalassemia using deep neural network based on red blood cell indices. *Clinica Chimica Acta*, 543, 117329.
- Mohammed, M. Q., & Al-Tuwaijari, J. M. (2021). A Survey on various Machine Learning Approaches for thalassemia detection and classification. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(13), 7866-7871.
- Noferesti, S., Shemshadi Nejad, N., & Heydari, F. (2019). Using data mining models for differential diagnosis of iron deficiency anemia and β -thalassemia minor. *Journal of Health and Biomedical Informatics*, 5(4), 435-446.
- Nti, I. K., Nyarko-Boateng, O., & Aning, J. (2021). Performance of machine learning algorithms with different K values in K-fold CrossValidation. *International Journal of Information Technology and Computer Science*, 13(6), 61-71.
- Ospanov, B., & Demirci, M. F. (2023). Pekiştirmeli Öğrenme ile Tenis Oyunu Simülasyonu Gerçekleştirimi. *Journal of Turkish Operations Management*, 7(2), 1724-1735.
- Ozcift, A., & Gulen, A. (2011). Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. *Computer methods and programs in biomedicine*, 104(3), 443-451.
- Özdemir, Ş., & Örsül, S. (2019). Makine öğrenmesinde yeni bir bakış açısı: otomatik makine öğrenmesi (AutoML). *Journal of Information Systems and Management Research*, 1(1), 23-30.
- Paokanta, P., Ceccarelli, M., & Srichairatanakool, S. (2010). The efficiency of data types for classification performance of Machine Learning Techniques for screening β -Thalassemia. 2010 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL 2010),
- Phirom, K., Charoenkwan, P., Shoombuatong, W., Charoenkwan, P., Sirichotiyakul, S., & Tongsong, T. (2022). DeepThal: A Deep Learning-Based Framework for the Large-Scale Prediction of the α^+ -Thalassemia Trait Using Red Blood Cell Parameters. *Journal of Clinical Medicine*, 11(21), 6305.

- Sadiq, S., Khalid, M. U., Ullah, S., Aslam, W., Mehmood, A., Choi, G. S., & On, B.-W. (2021). Classification of β -thalassemia carriers from red blood cell indices using ensemble classifier. *IEEE access*, 9, 45528-45538.
- Saeyns, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19), 2507-2517.
- Sel, A. (2021). Hibrit Regresyon Modelleri İle BİST'e Etki Eden G20 Endekslerinin Belirlenmesi. *Süleyman Demirel Üniversitesi Vizyoner Dergisi*, 12(31), 870-884.
- Sevli, O. (2022). Farklı sınıflandırıcılar ve yeniden örnekleme teknikleri kullanılarak kalp hastalığı teşhisine yönelik karşılaştırmalı bir çalışma. *Journal of Intelligent Systems: Theory and Applications*, 5(2), 92-105.
- Sözer, A. T. (2022). Denetimsiz Anomali Tespiti Yaklaşımı ile Düşme Algılama. *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, 11(1), 88-98.
- Şapcı, B., & Pektaş, Ş. T. (2021). Makine Öğrenmesi Aracılığı ile Kullanıcı Deneyimi Bilgilerinin Erken Mimari Tasarım Süreçleriyle Bütünleştirilmesi. *Journal of Computational Design*, 2(1), 67-94.
- Veranyurt, Ü., Deveci, A., Esen, M. F., & Veranyurt, O. (2020). Makine Öğrenmesi Teknikleriyle Hastalık Sınıflandırması: Random Forest, K-Nearest Neighbour Ve Adaboost Algoritmaları Uygulaması. *Uluslararası Sağlık Yönetimi ve Stratejileri Araştırma Dergisi*, 6(2), 275-286.
- Yağmur, N., Temurtaş, H., & İdris, D. (2023). Anemi Hastalığının Yapay Sinir Ağları Yöntemleri Kullanılarak Sınıflandırılması. *Journal of Scientific Reports-B(008)*, 20-34.

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı:	Ece Gülşah ABBASOĞULLARI
Doğum tarihi:	
Doğum Yeri:	
Uyruğu:	
Adres:	
Tel:	
E-mail:	
Eğitim	
Lise:	Erzurum Mecidiye Anadolu Lisesi (4 yıl)
Lisans:	Atatürk Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü
Yabancı Dil Bilgisi	
İngilizce:	İyi
Üye Olunan Mesleki Kuruluşlar	
Türkiye Bilişim Derneği	