

**ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

DOKTORA TEZİ

Uğur Engin EŞSİZ

**LABORATUVAR TETKİK SONUÇLARI ÜZERİNDE MAKİNE
ÖĞRENMESİ UYGULAMALARI**

ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

ADANA-2022



ÖZ

DOKTORA TEZİ

LABORATUVAR TETKİK SONUÇLARI ÜZERİNDE MAKİNE ÖĞRENMESİ UYGULAMALARI

Uğur Engin EŞSİZ

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
ENDÜSTRİ MÜHENDİSLİĞİ ANABİLİM DALI

Danışman : Doç. Dr. Hacire Oya YÜREGİR
Yıl:2022, Sayfa:172
Jüri : Doç. Dr. Oya Hacire YÜREGİR
: Prof. Dr. Ali KOKANGÜL
: Doç. Dr. Murat OTURAKÇI
: Dr. Öğr. Üyesi Ebru YILMAZ
: Dr. Öğr. Üyesi Mümine KAYA KELEŞ

Vitamin D hem yetişkin hem de çocuk sağlığı için gerekli vitaminler arasındadır ve eksikliği birçok hastalığın habercisi olarak görülmektedir. Diyabet ise birçok ölümcül hastalığın oluşumunda öncü rol oynayan kronik bir sağlık durumudur. Bu çalışmada, nitelik seçimi ile birleştirilmiş veri madenciliği tekniklerine dayalı olarak vitamin D eksikliği teşhisi önerilmiştir. Yapılan çalışma kan testi sonuçları üzerinden vitamin D değerini doğru tahminlemeye çalışırken aynı zamanda tahmin için gerekli parametreleri de azaltmayı hedeflemektedir. Makine öğrenmesi ile yapılan tahminler, laboratuvar ortamında yapılan vitamin D ölçüm sürecinden kıyaslanamayacak derecede hızlı sonuç vermektedir. En iyi sonuç diyabet hastalarının veri seti üzerinde tamamlanan deneyler sonucunda alınmış ve relif-f nitelik seçimi algoritmasıyla seçilen 19 nitelikten oluşan destek vektör makineleri modeli için % 98,030 ile en yüksek sınıflandırma doğruluğunun elde edildiği görülmüştür. Çalışma sonucunda, yapılan testlerin azaltılabileceği görülmüştür. Yöntemlerin performansları sınıflandırma doğruluğu, duyarlılık, özgüllük, kesinlik, f-measure ve kappa gibi performans ölçütleri kullanılarak değerlendirilmiştir. Ayrıca bu çalışmayla tüm parametrik ve parametrik olmayan algoritmalar için yeni bir kıyaslama veri seti hazırlanmıştır.

Anahtar Kelimeler: Vitamin D, Diyabet, Veri Madenciliđi, Makine Öğenmesi, Karar Destek Sistemi

ABSTRACT

PhD THESIS

**MACHINE LEARNING APPLICATIONS ON LABORATORY TEST
RESULTS**

Uğur Engin EŞSİZ

**ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF INDUSTRIAL ENGINEERING**

Supervisor : Assoc. Prof. Dr. Oya Hacire YÜREGİR

Year:2022, Page:172

Jury : Assoc. Prof. Dr. Hacire Oya YÜREGİR
: Prof. Dr. Ali KOKANGÜL
: Assoc. Prof. Dr. Murat OTURAKÇI
: Asst. Prof. Dr. Ebru YILMAZ
: Asst. Prof. Dr. Mümine KAYA KELEŞ

Vitamin D is among the essential vitamins for both adult and child health, and its deficiency is seen as a precursor to many diseases. Diabetes is a chronic health condition that plays a leading role in the formation of many fatal diseases. In this study, a diagnosis of vitamin D deficiency is proposed based on data mining techniques combined with feature selection. The study aims to accurately estimate the vitamin D value through blood test results, while also aiming to reduce the necessary parameters for estimation. Predictions made with machine learning give results incomparably faster than the vitamin D measurement process made in the laboratory. The best results were obtained as a result of the experiments completed on the data set of diabetes patients and it was seen that the highest classification accuracy was obtained with 98.030 % for the support vector machine model consisting of 19 features selected with the relief-f feature selection algorithm. As a result of the study, it has been seen that the number of the tests can be reduced. The performance of the methods was evaluated using performance metrics such as classification accuracy, sensitivity, specificity, precision, f-measure, and kappa. In addition, with this study, a novel benchmark data set was prepared for all parametric and non-parametric algorithms.

Keywords: Vitamin D, Diabetes, Data Mining, Machine Learning, Decision Support System

GENİŞLETİLMİŞ ÖZET

Vitamin D hem yetişkin hem de çocuk sağlığı için gerekli vitaminler arasındadır ve eksikliği birçok hastalığın habercisi olarak görülmektedir. Kalsiyum emilimi, bağışıklık fonksiyonu, kemik koruması, kas sağlığı, kalp sağlığı ve glikoz metabolizması üzerinde önemli rol oynar. Diyabet ise birçok ölümcül hastalığın oluşumunda öncü rol oynayan kronik bir sağlık durumudur.

Vitamin D eksikliği tahmini, bir hastanın vitamin D değerlerini dış ve iç verilere dayalı olarak bulmayı amaçlar. Vitamin D eksikliği birçok hastalığın habercisi olarak görülmektedir. Bu sebeple bu eksikliğin doğru tahmini hem hastalar hem de klinisyen için önemli bir yere sahiptir. Bu tahmin sayesinde laboratuvar testi yapılmaksızın hızlı sonuç alınabilmesi doktorun hastayı etkin yönlendirmesine ve hastanın gereken önlemi vaktinde alabilmesine olanak sağlamaktadır. Artan doğruluk için çok değişkenli girdiler tercih edildiğinden, tahmin yapmak için daha esnek modellere ihtiyaç vardır.

Makine öğrenmesi yöntemleri bu anlamda çok kullanışlıdır. Kan tahlili sonuçları üzerinden kandaki glukoz değeri, COVID-19 test sonucu gibi birçok değer tahmini araştırmacılar tarafından makine öğrenmesi yöntemleri yardımıyla çalışılmış ve farklı algoritmaların karşılaştırılması bu çalışmalar dahilinde yapılmıştır.

Literatüre bakıldığında vitamin D tahmini üzerinde makine öğrenmesi yöntemleri ile yapılmış çalışmalara nadiren rastlanmaktadır. Tahmin çalışmaları sırasında en büyük zorluklardan biri hatasız ve gürültüsüz veri toplamaktır. Bu tez çalışmasıyla tüm parametrik ve parametrik olmayan algoritmalar için bir kıyaslama (benchmark) veri seti hazırlanmıştır. Makine öğrenmesi ile yapılan tahminler, laboratuvar ortamında yapılan vitamin D ölçüm sürecinden kıyaslanamayacak derecede hızlı sonuç vermektedir.

Bu çalışmada, nitelik seçimi ile birleştirilmiş veri madenciliği tekniklerine dayalı olarak vitamin D eksikliği teşhisi önerilmiştir. Yapılan çalışma kan testi sonuçları üzerinden vitamin D değerini doğru tahminlemeye çalışırken aynı zamanda tahmin için gerekli parametreleri de azaltmayı hedeflemektedir. Bu hedef doğrultusunda analiz kısmında farklı sınıflandırma, kümeleme ve nitelik seçimi yöntemleri kullanılarak yöntemlerin bireysel ve hibrit karşılaştırmaları yapılmıştır.

Empirik olarak çalışmanın tahminleme kısmında kullanılan sınıflandırma yöntemlerinin destek vektör makineleri (DVM), k-en yakın komşuluk algoritması (KNN), karar ağaçları (KA), naive bayes (NB), yapay sinir ağları (YSA), rastgele orman (RO) ve lojistik regresyon (LR) olmasına karar verilmiştir. Kümeleme için k-ortalama algoritması; nitelik seçimi için de relief-f, info gain, gain ratio, faktör analizi (FA) ve lojistik regresyon (LR) yöntemlerinin kullanılması uygun görülmüştür.

Makine öğrenmesi ile yapılan tahminler, laboratuvar ortamında yapılan vitamin D ölçüm sürecinden kıyaslanamayacak derecede hızlı sonuç vermektedir. En iyi sonuç diyabet hastalarının veri seti üzerinde tamamlanan deneyler sonucunda alınmış ve relief-f nitelik seçimi algoritmasıyla seçilen 19 nitelikten oluşan destek vektör makineleri modeli için % 98,030 ile en yüksek sınıflandırma doğruluğunun elde edildiği görülmüştür. Çalışma sonucunda, yapılan testlerin azaltılabileceği görülmüştür. Yöntemlerin performansları sınıflandırma doğruluğu, duyarlılık, özgüllük, kesinlik, f-measure ve kappa gibi performans ölçütleri kullanılarak değerlendirilmiştir.

Ayrıca bu çalışmayla tüm parametrik ve parametrik olmayan algoritmalar için yeni bir kıyaslama (benchmark) veri seti hazırlanmıştır. Hazırlanan özgün veri seti, tahmin çalışmaları sırasında en büyük zorluklardan biri olan hatasız ve gürültüsüz veri toplama sorununa da katkı sağlamaktadır. Veri setinin konu üzerinde çalışacak olan araştırmacılara açık hale getirilmesi literatüre büyük katkılar sağlayacaktır.

TEŞEKKÜR

Tüm doktora sürecinde deneyim ve bilgisiyle bana sonsuz katkıda bulunan, ufkun ötesini görmemi sağlayan değerli danışmanım Doç. Dr. Hacıre Oya YÜREGİR'e; kritik noktalarda müdahaleleriyle beni yönlendiren ikinci danışmanım Dr. Öğr. Üyesi Birsen İrem SELAMOĞLU'na; destekleri ve pozitif enerjisiyle yoluma ışık tutan Prof. Dr. Ali KOKANGÜL'e; bana her an yardımcı olan, desteğini esirgemeyen Doç. Dr. Murat OTURAKÇI'ya; katkılarından ötürü jüri üyelerim Dr. Öğr. Üyesi Ebru YILMAZ ve Dr. Öğr. Üyesi Mümine KAYA KELEŞ'e teşekkürlerimi sunarım.

Sevgili eşim Dr. Öğr. Üyesi Esra SARAÇ EŞSİZ'e sonsuz desteği ve sabrından ötürü; biricik oğlum Kaya EŞSİZ'e hayatıma kattıklarından dolayı teşekkürlerimi sunarım.

Hayatımın her anında bana destek olan babam Dr. Mehmet Zeki EŞSİZ, annem Nurten ÖZDEMİR EŞSİZ, kardeşlerim Hasan Erdem EŞSİZ ve Cem Erman EŞSİZ'e teşekkürlerimi sunarım.

Prof. Dr. Mehmet Fatih AKAY, Prof. Dr. Selma Ayşe ÖZEL, Prof. Dr. Nuri Selim KADIOĞLU, Doç. Dr. Sami ARICA, Doç. Dr. Çiğdem ACI, Dr. Öğr. Üyesi Mehmet ACI, Baturay Fatih ÇOLAKOĞLU, Murat ULUBAŞ ve Yunus TENTİF'e desteklerinden dolayı teşekkürlerimi sunarım.

Adana Su ve Kanalizasyon İdaresi'nden Aykut AVŞAR, Mevlüt AKYILDIZ, Mehmet IŞIK, Ferit IŞIK, Serhat COŞKUN, Zeliha TEKİN, Aydın CİNT, Kürşad Kağan BATUR, Hasan Gökhan ÜNAL, Gökhan YÜKSEL, Doğan ERGEZER, Onat TUFAN, Erdem TORUK, Mustafa DEĞİRMENCİ, Selahattin AYSAN, Mustafa TÜRKYAŞAR, Tuğrul ODABAŞ, Aysun ÖZVEREN, Mehmet BEKTAŞ ve Ramazan ANAR'a yardımlarından dolayı teşekkürlerimi sunarım.

İÇİNDEKİLER

SAYFA

ÖZ	
I ABSTRACT	
II GENİŞLETİLMİŞ ÖZET	
III	TEŞEKKÜR
.....	V
İÇİNDEKİLER	
VI ÇİZELGELER DİZİNİ	
XII ŞEKİLLER DİZİNİ	
XVI	
SİMGELER VE KISALTMALAR	XVIII
1. GİRİŞ	
1	
1.1. Makine Öğrenmesi	
2	
1.2. Vitamin D	2
1.3. Diyabet	
5	
1.4. Çalışmanın Önemi ve Amacı	
5	
1.5. Çalışmada Kullanılan Materyal	6
1.6. Çalışmada Kullanılan Yöntemler	
7	
1.7. Çalışmanın Aşamaları	
8	
2. ÖNCEKİ ÇALIŞMALAR	9
2.1. Vitamin D Eksikliğinin Literatür Araştırması	9
2.3. Tıpta Makine Öğrenmesi Çalışmalarının Literatür Araştırması	12
3. MATERYAL VE METOT	15
3.1. Materyal	
15	

3.2. Metod	18
3.2.1 Makine Öğrenmesi	21
3.2.1.1. Sınıflandırma	21
3.2.1.1.(1). Destek Vektör Makineleri	23
3.2.1.1.(1).(a) Doğrusal Olarak Ayrılabilen Destek Vektör Makineleri	23
3.2.1.1.(1).(b) Doğrusal Olarak Ayrılamayan Destek Vektör Makineleri	26
3.2.1.1.(2) K-En Yakın Komşuluk Algoritması	28
3.2.1.1.(2).(a) K- En Yakın Komşuluk Algoritmasında Kullanılan Uzaklıklar ve Komşu Sayısı	29
3.2.1.1.(3) Karar Ağaçları	30
3.2.1.1.(4) Naive Bayes	32
3.2.1.1.(5) Yapay Sinir Ağları	34
3.2.1.1.(6) Rastgele Orman	37
3.2.1.2. Nitelik Seçimi	37
3.2.1.2.(1) Bilgi Kazanımı (Info Gain)	38
3.2.1.2.(2) Ki-Kare (Chi Square)	39
3.2.1.2.(3) Kazanım Oranı (Gain Ratio)	40
3.2.1.2.(4) Relief-f	40
3.2.1.3. Hata Matrisi (Confusion Matrix)	41
3.2.1.4. Kümeleme	43
3.2.1.4.(1) K-Ortalama Algoritması	43
3.2.1.4.(2) Optimum Küme Sayısının Belirlenmesi	44
3.2.3. Çalışmada Kullanılan İstatistiki Yöntemler	45
3.2.3.1. Örneklem Büyüklüğü	45

3.2.3.2. Faktör Analizi	46
3.2.3.2.(1) Veri Setinin Faktör Analizi İçin Uygunluğunun Test Edilmesi	47
3.2.3.2.(2) Faktörlerin Elde Edilmesi	48
3.2.3.2.(3) Faktörlerin Rotasyonu	48
3.2.3.2.(4) Örneklemin Genel Özellikleri ve Veri Ön İşleme	48
3.2.3.2.(5) Eksik Veri Analizi	49
3.2.3.3. Lojistik Regresyon	49
3.2.3.3.(1) Modelin Oluşturulması	50
3.2.3.3.(2) Modelin Uyum İyiliği	51
4. BULGULAR VE TARTIŞMA	55
4.1. Tüm Veri Üzerinde Yapılan Çalışmalar	55
4.1.1 Sınıflandırma Sonuçları	55
4.1.2. Nitelik Seçimi Sonuçları	56
4.1.3. Nitelik Seçimi Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları	56
4.1.4. Kümeleme Sonuçları	59
4.1.5. Kümeleme Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları	62
4.1.6. Faktör Analizi Sonuçları	63
4.1.7. Lojistik Regresyon Analizi Sonuçları	66
4.2. Diyabet Hastalarının Veri Seti Üzerinde Yapılan Çalışmalar	69
4.2.1. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Sınıflandırma Sonuçları	69
4.2.2. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Nitelik	

Seçimi Sonuçları	71
4.2.3. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Nitelik Seçimi Sonuçlarına Göre Tekrarlanan Sınıflandırma	
Sonuçları	71
4.2.4. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Kümeleme	
Sonuçları	74
4.2.5. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Kümeleme Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları	75
4.2.6. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Faktör	
Analizi Sonuçları	77
4.2.7. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Lojistik Regresyon Sonuçları	78
4.3. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Yapılan Çalışmalar	80
4.3.1. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Sınıflandırma Yöntemleri Sonuçları	80
4.3.2. Diyabet Hastalığı Olmayan Hastaların Üzerinde Çalışılan Nitelik Seçimi Sonuçları	81
4.3.3. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Nitelik Seçimi Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları	82
4.3.4. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Kümeleme Sonuçları	85
4.3.5. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Kümeleme Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları	85
4.3.6. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Faktör Analizi Sonuçları	87
4.3.7. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde	

Çalışılan Lojistik Regresyon Sonuçları	87
4.4. Tartışma	88
4.4.1 Tüm Veri Üzerinde Yapılan Çalışmaların Özeti	88
4.4.2. Diyabet Hastalarının Veri Seti Üzerinde Yapılan Çalışmaların Özeti	88
4.4.3. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Yapılan Çalışmaların Özeti	89
4.4.4. Çalışmada Kullanılan Yöntemler	90
4.4.5. Çalışmanın Literatür ile Kıyaslanması	90
4.4.6. En İyi Sınıflandırma Sonuçlarının Karşılaştırılması	93
4.4.7. En İyi Nitelik Seçimi Sonuçlarının Karşılaştırılması	93
4.4.8. Çalışmada En İyi Sonucu Veren Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Destek Vektör Makineleri Kullanarak Elde Edilen En İyi Sonuca Göre Seçilen Niteliklerin Literatür Çalışması	96
5. SONUÇLAR VE ÖNERİLER	107
KAYNAKLAR	109
ÖZGEÇMİŞ	129
EKLER	130



ÇİZELGELER DİZİNİ

SAYFA

Çizelge 1.1. Kandaki D Vitamini Seviyeleri	5
Çizelge 3.1. Veri Setinin Dağılımı	16
Çizelge 3.2. Nitelikler ve Niteliklerin Referans Aralıkları (Halk Sağlığı Laboratuvarı)	17
Çizelge 3.3. Sınıflandırma Yöntemleri Karşılaştırılması	22
Çizelge 3.4. Hata Matrisi Gösterimi	42
Çizelge 4.1. Sınıflandırma Sonuçları	55
Çizelge 4.2. Sınıflandırma Sonuçlarına Göre Oluşan Performans Ölçütleri	56
Çizelge 4.3. Nitelik Seçimi Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları (Parantez içindeki değerler seçilen nitelik sayılarını göstermektedir.)	57
Çizelge 4.4. Nitelik Seçimi Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçlarına Göre Oluşan Performans Ölçütleri.....	59
Çizelge 4.5. Optimum Küme Sayısının Belirlenmesi	60
Çizelge 4.6. Kümeleme Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları (Parantez içindeki değerler seçilen nitelik sayılarını göstermektedir.)	62
Çizelge 4.7. Kümeleme Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçlarına Göre Oluşan Performans Ölçütleri.....	63
Çizelge 4.8. KMO and Bartlett's Test	64

Çizelge 4.9. Minimum Faktör Yükleme Değerleri	64
Çizelge 4.10. Faktör Grupları	66
Çizelge 4.11. Vaka İşleme Özeti	67
Çizelge 4.12. Sınıflandırma Tablosu 1	67
Çizelge 4.13. Model Özeti	67
Çizelge 4.14. Sınıflandırma Tablosu 2	68
Çizelge 4.15. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Sınıflandırma Sonuçları	70
Çizelge 4.16. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Sınıflandırma Sonuçlarına Göre Oluşan Performans Ölçütleri.....	70
Çizelge 4.17. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Nitelik Seçimi Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları (Parantez içindeki değerler seçilen nitelik sayılarını göstermektedir.)	72
Çizelge 4.18. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Nitelik Seçimi Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçlarına Göre Oluşan Performans Ölçütleri.....	74
Çizelge 4.19. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Kümeleme Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları (Parantez içindeki değerler seçilen nitelik sayılarını göstermektedir.)	76
Çizelge 4.20. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Kümeleme Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçlarına Göre Oluşan Performans Ölçütleri.....	77
Çizelge 4.21. KMO and Bartlett's Test	78
Çizelge 4.22. Sınıflandırma Tablosu 1	78
Çizelge 4.23. Model Özeti	78
Çizelge 4.24. Sınıflandırma Tablosu 2	79
Çizelge 4.25. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Sınıflandırma Sonuçları	80

Çizelge 4.26. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Sınıflandırma Sonuçlarına Göre Oluşan Performans Ölçütleri	81
Çizelge 4.27. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Nitelik Seçimi Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları	82
Çizelge 4.28. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Nitelik Seçimi Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçlarına Göre Oluşan Performans Ölçütleri	84
Çizelge 4.29. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Kümeleme Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları (Parantez içindeki değerler seçilen nitelik sayılarını göstermektedir.)	86
Çizelge 4.30. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Kümeleme Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçlarına Göre Oluşan Performans Ölçütleri	87
Çizelge 4.31. Çalışmada Kullanılan Yöntemler	90
Çizelge 4.32. Çalışmanın Literatür ile Kıyaslanması	92
Çizelge 4.33. En İyi Sınıflandırma Sonuçlarının Karşılaştırılması	93
Çizelge 4.34. En İyi Nitelik Seçimi Sonuçlarının Karşılaştırılması	95
Çizelge 4.35. Yazılım Validasyonu	102



ŞEKİLLER DİZİNİ

SAYFA

Şekil 3.1. Verinin toplanma aşamaları ve çalışmanın akış diyagramı	20
Şekil 3 .2. İki sınıflı bir problem için hiper-düzlemler	25
Şekil 3 .3. Optimum hiper-düzlem ve destek vektörler	25
Şekil 3 .4. Doğrusal olarak ayrılabilen veri setleri için hiper-düzlemin belirlenmesi	26
Şekil 3 .5. Doğrusal olarak ayrılamayan veri seti	27
Şekil 3 .6. Doğrusal olarak ayrılamayan veri setleri için hiper-düzlemin belirlenmesi	27
Şekil 3 .7. Kernel fonksiyonu ile verinin daha yüksek bir boyuta dönüştürülmesi	28
Şekil 3 .8. Dört boyutlu özellik uzayına sahip 3 sınıftan oluşan bir karar ağacı yapısı	32
Şekil 3.9. Tipik bir ileri beslemeli yapay sinir ağları	35
Şekil 3.10. K-Ortalama Algoritması ile kümeleme aşamaları	44
Şekil 3 .11. Optimum Küme Sayısının Belirlenmesi	45
Şekil 3 .12. Karar destek sistemi yazılımının akış diyagramı	51
Şekil 4.1. Optimum Küme Sayısının Belirlenmesi	61
Şekil 4 .2. Yamaç Grafiği	65
Şekil 4 .3. Yazılım Hiyerarşisi	99
Şekil 4 .4. Yazılım Ana Menüsü	100
Şekil 4.5. Veri Seti Giriş Ekranı	100
Şekil 4 .6. Diyabet Durumu Bilinmeyen Hastanın Tahlil Giriş Ekranı	101
Şekil 4 .7. Diyabet Durumu Bilinmeyen Hasta Sonucu	101
Şekil 4 .8. Diyabet Hastasının Tahlil Giriş Ekranı	102
Şekil 4 .9. Diyabet Hastası Sonucu	102
Şekil 4.10. Diyabet Hastalığı Olmayan Hastanın Tahlil Giriş Ekranı	103

SİMGELER VE KISALTMALAR

UVB	: Ultraviyole B
IU	: Uluslararası birim
25OHD	: 25-hidroksivitamin D
DVM	: Destek vektör makineleri
KNN	: K-en yakın komşuluk algoritması
KA	: Karar ağaçları
NB	: Naive bayes
YSA	: Yapay sinir ağları
RO	: Rastgele orman
LR	: Lojistik regresyon
FA	: Faktör analizi
PTH	: Parathormon
T1D	: Tip 1 diyabet
7-DHC	: 7-dehidrokolesterol
DBP	: Vitamin D bağlayıcı proteinin
BMI	: Vücut kitle endeksi
T2D	: Tip 2 diyabet
LDL	: Düşük yoğunluklu lipoprotein
HDL	: Yüksek yoğunluklu lipoprotein



1. GİRİŞ

Vitamin D eksikliği tahmini, bir hastanın vitamin D değerlerini dış ve iç verilere dayalı olarak bulmayı amaçlar. Vitamin D eksikliği birçok hastalığın habercisi olarak görülmektedir. Bu sebeple bu eksikliğin doğru tahmini hem hastalar hem de klinisyen için önemli bir yere sahiptir. Bu tahmin sayesinde laboratuvar testi yapılmaksızın hızlı sonuç alınabilmesi doktorun hastayı etkin yönlendirmesine ve hastanın gereken önlemi vaktinde alabilmesine olanak sağlamaktadır. Artan doğruluk için çok değişkenli girdiler tercih edildiğinden, tahmin yapmak için daha esnek modellere ihtiyaç vardır. Makine öğrenmesi yöntemleri bu anlamda çok kullanışlıdır. Kan tahlili sonuçları üzerinden kandaki glukoz değeri, COVID-19 test sonucu gibi birçok değerin tahmini araştırmacılar tarafından makine öğrenmesi yöntemleri yardımıyla çalışılmış ve farklı algoritmaların karşılaştırılması bu çalışmalar dahilinde yapılmıştır (Bozkurt, 2021; Kavakiotis ve ark., 2017). Ancak literatüre bakıldığında vitamin D tahmini üzerinde makine öğrenmesi yöntemleri ile yapılmış çalışmalara nadiren rastlanmaktadır. Tahmin çalışmaları sırasında en büyük zorluklardan biri hatasız ve gürültüsüz veri toplamaktır. Bu tez çalışmasıyla tüm parametrik ve parametrik olmayan algoritmalar için bir kıyaslama (benchmark) veri seti hazırlanmıştır. Makine öğrenmesi ile yapılan tahminler, laboratuvar ortamında yapılan vitamin D ölçüm sürecinden kıyaslanamayacak derecede hızlı sonuç vermektedir.

Bu tez çalışmasının bölümleri şu şekilde organize edilmiştir. Birinci bölümde; teze genel bir bakış açısı kazandırmak için temel bilgilere giriş yapılmış ve tezin amaç ve kapsamı hakkında bilgi verilmiştir. İkinci bölümde; literatürde bulunan rutin laboratuvar test sonuçları kullanılarak yapılan mevcut çalışmalar ile bu tezin literatüre yapacağı katkılardan bahsedilmiştir. Üçüncü bölümde; veri madenciliğinin tanımından başlanarak tez çalışmasında kullanılan veri seti tanımlanması, tez çalışmasında kullanılacak veri setinin oluşturulması için

kullanılan veri ön işlem adımları, veri setinin sınıflandırılması ve modelleri oluşturmak için kullanılan destek vektör makineleri, k-en yakın komşuluk algoritması, karar ağaçları, naive bayes, yapay sinir ağları ve rastgele orman algoritmaları ile ilgili bilgiler ile model değerlendirme ölçütleri hakkında bilgiler verilmiştir. Dördüncü bölümde; vitamin D veri seti analizinden başlayarak, veri setinin hazırlanması ve modellenmesi için yapılan deneysel çalışmalardan ve değerlendirme için ise makine öğrenmesi algoritmalarının başarılarından elde edilen bulgular açıklanmış; tez çalışmasında yapılan uygulamaların sonuçları, tez çalışmasında uygulanan adımların literatürle karşılaştırılması ve öneriler hakkında bilgi verilmiştir. Son aşama olarak vitamin D eksikliğinin tespiti için bir prototip yazılım önerilmiştir.

1.1. Makine Öğrenmesi

Yapay zekanın geniş bir alt alanı olan makine öğrenmesi, bilgisayarların öğrenmesini sağlayan algoritmalar ve teknikler tasarlar. Makine öğrenmesi araştırmasının temel odağı, hesaplamalı ve istatistiksel yöntemlerle veriden otomatik olarak bilgi çıkarmaktır. Dolayısıyla, makine öğrenmesi sadece veri madenciliği ve istatistikle değil aynı zamanda teorik bilgisayar bilimleriyle de yakından ilişkilidir. Makine öğrenmesi, bilginin özerk edinimi ve entegrasyonu yeteneğine sahip bir sistemi ifade eder. Deneyimden, analitik gözlem ve diğer araçlardan kendini sürekli iyileştiren ve böylece verimliliği ve etkinliği artıran bir sistemle sonuçlanır. Makine Öğrenmesi, bilginin akıllı sistemlerde biriktirilmesi, değiştirilmesi ve güncellenmesi için hesaplama yöntemleri ve bilgi üretmemize yardımcı olacak belirli öğrenme mekanizmalarının sağlanmasını amaçlamaktadır (Magoulas ve Prentza, 2001). Tıbbi uygulamalarda makine öğrenimi kullanılarak, erken teşhisin doğruluğu geliştirilebilir.

1.2. Vitamin D

Vitamin D esas olarak iskelet sağlığının korunmasındaki önemi ile bilinir. Ancak araştırmalar vitamin D'nin üremeden, yetişkin kronik hastalıklarına kadar çok çeşitli roller oynadığını kanıtlamaktadır. Doğal olarak sadece birkaç gıdada bulunmaktadır ve gıda takviyesi vitamin D durumunu yükseltmeyi sağlayabilmektedir. Vitamin D takviyesinin raşitizm riskini azaltmadaki olumlu etkisi yaklaşık 100 yıldır bilinmektedir. Vitamin D kalsiyum ve fosfat emilimi ve metabolizması ile ilişkilendirilmiştir.

Vitamin D insanlarda önemli olan iki formda bulunmaktadır: D2 vitamini (bitkilerden, mayalardan ve mantarlardan elde edilen ergokalsiferol) ve D3 vitamini (cildi güneş ışığına maruz bırakarak veya yağlı balık, morina karaciğeri yağı, yumurta gibi gıdaların alımından elde edilen kolekalsiferol). Vitamin D'nin her iki formu da diyetle sınırlı miktarlarda bulunmaktadır. Ultraviyole B (UVB) radyasyonuna (270-300 nm dalga boyları) maruz kalındığında deride endojen olarak sentezlenen vitamin D3 insanlarda birincil kaynaktır. UVB radyasyonu cilt epidermisine nüfuz eder ve 7-dehidrokolesterolü previtamin D3'e dönüştürür. Cilt melanin derecesindeki farklılıklar, UVB'nin neden olduğu vitamin D üretimini % 90'ın üzerinde değiştirebilir ve daha koyu tenli insanları daha yüksek bir eksiklik düzeyine sokabilir. Dermatologlar, kapalı alanlarda geçirilen zamanın, güneş kremi uygulamasının, coğrafi konumun ve mevsimsel değişikliklerin UVB radyasyonuna maruz kalmayı azaltabileceğini öne sürmektedir (Armas ve ark., 2004; Holick ve ark., 2008; Holick, 2006; Uriu-Adams, 2013).

Vitamin D eksikliği birçok hastalığın oluşmasına ortam hazırlamaktadır. Vitamin D'nin rolü, bağışıklık sistemini, hücre proliferasyonunu ve farklılaşmasını ve glukoz metabolizmasını modüle etme dahil olmak üzere genişlemiştir (Goldsmith, 2015; Bhalla ve ark., 1983; Provvedini ve ark., 1983; Bikle, 2009; Weyland ve ark., 2014). Güncel çalışmalar, vitamin D'nin hastalık patogenezi, şiddeti ve tedavisi ile potansiyel ilişkisini araştırmaktadır (Ross ve ark., 2011;

Wabke ve ark., 2014). Fransa'da, 65 yaş üstü zayıf bir yetişkin popülasyonunu karakterize etmeyi amaçlayan bir çalışma, hemen hemen herkesin klinik vitamin D eksikliğine sahip olduğunu ortaya koymuştur (Leslie ve Hankey, 2015). Çok sayıda kanıt, vitamin D'nin anormal glukoz metabolizması, glukoz intoleransı, değişen insülin sekresyonu ve diyabette rol oynadığını göstermiştir (Melamed ve ark., 2008; Palomer ve ark., 2008; Chiu ve ark., 2004; Pittas ve ark., 2010; Thorand ve ark., 2011). Gözlemsel çalışmalar, diyabet riskinin düşük vitamin D konsantrasyonları ile pozitif ilişkili olduğunu göstermiştir. Başka bir sistematik inceleme, vitamin D alımını günde 500 uluslararası birimden (IU) daha fazla artırarak bu kanıtı doğrulamış ve diyabet riskini 200 IU/güne kıyasla % 13 azalttığını göstermiştir. Bir başka çalışmada 25 ng/ml'den yüksek 25hidroksivitamin D (25OHD) statüsüne sahip kişilerin diyabet geliştirme riski, 14 ng/ml'den düşük 25OHD durumuna kıyasla % 43 daha düşük olduğunu göstermiştir (Pittas ve Dawson-Hughes, 2010; Sowell ve ark., 2015). Ayrıca vitamin D eksikliği ve yetersizliğinin yaygın kanserler, kardiyovasküler hastalıklar, metabolik sendrom, enfeksiyöz ve otoimmün hastalıkların dahil olduğu birçok kronik hastalıkla ilişki içinde olduğu bulunmuştur. Bu yaygın hastalıkların spektrumu endişe vericidir. Vitamin D eksikliği güneş ışığına yetersiz maruziyet, diyetle yetersiz alım, absorpsiyon problemleri nedeniyle ortaya çıkabilmektedir. İngiltere'de yakın zamanda yapılan bir çalışmada; kış ve bahar dönemlerinde erişkin popülasyonun % 50'sinden fazlasında vitamin D yetersizliği, % 16'sında da ciddi vitamin D eksikliği saptandığı bildirilmiştir. Ülkemizden Uçar ve ark. son yıllarda Ankara bölgesinde yaptıkları bir çalışmada; oldukça yüksek oranda (% 51,8) vitamin D eksikliği ve % 20,7 oranında vitamin D yetersizliği tespit edilmiştir (Uçar ve Ark., 2014).

Vitamin D seviyelerine göre hasta durumunun nasıl değerlendirileceği çizelge

1.1.'de gösterilmiştir.

Çizelge 1.1. Kandaki D Vitamini Seviyeleri (Fidan ve ark., 2014)

25-hidroksivitamin D Seviyesi	Hasta Durumu
>30 µg/ml	D vitamini yeterli
21-29 µg/ml	D vitamini yetersizliği
<20 µg/ml	D vitamini eksikliği
11-20 µg/ml	Hafif D vitamini eksikliği
5-10 µg/ml	Orta D vitamini eksikliği
<5 µg/ml	Ciddi D vitamini eksikliği

1.3. Diyabet

Diyabet, 387 milyon insanı etkileyen küresel bir halk sağlığı sorunudur ve 2035 yılına kadar dünya genelinde bu sayının 592 milyona çıkacağı tahmin edilmektedir. Dünyada her gün yaklaşık 700 kişiye diyabet teşhisi konulmaktadır ve bu her iki dakikada bir diyabet teşhisine denk gelmektedir (Guariguata, 2014). Diyabet tanısı almış yetişkinlerin miyokard enfarktüsü nedeniyle hastaneye yatışları 1,8 kat, beyin damar hastalıkları nedeniyle hastaneye yatışları 1,5 kat daha fazladır. Uluslararası Diyabet Vakfı'nın raporlarına göre, HIV/AIDS, tüberküloz ve sıtma hastalarının yıllık toplamından daha fazla insan, diyabet ve komplikasyonlardan ölmektedir. Diyabetli hastaların bakımı için tıbbi maliyetler 2,3 kat daha fazladır (CDC, 2020; International Diabetes Federation, 2015; Signorello ve ark., 2007; American Diabetes Association, 2010). Kötü kontrol edilen diyabet, hastaneye yatış, sakatlık ve erken ölümle sonuçlanan sağlık komplikasyonlarına neden olabilmektedir (Napoli ve ark., 2017). Son yıllarda yapılan araştırmalar diyabet riski ile ilişkilendirilebilecek faktörlerin olduğunu göstermektedir. Bu faktörlerden biri vitamin D'dir.

1.4. Çalışmanın Önemi ve Amacı

Bu çalışmada, hastaların kan testi sonuçları kullanılarak vitamin D eksikliğinin doğru teşhisi için sağlık personeline yardımcı olabilecek çeşitli makine öğrenmesi tekniklerinin analizi anlatılmaktadır. Literatür incelendiğinde çalışmanın başladığı yıllarda vitamin D ile ilgili çalışmaların azlığı dikkat çekmiş ancak son yıllarda literatürdeki çalışmaların arttığı gözlemlenmiştir. Çalışmanın başladığı 2016 yılından 2020 yılına kadar Aile Sağlığı Merkezlerinde vitamin D testi yapılabilirken, 2020 itibarıyla Aile Sağlığı Merkezlerinde vitamin D testi yapılma imkanı alınan kararlar kaldırılmıştır. Bu bilgi ışığında çalışmanın önemi daha da artmaktadır. Çünkü çalışma Aile Sağlığı Merkezlerinde şu an olmayan bir test imkanı sunmaktadır.

Çalışmanın amacı, hastaların yıllar içinde elde ettikleri kan testi sonuçlarını kullanarak vitamin D eksikliğini makine öğrenmesi yöntemleri ile tahmin etmektir. Çalışma, yapılan testleri azaltması açısından değerli bir çalışmadır. Yapılan literatür araştırmasında benzer çalışmalardan daha iyi performans göstermesi nedeniyle bilimsel anlamda kıymetli bir çalışma olmuştur. Ayrıca, hazırlanan özgün veri seti, tahmin çalışmaları sırasında en büyük zorluklardan biri olan hatasız ve gürültüsüz veri toplama sorununa da katkı sağlamaktadır. Veri setinin konu üzerinde çalışacak olan araştırmacılara açık hale getirilmesi literatüre büyük katkılar sağlayacaktır.

Bunların yanında hasta sayıları ve her bir hasta için harcanan zaman göz önünde bulundurulduğunda hızlı teşhisin önemi ortaya çıkmaktadır. Hasta yoğunluğunun olduğu dönemlerde doktorlar yorum sürelerinin kısıtlanmasından ötürü hastalardan bazı tahlilleri istemeyi gözden kaçırabilmektedir. Ayrıca hastaların da sıkıntılarını yeterince anlatamamalarından kaynaklı olarak doktorlar doğru tahlillere yönelmede sorun yaşayabilmektedir. Önerilen sistemin hastanın tahlil sonuçlarını ele alarak kısa sürede daha doğru teşhisler ortaya koyup hem doktorların iş yükünü hafifletmesi hem de hasta sağlığı açısından daha detaylı incelemelerde bulunması beklenmektedir (Matheny ve Machado, 2014).

1.5. Çalışmada Kullanılan Materyal

Çalışmada kullanılan veri seti gerçek kan testi sonuçlarından oluşmaktadır. Detayları bölüm 3.1’de verilen veri seti, çalışma kapsamında özgün olarak oluşturulmuş olup üç farklı şekilde analizi sağlanmıştır. Çalışmanın birinci adımında veri seti bütünüyle ele alınmış ve analizler bu şekilde gerçekleştirilmiştir. İkinci adımda ise literatürden yola çıkarak, veri seti içerisinde diyabet hastalarına ait olan kan tahlil sonuçları seçilmiş ve diyabet hastaları özelinde analizler gerçekleştirilmiştir. Son adımda da ikinci adımda yapılan analizlerin kıyaslamasını yapmak üzere, diyabet hastası olmayan kişilerin kan testi sonuçları alınarak diyabet hastası olmayan kişiler özelinde analizler gerçekleştirilmiştir.

1.6. Çalışmada Kullanılan Yöntemler

Yapılan çalışma kan testi sonuçları üzerinden vitamin D değerini doğru tahminlemeye çalışırken aynı zamanda tahmin için gerekli parametreleri de azaltmayı hedeflemektedir. Bu hedef doğrultusunda analiz kısmında farklı sınıflandırma, kümeleme ve nitelik seçimi yöntemleri kullanılarak yöntemlerin bireysel ve hibrit karşılaştırmaları yapılmıştır. Empirik olarak çalışmanın tahminleme kısmında kullanılan sınıflandırma yöntemlerinin destek vektör makineleri (DVM), k-en yakın komşuluk algoritması (KNN), karar ağaçları (KA), naive bayes (NB), yapay sinir ağları (YSA), rastgele orman (RO) ve lojistik regresyon (LR) olmasına karar verilmiştir. 2020 yılı itibarıyla Aile Sağlığı Merkezlerinde vitamin D testlerinin yapılmamasıyla beraber etiketli veriye ulaşımın zorluk yaşanabileceği öngörülerek, çalışmada farklı bir adım daha belirlenmiştir. Çalışmanın kümeleme adımı, etiketsiz veri olması halinde tahminleme performansını ölçmek için tasarlanmıştır. Hali hazırda sınıf etiketi bulunan veriye etiketsiz veri seti gibi yaklaşılarak kümeleme ile sınıf etiketleri verilmiş ve sınıflama ve kümeleme performanslarının karşılaştırılması yapılmıştır. Burada amaç sonuçların

birbirine yakın çıkması halinde etiketsiz veri üzerinde de vitamin D eksikliğinin yüksek doğrulukla tahminlenebileceğini göstermektedir. Kümeleme için k-ortalama algoritması; nitelik seçimi için de relief-f, info gain, gain ratio, faktör analizi (FA) ve lojistik regresyon (LR) yöntemlerinin kullanılması uygun görülmüştür. Sınıflandırma yöntemleri karşılaştırıldığında destek vektör makineleri, yapay sinir ağları ve rastgele orman yöntemlerinin ortalama tahmin doğruluğunda yüksek performans verdiği ancak yapay sinir ağları ve rastgele orman yöntemlerinin eğitim hızında düşük performans verirken destek vektör makineleri yönteminin eğitim hızında yüksek performans verdiği gözlenmiştir. Ayrıca destek vektör makinelerinin anlatımının kolay ve tahmin hızının yüksek olduğu gözlenmiştir. Dolayısıyla performansı en yüksek algoritmanın destek vektör makineleri olduğu görülmektedir.

1.7. Çalışmanın Aşamaları

3 farklı veri setinin kullanıldığı çalışmada sırasıyla sınıflandırma (destek vektör makineleri, k-en yakın komşuluk algoritması, karar ağaçları, naive bayes, yapay sinir ağları ve rastgele orman), kümeleme (k-ortalama algoritması), faktör analizi ve lojistik regresyon işlemleri tamamlanmıştır.

Sınıflandırma, kümeleme, faktör analizi ve lojistik regresyon işlemleri tamamlandıktan sonra veri setlerinde nitelik seçimi sonuçları sürdürülmüştür. Sonrasında veri setleri nitelik seçimi ve kümeleme sonuçlarına göre güncellenmiştir. Veri setlerinin güncellenmesinden sonra sınıflandırma işlemleri tekrarlanmıştır. Her aşamada sonuçların kaydedilmesinden sonra literatür araştırmasında elde edilen bulgular kıyaslanmıştır.



2. ÖNCEKİ ÇALIŞMALAR

Tezin bu bölümünde önceki çalışmalar kendi aralarında sınıflandırılarak aktarılmıştır.

2.1. Vitamin D Eksikliğinin Literatür Araştırması

Tıp literatürü incelendiğinde vitamin D ile ilgili birçok çalışma dikkat çekmektedir. Bunların bazılarında aşağıda bahsedilmiştir.

Carnevale ve ark. buldukları cinsiyet farklılıklarının erkekler tarafından dışarıda geçirilen daha fazla zamana bağlanabildiğinden; cinsiyet farkının, vitamin D bağlayıcı protein seviyelerindeki androjene bağlı farklılıklara veya deri tarafından öncü üretimine ya da karaciğer tarafından 25-hidroksilasyonuna bağlanabildiğinden bahsetmiştir. 25-hidroksi D vitamini (25OHD) ve parathormon (PTH) seviyeleri arasındaki ters korelasyon nedeniyle, kış aylarında erkeklerde bulunan daha yüksek serum 25OHD seviyelerinin, erkeklerin kadın deneklere göre daha düşük PTH artışını açıklayabildiğinden bahsetmiştir (Carnevale ve ark., 2001).

Bolland ve ark. toplam 25OHD seviyeleri ile vücut ağırlığı, yağ kütlesi, yaş ve toplam 25OHD düzeylerindeki cinsiyet farklılıkları arasında ters ilişkiler olduğundan; toplam 25OHD düzeylerindeki yaş ve cinsiyet farklılıkları, yaşlanma ile ciltte vitamin D sentezleme yeteneğinde azalma ve cinsiyetler arasındaki davranışsal ve kültürel farklılıklar nedeniyle güneşe maruz kalma farklılıklarından kaynaklandığından bahsetmiştir. Toplam 25OHD'nin diyabet, iskemik kalp hastalığı, konjestif kalp yetmezliği, hipertansiyon ve çeşitli kanserler de dahil olmak üzere birçok durumda seviyesinin azaldığından; vitamin D eksikliğinin artmış tip 1 diyabet (T1D), multipl skleroz ve romatoid artrit riski ile ilişkili olabileceğinden toplam 25OHD düzeyinde cinsiyet farklılıkları bildiren çalışmaların çoğunda erkeklerde kadınlardan daha yüksek düzeyler rapor edildiğinden bahsetmiştir (Bolland ve ark., 2018).

Carbone ve ark. çalışmalarında kolesterolün cildin bariyer fonksiyonunun önemli bir bileşeni olduğundan ve stres zamanlarında ek kolesterol gereksinimi arttığından bahsetmiştir. Bu gereksinim kolesterolün endojen sentezi veya plazmadan karşılanabildiğinden; endojenin kolesterol sentezinin son öncüsü, aynı zamanda vitamin D üretmek için ultraviyole B (UVB) ile reaksiyona giren substrat olan 7-dehidrokolesterol (7-DHC) olduğundan bahsetmiştir. Vitamin D sentezinin 7-DHC konsantrasyonuna bağlı olduğundan; hücrel kolesterol konsantrasyonundaki bir azalmanın 7-DHC redüktazını (7-DHC'yi kolesterole dönüştüren enzim) arttırdığına dair haberci ribonükleik asit (RNA) ve aktivitesinin var olduğundan; plazma lipoproteinlerinden bir kolesterol akışının 7-DHC konsantrasyonlarını artırabildiğinden ve bunun da vitamin D sentezini artırabildiğinden bahsetmiştir (Carbone ve ark., 2008).

Hagenau ve ark. mevcut ekolojik meta-regresyon analiziyle, vitamin D durumunun küresel olarak yaşa, Cinsiyete ve cilt rengine bağlı olduğundan bahsetmiştir (Hagenau ve ark., 2009).

Bogh ve ark. kolesterol ve vitamin D'nin 7-DHC yolunu paylaştığından; UVB'ye maruz kaldıktan sonra vitamin D üretiminde cilt pigmentasyonu, başlangıç 25OHD seviyeleri ve toplam kolesterol seviyesinin önemli olduğundan bahsetmiştir. Vitamin D sentezinin bağırsak epitelinde, kolesterolün gıda veya safradan pro-vitamin D3'e 7-DHC oksidasyonu ile başladığından, daha sonra cilde, özellikle epidermise önlendiğinden; burada ön vitaminlere izomerleştirildiğinden bahsetmiştir. Vitamin D ile toplam kolesterol seviyesi arasında anlamlı bir pozitif ilişki olduğundan ve toplam kolesterol ile vitamin D üretimi arasında bir ilişki olduğundan bahsetmiştir (Bogh ve ark., 2009).

Lagunova ve ark. vitamin D eksikliğinin yaygınlığının, vitamin D eksikliği prevalansının tüm yaş ve cinsiyet gruplarında mevsimsel değişikliklerle benzer eğilimlere sahip olduğundan; genç erkekler ve kadınlar ve obez kişilerin, sadece kış aylarında değil, aynı zamanda yaz aylarında da düşük vitamin D durumuna daha

yatkın olduklarından bahsetmiştir. Vitamin D bağlayıcı proteinin (DBP) vitamin D durumundaki cinsiyet farklılıklarında da rol oynayabileceğinden; DBP ve adipozite arasındaki korelasyonun erkeklerde negatif, kadınlarda pozitif olduğundan ve DBP ile vücut ağırlığı, vücut kitle endeksi (BMI) ve yaş arasında anlamlı bir ilişki bulunmadığından bahsetmiştir (Lagunova ve ark., 2009).

Reboul ve ark. vitamin D'nin diğer yağsız (mikro) besinler gibi pasif bir süreçle emildiği varsayıldığından ve bununla birlikte, vitamin D yapısına çok yakın kimyasal bir yapı sergileyen kolesterolün, basit bir difüzyon yoluyla emilmediğinden bahsetmiştir (Reboul ve ark., 2011).

Batieha ve ark. düşük vitamin D durumunun yaygınlık oranının yaşa ve cinsiyete göre elde edildiğinden bahsetmiştir (Batieha ve ark., 2011).

Turhan ve Altınova son yıllarda yapılan epidemiyolojik çalışmalarda vitamin D eksikliği ile tip 1 diyabet (T1D) ve tip 2 diyabet (T2D) gelişimi arasında ilişki gösterildiğinden ve deneysel çalışmalarda vitamin D'nin diyabet patogenezinde rolü olabileceği görüşünü desteklediğinden bahsetmiştir. Ayrıca vitamin D replasmanının uzun dönemde diyabet gelişme riskini azaltabileceğini ve vitamin D'nin koruyucu etkilerini immün sistem ve kalsiyum metabolizması üzerinden gerçekleştirdiğinden bahsetmiştir (Turhan ve Altınova, 2012).

Ponda ve ark. düşük yoğunluklu lipoprotein (LDL) kolesterol seviyelerinin tipik olarak doğrudan ölçülemediğinden; bunun yerine Friedewald denkleminden hesaplandığından ve daha yüksek 25OHD seviyeleri ile düşük toplam kolesterol, düşük seviye LDL kolesterol, yüksek seviye yüksek yoğunluklu lipoprotein (HDL) kolesterol ve düşük trigliseridler arasında güçlü kesitsel ilişkiler kurulduğundan bahsetmiştir (Ponda ve ark., 2012).

Verdoia ve ark. vitamin D durumu için cinsiyet farklılığının özellikle menopoza sonrası kadınlarda daha yüksek eksiklik oranının, kemik kırıkları ve osteoporoz riskini arttırdığından ve cinsiyetin, vitamin D durumunu önemli ölçüde etkilediğinden bahsetmiştir (Verdoia ve ark., 2015).

Karagöl ve Atak ekolojik kanıtlarla, vitamin D yetersizliğinin belirgin olduğu az güneş alan bölgelerde, diyabet sıklığının arttığını göstermiştir. Vitamin D son zamanlarda T2D etiopatogenezi ve önlenmesinde giderek artan bir şekilde gündeme geldiğinden; yetersizliğinin T2D riskini arttırdığı belirtmiştir. Ek olarak 25OHD düzeyinin artırılması, T2D'nin önlenmesi ve tedavisinde umut vadeden bir yaklaşım olabileceğinden bahsetmiştir (Karagöl ve Atak, 2016).

2.2. Tıpta Makine Öğrenmesi Çalışmalarının Literatür Araştırması

Literatür araştırması sırasında vitamin D ile ilgili veri setinde makine öğrenmesi tekniklerinin kullanıldığı çalışmaların sayısının azlığı gözlemlenmiştir. Bu nedenle kanla ilgili veri kümeleri üzerinde makine öğrenmesi tekniklerini kullanan diğer çalışmalar da sunulmuştur. Çalışma, vitamin D veri setini kullanarak makine öğrenmesi yaklaşımını kullanarak bu boşluğu doldurmayı amaçlamaktadır.

Palaniappan ve Awang'ın (2008) çalışmasında naive bayes'e bağlı kalp hastalığı öngörmeyi önermiş ve % 86,120 doğruluk elde etmiştir (Palaniappan ve Awang, 2008). Konda ve ark. (2010) naive bayes'i kalp krizi % 84,050 doğrulukla tahmin etmek için kullanmıştır (Konda ve ark., 2010). Methaila ve ark. (2014), karar ağacı'nın kalp hastalığı tahmini için yararlı bir yöntem olduğunu vurgulamış ve çalışma doğruluk sağlamıştır (Methaila ve ark., 2014). Komi ve ark. (2014) diyabeti % 89,000 doğrulukla tahmin etmek için yapay sinir ağları kullanmıştır (Komi ve ark. 2014). Saichanma ve ark. (2014) kırmızı kan hücresi morfolojisini % 94,300 doğrulukla tahmin etmek için karar ağacı kullanmıştır (Saichanma ve ark., 2014). Goodnough ve Neil (2014) kan transfüzyonu örneklenmesi ve elektronik tıbbi kayıtlar aracılığıyla klinik karar desteği uyguladıkları çalışmalarında yıllık transfüzyonları % 24 oranında azaltmışlardır (Goodnough ve Neil, 2014). Portela ve ark. (2015), kan basıncını % 71,520 doğrulukla tahmin etmek için karar ağacı kullanmıştır (Portela ve ark., 2015). GeethaRamani ve Balasubramanian (2016) retinal kan damarı segmentasyonunu % 95,360 doğrulukla tahmin etmek için kök

kılavuzlu karar ağacı kullanmıştır (GeethaRamani ve Balasubramanian, 2016). ElHalees ve Shurrah (2017) kan tümörü tahmini için bir derin öğrenme yöntemi önermiştir. Bu çalışmanın doğruluğu % 79,450 olarak hesaplanmıştır (El-Halees ve Shurrah, 2017). Celik ve ark. (2018) extreme learning machine yöntemini uygulamak için meme kanseri tanısını % 83,870 doğrulukla araştırmıştır (Celik ve ark., 2018). Thomas ve Princy (2018) kalp hastalığı tahminini % 80,600 doğrulukla tahmin etmek için karar ağacı kullanmıştır (Thomas ve Princy, 2018). Lippeveld ve ark. (2019) beyaz kan hücrelerinin sınıflandırılması için gradient boost'u kullanmış ve genel doğruluk % 87,100 olarak bulunmuştur (Lippeveld ve ark., 2019). Agarwal ve ark. (2019) kan transfüzyon sistemi için % 93,200 doğrulukla bir rastgele orman yöntemi önermiştir (Agarwal ve ark., 2019). Akben (2019) kan, hormon ve obezite değer aralıklarının belirlenmesi için karar ağacı yaklaşımı kullanarak % 90,500 doğruluk elde etmiştir (Akben, 2019). Gonoodi ve ark. (2019) vitamin D eksikliğinin risk faktörlerinin değerlendirildiği çalışmalarında karar ağacı kullanmış ve % 77,800 doğruluk oranını yakalamışlardır (Gonoodi ve ark., 2019). Marcus ve ark. (2020) kan glikoz düzeyi tahmini için kernel ridge regression tabanlı bir yöntem önermiştir ve çalışmasının doğruluğu % 93,400 bulmuştur (Marcus ve ark., 2020). Brinati ve ark. (2020) COVID-19 enfeksiyonunun tespiti için % 89,100 doğrulukla rastgele orman geliştirmiştir (Brinati ve ark., 2020). Wu ve ark. (2020) COVID-19 enfeksiyonunun tanımlanması için % 97,950 doğruluk sağlayan bir rastgele orman metodolojisi geliştirmiştir (Wu ve ark., 2020a). Wu ve ark. (2020) sistolik kan basıncını % 95,500 doğrulukla tahmin etmek için yapay sinir ağları kullanmıştır (Wu ve ark., 2020b). Yao ve ark. (2020) COVID-19 için önem tespitini % 81,480 doğrulukla tahmin etmek için destek vektör makineleri kullanmıştır (Yao ve ark., 2020). Sambasivam ve ark. (2020) vitamin D eksikliğinin öneminin performans analizini yaptıkları çalışmalarında rastgele orman kullanmış ve % 96,000 doğruluk oranını yakalamışlardır (Sambasivam ve ark., 2020). Amiri ve ark. (2021) vitamin D'nin takviyeye cevabını etkileyen faktörlerin belirlendiği çalışmalarında rastgele

orman kullanmış ve % 93,000 doğruluk oranını yakalamışlardır (Amiri ve ark., 2021). Padmajaa ve ark. (2021) vitamin D eksikliğinin önemini tahmin edildiği çalışmalarında extra trees classifier kullanmış ve % 73,300 doğruluk oranını yakalamışlardır (Padmajaa ve ark., 2021). Carretero ve ark. (2021) vitamin D eksikliğinin tahmin ettikleri çalışmalarında destek vektör makineleri kullanmış ve % 73,000 doğruluk oranını yakalamışlardır (Carretero ve ark., 2021). Karthikeyan ve ark. (2021) COVID-19 ölüm oranının erken teşhis edilmesini amaçladıkları çalışmalarında tespitlerini 16 gün öncesinden saptamış ve % 96 doğruluk oranını yakalamışlardır (Karthikeyan ve ark., 2021).

Ek 1’de tıpta makine öğrenmesi çalışmalarının literatür araştırması çizelgesinde yukarıda sözü edilen çalışmalar listelenmiştir.

3. MATERYAL VE METOT

Çalışmada kullanılan materyal ve metodlar bu bölümde detaylandırılmaktadır.

3.1. Materyal

Çalışmada kullanılan veri, Şubat 2016 - Nisan 2019 tarihleri arasında Türkiye'nin güneyindeki iki tıp merkezinden toplanan kan tahlil kayıtlarından oluşmaktadır.

Çalışmanın birinci kısmında veride diyabet hastalarının kan tahlillerinin yanı sıra, diyabet hastalığı olmayan hastaların kan tahlillerini de içermektedir. Veri 684 hastanın kan tahlillerinden oluşmaktadır. Bunlardan 406 tanesi diyabet hastasıyken 278 tanesi diyabet hastası değildir. Bunların 407 tanesi kadın; 277 tanesi erkek hastaların kan tahlillerinden oluşmaktadır. Yaş aralığı 2 ile 85 arasında değişmektedir. 447 tane kan tahlilinde vitamin D değeri normal değerinin altında; 237 tane kan tahlilinde vitamin D değeri normal değerindedir.

Çalışmanın ikinci kısmındaki veri diyabet hastalarının kan tahlillerini içermektedir. Veri 406 hastanın kan tahlillerinden oluşmaktadır. Bunların 291 tanesi kadın; 115 tanesi erkek hastaların kan tahlillerinden oluşmaktadır. Yaş aralığı 15 ile 85 arasında değişmektedir. 298 tane kan tahlilinde vitamin D değeri normal değerinin altında; 108 tane kan tahlilinde vitamin D değeri normal değerindedir.

Çalışmanın üçüncü kısmında veri diyabet hastalığı olmayan hastaların kan tahlillerini içermektedir. Veri 278 hastanın kan tahlillerinden oluşmaktadır. Bunların 116 tanesi kadın; 162 tanesi erkek hastaların kan tahlillerinden oluşmaktadır. Yaş aralığı 2 ile 77 arasında değişmektedir. 149 tane kan tahlilinde vitamin D değeri normal değerinin altında; 129 tane kan tahlilinde vitamin D değeri normal değerindedir.

Çizelge 3.1'de veri setinin cinsiyet, vitamin D seviyesi ve yaşa göre dağılımı gösterilmiştir.

Çizelge 3.1. Veri Setinin Dağılımı

	Verinin Özelliği	Toplam	Kadın	Erkek	Vitamin D Seviyesi Düşük	Vitamin D Seviyesi Normal	Yaş Aralığı
Veri Seti 1	Tüm Veri	684	407	277	447	237	2-85
Veri Seti 2	Diyabet Hastalarının Veri Seti	406	291	115	298	108	15-85
Veri Seti 3	Diyabet Hastalığı Olmayan Hastaların Veri Seti	278	116	162	149	129	2-77

Çalışmada kullanılan veride bulunan nitelikler ve niteliklerin referans aralıkları çizelge 3.2’de listelenmiştir.

Çizelge 3.2. Nitelikler ve Niteliklerin Referans Aralıkları
(Halk Sağlığı Laboratuvarı)

3. MATERYAL VE METOT

Uğur Engin EŞSİZ

NİTELİKLER	REFERANS ARALIĞI		BİRİM	NİTELİKLER	REFERANS ARALIĞI		BİRİM
Diyabet	1	0		C-Reaktif Protein (CRP)	0	0,8	mg/dL
Vitamin D	1	0		Prokalsitonin (PCT)	0,12	0,36	%
Cinsiyet	Kadın	Erkek		Kırmızı Kan Hücresi (RBC)	3,8	5,8	$\times 10^6/\mu\text{L}$
Yaş	2	85		Hemoglobin (HGB)	11,5	16,5	g/dL
Glukoz	70	110	mg/dL	Hematokrit (HCT)	37	47	%
Glycated Hemoglobin (HbA1c)	4	6,2	%	Ortalama Korpuskuler Hacim (MCV)	78	100	fL
Kreatinin	0,84	1,25	mg/dL	Ortalama Korpuskuler Hemoglobin (MCH)	27	32	pg
Glycated Hemoglobin International Federation of Clinical Chemistry Unit (HbA1c (IFCC))	20	42	mmol/mol	Ortalama Korpuskuler Hemoglobin Konsantrasyonu (MCHC)	30	35	g/dL
Gama-glutamil Transferaz (GGT)	0	55	U/L	Kırmızı Kan Hücresi Dağılım Genişliği (RDW-CV)	11,5	14,5	%
Ürik Asit	3,5	7,2	mg/dL	Trombosit (PLT)	150	400	$\times 10^3/\mu\text{L}$
Alanin Transaminaz (ALT)	7	40	U/L	Ortalama Trombosit Hacmi (MPV)	8	12	fL
Aspartat Aminotransferaz (AST)	9	45	U/L	Beyaz Kan Hücresi (WBC)	4	11	$\times 10^3/\mu\text{L}$
Alkalın Fosfataz (ALP)	30	120	U/L	Lenfositler (LY)	1,5	4	$\times 10^3/\mu\text{L}$
Laktat Dehidrojenaz (LDH)	120	250	U/L	Monositler (MO)	0	0,8	$\times 10^3/\mu\text{L}$
Üre	17	43	mg/dL	Nötrofiller (NE)	2	7,5	$\times 10^3/\mu\text{L}$
Albümin	3,5	5,2	g/dL	Eozinofil (EO)	0	0,4	$\times 10^3/\mu\text{L}$
Bilirübin	0	0,5	mg/dL	Ba sophils (BA)	0	0,1	$\times 10^3/\mu\text{L}$
Sodyum	136	146	mEq/L	MO%	0	10	%
Potasyum	3,5	5,1	mEq/L	NE%	40	75	%
Total Kolesterol	120	200	mg/dL	EO%	0	6	%
Magnezyum	1,3	2,7	mg/dL	BA%	0	2	%
Düşük Yoğunluklu Lipoprotein (LDL)	0	130	mg/dL	Yüksek Yoğunluklu Lipoprotein (HDL)	40	60	mg/dL
Trigliserid	40	160	mg/dL	LY%	20	50	%
Demir	40	160	$\mu\text{g/dL}$	Sedimentasyon	0	20	mm
Demir Bağlama	155	355	$\mu\text{g/dL}$	İnsülin	2,6	37,6	$\mu\text{IU/mL}$
Kalsiyum	8,8	10,6	mg/dL	Triiodotironin (T3)	1,71	3,71	0,1
Fosfor	2,3	4,7	mg/dL	Tiroksin (T4)	0,7	1,48	ng/dL

Trombosit Dağılım Genişliği (PDW)	25	65	%	Tiroid Uyarıcı Hormon (TSH)	0,35	4,94	uIU/mL
Kreatin Kinaz (CK)	30	294	U/L	Ferritin	10	291	ng/mL
Antistreptolizin O (ASO)	0	200	IU/mL	Vitamin B12 (VB12)	187	883	pg/mL

3.2. Metot

Verinin toplanma aşamaları ve çalışmanın akış diyagramı şekil 3.1’de gösterilmiştir.

Çalışmada kullanılan veri yaklaşık 5000 hastanın yıllar içinde arşivlenmiş kan tahlillerini içermektedir. Veride yaklaşık 10000 farklı tahlil sonucu bulunmaktadır.

Çalışmanın başlangıcındaki bu ham verinin işlenebilmesi için içerisinde vitamin D tahlili olanlar manuel olarak ayıklanmıştır. Sonrasında tahliller Excel ortamına yine manuel olarak girilmiştir.

Sonraki aşamada literatür çalışmasından yola çıkarak verimiz diyabet hastalarının ve diyabet hastalığı olmayan hastalarının kan tahlillerini içerecek şekilde bölünmüştür. Dolayısıyla işlemleri 3 farklı veri setinde deneme şansı oluşmuştur.

Veri 3 farklı veri setine dönüştürüldükten sonra öncelikle sınıflandırma (destek vektör makineleri, k-en yakın komşuluk algoritması, karar ağaçları, naive bayes, yapay sinir ağları ve rastgele orman) işlemleri tamamlanmıştır. Sınıflandırma işlemlerinde model, eğitim kümesinden hangi veriyi hangi sınıfa atayacağını öğrenmiş ve test verilerini doğru sınıflara atamaya çalışmıştır.

Sınıflandırma işlemleri tamamlandıktan sonra veride nitelik seçimi ve işlemleri sürdürülmüş ve veri bu sonuçlara göre güncellenmiştir. Nitelik seçiminde amaç, veri içerisinde vitamin D eksikliğini etkileyen en faydalı nitelikleri bulmak ve önem sırasına göre listelemektir.

Nitelik seçimi sonuçlarına göre veri güncellenmesinden sonra sınıflandırma işlemleri daraltılmış veride tekrarlanmıştır. Bu işlemde amaç verideki nitelik sayısını azaltmanın yanında modelin başarı oranını da arttırmaktır.

Bu işlemlerden sonra, veride eğer etiket bilgisi olmasaydı nasıl bir yol izlenip sonuç alınabilirdi sorusuna cevap bulmak ve vitamin D eksikliğini etkileyen nitelikleri bulmak amacıyla kümeleme (k-ortalama algoritması) işlemleri sürdürülmüştür. Kümelemede amaç, etiketleri kesin olarak bilinmeyen nitelikleri benzer kümelere ayırmaktır.

Kümeleme sonuçlarına göre veri güncellenmesinden sonra sınıflandırma işlemleri daraltılmış veride tekrarlanmıştır. Bu işlemde amaç, nitelik seçimi sonuçlarına göre güncellenen verideki gibi nitelik sayısını azaltmanın yanında modelin başarı oranını da arttırmaktır.

Bir sonraki aşamada, nitelikler arasında korelasyonlu olanları belirlemek ve nitelik sayısını azaltmak amacıyla faktör analizi işlemi sürdürülmüştür.

Son analiz aşamasında, en az değişkeni kullanarak en iyi uyuma sahip olacak şekilde bağımlı ve bağımsız değişkenler arasındaki ilişkiyi tanımlayabilmek amacıyla lojistik regresyon işlemleri sürdürülmüştür.

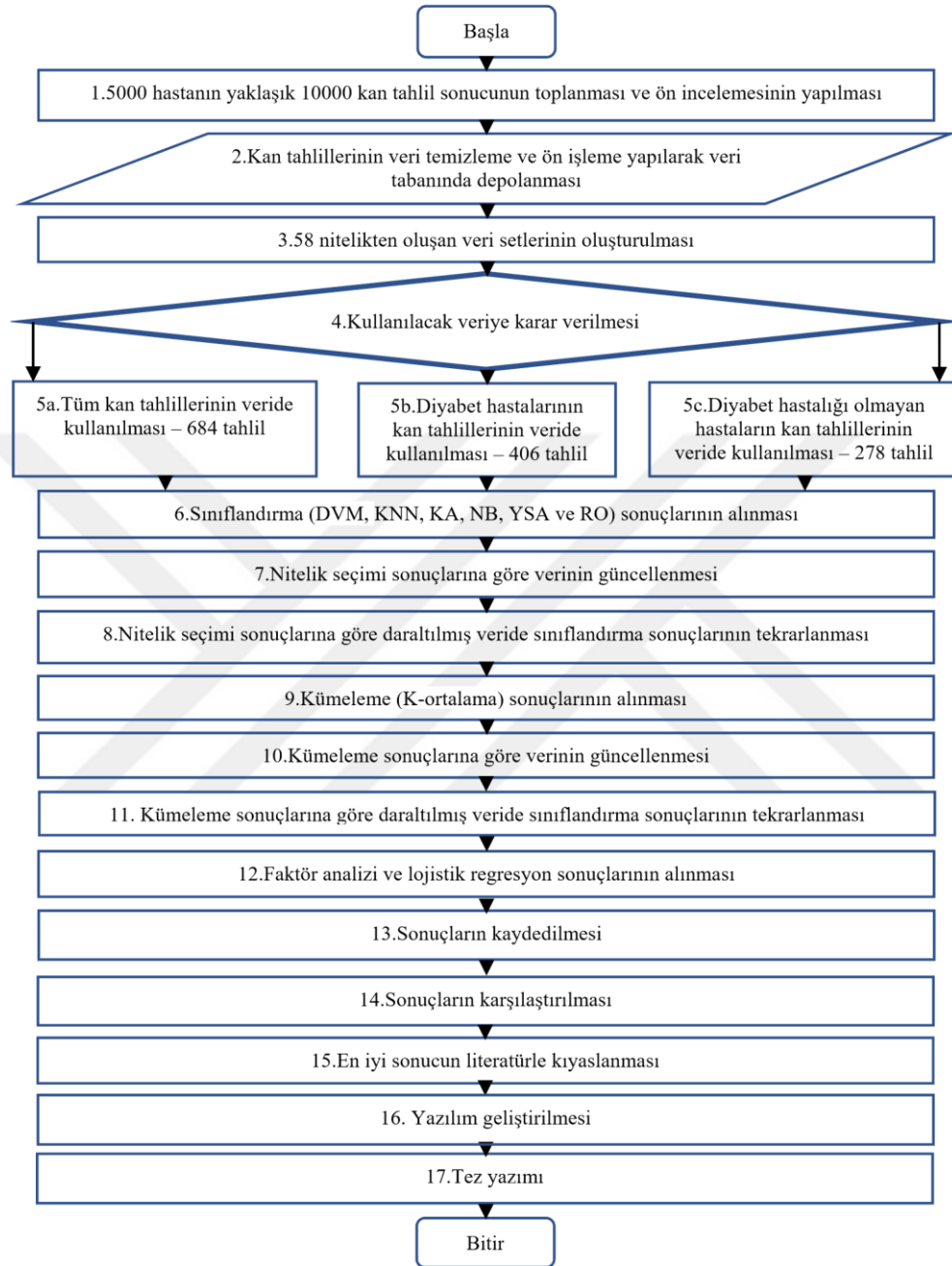
Tüm bu işlemler sırasında sonuçlar kaydedilmiştir.

Sonuçların kaydedilmesinden sonra sonuçlar ve literatür araştırmasında elde edilen bulgular kıyaslanmıştır.

Sonuçların literatürle kıyaslanmasının ardından yazılım geliştirilmesi aşamasına geçilmiştir.

Son olarak tez yazım süreci ile tez tamamlanmıştır.





Şekil 3.1. Verinin toplanma aşamaları ve çalışmanın akış diyagramı

3.2.1 Makine Öğrenmesi

Makine öğrenmesi, bilgisayar yazılımlarıyla mevcut verilerden elde edilen deneyimlerin, gelecekteki olayları tahmin etmesine ve modelleme yapmasına imkân veren bir yapay zekâ alanıdır. Makine öğrenmesi yöntemlerinin klinik analizde kullanımı gün geçtikçe artmaktadır. Hastalardan alınan bilgilerin değerlendirilmesi ve uzmanlar tarafından alınan kararlar tanı için önemli faktörlerdir. Tahmin için uzman sistemler ve makine öğrenmesi yöntemlerinin kullanılması uzmanlara yardımcı olur ve olası hataları en aza indirir (Amiri ve Armano, 2013). Çalışmanın deneysel tasarımında, tahmin modelini oluşturmak için destek vektör makineleri, k-en yakın komşuluk algoritması, karar ağaçları, naive bayes, yapay sinir ağları, rastgele orman, faktör analizi ve lojistik regresyon dahil olmak üzere makine öğrenme teknikleri kullanılmıştır. Bu tahmin modelleri, farklı sınıflandırıcılar oluşturmak için ayrı ayrı kullanılmış ve karşılaştırma için bu tekniklerin tahmin sonuçları değerlendirilmiştir. Çalışmada, her sınıflandırıcının performansının incelendiği deneyler için 10 kat çapraz doğrulama benimsenmiş ve bu model sonuçlarının karşılaştırılmasında doğruluk oranı kullanılmıştır. Performans sonuçlarına göre ortalamalar hesaplanmış ve sınıflandırıcıların performansları listelenmiştir. Çalışmada, makine öğrenmesi tekniklerini kullanarak vitamin D eksikliğini sınıflandırmak için bir model oluşturulmaya çalışılmıştır.

Çalışmada kullanılan makine öğrenmesi yöntemleri sınıflandırma ve kümeleme olarak ikiye ayrılmaktadır.

3.2.1.1. Sınıflandırma

Sınıflandırma benzer özellikteki nesnelerin önceden belirlenmiş alt gruplara atanması işlemidir. Temel amacı veriyi sadeleştirerek kullanıcıya daha anlaşılır bilgiler sunabilmektir. Çalışmada kullanılan sınıflandırma yöntemleri şunlardır:

- Destek Vektör Makineleri,
- K-En Yakın Komşuluk,
- Karar Ağaçları,
- Naive Bayes, □ Yapay Sinir Ağları,
- Rastgele Orman.

Çalışmada kullanılan sınıflandırma yöntemleri karşılaştırılması Çizelge 3.3'te gösterilmiştir. Destek vektör makineleri, yapay sinir ağları ve rastgele orman ortalama tahmin doğruluğunda yüksek performans vermektedir. Ancak yapay sinir ağları ve rastgele orman eğitim hızında düşük performans verirken destek vektör makineleri eğitim hızında yüksek performans vermektedir. Ayrıca destek vektör makinelerinin anlatımı kolay ve tahmin hızı da yüksektir. Dolayısıyla performansı en yüksek algoritmanın destek vektör makineleri olduğu görülmektedir.

Çizelge 3.3. Sınıflandırma Yöntemleri Karşılaştırılması (Han ve ark., 2009)

Sınıflandırma Algoritması	Anlatım Kolaylığı	Ortalama Tahmin Doğruluğu	Eğitim Hızı	Tahmin Hızı
Destek Vektör Makineleri	Kolay	Yüksek	Hızlı	Hızlı
K-En Yakın Komşuluk	Kolay	Düşük	Hızlı	Komşuluk sayısına bağlı
Karar Ağaçları	Orta	Düşük	Hızlı	Hızlı
Naive Bayes	Orta	Düşük	Hızlı	Hızlı
Yapay Sinir Ağları	Zor	Yüksek	Yavaş	Hızlı
Rastgele Orman	Zor	Yüksek	Yavaş	Orta

3.2.1.1.(1). Destek Vektör Makineleri

Destek vektör makinelerinin sahip olduğu matematiksel algoritmalar başlangıçta iki sınıflı doğrusal verilerin sınıflandırılması problemi için tasarlanmıştır. Sonrasında çok sınıflı ve doğrusal olmayan verilerin sınıflandırılması için genelleştirilmiştir. Destek vektör makinelerinin çalışma prensibi iki sınıfı birbirinden ayırabilen en uygun karar fonksiyonun tahmin edilmesidir. Başka bir ifadeyle iki sınıfı birbirinden en uygun şekilde ayırabilen hiper-düzlemin tanımlanması esasına dayanmaktadır (Cortes ve Vapnik, 1995).

Destek vektör makineleri denetimli öğrenme yöntemleridir ve tahmin için güçlü bir istatistiksel öğrenme yöntemi olarak önerilmiştir. Destek vektör makineleri çeşitli sınıflar arasındaki karar sınırlarını belirlemek için ve yüksek boyutlu bir özellik uzayında doğrusal bir makineye dayalı olarak destek vektörlerini kullanır. Destek vektör makineleri, çok sayıda girdi değişkeni ile hızlı eğitim tekniklerinin geliştirilmesini sağlamıştır (Amiri ve Armano, 2013). Destek vektör makineleri algoritması, eğitim verilerindeki her biri iki farklı sınıftan birine ait olan kalıpları tanıyan ve ardından verilerin yeni örneklerini tahmin eden bir model oluşturur. Basitçe söylemek gerekirse, destek vektör makineleri ayrımcı sınıflandırıcılardır ve destek vektör makineleri algoritması yeni örnekleri en uygun ayırma kararı sınırına göre sınıflandırır (Han ve Kamber, 2006; Scholkopf ve ark., 1997).

Destek vektör makineleri doğrusal olarak ayrılıp ayrılamamasına göre ikiye ayrılır.

3.2.1.1.(1).(a) Doğrusal Olarak Ayrılabilen Destek Vektör Makineleri

Eğitim veri seti doğrusal olarak ayrılabilirse, örnekleri ayıran iki sınır seçilerek doğrusal destek vektör makineleri kullanılabilir. Doğrusal olarak ayrılabilir iki sınıfa ait eğitim vektörleri kümesini ayırma sorununu düşünüldüğünde,

$$x_i, y_i \in \mathbb{R}^n, y_i \in \{+1, -1\}, i = 1, \dots, n \quad (1)$$

Burada x_i , gerçek değerli n boyutlu bir girdi vektörü ve y_i , x_i sınıfını belirleyen bir etikettir. Ayırıcı bir hiper düzlem şu şekilde yazılabilir:

$$w \cdot x + b = 0 \quad (2)$$

Burada w bir ağırlık vektörüdür, yani, $w_1, w_2, \dots, w_n$; n , öz niteliklerin sayısıdır; ve b skalerdir (Han ve Kamber, 2006).

w ve b parametreleri aşağıdakilerle sınırlandırılmıştır:

$$\|w\| = 1 \quad (3)$$

Kanonik formdaki ayırıcı bir alt düzlem, aşağıdaki kısıtlamaları karşılamalıdır:

$$w \cdot x_i + b \geq 1 \quad (4)$$

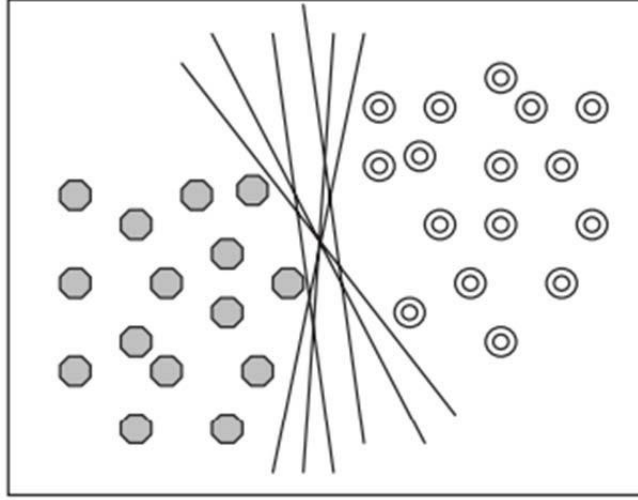
Verileri en iyi şekilde ayıran hiper düzlem, en aza indirgeyendir.

$$\min_{w,b} \sum_{i=1}^n (w \cdot x_i + b - 1) \quad (5)$$

Optimum fonksiyon şu şekilde verilir:

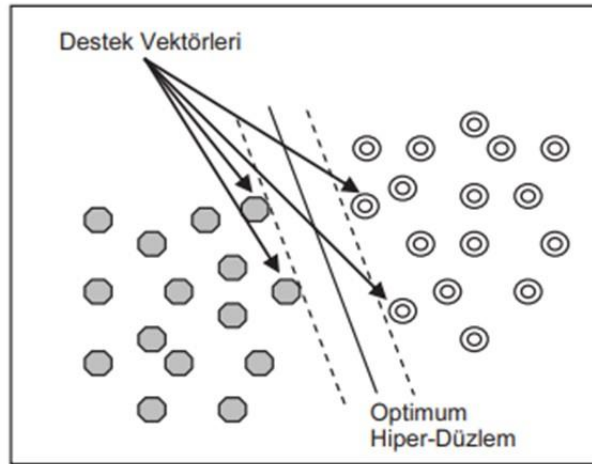
$$\max_{\alpha} \sum_{i=1}^n \alpha_i - \sum_{i=1}^n \alpha_i \quad (6)$$

İki sınıflı bir problem için hiper-düzlemler şekil 3.2’de gösterilmiştir.



Şekil 3.2. İki sınıflı bir problem için hiper-düzlemler (Kavzoğlu ve Çölkesen, 2010a)

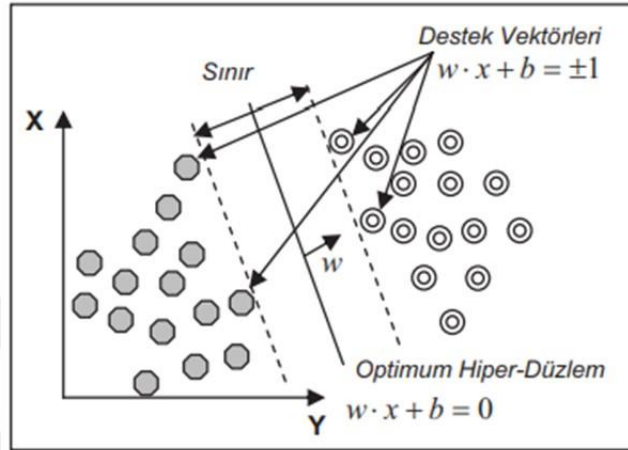
Optimum hiper-düzlem ve destek vektörler şekil 3.3'te gösterilmiştir.



Şekil 3.3. Optimum hiper-düzlem ve destek vektörler (Kavzoğlu ve Çölkesen, 2010a)

Veri setinin doğrusal olarak ayrılma durumuna ilişkin destek vektör makinelerinin geometrik yapısı şekil 3.4'te gösterilmiştir. Hiper düzlemin sol

tarafında bir sınıf; sağ tarafında ise diğer sınıf bulunmaktadır. Optimum hiper düzlem sınıfların sınırlarına arasında bulunmaktadır.



Şekil 3.4. Doğrusal olarak ayrılabilen veri setleri için hiper-düzlemin belirlenmesi (Kavzoğlu ve Çölkesen, 2010a)

3.2.1.1.(1).(b) Doğrusal Olarak Ayrılamayan Destek Vektör Makineleri

Eğitim veri kümesi doğrusal olarak ayrılamıyorsa, Destek Vektör Makineleri giriş vektörünü yüksek boyutlu bir örnek uzayına eşler. Doğrusal olmayan bir haritalama tanımlayarak, destek vektör makineleri bu yüksek boyutlu uzayda optimum bir ayırma hiper düzlem oluşturur. Doğrusal olmayan eşleme genellikle şu şekilde tanımlanır:

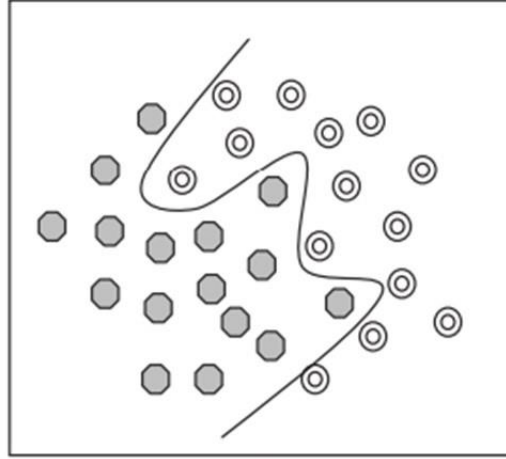
$$\therefore \rightarrow \quad (7)$$

$$\max_{\alpha} \sum \alpha - \sum \alpha \quad (8)$$

$$, \quad (9)$$

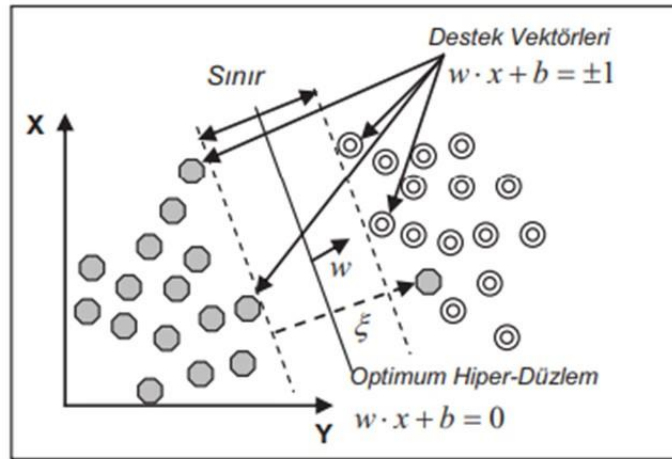
Bu durumda, optimum işlev öznitelik uzayına doğrusal olmayan eşlemeyi gerçekleştiren çekirdek işlevidir (Maria ve Yassine, 2020).

Doğrusal olarak ayrılamayan veri seti şekil 3.5’te gösterilmiştir.



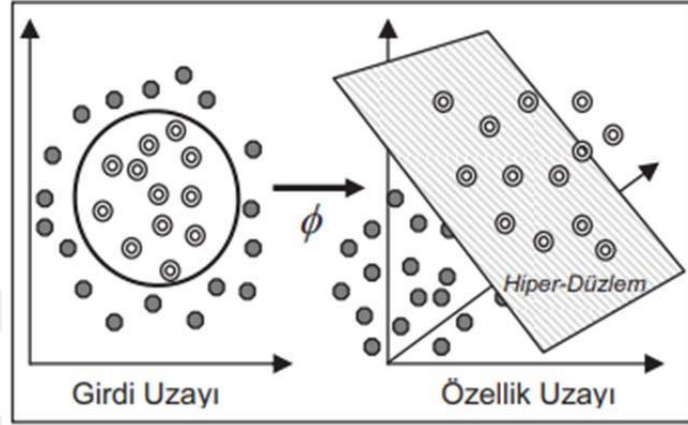
Şekil 3.5. Doğrusal olarak ayrılamayan veri seti (Kavzoğlu ve Çölkesen, 2010a)

Doğrusal olarak ayrılamayan veri setleri için hiper-düzlemin belirlenmesi şekil 3.6’da gösterilmiştir.



Şekil 3.6. Doğrusal olarak ayrılamayan veri setleri için hiper-düzlemin belirlenmesi (Kavzoğlu ve Çölkesen, 2010a)

Kernel fonksiyonu ile verinin daha yüksek bir boyuta dönüştürülmesi şekil 3.7’de gösterilmiştir.



Şekil 3.7. Kernel fonksiyonu ile verinin daha yüksek bir boyuta dönüştürülmesi (Kavzoğlu ve Çölkesen, 2010a)

3.2.1.1.(2) K-En Yakın Komşuluk Algoritması

K-en yakın komşuluk algoritması örnek veri noktasının bulunduğu sınıfın ve en yakın komşunun, k değerine göre belirlendiği bir sınıflandırma yöntemidir (Cover ve Hart, 1967). K-en yakın komşuluk algoritması en iyi bilinen sınıflandırma yöntemlerinden biridir ve makine öğrenme algoritmaları arasında popüler olarak kullanılmaktadır (Qiu ve ark., 2008; Batista ve Silva, 2009). Ayrıca veri madenciliği, örüntü tanıma, yapay zekâ, istatistik, tıp, biyoinformatik gibi çok çeşitli alanlarda uygulanmaktadır (Keller ve ark., 1985; Mao ve ark., 2015). K-en yakın komşuluk algoritmasının bazı avantajları: kolay gerçekleştirilebili, analitik olarak izlenebilir, yerel bilgilere uyarlanabilir, paralel gerçekleştirime uygun, gürültülü eğitim verilerine karşı dirençli olması şeklinde sıralanabilir. Yüksek miktarda bellek alanına gereksinim duyması, veri seti boyutu arttıkça işlem yükünün artması, performansın k komşu sayısı, uzaklık ölçütü ve öznitelik sayısı gibi parametrelere bağlı olarak etkilenmesi gibi dezavantajları bulunmaktadır (Bhatia ve Vandana, 2010; Liu ve Zhang, 2012).

K-en yakın komşuluk algoritmasında öğrenme işlemi eğitim setinde tutulan verilere dayalı olarak tamamlanmaktadır. İlk kez karşılaşılan bir örnek, eğitim setinde yer alan örnekler ile arasındaki benzerliğe göre sınıflandırılmaktadır (Mitchell, 1997). Eğitim setinde yer alan örnekler n boyutlu sayısal nitelikler ile belirtilir. Her örnek n boyutlu uzayda bir noktayı temsil edecek biçimde tüm eğitim örnekleri, örnek uzayında tutulur. Bir örnekle ilk kez karşılaşıldığında, eğitim setinden ilgili örneğe en yakın k tane örnek belirlenerek yeni örneğin sınıf etiketi, k -en yakın komşusunun sınıf etiketlerinin çoğunluk oylamasına göre atanır (Han ve Kamber, 2006).

K-en yakın komşuluk algoritmasının performansı için önemli noktalardan birisi örnekler arası yakınlığın nasıl ölçümleneceğidir. Sınıf etiketinin çoğunluk oylamasına dayalı olarak belirlenmesi, simetrik olmayan dağılıma sahip veri setlerinde sıklıkla görülen sınıfların, yeni örneklerin sınıf etiketlerinin belirlenmesinde daha baskın bir role sahip olmalarına neden olmaktadır (Coomans ve Massart, 1982). Bu nedenle, k -en yakın komşuluk algoritmasının uzaklık ölçütünün etki değerine farklı şekillerde ağırlık değeri atayan yöntemler bulunmaktadır (Gartner ve ark., 2004; Aha ve ark., 1991; Duda ve ark., 2000).

3.2.1.1.(2).(a) K- En Yakın Komşuluk Algoritmasında Kullanılan Uzaklıklar ve Komşu Sayısı

K-en yakın komşuluk algoritmasında Minkowski, Öklid ve Manhattan olmak üzere üç farklı metrik kullanılmıştır.

□ Minkowski Uzaklığı

Minkowski uzaklığı, öklid uzayında tanımlı bir dizidir. Sınıflandırma, kümeleme gibi algoritmalarda sıklıkla kullanılmaktadır (Taşçı ve Onan, 2015).

$$\sum_{i=1}^n |x_i - y_i|^p \quad (10)$$

□ Öklid Uzaklığı

Öklid uzaklığı, iki nokta arasındaki doğrusal uzaklığı belirtir ve sınıflandırma, kümeleme algoritmalarında en sık kullanılan uzaklık ölçütüdür (Taşçı ve Onan, 2015).

$$\sqrt{\sum (11)}$$

□ Manhattan Uzaklığı

Manhattan uzaklığı, n boyutlu iki nokta arasındaki farkların mutlak değerlerinin toplamı şeklinde tanımlanmaktadır (Taşçı ve Onan, 2015).

$$\sum (12)$$

□ Komşu Sayısı (k)

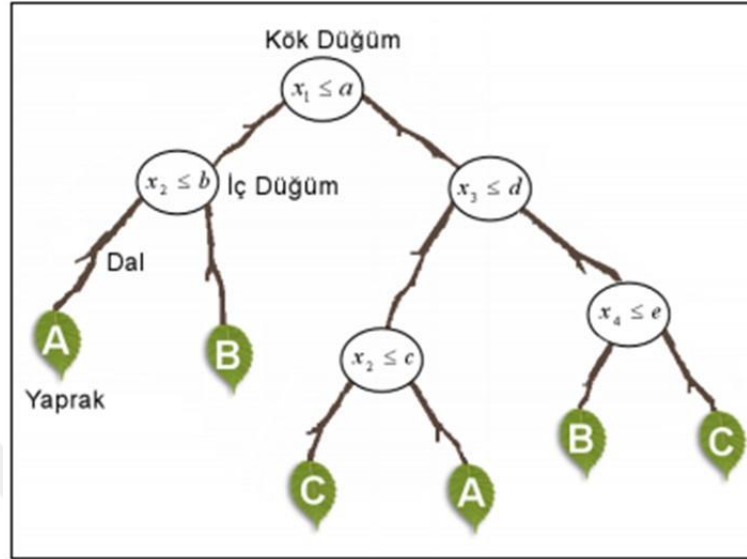
K-en yakın komşuluk algoritmasında, sınıflandırma komşu sayısı k parametresinin değerine dayalı olarak yapılmaktadır. Sınıflandırma sürecinde, k 1 için, sadece en yakın komşunun bulunduğu sınıfa atanırken, k sayısı örnek sayısına yaklaştıkça veri setinde yer alan tüm veriler dikkate alınmaktadır (Taşçı ve Onan, 2015).

3.2.1.1.(3) Karar Ağaçları

Karar ağaçları son yıllarda literatürde yaygın olarak kullanılmaktadır. Bunun en önemli nedeni ağaç yapılarının oluşturulmasında kullanılan kuralların anlaşılabilir ve sade olmasıdır. Karar ağacı sınıflandırma işleminin gerçekleştirilmesinde çok aşamalı veya ardışık bir yaklaşım kullanılmaktadır (Safavian ve Landgrebe, 1991).

Bir karar ağacı temel yapısı düğüm, dal ve yaprak olarak adlandırılan 3 temel kısımdan oluşmaktadır. Bu ağaç yapısında her bir öznitelik bir düğüm tarafından

temsil edilmektedir. Dallar ve yapraklar ise ağaç yapısının diğer elemanlarıdır (Quinlan, 1993). Bir ağacın yapısı; verileri temsilen bir kök düğüm, iç düğümleri temsilen dallar ve uç düğümleri temsilen yapraklardan oluşmaktadır. Eğitim verilerine ait öznitelik bilgilerinden yararlanılarak bir karar ağacı yapısı oluşturulurken temel prensip verilere ilişkin bir dizi sorular sorulması ve cevaplar doğrultusunda en kısa sürede sonuca gidilmesi olarak ifade edilebilmektedir. Başka bir deyişle, karar ağacı sorulara aldığı cevapları toplayarak karar kuralları oluşturmaktadır. Kök düğümünde verilerin sınıflandırılması ve ağaç yapısının oluşturulması için sorular sorulmaya başlanmaktadır. Dalları olmayan düğümler ya da yapraklar bulunana kadar bu işlem devam etmektedir (Pal ve Mather, 2003). Oluşturulan ağacın yeni bir veri seti için genelleme kabiliyetinin belirlenmesi için test veri seti kullanılmaktadır. Eğitim veri seti ile oluşturulan ağaç yapısına yeni gelen bir test verisi, ağacın kökünden girmekte ve kökte test edilen bu yeni veri test sonucuna göre bir alt düğüme gönderilmektedir. Belirli bir yaprağa gelene kadar bu işleme devam edilmektedir. Kökten her bir yaprağa giden tek bir yol veya tek bir karar kuralı oluşmaktadır. Şekil 3.8’de 3 sınıfa ait dört boyutlu öznitelik değerlerinden oluşan basit bir ağaç yapısı görülmektedir. Şekilde x_i öznitelik değerlerini a , b , c , d ve e değerleri dallanma için esas olan eşik değerleri ve A , B ve C ise sınıf etiketlerini göstermektedir.



Şekil 3.8. Dört boyutlu özellik uzayına sahip 3 sınıftan oluşan bir karar ağacı yapısı (Kavzoğlu ve Çölkesen, 2010b)

Karar ağacı oluşturulmasındaki en önemli adım ağaçtaki dallanmanın hangi kritere göre yapılacağı ya da hangi öznelilik değerlerine göre ağaç yapısının oluşturulacağıdır. Bu yöntemle göre karar ağacında hangi özelliğe göre dallanmanın yapılacağını belirlemek üzere entropi kurallarını içeren bilgi teorisi kullanılmıştır. Entropi bir sistemdeki düzensizliğin ya da belirsizliğin ölçüsüdür.

$$\Sigma \quad (13)$$

Her değişken için entropi değerleri hesaplanır ve bu değerler diğerinden çıkarılarak bilgi değerleri elde edilebilir (Tanner, 2005).

3.2.1.1.(4) Naive Bayes

Olasılık teorisinin iki temel kuralı olan toplam kuralı eşitlik 14'te, ürün kuralı eşitlik 15'te belirtilmektedir

$$\sum_{j=1}^n P_{j|y} \quad (14)$$

$$P_{j|y} = \frac{P_{j|y} P_y}{\sum_{i=1}^n P_{i|y} P_y} \quad (15)$$

Eşitliklerde $P_{x,y}$ ortak olasılık, $P_{y|x}$ koşullu olasılık, P_x marjinal olasılıktır. Ürün kuralına dayanarak, simetri özelliği, , , uygulandığında, eşitlik 16'da belirtilen Bayes kuralının elde edilmesi kolaylaşmaktadır.

$$P_{x|y} = \frac{P_{y|x} P_x}{\sum_{j=1}^n P_{j|y} P_y} \quad (16)$$

Toplam kuralının kullanılması ile de Bayes teoreminde pay olarak görünen eşitlik payda içinde de ifade edilebilmektedir (Deng ve ark., 2015).

$$\sum_{j=1}^n P_{j|y} P_y \quad (17)$$

Verilen veri örneklerinde sınıf olasılıklarının bulunmasında y değişkeni, veri örnekleri için sınıf etiketini temsil etmektedir ve y_j sınıfı için önsel olasılık olarak ifade edilmektedir. x değişkeni, veri örneklerinin özelliğini temsil etmektedir. Veri setindeki x özelliğine karşılık y sınıfının olma olasılığını hesaplayan $|$ sonsal olasılık değerinin bulunmasında, farklı sınıflar için gözlenen x verisinin olasılığı $|$ ve önsel olasılık değerlerine ihtiyaç vardır (Deng ve ark., 2015; Arpacı ve Kalıpsız, 2018).

$$P_{x|y} = \frac{P_{y|x} P_x}{\sum_{j=1}^n P_{j|y} P_y} \quad (18)$$

Naive bayes sınıflayıcı, en iyi olasılık hangi sınıf için hesaplandıysa karar sınıfı olarak o sınıfı seçmektedir. Bu sınıflandırıcı niteliklerin birbirinden bağımsız olduğuna ve niteliklerin hepsinin aynı derecede önemli olduğunu varsaymaktadır (Sarıkoc, 2012; Solmaz ve ark., 2014).

Öznitelik sayısı birden fazla olması durumunda

$$| \quad \Pi \quad |c \quad (19)$$

Bayes teoreminde koşullu olasılıklar arasında bağıntı kurar. Bir c sınıfının bilinmesi koşulunda x özniteliğinin var olma olasılığı ile x özniteliğinin bilinmesi koşulunda c sınıfının var olma olasılığı arasında Bayes teoremine göre eşitlikteki bağıntı mevcuttur (Solmaz ve ark., 2014).

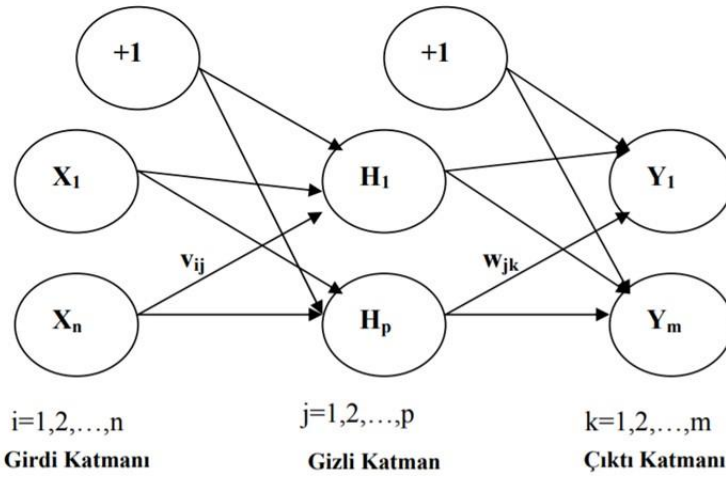
$$|x \quad \text{_____} | \quad (20)$$

3.2.1.1.(5) Yapay Sinir Ağları

Yapılan birçok çalışma geleceği tahminde, yapay sinir ağlarının iyi neticeler verdiğini göstermiştir (Hill ve ark., 1996; Tang ve ark., 1991). Yapay sinir ağlarının özellikle doğrusal olmayan zaman serilerinde başarı göstermektedir ve bu yüzden bir tahmin aracı olarak tercih edilmektedir (Zhang ve ark., 1998). Yapay sinir ağları insan beyin fonksiyonundan esinlenmiş ve deneme yolu ile öğrenme ve genelleştirme yapabilmektedir. Yapay sinir ağlarının kullanıldığı önemli alanlardan biri de geleceği tahmindir. Yapay sinir ağları veriler arasındaki bilinmeyen ve fark edilmesi güç ilişkileri ortaya çıkartabilmektedir. Yapay sinir ağları girdi ve çıktı değişkenleri arasındaki herhangi bir ön bilgiye ihtiyaç duymadan, herhangi bir varsayımda bulunmadan, doğrusal olmayan modellemeyi sağlayabilmektedir (Kaastra ve Boyd, 1996). Ağa, girdi bilgileri ve bu girdilere karşılık gelen çıktı bilgileri verilerek ağı

girdi-çıkı arasında ilişkiyi öğrenmesi sağlanmakta ve ağı eğitimi gerçekleştirilmektedir (Haykin, 1999).

İleri beslemeli tipik bir yapay sinir ağı, girdi katmanı, ara katman (gizli katman) ve çıktı katmanından oluşmaktadır. Her katmanda ilgilenilen probleme göre değişen sayıda nöron bulunmaktadır. Şekil 3.9’da ileri beslemeli tek gizli katmanlı bir yapay sinir ağı görülmektedir. Girdi katmanında n , gizli katmanda p ve çıktı katmanında m adet nöron bulunmaktadır. Her bir katmandaki nöronlar arasındaki bağlantıların ağırlıklarının düzenlenmesi ile ağı eğitimi gerçekleştirilmektedir.



Şekil 3.9. Tipik bir ileri beslemeli yapay sinir ağı (Hamzaçebi ve Kutay, 2010)

Bir katmandaki her nöron, bir üst katmandaki tüm nöronlara değişik sayısal ağırlıklarla (V_{ij}) bağlıdır. Ağırlıklar, öncül katmandaki nöronun (W_{jk}), üst katmandaki nöron üzerindeki etkisini belirtmektedir. Pozitif ağırlık değerleri pekiştirme, negatif değerler ise engel olma durumunu belirtmektedir (Gallant, 1993). Ağ üzerindeki her birim alt düzeyden gelen girdi değerini almakta ve ağırlıklı toplamı hesaplamaktadır. Girdi verileri bağlantı ağırlıkları ile çarpılarak ağ üzerinde ilerlemektedir. Ağırlıklar başlangıçta rasgele olarak belirlenmekte ve böylece sistemin başlangıç noktasında kilitlenmemesi sağlanmaktadır (Baxter, 2004).

Çarpım sonuçları ilk gizli katmanda toplanmaktadır. Sonuç bir transfer fonksiyonundan geçirilmekte, 0 ile 1 arasında değişen örneksel çıktı elde edilmektedir (Deboeck, 1994). Sinyaller gizli katmanlar arasında ve çıktı katmanına aynı şekilde iletilmektedir. Teorik olarak herhangi bir tahmin fonksiyonu 1 gizli katman ile çözülebilmekte ancak 2 katman kullanıldığında çözüme daha kolay ulaşılabilmektedir (Partovi ve Anandrajan, 2002; Carsten ve Thorstein, 1993). Gizli katmandaki nöron sayısı genellikle deneme ve yanılma yöntemi ile belirlenmektedir (Nehdi ve ark., 2001).

Bağlantıların girdiden çıktıya doğru tek yönlü olarak geliştiği ağlara ileri beslemeli ağlar denilmektedir (Rosenblatt, 1961). Geri dağılım kullanılırken ağ her girdi deseni ile hedef çıktıyı karşılaştırarak öğrenmekte, hatayı hesaplamakta ve bir hata fonksiyonunu ağ üzerinde geriye doğru çalıştırmaktadır. Bu tip ağların en önemli avantajı kolay kullanım ve girdi/çıktı haritası için yakınsamadır. En önemli dezavantajı ise yavaş öğrenmesi ve öğrenme süreci için çok fazla veri talep etmesidir (Rumelhart ve ark., 1986).

Öğrenme ağın serbest parametrelerinin optimal değerlerini bulması işlemi olarak tanımlanmaktadır. Öğrenme süreci tamamlandığında ağırlık değerleri belirlenmiş olan ağ, öğrenme işlemi için kullanılan veri dışında kalan veri kümesi üzerinde test edilmektedir. Böylece ağın öğrenme kümesi verisine büyük ölçüde benzeyen örüntüler için doğru yanıtlar üreteceğinin geçerliliği sağlanmaktadır. Genelleme yapılmak istendiğinde ağ yeterli büyüklükteki giriş veri kümesi ile eğitilmelidir (Nagendra, 1998). Algoritmanın performansını iyileştirmek amacı ile ağ üzerinde birden fazla öğrenme ve test işlemi gerçekleştirilmelidir. Öğrenme kümesinin test kümesine oranı bu çalışmada 9:1 oranı kullanılmaktadır (Tanner, 2005).

3.2.1.1.(6) Rastgele Orman

Karar ağaçlarının birleşiminden oluşan rastgele orman algoritmasında kullanılan karar ağaçları arasında doğruluğu ve bağımsızlığı en yüksek ağaçlar tercih

edilmektedir. Ağaçların her birisi veri setindeki özelliklere göre dallanmakta ve bu özelliklerde karar noktalarına bağlanmaktadır. Karar ağaçları kullanan modellerdeki en büyük problem, verinin az olması durumunda aşırı oranda uyumun oluşmasıdır (Liao, 2016). Rastgele orman sınıflandırıcısı, veri kümesinin çeşitli alt örneklerine bir dizi karar ağacı sınıflandırmasına uyan ve tahmin doğruluğunu iyileştirmek ve aşırı uyumu kontrol etmek için ortalamayı kullanan bir meta tahmincidir. Alt örnek boyutu her zaman orijinal giriş örneklem boyutuyla aynı olmaktadır (Veranyurt ve ark., 2020). Rastgele orman düğümleri ayırırken en belirgin özelliği aramak yerine, rastgele bir özellik alt kümesi arasında en iyi özelliği arar. Her bir karar ağacı tahmini düşük olabilir, ancak tahminler birleştirildikçe, rastgele orman doğruluğu ve kararlılığı artar. Rastgele ormanın hem özellik hem de örnek için rastgeleliği olduğundan, aşırı uyum bir sorun değildir. Rastgele orman gürültüden ve aykırı değerlerden daha az etkilenir ve genelleme hatalarını azaltarak yüksek doğruluğa sahiptir (Witten ve ark., 2011).

3.2.1.2. Nitelik Seçimi

Nitelik seçimi, bölme ölçütünü seçmek için en iyi sınıf etiketli eğitim kümelerinin verisini bireysel olarak ayırır. Niteliği tanımlayan her özellik için bir sıralama sağlar. Aynı zamanda bölme kuralları olarak da bilinir. Kullanılan nitelik seçimi yöntemleri aşağıda anlatılmaya çalışılmıştır.

Klinik tanı alanında, küçük bir özellik alt kümesi daha düşük test masrafları anlamına gelir. Bir sınıflandırıcıdaki girdi özelliklerinin sayısını sınırlamanın avantajı, düzgün, öngörücü ve daha az yoğun bir modele sahip olmaktır.

3.2.1.2.(1) Bilgi Kazanımı

Bu yöntem, mesajların değerini veya bilgi içeriğini inceleyen bilgi teorisi üzerine Claude Shannon'un öncü çalışmalarına dayanmaktadır. Bir nitelik için sınıflandırma sonuçlarının ne kadar değer ile kazanılabileceğini gösterir. Bir niteliğin değerini sınıfa göre edindiği bilgileri ölçerek değerlendirir. Özniteliklere göre bölünen veri

setindeki tahmini kaybı hesaplamada kullanılan entropiye dayalı yöntemlerden biridir. Entropi sistemin düzensizliğini ya da belirsizliğini belirleyen 0 ile 1 arasında bir değerdir. Entropi değerinin 1'e yaklaşması sistemin daha çok bilgi içerdiğini göstermekte ve entropi değerini kullanarak her niteliğin karar verilecek sınıfın ayırt ediciliğini hesaplamaya çalışmaktadır. Bir niteliğin bilgi kazanımını hesaplamak için öncelikle niteliğin her sınıf için ayrı ayrı entropisi hesaplanmaktadır. Bu değer beklenen bilgi değerinden çıkarılmakta ve herhangi bir A niteliğinin bilgi kazanımı, niteliğin gözlenmeden önceki ve sonraki entropi değerleri arasındaki fark olarak aşağıdaki gibi hesaplanmaktadır. Bilgi kazanımı en yüksek olan nitelikler sınıflandırma işleminde kullanılmaktadır. Bilgi kazanımı yöntemi ile öznelik seçimi yapılırken sistemi tanımlamada yetersiz kalan değişkenler veri setinden çıkarılır ve kalan değişkenler sistemi eğitmek için kullanılır (Kaynar ve ark., 2018).

$$\Sigma \quad (21)$$

Yukarıdaki denklemde p_i , i sınıfına ait eğitim kümesindeki küme sayısının veri kümesindeki tüm kümelerin sayısına bölünmesiyle elde edilen i sınıfının olasılığıdır. , aynı zamanda veri kümesi D 'nin entropisi olarak da adlandırılır (Han ve Kamber, 2006).

$$\Sigma \quad (22) \text{ Burada } D_j, \text{ olduğu } D \text{'nin}$$

bir alt kümesidir ve $|D_j|$ bu altkümenin kardinalitesidir.

$$(23)$$

Kazanç değeri, veri kümesindeki tüm öznitelikler için hesaplanır. Daha sonra bilgi kazanımı öznitelik seçim yönteminde en yüksek kazanç değerlerine sahip öznitelikler seçilir. Diğer bir deyişle, bize A üzerinde dallanıldığında ne elde edileceğini söyler. Bu, A'nın değerini bilmenin neden olduğu veri ihtiyaçlarında beklenen azalmadır.

3.2.1.2.(2) Ki-Kare

Ki-kare yöntemi değişkenlerin veri setini tanımlamaya uygun olup olmadığını belirlemek içinde kullanılabilir. Ki-kare testinde veri setindeki değişkenlerin uygun olduğu H_0 ve veri setindeki değişkenlerin uygun olmadığı H_1 olmak üzere iki hipotez bulunmaktadır. İlk aşamada gözlenen değerlerin gerçek sınıflara göre ki-kare istatistiği hesaplanmaktadır. Ki-kare değeri sıfır ile pozitif sonsuz arasında değerler alabilmekte ve bu değer sıfıra yaklaşması gözlenen frekans değerleri ile beklenen frekans değerlerinin daha uyumlu olduğunu göstermektedir. Bu değer çok büyük olması uyumsuzluğu işaret etmektedir. Bu nedenle ikinci aşamada ilk aşamada hesaplanan ki-kare değeri ki-kare dağılımındaki belirlenen eşik değeri ile kıyaslanmaktadır. Bu eşik değeri önemlilik seviyesine ve serbestlik derecesine bakılarak bulunmaktadır. Önemlilik seviyesi testi yapan kişi tarafından belirlenen yüzde değeridir, serbestlik derecesi ise veri setindeki öznitelik sayısının bir eksigidir. Hesaplanan değer belirlenen değerden büyük ise H_1 hipotezi, küçük ise H_0 hipotezi kabul edilir (Kaynar ve ark., 2018).

$$2 \sum \sum \frac{O_{ij}^2}{e_{ij}} \quad (24)$$

Burada O_{ij} , ortak olayın gözlemlenen sıklığıdır ve e_{ij} , ortak olayın beklenen sıklığıdır (Han ve Kamber, 2006).

3.2.1.2.(3) Kazanım Oranı

Bilgi kazanımı birçok sonucu olan testlerle doğru orantılıdır. Başka bir deyişle, çok sayıda değer içeren nitelikleri seçmeyi tercih eder. Kazanım oranı benzer şekilde tanımlanmış bir bölünmüş bilgi değeri kullanarak bilgi kazanımına bir çeşit normalizasyon uygular. Sınıfa göre kazanç oranını ölçerek bir niteliği değerlendirir. Kazanım oranı, bilgi kazanımı yöntemine alternatif olarak kullanılan ve aynı amaç doğrultusunda öznitelik seçimi yapan yöntemlerden biridir. Veri setinde bir öznitelik çok fazla farklı değere sahip olduğunda o öznitelik için her bir farklı değere düşen örnek sayısı düşük olduğundan o öznitelik için hesaplanan entropi değeri küçük, bilgi kazancı ise büyük çıkmaktadır. Bilgi kazanımı yönteminde de anlatıldığı gibi bu değerın büyük çıkması o değişkenin veri setini tanımlamada iyi olduğunu göstermektedir. Özniteliğin alabileceği farklı değer sayısının çok olması durumunda bilgi kazanımı yönteminin o özniteliği iyi bir seçici olarak seçmesi sistemin ezberleme yapmasına neden olabilmektedir. Kazanım oranı bu probleme çözüm olarak her bir öznitelik için bölünme bilgisini hesaplamakta ve elde edilen bilgi kazancını bölünme bilgisine bölerek kazanım oranını hesaplamaktadır. Bu oranın 1'e yaklaşması o değişkenin veri setini tanımlamada başarılı olduğunu göstermektedir (Kaynar ve ark., 2018).

(25)

3.2.1.2.(4) Relief-f

Kira ve arkadaşları tarafından 1992 yılında geliştirilen relief-f algoritması öznitelikleri aralarındaki ilişkiye göre ağırlıklandıran bir öznitelik seçim yöntemidir. Bu yöntem iki sınıfı olan veri setleri için başarılı sonuç verse de çoklu sınıfa sahip veri setleri için çalışmadığından ötürü bu problemi çözmek için 1994 yılında Kononenko tarafından çoklu sınıfı olan veri setleri içinde çalışan relief-f

algoritması geliştirilmiştir. Bu yöntemin ilk aşamasında tüm özneliklerin ağırlıkları 0 olarak belirlenmektedir. Daha sonra her bir adımda veri setinden rastgele bir veri seçilmekte ve bu veri ile aynı sınıfa ait en yakın k adet veri bulunarak her bir farklı sınıfa ait en yakın veriler bulunmaktadır. Sonrasında her bir özelliğe ait ağırlıklar bu veriler kullanılarak güncellenmektedir. Son aşamada belirlenen koşulu sağlamayan özellikler veri setinden atılarak yeni veri seti oluşturulmaktadır (Kaynar ve ark., 2018).

Relief-f algoritması, makine öğrenimi alanındaki en başarılı özellik seçme algoritmalarından biridir. Çok sayıda alt küme özellik seçme görevinde kullanılmıştır. Hem bağımlı hem de bağımsız özelliklerin kalitesini tahmin eder ve her problem için tüm anlamlı özellikleri bulmaya çalışır (Wang ve Makedon, 2004; Florez-Lopez, 2002).

Relief-f algoritmasının pseudo kodu aşağıdaki gibi özetlenebilir

- initialize vector W
- for $i=1$ to m do
- randomly select a target instance R_i
- find a nearest hits H and nearest miss M
- for $A=1$ to a do
- $W[A] = W[A] - \text{diff}(A, R_i, H)/m + \text{diff}(A, R_i, M)/m$
- end
- end
- return W

(RobnikSikonja ve Kononenko, 2003).

3.2.1.3. Hata Matrisi (Confusion Matrix)

Bir hata matrisi, gerçek değerlerin bilinmekte olduğu bir dizi test verisi üzerinde makine öğrenmesinde kullanılan sınıflandırma modellerinin performansını değerlendirmek için hedef niteliğe ait tahminlerin ve gerçek değerlerin karşılaştırmak için kullanılan tablodur.

Çizelge 3.4 bir sınıflandırma sistemi tarafından yürütülen gerçek ve tahmin edilen sınıflandırmalar hakkında bilgi içeren iki sınıf sınıflandırıcı için hata matrisini göstermektedir.

Çizelge 3.4. Hata Matrisi Gösterimi (Kohavi ve Provost, 1998)

Gerçek	Tahmin	
	Pozitif	Negatif
Pozitif	Doğru Pozitif (TP)	Yanlış Negatif (FN)
Negatif	Yanlış Pozitif (FP)	Doğru Negatif (TN)

Sınıflandırma doğruluğu, duyarlılık, özgüllük, kesinlik, f-measure ve kappa değerleri hata matrisinin öğeleri kullanılarak tanımlanabilir:

$$\text{Doğruluk} = \frac{TP + TN}{TP + FP + FN + TN} \quad (26)$$

$$\text{Duyarlılık} = \frac{TP}{TP + FN} \quad (27)$$

$$\text{Özgüllük} = \frac{TN}{FP + TN} \quad (28)$$

$$\text{Kesinlik} = \frac{TP}{TP + FP} \quad (29)$$

$$\text{F-Measure} = \frac{2 \cdot \text{Doğruluk} \cdot \text{Kesinlik}}{\text{Doğruluk} + \text{Kesinlik}} \quad (30)$$

3.2.1.4. Kümeleme

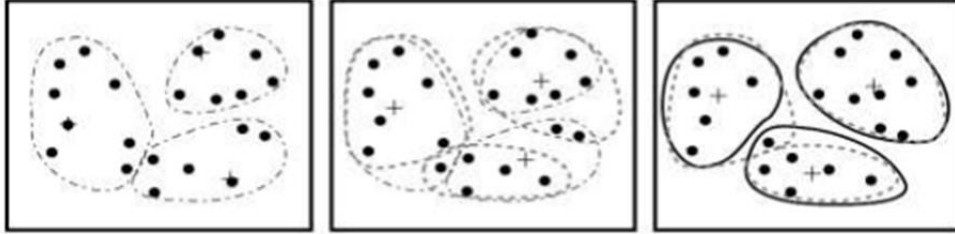
Bir veri kümesindeki bilgileri belirli yakınlık kriterlerine göre gruplara ayırma işlemi kümeleme analizi olarak tanımlanmaktadır. Kümeleme işleminde küme içindeki elemanların benzerliği fazlayken, kümeler arası benzerlik ise az olmalıdır. Her bir kümede yer alan elemanlar grubu diğer gruplardan ayıran ortak özelliklere sahiptir. Kümeleme veri madenciliği tekniklerinden tanımlayıcı modeller olan gözetimsiz sınıflandırmaya girmektedir. Gözetimsiz sınıflamada amaç, başlangıçta verilen ve henüz sınıflandırılmamış bir veriyi anlamlı alt kümeler oluşturacak şekilde öbeklemektir (Dinçer, 2006). Kümeleme analizi yöntemleri: tek bağlantı, tam bağlantı, ortalama bağlantı, ağırlıklı ortalama bağlantı, merkezi bağlantı, medyan bağlantı, Ward's bağlantı, k-ortalama, en çok olabilirlik yöntemleri olarak sıralanabilir. Bu çalışmada k-ortalama yöntemi kullanılmıştır.

3.2.1.4.(1) K-Ortalama Algoritması

En yaygın kullanılan gözetimsiz öğrenme yöntemlerinden biri olan k-ortalama algoritmasının atama mekanizması her verinin sadece bir kümeye ait olabilmesine izin vermektedir. Bu nedenle, keskin bir kümeleme algoritması olan k-ortalama algoritmasının genel mantığı n adet veri nesnesinden oluşan bir veri kümesini, giriş parametresi olarak verilen k adet kümeye bölümlenektir. Bölümleme işlemi sonunda elde edilen kümelerin, küme içi benzerliklerinin maksimum ve kümeler arası benzerliklerinin minimum olması amaçlanmaktadır. Çalışma yönteminde, öklit uzaklığı formülü temel alınarak kümeleme yapılmaktadır.

K-ortalama algoritması rastgele seçilen k adet merkez noktayla başlamakta ve veri kümesindeki her nokta kendisine en yakın merkez noktanın kümesine atanmaktadır. Küme merkezinin değeri kendine ait noktaların ortalaması alınarak hesaplanmakta ve bu işlem merkezlerin değerleri değişmeyinceye kadar devam etmektedir (Han ve ark., 2009).

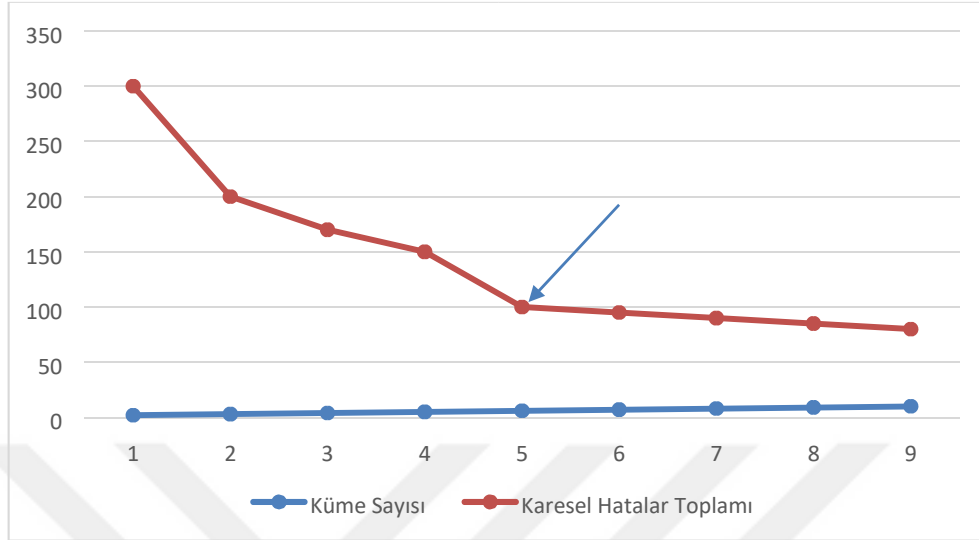
Şekil 3.10'da k-ortalama algoritması ile kümeleme aşamalarını gösterilmektedir.



Şekil 3.10. K-Ortalama Algoritması ile kümeleme aşamaları (Han ve Kamber, 2006)

3.2.1.4.(2) Optimum Küme Sayısının Belirlenmesi

Kümeler için karesel hatanın hesaplanması yardımıyla gerçekleştirilen dirsek yöntemi, bir veri kümesinde uygun sayıda kümenin bulunmasına yardımcı olmak üzere tasarlanan kümeleme analizi içindeki tutarlılığın yorumlanması ve doğrulanması için bir yöntemdir. Karesel hatanın düşük olması kümeleme modelinin daha iyi olduğu anlamına gelmektedir. Yöntemde her bir küme sayısı için karesel hata hesaplanmakta ve yatay eksene küme sayısı, dikey eksene karesel hata gelecek şekilde grafiğe yerleştirilmekte ve şekilde karesel hatanın azalma hızının düştüğü nokta ideal küme sayısı olarak belirlenmektedir. Şekil 3.11'de, çalışmada farklı küme sayıları için elde edilen karesel hatalar gösterilmiştir (Ayyıldız ve Yalçın, 2018).



Şekil 3.11. Optimum Küme Sayısının Belirlenmesi

3.2.3. Çalışmada Kullanılan İstatistiki Yöntemler

3.2.3.1. Örneklem Büyüklüğü

Bir evrenin bütün özelliklerini gösteren kısmının seçilmesi örneklem işlemini tanımlamaktadır. Bir bütünün daha küçük boyuttaki örneği ise örneklem ifadesini tanımlamaktadır. Örneklem grubunun alındığı yer ise evren ifadesidir. Bir evrenden örneklem seçilirken, seçilen örneklemin seçildiği ana kütleyi temsil edecek özellikte ve büyüklükte olması gerekmektedir. Örneklem seçilerek yapılan araştırmalar zaman ve maliyet yönünden ekonomiktir ve bütün evrenin incelenmesiyle elde edilen sonuçlar kadar geçerli, sağlıklı ve güvenilir olabilir (Gökçe, 1988). Örnek büyüklüğünün ne kadar büyüklükte olması gerektiğine ilişkin bazı hesaplamalar söz konusudur. Alınan örneklem eğer alındığı evreni temsil edemiyorsa, sonuçlarda ve analizlerde hata ortaya çıkar (Bailey, 1987). Yeterli bir örneklem, güvenilir sonuçlar sağlayacak kadar eleman kapsayan örneklemidir (Young, 1968). Örneklem büyüklüğü merkezi limit teoremi aracılığıyla standart hata formülü ile hesaplanmaktadır (Kurtuluş, 1998).

$$\frac{\text{---}}{\text{---}} = \frac{\text{---}}{\text{---}} 380,25 \quad (31)$$

Eşitliğe göre; n örnek hacmini, p incelenen olayın görülüş sıklığını, e hata payı oranı, z ise tablo değerini temsil eder. Formülde popülasyon oranı p 0,50 olarak kabul edilmektedir. Bu değer örnek kütle miktarının % 50'lik kısmını ifade etmektedir. Formülde yerine koyulan değerlerden açıkça görüldüğü üzere hata payı e 0,05 olarak kabul edilmektedir. Bu durum güven aralığının % 95 olarak kabul edildiğini göstermektedir. Formülde değerlerin yerine koyulması sonucunda ulaşılan sonuç 380'dir.

3.2.3.2. Faktör Analizi

Faktör analizi, birbirleriyle ilişkili çok sayıdaki karmaşık değişkeni bir araya getirir. Daha az sayıda anlamlı ve birbirinden bağımsız faktör adı verilen yeni değişkenler oluşturan çok değişkenli istatistiksel analiz yöntemidir (Kalaycı, 2008).

Birçok ilişkili tekniği içerisinde bulunduran faktör analizinde iki temel amaç söz konusudur. Bunlardan ilki, değişkenler arasındaki korelasyonları en iyi açıklayan en az sayıdaki ortak faktör sayısını belirlemektir. Diğeri ise karmaşık yapıda çok sayıda değişken analiz edilerek, en az bilgi kaybıyla olayı açıklayan daha az faktör adı verilen değişkenler türetilmektedir (Albayrak, 2006). Temel bileşenler analizi (principal component analysis - PCA) en sık kullanılan teknik olarak bilinmektedir.

Temel bileşenler analizi orijinal değişkenlerin birkaç doğrusal kombinasyonu ile varyans-kovaryans yapısını açıklamakta ve değişkenler arasındaki bağımlılık yapısını yok etmek, birbiriyle ilişkisiz ve daha az sayıda yeni değişken elde etmek, tahminleme yapmak ve veri kümesini bazı yöntemlerin analiz edebileceği şekle dönüştürme amacına hizmet etmektedir (Süzülmüş, 2005).

Faktör analizinde dört aşama mevcuttur. Veri setinin faktör analizi için uygunluğunun test edilmesi, faktörlerin elde edilmesi, faktörlerin rotasyonu ve faktörlerin isimlendirilmesi olarak belirlenmiştir (Kalaycı, 2008).

3.2.3.2.(1) Veri Setinin Faktör Analizi İçin Uygunluğunun Test Edilmesi

Faktör analizi uygulanacak bir veri setinin uygunluk testi için birden fazla yöntem söz konusudur. Bu yöntemlerden ilki korelasyon matrislerinin oluşturulmasıdır. Bu yöntemde iki değişken arasındaki korelasyon katsayısına bakılarak gruplamalar yapılabilmektedir. İki değişken arasındaki korelasyonun yüksek olması, bu iki değişkenin aynı grupta değerlendirilebileceği anlamına gelirken, düşük korelasyon olması, aynı faktör grubu altında değerlendirilemeyecekleri anlamı taşımaktadır

Veri setinin uygunluğu için ikinci yöntem Barlett küresellik testi olarak adlandırılır. Korelasyon matrisinin birim matris olup olmadığının sınındığı bu testte, korelasyon matrisi birim matris değilse değişkenler arasında yüksek korelasyonların olduğu ifade edilir. Bu sayede ilgili değişkenlerin faktör analizi için uygun veriler taşıdığı ifade edilmektedir (Hair ve ark. 1998).

Veri setinin faktör analizine uygun olup olmadığını değerlendirmek için en yaygın kullanılan diğer bir yöntem Kaiser-Meyer-Olkin (KMO) örneklem yeterliliği ölçütüdür. Bu ölçüt incelenen korelasyon katsayılarının büyüklüğü ile kısmi korelasyon katsayılarını karşılaştırır. Oransal olarak sonuçlanan KMO oranının yüksek olması, veri setinin faktör analizine uygunluğu ile doğru orantılıdır. KMO oranının 0,5'in üzerinde olması gerekir. Oran ne kadar yüksek olursa veri setinin faktör analizi yapmak için o kadar iyi olduğu söylenebilir. KMO değeri 0,5-0,6 aralığında yeterli; 0,6-0,7 aralığında ise orta; 0,7-0,8 arasında iyi; 0,8-0,9 arasında çok iyi; 0,9 üzerinde ise mükemmel seviyede uygun olarak yorumlanır (Sharma ve Kumar, 2006).

3.2.3.2.(2) Faktörlerin Elde Edilmesi

Faktör analizinde faktörlerin en az sayıda olması dikkate alınmaktadır. Faktör sayısına karar verebilmek için birçok kriter söz konusudur. Özdeğer (eigenvalues) istatistiği olarak adlandırılan ilk kriter, ilgili özdeğer istatistiği 1' den büyük olan faktörlerin anlamlı kabul edildiğini; küçük olanların ise dikkate alınmaması gereken

faktörler olduğunu belirtir. Çizgi grafiği, toplam varyansın yüzdesi yöntemi, Joliffe kriteri, açıklanan varyans kriteri de ilgili veri setinden kaç adet faktör elde edilmesine yardımcı diğer ölçütler olarak adlandırılmaktadır (Dunteman, 1989).

3.2.3.2.(3) Faktörlerin Rotasyonu

Faktör matrislerinin daha kolay yorumlanabilmesi için rotasyon gerekmektedir (Dunteman, 1989). Faktörlerin rotasyona tabi tutulmasının, daha basit bir yapı ve anlaşılır bir matris elde edilmesinin gereksinimi sebebiyle faktörlerin önceden açıkladığı varyansın faktörler arasında yeniden dağıtma işlemi olarak da adlandırılabilir (Polat, 2012). Bu sayede toplam varyans değişmemekte, yalnızca her bir faktörün açıkladığı toplam varyansta değişiklik meydana gelmektedir (Albayrak, 2006). Ağırlıkların gücüne göre sıralanmış olan değişkenler analizin sonucunu vermektedir (Akgül, 2003). Rotasyonda kullanılan varimax tekniği en çok kullanılan teknik olmasının yanı sıra, equamax ve quartimax teknikleri de kullanılabilir. Rotasyonu yapılmış matriste orijinal değişkenlerle birlikte faktörü arasındaki korelasyonlar görülür. Mutlak değerce yüksek korelasyona sahip değişken, hangi faktör grubunun içerisinde görünüyorsa yakın ilişkidir anlamına gelmektedir.

3.2.3.2.(4) Örneklemin Genel Özellikleri ve Veri Ön İşleme

Veriler analiz edilmeden önce, bazı ön işlemlerden geçirilmesi gerekmektedir. Bunun sebebi verilerin standart bir yapıya uygun hale getirilerek, baskın veya önemsiz verilerle karşılaşılmasını engellemektir. Özellikle metin tipindeki verilerin sayısal tipteki veriler ile karşılaştırmada güçlük yaşanmaması için verilerin ön işleminden geçirilmesi bir zorunluluk haline gelmiştir. Eldeki verilerin mevcut halleri yerine yapılabilecek gruplamalar ile analizler yapıp yorumlanabilmektedir.

Veriler MS Excel üzerinde tutulmuş ve ilgili ön işlemlerden geçirilmiştir. Faktör analizi ile ilgili bütün uygulamalar IBM SPSS Statistics kullanılarak yapılmıştır.

3.2.3.2.(5) Eksik Veri Analizi

Eksik verileri incelerken expectation-maksimization yöntemi kullanılmıştır. Serinin ortalamasını alarak boş olan eksik verilerin yerlerine bu ortalama konmuştur. Significance değeri 0,05'ten büyük olduğundan kayıp verinin rassal dağıldığını göstermektedir.

3.2.3.3. Lojistik Regresyon

Lojistik regresyon analizinin kullanım amacı, en az değişkeni kullanarak en iyi uyuma sahip olacak şekilde bağımlı ile bağımsız değişkenler arasındaki ilişkiyi tanımlayabilen bir model kurmaktır. Lojistik regresyon analizi son yıllarda biyoloji, tıp, ekonomi, tarım, veterinerlik ve taşıma sahalarında sıkça kullanılmaktadır (Oğuzlar, 2010).

İki mümkün ve farklı değer içeren verilere iki değerli veriler denilmektedir. İki değerli değişkenler ile çalışan bir araştırmacının hedefi, bağımsız değişkenlerin koşullu bir kümesine bağımlı olarak, başarı veya başarısızlık olasılığının kestirilmesidir (Powers ve Xie, 2000). Neden sonuç ilişkilerinin ortaya konulması amacıyla yapılan çoğu sosyo-ekonomik araştırmada, incelenen değişkenlerden bazıları olumlu-olumsuz, başarılı-başarısız, evet-hayır, memnun-memnun değil şeklinde iki düzeyli verilerden oluşmaktadır. Bu türde bağımlı değişkenin iki düzeyli ya da çok düzeyli kategorik verilerden oluşması durumunda; bağımlı değişken ile bağımsız değişkenler arasındaki neden-sonuç ilişkisinin incelenmesinde lojistik regresyon analizi önemli bir yere sahiptir (Agresti, 1996).

Lojistik regresyon analizinde amaç, kategorik bağımlı değişkenin değerini tahmin etmektir. Bir başka deyişle iki ya da daha fazla gruba ilişkin “üyelik” tahmini yapmaktır. Buna göre analizin amaçlarından birinin sınıflandırma, diğerinin ise bağımlı ve bağımsız değişkenler arasındaki ilişkileri araştırmak olduğu ifade edilebilmektedir. Bağımlı değişken kategorik veri oluşturmakta ve kesikli değerler almaktayken, bağımsız değişkenlerin hepsinin veya bir kısmının sürekli ya da

kategorik değişkenler olmasına ilişkin bir zorunluluk bulunmamaktadır (Mertler ve Vannatta,2005).

Lojistik regresyon denkleminde P incelenen olayın gözlenme olasılığını göstermektedir. İncelenen bir olayın olasılığının kendi dışında kalan diğer olayların olasılığına oranına odds değeri denilmektedir. İncelenen iki farklı olayın odds değerlerinin birbirine oranına ise odds oranı denilmektedir. Lojistik regresyon denkleminde odds oranı, $\exp \beta$ olarak ifade edilmektedir. Odds oranı, bir olayın meydana gelme olasılığının meydana gelmeme olasılığına oranı olduğuna göre; $\exp \beta$

β_p y değişkeninin x_p değişkeninin etkisi ile kaç kat daha fazla ya da % kaç

oranında fazla gözlenme olasılığına sahip olduğunu belirtmektedir (Gujarati, 1999).

$$\frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}} \quad (32)$$

$$\frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}} \quad (33)$$

3.2.3.3.(1) Modelin Oluşturulması

Model oluşturmadaki amaç, en az bağımsız değişken yardımıyla bağımlı değişkendeki değişimi en fazla biçimde açıklamaktır. Eğer fazla sayıda değişken modele dahil edilirse standart hata tahminleri büyümektedir. Aynı zamanda fazla sayıda bağımsız değişken ile modelin kurulması ve geliştirilmesi süreçleri daha karmaşık hale getirmektedir (Oğuzlar, 2010).

3.2.3.3.(2) Modelin Uyum İyiliği

Modelin uyum iyiliği, bağımlı değişkeni açıklamak için oluşturulan en iyi modelin etkinliğinin bir ölçüsünü bize göstermektedir. Sınıflandırma tabloları, uyum iyiliği için kullanılan bir ölçüttür. Bu tablo, bağımlı değişkenin çapraz sınıflandırılması ile elde edilmektedir. Sınıflandırma tablosunda, bağımlı

değişkeninin gözlenen ve kestirilen lojistik olasılıklarından türetilen (0 veya 1) değerleri yer almaktadır. Türetilen bağımlı değişken değerlerinin elde edilmesinde bir kesim değerinin (c) tanımlanması zorunludur. Kestirilen olasılık değeri c' yi aştığında türetilen bağımlı değişken 1, aksi durumda 0 değerini almaktadır. Kesim değeri için en çok kullanılan değer 0,5'tir. Ele alınan kestirim değeri 0,5 değerini geçtiğinde 1, aksi durumda 0 grubuna atama yapılmaktadır.

Lojistik regresyon analizinde katsayıların elde edilmesinde en küçük kareler yönteminden farklı olarak, en yüksek olabilirlik yöntemi kullanılmakta ve dolayısıyla az sayıda gözlemle çalışılmaması gerekmektedir. Az sayıda gözlemle yapılan tahminlerde modelin güvenilirliği azalmaktadır. Bir başka deyişle lojistik regresyon modelinin güvenilirliğinin büyük örneklerle çalışılması durumunda arttığı belirtilebilmektedir.

Kararlı ve anlamlı sonuçların elde edilmesi, daha fazla veri gerektirmektedir. Standart bir regresyon analizi ya da diskriminant analizinde tipik olarak her bağımsız (yordayıcı) değişken için 20 kişilik bir grup alt sınır olarak kabul edilebilmektedir. Ancak lojistik regresyonda kararlı sonuçlara ulaşmak için, her bağımsız değişkende en az 50 kişilik grupların olmasına ihtiyaç vardır. Bir diğer önemli nokta, kategorilerin birbirini dışta tutması ve bağımsız olmasıdır (Cook, 2008).

3.2.4. Karar Destek Sistemi Yazılımının Akış Diyagramı

Tez çalışmasının önerilen modelleri kullanılarak vitamin D eksikliğinin tespiti için Eclipse ortamında Java programlama dili kullanılarak bir prototip yazılım önerilmiştir. Yazılımın nitelik seçimi ve sınıflandırma aşamalarında Weka programının 3.8.6. versiyonu çağrılarak ilgili değerler elde edilmiştir. Açık kaynaklı bir yazılım olan Weka, makine öğrenimi teknikleri geliştirebilmek ve bunları gerçek dünyadaki veri madenciliği sorunlarına uygulayabilmek için veri ön işleme ile çeşitli makine öğrenmesi algoritmalarının uygulanması için araçlar sağlamaktadır.

Şekil 3.12'de karar destek sistemi yazılımının akış diyagramı gösterilmektedir.

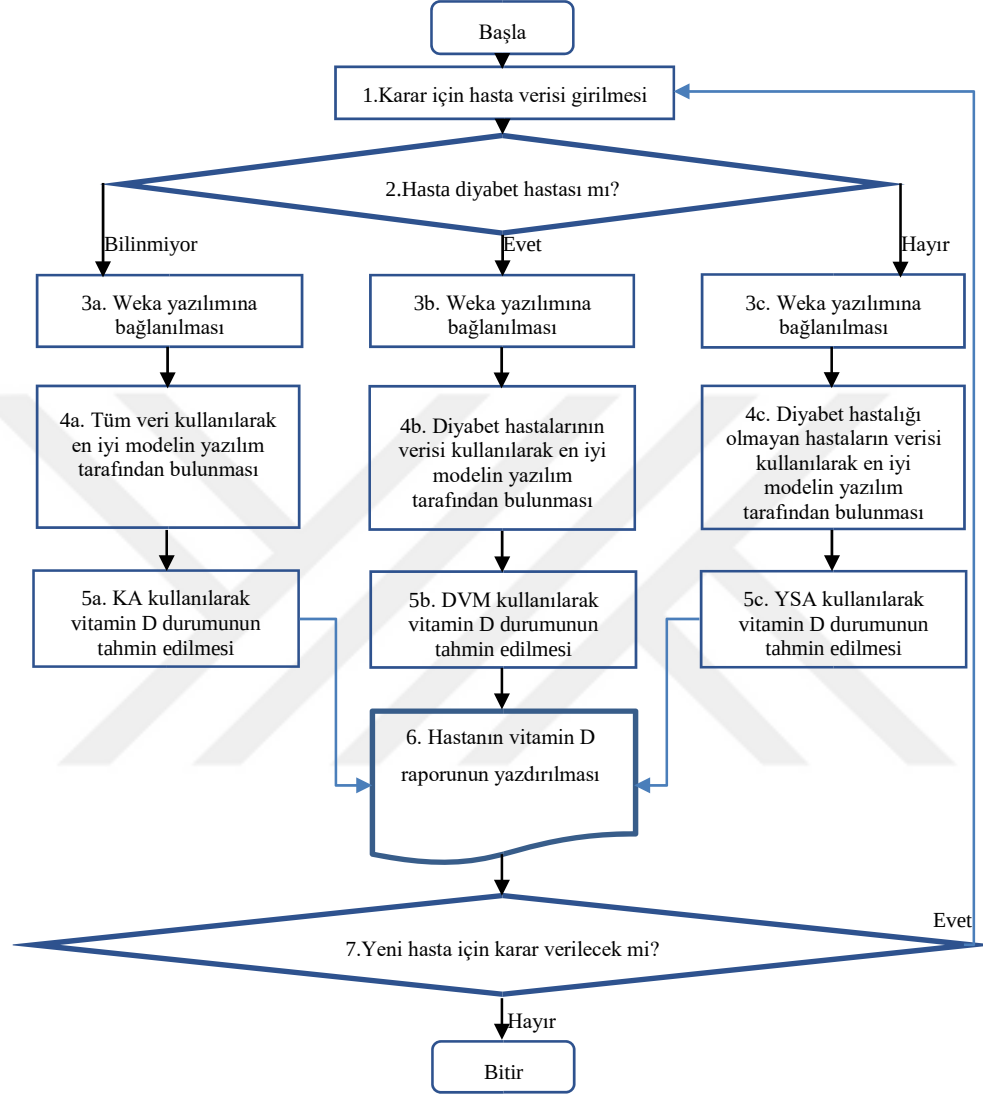
İlk olarak sisteme hasta bilgilerinin girişi yapılmaktadır.

İkinci aşamada hastanın diyabet durumu göz önünde bulundurularak yazılımın hangi modülünün çalışacağına karar verilmektedir. Üçüncü aşamada Weka yazılımına bağlanılmaktadır.

Dördüncü ve beşinci aşamalarda eğer hastanın diyabet durumu bilinmiyorsa, 684 hastanın tahlil sonuçlarının bulunduğu tüm veri üzerinde işlem yapılarak sonuç alınmaya çalışılmaktadır. Bu aşamada karar ağaçları yöntemi kullanılmakta ve sonuç almak için 36 farklı nitelik bilgisine ihtiyaç duyulmaktadır. Hastanın diyabet durumu biliniyorsa ve hasta diyabet hastasıysa, 406 hastanın tahlil sonuçlarının bulunduğu diyabet hastalarının verisi üzerinde işlem yapılarak sonuç alınmaya çalışılmaktadır. Bu aşamada destek vektör makineleri yöntemi kullanılmakta ve sonuç almak için 19 farklı nitelik bilgisine ihtiyaç duyulmaktadır. Hastanın diyabet durumu biliniyorsa ve hasta diyabet hastası değilse, 278 hastanın tahlil sonuçlarının bulunduğu diyabet hastalığı olmayan hastaların verisi üzerinde işlem yapılarak sonuç alınmaya çalışılmaktadır. Bu aşamada yapay sinir ağı yöntemi kullanılmakta ve sonuç almak için 6 farklı nitelik bilgisine ihtiyaç duyulmaktadır.

Altıncı aşamada hastanın diyabet durumuna göre yapılan işlemler sonucunda hastanın vitamin D raporu ekranda yazdırılabilmektedir.

Farklı bir hasta ile işleme devam edilmek istenirse işlemler ilk aşamadan başlanarak tekrarlanmaktadır.



Şekil 3.12. Karar destek sistemi yazılımının akış diyagramı





4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

4. BULGULAR VE TARTIŞMA

Bu çalışmada vitamin D eksik ve vitamin D yeterli olmak üzere sırasıyla 0 ve 1 şeklinde tanımlanan iki sınıf bulunmaktadır. Makine öğrenmesi teknikleri, Weka programının 3.8.6. versiyonunda destek vektör makineleri, k-en yakın komşuluk algoritması, karar ağaçları, naive bayes, yapay sinir ağları ve rastgele orman sınıflandırıcı modülleri uygulanarak gerçekleştirilmiştir.

4.1. Tüm Veri Üzerinde Yapılan Çalışmalar

Çalışmanın birinci aşamasının sonuçları bu kısımda verilmiştir.

4.1.1. Sınıflandırma Sonuçları

684 hastanın kan tahlilinin bulunduğu veri setinin sınıflandırma sonuçları çizelge 4.1’de gösterilmiştir. Yenilenen sınıflandırma sonuçları incelendiğinde karar ağaçları, yapay sinir ağları, rastgele orman, destek vektör makineleri, naive bayes ve k-en yakın komşuluk algoritması modellerinin doğruluk oranları sırasıyla %78,655; % 78,216; % 77,778; % 70,029; % 68,275 ve % 45,614 şeklinde sıralanmaktadır. Karar ağaçları veri setinde % 78,655 doğruluk oranıyla en iyi performansı ortaya koymuştur.

Çizelge 4.1. Sınıflandırma Sonuçları

Sınıflandırma Algoritması	Sınıflandırma Doğruluğu (%)
Karar Ağaçları (Random Committee)	78,655
Yapay Sinir Ağları (Multilayer Perceptron)	78,216
Rastgele Orman (Random Forest)	77,778
Destek Vektör Makineleri (LibSVM)	70,029
Naive Bayes (Bayes Net)	68,275
K-En Yakın Komşuluk Algoritması (IBk)	45,614

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Bu model için duyarlılık, özgüllük, kesinlik, f-measure ve kappa gibi diğer performans ölçütleri çizelge 4.2’de sunulmuştur.

Çizelge 4.2. Sınıflandırma Sonuçlarına Göre Oluşan Performans Ölçütleri

Sınıflandırma Algoritması	Duyarlılık	Özgüllük	Kesinlik	F-Measure	Kappa
Yapay Sinir Ağları	0,782	0,709	0,785	0,783	0,524
Karar Ağaçları	0,787	0,510	0,789	0,772	0,485
Rastgele Orman	0,778	0,439	0,792	0,755	0,447
Destek Vektör Makineleri	0,700	0,135	0,795	0,614	0,169
Naive Bayes	0,683	0,422	0,668	0,671	0,257
K-En Yakın Komşuluk Algoritması	0,456	0,738	0,575	0,445	0,036

Yapay sinir ağları bu performans ölçütlerinde en yüksek performansı göstermiş duyarlılık, özgüllük, kesinlik, f-measure ve kappa değerleri sırasıyla 0,782; 0,709; 0,785; 0,783 ve 0,524 olarak hesaplanmıştır. Bu nedenle her ne kadar sınıflandırma doğruluğunda karar ağaçları iyi sonuç vermişse de yapay sinir ağları diğer performans ölçütlerinde daha başarılı bir yöntem olmuştur.

Sınıflandırma sonuçlarının Weka çıktıları ek 2, ek 3, ek 4, ek 5, ek 6 ve ek 7’ de gösterilmiştir.

4.1.2.Nitelik Seçimi Sonuçları

Sınıflandırma yöntemlerinin sonuçlarını geliştirmek için nitelik seçici yöntemler veri seti üzerinde uygulanmıştır.

684 hastanın kan tahlilinin bulunduğu veri setinin nitelik seçimi sonuçları ek 8’de gösterilmiştir. Sonuçlarda bilgi kazanımı, ki-kare ve kazanım oranı yöntemleri uygulandığında benzer sonuçlar göstermiştir.

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

4.1.3. Nitelik Seçimi Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları

684 hastanın kan tahlilinin bulunduğu veri setinin nitelik seçimi sonuçlarına göre tekrarlanan sınıflandırma sonuçları çizelge 4.3'te gösterilmiştir.

Çizelge 4.3. Nitelik Seçimi Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları
(Parantez içindeki değerler seçilen nitelik sayılarını göstermektedir.)

Sınıflandırma Algoritması	Nitelik Seçimi Öncesi (%)	Bilgi Kazanımı (%)	Ki-Kare (%)	Kazanım Oranı (%)	Relief-f (%)
Karar Ağaçları	78,655 (59)	83,479 (36)	83,479 (36)	83,479 (36)	80,409 (58)
Rastgele Orman	77,778 (59)	81,286 (32)	81,286 (32)	81,286 (32)	78,509 (58)
Destek Vektör Makineleri	70,029 (59)	80,263 (32)	80,263 (32)	80,263 (32)	77,485 (15)
Yapay Sinir Ağları	78,216 (59)	78,801 (57)	78,801 (57)	78,801 (57)	78,216 (59)
K-En Yakın Komşuluk Algoritması	45,614 (59)	70,614 (7)	70,614 (7)	70,614 (7)	72,222 (21)
Naive Bayes	68,275 (59)	70,029 (5)	70,029 (5)	70,029 (5)	68,275 (58)

Görüldüğü gibi her bir nitelik seçimi yöntemi sonucunda elde edilen sınıflandırma sonuçları, nitelik seçimi yapılmadan önceki sonuçlara göre başarı oranlarında yükselme sağlamıştır. Nitelik seçimi yöntemleri başarı oranını yükseltmenin yanında nitelik sayılarını da düşürerek daha iyi sonuçlar elde edilmesini sağlamıştır. Bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri veri seti üzerinde benzer sonuçlar vermiştir.

Karar ağaçları sınıflandırma yöntemi bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri sonrası yapılan sınıflandırmada % 83,479 başarı

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

oranını sağlamıştır. Bu oranı sağlarken bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri nitelik sayısını 36'ya düşürmüştür.

Rastgele orman sınıflandırma sınıflandırma yöntemi bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri sonrası yapılan sınıflandırmada % 81,286 başarı oranını sağlamıştır. Bu oranı sağlarken bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri nitelik sayısını 32'ye düşürmüştür.

Destek vektör makineleri sınıflandırma yöntemi bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri sonrası yapılan sınıflandırmada % 80,263 başarı oranını sağlamıştır. Bu oranı yakalarken bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri nitelik sayısını 32'ye düşürmüştür.

Yapay sinir ağları sınıflandırma yöntemi bilgi kazanımı, ki-kare, kazanım oranı nitelik seçimi yöntemleri sonrası yapılan sınıflandırmada % 78,801 başarı oranını sağlamıştır. Bu oranı yakalarken bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri nitelik sayısını 57'ye düşürmüştür.

K-en yakın komşuluk algoritması sınıflandırma yöntemi relief-f nitelik seçimi yöntemi sonrası yapılan sınıflandırmada % 72,222 başarı oranını sağlamıştır. Bu oranı sağlarken bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri nitelik sayısını 21'e düşürmüştür.

Naive bayes sınıflandırma yöntemi bilgi kazanımı, ki-kare, kazanım oranı nitelik seçimi yöntemleri sonrası yapılan sınıflandırmada % 70,029 başarı oranını sağlamıştır. Bu oranı sağlarken bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri nitelik sayısını 5'e düşürmüştür.

Nitelik seçimi sonrası oluşan model için duyarlılık, özgüllük, kesinlik, fmeasure ve kappa gibi diğer performans ölçütleri çizelge 4.4'te sunulmuştur.

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Çizelge 4.4. Nitelik Seçimi Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçlarına Göre Oluşan Performans Ölçütleri

Sınıflandırma Algoritması	Duyarlılık	Özgüllük	Kesinlik	F-Measure	Kappa
Karar Ağaçları	0,835	0,633	0,837	0,828	0,612
Rastgele Orman	0,813	0,570	0,817	0,802	0,553
Destek Vektör Makineleri	0,803	0,430	0,848	0,776	0,497
Yapay Sinir Ağları	0,788	0,717	0,791	0,789	0,537
K-En Yakın Komşuluk Algoritması	0,722	0,253	0,749	0,670	0,268
Naive Bayes	0,700	0,342	0,685	0,673	0,260

Karar ağaçları bu performans ölçütlerinde en yüksek performansı göstermiş; duyarlılık, özgüllük, kesinlik, f-measure ve kappa değerleri sırasıyla 0,835; 0,633; 0,837; 0,828 ve 0,612 olarak hesaplanmıştır. Bu nedenle karar ağaçları optimum tahmin modeli olmuştur.

Nitelik seçimi sonuçlarına göre tekrarlanan sınıflandırma sonuçlarının Weka çıktıları ek 9, ek 10, ek 11, ek 12, ek 13 ve ek 14'te gösterilmiştir.

4.1.4. Kümeleme Sonuçları

Kümeleme sonuçları incelenmeden önce küme sayısını belirlemede karesel hatalar ve dirsek yöntemlerini kullanılmış ve optimum küme sayısını 5 olduğu tespit

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

edilmiştir. K-ortalama yöntemi uygulanırken farklı başlangıç noktaları çalıştırılmış ve hatanın en aza indirgendiği nokta belirlenmiştir. Kullanılan verilerde aralıkların birbirinden farklı olması nedeniyle normalizasyon uygulanan yöntemde $k=1,2,3,4,5,6,7,8,9,10$ değerleri için denemeler yapılmış ve sonuçlar çizelge 4.5'te gösterilmiştir.

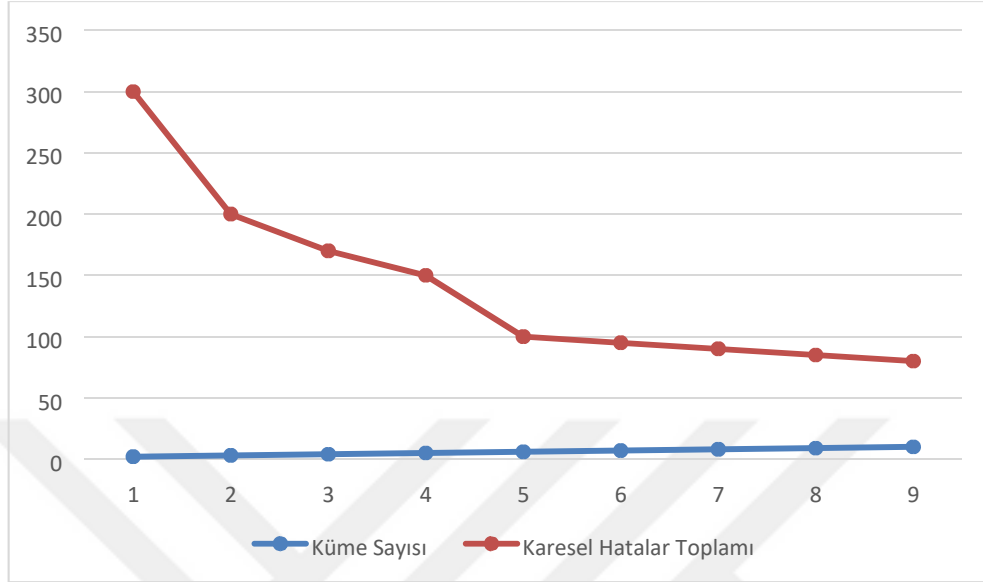
Çizelge 4.5. Optimum Küme Sayısının Belirlenmesi

Küme Sayısı	Karesel Hatalar Toplamı
2	785
3	772
4	763
5	496
6	490
7	486
8	411
9	409
10	376

Karesel hatalar toplamının sonuçları şekil 4.1'de işaretlenmiş ve dirsek yöntemine göre optimum küme sayısı belirlenmeye çalışılmıştır. Şekil üzerinde işaretlenmiş toplamlar arasındaki farkın azalmaya başladığı nokta dirsek noktasıdır. Şekil incelendiğinde eğimin 5. noktadan sonra azaldığı görülmektedir. Dolayısıyla optimum küme sayısı 5 olarak belirlenmiştir.

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ



Şekil 4.1. Optimum Küme Sayısının Belirlenmesi

684 hastanın kan tahlilinin bulunduğu veri setinin k-ortalama yöntemiyle tamamlanan kümeleme sonucu ek 15'te gösterilmiştir.

Yapılan kümeleme işlemi sonucunda diyabet hastalarında vitamin D eksikliğini etkileyen faktörler glukoz, HbA1c, HbA1c(IFCC), albümin, total kolesterol, trigliserid ve kalsiyum olarak belirlenmiştir. Glukoz, HbA1c, HbA1c(IFCC), albümin, total kolesterol, trigliserid ve kalsiyum'un normal değerlerden yüksek olduğu durum vitamin D değerinin normal değer altında çıkmasına neden olmuştur. Ayrıca bu koşulların sadece kadın hastalarda geçerli olduğu görülmüştür.

2 ve 3 numaralı küme incelendiğinde vitamin D değeri normal değerlerden düşük olan hastaların glukoz, HbA1c, HbA1c(IFCC), albümin, total kolesterol, trigliserid ve kalsiyum değerlerinin yüksek olmasının yanı sıra CRP değerinin de normal değerlerden yüksek olduğu tespit edilmiştir.

3 ve 4 numaralı küme incelendiğinde vitamin D değeri normal değerlerden düşük olan kadın hastaların glukoz, HbA1c, HbA1c(IFCC), albümin,

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

total kolesterol, trigliserid ve kalsiyum değerlerinin yüksek olmasının yanı sıra kreatinin değerinin normal değerlerden düşük olduğu tespit edilmiştir.

Diyabet hastası olmayan hastalarda vitamin D eksikliğini etkileyen faktörler albümin, trigliserid, kalsiyum ve CRP olarak belirlenmiştir. Albümin, trigliserid, kalsiyum ve CRP 'nin normal değerlerden yüksek olduğu durum vitamin D değerinin normal değerin altında çıkmasına neden olmuştur. Ayrıca bu koşulların sadece erkek hastalarda geçerli olduğu görülmüştür.

4.1.5. Kümeleme Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları

684 hastanın kan tahlilinin bulunduğu veri setinin kümeleme sonuçlarına göre tekrarlanan sınıflandırma sonuçları çizelge 4.6'da gösterilmiştir.

Çizelge 4.6. Kümeleme Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları

Sınıflandırma Algoritması	Kümeleme Öncesi (%)	Kümeleme Sonrası (%)
Destek Vektör Makineleri	70,029	78,947
Karar Ağaçları	78,655	71,053
Rastgele Orman	77,778	70,322
Naive Bayes	68,275	66,521
Yapay Sinir Ağları	78,216	64,181
K-En Yakın Komşuluk Algoritması	45,614	47,515

Görüldüğü gibi kümeleme sonucuna göre tekrarlanan destek vektör makineleri ve k-en yakın komşuluk algoritması ile elde edilen sınıflandırma sonuçları, kümeleme yapılmadan önceki sonuçlara göre başarı oranlarında yükselme sağlamıştır. Kümeleme sonucu başarı oranını yükseltmenin yanında nitelik sayılarını da düşürerek daha iyi sonuçlar elde edilmesini sağlamıştır. Kümeleme sonucunda nitelik sayısı 7'ye düşmüştür.

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Destek vektör makineleri sınıflandırma yöntemi kümeleme sonrası yapılan sınıflandırmada % 78,947 başarı oranını sağlamıştır.

K-en yakın komşuluk algoritması sınıflandırma yöntemi kümeleme sonrası yapılan sınıflandırmada % 47,515 başarı oranını sağlamıştır.

Kümeleme sonrası oluşan model için duyarlılık, özgüllük, kesinlik, fmeasure ve kappa gibi diğer performans ölçütleri çizelge 4.7’de sunulmuştur.

Çizelge 4.7. Kümeleme Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçlarına Göre Oluşan Performans Ölçütleri

Sınıflandırma Algoritması	Duyarlılık	Özgüllük	Kesinlik	F-Measure	Kappa
Destek Vektör Makineleri	0,789	0,481	0,800	0,771	0,483
Karar Ağaçları	0,711	0,414	0,697	0,693	0,306
Rastgele Orman	0,703	0,363	0,689	0,679	0,273
Naive Bayes	0,665	0,392	0,649	0,652	0,215
Yapay Sinir Ağları	0,642	0,405	0,629	0,634	0,179
K-En Yakın Komşuluk Algoritması	0,475	0,363	0,510	0,487	-0,079

Destek vektör makineleri bu performans ölçütlerinde en yüksek performansı göstermiş; duyarlılık, özgüllük, kesinlik, f-measure ve kappa değerleri sırasıyla 0,789; 0,481; 0,800; 0,771 ve 0,483 olarak hesaplanmıştır. Bu nedenle destek vektör makineleri optimum tahmin modeli olmuştur.

Kümeleme sonuçlarına göre tekrarlanan sınıflandırma sonuçlarının Weka çıktıları ek 16, ek 17, ek 18, ek 19, ek 20 ve ek 21’de gösterilmiştir.

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

4.1.6. Faktör Analizi Sonuçları

684 hastanın kan tahlilinin bulunduğu veri setiyle tamamlanan faktör analizi sonuçları aşağıda listelenmiştir.

Öncelikle veri setinin faktör analizi için uygun olup olmadığı belirlemek için Kaiser-Meyer-Olkin (KMO) örneklem yeterliliği ölçütü incelemesi yapılmıştır. KMO oranı 0,5'in üzerinde olmalıdır. Oran ne kadar yüksek olursa veri seti faktör analizi yapmak için o kadar uygundur. Çizelge 4.8'de görüldüğü gibi KMO oranı 0,513'tür. Bu sonuç faktör analizine devam etmek için yeterlidir.

Çizelge 4.8. KMO and Bartlett's Test

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0,513
Bartlett's Test of Sphericity	Approx. Chi-Square	13512,430
	df	1711
	Sig.	,000

Faktör analizlerinin yorumlanmasından önce örneklem büyüklüğüne göre faktör değerlerinin değerlendirmeye alınabileceği alt limitler literatürden elde edilmiştir. Yapılan çalışmalara göre 50 örnek büyüklüğüne sahip olan bir veri setinde, en az 0,75 faktör yükleme karşılanmalıdır. Örnek büyüklüğüne göre minimum faktör yükleme değerleri çizelge 4.9'da verilmiştir. Bu değerler ışığında örneklem büyüklüğü 350'den fazla olduğu için veriden çıkarmamız gereken nitelik değeri 0,3'ten düşük olmalıdır.

Çizelge	4.9.	Minimum	Faktör	Yükleme	Değerleri
Minimum Faktör Yükleme Değeri			Yeterli Örnek Büyüklüğü		
0,3			350		
0,35			250		
0,4			200		

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

0,45	150
0,5	120
0,55	100
0,6	85
0,65	70
0,7	60
0,75	50

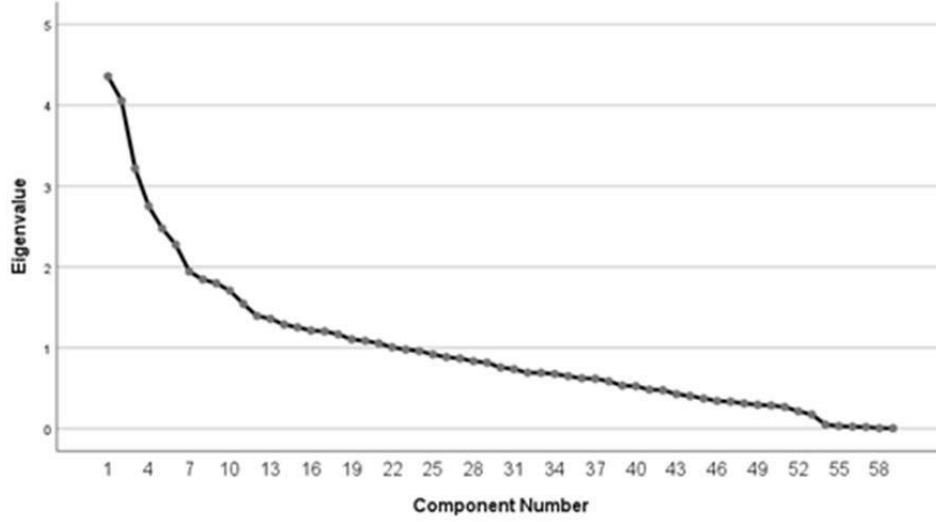
Ek 22’de ortak yükler incelediğinde veri setinde nitelik değeri 0,3’ten küçük nitelik olmadığı görülmüştür. Dolayısıyla veri setinden nitelik çıkarılmasına gerek olmadığı anlaşılmıştır.

Ek 23’teki açıklanan toplam varyans incelendiğinde faktör sayısı belirlenirken özdeğer istatistiği 1’den büyük olan faktörler anlamlı olarak belirlenmiştir. Faktör analizi sonucunda 22 faktörün toplam varyansı % 67,890 oranında açıkladığı görülmüştür.

Şekil 4.2’de gösterildiği gibi faktör sayısı belirlenirken faktör analizi çizgi grafiğinde eğimin kaybolmaya başladığı noktanın işaret ettiği sayıda faktör belirlenebilir.

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ



Şekil 4.2. Yamaç Grafiği

Ek 24'te döndürülmüş bileşen matrisi incelemesi sonucuna göre faktör grupları belirlenmiştir. İnceleme sonucunda nitelikler arasında korelasyonlu olanlar belirlenmiş ve nitelik sayısı azaltılmıştır. Çizelge 4.10'da gösterilen faktör gruplarına göre vitamin D'nin de içinde bulunduğu grupta ayrıca albümin, CRP ve insülin de bulunmaktadır.

Çizelge 4.10. Faktör Grupları

Grup 1	NE#	Grup 8	BA	Grup 15	MO
	WBC		BA#		MO#
	NE	Grup 9	AST	Grup 16	Albümin
Grup 2	HCT		ALT		CRP
	HGB		CK		25-Hidroksi Vitamin D

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

	RBC	Grup 10	Total Kolesterol		İnsülin
Grup 3	MCV		LDL	Grup 17	LY#
	MCH		Trigliserid	Grup 18	Kalsiyum
	RDWCV	Grup 11	Kreatinin		LDH
Grup 4	GGT		Üre	Grup 19	Potasyum
	ALP		Cinsiyet		Bilirübin
	SODYUM		Sedimentasyon		Fosfor
	T4		Ürik asit	Grup 20	Magnezyum
Grup 5	YAŞ	Grup 12	Glukoz		TSH
	T3		HbA1c(IFCC)	Grup 21	ASO
	Diyabet		HbA1c		LY
Grup 6	PLT	Grup 13	Demir Bağlama	Grup 22	MCHC
	PCT		Ferritin		MPV
	PDW		Demir		
Grup 7	EO#	Grup 14	VB12		
	EO		HDL		

4.1.7. Lojistik Regresyon Analizi Sonuçları

684 hastanın kan tahlilinin bulunduğu veri setiyle tamamlanan lojistik regresyon analizi sonuçları aşağıda listelenmiştir.

Çizelge 4.11 incelendiğinde 18 hastanın verisinin işleme alınmadığı görülmektedir.

Çizelge 4.11. Vaka İşleme Özeti

Unweighted Cases ^a	N	Percent
-------------------------------	---	---------

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Selected Cases	Included in Analysis	666	97,4
	Missing Cases	18	2,6
	Total	684	100,0
Unselected Cases		0	,0
Total		684	100,0

Çizelge 4.12’de sınıflandırma tablosu 1 incelendiğinde gözlenen durum (gerçek durum) ve ilk sınıflandırma sonuçları görülmektedir. Programın öncelikle tüm örnekleri düşük vitamin D kategorisinde sınıflandırdığı ve bu haliyle doğru sınıflandırma yüzdesinin % 64,400 olduğu görülmektedir.

Çizelge 4.12. Sınıflandırma Tablosu 1

	Observed		Predicted		
			25-Hidroksi Vitamin D		Percentage Correct
			0	1	
Step 0	25-Hidroksi Vitamin D	0	429	0	100,0
		1	237	0	,0
	Overall Percentage				64,4

Çizelge 4.13’te model özeti incelendiğinde bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin oranını gösteren Nagelkerke R^2 istatistiğinin % 43,5 olarak elde edildiği görülmüştür. Çeşitli kaynaklarda R^2 istatistiklerinin 0,20 ile 0,40 arasında çıkması yeterli olduğu açıklanmaktadır (Alpar, 2011).

Çizelge 4.13. Model Özeti

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	613,317 ^a	,317	,435

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Çizelge 4.14'te sınıflandırma tablosu 2'de lojistik regresyon modeli sonucunda elde edilen sınıflandırma sonucu incelendiğinde ise, doğru sınıflandırılma oranının % 81,100 olduğu görülmektedir.

Çizelge 4.14. Sınıflandırma Tablosu 2

	Observed		Predicted		
			25-Hidroksi Vitamin D		Percentage Correct
			0	1	
Step 1	25-Hidroksi Vitamin D	0	394	35	91,8
		1	91	146	61,6
	Overall Percentage				81,1

Ek 25'te denklemdaki değişkenler incelediğinde başlangıç modelini oluşturan sabit terim, sabit terime ilişkin standart hata, değişkenin anlamlılığını test eden Wald istatistiği, Wald istatistiğinin serbestlik derecesi ve anlamlılık düzeyi ile $\text{Exp}(b)$, yani üstel lojistik regresyon katsayısı yer almaktadır. Üstel lojistik regresyon katsayısı odds oranını temsil etmektedir. Significance değeri bir karşılaştırmada “istatistiksel anlamlı fark vardır” kararı verileceği zaman yapılacak olası hata miktarını gösterir. Fisher tarafından bu hatanın maksimum kabul edilebilir düzeyi 0,05 olarak önerilmiş ve kabul görmüştür. Bir test sonucunda bulunan Significance değeri 0,05'in altında bir değer ise karşılaştırma sonucunda anlamlı farklılık bulunduğu anlamına gelir (Önder ve Cebeci, 2002).

Ek 25'te denklemdaki değişkenlerden elde edilen sonuçlara göre; diyabet değişkeni bakımından vitamin D eksikliği oddsunda 12,680 kat artışa yol açtığı görülmektedir. Cinsiyet değişkeni bakımından vitamin D eksikliği oddsunda 15,668 kat artışa yol açtığı görülmektedir. Yaş değişkeni bakımından vitamin D eksikliği oddsunda 1,024 kat artışa yol açtığı görülmektedir. Ürik asit değişkeni bakımından vitamin D eksikliği oddsunda 1,342 kat artışa yol açtığı görülmektedir. ALT değişkeni bakımından vitamin D eksikliği oddsunda 0,950 kat azalışa yol açtığı görülmektedir. GGT değişkeni bakımından vitamin D eksikliği oddsunda 1,021 kat

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

artışa yol açtığı görülmektedir. Albümin değişkeni bakımından vitamin D eksikliği oddsunda 1,099 kat artışa yol açtığı görülmektedir. Trigliserid değişkeni bakımından vitamin D eksikliği oddsunda 0,996 kat azalışa yol açtığı görülmektedir. Fosfor değişkeni bakımından vitamin D eksikliği oddsunda 61,607 kat artışa yol açtığı görülmektedir. CK değişkeni bakımından vitamin D eksikliği oddsunda 1,004 kat artışa yol açtığı görülmektedir. PDW değişkeni bakımından vitamin D eksikliği oddsunda 1,044 kat artışa yol açtığı görülmektedir. Sedimentasyon değişkeni bakımından vitamin D eksikliği oddsunda 0,944 kat azalışa yol açtığı görülmektedir. TSH değişkeni bakımından vitamin D eksikliği oddsunda 0,849 kat azalışa yol açtığı görülmektedir.

Lojistik regresyon modelinin denklemi şu şekilde gösterilmektedir.

$$\begin{aligned} Y &= -42,112 + 2.54 \cdot \text{DIA} + 2.752 \cdot \text{CİNSİYET} \\ &+ 0.023 \cdot \text{YAŞ} + 0.295 \cdot \text{SMEANÜRİKASİT} - \\ &0.051 \cdot \text{SMEANALT} - 0.021 \cdot \text{SMEANGGT} \\ &0.094 \cdot \text{SMEANALBÜMİN} - 0.004 \cdot \text{SMEANTRİGLİSERİD} \\ &4.121 \cdot \text{SMEANFOSFOR} - 0.004 \cdot \text{SMEANCK} \\ &0.043 \cdot \text{SMEANPDW} - \\ &0.057 \cdot \text{SMEANSEDİMENTASYON} - 0.163 \cdot \text{SMEANTSH} \end{aligned}$$

4.2. Diyabet Hastalarının Veri Seti Üzerinde Yapılan Çalışmalar

Çalışmanın ikinci aşamasının sonuçları bu kısımda verilmiştir.

4.2.1. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Sınıflandırma

Sonuçları

Diyabet hastalarının vitamin D eksikliğini tahmin etme yönteminin etkinliğini değerlendirmek için diyabet hastalarının vitamin D veri seti üzerinde deneyler yapılmıştır.

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

406 hastanın kan tahlilinin bulunduğu veri setinin sınıflandırma sonuçları çizelge 4.15'te gösterilmiştir. Destek vektör makineleri, karar ağaçları, rastgele orman, yapay sinir ağları, naive bayes ve k-en yakın komşuluk algoritması modellerinin doğruluk oranları sırasıyla % 98,030; % 96,800; % 95,320; % 81,773; % 73,150 ve % 41,870 şeklindedir. Destek vektör makineleri, karar ağaçları, rastgele orman, yapay sinir ağları, naive bayes ve k-en yakın komşuluk algoritması sınıflandırıcıları için sınıflandırma doğruluk oranlarını karşılaştıran çalışma, diyabet hastalarının vitamin D eksikliğini tahmin etmek için optimum tahmin modelini seçmiştir. Bu sınıflandırıcılar arasında Destek vektör makineleri için sınıflandırma doğruluk oranı tüm sınıflandırma yöntemlerinden daha yüksek olduğu tespit edilmiştir.

Çizelge 4.15. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Sınıflandırma Sonuçları

Sınıflandırma Algoritması	Sınıflandırma Doğruluğu (%)
Destek Vektör Makineleri (LibSVM)	98,030
Karar Ağaçları (Random Committee)	96,800
Rastgele Orman (Random Forest)	95,320
Yapay Sinir Ağları (Multilayer Perceptron)	81,773
Naive Bayes (Bayes Net)	73,150
K-En Yakın Komşuluk Algoritması (IBk)	41,870

Bu model için duyarlılık, özgüllük, kesinlik, f-measure ve kappa gibi diğer performans ölçütleri çizelge 4.16'da sunulmuştur.

Çizelge 4.16. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Sınıflandırma Sonuçlarına Göre Oluşan Performans Ölçütleri

Sınıflandırma Algoritması	Duyarlılık	Özgüllük	Kesinlik	F-Measure	Kappa
---------------------------	------------	----------	----------	-----------	-------

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Destek Vektör Makineleri	0,980	0,926	0,981	0,980	0,948
Karar Ağaçları	0,968	0,889	0,969	0,967	0,915
Rastgele Orman	0,953	0,824	0,956	0,952	0,873
Yapay Sinir Ağları	0,818	0,815	0,843	0,825	0,576
Naive Bayes	0,732	0	0,538	0,620	-0,005
K-En Yakın Komşuluk Algoritması	0,419	0,583	0,583	0,442	-0,040

Destek vektör makineleri bu performans ölçütlerinde en yüksek performansı göstermiş; duyarlılık, özgüllük, kesinlik, f-measure ve kappa değerleri sırasıyla 0,980; 0,926; 0,981; 0,980 ve 0,948 olarak hesaplanmıştır. Bu nedenle destek vektör makineleri diyabet hastalarının veri setinde optimum tahmin modeli olmuştur.

Diyabet hastalarının veri seti üzerinde çalışılan sınıflandırma sonuçlarının Weka çıktıları ek 26, ek 27, ek 28, ek 29, ek 30 ve ek 31’de gösterilmiştir.

4.2.2. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Nitelik Seçimi Sonuçları

Sınıflandırma yöntemlerinin sonuçları her ne kadar % 90 başarı oranının üzerinde olsa da sonraki aşamada bu oranlar da geliştirilmeye çalışılmıştır.

406 hastanın kan tahlilinin bulunduğu veri setinin nitelik seçimi sonuçları ek 32’de gösterilmiştir. Sonuçlarda bilgi kazanımı, ki-kare ve kazanım oranı yöntemleri benzer sonuçlar göstermiştir.

4.2.3. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Nitelik Seçimi Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları

406 hastanın kan tahlilinin bulunduğu veri setinin nitelik seçimi sonuçlarına göre tekrarlanan sınıflandırma sonuçları çizelge 4.17’de gösterilmiştir.

4. BULGULAR VE TARTIŞMA
Uğur Engin EŞSİZ

Çizelge 4.17. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Nitelik Seçimi Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları (Parantez içindeki değerler seçilen nitelik sayılarını göstermektedir.)

Sınıflandırma Algoritması	Nitelik Seçimi Öncesi (%)	Bilgi Kazanımı (%)	Ki-Kare (%)	Kazanım Oranı (%)	Relief-f (%)
Destek Vektör Makineleri	98,030 (58)	98,030 (31)	98,030 (31)	98,030 (31)	98,030 (19)
Karar Ağaçları	96,800 (58)	97,537 (42)	97,537 (42)	97,537 (42)	97,040 (32)
Rastgele Orman	95,320 (58)	97,044 (51)	97,044 (51)	97,044 (51)	96,798 (42)
Yapay Sinir Ağları	81,773 (58)	85,222 (23)	85,222 (23)	85,222 (23)	85,222 (29)
K-En Yakın Komşuluk Algoritması	41,870 (58)	81,281 (4)	81,281 (4)	81,281 (4)	76,600 (25)
Naive Bayes	73,150 (58)	73,400 (1)	73,400 (1)	73,400 (1)	73,400 (53)

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Görüldüğü gibi nitelik seçimi yöntemleri sonuçlarından elde edilen sınıflandırma sonuçları, nitelik seçimi yapılmadan önceki sonuçlara göre başarı oranlarında, destek vektör makineleri yöntemi hariç yükselme sağlamıştır. Nitelik seçimi yöntemleri başarı oranını yükseltmenin yanında nitelik sayılarını da düşürerek daha iyi sonuçlar elde edilmesini sağlamıştır. Bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri veri seti üzerinde benzer sonuçlar vermiştir.

Destek vektör makineleri sınıflandırma yöntemi her bir nitelik seçimi yöntemi sonrası yapılan sınıflandırmada % 98,030 başarı oranını sağlamıştır. Bu oranı yakalarken bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri nitelik sayısını 31'e; relief-f nitelik seçimi yöntemi ise nitelik sayısını 19'a düşürmüştür.

Karar ağaçları sınıflandırma yöntemi bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri sonrası yapılan sınıflandırmada % 97,537 başarı oranını sağlamıştır. Bu oranı sağlarken bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri nitelik sayısını 42'ye düşürmüştür.

Rastgele Orman sınıflandırma yöntemi bilgi kazanımı, kikare ve kazanım oranı nitelik seçimi yöntemleri sonrası yapılan sınıflandırmada % 97,044 başarı oranını sağlamıştır. Bu oranı sağlarken bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri nitelik sayısını 51'e düşürmüştür.

Yapay sinir ağları sınıflandırma yöntemi her bir nitelik seçimi yöntemi sonrası yapılan sınıflandırmada % 85,222 başarı oranını sağlamıştır. Bu oranı yakalarken bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri nitelik sayısını 23'e; relief-f nitelik seçimi yöntemi ise nitelik sayısını 29'a düşürmüştür.

K-en yakın komşuluk algoritması sınıflandırma yöntemi bilgi kazanımı, kikare ve kazanım oranı nitelik seçimi yöntemleri sonrası yapılan sınıflandırmada % 81,281 başarı oranını sağlamıştır. Bu oranı sağlarken bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri nitelik sayısını 4'e düşürmüştür.

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Naive bayes sınıflandırma yöntemi bilgi kazanımı, ki-kare, kazanım oranı ve relief-f nitelik seçimi yöntemleri sonrası yapılan sınıflandırmada % 73,400 başarı oranını sağlamıştır. Bu oranı sağlarken bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri nitelik sayısını 1'e; relief-f nitelik seçimi yöntemi ise nitelik sayısını 53'e düşürmüştür.

Nitelik seçimi sonrası oluşan model duyarlılık, özgüllük, kesinlik, fmeasure ve kappa gibi diğer performans ölçütleri çizelge 4.18'de sunulmuştur.

Çizelge 4.18. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Nitelik Seçimi Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçlarına Göre Oluşan Performans Ölçütleri

Sınıflandırma Algoritması	Duyarlılık	Özgüllük	Kesinlik	F-Measure	Kappa
Destek Vektör Makineleri	0,980	0,926	0,981	0,980	0,948
Karar Ağaçları	0,975	0,926	0,975	0,975	0,936
Rastgele Orman	0,970	0,889	0,972	0,970	0,921
Yapay Sinir Ağları	0,852	0,843	0,869	0,857	0,648
K-En Yakın Komşuluk Algoritması	0,813	0,648	0,813	0,813	0,521
Naive Bayes	0,734	0	?	?	0

Destek vektör makineleri bu performans ölçütlerinde en yüksek performansı göstermiş; duyarlılık, özgüllük, kesinlik, f-measure ve kappa değerleri sırasıyla 0,980; 0,926; 0,981; 0,980 ve 0,948 olarak hesaplanmıştır. Bu nedenle destek vektör makineleri diyabet hastalarının veri setinde optimum tahmin modeli olmuştur.

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Diyabet hastalarının veri seti üzerinde çalışılan nitelik seçimi sonuçlarına göre tekrarlanan sınıflandırma sonuçlarının Weka çıktıları ek 33, ek 34, ek 35, ek 36, ek 37 ve ek 38’de gösterilmiştir.

4.2.4. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Kümeleme Sonuçları

406 hastanın kan tahlilinin bulunduğu veri setinin k-ortalama kümeleme işlemi sonucu ek 39’da gösterilmiştir.

Diyabet hastalarında vitamin D eksikliğini etkileyen faktörler glukoz, HbA1c, HbA1c(IFCC), total kolesterol ve trigliserid şeklinde belirlenmiştir. Glukoz, HbA1c, HbA1c(IFCC), total kolesterol ve trigliserid’in normal değerlerden yüksek olduğu durumda vitamin D değerinin normal değerin altında çıkmasına neden olmuştur. Ayrıca bu koşulların hem kadın hem de erkek hastalarda geçerli olduğu görülmüştür.

1 numaralı küme incelendiğinde vitamin D değeri normal değerlerden düşük erkek hastaların glukoz, HbA1c, HbA1c(IFCC), total kolesterol ve trigliserid değerlerinin normal değerlerden yüksek olduğu tespit edilmiştir.

2 numaralı küme incelendiğinde vitamin D değeri normal değerlerden düşük kadın hastaların glukoz, HbA1c, HbA1c(IFCC), total kolesterol ve trigliserid değerlerinin yanı sıra CRP ve MCHC değerlerinin de normal değerlerden yüksek olduğu tespit edilmiştir.

3 numaralı küme incelendiğinde vitamin D değeri normal değerlerden düşük kadın hastaların HbA1c, HbA1c(IFCC), total kolesterol, LDL ve trigliserid değerlerinin normal değerlerden yüksek olduğu tespit edilmiştir.

4 numaralı küme incelendiğinde vitamin D değeri normal değerlerden düşük kadın hastaların HbA1c(IFCC) ve total kolesterol değerlerinin yanı sıra CK, CRP, PLT, PCT, EO1, EO2 ve TSH değerlerinin de normal değerlerden yüksek olduğu tespit edilmiştir.

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

4.2.5. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Kümeleme Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları

406 hastanın kan tahlilinin bulunduğu veri setinin kümeleme sonuçlarına göre tekrarlanan sınıflandırma sonuçları çizelge 4.19’da gösterilmiştir.

Çizelge 4.19. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Kümeleme Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları

Sınıflandırma Algoritması	Kümeleme Öncesi (%)	Kümeleme Sonrası (%)
Destek Vektör Makineleri	98,030	95,567
Karar Ağaçları	96,800	91,872
Rastgele Orman	95,320	91,133
Naive Bayes	73,150	73,399
K-En Yakın Komşuluk Algoritması	41,870	72,660
Yapay Sinir Ağları	81,773	71,675

Görüldüğü gibi kümeleme sonucuna göre tekrarlanan naive bayes ve k-en yakın komşuluk algoritması ile elde edilen sınıflandırma sonuçları, kümeleme yapılmadan önceki sonuçlara göre başarı oranlarında yükselme sağlamıştır. Kümeleme sonucu başarı oranını yükseltmenin yanında nitelik sayılarını da düşürerek daha iyi sonuçlar elde edilmesini sağlamıştır. Kümeleme sonucunda nitelik sayısı 5’e düşmüştür.

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Naive bayes sınıflandırma yöntemi kümeleme sonrası yapılan sınıflandırmada % 73,399 başarı oranını sağlamıştır.

K-en yakın komşuluk algoritması sınıflandırma yöntemi kümeleme sonrası yapılan sınıflandırmada % 72,660 başarı oranını sağlamıştır.

Kümeleme sonrası oluşan model için duyarlılık, özgüllük, kesinlik, fmeasure ve kappa gibi diğer performans ölçütleri çizelge 4.20’de sunulmuştur.

Çizelge 4.20. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Kümeleme Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçlarına Göre Oluşan Performans Ölçütleri

Sınıflandırma Algoritması	Duyarlılık	Özgüllük	Kesinlik	F-Measure	Kappa
Destek Vektör Makineleri	0,956	0,852	0,957	0,955	0,882
Karar Ağaçları	0,919	0,843	0,918	0,919	0,791
Rastgele Orman	0,911	0,759	0,910	0,909	0,762
Naive Bayes	0,734	0	?	?	0
K-En Yakın Komşuluk Algoritması	0,727	0,444	0,720	0,723	0,281
Yapay Sinir Ağları	0,717	0,065	0,630	0,639	0,024

Destek vektör makineleri bu performans ölçütlerinde en yüksek performansı göstermiş; duyarlılık, özgüllük, kesinlik, f-measure ve kappa değerleri sırasıyla

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

0,956; 0,852; 0,957; 0,955 ve 0,882 olarak hesaplanmıştır. Bu nedenle destek vektör makineleri optimum tahmin modeli olmuştur.

Diyabet hastalarının veri seti üzerinde çalışılan kümeleme sonuçlarına göre tekrarlanan sınıflandırma sonuçlarının Weka çıktıları ek 40, ek 41, ek 42, ek 43, ek 44 ve ek 45'te gösterilmiştir.

4.2.6. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Faktör Analizi Sonuçları

406 hastanın kan tahlilinin bulunduğu veri setiyle tamamlanan faktör analizi sonuçları aşağıda listelenmiştir.

Öncelikle veri setinin faktör analizi için uygun olup olmadığı belirlemek için KMO örneklem yeterliliği ölçütü incelemesi yapılmıştır. KMO oranı 0,5'in üzerinde olmalıdır. Oran ne kadar yüksek olursa veri seti faktör analizi yapmak için o kadar uygundur. Çizelge 4.21'den görüldüğü gibi KMO oranı 0,494'tür. Bu sonuç faktör analizine devam etmek için yeterli değildir.

Çizelge 4.21. KMO and Bartlett's Test

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,494
Bartlett's Test of Sphericity	Approx. Chi-Square	13212,486
	df	1711
	Sig.	,000

4.2.7. Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Lojistik Regresyon Sonuçları

406 hastanın kan tahlilinin bulunduğu veri setiyle tamamlanan lojistik regresyon analizi sonuçları aşağıda listelenmiştir.

Çizelge 4.22'de sınıflandırma tablosu 1 incelendiğinde gözlenen durum (gerçek durum) ve ilk sınıflandırma sonuçları görülmektedir. Programın öncelikle

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

tüm örnekleri düşük vitamin D kategorisinde sınıflandırdığı ve bu haliyle doğru sınıflandırma yüzdesinin % 73,400 olduğu görülmektedir.

Çizelge 4.22. Sınıflandırma Tablosu 1

	Observed		Predicted		
			25-Hidroksi Vitamin D		Percentage Correct
			0	1	
Step 0	25-Hidroksi Vitamin D	0	298	0	100,0
		1	108	0	,0
	Overall Percentage				73,4

Çizelge 4.23'te model özeti incelendiğinde bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin oranını gösteren Nagelkerke R² istatistiğinin % 27,2 olarak elde edildiği görülmüştür.

Çizelge 4.23. Model Özeti

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	386,636 ^a	,186	,272

Çizelge 4.24'te sınıflandırma tablosu 2'de lojistik regresyon modeli sonucunda elde edilen sınıflandırma sonucu incelendiğinde ise, doğru sınıflandırılma oranının % 76,100 olduğu görülmektedir.

Çizelge 4.24. Sınıflandırma Tablosu 2

	Observed		Predicted		
			25-Hidroksi Vitamin D		Percentage Correct
			0	1	
Step 1	25-Hidroksi Vitamin D	0	278	20	93,3
		1	77	31	28,7
	Overall Percentage				76,1

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Ek 46’da denklemdeki değişkenlerden elde edilen sonuçlara göre; yaş değişkeni bakımından vitamin D eksikliği oddsunda 1,032 kat artışa yol açtığı görülmektedir. GGT değişkeni bakımından vitamin D eksikliği oddsunda 1,015 kat artışa yol açtığı görülmektedir. Albümin değişkeni bakımından vitamin D eksikliği oddsunda 1,121 kat artışa yol açtığı görülmektedir. T3 değişkeni bakımından vitamin D eksikliği oddsunda 2,332 kat artışa yol açtığı görülmektedir. T4 değişkeni bakımından vitamin D eksikliği oddsunda 0,096 kat azalışa yol açtığı görülmektedir.

Lojistik regresyon modelinin denklemi şu şekilde gösterilmektedir.

$$Y = -30,849 + 0.032*YAŞ + 0,015*SMEANGGT + 0,114*SMEANALBÜMİN + 0,847*SMEANT3 - 2,341*SMEANT4$$

4.3. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Yapılan Çalışmalar

Çalışmanın üçüncü aşamasının sonuçları bu kısımda verilmiştir.

4.3.1. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Sınıflandırma Yöntemleri Sonuçları

278 hastanın kan tahlilinin bulunduğu veri setinin sınıflandırma sonuçları çizelge 4.25’te gösterilmiştir. Yapay sinir ağları, naive bayes, rastgele orman, karar ağaçları, destek vektör makineleri ve k-en yakın komşuluk algoritması modellerinin doğruluk oranları sırasıyla % 82,734, % 78,417, % 78,058, % 76,260, % 53,237 ve % 35,252 şeklinde sıralanmaktadır. Yapay sinir ağları, naive bayes, rastgele orman, karar ağaçları, destek vektör makineleri ve k-en yakın komşuluk algoritması sınıflandırıcıları için sınıflandırma doğruluk oranlarını karşılaştıran çalışma, diyabet hastalığı olmayanların vitamin D eksikliğini tahmin etmek için optimum tahmin modelini seçmiştir. Bu sınıflandırıcılar arasında yapay sinir ağları için sınıflandırma doğruluk oranı tüm sınıflandırma yöntemlerinden daha yüksek olduğu tespit

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

edilmiştir. Yapay sinir ağları veri setinde % 82,734 doğruluk oranıyla en iyi performansı ortaya koymuştur.

Çizelge 4.25. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Sınıflandırma Sonuçları

Sınıflandırma Algoritması	Sınıflandırma Doğruluğu (%)
Yapay Sinir Ağları (Multilayer Perceptron)	82,734
Naive Bayes (Bayes Net)	78,417
Rastgele Orman (Random Forest)	78,058
Karar Ağaçları (Random Committee)	76,260
Destek Vektör Makineleri (LibSVM)	53,237
K-En Yakın Komşuluk Algoritması (IBk)	35,252

Bu model için duyarlılık, özgüllük, kesinlik, f-measure ve kappa gibi diğer performans ölçütleri çizelge 4.26’da sunulmuştur.

Çizelge 4.26. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Sınıflandırma Sonuçlarına Göre Oluşan Performans Ölçütleri

Sınıflandırma Algoritması	Duyarlılık	Özgüllük	Kesinlik	F-Measure	Kappa
Yapay Sinir Ağları	0,827	0,822	0,828	0,827	0,653
Naive Bayes	0,784	0,705	0,786	0,783	0,562
Rastgele Orman	0,781	0,705	0,782	0,779	0,555
Karar Ağaçları	0,763	0,659	0,767	0,760	0,517
Destek Vektör Makineleri	0,532	0	0,286	0,372	-0,007
K-En Yakın Komşuluk Algoritması	0,353	0,535	0,344	0,332	-0,262

Yapay sinir ağları bu performans ölçütlerinde en yüksek performansı göstermiş; duyarlılık, özgüllük, kesinlik, f-measure ve kappa değerleri sırasıyla

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

0,827; 0,822; 0,828; 0,827 ve 0,653 olarak hesaplanmıştır. Bu nedenle yapay sinir ağları diyabet hastalığı olmayan hastaların veri setinde optimum tahmin modeli olmuştur.

Diyabet hastalığı olmayan hastaların veri seti üzerinde çalışılan sınıflandırma sonuçlarının Weka çıktıları ek 47, ek 48, ek 49, ek 50, ek 51 ve ek 52’te gösterilmiştir.

4.3.2. Diyabet Hastalığı Olmayan Hastaların Üzerinde Çalışılan Nitelik Seçimi Sonuçları

Yenilenen sınıflandırma yöntemlerinin sonuçlarını geliştirilmek amacıyla nitelik seçimi sonuçları da yenilenmiştir.

278 hastanın kan tahlilinin bulunduğu veri setinin nitelik seçimi sonuçları ek 53’te gösterilmiştir. Sonuçlarda ilk veri setinde olduğu gibi bilgi kazanımı, kıkare ve kazanım oranı yöntemleri uygulandığında benzer sonuçlar göstermiştir.

4.3.3. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Nitelik Seçimi Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları

278 hastanın kan tahlilinin bulunduğu veri setinin nitelik seçimi sonuçlarına göre tekrarlanan sınıflandırma sonuçları çizelge 4.27’de gösterilmiştir.

Çizelge 4.27. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Nitelik Seçimi Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları (Parantez içindeki değerler seçilen nitelik sayılarını göstermektedir.)

Sınıflandırma Algoritması	Nitelik Seçimi Öncesi (%)	Bilgi Kazanımı (%)	Ki-Kare (%)	Kazanım Oranı (%)	Relief-f (%)
Yapay Sinir Ağları	82,734 (58)	85,971 (6)	85,971 (6)	86,331 (6)	84,892 (5)
Destek Vektör Makineleri	53,237 (58)	77,338 (1)	77,338 (1)	77,338 (1)	84,173 (2)

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Karar Ağaçları	76,259 (58)	77,338 (1)	77,338 (1)	77,338 (1)	79,496 (2)
Naive Bayes	78,417 (58)	78,417 (8)	78,417 (8)	78,417 (8)	78,417 (58)
Rastgele Orman	78,058 (58)	78,058 (24)	78,058 (24)	78,058 (24)	78,058 (58)
K-En Yakın Komşuluk Algoritması	35,252 (58)	77,338 (1)	77,338 (1)	77,338 (1)	76,259 (10)

Görüldüğü gibi nitelik seçimi yöntemi sonuçlarından elde edilen sınıflandırma sonuçları, nitelik seçimi yapılmadan önceki sonuçlara göre başarı oranlarında, naive bayes ve rastgele orman yöntemleri hariç yükselme sağlamıştır. Nitelik seçimi yöntemleri başarı oranını yükseltmenin yanında nitelik sayılarını da düşürerek daha iyi sonuçlar elde edilmesini sağlamıştır. Bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri veri seti üzerinde benzer sonuçlar vermiştir. Yapay sinir ağları sınıflandırma yöntemi kazanım oranı nitelik seçimi yöntemi sonrası yapılan sınıflandırmada % 86,331 başarı oranını sağlamıştır. Bu oranı yakalarken kazanım oranı nitelik seçimi yöntemi nitelik sayısını 6'ya düşürmüştür.

Destek vektör makineleri sınıflandırma yöntemi relief-f nitelik seçimi yöntemi sonrası yapılan sınıflandırmada % 84,173 başarı oranını sağlamıştır. Bu oranı yakalarken relief-f nitelik seçimi yöntemi nitelik sayısını 2'ye düşürmüştür.

Karar ağaçları sınıflandırma yöntemi relief-f nitelik seçimi yöntemi sonrası yapılan sınıflandırmada % 79,496 başarı oranını sağlamıştır. Bu oranı sağlarken relief-f nitelik seçimi yöntemi nitelik sayısını 2'ye düşürmüştür.

Naive bayes sınıflandırma yöntemi her bir nitelik seçimi yöntemi sonrası yapılan sınıflandırmada % 78,417 başarı oranını sağlamıştır. Bu oranı yakalarken bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri nitelik sayısını 8'e düşürmüştür.

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Rastgele Orman sınıflandırma sınıflandırma yöntemi bilgi kazanımı, kıkare ve kazanım oranı nitelik seçimi yöntemleri sonrası yapılan sınıflandırmada % 78,058 başarı oranını sağlamıştır. Bu oranı sağlarken bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri nitelik sayısını 24'e düşürmüştür. K-en yakın komşuluk algoritması sınıflandırma yöntemi bilgi kazanımı, kıkare ve kazanım oranı nitelik seçimi yöntemleri sonrası yapılan sınıflandırmada % 77,338 başarı oranını sağlamıştır. Bu oranı sağlarken bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri nitelik sayısını 1'e düşürmüştür.

Nitelik seçimi sonrası oluşan model için duyarlılık, özgülük, kesinlik, fmeasure ve kappa gibi diğer performans ölçütleri çizelge 4.28'de sunulmuştur.

Çizelge 4.28. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Nitelik Seçimi Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçlarına Göre Oluşan Performans Ölçütleri

Sınıflandırma Algoritması	Duyarlılık	Özgüllük	Kesinlik	F-Measure	Kappa
Yapay Sinir Ağları	0,863	0,938	0,874	0,863	0,728
Destek Vektör Makineleri	0,842	0,845	0,842	0,842	0,682
Karar Ağaçları	0,795	0,705	0,799	0,793	0,584
Naive Bayes	0,784	0,713	0,786	0,783	0,563
Rastgele Orman	0,781	0,705	0,782	0,779	0,555
K-En Yakın Komşuluk Algoritması	0,773	0,705	0,774	0,772	0,541

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Yapay sinir ağları bu performans ölçütlerinde en yüksek performansı göstermiş; duyarlılık, özgüllük, kesinlik, f-measure ve kappa değerleri sırasıyla 0,863; 0,938; 0,874; 0,863 ve 0,728 olarak hesaplanmıştır. Bu nedenle yapay sinir ağları diyabet hastalığı olmayan hastaların veri setinde optimum tahmin modeli olmuştur.

Diyabet hastalığı olmayan hastaların veri seti üzerinde çalışılan nitelik seçimi sonuçlarına göre tekrarlanan sınıflandırma sonuçlarının Weka çıktıları ek 54, ek 55, ek 56, ek 57, ek 58 ve ek 59'da gösterilmiştir.

4.3.4. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Kümeleme Sonuçları

278 hastanın kan tahlilinin bulunduğu veri setinin kümeleme işlemi sonucu ek 60'da gösterilmiştir.

Diyabet hastalığı olmayan hastalarda vitamin D eksikliğini etkileyen faktörler albümin, trigliserid, kalsiyum ve CRP şeklinde belirlenmiştir. Albümin, trigliserid, kalsiyum ve CRP'nin normal değerlerden yüksek olduğu durumda vitamin D değerinin normal değer altında çıkmasına neden olmuştur. Ayrıca bu koşulların hem kadın hem de erkek hastalarda geçerli olduğu görülmüştür.

1 numaralı küme incelendiğinde vitamin D değeri normal değerlerden düşük erkek hastaların albümin, trigliserid, kalsiyum ve CRP değerlerinin normal değerlerden yüksek olduğu tespit edilmiştir.

4 numaralı küme incelendiğinde vitamin D değeri normal değerlerden düşük kadın hastaların albümin, trigliserid, kalsiyum ve CRP değerlerinin yanı sıra LDL değerlerinin de normal değerlerden yüksek olduğu tespit edilmiştir.

5 numaralı küme incelendiğinde vitamin D değeri normal değerlerden düşük erkek hastaların albümin, trigliserid, kalsiyum ve CRP değerlerinin yanı sıra total kolesterol değerlerinin de normal değerlerden yüksek olduğu tespit edilmiştir.

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

4.3.5. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Kümeleme Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları

278 hastanın kan tahlilinin bulunduğu veri setinin kümeleme sonuçlarına göre tekrarlanan sınıflandırma sonuçları çizelge 4.29’da gösterilmiştir.

Çizelge 4.29. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Kümeleme Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçları (Parantez içindeki değerler seçilen nitelik sayılarını göstermektedir.)

Sınıflandırma Algoritması	Kümeleme Öncesi (%)	Kümeleme Sonrası (%)
Yapay Sinir Ağları	82,734	55,396
Naive Bayes	78,417	53,597
Destek Vektör Makineleri	53,237	52,518
Rastgele Orman	78,058	50,719
K-En Yakın Komşuluk Algoritması	35,252	50,719
Karar Ağaçları	76,260	50,360

Görüldüğü gibi kümeleme sonucuna göre tekrarlanan k-en yakın komşuluk algoritması ile elde edilen sınıflandırma sonuçları, kümeleme yapılmadan önceki sonuçlara göre başarı oranlarında yükselme sağlamıştır. Kümeleme sonucu başarı oranını yükseltmenin yanında nitelik sayılarını da düşürerek daha iyi sonuçlar elde edilmesini sağlamıştır. Kümeleme sonucunda nitelik sayısı 4’e düşmüştür. K-en yakın komşuluk algoritması sınıflandırma yöntemi kümeleme sonrası yapılan sınıflandırmada % 50,719 başarı oranını sağlamıştır.

Kümeleme sonrası oluşan model için duyarlılık, özgüllük, kesinlik, fmeasure ve kappa gibi diğer performans ölçütleri çizelge 4.30’da sunulmuştur.

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Çizelge 4.30. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Kümeleme Sonuçlarına Göre Tekrarlanan Sınıflandırma Sonuçlarına Göre Oluşan Performans Ölçütleri

Sınıflandırma Algoritması	Duyarlılık	Özgüllük	Kesinlik	F-Measure	Kappa
Yapay Sinir Ağları	0,554	0,504	0,553	0,553	0,101
Naive Bayes	0,536	0	?	?	0
Destek Vektör Makineleri	0,525	0,519	0,527	0,526	0,049
Rastgele Orman	0,507	0,202	0,484	0,464	-0,028
K-En Yakın Komşuluk Algoritması	0,507	0,736	0,531	0,485	0,044
Karar Ağaçları	0,504	0,178	0,475	0,453	-0,038

Yapay sinir ağları bu performans ölçütlerinde en yüksek performansı göstermiş; duyarlılık, özgüllük, kesinlik, f-measure ve kappa değerleri sırasıyla 0,554; 0,504; 0,553; 0,553 ve 0,101 olarak hesaplanmıştır. Bu nedenle yapay sinir ağları optimum tahmin modeli olmuştur.

Diyabet hastalığı olmayan hastaların veri seti üzerinde çalışılan kümeleme sonuçlarına göre tekrarlanan sınıflandırma sonuçlarının Weka çıktıları ek 61, ek 62, ek 63, ek 64, ek 65 ve ek 66'da gösterilmiştir.

4.3.6. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Faktör Analizi Sonuçları

278 hastanın kan tahlilinin bulunduğu veri setinin faktör analizi sonuçları örneklem büyüklüğünün yeterli olmaması nedeniyle tamamlanamamıştır.

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

4.3.7. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Lojistik Regresyon Sonuçları

278 hastanın kan tahlilinin bulunduğu veri setinin lojistik regresyon sonuçları örneklem büyüklüğünün yeterli olmaması nedeniyle tamamlanamamıştır.

4.4. Tartışma

Elde edilen sonuçlar bu kısımda tartışılacaktır.

4.4.1. Tüm Veri Üzerinde Yapılan Çalışmaların Özeti

684 hastanın kan tahlilinin bulunduğu veri setinin sınıflandırma sonuçları incelendiğinde karar ağaçları % 78,655 ile en iyi sınıflandırma doğruluğu oranını yakalamıştır ancak yapay sinir ağları duyarlılık, özgüllük, kesinlik, f-measure ve kappa gibi diğer performans ölçütlerinde daha iyi sonuç vermiştir. Nitelik seçimi sonuçlarına göre tekrarlanan sınıflandırma sonuçları incelendiğinde karar ağaçları % 83,479 ile en iyi sınıflandırma doğruluğu oranının yanı sıra duyarlılık, özgüllük, kesinlik, f-measure ve kappa gibi diğer performans ölçütlerinde en yüksek performansı göstermiştir. Bu oranı sağlarken bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri nitelik sayısını 36'ya düşürmüştür. Yapılan kümeleme işlemi sonucunda vitamin D eksikliğini etkileyen faktörler glukoz, HbA1c, HbA1c(IFCC), albümin, total kolesterol, trigliserid, kalsiyum ve CRP olarak belirlenmiştir. Glukoz, HbA1c, HbA1c(IFCC), albümin, total kolesterol, trigliserid, kalsiyum ve CRP'nin normal değerlerden yüksek olduğu durum vitamin D değerinin normal değer altında çıkmasına neden olmuştur. Kümeleme sonucuna göre tekrarlanan sınıflandırma sonuçları destek vektör makineleri için % 70,029'den % 78,947'ye; k-en yakın komşuluk algoritması için % 45,614'den % 47,515'e başarı oranlarında yükselme sağlamıştır. Faktör analizi sonuçları incelendiğinde vitamin D'nin de içinde bulunduğu grupta ayrıca albümin, CRP ve insülin de bulunmaktadır. Lojistik regresyon sonuçları incelendiğinde vitamin D eksikliğini etkileyen faktörler

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

diyabet, cinsiyet, yaş, ürik asit, ALT, GGT, albümin, trigliserid, fosfor, CK, PDW, sedimentasyon ve TSH olarak sıralanmaktadır.

4.4.2. Diyabet Hastalarının Veri Seti Üzerinde Yapılan Çalışmaların Özeti

406 hastanın kan tahlilinin bulunduğu veri setinin sınıflandırma sonuçları incelendiğinde destek vektör makineleri % 98,030 ile en iyi sınıflandırma doğruluğu oranının yanı sıra duyarlılık, özgüllük, kesinlik, f-measure ve kappa gibi diğer performans ölçütlerinde en yüksek performansı göstermiştir. Nitelik seçimi sonuçlarına göre tekrarlanan sınıflandırma sonuçları incelendiğinde destek vektör makineleri % 98,030 ile en iyi sınıflandırma doğruluğu oranının yanı sıra duyarlılık, özgüllük, kesinlik, f-measure ve kappa gibi diğer performans ölçütlerinde en yüksek performansı göstermiştir. Bu oranı yakalarken bilgi kazanımı, ki-kare ve kazanım oranı nitelik seçimi yöntemleri nitelik sayısını 31'e; relief-f nitelik seçimi yöntemi ise nitelik sayısını 19'a düşürmüştür. Yapılan kümeleme işlemi sonucunda vitamin D eksikliğini etkileyen faktörler glukoz, HbA1c, HbA1c(IFCC), total kolesterol ve trigliserid olarak belirlenmiştir. Glukoz, HbA1c, HbA1c(IFCC), total kolesterol ve trigliserid'in normal değerlerden yüksek olduğu durumda vitamin D değerinin normal değer altında çıkmasına neden olmuştur. Kümeleme sonucuna göre tekrarlanan sınıflandırma sonuçları naive bayes için % 73,150'den % 73,399'a; k-en yakın komşuluk algoritması için % 41,870'den % 72,660'a başarı oranlarında yükselme sağlamıştır. Faktör analizi sonuçları KMO oranının 0,494 çıkması nedeniyle tamamlanamamıştır. Lojistik regresyon sonuçları incelendiğinde vitamin D eksikliği etkileyen faktörler yaş, GGT, albümin, T3 ve T4 olarak sıralanmaktadır.

4.4.3. Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Yapılan Çalışmaların Özeti

278 hastanın kan tahlilinin bulunduğu veri setinin sınıflandırma sonuçları incelendiğinde yapay sinir ağları % 82,734 ile en iyi sınıflandırma doğruluğu

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

oranının yanı sıra duyarlılık, özgüllük, kesinlik, f-measure ve kappa gibi diğer performans ölçütlerinde en yüksek performansı göstermiştir. Nitelik seçimi sonuçlarına göre tekrarlanan sınıflandırma sonuçları incelendiğinde yapay sinir ağları % 86,331 ile en iyi sınıflandırma doğruluğu oranının yanı sıra duyarlılık, özgüllük, kesinlik, f-measure ve kappa gibi diğer performans ölçütlerinde en yüksek performansı göstermiştir. Bu oranı yakalarken kazanım oranı nitelik seçimi yöntemi nitelik sayısını 6'ya düşürmüştür. Yapılan kümeleme işlemi sonucunda vitamin D eksikliğini etkileyen faktörler albümin, trigliserid, kalsiyum ve CRP olarak belirlenmiştir. Albümin, trigliserid, kalsiyum ve CRP'nin normal değerlerden yüksek olduğu durumda vitamin D değerinin normal değer altında çıkmasına neden olmuştur. Kümeleme sonucuna göre tekrarlanan sınıflandırma sonuçları k-en yakın komşuluk algoritması için % 35,252'den % 50,719'a başarı oranlarında yükselme sağlamıştır. Faktör analizi ve lojistik regresyon sonuçları örneklem büyüklüğünün yeterli olmaması nedeniyle tamamlanamamıştır.

4.4.4. Çalışmada Kullanılan Yöntemler

Çalışmada kullanılan yöntemler çizelge 4.31'de gösterilmiştir.

Çizelge 4.31. Çalışmada Kullanılan Yöntemler

Kullanılan Yöntemler	Tüm Veri	Diyabet Hastalarının Veri Seti	Diyabet Olmayan Hastaların Veri Seti
Sınıflandırma (DVM, KNN, KA, NB, YSA ve RO)	✓	✓	✓
Nitelik Seçimi	✓	✓	✓
Kümeleme	✓	✓	✓
Faktör Analizi	✓		
Lojistik Regresyon	✓	✓	

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

4.4.5. Çalışmanın Literatür ile Kıyaslanması

Çizelge 4.32 incelendiğinde çalışmaya benzer vitamin D veri setlerinin kullanıldığı diğer çalışmalara göre daha iyi sonuçlar verdiği gözlemlenmektedir. En iyi sonucu Sambasivam ve ark. 2020 yılındaki çalışmalarında vitamin D eksikliğinin öneminin performans analizi için rastgele orman kullanmış ve % 96,000 doğruluk oranını yakalamıştır (Sambasivam ve ark., 2020). Kanla ilgili veri setlerinin kullanıldığı diğer çalışmalar incelendiğinde en iyi sonucu Wu ve ark. 2020 yılındaki çalışmalarında COVID-19 enfeksiyonunun tanımlanması için yine rastgele orman kullanmış ve % 97,950 doğruluk oranını yakalamıştır (Wu ve ark., 2020a). Bu çalışma ise vitamin D eksikliğinin tahmin modelleri için en iyi sonucu destek vektör makineleri kullanarak 98,030 % doğruluk oranını yakalamıştır. Görüldüğü gibi bu çalışma literatürdeki benzer çalışmalarla kıyaslandığında daha iyi performans göstererek literatüre katkıda bulunmaktadır.

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Çizelge 4.32. Çalışmanın Literatür ile Kıyaslanması

Yazar	Çalışma	Metod	Sınıflandırma Doğruluğu (%)
Ghani and Awang (2008)	Kalp hastalığı tahmini	Naive bayes	% 86,120
Srinivas ve ark. (2010)	Kalp krizi tahmini	Naive bayes	% 84,050
Methaila ve ark. (2014)	Kalp hastalığı tahmini	Karar ağaçları	% 89,000
Komi ve ark. (2014)	Diyabet tahmini	Yapay sinir ağları	% 89,000
Chulsomlee ve ark. (2014)	Kırmızı kan hücresi morfolojisi	Karar ağaçları	% 94,300
Portela ve ark. (2015)	Kan basıncı tahmini	Karar ağaçları	% 71,520
GeethaRamani and Balasubramanian (2016)	Retinal kan damarı segmentasyonu	Root guided decision tree	% 95,360
El-Halees and Shurrab (2017)	Kan tümörü tahmini	Derin öğrenme	% 79,450
Aslan ve ark. (2018)	Meme kanseri tanısı	Extreme learning machine	% 83,870
Thomas and Princy (2018)	Kalp hastalığı tahmini	Karar ağaçları	% 80,600
Lippeveld ve ark. (2019)	Beyaz kan hücrelerinin sınıflandırılması	Gradient boosting	% 87,100
Agarwal ve ark. (2019)	Kan transfüzyon sistemi	Rastgele orman	% 93,200
Akben (2019)	Kan, hormon ve obezite değer aralıklarının belirlenmesi	Karar ağaçları	% 90,500
Gonoodi ve ark. (2019)	Vitamin D eksikliğinin risk faktörlerinin değerlendirilmesi	Karar ağaçları	% 77,800
Marcus ve ark. (2020)	Kan glikoz düzeyi tahmini	Kernel ridge regression	% 93,4

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Brinati ve ark. (2020)	COVID-19 enfeksiyonunun tespiti	Rastgele orman	% 89,1
Wu ve ark. (2020)	COVID-19 enfeksiyonunun tanımlanması	Rastgele orman	% 97,95
Wu ve ark. (2020)	Sistolik kan basıncı tahmini	Yapay sinir ağları	% 95,5
Yao ve ark. (2020)	COVID-19 önem tespiti	Destek vektör makineleri	% 81,48
Sambasivam ve ark. (2020)	Vitamin D eksikliğinin öneminin performans analizi	Rastgele orman	% 96
Amiri ve ark. (2021)	Vitamin D'nin takviyeye cevabını etkileyen faktörlerin belirlenmesi	Rastgele orman	% 93
Padmajaa ve ark. (2021)	Vitamin D eksikliğinin öneminin tahmini	Extra trees classifier	% 73,3
Carretero ve ark. (2021)	Vitamin D eksikliğinin tahmin modelleri	Destek vektör makineleri	% 73
Bu çalışma	Vitamin D eksikliğinin tahmin modelleri	Destek vektör makineleri	% 98,03

4.4.6. En İyi Sınıflandırma Sonuçlarının Karşılaştırılması

Çizelge 4.33'de tüm veri seti, diyabet hastalarının veri seti ve diyabet hastalığı olmayan hastaların veri seti üzerinde yapılan sınıflandırma ve en iyi nitelik seçimi sonucuna göre tekrarlanan sınıflandırma sonuçları gösterilmektedir. Çizelge incelendiğinde diyabet hastalarının veri seti üzerinde tamamlanan nitelik seçimi dışında her bir nitelik seçimi sonucunda doğruluk oranları iyileşirken nitelik sayıları da azalmıştır. Diyabet hastalarının veri seti üzerinde tamamlanan nitelik seçimi sonucunda ise doğruluk oranı sabit kalmasına rağmen nitelik sayısı belirli ölçüde azalmıştır. Nitelik sayısının azalması hem yapılması gereken tahlillerde hem de sonuç alma süresinde azalmayı sağlamaktadır.

Çizelge 4.33. En İyi Sınıflandırma Sonuçlarının Karşılaştırılması (Parantez içindeki değerler seçilen nitelik sayılarını göstermektedir.)

		DVM (%)	KNN (%)	KA (%)	NB (%)	YSA (%)	RO (%)
Tüm Veri Üzerinde	Nitelik Seçimi Yok	70,029 (59)	45,614 (59)	78,655 (59)	68,275 (59)	78,216 (59)	77,778 (59)

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Yapılan Çalışmalar	En İyi Nitelik Seçimi Sonucu	80,263 (32)	72,222 (21)	83,479 (36)	70,029 (5)	78,801 (57)	81,286 (32)
Diyabet Hastalarının Veri Seti Üzerinde Yapılan Çalışmalar	Nitelik Seçimi Yok	98,030 (58)	41,870 (58)	96,800 (58)	73,150 (58)	81,773 (58)	95,320 (58)
	En İyi Nitelik Seçimi Sonucu	98,030 (19)	81,281 (4)	97,537 (42)	73,400 (1)	85,222 (23)	97,044 (51)
Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Yapılan Çalışmalar	Nitelik Seçimi Yok	53,237 (58)	35,252 (58)	76,259 (58)	78,417 (58)	82,734 (58)	78,058 (58)
	En İyi Nitelik Seçimi Sonucu	84,173 (2)	77,338 (1)	79,496 (2)	78,417 (8)	86,331 (6)	78,058 (24)

4.4.7. En İyi Nitelik Seçimi Sonuçlarının Karşılaştırılması

Çizelge 4.34’de tüm veri seti, diyabet hastalarının veri seti ve diyabet hastalığı olmayan hastaların veri seti üzerinde yapılan en iyi nitelik seçimi, kümeleme, faktör analizi ve lojistik regresyon sonuçlarına göre sıralanan nitelikler gösterilmektedir. Nitelik seçimi sonuçlarına göre sıralanan nitelikler incelendiğinde cinsiyet, glukoz, VB12 ve trigliserid’in her 3 veri setinde vitamin D eksikliği açısından öne çıktığı gözlemlenmektedir. Kümeleme sonuçlarına göre sıralanan nitelikler incelendiğinde trigliserid’in her 3 veri setinde de vitamin D eksikliği açısından öne çıktığı gözlemlenmektedir. Lojistik regresyon tüm veri seti ve diyabet hastalarının veri seti üzerinde tamamlanabilmiştir. Lojistik regresyon sonuçlarına göre sıralanan nitelikler incelendiğinde yaş, GGT ve albümin’in vitamin D eksikliği açısından öne çıktığı gözlemlenmektedir. Faktör analizi sadece tüm veri seti üzerinde tamamlanabilmiştir. Faktör analizi sonuçlarına göre sıralanan nitelikler incelendiğinde albümin, CRP ve insülinin vitamin D eksikliği açısından öne çıktığı gözlemlenmektedir.

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Her 3 veri setinde tamamlanan yöntemler olan sınıflandırma ve kümeleme sonuçları incelendiğinde trigliserid'in vitamin D eksikliği açısından en önemli faktör olduğu ortaya çıkmıştır.



4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Çizelge 4.34. En İyi Nitelik Seçimi Sonuçlarının Karşılaştırılması

Nitelik Sırası	Tüm Veri Üzerinde Yapılan Çalışmalar				Diyabet Hastalarının Veri Seti Üzerinde Yapılan Çalışmalar			Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Yapılan Çalışmalar	
	En İyi Nitelik Seçimi Sonucu	Faktör Analizi Sonucu	Kümeleme Sonucu	Lojistik Regresyon Sonucu	En İyi Nitelik Seçimi Sonucu	Kümeleme Sonucu	Lojistik Regresyon Sonucu	En İyi Nitelik Seçimi Sonucu	Kümeleme Sonucu
1	T4	Albümin	Glukoz	Diyabet	<u>Cinsiyet</u>	Glukoz	Yaş	<u>Cinsiyet</u>	Albümin
2	Yaş	CRP	HbA1c	Cinsiyet	Magnez	HbA1c	GGT	<u>Glukoz</u>	<i>Triglise</i>
3	Diyabet	25-Hidroksi Vitamin D	HbA1c (IFCC)	Yaş	<u>Glukoz</u>	HbA1c (IFCC)	Albümin	<u>VB12</u>	Kalsiyum
4	<u>Glukoz</u>	İnsülin	Albümin	Ürik asit	MO2	Total Kolesterol	T3	MCV	CRP
5	<u>Cinsiyet</u>		Total Kolesterol	ALT	Fosfor	<i>Triglise</i>	T4	ALT	
6	<u>Triglise</u>		<i>Triglise</i>	GGT	Yaş			<u>Triglise</u>	
7	TSH		Kalsiyum	Albümin	Sedim				
8	ALT			Triglise	PCT				
9	BA2			Fosfor	DEMİR				
10	Potasyum			CK	T3				
11	HDL			PDW	PLT				
12	LDL			Sedim	LDL				
13	Total Kolesterol			TSH	RDW-CV				
14	Kalsiyum				AST				
15	Demir				Albümin				
16	Fosfor				HDL				
17	Bilirubin				<u>Triglise</u>				

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

Çizelge 4.34. (Devamı)

18	Demir Bağlama				MPV				
19	Sodyum				<u>VB12</u>				
20	ALP								
21	Albümin								
22	HbA1c (IFCC)								
23	HbA1c								
24	Üre								
25	Kreatinin								
26	Ürik Asit								
27	GGT								
28	LDH								
29	CK								
30	AST								
31	Magnez								

32	<u>VB12</u>								
33	ASO								
34	NE2								
35	BA1								
36	EO1								

4.4.8. Çalışmada En İyi Sonucu Veren Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Destek Vektör Makineleri Kullanarak Elde Edilen En İyi Sonuca Göre Seçilen Niteliklerin Literatür Çalışması

En iyi sınıflandırma doğruluğu sonucunu veren nitelik sıralaması ışığında yeni literatür taraması gerçekleştirilmiştir. Bu şekilde nitelik seçimi literatürle doğrulanmıştır.

Verdoia ve ark. (2015) cinsiyetin vitamin D durumunu önemli ölçüde etkilediğini vurgulamıştır (Verdoia ve ark., 2015). Uwitonze ve Razzaque (2018) vitamin D'nin optimal faydalarını elde etmek için önerilen miktarda magnezyum tüketiminin gerekli olduğunu öne sürmüştür (Uwitonze ve Razzaque, 2018). Tai ve ark. (2008) glukoz metabolizmasının anormalliği ile vitamin D eksikliği arasında bir bağlantı olduğuna dair bir inceleme geliştirmiştir (Tai ve ark., 2008). Nurminen ve ark. (2019) monositlerin bağışıklıkla ilgili temel işlevlerinin vitamin D tarafından desteklendiğini belirtmiştir (Nurminen ve ark., 2019). Deluca (2009) vitamin D'nin vücutta kullanılmasının kalsiyum ve fosfor düzeylerini yükselttiğini öne sürmüştür (Deluca, 2009). Baker ve ark. (1980) anormal derecede düşük plazma 25OHD düzeylerinin yaşlılarda daha yaygın olduğunu vurgulamıştır (Baker ve ark., 1980). Kaya ve ark. (2018) vitamin D eksikliği olan hastalarda eritrosit sedimentasyon oranının daha yüksek olduğunu ortaya koymuştur (Kaya ve ark., 2018). Wolf ve ark. (2019) vitamin D takviyesinin prokalsitonin kullanma kabiliyetine müdahale edebileceğini öne sürmüştür (Wolf ve ark., 2019). AziziSoleiman ve ark. (2016) demir durumu ile vitamin D arasında pozitif bir ilişki olduğunu belirtmiştir (Azizi-Soleiman ve ark., 2016). Colston ve Cleeve (1986) vitamin D kısıtlı diyet alan yetişkin dişi wistar sıçanlarının deri altı T3 enjeksiyonu ile tirotoksik hale getirildiğini belirtmişlerdir (Colston ve Cleeve, 1986). Cure ve ark. (2014) vitamin D eksikliğin yüksek yoğunluk trombosit hacmini artırabileceğini bulmuştur (Cure ve ark., 2014). Schwetz ve ark (2018) vitamin D takviyesinin düşük yoğunluklu lipoproteini önemli ölçüde azalttığını öne sürmüştür (Schwetz ve ark, 2018). Bujak

ve ark. (2015) yüksek RDW düzeylerinin olumsuz prognostik etkilerinin, vitamin D gibi risk faktörlerinin olumsuz etkilerine atfedilebileceğini tartışmıştır (Bujak ve ark., 2015). Seif ve Abdelwahed (2014) vitamin D ile tedavi edilen sıçanların serum aspartat aminotransferaz düzeylerinin önemli ölçüde daha düşük olduğunu vurgulamıştır (Seif ve Abdelwahed, 2014). Carter ve Ho (1994) albüminin vitamin D bağlayıcı proteini içeren çok genli bir protein ailesine ait olduğunu öne sürmüştür (Carter ve Ho, 1994). Kazlauskaite ve ark. (2010) düşük vitamin D düşük seviyelerde yüksek yoğunluklu lipoprotein ile ilişkili olduğunu öne sürmüştür (Kazlauskaite ve ark., 2010). Rodriguez ve ark. (2011) çocuklarda düşük serum vitamin D düzeylerinin yüksek trigliserid düzeyleri ile ilişkili olduğunu öne sürmüştür (Rodriguez ve ark., 2011). Tariq ve ark. (2017) vitamin D ile B12 vitamininin anlamlı negatif korelasyonunun olduğunu öne sürmüştür (Tariq ve ark., 2017).

4.5. Yazılım Önerisi

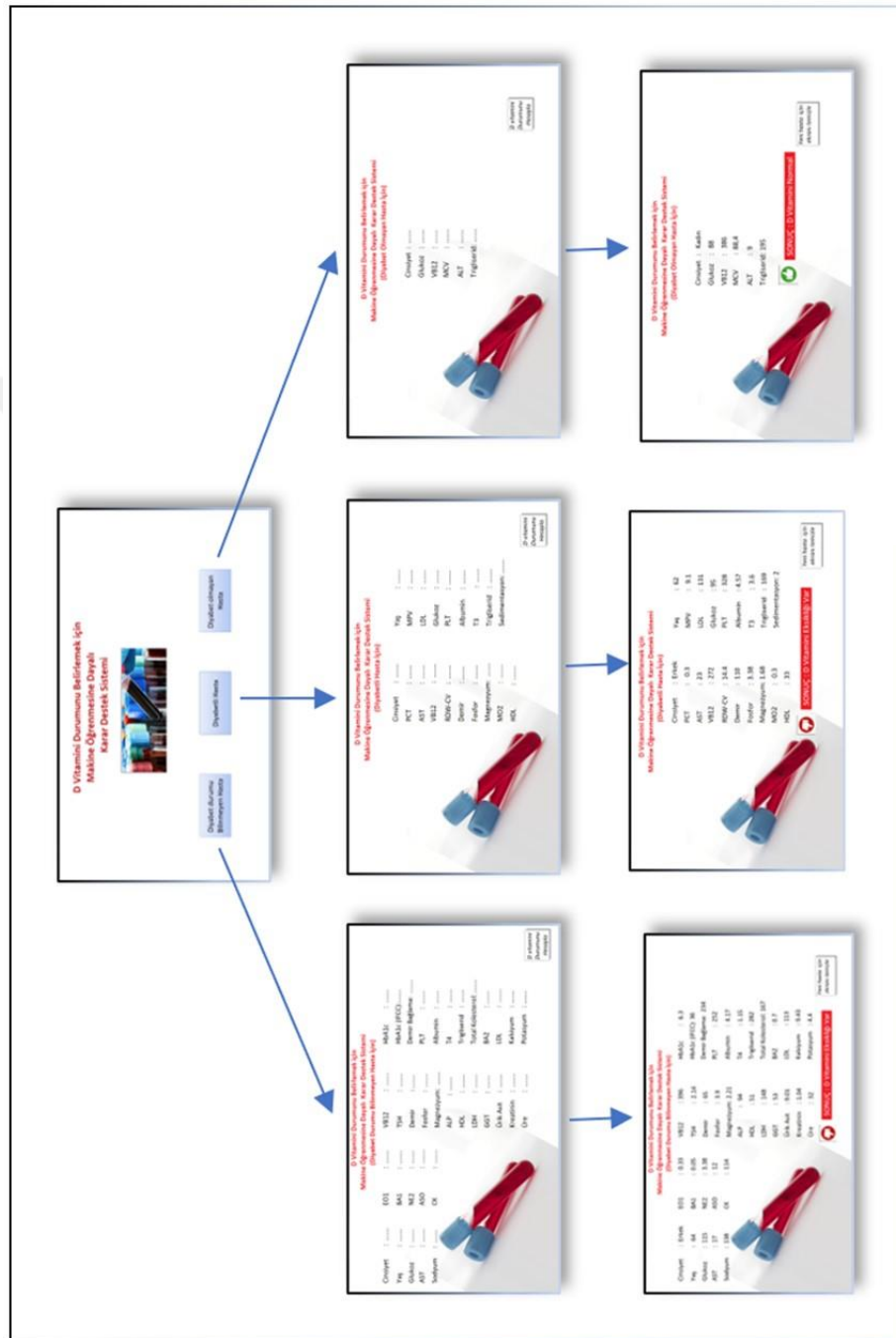
Tez çalışmasının önerilen modelleri kullanılarak vitamin D eksikliğinin tespiti için bir prototip yazılım önerilmiştir.

4.5.1. Yazılım Arayüzleri

Yapılan analizler sonucunda en iyi performans gösteren 3 farklı model bulunmaktadır. Bu modeller sırasıyla: tüm veri seti için karar ağaçları, diyabetli veri seti için destek vektör makineleri ve diyabetsiz veri seti için yapay sinir ağları şeklinde sıralanmaktadır. Şekil 4.3'te yazılım hiyerarşisi gösterilmektedir.

4. BULGULAR VE TARTIřMA
Uğur Engin EřSİZ



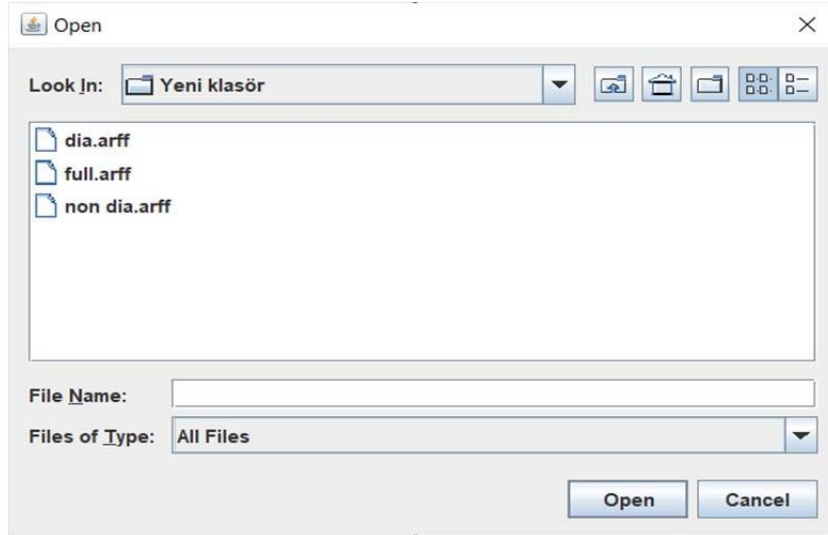


Şekil 4.4'te yazılım ana menüsü gösterilmektedir.



Şekil 4.4. Yazılım Ana Menüsü

Şekil 4.5'te veri seti giriş ekranı gösterilmektedir. Menüde işlem yapılacak veri seti yüklemesi yapılmaktadır. Yüklenen veri setine uygun olan model önerisi yazılım tarafından otomatik olarak karar verilmekte ve öneri olarak gösterilmektedir.



Şekil 4.5. Veri Seti Giriş Ekranı

4. BULGULAR VE TARTIŞMA
Uğur Engin EŞSİZ

Şekil 4.6'da diyabet durumu bilinmeyen bir hastanın kan tahlillerinin giriş ekranı gösterilmiştir.

D Vitamini Durumunu Belirlemek için
Makine Öğrenmesine Dayalı Karar Destek Sistemi
(Diyabet Durumu Bilinmeyen Hasta için)

Cinsiyet :	EO1 :	VB12 :	HbA1c :
Yaş :	BA1 :	TSH :	HbA1c (IFCC):.....
Glukoz :	NE2 :	Demir :	Demir Bağlama:
AST :	ASO :	Fosfor :	PLT :
Sodyum :	CK :	Magnezyum:	Albumin :
		ALP :	T4 :
		HDL :	Trigliserid :
		LDH :	Total Kolesterol:
		GGT :	BA2 :
		Ürik Asit :	LDL :
		Kreatinin :	Kalsiyum :
		Üre :	Potasyum :



D vitamini Durumunu Hesapla

Şekil 4.6. Diyabet Durumu Bilinmeyen Hastanın Tahlil Giriş Ekranı

Şekil 4.7'de sonuçlar sisteme girilmiş ve vitamin D durumuyla ilgili tahmin gösterilmiştir. Hastada vitamin D eksikliği tespit edilmiştir.

D Vitamini Durumunu Belirlemek için
Makine Öğrenmesine Dayalı Karar Destek Sistemi
(Diyabet Durumu Bilinmeyen Hasta için)

Cinsiyet : Erkek	EO1 : 0.33	VB12 : 396	HbA1c : 6.3
Yaş : 64	BA1 : 0.05	TSH : 2.14	HbA1c (IFCC): 36
Glukoz : 115	NE2 : 3.38	Demir : 65	Demir Bağlama: 234
AST : 17	ASO : 12	Fosfor : 3.9	PLT : 252
Sodyum : 138	CK : 114	Magnezyum: 2.21	Albumin : 4.17
		ALP : 94	T4 : 1.15
		HDL : 51	Trigliserid : 282
		LDH : 148	Total Kolesterol: 167
		GGT : 53	BA2 : 0.7
		Ürik Asit : 9.01	LDL : 113
		Kreatinin : 1.04	Kalsiyum : 9.43
		Üre : 32	Potasyum : 4.4



SONUÇ : D Vitamini Eksikliği Var

Yeni hasta için ekranı temizle

Şekil 4.7. Diyabet Durumu Bilinmeyen Hasta Sonucu

Şekil 4.8’de bir diyabet hastasının kan tahlillerinin giriş ekranı gösterilmiştir.

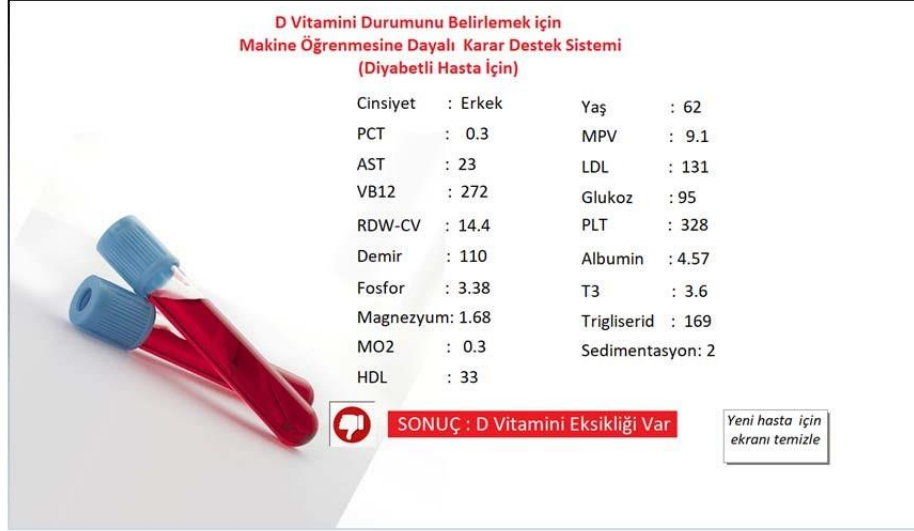
**D Vitamini Durumunu Belirlemek için
Makine Öğrenmesine Dayalı Karar Destek Sistemi
(Diyabetli Hasta için)**

Cinsiyet	:	Yaş	:
PCT	:	MPV	:
AST	:	LDL	:
VB12	:	Glukoz	:
RDW-CV	:	PLT	:
Demir	:	Albumin	:
Fosfor	:	T3	:
Magnezyum	:	Trigliserid	:
MO2	:	Sedimentasyon	:
HDL	:		

*D vitamini
Durumunu
Hesapla*

Şekil 4.8. Diyabet Hastasının Tahlil Giriş Ekranı

Şekil 4.9’da sonuçlar sisteme girilmiş ve vitamin D durumuyla ilgili tahmin gösterilmiştir. Hastada vitamin D eksikliği tespit edilmiştir.



Şekil 4.9. Diyabet Hastası Sonucu

Şekil 4.10’da diyabet hastalığı olmayan bir hastanın kan tahlillerinin giriş ekranı gösterilmiştir.

**D Vitamini Durumunu Belirlemek için
Makine Öğrenmesine Dayalı Karar Destek Sistemi
(Diyabet Olmayan Hasta için)**

Cinsiyet :
Glukoz :
VB12 :
MCV :
ALT :
Trigliserid :

D vitamini Durumunu Hesapla

Şekil 4.10. Diyabet Hastalığı Olmayan Hastanın Tahlil Giriş Ekranı

Şekil 4.11’de sonuçlar sisteme girilmiş ve vitamin D durumuyla ilgili tahmin gösterilmiştir. Hastada vitamin D eksikliği tespit edilmemiştir.



Şekil 4.11. Dişabet Hastalığı Olmayan Hasta Sonucu

4.5.2. Yazılım Validasyonu

Çizelge 4.35’deki kontrol listesi yazılan karar destek sistemi yazılımının geçerliliğini test etmek için hazırlanmıştır.

İlk 30 veri, rastgele seçilmiş tüm veri setine ait olup tüm veri analizine tabi tutulmuştur. İkinci 30 veri, rastgele seçilmiş dişabet hastalarının veri setine ait olup dişabet hastalarının analizine tabi tutulmuştur. Üçüncü 30 veri, rastgele seçilmiş dişabet hastalığı olmayan hastaların veri setine ait olup hem dişabet hastalığı olmayan hastaların analizine tabi tutulmuştur.

İkinci sütunda yer alan veriler hastanın vitamin D değerleri olarak girilmiş; yazılımdan “Yeterli” veya “Yetersiz” diye çıkan sonuçlar ise el ile hazırlanan kontrol listesinin üçüncü, dördüncü ve beşinci sütunlarına yazılmıştır. İkinci sütunda yer alan hastanın vitamin D değerleri 30’den büyük ise “Yeterli”; değil ise “Yetersiz” kabul edilmiştir. Üçüncü, dördüncü ve beşinci sütunlarda yazılımdan çıkan vitamin D tahmin sonuçlarını ve tahmininin doğru olup olmadığını göstermektedir.

4. BULGULAR VE TARTIŞMA
Uğur Engin EŞSİZ

Çıkan sonuca göre tüm veri analizinde 30 verinin 29’unda karar destek sistemi yazılımı doğru tahminde bulunurken; diyabetli hastalarının analizinde karar destek sistemi yazılımı 30 veriden 30’unda doğru tahmin yapmış, diyabet hastalığı olmayan hastaların analizinde ise karar destek sistemi yazılımı 30 veriden 27’sinde doğru tahmin yapmıştır.

Sonuç olarak, yazılım rastgele seçilmiş toplam 90 veri ile test edilmiş ve % 95,555 sınıflandırma doğruluğu oranını sağlamıştır.

Çizelge 4.35. Yazılım Validasyonu

Veri No	Hasta Vitamin D Değeri (µg/ml)	Tüm Veri Setine Göre Analiz Sonucu	Diyabet Hastalarının Veri Seti Göre Analiz Sonucu	Diyabet Hastalığı Olmayan Hastaların Veri Setine Göre Analiz Sonucu
1	16,5	Yetersiz ✓		
2	10,6	Yetersiz ✓		
3	27,2	Yetersiz ✓		
4	23,6	Yetersiz ✓		
5	26,7	Yetersiz ✓		
6	22,2	Yetersiz ✓		
7	15,1	Yetersiz ✓		
8	13,9	Yetersiz ✓		
9	18,2	Yetersiz ✓		
10	23,8	Yetersiz ✓		
11	24,2	Yetersiz ✓		
12	29,1	Yeterli X		
13	30,6	Yeterli ✓		

4. BULGULAR VE TARTIŞMA

Uğur Engin EŞSİZ

14	39,9	Yeterli ✓		
15	16,8	Yetersiz ✓		
16	17,6	Yetersiz ✓		
17	22,2	Yetersiz ✓		
18	5	Yetersiz ✓		
19	14,3	Yetersiz ✓		
20	14,6	Yetersiz ✓		
21	7,7	Yetersiz ✓		
22	8,1	Yetersiz ✓		
23	9,1	Yetersiz ✓		
24	10,8	Yetersiz ✓		
25	19,7	Yetersiz ✓		
26	23,4	Yetersiz ✓		
27	32	Yeterli ✓		
28	19,8	Yetersiz ✓		
29	21	Yetersiz ✓		
30	23,7	Yetersiz ✓		
31	16,5		Yetersiz ✓	
32	10,6		Yetersiz ✓	
33	27,2		Yetersiz ✓	
34	23,6		Yetersiz ✓	
35	26,7		Yetersiz ✓	
36	22,2		Yetersiz ✓	
37	15,1		Yetersiz ✓	
38	13,9		Yetersiz ✓	
39	18,2		Yetersiz ✓	
40	13,4		Yetersiz ✓	
41	23,8		Yetersiz ✓	
42	24,2		Yetersiz ✓	
43	29,1		Yetersiz ✓	
44	30,6		Yeterli ✓	
45	39,9		Yeterli ✓	

4. BULGULAR VE TARTIŞMA
Uğur Engin EŞSİZ

46	34,3		Yeterli ✓	
47	36,6		Yeterli ✓	
48	33,2		Yeterli ✓	
49	16,8		Yetersiz ✓	
50	17,6		Yetersiz ✓	
51	22,2		Yetersiz ✓	

Çizelge 4.35.(devamı)

52	5		Yetersiz ✓	
53	14,3		Yetersiz ✓	
54	14,6		Yetersiz ✓	
55	7,7		Yetersiz ✓	
56	8,1		Yetersiz ✓	
57	9,1		Yetersiz ✓	
58	10,8		Yetersiz ✓	
59	19,7		Yetersiz ✓	
60	23,4		Yetersiz ✓	
61	4,2			Yetersiz ✓
62	5,8			Yetersiz ✓
63	6,2			Yetersiz ✓
64	7,6			Yetersiz ✓
65	7,7			Yetersiz ✓
66	10,6			Yetersiz ✓
67	11			Yetersiz ✓
68	11,6			Yetersiz ✓
69	12,4			Yetersiz ✓
70	12,4			Yetersiz ✓
71	13,8			Yetersiz ✓
72	14,4			Yetersiz ✓
73	17			Yetersiz ✓
74	17,4			Yetersiz ✓
75	17,7			Yetersiz ✓
76	19,1			Yetersiz ✓
77	20,2			Yetersiz ✓

4. BULGULAR VE TARTIŞMA
Uğur Engin EŞSİZ

78	21,9			Yetersiz ✓
79	21,9			Yetersiz ✓
80	26			Yetersiz ✓
81	26,6			Yetersiz ✓
82	29,1			Yetersiz ✓
83	29,5			Yeterli X
84	29,6			Yeterli X
85	29,7			Yeterli X
86	31,1			Yeterli ✓
87	31,4			Yeterli ✓
88	31,9			Yeterli ✓
89	31,9			Yeterli ✓
90	32,2			Yeterli ✓
Toplam Doğru Tahmin		29	30	27
Toplam Veri		30	30	30

5. SONUÇLAR VE ÖNERİLER

Uğur Engin EŞSİZ

5. SONUÇLAR VE ÖNERİLER

Çalışmada diyabet hastalarının veri seti, diyabet hastalığı olmayan hastaların veri seti ve bunların birleşiminden oluşan tüm veri seti üzerinde vitamin D eksikliğini teşhis etmek için destek vektör makineleri, k-en yakın komşuluk algoritması, karar ağaçları, naive bayes, yapay sinir ağları ve rastgele orman sınıflandırıcılarının; bilgi kazanımı, ki-kare, kazanım oranı ve relief-f nitelik seçici algoritmalarının; k-ortalama algoritması kümeleme tekniğinin; faktör analizinin ve lojistik regresyonun kullanıldığı bir tıbbi karar verme sistemi uygulanmıştır. Model için sınıflandırma doğruluğu, duyarlılık, özgüllük, kesinlik, f-measure ve kappa gibi performans ölçütleri sunulmuştur.

Modellenen karar verme sistemi yoğun tempoda çalışan hekimlerin işlerini kolaylaştırmakla beraber test sayısını azaltmayı da hedeflemiştir. Literatür araştırması sırasında vitamin D ile ilgili veri setinde veri madenciliği tekniklerinin kullanıldığı çalışmaların sayısının azlığı gözlemlendiğinden çalışma bu boşluğu doldurmayı amaçlamıştır.

Diyabet hastalarının veri seti üzerinde destek vektör makineleri kullanılarak % 98,030; diyabet hastalığı olmayan hastaların veri seti üzerinde nitelik seçimi sonrası yapay sinir ağları kullanılarak % 86,331 ve tüm veri seti üzerinde nitelik seçimi sonrası karar ağaçları kullanılarak % 83,479 oranlarında sınıflandırma doğruluğu sonuçlarına ulaşılmıştır.

En iyi sınıflandırma doğruluğunun sağlandığı 19 öznitelik içeren bir alt küme için destek vektör makineleri kullanılarak % 98,030 oranındaki sonuç ile yöntem araştırması sürecinde öne çıkan destek vektör makineleri yönteminin veri üzerinde de ayrıca başarı sağlayabildiği doğrulanmıştır.

Söz edilen 19 öznitelik ve vitamin D ile ilgili literatür taraması yenilenmiş ve nitelik seçimi sonuçları literatürle doğrulanmıştır. Bunun yanı sıra sınıflandırma

sonuçlarında herhangi bir doğruluk kaybı olmadan öznitelik sayısını azaltarak veri setini sınıflandırmak için gereken süre azalmaktadır.

5. SONUÇLAR VE ÖNERİLER

Uğur Engin EŞSİZ

Ayrıca bu sonucun literatürdeki benzer çalışmaların performanslarından daha iyi olduğu gözlemlenmiştir.

Bu sonuçlar göz önüne alındığında, öznitelik seçimi ile birleştirilmiş destek vektör makineleri sınıflandırıcı tabanlı model, diyabet hastalarının vitamin D eksikliğini sınıflandırmada çok başarılı sonuçlar vermiştir. Yöntemin, hekimlerin erken aşamada daha iyi klinik kararlar almasına yardımcı olmak için bir bilgi platformu olarak yardımcı olabileceği düşünülmektedir. Çalışma ayrıca laboratuvar kan testi sürecini kısaltmakta, doğrulanan vakaların tedavisine yönlendirmekte ve kan testlerinde azalma sağlamaktadır.

Gelecek çalışmalarda verinin genişletilmesi amaçlanmaktadır. Ayrıca veri üzerinde farklı sezgisel veya hibrit yöntemler denenerek sonuçların iyileştirilebileceği düşünülmektedir. Daha da önemlisi uygun veriye ulaşılabildiği takdirde farklı hastalıkların erken teşhisinin yapılabileceği öngörülmektedir.

KAYNAKLAR

Agarwal, K., Gupta, M., Gupta, K., Khan, A., & Nallakaruppan, M. K. (2019). Blood Transfusion System Using Data Mining Techniques and GRA. 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), 1143-1147.

<https://doi.org/10.1109/ICACCS.2019.8728401>

- Agresti, A. (1996). An Introduction to Categorical Data Analysis, John Wiley and Sons. Inc.
- Aha, D.W., Kibler, D. and Goldstone, R.L., "Instance-based learning algorithms", Machine Learning, 6:37-66 (1991)
- Akben, S. B. (2019). Determination of the Blood, Hormone and Obesity Value Ranges that Indicate the Breast Cancer, Using Data Mining Based Expert System. IRBM, 40(6), 355-360. <https://doi.org/10.1016/j.irbm.2019.05.007>
- Akgül, A., 2003. Tıbbi Araştırmalarda İstatistiksel Analiz Teknikleri, Emek Ofset Ltd.Şti., Ankara, 467s.
- Albayrak, A.S., 2006. Uygulamalı Çok Değişkenli İstatistik Teknikleri, Asil Yayın Dağıtım Ltd. Şti., Ankara, 499s.
- Alpar Reha. (2011), Uygulamalı Çok Değişkenli İstatistiksel Yöntemler, 3. Baskı, Ankara, Detay Yayınları, S.616
- American Diabetes Association. (2010). Diagnosis and Classification of Diabetes Mellitus. Diabetes Care, 33(Supplement_1), S62-S69. <https://doi.org/10.2337/dc10-S062>
- Amiri, A. M., & Armano, G. (2013). Segmentation and Feature Extraction of Heart Murmurs in Newborns. Journal of Life Sciences and Technologies, 107-111. <https://doi.org/10.12720/jolst.1.2.107-111>
- Amiri, Z., Nosrati, M., Sharifan, P., Saffar Soflaei, S., Darroudi, S., Ghazizadeh, H., Mohammadi Bajgiran, M., Moafian, F., Tayefi, M., Hasanzade, E., Rafiee, M., Ferns, G. A., Esmaily, H., Amini, M., & Ghayour-Mobarhan, M. (2021). Factors determining the serum 25-hydroxyvitamin D response to vitamin D supplementation: Data mining approach. BioFactors, biof.1770. <https://doi.org/10.1002/biof.1770>
- Armas, L. A. G., Hollis, B. W., & Heaney, R. P. (2004). Vitamin D₂ Is Much Less Effective than Vitamin D₃ in Humans. The Journal of Clinical

- Endocrinology & Metabolism, 89(11), 5387-5391.
<https://doi.org/10.1210/jc.2004-0360>
- Arpacı, S. A. & Kalıpsız, O. (2018). Yazılım Hata Sınıflandırmasında Farklı Naive Bayes Tekniklerin Kıyaslanması. Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi, 7 (1), 1-13. DOI: 10.28948/ngumuh.383709
- Safavian S.R., Landgrebe D., 1991, “A survey of decision tree classifier methodology”, IEEE Transactions on Systems Man and Cybernetics, 21, 660-674
- Ayyildiz, E., & Yalcin, S. (2018). Analysis of airports using clustering methods: Case study in Turkey. Pressacademia, 5(3), 194-205.
<https://doi.org/10.17261/Pressacademia.2018.963>
- Azizi-Soleiman, F., Vafa, M., Abiri, B., & Safavi, M. (2016). Effects of iron on Vitamin D metabolism: A systematic review. International Journal of Preventive Medicine, 7(1), 126. <https://doi.org/10.4103/2008-7802.195212>
- Bailey, K., 1987, Methods of Social Research, 3.baskı, New York:The Free PressLondon: Collier McMillan Publisher
- Baker, M. R., Peacock, M., & Nordin, B. E. C. (1980). The decline in Vitamin D status with Age. Age and Ageing, 9(4), 249-252.
<https://doi.org/10.1093/ageing/9.4.249>
- Batieha A, Khader Y, Jaddou H, Hyassat D, Batieha Z, Khateeb M, Belbisi A, Ajlouni K. Vitamin D status in Jordan: dress style and gender discrepancies. Ann Nutr Metab. 2011;58(1):10-8. doi: 10.1159/000323097. Epub 2011 Jan 19. PMID: 21252499.
- Batista, G.E.A.P.A. and Silva, D.F., “How k-nearest neighbor parameters affect its performance”, Simposio Argentino de Inteligencia Artificial (ASAI 2009), 95–106 (2009)
- Baxter, C.W., “Modelling Heuristics from Literature”, CIV E 729 Course Notes, Dept. of Civil and Environmental Engineering. University of Alberta, Edmonton, Canada, www.civil.ualberta.ca/courses/cive729

/lectures/lecture5.ppt, retrieved on 2004.

- Bhalla, Ashok. K., Amento, E. P., Clemens, T. L., Holick, M. F., & Krane, S. M. (1983). Specific High-Affinity Receptors for 1,25-dihydroxyvitamin D₃ in Human Peripheral Blood Mononuclear Cells: Presence in Monocytes and Induction in T Lymphocytes Following Activation. *The Journal of Clinical Endocrinology & Metabolism*, 57(6), 1308-1310. <https://doi.org/10.1210/jcem-57-6-1308>
- Bhatia, N. and Vandana, "Survey of nearest neighbor techniques", *International Journal of Computer Science and Information Security*, 8(2):302-305 (2010)
- Bikle, D. (2009). Nonclassic Actions of Vitamin D. *The Journal of Clinical Endocrinology & Metabolism*, 94(1), 26-34. <https://doi.org/10.1210/jc.2008-1454>
- Bogh MK, Schmedes AV, Philipsen PA, Thieden E, Wulf HC. Vitamin D production after UVB exposure depends on baseline vitamin D and total cholesterol but not on skin pigmentation. *J Invest Dermatol*. 2010 Feb;130(2):546-53. doi: 10.1038/jid.2009.323. Epub 2009 Oct 8. Erratum in: *J Invest Dermatol*. 2010 Jun;130(6):1751. PMID: 19812604.
- Bolland MJ, Grey A, Avenell A. Effects of vitamin D supplementation on musculoskeletal health: a systematic review, meta-analysis, and trial sequential analysis. *Lancet Diabetes Endocrinol*. 2018 Nov;6(11):847-858. doi: 10.1016/S2213-8587(18)30265-1. Epub 2018 Oct 4. PMID: 30293909.
- Bozkurt, Aybüke. (2021). Veri Madenciliği Yöntemleri ile Tam Kan Sayımı Sonuçlarından COVID-19 Test Sonuçlarının Tahmini.
- Brinati, D., Campagner, A., Ferrari, D., Locatelli, M., Banfi, G., & Cabitza, F. (2020). Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. *Journal of Medical Systems*, 44(8), 135. <https://doi.org/10.1007/s10916-020-01597-4>

- Bujak, K., Wasilewski, J., Osadnik, T., Jonczyk, S., Kołodziejska, A., Gierlotka, M., & Gąsior, M. (2015). The Prognostic Role of Red Blood Cell Distribution Width in Coronary Artery Disease: A Review of the Pathophysiology. *Disease Markers*, 2015, 1-12. <https://doi.org/10.1155/2015/824624>
- Carbone LD, Rosenberg EW, Tolley EA, Holick MF, Hughes TA, Watsky MA, Barrow KD, Chen TC, Wilkin NK, Bhattacharya SK, Dowdy JC, Sayre RM, Weber KT. 25-Hydroxyvitamin D, cholesterol, and ultraviolet irradiation. *Metabolism*. 2008 Jun;57(6):741-8. doi: 10.1016/j.metabol.2008.01.011. PMID: 18502255.
- Carnevale V, Modoni S, Pileri M, Di Giorgio A, Chiodini I, Minisola S, Vieth R, Scillitani A. Longitudinal evaluation of vitamin D status in healthy subjects from southern Italy: seasonal and gender differences. *Osteoporos Int*. 2001 Dec;12(12):1026-30. doi: 10.1007/s001980170012. PMID: 11846328.
- Carsten, P., Thorstein R., 1993, “Jetnet 3.0: A Versatile Artificial Neural Network Package”, www.thep.lu.se/pub/Preprints/93/lutp-93-29.ps.gz, retrieved on 2004, Sweden.
- Carter, D. C., & Ho, J. X. (1994). Structure of Serum Albumin. *İçinde Advances in Protein Chemistry* (C. 45, ss. 153-203). Elsevier. [https://doi.org/10.1016/S0065-3233\(08\)60640-3](https://doi.org/10.1016/S0065-3233(08)60640-3)
- Celik, Y., Sabanci, K., Durdu, A., & Aslan, M. F. (2018). Breast Cancer Diagnosis by Different Machine Learning Methods Using Blood Analysis Data. *International Journal of Intelligent Systems and Applications in Engineering*, 6(4), 289-293. <https://doi.org/10.18201/ijisae.2018648455>
- Centers for Disease Control and Prevention. National Diabetes Statistics Report, 2020. Atlanta, GA: Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services; 2020.
- Chiu, K. C., Chu, A., Go, V. L. W., & Saad, M. F. (2004). Hypovitaminosis D is associated with insulin resistance and β cell dysfunction. *The American Journal of Clinical Nutrition*, 79(5), 820-825.

<https://doi.org/10.1093/ajcn/79.5.820>

- Colston, K., & Cleeve, H. J. W. (1986). Effect of triiodothyronine on intestinal and kidney isoenzymes of alkaline phosphatase and on vitamin D metabolism in adult female rats. *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry*, 83(3), 681-684. [https://doi.org/10.1016/0305-0491\(86\)90317-2](https://doi.org/10.1016/0305-0491(86)90317-2)
- Cook, D. (2008). Binary response and logistic regression analysis. www.public.iastate.edu/~stat415/stephenson/stat415_chapter3.pdf
- Coomans, D and Massart, D.L., “Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. Knearest neighbour classification by using alternative voting rules”, *Analytica Chimica Acta*, 136: 15-27 (1982)
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297. <https://doi.org/10.1007/BF00994018>
- Coşkun Hamzaçebi ve Fevzi Kutay Yapay Sinir Ağları ile Türkiye Elektrik Enerjisi Tüketiminin 2010 Yılına Kadar Tahmini Endüstri Mühendisliği Bölümü, Mühendislik Mimarlık Fakültesi, Gazi Üniversitesi Maltepe, Ankara
- Cover, T.M. and Hart, P.E., “Nearest neighbor pattern classification”. *IEEE Transactions on Information Theory*, IT13(1):21–27 (1967)
- Cure, M. C., Cure, E., Yuce, S., TarkanYazici, Karakoyun, I., & Efe, H. (2014). Mean Platelet Volume and Vitamin D Level. *Annals of Laboratory Medicine*, 34(2), 98-103. <https://doi.org/10.3343/alm.2014.34.2.98>
- Deboeck, G. J., 1994, *Trading on the Edge*, John Wiley & Sons Inc., New York.
- DeLuca, H. F. (2009). The Vitamin D System in the Regulation of Calcium and Phosphorus Metabolism. *Nutrition Reviews*, 37(6), 161-193. <https://doi.org/10.1111/j.1753-4887.1979.tb06660.x>
- Deng, H., Sun, Y., Chang, Y., Han, J., Probabilistic Models for Classification. C.C. AGGARWAL (Eds.), *Data Classification Algorithms and Applications* (pp. 67-70), CRC Press, New York, USA, 2015.

- Dinçer, E.Ş. Veri Madenciliğinde K-means Algoritması ve Tıp Alanında Uygulanması, Yüksek Lisans Tezi, Kocaeli Üniversitesi Fen Bilimleri Enstitüsü, Kocaeli, 2006.
- Duda, R.O., Hart, P.E. and Stork, D.G., “Pattern Classification”, John Wiley & Sons, New Jersey, (2000)
- Dunteman, G. H., 1989. Principal component analysis (Quantitative applications in the social sciences).
- Fidan, F., Alkan, B. M., & Tosun, A. (2014). Çağın Pandemisi: D Vitamini Eksikliği ve Yetersizliği. *Türk Osteoporoz Dergisi*, 20(2), 71-74.
<https://doi.org/10.4274/tod.94830>
- Florez-Lopez, R. (2002). Reviewing Relief and Its Extensions: A New Approach for Estimating Attributes Considering High-Correlated Features. 2002 IEEE International Conference on Data Mining, 2002. Proceedings., 605-608. <https://doi.org/10.1109/ICDM.2002.1184009>
- Gallant, S.I., 1993, Neural Network Learning and Expert Systems, MIT Press, Cambridge.
- Garcia Carretero, R., Vigil-Medina, L., Barquero-Perez, O., Mora-Jimenez, I., Soguero-Ruiz, C., & Ramos-Lopez, J. (2021). Machine learning approaches to constructing predictive models of vitamin D deficiency in a hypertensive population: A comparative study. *Informatics for Health and Social Care*, 1-15. <https://doi.org/10.1080/17538157.2021.1896524>
- Gartner, T., Lloyd, J.W. and Flach, P.A., “Kernels and Distances for Structured Data”, *Machine Learning*, 57(3):205-232 (2004)
- GeethaRamani, R., & Balasubramanian, L. (2016). Retinal blood vessel segmentation employing image processing and data mining techniques for computerized retinal image analysis. *Biocybernetics and Biomedical Engineering*, 36(1), 102-118. <https://doi.org/10.1016/j.bbe.2015.06.004>

- Goldsmith, J. (2015). Vitamin D as an Immunomodulator: Risks with Deficiencies and Benefits of Supplementation. *Healthcare*, 3(2), 219-232. <https://doi.org/10.3390/healthcare3020219>
- Gonoodi, K., Tayefi, M., Saberi-Karimian, M., Amirabadi zadeh, A., Darroudi, S., Farahmand, S. K., Abasalti, Z., Moslem, A., Nematy, M., Ferns, G. A., Eslami, S., & Mobarhan, M. G. (2019). An assessment of the risk factors for vitamin D deficiency using a decision tree model. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 13(3), 1773-1777. <https://doi.org/10.1016/j.dsx.2019.03.020>
- Goodnough, L. T., & Shah, N. (2014). The Next Chapter in Patient Blood Management. *American Journal of Clinical Pathology*, 142(6), 741-747. <https://doi.org/10.1309/AJCP4W5CCFOZUJFU>
- Gökçe, B. (1988). Toplumsal Bilimlerde Araştırma. Ankara: Savaş Yayınları.
- Guariguata, L., Whiting, D. R., Hambleton, I., Beagley, J., Linnenkamp, U., & Shaw, J. E. (2014). Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes Research and Clinical Practice*, 103(2), 137-149. <https://doi.org/10.1016/j.diabres.2013.11.002>
- Gujarati, D.N. (1999). Temel ekonometri. (Çev: Ü. Şenesen ve G.G. Şenesen), Literatür Yayıncılık, İstanbul.
- Hagenau T, Vest R, Gissel TN, Poulsen CS, Erlandsen M, Mosekilde L, Vestergaard P. Global vitamin D levels in relation to age, gender, skin pigmentation and latitude: an ecologic meta-regression analysis. *Osteoporos Int*. 2009 Jan;20(1):133-40. doi: 10.1007/s00198-008-0626-y. Epub 2008 May 6. PMID: 18458986.
- Halk Sağlığı Laboratuvarı İl Sağlık Müdürlüğü Ek Hizmet Binası Halk Sağlığı Hizmetleri Başkanlığı Reşatbey Mah. 5 Ocak Cad. No:53 01120 Seyhan/ADANA
- Hair J. F., Anderson, R. E., Tatham, R. L., William, C. B., 1998, Multivariate Data Analysis, 5th edition. Sharma, S., Kumar, A., 2006. Factor analysis: The

Handbook of Marketing Research: Uses, Misuses, and Future Advances,
365

Han, J. and Kamber, M., "Data mining: concepts and techniques", Morgan Kaufmann Publishers, Burlington, (2006)

HAN, J., LEE, J.G., KAMBER, M., An Overview of Clustering Methods in Geographic Data Analysis, In Miller H.J., Han H. (Eds.) Geographic Data Mining and Knowledge Discovery, Taylor & Francis Group, LLC, 2009.

Haykin, S., Neural Networks: A Comprehensive Foundation, Perenctice Hall, New Jersey, 1999.

Hill, T., O'Connor, M., Remus, W., "Neural Networks Models for Time Series Forecasts", Management Sciences, Cilt 42, No 7, 1082-1092, 1996

Holick, M. F. (2006). High Prevalence of Vitamin D Inadequacy and Implications for Health. Mayo Clinic Proceedings, 81(3), 353-373.
<https://doi.org/10.4065/81.3.353>

Holick, M. F., Biancuzzo, R. M., Chen, T. C., Klein, E. K., Young, A., Bibuld, D., Reitz, R., Salameh, W., Ameri, A., & Tannenbaum, A. D. (2008). Vitamin D2 Is as Effective as Vitamin D3 in Maintaining Circulating Concentrations of 25-Hydroxyvitamin D. The Journal of Clinical Endocrinology & Metabolism, 93(3), 677-681.
<https://doi.org/10.1210/jc.2007-2308>

Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda, Machine Learning and Data Mining Methods in Diabetes Research, Computational and Structural Biotechnology Journal, Volume 15, 2017, Pages 104-116, ISSN 2001-0370,
<https://doi.org/10.1016/j.csbj.2016.12.005>.

International Diabetes Federation. (2015). IDF diabetes atlas. International Diabetes Federation.

Kaastra, I., Boyd, M., "Designing a Neural Network for Forecasting Financial and Economic Time Series", Neurocomputing, Cilt 10, 215- 236, 1996.

- Kalaycı, Ş., 2008. SPSS Uygulamalı Çok Değişkenli İstatistik Teknikleri. Ankara: Asil Yayıncılık
- Karagöl A, Atak N. ;D vitamini ve Tip 2 Diyabet. Turkish Journal of Public Health, 2016.
- Karthikeyan A, Garg A, Vinod PK and Priyakumar UD (2021) Machine Learning Based Clinical Decision Support System for Early COVID-19 Mortality Prediction. Front. Public Health 9:626697. doi: 10.3389/fpubh.2021.626697
- Kavzoğlu, T. & Çölkesen, İ. (2010). Karar Ağaçları İle Uydu Görüntülerinin Sınıflandırılması: Harita Teknolojileri Elektronik Dergisi, 2 (1), 36-45.
- Kavzoğlu T, Çölkeşen İ (2010). Destek vektör makineleri ile uydu görüntülerinin sınıflandırılmasında kernel fonksiyonlarının etkilerinin incelenmesi. Harita Dergisi, 76(144), 73 - 82.
- Kaya, T., Akçay, E. Ü., Ertürk, Z., Ergenç, H., & Tamer, A. (2018). The relationship between vitamin D deficiency and erythrocyte sedimentation rate in patients with diabetes. Turkish Journal of Medical Sciences, 48(2), 424-429. <https://doi.org/10.3906/sag-1712-28>
- Kaynar, O., Arslan, H., Görmez, Y., & Işık, Y. E. (2018). Makine Öğrenmesi ve Öznitelik Seçim Yöntemleriyle Saldırı Tespiti. Bilişim Teknolojileri Dergisi. <https://doi.org/10.17671/gazibtd.368583>
- Kazlauskaitė, R., Powell, L. H., Mandapakala, C., Cursio, J. F., Avery, E. F., & Calvin, J. (2010). Vitamin D is associated with atheroprotective highdensity lipoprotein profile in postmenopausal women. Journal of Clinical Lipidology, 4(2), 113-119. <https://doi.org/10.1016/j.jacl.2010.01.006>
- Keller, J. M., Gray, M.R. and Givens, J.A., “A fuzzy K-nearest neighbor algorithm”, IEEE Transactions on Systems, Man and Cybernetics, SMC15(4):580-585 (1985)
- Komi, M., Jun Li, Yongxin Zhai, & Xianguo Zhang. (2014). Application of data mining methods in diabetes prediction. 2014 2nd International Conference on Image, Vision and Computing (ICIVC), 1006-1010.

<https://doi.org/10.1109/ICIVC.2017.7984706>

- Konda, S., B.Kavihta, R., & Govardhan, D. (2010). Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. *International Journal on Computer Science and Engineering*.
- Kurtuluş, Kemal (1998), “Pazarlama Araştırmaları”, Avcıol Basım-Yayın, 6. bası, s. 495- 506.
- Lagunova Z, Porojnicu AC, Lindberg F, Hexeberg S, Moan J. The dependency of vitamin D status on body mass index, gender, age and season. *Anticancer Res.* 2009 Sep;29(9):3713-20. PMID: 19667169.
- Leslie, W., & Hankey, C. (2015). Aging, Nutritional Status and Health. *Healthcare*, 3(3), 648-658. <https://doi.org/10.3390/healthcare3030648>
- Liao, Z., Ju, Y., and Zou, Q. (2016). Prediction of G protein-coupled receptors with SVM-Prot features and random forest. *Scientifica*, 8309253. doi: 10.1155/2016/8309253.
- Lippeveld, M., Knill, C., Ladlow, E., Fuller, A., Michaelis, L. J., Saeys, Y., Filby, A., & Peralta, D. (2019). Classification of Human White Blood Cells Using Machine Learning for Stain-Free Imaging Flow Cytometry. *Cytometry Part A*, 97(3), 308-319. <https://doi.org/10.1002/cyto.a.23920>
- Liu, H. and Zhang, S., “Noisy data elimination using mutual k-nearest neighbor for classification mining”, *Journal of Systems and Software*, 85(5):1067-1074 (2012)
- M. El-Halees, A., & H. Shurrah, A. (2017). Blood Tumor Prediction Using Data Mining Techniques. *Health Informatics - An International Journal*, 6(2), 23-30. <https://doi.org/10.5121/hij.2017.6202>
- Magoulas G.D., Prentza A. (2001) Machine Learning in Medical Applications. In: Paliouras G., Karkaletsis V., Spyropoulos C.D. (eds) *Machine Learning and Its Applications*. ACAI 1999. Lecture Notes in Computer Science, vol 2049. Springer, Berlin, Heidelberg. <https://doi.org/10.1007/3-540-44673->

- Mao, C., Hu, B., Wang, M. and Moore, P., "Learning from neighborhood for classification with local distribution characteristics", IEEE International Joint Conference on Neural Networks (IJCNN), 1-8 (2015)
- Marcus, Y., Eldor, R., Yaron, M., Shaklai, S., Ish-Shalom, M., Shefer, G., Stern, N., Golan, N., Dvir, A. Z., Pele, O., & Gonen, M. (2020). Improving blood glucose level predictability using machine learning. *Diabetes/Metabolism Research and Reviews*, 36(8). <https://doi.org/10.1002/dmrr.3348>
- Maria & Yassine. (2020). Machine Learning Based Approaches for Modeling the Output Power of Photovoltaic Array in Real Outdoor Conditions. *Electronics*, 9(2), 315. <https://doi.org/10.3390/electronics9020315>
- Matheny, M & Ohno-Machado, L. (2014). Generation of Knowledge for Clinical Decision Support. 227-248. 10.1016/B978-012369377-8/50011-8.
- Melamed, M. L., Michos, E. D., Post, W., & Astor, B. (2008). 25-Hydroxyvitamin D Levels and the Risk of Mortality in the General Population. *Archives of Internal Medicine*, 168(15), 1629-1637. <https://doi.org/10.1001/archinte.168.15.1629>
- Mertler, C. A., & Vannatta, R. A. (2005). *Advanced and multivariate statistical methods: Practical application and interpretation* (3rd ed.). Glendale, CA: Pyrczak Publishing.
- Methaila, A., Kansal, P., Arya, H., & Kumar, P. (2014). Early Heart Disease Prediction Using Data Mining Techniques. *Computer Science & Information Technology (CS & IT)*, 53-59. <https://doi.org/10.5121/csit.2014.4807>
- Mitchell, T., "Machine Learning", McGraw Hill, New York, (1997)
- Nagendra, S., 1998, "Practical Aspects of Using Neural Networks: Necessary Preliminary Specifications", GE Research and Development Center Technical Paper.
- Nehdi, M., Debbar, Y., Khn, A., 2001b, "Neural Network Model for PreformedFoam Cellular Concrete", *ACI Materials Journal*, Vol:98, No:5, s:402-410.

- Nurminen, V., Seuter, S., & Carlberg, C. (2019). Primary Vitamin D Target Genes of Human Monocytes. *Frontiers in Physiology*, 10, 194.
<https://doi.org/10.3389/fphys.2019.00194>
- Napoli, N., Chandran, M., Pierroz, D. D., Abrahamsen, B., Schwartz, A. V., & Ferrari, S. L. (2017). Mechanisms of diabetes mellitus-induced bone fragility. *Nature Reviews Endocrinology*, 13(4), 208-219.
<https://doi.org/10.1038/nrendo.2016.153>
- Oğuzlar, A. (2010). Lojistik regresyon analizi yardimiyla suçlu profilinin belirlenmesi . *Atatürk Üniversitesi İktisadi ve İdari Bilimler Dergisi* , 19 (1) , 21-35.
- Önder, Hasan & Cebeci, Zeynel. (2002). Lojistik Regresyonlarda Değişken Seçimi. *Çukurova Üniv. Ziraat Fakültesi Dergisi*. 17. 105–114
- Pal M., Mather P.M., 2003, “An assessment of the effectiveness of decision tree methods for land cover classification”, *Remote Sensing of Environment*, 86, 554-565
- Palaniappan, S., & Awang, R. (2008). Intelligent heart disease prediction system using data mining techniques. 2008 IEEE/ACS International Conference on Computer Systems and Applications, 108-115.
<https://doi.org/10.1109/AICCSA.2008.4493524>
- Palomer, X., González-Clemente, J. M., Blanco-Vaca, F., & Mauricio, D. (2008). Role of vitamin D in the pathogenesis of type 2 diabetes mellitus. *Diabetes, Obesity and Metabolism*, 10(3), 185-197. <https://doi.org/10.1111/j.1463-1326.2007.00710.x>
- Parvoti, F.Y., Anandrajan, M., 2002, “Classifying Inventory Using an Artificial Neural Network Approach”, *Computers and Industrial Engineering*, Vol:41, s:389-404
- Pittas, A. G., & Dawson-Hughes, B. (2010). Vitamin D and diabetes. *The Journal of Steroid Biochemistry and Molecular Biology*, 121(1-2), 425-429.

<https://doi.org/10.1016/j.jsbmb.2010.03.042>

Pittas, A. G., Sun, Q., Manson, J. E., Dawson-Hughes, B., & Hu, F. B. (2010). Plasma 25-Hydroxyvitamin D Concentration and Risk of Incident Type 2 Diabetes in Women. *Diabetes Care*, 33(9), 2021-2023. <https://doi.org/10.2337/dc10-0790>

Polat, Y., 2012. Faktör Analizi Yöntemlerinin Karşılaştırmalı Olarak İncelenmesi ve Hayvancılık Denemesine Uygulanışı. Yayınlanmamış Doktora Lisans Tezi, Çukurova Üniversitesi Fen Bilimleri Enstitüsü, Adana

Ponda, M. P., Dowd, K., Finkelstein, D., Holt, P. R., & Breslow, J. L. (2012). The Short-Term Effects of Vitamin D Repletion on Cholesterol. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 32(10), 2510–2515.

<https://doi.org/10.1161/atvbaha.112.254110>

Portela, F., Santos, M. F., Machado, J., Abelha, A., Rua, F., & Silva, Á. (2015). Real-Time Decision Support Using Data Mining to Predict Blood Pressure Critical Events in Intensive Medicine Patients. İçinde J. Bravo, R. Hervás, & V. Villarreal (Ed.), *Ambient Intelligence for Health* (C. 9456, ss. 77-90). Springer International Publishing. https://doi.org/10.1007/978-3-319-26508-7_8

Padmaja B , Battu Ramya Reddy , R Vikrant Sagar , Heetesh Kumar Pradhan , G Chandra Sekhar , E Krishna Rao Patro (2021) Prognosis of Vitamin D Deficiency Severity using SMOTE optimized Machine Learning Models Assistant Professor, Department of CSE, Institute of Aeronautical Engineering, Hyderabad, India b, c, d Student, Department of CSE, Institute of Aeronautical Engineering, Hyderabad, India

Powers Daniel A. ve Xie Yu (2000), *Statistical Methods for Categorical Data Analysis*, Academic Press, ABD.

Provvedini, D., Tsoukas, C., Deftos, L., & Manolagas, S. (1983). 1,25dihydroxyvitamin D₃ receptors in human leukocytes. *Science*, 221(4616), 1181-1183. <https://doi.org/10.1126/science.6310748>

- Qiu, X.Y., Kang, K. and Zhang, H.X., “Selection of kernel parameters for K-NN”, IEEE International Joint Conference on Neural Networks (IJCNN), 61-65 (2008)
- Quinlan J.R., 1993, “C4.5: Programs for Machine Learning”, Morgan Kaufmann, San Mateo, CA, 302 s
- R. Kohavi and F. Provost Glossary of Terms. (1998). Machine Learning, 30(2), 271-274. <https://doi.org/10.1023/A:1017181826899>
- Reboul E, Goncalves A, Comera C, Bott R, Nowicki M, Landrier JF, JourdeuilRahmani D, Dufour C, Collet X, Borel P. Vitamin D intestinal absorption is not a simple passive diffusion: evidences for involvement of cholesterol transporters. Mol Nutr Food Res. 2011 May;55(5):691-702. doi: 10.1002/mnfr.201000553. Epub 2011 Jan 31. PMID: 21280209.
- Robnik-Sikonja M., & Kononenko I. (2003). Theoretical and Empirical Analysis of ReliefF and ReliefF. Machine Learning Journal, 23-69
- Rodríguez-Rodríguez, E., Ortega, R. M., González-Rodríguez, L. G., & LópezSobaler, A. M. (2011). Vitamin D deficiency is an independent predictor of elevated triglycerides in Spanish school children. UCM Research Group Valornut (920030), European Journal of Nutrition, 50(5), 373-378. <https://doi.org/10.1007/s00394-010-0145-4>
- Rosenblatt, F., 1961, Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms, Spartan Press, Washington DC.
- Ross, A. C., Manson, J. E., Abrams, S. A., Aloia, J. F., Brannon, P. M., Clinton, S. K., Durazo-Arvizu, R. A., Gallagher, J. C., Gallo, R. L., Jones, G., Kovacs, C. S., Mayne, S. T., Rosen, C. J., & Shapses, S. A. (2011). The 2011 Report on Dietary Reference Intakes for Calcium and Vitamin D from the Institute of Medicine: What Clinicians Need to Know. The Journal of Clinical Endocrinology & Metabolism, 96(1), 53-58. <https://doi.org/10.1210/jc.2010-2704>
- Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986, “Learning Internal Representations by Error Propagation”, Parallel Distributed Processing:

- Explorations in the Microstructures of Cognition, Rumelhart DE., McClelland, J.L. (eds.), Vol:1, s:318-362. MIT Press, Cambridge.
- Saichanma, S., Chulsomlee, S., Thangrua, N., Pongsuchart, P., & Sanmun, D. (2014). The Observation Report of Red Blood Cell Morphology in Thailand Teenager by Using Data Mining Technique. *Advances in Hematology*, 2014, 1-5. <https://doi.org/10.1155/2014/493706>
- Sambasivam, G., Amudhavel, J., & Sathya, G. (2020). A Predictive Performance Analysis of Vitamin D Deficiency Severity Using Machine Learning Methods. *IEEE Access*, 8, 109492-109507. <https://doi.org/10.1109/ACCESS.2020.3002191>
- Sarıkoç, F., “Meme Kanseri Biyopsi Örneklerinde Östrojen Hormonu Alıcı Varlığının Zeki Yöntemlerle Tespiti”, Doktora Tezi, Erciyes Üniversitesi, Fen Bilimleri Enstitüsü, Kayseri, 148s, (2012)
- Scholkopf, B., Kah-Kay, S., Burges, C. J., Girosi, F., Niyogi, P., Poggio, T., et al. (1997). Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing*, 45(11), 2758–2765
- Schwetz, V., Scharnagl, H., Trummer, C., Stojakovic, T., Pandis, M., Grüber, M. R., Verheyen, N., Gaksch, M., Zittermann, A., Aberer, F., Lerchbaum, E., Obermayer-Pietsch, B., Pieber, T. R., März, W., Tomaschitz, A., & Pilz, S. (2018). Vitamin D supplementation and lipoprotein metabolism: A randomized controlled trial. *Journal of Clinical Lipidology*, 12(3), 588-596.e4. <https://doi.org/10.1016/j.jacl.2018.03.079>
- Seif, A. A., & Abdelwahed, D. M. (2014). Vitamin D ameliorates hepatic ischemic/reperfusion injury in rats. *Journal of Physiology and Biochemistry*, 70(3), 659-666. <https://doi.org/10.1007/s13105-014-0335-2>
- Signorello, L. B., Schlundt, D. G., Cohen, S. S., Steinwandel, M. D., Buchowski, M. S., McLaughlin, J. K., Hargreaves, M. K., & Blot, W. J. (2007). Comparing Diabetes Prevalence Between African Americans and Whites of Similar

- Socioeconomic Status. American Journal of Public Health, 97(12), 2260-2267. <https://doi.org/10.2105/AJPH.2006.094482>
- Solmaz Ramazan, Günay Mücahid, Alkan Ahmet (2014) Fonksiyonel Tiroit Hastalığı Tanısında Naive Bayes Sınıflandırıcının Kullanılması Elektrik Elektronik Mühendisliği Bölümü , KSÜ, Kahramanmaraş
- Sowell, K., Keen, C., & Uriu-Adams, J. (2015). Vitamin D and Reproduction: From Gametes to Childhood. Healthcare, 3(4), 1097-1120. <https://doi.org/10.3390/healthcare3041097>
- Süzülmüş, S., 2005. Faktör Analizi Modellerinin Belirlenebilirliği Ve Genelleştirilmiş İnverslerin Kullanımı, Doktora Tezi, Çukurova Üniversitesi Fen Bilimleri Enstitüsü, Adana, 118s.
- Tai, K., Need, A. G., Horowitz, M., & Chapman, I. M. (2008). Vitamin D, glucose, insulin, and insulin sensitivity. Nutrition, 24(3), 279-285. <https://doi.org/10.1016/j.nut.2007.11.006>
- Tang, Z., Almeida, C., Fishwick, P.A., “Time Series Forecasting Using Neural Networks vs Box-Jenkins Methodology”, Simulation, Cilt 57, No 5, 303-310, 1991
- Tanner, C., 2005, “Accelerating Artificial Neural Network Learning via Weight Predictions”, www.cs.ucla.edu/~ctanner/papers/01Accelerating_NN_Learning_04_2005.pdf, retrieved on 2009.
- Tariq, S., Tariq, S., & Lone, K. P. (2018). Interplay of vitamin D, vitamin B₁₂, homocysteine and bone mineral density in postmenopausal females. Health Care for Women International, 39(12), 1340-1349. <https://doi.org/10.1080/07399332.2018.1444042>
- Taşcı, E., & Onan, A. (2017). K-En Yakın Komşu Algoritması Parametrelerinin Sınıflandırma Performansı Üzerine Etkisinin İncelenmesi .
- Thomas, J., & Princy, R. T. (2018). Human heart disease prediction system using data mining techniques. 2018 International Conference on Circuit, Power and Computing Technologies (ICCPCT), 1-5.

- Thorand, B., Zierer, A., Huth, C., Linseisen, J., Meisinger, C., Roden, M., Peters, A., Koenig, W., & Herder, C. (2011). Effect of Serum 25-Hydroxyvitamin D on Risk for Type 2 Diabetes May Be Partially Mediated by Subclinical Inflammation: Results from the MONICA/KORA Augsburg study. *Diabetes Care*, 34(10), 2320-2322. <https://doi.org/10.2337/dc11-0775>
- Turhan, Ozlem & Eroğlu Altınova, Alev. (2012). Vitamin D and Diabetes Mellitus. *Turkish Journal of Endocrinology and Metabolism*. 16. 89-94. 10.4274/Tjem.1968.
- Uçar F, Taşlıpınar MY, Soydaş AÖ, Özcan N. Ankara Etlik İhtisas Eğitim Araştırma Hastanesi'ne Başvuran Hastalarda 25-OH Vitamin D Düzeyleri. *Eur J Basic Med Sci* 2012;2:12-5
- Uriu-Adams, J. Y., Obican, S. G., & Keen, C. L. (2013). Vitamin D and maternal and child health: Overview and implications for dietary requirements: Vitamin D and Maternal and Child Health. *Birth Defects Research Part C: Embryo Today: Reviews*, 99(1), 24-44. <https://doi.org/10.1002/bdrc.21031>
- Uwitonze, A. M., & Razzaque, M. S. (2018). Role of Magnesium in Vitamin D Activation and Function. *The Journal of the American Osteopathic Association*, 118(3), 181. <https://doi.org/10.7556/jaoa.2018.037>
- Veranyurt Ü., Deveci A. F., Esen M. F. (2020) Makine öğrenmesi teknikleriyle hastalık sınıflandırması: random forest, k-nearest neighbour ve adaboost algoritmaları uygulaması
- Verdoia, M., Schaffer, A., Barbieri, L., Di Giovine, G., Marino, P., Suryapranata, H., & De Luca, G. (2015). Impact of gender difference on vitamin D status and its relationship with the extent of coronary artery disease. *Nutrition, Metabolism and Cardiovascular Diseases*, 25(5), 464-470.

<https://doi.org/10.1016/j.numecd.2015.01.009>

Wabke, T. K., Sorg, B. L., & Steinhilber, D. (2014). Vitamin D in inflammatory diseases. *Frontiers in Physiology*, 5.

<https://doi.org/10.3389/fphys.2014.00244>

Weyland, P., Grant, W., & Howie-Esquivel, J. (2014). Does Sufficient Evidence Exist to Support a Causal Association between Vitamin D Status and Cardiovascular Disease Risk? An Assessment Using Hill's Criteria for Causality. *Nutrients*, 6(9), 3403-3430. <https://doi.org/10.3390/nu6093403>

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed). Morgan Kaufmann.

Wolf, T. A., Wimalawansa, S. J., & Razzaque, M. S. (2019). Procalcitonin as a biomarker for critically ill patients with sepsis: Effects of vitamin D supplementation. *The Journal of Steroid Biochemistry and Molecular Biology*, 193, 105428. <https://doi.org/10.1016/j.jsbmb.2019.105428>

Wu, J., Zhang, P., Zhang, L., Meng, W., Li, J., Tong, C., Li, Y., Cai, J., Yang, Z., Zhu, J., Zhao, M., Huang, H., Xie, X., & Li, S. (2020). Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results [Preprint]. *Infectious Diseases (except HIV/AIDS)*. <https://doi.org/10.1101/2020.04.02.20051136>

Wu, T. H., Pang, G. K.-H., & Kwong, E. W.-Y. (2020). Predicting Systolic Blood Pressure Using Machine Learning. *7th International Conference on Information and Automation for Sustainability*, 1-6. <https://doi.org/10.1109/ICIAFS.2014.7069529>

Yao, H., Zhang, N., Zhang, R., Duan, M., Xie, T., Pan, J., Peng, E., Huang, J., Zhang, Y., Xu, X., Xu, H., Zhou, F., & Wang, G. (2020). Severity Detection for the Coronavirus Disease 2019 (COVID-19) Patients Using a Machine Learning Model Based on the Blood and Urine Tests. *Frontiers in Cell and Developmental Biology*, 8, 683.

<https://doi.org/10.3389/fcell.2020.00683>

Young, P. V., 1968., Bilimsel Sosyal İncelemeler ve Araştırma (Çeviri. G. Bingöl ve N. İşçil). Ankara: Ege Matbaası.

Yuhang Wang, & Makedon, F. (2004). Application of relief-F feature filtering algorithm to selecting informative genes for cancer classification using microarray data. Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004., 477-478.

<https://doi.org/10.1109/CSB.2004.1332474>

Zhang, G., Patuwo, B.E., Hu, M.Y., “Forecasting with Artificial Neural Networks: The State of the Art”, Inter. Journal of Forecasting, Cilt 14, 35- 62, 1998.

ÖZGEÇMİŞ

Uğur Engin EŞSİZ, İlkokul, ortaokul ve lise eğitimlerini Adana'da tamamladı. 2008 yılında Çukurova Üniversitesi Elektrik Elektronik Mühendisliği Bölümü'nden mezun oldu. 2013 yılında Çukurova Üniversitesi İşletme Yüksek Lisans Programı'ndan mezun oldu. 2009-2010 yılları arasında bir yapı denetim firmasında kontrol mühendisi olarak görev aldı. 2010 yılında merkezi atama ile Adana Büyükşehir Belediyesi'nde memuriyet hayatına başladı. 2011 yılı itibariyle Adana Su ve Kanalizasyon İdaresi'nde çalışmaktadır. Evli ve bir erkek çocuk babasıdır.



EKLER



Ek 1: Tıpta makine öğrenmesi çalışmalarının literatür araştırması

Yazar	Çalışma	Metod	Sınıflandırma Doğruluğu (%)
Ghani and Awang (2008)	Kalp hastalığı tahmini	Naive bayes	86,120 %
Srinivas ve ark. (2010)	Kalp krizi tahmini	Naive bayes	84,050 %
Methaila ve ark. (2014)	Kalp hastalığı tahmini	Karar ağaçları	89,000 %
Komi ve ark. (2014)	Diyabet tahmini	Yapay sinir ağları	89,000 %
Chulsomlee ve ark. (2014)	Kırmızı kan hücresi morfolojisi	Karar ağaçları	94,300 %
Portela ve ark. (2015)	Kan basıncı tahmini	Karar ağaçları	71,520 %
GeethaRamani and Balasubramanian (2016)	Retinal kan damarı segmentasyonu	Root guided decision tree	95,360 %
El-Halees and Shurrah (2017)	Kan tümörü tahmini	Derin öğrenme	79,450 %
Aslan ve ark. (2018)	Meme kanseri tanısı	Extreme learning machine	83,870 %
Thomas and Princy (2018)	Kalp hastalığı tahmini	Karar ağaçları	80,600 %
Lippeveld ve ark. (2019)	Beyaz kan hücrelerinin sınıflandırılması	Gradient boosting	87,100 %
Agarwal ve ark. (2019)	Kan transfüzyon sistemi	Rastgele orman	93,200 %
Akben (2019)	Kan, hormon ve obezite değer aralıklarının belirlenmesi	Karar ağaçları	90,500 %
Gonoodi ve ark. (2019)	Vitamin D eksikliğinin risk faktörlerinin değerlendirilmesi	Karar ağaçları	77,800 %
Marcus ve ark. (2020)	Kan glikoz düzeyi tahmini	Kernel ridge regression	93,400 %
Brinati ve ark. (2020)	COVID-19 enfeksiyonunun tespiti	Rastgele orman	89,100 %
Wu ve ark. (2020)	COVID-19 enfeksiyonunun tanımlanması	Rastgele orman	97,950 %
Wu ve ark. (2020)	Sistolik kan basıncı tahmini	Yapay sinir ağları	95,500 %
Yao ve ark. (2020)	COVID-19 önem tespiti	Destek vektör makineleri	81,480 %
Sambasivam ve ark. (2020)	Vitamin D eksikliğinin öneminin performans analizi	Rastgele orman	96,000 %
Amiri ve ark. (2021)	Vitamin D'nin takviyeye cevabını etkileyen faktörlerin belirlenmesi	Rastgele orman	93,000 %

Padmajaa ve ark. (2021)	Vitamin D eksikliğinin öneminin tahmini	Extra trees classifier	73,300 %
Carretero ve ark. (2021)	Vitamin D eksikliğinin tahmin modelleri	Destek vektör makineleri	73,000 %

Ek 2: Yapay Sinir Ağları Çıktısı

```

Correctly Classified Instances      535          78.2164 %
Kappa statistic                    0.5242
Mean absolute error                0.2276
Root mean squared error            0.4363
Relative absolute error             50.2476 %
Root relative squared error        91.687 %
Total Number of Instances         684

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,821	0,291	0,842	0,821	0,831	0,524	0,829	0,884	0
	0,709	0,179	0,677	0,709	0,693	0,524	0,829	0,731	1
Weighted Avg.	0,782	0,252	0,785	0,782	0,783	0,524	0,829	0,831	

```

=== Confusion Matrix ===
      a    b  <-- classified as
367  80 |   a = 0
 69 168 |   b = 1

```

Ek 3: Karar Ağaçları Çıktısı

```

Correctly Classified Instances      538          78.655 %
Kappa statistic                    0.4848
Mean absolute error                0.3562
Root mean squared error            0.3988
Relative absolute error             78.6253 %
Root relative squared error        83.8135 %
Total Number of Instances         684

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,933	0,489	0,782	0,933	0,851	0,509	0,834	0,877	0
	0,511	0,067	0,801	0,511	0,624	0,509	0,834	0,764	1
Weighted Avg.	0,787	0,343	0,789	0,787	0,772	0,509	0,834	0,838	

```

=== Confusion Matrix ===
      a    b  <-- classified as
417  30 |   a = 0
116 121 |   b = 1

```

Ek 4: Rastgele Orman Çıktısı

```

Correctly Classified Instances      532          77.7778 %
Kappa statistic                    0.4468
Mean absolute error                0.3735
Root mean squared error            0.4077
Relative absolute error             82.4355 %
Root relative squared error        85.6772 %
Total Number of Instances         684

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,957	0,561	0,763	0,957	0,849	0,491	0,842	0,883	0
	0,439	0,043	0,846	0,439	0,578	0,491	0,842	0,772	1
Weighted Avg.	0,778	0,381	0,792	0,778	0,755	0,491	0,842	0,844	

```

=== Confusion Matrix ===
      a    b  <-- classified as
428  19 |   a = 0
133 104 |   b = 1

```

Ek 5: Destek Vektör Makineleri Çıktısı

```
Correctly Classified Instances      479          70.0292 %
Kappa statistic                    0.1695
Mean absolute error                 0.2997
Root mean squared error             0.5475
Relative absolute error             66.1563 %
Root relative squared error        115.0454 %
Total Number of Instances          684

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          1,000    0,865    0,686     1,000    0,813    0,304    0,568    0,686    0
          0,135    0,000    1,000     0,135    0,238    0,304    0,568    0,435    1
Weighted Avg.    0,700    0,565    0,795     0,700    0,614    0,304    0,568    0,599

=== Confusion Matrix ===

  a  b  <-- classified as
447  0 |  a = 0
205 32 |  b = 1
```

Ek 6: Naive Bayes Çıktısı

```
Correctly Classified Instances      467          68.2749 %
Kappa statistic                    0.2575
Mean absolute error                 0.4004
Root mean squared error             0.4637
Relative absolute error             88.3865 %
Root relative squared error        97.4436 %
Total Number of Instances          684

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          0,821    0,578    0,728     0,821    0,772    0,263    0,652    0,736    0
          0,422    0,179    0,556     0,422    0,480    0,263    0,652    0,559    1
Weighted Avg.    0,683    0,440    0,668     0,683    0,671    0,263    0,652    0,675

=== Confusion Matrix ===

  a  b  <-- classified as
367  80 |  a = 0
137 100 |  b = 1
```

Ek 7: K-En Yakın Komşuluk Algoritması Çıktısı

```
Correctly Classified Instances      312          45.614 %
Kappa statistic                    0.036
Mean absolute error                 0.5437
Root mean squared error             0.7363
Relative absolute error            120.0181 %
Root relative squared error        154.725 %
Total Number of Instances          684

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
          0,306    0,262    0,688     0,306    0,424    0,047    0,534    0,677    0
          0,738    0,694    0,361     0,738    0,485    0,047    0,534    0,364    1
Weighted Avg.    0,456    0,411    0,575     0,456    0,445    0,047    0,534    0,568

=== Confusion Matrix ===

  a  b  <-- classified as
137 310 |  a = 0
 62 175 |  b = 1
```

Ek 8: Nitelik Seçimi Sonuçları

Nitelik Sırası	Bilgi Kazanımı	Ki-Kare	Kazanım Oranı	Relief-f
1	T4	T4	T4	Diyabet
2	Yaş	Yaş	TSH	MO2
3	Diyabet	Diyabet	Diyabet	Magnezyum
4	Glukoz	Glukoz	Yaş	Kreatinin
5	Cinsiyet	Cinsiyet	Glukoz	EO1
6	Trigliserid	Trigliserid	Cinsiyet	T4
7	TSH	ALT	BA2	Glukoz
8	ALT	TSH	Trigliserid	Total Kolesterol
9	BA2	BA2	ALT	EO2
10	Potasyum	Potasyum	Potasyum	BA1
11	HDL	HDL	LDL	BA2
12	LDL	LDL	HDL	AST
13	Total Kolesterol	Total Kolesterol	Total Kolesterol	Yaş
14	Kalsiyum	Kalsiyum	Kalsiyum	Ferritin
15	Demir	Demir	Demir	Sedimentasyon
16	Fosfor	Fosfor	Fosfor	PLT
17	Bilirubin	Bilirubin	Bilirubin	GGT
18	Demir Bağlama	Demir Bağlama	Demir Bağlama	HCT
19	Sodyum	Sodyum	Sodyum	HGB
20	ALP	ALP	ALP	MCV
21	Albümin	Albümin	Albümin	MPV
22	HbA1c(IFCC)	HbA1c(IFCC)	HbA1c(IFCC)	PDW
23	HbA1c	HbA1c	HbA1c	ALP
24	Üre	Üre	Üre	PCT

25	Kreatinin	Kreatinin	Kreatinin	HbA1c
26	Ürik Asit	Ürik Asit	Ürik Asit	LDH
27	GGT	GGT	GGT	ALT
28	LDH	LDH	LDH	LDL
29	CK	CK	CK	Demir
30	AST	AST	AST	Albümin
31	Magnezyum	Magnezyum	Magnezyum	NE2

Ek 8: Nitelik Seçimi Sonuçları (devamı)

32	VB12	VB12	VB12	HbA1c(IFCC)
33	ASO	ASO	ASO	RDW-CV
34	NE2	NE2	NE2	Demir Bağlama
35	BA1	BA1	BA1	Trigliserid
36	EO1	EO1	EO1	Kalsiyum
37	NE1	NE1	NE1	MO1
38	MO2	MO2	MO2	VB12
39	EO2	EO2	EO2	ASO
40	LY1	LY1	LY1	LY2
41	LY2	LY2	LY2	MCHC
42	T3	T3	T3	NE1
43	İnsülin	İnsülin	İnsülin	HDL
44	Sedimentasyon	Sedimentasyon	Sedimentasyon	CK
45	MO1	MO1	MO1	CRP
46	PDW	PDW	PDW	WBC
47	Ferritin	Ferritin	Ferritin	TSH
48	MCV	MCV	MCV	MCH
49	HGB	HGB	HGB	T3

50	RBC	RBC	RBC	LY1
51	WBC	WBC	WBC	RBC
52	HCT	HCT	HCT	Bilirubin
53	MCH	MCH	MCH	Fosfor
54	PCT	PCT	PCT	Sodyum
55	MCHC	MCHC	MCHC	Üre
56	MPV	MPV	MPV	İnsülin
57	PLT	PLT	PLT	Ürik Asit
58	RDW-CV	RDW-CV	RDW-CV	Cinsiyet
59	CRP	CRP	CRP	Potasyum

Ek 9: Karar Ağaçları Bilgi Kazanımı Nitelik Seçimi Çıktısı

```

Correctly Classified Instances      571          83.4795 %
Kappa statistic                    0.6117
Mean absolute error                 0.3212
Root mean squared error             0.3764
Relative absolute error             70.8993 %
Root relative squared error         79.1052 %
Total Number of Instances          684

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
      0.942    0.367    0.829    0.942    0.882    0.626    0.866    0.911    0
      0.633    0.058    0.852    0.633    0.726    0.626    0.866    0.790    1
Weighted Avg.    0.835    0.260    0.837    0.835    0.828    0.626    0.866    0.869

=== Confusion Matrix ===

  a  b  <-- classified as
421 26 | a = 0
 87 150 | b = 1

```

Ek 10: Rastgele Orman Bilgi Kazanımı Nitelik Seçimi Çıktısı

```
Correctly Classified Instances      556          81.2865 %
Kappa statistic                    0.5531
Mean absolute error                 0.3373
Root mean squared error             0.3847
Relative absolute error             74.4436 %
Root relative squared error         80.8425 %
Total Number of Instances          684

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,942    0,430    0,805    0,942    0,868    0,574    0,867    0,910    0
      0,570    0,058    0,839    0,570    0,678    0,574    0,867    0,772    1
Weighted Avg.    0,813    0,301    0,817    0,813    0,802    0,574    0,867    0,862

=== Confusion Matrix ===

  a  b  <-- classified as
421 26 |  a = 0
102 135 | b = 1
```

Ek 11: Destek Vektör Makineleri Bilgi Kazanımı Nitelik Seçimi Çıktısı

```
Correctly Classified Instances      549          80.2632 %
Kappa statistic                    0.4969
Mean absolute error                 0.1974
Root mean squared error             0.4443
Relative absolute error             43.5663 %
Root relative squared error         93.3597 %
Total Number of Instances          684

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      1,000    0,570    0,768    1,000    0,869    0,575    0,715    0,768    0
      0,430    0,000    1,000    0,430    0,602    0,575    0,715    0,628    1
Weighted Avg.    0,803    0,372    0,848    0,803    0,776    0,575    0,715    0,719

=== Confusion Matrix ===

  a  b  <-- classified as
447  0 |  a = 0
135 102 | b = 1
```

Ek 12: Yapay Sinir Ağları Bilgi Kazanımı Nitelik Seçimi Çıktısı

```
Correctly Classified Instances      539          78.8012 %
Kappa statistic                    0.5369
Mean absolute error                 0.2247
Root mean squared error             0.4365
Relative absolute error             49.5924 %
Root relative squared error         91.7277 %
Total Number of Instances          684

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,826    0,283    0,846    0,826    0,836    0,537    0,823    0,869    0
      0,717    0,174    0,685    0,717    0,701    0,537    0,823    0,739    1
Weighted Avg.    0,788    0,245    0,791    0,788    0,789    0,537    0,823    0,824

=== Confusion Matrix ===

  a  b  <-- classified as
369 78 |  a = 0
67 170 | b = 1
```

Ek 13: K-En Yakın Komşuluk Algoritması Relief-f Nitelik Seçimi Çıktısı

```
Correctly Classified Instances      494      72.2222 %
Kappa statistic                    0.2676
Mean absolute error                0.2785
Root mean squared error            0.5262
Relative absolute error            61.4744 %
Root relative squared error        110.5775 %
Total Number of Instances         684

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,971    0,747    0,710     0,971    0,820     0,345    0,618    0,713    0
      0,253    0,029    0,822     0,253    0,387     0,345    0,618    0,488    1
Weighted Avg.   0,722    0,498    0,749     0,722    0,670     0,345    0,618    0,635

=== Confusion Matrix ===

  a    b  <-- classified as
434  13 |   a = 0
177  60 |   b = 1
```

Ek 14: Naive Bayes Bilgi Kazanımı Nitelik Seçimi Çıktısı

```
Correctly Classified Instances      479      70.0292 %
Kappa statistic                    0.2597
Mean absolute error                0.4089
Root mean squared error            0.4587
Relative absolute error            90.2636 %
Root relative squared error        96.3885 %
Total Number of Instances         684

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,890    0,658    0,718     0,890    0,795     0,282    0,644    0,724    0
      0,342    0,110    0,623     0,342    0,441     0,282    0,644    0,565    1
Weighted Avg.   0,700    0,468    0,685     0,700    0,673     0,282    0,644    0,669

=== Confusion Matrix ===

  a    b  <-- classified as
398  49 |   a = 0
156  81 |   b = 1
```

Ek 15: Kümeleme Sonuçları

	Küme 1	Küme 2	Küme 3	Küme 4	Küme 5	Min	Max
Diyabet	0	1	1	1	0		
Vitamin D	1	0	0	0	0		
Cinsiyet	X	Y	X	X	Y		
Yaş	36,1905	58,1913	51,6418	58,8371	34,0581		
Glukoz	91,526	126,459	135,994	112,716	99,4406	70	110
HbA1c	6,1924	6,7917	6,9973	6,6753	6,0859	4	6,2
HbA1c(IFCC)	41,7493	48,878	52,0488	43,9711	41,2549	20	42
Kreatinin	0,7354	0,9149	0,7207	0,7291	0,8158	0,84	1,25
Üre	28,645	31,2362	28,8509	30,2564	29,1786	17	43
Ürik Asit	5,2959	5,8663	5,4864	5,2853	5,5617	3,5	7,2

ALT	19,0308	27,8926	21,9727	19,6703	22,7027	7	40
AST	20,4628	23,5723	19,6915	19,3958	22,6898	9	45
ALP	70,4811	79,1583	74,6017	69,8475	71,8276	30	120
LDH	161,4373	161,8053	163,63	160,6163	162,0161	120	250
GGT	24,6378	43,1972	24,716	19,7502	25,7027	0	55
Albümün	6,8661	6,709	6,6021	7,0961	7,4201	3,5	5,2
Bilirübin	0,2789	0,2776	0,2789	0,2777	0,2815	0	0,5
Sodyum	138,7133	138,5149	138,7981	138,889	138,8379	136	146
Potasyum	4,5037	4,5504	4,5969	4,5095	4,5164	3,5	5,1
Total Kolesterol	197,4767	200,475	208,014	203,625	196,3664	120	200
HDL	49,4699	41,7012	44,9505	52,005	44,7808	40	60
LDL	121,663	116,0969	126,6265	133,366	120,0714	0	130
Trigliserid	155,9246	185,427	233,077	160,964	191,543	40	160
Demir	69,6508	77,9783	55,3582	76,6493	85,4839	40	160
Demir Bağlama	271,4068	274,229	278,8242	259,5111	263,111	155	355
Kalsiyum	112,941	27,3924	34,6386	27,1793	40,8413	8,8	10,6
Fosfor	3,8497	3,8444	3,8221	3,8261	3,8323	2,3	4,7
Magnezyum	2,037	2,0482	2,0167	2,0476	2,0599	1,3	2,7
CK	100,5579	142,5082	100,4976	89,6791	102,8199	30	294
ASO	142,6161	137,7761	141,2067	140,9831	142,925	0	200
CRP	1,0748	0,8278	1,1728	0,796	1,1601	0	0,8
WBC	6,9956	7,4443	8,3526	6,8143	7,1808	4	11

Ek 15: Kümeleme Sonuçları (devamı)

RBC	4,6008	4,8635	4,639	4,5705	4,7527	3,8	5,8
HGB	12,5868	13,9094	12,0882	13,0503	13,4551	11,5	16,5
HCT	38,7154	42,1661	37,6882	39,9081	41,0625	37	47
MCV	84,611	86,7623	81,5131	87,574	86,6097	78	100
MCH	27,5293	28,7486	26,1672	29,0105	28,523	27	32
MCHC	32,6326	33,2447	32,0496	34,4219	33,3561	30	35

RDW-CV	14,2143	13,9697	13,9682	13,9955	13,9945	11,5	14,5
PLT	274,9659	244,587	334,0461	258,5551	260,5898	150	400
MPV	9,3858	9,3916	9,7599	9,573	9,5031	8	12
PCT	0,2539	0,2302	0,3226	0,2452	0,2454	0,12	0,36
PDW	49,2616	50,0373	48,1233	48,8151	49,8001	25	65
LY1	2,1779	2,6404	2,3841	2,2645	2,3287	1,5	4
MO1	0,3828	0,4497	0,4285	0,4002	0,4196	0	0,8
NE1	4,1086	4,3882	5,115	3,8561	4,1527	2	7,5
EO1	0,1908	0,2118	0,2439	0,1913	0,1999	0	0,4
BA1	0,0317	0,0407	0,0485	0,037	0,0366	0	0,1
MO2	5,5645	6,0992	5,2352	5,7686	5,8656	0	10
NE2	57,9991	58,1721	60,2973	55,8124	56,9952	40	75
EO2	2,7786	2,9005	2,9783	2,8375	2,8125	0	6
BA2	0,4584	0,5614	0,6215	0,5659	0,5312	0	2
LY2	31,5553	30,7304	29,286	35,0643	32,5938	20	50
Sedimentasyon	12,1954	9,2263	14,8349	12,9685	11,1929	0	20
İnsülin	11,5992	11,5176	11,5862	11,3715	11,7807	2,6	37,6
T3	3,1227	3,0859	2,8677	2,9658	3,398	1,71	3,71
T4	1,1409	1,1739	1,1199	1,139	1,1799	0,7	1,48
TSH	2,1358	2,0654	4,6848	2,6263	2,0224	0,35	4,94
Ferritin	37,3912	60,7007	37,6058	44,4474	62,494	10	291
VB12	438,3301	367,186	409,1955	408,6867	302,24	187	883

Ek 16: Destek Vektör Makineleri K-Ortalama Algoritması Kümeleme Sonrası Çıktısı

```
Correctly Classified Instances      540          78.9474 %
Kappa statistic                    0.4828
Mean absolute error                0.2105
Root mean squared error            0.4588
Relative absolute error            46.4707 %
Root relative squared error        96.4215 %
Total Number of Instances         684

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,953    0,519    0,776    0,953    0,855    0,519    0,717    0,770    0
      0,481    0,047    0,844    0,481    0,613    0,519    0,717    0,586    1
Weighted Avg.    0,789    0,355    0,800    0,789    0,771    0,519    0,717    0,706

=== Confusion Matrix ===

  a  b  <-- classified as
426 21 | a = 0
123 114 | b = 1
```

Ek 17: Karar Ağaçları K-Ortalama Algoritması Kümeleme Sonrası Çıktısı

```
Correctly Classified Instances      486          71.0526 %
Kappa statistic                    0.3058
Mean absolute error                0.3577
Root mean squared error            0.4177
Relative absolute error            78.9588 %
Root relative squared error        87.7684 %
Total Number of Instances         684

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,868    0,586    0,736    0,868    0,797    0,319    0,796    0,873    0
      0,414    0,132    0,624    0,414    0,497    0,319    0,796    0,620    1
Weighted Avg.    0,711    0,429    0,697    0,711    0,693    0,319    0,796    0,786

=== Confusion Matrix ===

  a  b  <-- classified as
388 59 | a = 0
139 98 | b = 1
```

Ek 18: Rastgele Orman K-Ortalama Algoritması Kümeleme Sonrası Çıktısı

```
Correctly Classified Instances      481          70.3216 %
Kappa statistic                    0.2734
Mean absolute error                0.3729
Root mean squared error            0.4242
Relative absolute error            82.3168 %
Root relative squared error        89.1456 %
Total Number of Instances         684

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,884    0,637    0,723    0,884    0,796    0,292    0,786    0,867    0
      0,363    0,116    0,623    0,363    0,459    0,292    0,786    0,590    1
Weighted Avg.    0,703    0,457    0,689    0,703    0,679    0,292    0,786    0,771

=== Confusion Matrix ===

  a  b  <-- classified as
395 52 | a = 0
151 86 | b = 1
```

Ek 19: Naive Bayes K-Ortalama Algoritması Kümeleme Sonrası Çıktısı

```
Correctly Classified Instances      455          66.5205 %
Kappa statistic                    0.2148
Mean absolute error                0.4207
Root mean squared error            0.471
Relative absolute error            92.8645 %
Root relative squared error        98.9719 %
Total Number of Instances         684

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,810    0,608    0,715    0,810    0,760    0,219    0,620    0,735    0
      0,392    0,190    0,522    0,392    0,448    0,219    0,620    0,443    1
Weighted Avg.    0,665    0,463    0,649    0,665    0,652    0,219    0,620    0,634

=== Confusion Matrix ===

  a  b  <-- classified as
362 85 | a = 0
144 93 | b = 1
```

Ek 20: Yapay Sinir Ağları K-Ortalama Algoritması Kümeleme Sonrası Çıktısı

```
Correctly Classified Instances      439          64.1813 %
Kappa statistic                    0.179
Mean absolute error                0.4164
Root mean squared error            0.484
Relative absolute error            91.918 %
Root relative squared error        101.7205 %
Total Number of Instances         684

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,767    0,595    0,709    0,767    0,737    0,180    0,620    0,731    0
      0,405    0,233    0,480    0,405    0,439    0,180    0,620    0,446    1
Weighted Avg.    0,642    0,469    0,629    0,642    0,634    0,180    0,620    0,632

=== Confusion Matrix ===

  a  b  <-- classified as
343 104 | a = 0
141 96 | b = 1
```

Ek 21: K-En Yakın Komşuluk Algoritması K-Ortalama Algoritması Kümeleme Sonrası Çıktısı

```
Correctly Classified Instances      325          47.5146 %
Kappa statistic                    -0.0788
Mean absolute error                0.5248
Root mean squared error            0.7233
Relative absolute error            115.8364 %
Root relative squared error        151.9975 %
Total Number of Instances         684

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,515    0,599    0,618    0,515    0,562    -0,081    0,461    0,635    0
      0,401    0,485    0,304    0,401    0,346    -0,081    0,461    0,327    1
Weighted Avg.    0,475    0,560    0,510    0,475    0,487    -0,081    0,461    0,528

=== Confusion Matrix ===

  a  b  <-- classified as
230 217 | a = 0
142 95 | b = 1
```

Ek 22: Ortak yükler

Communalities

	Initial	Extraction
DIA	1,000	,699
CİNSİET	1,000	,730
YAŞ	1,000	,750
SMEAN(GLUKOZ)	1,000	,630
SMEAN(HbA1c)	1,000	,385
SMEAN(HbA1cIFCC)	1,000	,593
SMEAN(KREATİNİN)	1,000	,672
SMEAN(ÜRE)	1,000	,570
SMEAN(ÜRİKASİT)	1,000	,544
SMEAN(ALT)	1,000	,782
SMEAN(AST)	1,000	,771
SMEAN(ALP)	1,000	,689
SMEAN(LDH)	1,000	,571
SMEAN(GGT)	1,000	,805
SMEAN(ALBÜMİN)	1,000	,592
SMEAN(BİLİRUBİN)	1,000	,426
SMEAN(SODYUM)	1,000	,607
SMEAN(POTASYUM)	1,000	,572
SMEAN(TOTALKOLESTEROL)	1,000	,770
SMEAN(HDL)	1,000	,681
SMEAN(LDL)	1,000	,723
SMEAN(TRİGLİSERİD)	1,000	,640
SMEAN(DEMİR)	1,000	,526
SMEAN(DEMİRBAĞLAMA)	1,000	,698
SMEAN(KALSİYUM)	1,000	,587
SMEAN(FOSFOR)	1,000	,511
SMEAN(MAGNEZYUM)	1,000	,480

SMEAN(CK)	1,000	,573
SMEAN(ASO)	1,000	,699
Ek 22: Ortak yükler (devamı)		
SMEAN(CRP)	1,000	,474
SMEAN(WBC)	1,000	,894
SMEAN(RBC)	1,000	,911
SMEAN(HGB)	1,000	,933
SMEAN(HCT)	1,000	,937
SMEAN(MCV)	1,000	,802
SMEAN(MCH)	1,000	,617
SMEAN(MCHC)	1,000	,600
SMEAN(RDWCV)	1,000	,484
SMEAN(PLT)	1,000	,888
SMEAN(MPV)	1,000	,686
SMEAN(PCT)	1,000	,849
SMEAN(PDW)	1,000	,564
SMEAN(LY#)	1,000	,652
SMEAN(MO#)	1,000	,772
SMEAN(NE#)	1,000	,948
SMEAN(EO#)	1,000	,964
SMEAN(BA#)	1,000	,925
SMEAN(MO)	1,000	,777
SMEAN(@NE)	1,000	,766
SMEAN(EO)	1,000	,936
SMEAN(BA)	1,000	,935
SMEAN(LY)	1,000	,593
SMEAN(SEDİMENTASYON)	1,000	,583
SMEAN(İNSULİN)	1,000	,566
SMEAN(T3)	1,000	,634

SMEAN(T4)	1,000	,532
SMEAN(TSH)	1,000	,461
SMEAN(FERRITIN)	1,000	,600
SMEAN(VB12)	1,000	,528
25-Hidroksi Vitamin D	1,000	,647



Ek 23: Açıklanan Toplam Varyans

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared			Rotation Sums of Squared		
	Total	Loadings		Total	Loadings		Total	Loadings	
		% of Variance	Cumulative %		% of Variance	Cumulative %		% of Variance	Cumulative %
1	4,258	7,097	7,097	4,258	7,097	7,097	3,070	5,116	5,116
2	3,750	6,250	13,347	3,750	6,250	13,347	2,764	4,606	9,723
3	3,140	5,233	18,581	3,140	5,233	18,581	2,517	4,195	13,918
4	2,795	4,658	23,239	2,795	4,658	23,239	2,351	3,918	17,835
5	2,523	4,205	27,444	2,523	4,205	27,444	2,322	3,870	21,706
6	2,267	3,778	31,222	2,267	3,778	31,222	2,086	3,477	25,183
7	1,953	3,255	34,477	1,953	3,255	34,477	2,072	3,453	28,636
8	1,840	3,066	37,544	1,840	3,066	37,544	2,029	3,382	32,018
9	1,704	2,839	40,383	1,704	2,839	40,383	2,007	3,345	35,362
10	1,681	2,801	43,184	1,681	2,801	43,184	1,839	3,066	38,428
11	1,499	2,498	45,682	1,499	2,498	45,682	1,831	3,051	41,479
12	1,472	2,454	48,136	1,472	2,454	48,136	1,827	3,045	44,525
13	1,408	2,346	50,482	1,408	2,346	50,482	1,793	2,988	47,512
14	1,319	2,198	52,680	1,319	2,198	52,680	1,708	2,847	50,359
15	1,278	2,131	54,811	1,278	2,131	54,811	1,534	2,557	52,916
16	1,232	2,053	56,864	1,232	2,053	56,864	1,439	2,399	55,315
17	1,185	1,975	58,838	1,185	1,975	58,838	1,366	2,276	57,591
18	1,169	1,948	60,786	1,169	1,948	60,786	1,359	2,265	59,856
19	1,113	1,856	62,642	1,113	1,856	62,642	1,294	2,157	62,013
20	1,100	1,833	64,474	1,100	1,833	64,474	1,197	1,995	64,008
21	1,033	1,721	66,195	1,033	1,721	66,195	1,181	1,968	65,976
22	1,020	1,699	67,894	1,020	1,699	67,894	1,151	1,918	67,894
23	,991	1,651	69,546						
24	,945	1,574	71,120						
25	,935	1,559	72,679						
26	,925	1,542	74,221						
27	,886	1,476	75,697						
28	,878	1,464	77,162						
29	,870	1,450	78,612						
30	,835	1,391	80,003						
31	,784	1,306	81,309						
32	,778	1,296	82,605						
33	,757	1,262	83,868						
34	,733	1,222	85,090						
35	,718	1,197	86,286						

Ek 23: Açıklanan Toplam Varyans (devamı)

36	,671	1,118	87,404						
----	------	-------	--------	--	--	--	--	--	--

37	,651	1,086	88,489				
38	,624	1,040	89,529				
39	,585	,975	90,504				
40	,554	,923	91,427				
41	,539	,899	92,326				
42	,532	,887	93,213				
43	,482	,804	94,017				
44	,445	,742	94,758				
45	,415	,691	95,450				
46	,390	,650	96,100				
47	,360	,600	96,700				
48	,347	,578	97,278				
49	,320	,533	97,811				
50	,315	,526	98,337				
51	,277	,462	98,798				
52	,214	,356	99,155				
53	,197	,328	99,483				
54	,165	,275	99,757				
55	,057	,095	99,852				
56	,037	,061	99,913				
57	,021	,034	99,947				
58	,019	,032	99,979				
59	,007	,012	99,991				
60	,006	,009	100,000				

Extraction Method: Principal Component Analysis.

Ek 24: Döndürülmüş Bileşen Matrisi

Rotated Component Matrix ^a																						
	Component																					
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
SMEAN(NE#)	,950	,071	-,010	,114	,033	,031	-,016	-,055	,003	-,061	-,002	,086	,007	,051	-,033	,018	-,018	-,030	-,014	,001	-,020	-,069
SMEAN(WBC)	,880	,120	-,086	,105	-,001	,066	,124	,018	,030	-,002	-,019	,048	-,028	,032	,008	,012	,216	-,020	,007	-,042	,045	-,101
SMEAN(@NE)	,679	-,015	,108	,062	,075	-,013	-,209	-,160	-,044	-,117	,058	,100	,067	,073	-,087	-,009	-,377	-,032	-,015	,075	-,129	,041
SMEAN(HCT)	,089	,890	,290	-,028	-,014	-,114	-,054	,092	,045	,044	,085	,000	-,084	,059	,052	,023	,017	,032	,010	,003	,027	,017
SMEAN(HGB)	,066	,850	,376	-,006	-,031	-,158	-,021	,057	,037	,046	,095	-,004	-,119	,052	,038	,000	,006	,035	-,006	,013	,029	-,042
SMEAN(RBC)	,077	,844	-,373	-,032	-,093	-,030	-,038	,131	,029	,041	,037	,078	-,002	,067	,051	-,037	,036	-,028	,020	,004	,046	,033
SMEAN(MCV)	,009	,045	,857	,008	,102	-,083	-,019	-,054	,026	,000	,062	-,118	-,108	-,018	-,002	,086	-,036	,069	-,020	-,001	-,012	-,028
SMEAN(MCH)	-,005	,012	,771	,022	,048	-,028	-,041	-,076	,018	,008	,048	-,033	-,039	-,001	-,037	,043	-,013	,034	-,028	-,026	-,012	-,016
SMEAN(RDWCV)	,036	-,215	-,574	,014	,060	,043	-,117	-,006	,005	-,042	,044	,068	,198	-,038	,061	,109	,018	-,050	-,088	-,111	-,007	,029
SMEAN(GGT)	,128	,013	-,010	,850	,062	-,063	-,030	,017	,125	-,080	,156	-,037	-,019	,018	,006	-,032	,028	,005	,051	-,037	-,042	-,033
SMEAN(ALP)	,101	,049	-,048	,757	,081	,038	,044	-,007	,030	-,014	,062	-,048	,126	,123	,033	-,028	-,071	,181	,111	,013	-,017	,014
SMEAN(SODYUM)	-,044	,100	-,078	-,700	,075	-,031	-,011	-,079	-,029	-,020	,035	-,035	,056	,136	-,065	-,049	,098	,175	,038	,099	,000	,046
SMEAN(T4)	-,025	-,002	,014	,368	-,275	-,004	-,094	,018	,039	-,037	,018	,197	-,017	-,121	-,015	,120	,257	-,051	-,074	,223	,254	,225
YAŞ	-,030	,034	,129	,037	,767	,012	-,016	,064	-,103	,091	,179	,155	-,063	-,132	,068	-,038	,075	,006	-,144	-,042	,055	,062
SMEAN(T3)	-,068	,152	,056	-,016	-,737	-,015	,124	,007	,043	-,032	-,058	-,037	,032	-,062	,039	-,068	,099	,066	-,075	,039	-,060	-,019
DIA	,049	,051	,100	,003	,647	,012	,062	,141	,053	,019	-,059	,286	,181	-,187	,047	-,214	,103	,114	-,040	-,057	,061	-,066
SMEAN(PLT)	,087	-,157	-,199	,010	,017	,872	,067	,028	-,017	,047	,004	-,021	,084	,041	-,118	,013	-,022	,062	,018	,115	,029	-,072
SMEAN(PCT)	,218	-,239	-,135	-,029	,016	,789	,001	,185	-,049	,086	-,073	,008	,036	,028	,034	,045	,053	,058	,067	-,003	,043	,186
SMEAN(PDW)	,177	-,063	-,133	-,019	-,008	-,633	-,042	,124	-,062	,073	,020	-,002	,047	,101	,020	,078	-,048	,131	,036	,188	,087	,000
SMEAN(EO#)	,153	-,017	-,001	,022	-,039	,066	,954	,105	,015	,047	,033	,015	-,018	,030	,047	-,023	,060	,018	-,013	-,008	,019	,013
SMEAN(EO)	-,138	-,072	,025	-,009	-,042	,029	,940	,111	,017	,058	,031	,002	-,021	,024	,047	-,023	-,032	,034	-,026	,008	-,002	,029
SMEAN(BA)	-,244	,094	-,065	,024	,069	,017	,085	,914	-,016	,080	-,011	,042	-,014	,038	,018	-,015	-,017	-,001	-,012	,048	,010	,012
SMEAN(BA#)	,090	,141	-,089	,068	,059	,042	,154	,913	-,015	,063	-,008	,076	-,018	,053	,011	-,009	,075	,011	,010	,028	,028	,000
SMEAN(AST)	,005	,000	-,003	,040	-,159	-,010	,030	-,031	,852	,012	-,028	-,024	,028	-,065	,078	-,005	,003	-,028	,036	,028	,005	,029
SMEAN(ALT)	,001	,089	,044	,140	,044	,002	-,003	-,015	,849	,016	,064	,028	-,060	,085	-,041	-,061	,010	,035	-,074	,032	,009	,032
SMEAN(CK)	,029	-,067	,046	-,179	-,096	,076	-,004	,127	,363	-,039	,325	-,082	-,014	-,124	,027	,033	-,164	,207	,331	-,069	,099	-,213
SMEAN(TOTALKOLESTEROL)	-,056	,045	-,025	,021	,074	-,001	,006	,087	-,001	,860	,037	,016	-,007	-,015	-,059	-,021	,011	,014	,019	-,040	-,018	-,060
SMEAN(LDL)	-,079	,036	,099	-,075	,062	,044	,089	,024	-,008	,803	-,134	-,056	,046	-,022	,048	,012	,027	-,056	,055	,084	,018	-,007
SMEAN(TRİGLİSERİD)	,060	,017	-,105	-,045	-,034	-,008	,028	,052	,094	,499	,032	,250	-,015	,462	-,085	-,049	,020	-,039	-,084	-,235	-,080	,050
SMEAN(KREATİNİN)	,036	,130	,072	,167	,187	-,066	,044	-,018	,035	-,054	,739	,038	-,074	,103	,086	,029	,035	-,016	,001	,024	-,024	-,018

SMEAN(ÜRE)	-,096	-,034	,060	,163	,342	,007	,119	-,036	-,112	,020	,480	-,043	,089	-,001	,059	,248	,083	-,092	,087	,010	-,010	,239
CİNSİİYET	-,013	,291	,013	,116	-,386	-,124	-,019	-,014	,212	-,149	,437	,026	-,125	,379	,183	-,042	,060	-,080	-,019	,055	-,021	-,022
SMEAN(SEDİMENTASYON)	,002	-,213	-,016	,273	,191	,006	,108	-,011	-,016	,080	-,424	-,032	,266	,237	,023	,239	,154	,068	-,068	,033	,019	,093
SMEAN(ÜRİKASİT)	,063	,082	-,112	,095	,073	,003	,052	-,072	,194	-,008	,384	-,092	-,122	,163	-,097	,141	,120	,314	-,031	-,347	,011	-,011
SMEAN(GLUKOZ)	,082	,063	-,060	-,001	,190	-,077	-,040	,117	-,011	,082	,071	,715	,061	,083	-,037	-,081	,004	,035	-,082	-,017	,070	,059
SMEAN(HbA1cIFCC)	,131	-,031	-,141	-,036	,142	,003	,045	,031	,051	-,096	,022	,699	,068	,094	-,013	,104	,015	,023	-,003	,037	,024	,041
SMEAN(HbA1c)	-,014	,023	-,016	,015	-,024	,038	,005	-,027	-,050	,031	-,057	,564	-,105	-,106	,016	,042	-,010	,016	,143	-,025	-,063	-,078
SMEAN(DEMİRBAĞLAMA)	-,006	,029	-,165	,111	,015	,055	-,001	-,018	,221	,007	,015	-,052	,674	,167	-,038	,033	-,080	-,323	,046	,016	,059	-,013
SMEAN(FERRİTİN)	,036	,136	,053	,124	,103	-,029	,051	,015	,248	,032	,092	-,048	-,654	,077	-,031	-,022	,012	,042	-,133	,141	,010	,012
SMEAN(DEMİR)	-,067	,086	,275	-,088	-,161	-,007	,000	,017	-,011	-,059	,068	-,005	-,589	,045	-,001	-,119	,024	-,072	-,015	-,036	,152	,053
SMEAN(VB12)	-,019	-,097	-,074	,057	,074	-,018	-,029	-,024	,157	-,041	,049	,093	-,010	-,644	-,015	-,035	,035	,019	-,116	-,137	-,087	,091

SMEAN(HDL)	-,161	-,061	,085	-,077	,111	,066	-,038	-,067	-,202	,083	-,249	-,113	,002	-,623	,081	-,048	-,033	-,242	-,069	,167	,057	-,032
SMEAN(MO)	-,259	,066	-,062	,050	,029	-,102	,032	,025	-,008	-,059	,116	-,033	,022	-,007	,793	,094	-,102	,124	,037	,019	-,008	,046
SMEAN(MO#)	,461	,130	-,116	,119	,025	-,004	,145	-,003	,072	,005	,053	-,030	-,027	-,033	,671	,057	,151	,028	,017	-,062	,038	-,011
SMEAN(ALBÜMİN)	-,098	-,027	,053	-,004	-,047	-,047	,032	,003	-,069	,022	,127	,056	,130	-,007	,056	,713	,129	-,014	-,039	,004	,045	,007
SMEAN(CRP)	,250	,006	,025	,014	-,034	,056	-,137	,000	,040	-,062	-,086	,052	,008	,085	,091	,546	-,170	,055	-,153	-,002	,002	,055
25-Hidroksi Vitamin D	,005	,074	-,237	,041	-,048	-,127	,036	-,132	-,128	,013	,056	-,245	-,069	-,373	-,311	,410	-,141	,087	,173	,019	-,017	,021
SMEAN(İNSULİN)	,089	-,014	,007	,065	-,200	-,005	-,017	-,002	-,051	,071	,136	-,125	,295	-,033	-,123	-,321	,064	,320	-,178	-,001	,185	,304
SMEAN(LY#)	,037	,031	-,040	-,088	,014	,040	,021	,036	-,008	,026	,052	,021	-,039	,023	,000	-,003	,785	,014	,071	,082	-,042	-,039
SMEAN(KALSİYUM)	,073	-,046	-,155	-,008	-,017	,054	-,035	-,056	,007	,115	,053	-,056	,120	-,031	-,051	-,015	-,048	-,697	-,055	,111	,080	,018
SMEAN(LDH)	-,011	-,064	-,005	-,028	-,008	,117	,026	-,106	,079	,146	-,007	,087	-,028	,076	,223	,027	-,136	,516	-,110	,384	,035	-,051
SMEAN(POTASYUM)	-,031	,019	,086	,051	,012	,044	-,004	,028	,072	,033	,066	,054	,230	-,002	,039	-,018	,194	,024	,665	-,083	-,032	-,031
SMEAN(BİLİRUBİN)	,034	-,057	-,038	,034	-,116	,006	-,054	-,036	-,190	,053	,050	,077	,009	,095	,067	-,201	-,106	-,075	,499	,125	,085	,107
SMEAN(FOSFOR)	-,037	,141	-,121	,112	,098	-,048	,007	-,023	,048	-,035	-,327	-,051	-,239	,181	-,164	,062	,009	,021	,454	,006	-,019	,097
SMEAN(MAGNEZYUM)	,018	,097	,051	-,256	,031	,097	,005	,109	,055	-,031	-,016	-,224	-,029	,102	,005	-,043	,103	-,068	,057	,533	-,097	-,027
SMEAN(TSH)	-,005	,033	,001	-,103	,215	,268	-,004	,001	-,038	,030	-,022	-,140	,105	,082	,110	-,065	-,089	-,019	,033	-,512	-,034	-,033
SMEAN(ASO)	-,049	-,062	-,007	,020	-,112	,019	-,054	-,028	-,017	,034	-,018	-,049	,016	-,041	-,035	-,027	,119	,042	-,031	,005	-,807	,009
SMEAN(LY)	-,367	,010	-,054	-,016	,035	,007	-,083	,009	,013	,032	-,104	-,098	-,109	-,009	-,047	-,012	,342	-,010	,002	-,066	,531	-,089
SMEAN(MCHC)	,172	-,039	,101	,029	-,054	-,105	-,071	,035	-,064	,080	-,020	,000	,037	,038	,014	-,025	,098	,025	-,048	-,036	,059	-,711
SMEAN(MPV)	,239	-,137	,133	-,069	-,016	-,204	-,129	,284	-,061	,066	-,141	,024	-,104	-,055	,299	,079	,118	-,017	,082	-,213	,027	,493



Ek 25: Denklemdaki Değişkenler

		Variables in the Equation					
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	DİA	2,540	,334	57,726	1	,000	12,680
	CİNSİYET	2,752	,355	60,119	1	,000	15,668
	YAŞ	,023	,008	7,682	1	,006	1,024
	SMEAN(GLUKOZ)	-,005	,004	1,712	1	,191	,995
	SMEAN(HbA1c)	-,021	,045	,222	1	,637	,979
	SMEAN(HbA1cIFCC)	,015	,013	1,309	1	,253	1,015
	SMEAN(KREATİNİN)	,151	,843	,032	1	,858	1,163
	SMEAN(ÜRE)	,044	,025	2,995	1	,084	1,045
	SMEAN(ÜRİKASİT)	,295	,149	3,900	1	,048	1,342
	SMEAN(ALT)	-,051	,019	7,649	1	,006	,950
	SMEAN(AST)	,024	,026	,829	1	,363	1,024
	SMEAN(ALP)	-,001	,011	,009	1	,925	,999
	SMEAN(LDH)	-,011	,009	1,468	1	,226	,989
	SMEAN(GGT)	,021	,007	9,077	1	,003	1,021
	SMEAN(ALBÜMİN)	,094	,039	5,699	1	,017	1,099
	SMEAN(BİLİRUBİN)	-6,822	12,010	,323	1	,570	,001
	SMEAN(SODYUM)	,106	,071	2,189	1	,139	1,111
	SMEAN(POTASYUM)	-,423	,416	1,033	1	,309	,655
	SMEAN(TOTALKOLESTEROL)	,009	,005	2,908	1	,088	1,009
	SMEAN(HDL)	-,002	,015	,014	1	,906	,998
	SMEAN(LDL)	-,009	,005	2,998	1	,083	,991
	SMEAN(TRİGLİSERİD)	-,004	,002	4,378	1	,036	,996
	SMEAN(DEMİR)	-,005	,004	1,467	1	,226	,995
	SMEAN(DEMİRBAĞLAMA)	-,003	,003	,824	1	,364	,997
	SMEAN(KALSİYUM)	,000	,001	,379	1	,538	1,000
	SMEAN(FOSFOR)	4,121	1,866	4,878	1	,027	61,607
	SMEAN(MAGNEZYUM)	-,590	,793	,553	1	,457	,554
	SMEAN(CK)	,004	,002	6,477	1	,011	1,004
	SMEAN(ASO)	-,001	,005	,018	1	,893	,999

SMEAN(CRP)	,054	,071	,582	1	,446	1,056
SMEAN(WBC)	,755	,692	1,191	1	,275	2,129
SMEAN(RBC)	2,908	2,355	1,524	1	,217	18,313
SMEAN(HGB)	,723	,394	3,379	1	,066	2,061

Ek 25: Denklemdaki Değişkenler(devamı)

SMEAN(HCT)	-,403	,301	1,800	1	,180	,668
SMEAN(MCV)	,085	,135	,393	1	,531	1,088
SMEAN(MCH)	-,043	,121	,127	1	,721	,958
SMEAN(MCHC)	-,014	,026	,292	1	,589	,986
SMEAN(RDWCV)	,004	,108	,002	1	,969	1,004
SMEAN(PLT)	,013	,015	,755	1	,385	1,013
SMEAN(MPV)	,152	,415	,133	1	,715	1,164
SMEAN(PCT)	-16,048	16,024	1,003	1	,317	,000
SMEAN(PDW)	,043	,019	4,928	1	,026	1,044
SMEAN(LY#)	-,037	,240	,023	1	,878	,964
SMEAN(MO#)	-2,239	4,343	,266	1	,606	,107
SMEAN(NE#)	-,872	,843	1,070	1	,301	,418
SMEAN(EO#)	-5,287	5,661	,873	1	,350	,005
SMEAN(BA#)	4,368	28,180	,024	1	,877	78,886
SMEAN(MO)	-,012	,370	,001	1	,974	,988
SMEAN(@NE)	-,023	,178	,017	1	,896	,977
SMEAN(EO)	,315	,454	,481	1	,488	1,370
SMEAN(BA)	-1,567	1,958	,640	1	,424	,209
SMEAN(LY)	-,070	,176	,157	1	,692	,933
SMEAN(SEDİMENTASYON)	-,057	,026	4,801	1	,028	,944
SMEAN(İNSULİN)	-,028	,054	,265	1	,607	,973
SMEAN(T3)	,478	,303	2,488	1	,115	1,613
SMEAN(T4)	-,751	,654	1,319	1	,251	,472
SMEAN(TSH)	-,163	,072	5,146	1	,023	,849
SMEAN(FERRİTİN)	-,003	,003	,981	1	,322	,997
SMEAN(VB12)	,000	,000	,883	1	,348	1,000
Constant	-42,112	27,583	2,331	1	,127	,000

Ek 26: Destek Vektör Makineleri Çıktısı

```
Correctly Classified Instances      398          98.0296 %
Kappa statistic                    0.9483
Mean absolute error                0.0197
Root mean squared error            0.1404
Relative absolute error             5.0381 %
Root relative squared error        31.7664 %
Total Number of Instances         406

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      1,000    0,074    0,974    1,000    0,987    0,950    0,963    0,974    0
      0,926    0,000    1,000    0,926    0,962    0,950    0,963    0,946    1
Weighted Avg.    0,980    0,054    0,981    0,980    0,980    0,950    0,963    0,966

=== Confusion Matrix ===

  a  b  <-- classified as
298  0  |  a = 0
 8 100 |  b = 1
```

Ek 27 Karar Ağaçları Çıktısı

```
Correctly Classified Instances      393          96.798 %
Kappa statistic                    0.9153
Mean absolute error                0.1969
Root mean squared error            0.2437
Relative absolute error             50.3331 %
Root relative squared error        55.1542 %
Total Number of Instances         406

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,997    0,111    0,961    0,997    0,979    0,918    0,967    0,981    0
      0,889    0,003    0,990    0,889    0,937    0,918    0,967    0,963    1
Weighted Avg.    0,968    0,082    0,969    0,968    0,967    0,918    0,967    0,976

=== Confusion Matrix ===

  a  b  <-- classified as
297  1  |  a = 0
12 96  |  b = 1
```

Ek 28: Rastgele Orman Çıktısı

```
Correctly Classified Instances      387          95.3202 %
Kappa statistic                    0.873
Mean absolute error                0.2625
Root mean squared error            0.2948
Relative absolute error             67.111 %
Root relative squared error        66.7211 %
Total Number of Instances         406

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      1,000    0,176    0,940    1,000    0,969    0,880    0,972    0,983    0
      0,824    0,000    1,000    0,824    0,904    0,880    0,972    0,969    1
Weighted Avg.    0,953    0,129    0,956    0,953    0,952    0,880    0,972    0,979

=== Confusion Matrix ===

  a  b  <-- classified as
298  0  |  a = 0
19 89  |  b = 1
```

Ek 29: Yapay Sinir Ağları Çıktısı

```
Correctly Classified Instances      332          81.7734 %
Kappa statistic                    0.5758
Mean absolute error                 0.1914
Root mean squared error            0.3928
Relative absolute error            48.9502 %
Root relative squared error       88.9018 %
Total Number of Instances         406

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,819   0,185   0,924     0,819   0,868     0,587   0,805     0,858     0
      0,815   0,181   0,620     0,815   0,704     0,587   0,805     0,591     1
Weighted Avg.   0,818   0,184   0,843     0,818   0,825     0,587   0,805     0,787

=== Confusion Matrix ===

  a  b  <-- classified as
244 54 | a = 0
 20 88 | b = 1
```

Ek 30: Naive Bayes Çıktısı

```
Correctly Classified Instances      297          73.1527 %
Kappa statistic                    -0.0049
Mean absolute error                 0.3919
Root mean squared error            0.448
Relative absolute error           100.2094 %
Root relative squared error       101.3769 %
Total Number of Instances         406

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,997   1,000   0,733     0,997   0,845    -0,030   0,476     0,721     0
      0,000   0,003   0,000     0,000   0,000    -0,030   0,476     0,253     1
Weighted Avg.   0,732   0,735   0,538     0,732   0,620    -0,030   0,476     0,597

=== Confusion Matrix ===

  a  b  <-- classified as
297  1 | a = 0
 108  0 | b = 1
```

Ek 31: K-En Yakın Komşuluk Algoritması Çıktısı

```
Correctly Classified Instances      170          41.8719 %
Kappa statistic                    -0.0403
Mean absolute error                 0.5812
Root mean squared error            0.761
Relative absolute error           148.6029 %
Root relative squared error       172.2135 %
Total Number of Instances         406

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,359   0,417   0,704     0,359   0,476    -0,053   0,458     0,721     0
      0,583   0,641   0,248     0,583   0,348    -0,053   0,458     0,244     1
Weighted Avg.   0,419   0,476   0,583     0,419   0,442    -0,053   0,458     0,594

=== Confusion Matrix ===

  a  b  <-- classified as
107 191 | a = 0
 45  63 | b = 1
```

Ek 32: Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Nitelik Seçimi Sonuçları

Nitelik Sırası	Bilgi Kazanımı	Ki-Kare	Kazanım Oranı	Relief-f
1	Yaş	Yaş	Yaş	Cinsiyet
2	MPV	MPV	EO2	Magnezyum
3	EO2	EO2	MPV	Glukoz
4	Cinsiyet	Cinsiyet	Cinsiyet	MO2
5	Potasyum	Potasyum	Potasyum	Fosfor
6	Total Kolesterol	Total Kolesterol	Total Kolesterol	Yaş
7	HDL	HDL	HDL	Sedimentasyon
8	LDL	LDL	LDL	PCT
9	Trigliserid	Trigliserid	Trigliserid	Demir
10	Bilirübin	Bilirübin	Bilirübin	T3
11	Fosfor	Fosfor	Fosfor	PLT
12	Magnezyum	Magnezyum	Magnezyum	LDL
13	Kalsiyum	Kalsiyum	Kalsiyum	RDW-CV
14	Demir	Demir	Demir	AST
15	Demir Bağlama	Demir Bağlama	Demir Bağlama	Albümin
16	Sodyum	Sodyum	Sodyum	HDL
17	Albümin	Albümin	Albümin	Trigliserid
18	ASO	ASO	ASO	MPV
19	Üre	Üre	Üre	VB12
20	HbA1c(IFCC)	HbA1c(IFCC)	HbA1c(IFCC)	ALT
21	HbA1c	HbA1c	HbA1c	T4
22	Glukoz	Glukoz	Glukoz	MO1
23	Kreatinin	Kreatinin	Kreatinin	NE1
24	Ürik Asit	Ürik Asit	Ürik Asit	BA1
25	GGT	GGT	GGT	BA2
26	ALT	ALT	ALT	WBC
27	LDH	LDH	LDH	GGT
28	ALP	ALP	ALP	HCT
29	AST	AST	AST	ALP
30	CK	CK	CK	HGB
31	VB12	VB12	VB12	EO1
32	Ferritin	Ferritin	Ferritin	ASO
33	LY2	LY2	LY2	LY2
34	NE2	NE2	NE2	MCHC
35	MO2	MO2	MO2	Total Kolesterol
36	BA1	BA1	BA1	MCH
37	BA2	BA2	BA2	HbA1c
38	Sedimentasyon	Sedimentasyon	Sedimentasyon	MCV
39	NE1	NE1	NE1	Ferritin
40	İnsülin	İnsülin	İnsülin	Kreatinin
41	TSH	TSH	TSH	TSH
42	T4	T4	T4	CK

43	T3	T3	T3	NE2
44	EO1	EO1	EO1	RBC
45	MO1	MO1	MO1	HbA1c(IFCC)
46	WBC	WBC	WBC	LY1
47	MCH	MCH	MCH	Kalsiyum
48	HCT	HCT	HCT	EO2
49	HGB	HGB	HGB	CRP
50	RBC	RBC	RBC	PDW
51	MCV	MCV	MCV	Ürik Asit
52	MCHC	MCHC	MCHC	Bilirubin
53	LY1	LY1	LY1	Üre
54	RDW-CV	RDW-CV	RDW-CV	LDH
55	PDW	PDW	PDW	Demir Bağlama
56	PCT	PCT	PCT	Sodyum
57	PLT	PLT	PLT	İnsülin
58	CRP	CRP	CRP	Potasyum

Ek 33: Relief-f Nitelik Seçimi Sonrası Destek Vektör Makineleri Çıktısı

```

Correctly Classified Instances      398          98.0296 %
Kappa statistic                    0.9483
Mean absolute error                 0.0197
Root mean squared error             0.1404
Relative absolute error             5.0381 %
Root relative squared error        31.7664 %
Total Number of Instances         406

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      1,000    0,074    0,974      1,000    0,987      0,950    0,963    0,974      0
      0,926    0,000    1,000      0,926    0,962      0,950    0,963    0,946      1
Weighted Avg.   0,980    0,054    0,981      0,980    0,980      0,950    0,963    0,966

=== Confusion Matrix ===

  a    b  <-- classified as
298  0 |  a = 0
 8 100 |  b = 1

```

Ek 34: Bilgi Kazanımı Nitelik Seçimi Sonrası Karar Ağaçları Çıktısı

```

Correctly Classified Instances      396          97.5369 %
Kappa statistic                    0.9358
Mean absolute error                 0.1904
Root mean squared error             0.239
Relative absolute error            48.6916 %
Root relative squared error        54.0942 %
Total Number of Instances         406

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,993    0,074    0,974      0,993    0,983      0,936    0,967    0,980      0
      0,926    0,007    0,980      0,926    0,952      0,936    0,967    0,965      1
Weighted Avg.   0,975    0,056    0,975      0,975    0,975      0,936    0,967    0,976

=== Confusion Matrix ===

  a    b  <-- classified as
296  2 |  a = 0
 8 100 |  b = 1

```

Ek 35: Bilgi Kazanımı Nitelik Seçimi Sonrası Rastgele Orman Çıktısı

```
Correctly Classified Instances      394          97.0443 %
Kappa statistic                    0.9215
Mean absolute error                0.2561
Root mean squared error            0.2898
Relative absolute error             65.4694 %
Root relative squared error        65.5931 %
Total Number of Instances         406

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      1,000    0,111    0,961    1,000    0,980    0,924    0,967    0,974    0
      0,889    0,000    1,000    0,889    0,941    0,924    0,967    0,969    1
Weighted Avg.    0,970    0,082    0,972    0,970    0,970    0,924    0,967    0,973

=== Confusion Matrix ===

  a  b  <-- classified as
298  0  |  a = 0
 12 96 |  b = 1
```

Ek 36: Bilgi Kazanımı Nitelik Seçimi Sonrası Yapay Sinir Ağları Çıktısı

```
Correctly Classified Instances      346          85.2217 %
Kappa statistic                    0.6485
Mean absolute error                0.1745
Root mean squared error            0.3583
Relative absolute error             44.6162 %
Root relative squared error        81.0899 %
Total Number of Instances         406

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,856    0,157    0,938    0,856    0,895    0,656    0,840    0,909    0
      0,843    0,144    0,679    0,843    0,752    0,656    0,840    0,632    1
Weighted Avg.    0,852    0,154    0,869    0,852    0,857    0,656    0,840    0,835

=== Confusion Matrix ===

  a  b  <-- classified as
255 43 |  a = 0
 17 91 |  b = 1
```

Ek 37: Bilgi Kazanımı Nitelik Seçimi Sonrası K-En Yakın Komşuluk Algoritması Çıktısı

```
Correctly Classified Instances      330          81.2808 %
Kappa statistic                    0.5206
Mean absolute error                0.1882
Root mean squared error            0.4302
Relative absolute error             48.1174 %
Root relative squared error        97.3475 %
Total Number of Instances         406

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,872    0,352    0,872    0,872    0,872    0,521    0,726    0,856    0
      0,648    0,128    0,648    0,648    0,648    0,521    0,726    0,476    1
Weighted Avg.    0,813    0,292    0,813    0,813    0,813    0,521    0,726    0,755

=== Confusion Matrix ===

  a  b  <-- classified as
260 38 |  a = 0
 38 70 |  b = 1
```

Ek 38: Bilgi Kazanımı Nitelik Seçimi Sonrası Naive Bayes Çıktısı

```

Correctly Classified Instances      298          73.399 %
Kappa statistic                    0
Mean absolute error                0.3908
Root mean squared error            0.4419
Relative absolute error            99.924 %
Root relative squared error        99.9997 %
Total Number of Instances         406

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
      1,000    1,000    0,734     1,000    0,847     ?       0,490    0,730    0
      0,000    0,000    ?         0,000    ?         ?       0,490    0,262    1
Weighted Avg.    0,734    0,734    ?         0,734    ?         ?       0,490    0,605

=== Confusion Matrix ===

  a  b  <-- classified as
298  0 |  a = 0
108  0 |  b = 1

```

Ek 39: Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Kümeleme Sonuçları

	Küme 1	Küme 2	Küme 3	Küme 4	Küme 5	Min	Max
Sınıf	0	0	0	0	1		
Vitamin D	Y	X	X	X	X		
Yaş	58,4234	49,7467	57,7956	65	57,9024		
Glukoz	128,4854	127,6792	116,4248	105	113,3735	70	110
HbA1c	6,8502	6,78	7,0127	5,9	6,473	4	6,2
HbA1c(IFCC)	50,4575	50,6811	45,2476	47,9752	46,697	20	42
Kreatinin	0,9192	0,7245	0,7235	1,13	0,741	0,84	1,25
Üre	32,4976	29,9808	31,7469	31,5785	31,5144	17	43
Ürik Asit	5,9352	5,6117	5,2911	5,9	5,3336	3,5	7,2
ALT	28,2597	21,8369	20,3661	23	17,7052	7	40
AST	23,6155	19,3846	19,5517	20,4737	18,7574	9	45
ALP	80,5151	75,3357	69,3633	75	71,1167	30	120
LDH	160,8094	163,4647	159,1019	160,2396	158,4192	120	250
GGT	45,0285	26,9351	20,688	31	20,6402	0	55
Albümin	6,1834	5,9869	6,0706	6,4308	7,7736	3,5	5,2
Bilirübin	0,2128	0,2112	0,211	0,2117	0,2117	0	0,5
Sodyum	138,3426	138,6575	138,9301	139	138,8943	136	146
Potasyum	4,5479	4,5816	4,5328	4,6	4,528	3,5	5,1
Total Kolesterol	204,2458	207,0004	207,7295	207,2571	210,7791	120	200
HDL	41,4507	44,3629	52,7485	46	52,601	40	60

LDL	116,27	124,0805	134,6192	175	127,466	0	130
Trigliserid	188,4785	228,3165	160,9435	129	147,2547	40	160
Demir	77,9396	56,8584	81,7168	41	63,3084	40	160
Demir Bağlama	274,6963	281,0062	254,8824	267,2012	265,0104	155	355
Kalsiyum	9,5883	9,6683	9,6736	9,54	9,7114	8,8	10,6
Fosfor	3,7465	3,7125	3,7253	3,7307	3,7348	2,3	4,7
Magnezyum	2,0409	2,0215	2,0504	1,95	2,0141	1,3	2,7
CK	148,3641	105,8788	93,53	336	89,7994	30	294
ASO	51,2654	52,42	53,3555	52,42	52,42	0	200
CRP	0,7633	1,1427	0,6337	1,94	0,7838	0	0,8
WBC	7,5313	8,4358	6,6138	6,72	7,3389	4	11
RBC	4,8923	4,6428	4,5509	4,32	4,6278	3,8	5,8
HGB	14,1039	12,4083	13,0326	12,2	12,8702	11,5	16,5
HCT	42,619	38,5893	39,7767	36,1	39,4871	37	47
MCV	87,1487	83,3772	87,6325	83,8	85,6111	78	100

Ek 39: Diyabet Hastalarının Veri Seti Üzerinde Çalışılan Kümeleme Sonuçları(devamı)

MCH	28,9706	26,8418	29,2659	28,2	28,0007	27	32
MCHC	33,3304	36,3099	33,078	33,7	32,9215	30	35
RDW-CV	13,9828	14,6865	13,936	13,6	14,3126	11,5	14,5
PLT	240,1272	323,1068	252,4831	431	271,0339	150	400
MPV	9,3164	9,5591	9,7211	8,6	9,3939	8	12
PCT	0,2246	0,307	0,2434	0,37	0,2526	0,12	0,36
PDW	50,2905	47,4109	49,3834	44,9	50,2879	25	65
LY1	2,6555	2,3944	2,2327	1,91	2,373	1,5	4
MO1	0,4599	0,4426	0,3964	0,46	0,3953	0	0,8
NE1	4,4572	5,1549	3,7124	3,75	4,2679	2	7,5
EO1	0,2131	0,2668	0,1787	0,41	0,1917	0	0,4
BA1	0,0426	0,0479	0,0368	0,07	0,0412	0	0,1
MO2	6,1604	5,3618	5,7809	6,9	5,4581	0	10
NE2	58,3656	60,1256	55,4445	55,8	57,0859	40	75

EO2	2,897	3,2561	2,7475	6,1	2,6414	0	6
BA2	0,5778	0,6155	0,5747	1	0,5859	0	2
LY2	30,4507	29,1151	36,3913	28,5	32,9672	20	50
Sedimentasyon	9,3961	13,9626	13,2237	13	11,8992	0	20
İnsülin	10,8631	10,9844	10,6982	10,8527	10,9761	2,6	37,6
T3	3,0283	2,8793	2,8968	2,9452	2,974	1,71	3,71
T4	1,2004	1,1137	1,1594	0,5	1,183	0,7	1,48
TSH	2,011	3,6379	2,4744	52,07	2,4175	0,35	4,94
Ferritin	61,1289	39,8506	44,5375	37,6	35,9642	10	291
VB12	369,0667	397,1893	400,5066	481	460,4341	187	883

Ek 40: K-Ortalama Algoritması Kümeleme Sonrası Destek Vektör Makineleri Çıktısı

```

Correctly Classified Instances      388          95.5665 %
Kappa statistic                    0.8816
Mean absolute error                 0.0443
Root mean squared error             0.2106
Relative absolute error             11.3358 %
Root relative squared error         47.6497 %
Total Number of Instances          406

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,993    0,148    0,949    0,993    0,970    0,885    0,923    0,947    0
      0,852    0,007    0,979    0,852    0,911    0,885    0,923    0,873    1
Weighted Avg.   0,956    0,111    0,957    0,956    0,955    0,885    0,923    0,928

=== Confusion Matrix ===
      a  b  <-- classified as
296  2  |  a = 0
 16 92  |  b = 1

```

Ek 41: K-Ortalama Algoritması Kümeleme Sonrası Karar Ağaçları Çıktısı

```
Correctly Classified Instances      373          91.8719 %
Kappa statistic                    0.7912
Mean absolute error                0.1883
Root mean squared error           0.2729
Relative absolute error            48.1577 %
Root relative squared error        61.7534 %
Total Number of Instances          406

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                0,946   0,157   0,943      0,946   0,945      0,791   0,934    0,961    0
                0,843   0,054   0,850      0,843   0,847      0,791   0,933    0,902    1
Weighted Avg.   0,919   0,130   0,918      0,919   0,919      0,791   0,934    0,945

=== Confusion Matrix ===

  a  b  <-- classified as
282 16 | a = 0
 17 91 | b = 1
```

Ek 42: K-Ortalama Algoritması Kümeleme Sonrası Rastgele Orman Çıktısı

```
Correctly Classified Instances      370          91.133 %
Kappa statistic                    0.7617
Mean absolute error                0.2544
Root mean squared error           0.3116
Relative absolute error            65.0558 %
Root relative squared error        70.5044 %
Total Number of Instances          406

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                0,966   0,241   0,917      0,966   0,941      0,766   0,935    0,964    0
                0,759   0,034   0,891      0,759   0,820      0,766   0,935    0,878    1
Weighted Avg.   0,911   0,186   0,910      0,911   0,909      0,766   0,935    0,941

=== Confusion Matrix ===

  a  b  <-- classified as
288 10 | a = 0
 26 82 | b = 1
```

Ek 43: K-Ortalama Algoritması Kümeleme Sonrası Naive Bayes Çıktısı

```
Correctly Classified Instances      298          73.399 %
Kappa statistic                    0
Mean absolute error                0.3895
Root mean squared error           0.4429
Relative absolute error            99.5988 %
Root relative squared error        100.2234 %
Total Number of Instances          406

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                1,000   1,000   0,734      1,000   0,847      ?      0,494    0,732    0
                0,000   0,000   ?          0,000   ?          ?      0,494    0,265    1
Weighted Avg.   0,734   0,734   ?          0,734   ?          ?      0,494    0,607

=== Confusion Matrix ===

  a  b  <-- classified as
298  0 | a = 0
108  0 | b = 1
```

Ek 44: K-Ortalama Algoritması Kümeleme Sonrası K-En Yakın Komşuluk Algoritması Weka Çıktısı

```
Correctly Classified Instances      295          72.6601 %
Kappa statistic                    0.2808
Mean absolute error                 0.2747
Root mean squared error             0.5217
Relative absolute error             70.2382 %
Root relative squared error         118.0643 %
Total Number of Instances          406

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0.829    0.556    0.805    0.829    0.817    0.281    0.607    0.786    0
      0.444    0.171    0.485    0.444    0.464    0.281    0.607    0.329    1
Weighted Avg.   0.727    0.453    0.720    0.727    0.723    0.281    0.607    0.664

=== Confusion Matrix ===

  a  b  <-- classified as
247 51 | a = 0
 60 48 | b = 1
```

Ek 45: K-Ortalama Algoritması Kümeleme Sonrası Yapay Sinir Ağları Çıktısı

```
Correctly Classified Instances      291          71.6749 %
Kappa statistic                    0.024
Mean absolute error                 0.3584
Root mean squared error             0.4451
Relative absolute error             91.64 %
Root relative squared error         100.7218 %
Total Number of Instances          406

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0.953    0.935    0.738    0.953    0.832    0.036    0.593    0.782    0
      0.065    0.047    0.333    0.065    0.109    0.036    0.593    0.338    1
Weighted Avg.   0.717    0.699    0.630    0.717    0.639    0.036    0.593    0.664

=== Confusion Matrix ===

  a  b  <-- classified as
284 14 | a = 0
101  7 | b = 1
```

Ek 46: Denklemdaki Değişkenler

		Variables in the Equation					95% C.I. for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower Upper
Step 1 ^a	CİNSİYET(1)	,642	,473	1,842	1	,175	1,899	,752 4,797
	YAŞ	,032	,013	6,097	1	,014	1,032	1,007 1,059
	SMEAN(GLUKOZ)	-,004	,005	,770	1	,380	,996	,987 1,005
	SMEAN(HbA1c)	-,041	,082	,255	1	,614	,960	,818 1,126
	SMEAN(HbA1cIFCC)	,013	,016	,655	1	,418	1,013	,982 1,044
	SMEAN(KREATİNİN)	-1,119	1,210	,855	1	,355	,327	,030 3,501
	SMEAN(ÜRE)	,058	,034	2,874	1	,090	1,060	,991 1,134
	SMEAN(ÜRİKASİT)	,243	,163	2,216	1	,137	1,275	,926 1,757
	SMEAN(ALT)	-,034	,026	1,661	1	,197	,967	,918 1,018
	SMEAN(AST)	,016	,040	,165	1	,684	1,016	,940 1,099
	SMEAN(ALP)	,000	,011	,001	1	,981	1,000	,979 1,021
	SMEAN(LDH)	-,007	,009	,574	1	,449	,993	,975 1,011
	SMEAN(GGT)	,015	,007	4,410	1	,036	1,015	1,001 1,030
	SMEAN(ALBÜMİN)	,114	,048	5,686	1	,017	1,121	1,021 1,231
	SMEAN(BİLİRUBİN)	3,308	12,618	,069	1	,793	27,333	,000 1504592487252,679
	SMEAN(SODYUM)	,050	,082	,368	1	,544	1,051	,894 1,236
	SMEAN(POTASYUM)	-,205	,465	,194	1	,660	,815	,328 2,026
	SMEAN(TOTALKOLESTEROL)	,008	,007	1,367	1	,242	1,008	,995 1,021
	SMEAN(HDL)	,003	,019	,025	1	,874	1,003	,966 1,042
	SMEAN(LDL)	-,008	,006	1,742	1	,187	,992	,980 1,004
	SMEAN(TRİGLİSERİD)	-,005	,003	3,678	1	,055	,995	,990 1,000

SMEAN(DEMİR)	-,005	,006	,732	1	,392	,995	,983	1,007
SMEAN(DEMİRBAĞLAMA)	-,005	,004	1,281	1	,258	,995	,988	1,003
SMEAN(KALSİYUM)	,227	,513	,195	1	,659	1,254	,459	3,428
SMEAN(FOSFOR)	2,964	1,834	2,611	1	,106	19,366	,532	705,108
SMEAN(MAGNEZYUM)	-,202	,845	,057	1	,812	,817	,156	4,287
SMEAN(CK)	,003	,002	2,629	1	,105	1,003	,999	1,006
SMEAN(ASO)	-,005	,015	,090	1	,764	,995	,966	1,026
SMEAN(CRP)	-,023	,192	,015	1	,902	,977	,671	1,422
SMEAN(WBC)	,235	,767	,094	1	,759	1,265	,282	5,682

160

SMEAN(RBC)	1,310	2,861	,210	1	,647	3,706	,014	1010,326
SMEAN(HGB)	,383	,411	,868	1	,352	1,466	,655	3,280
SMEAN(HCT)	-,189	,352	,288	1	,591	,828	,415	1,651
SMEAN(MCV)	,051	,158	,105	1	,746	1,052	,773	1,433
SMEAN(MCH)	-,017	,050	,110	1	,740	,984	,892	1,085
SMEAN(MCHC)	-,007	,011	,344	1	,557	,993	,972	1,016
SMEAN(RDWCV)	,094	,189	,249	1	,618	1,099	,759	1,591
SMEAN(PLT)	,001	,020	,005	1	,942	1,001	,962	1,042
SMEAN(MPV)	-,215	,595	,130	1	,718	,807	,252	2,587
SMEAN(PCT)	-,545	22,270	,001	1	,980	,580	,000	5246973212525914100,000
SMEAN(PDW)	,051	,026	3,811	1	,051	1,052	1,000	1,107
SMEAN(LY#)	,004	,080	,003	1	,960	1,004	,858	1,175
SMEAN(MO#)	-,091	1,903	,002	1	,962	,913	,022	38,020
SMEAN(NE#)	-,111	1,125	,010	1	,922	,895	,099	8,114
SMEAN(EO#)	-7,235	7,118	1,033	1	,309	,001	,000	825,106
SMEAN(BA#)	7,775	30,116	,067	1	,796	2380,007	,000	1026852847006945000000000000000,000

SMEAN(MO)	-,280	,206	1,844	1	,175	,756	,504	1,132
SMEAN(@NE)	-,001	,085	,000	1	,990	,999	,846	1,180
SMEAN(E0)	,448	,501	,798	1	,372	1,565	,586	4,180
SMEAN(BA)	-,321	2,024	,025	1	,874	,725	,014	38,320
SMEAN(LY)	-,002	,010	,031	1	,860	,998	,978	1,018
SMEAN(SEDİMENTASYON)	-,033	,029	1,300	1	,254	,968	,914	1,024
SMEAN(İNSULİN)	,030	,102	,085	1	,771	1,030	,843	1,260
SMEAN(T3)	,847	,420	4,055	1	,044	2,332	1,023	5,316
SMEAN(T4)	-2,341	,909	6,638	1	,010	,096	,016	,571
SMEAN(TSH)	-,157	,084	3,460	1	,063	,855	,724	1,008
SMEAN(FERRİTİN)	-,009	,005	2,827	1	,093	,991	,982	1,001
SMEAN(VB12)	,000	,001	,143	1	,705	1,000	,999	1,001
Constant	-30,849	22,468	1,885	1	,170	,000		

Ek 47: Yapay Sinir Ağları Çıktısı

```
Correctly Classified Instances      230          82.7338 %
Kappa statistic                    0.6532
Mean absolute error                 0.1806
Root mean squared error             0.3984
Relative absolute error             36.3138 %
Root relative squared error         79.8922 %
Total Number of Instances          278

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                0,832    0,178    0,844    0,832    0,838      0,653    0,855    0,824    0
                0,822    0,168    0,809    0,822    0,815      0,653    0,855    0,810    1
Weighted Avg.   0,827    0,173    0,828    0,827    0,827      0,653    0,855    0,817

=== Confusion Matrix ===

  a  b  <-- classified as
124 25 |  a = 0
 23 106 | b = 1
```

Ek 48: Naive Bayes Çıktısı

```
Correctly Classified Instances      218          78.4173 %
Kappa statistic                    0.5625
Mean absolute error                 0.3123
Root mean squared error             0.4194
Relative absolute error             62.7797 %
Root relative squared error         84.0898 %
Total Number of Instances          278

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                0,852    0,295    0,770    0,852    0,809      0,566    0,794    0,777    0
                0,705    0,148    0,805    0,705    0,752      0,566    0,794    0,748    1
Weighted Avg.   0,784    0,226    0,786    0,784    0,783      0,566    0,794    0,764

=== Confusion Matrix ===

  a  b  <-- classified as
127 22 |  a = 0
 38 91 |  b = 1
```

Ek 49: Rastgele Orman Çıktısı

```
Correctly Classified Instances      217          78.0576 %
Kappa statistic                    0.5554
Mean absolute error                 0.3716
Root mean squared error             0.4178
Relative absolute error             74.6914 %
Root relative squared error         83.7809 %
Total Number of Instances          278

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                0,846    0,295    0,768    0,846    0,805      0,559    0,807    0,799    0
                0,705    0,154    0,798    0,705    0,749      0,559    0,807    0,768    1
Weighted Avg.   0,781    0,230    0,782    0,781    0,779      0,559    0,807    0,785

=== Confusion Matrix ===

  a  b  <-- classified as
126 23 |  a = 0
 38 91 |  b = 1
```

Ek 50: Karar Ağaçları Çıktısı

```
Correctly Classified Instances      212          76.259 %
Kappa statistic                    0.5172
Mean absolute error                0.3618
Root mean squared error            0.4179
Relative absolute error             72.7337 %
Root relative squared error        83.7853 %
Total Number of Instances         278

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                0,852    0,341    0,743      0,852    0,794      0,524    0,818    0,810    0
                0,659    0,148    0,794      0,659    0,720      0,524    0,818    0,779    1
Weighted Avg.   0,763    0,251    0,767      0,763    0,760      0,524    0,818    0,796

=== Confusion Matrix ===

  a  b  <-- classified as
127 22 | a = 0
 44 85 | b = 1
```

Ek 51: Destek Vektör Makineleri Çıktısı

```
Correctly Classified Instances      148          53.2374 %
Kappa statistic                    -0.0072
Mean absolute error                0.4676
Root mean squared error            0.6838
Relative absolute error            94.0052 %
Root relative squared error        137.1175 %
Total Number of Instances         278

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                0,993    1,000    0,534      0,993    0,695     -0,056    0,497    0,534    0
                0,000    0,007    0,000      0,000    0,000     -0,056    0,497    0,464    1
Weighted Avg.   0,532    0,539    0,286      0,532    0,372     -0,056    0,497    0,502

=== Confusion Matrix ===

  a  b  <-- classified as
148  1 | a = 0
129  0 | b = 1
```

Ek 52: K-En Yakın Komşuluk Algoritması Çıktısı

```
Correctly Classified Instances      98          35.2518 %
Kappa statistic                    -0.2623
Mean absolute error                0.6463
Root mean squared error            0.8015
Relative absolute error            129.926 %
Root relative squared error        160.7067 %
Total Number of Instances         278

=== Detailed Accuracy By Class ===

                TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                0,195    0,465    0,326      0,195    0,244     -0,289    0,350    0,477    0
                0,535    0,805    0,365      0,535    0,434     -0,289    0,350    0,398    1
Weighted Avg.   0,353    0,623    0,344      0,353    0,332     -0,289    0,350    0,441

=== Confusion Matrix ===

  a  b  <-- classified as
29 120 | a = 0
 60  69 | b = 1
```

Ek 53: Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Nitelik Seçimi Sonuçları

Nitelik Sırası	Bilgi Kazanımı	Ki-Kare	Kazanım Oranı	Relief-f
1	Cinsiyet	Cinsiyet	Cinsiyet	BA1
2	VB12	VB12	Glukoz	NE2
3	Glukoz	ALT	VB12	BA2
4	ALT	Glukoz	MCV	Total Kolesterol
5	MCV	MCV	ALT	Kreatinin
6	HDL	HDL	Trigliserid	Yaş
7	Trigliserid	Trigliserid	HDL	MO2
8	PDW	PDW	PDW	PDW
9	Bilirübin	Bilirübin	Bilirübin	RDW-CV
10	Sodyum	Sodyum	Total Kolesterol	MCHC
11	Demir	Demir	Sodyum	AST
12	Total Kolesterol	Total Kolesterol	Potasyum	PLT
13	Potasyum	Potasyum	Demir	T4
14	Fosfor	Fosfor	LDL	Glukoz
15	Demir Bağlama	Demir Bağlama	Demir Bağlama	LY2
16	LDL	LDL	Albümin	Albümin
17	Kalsiyum	Kalsiyum	Kalsiyum	VB12
18	TSH	TSH	T4	HDL
19	Albümin	Albümin	Magnezyum	RBC
20	CK	CK	Kreatinin	TSH
21	GGT	GGT	HbA1c	Ferritin
22	HbA1c(IFCC)	HbA1c(IFCC)	TSH	MO1

23	HbA1c	HbA1c	Yaş	ASO
24	Yaş	Yaş	HbA1c(IFCC)	Trigliserid
25	Kreatinin	Kreatinin	Üre	Bilirubin
26	Üre	Üre	GGT	ALT
27	Ürik Asit	Ürik Asit	Ürik Asit	İnsülin
28	LDH	LDH	LDH	Magnezyum
29	ALP	ALP	ALP	Demir
30	AST	AST	AST	Kalsiyum
31	Magnezyum	Magnezyum	Fosfor	T3

Ek 53: Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Nitelik Seçimi Sonuçları (devamı)

32	ASO	ASO	ASO	CRP
33	T4	T4	CK	HCT
34	EO2	EO2	NE2	EO2
35	MO2	MO2	BA1	WBC
36	BA1	BA1	EO1	Fosfor
37	EO1	EO1	NE1	HGB
38	NE2	NE2	MO2	NE1
39	BA2	BA2	EO2	Sedimentasyon
40	Ferritin	Ferritin	T3	LDL
41	LY2	LY2	BA2	MPV
42	T3	T3	İnsülin	LY1
43	İnsülin	İnsülin	Sedimentasyon	MCV
44	Sedimentasyon	Sedimentasyon	LY2	Potasyum
45	NE1	NE1	MO1	HbA1c
46	MO1	MO1	LY1	Üre
47	LY1	LY1	PCT	LDH

48	PCT	PCT	MPV	PCT
49	HGB	HGB	RBC	EO1
50	RBC	RBC	WBC	GGT
51	WBC	WBC	Ferritin	ALP
52	HCT	HCT	HGB	Demir Bağlama
53	MCH	MCH	HCT	MCH
54	MCHC	MCHC	MCH	CK
55	MPV	MPV	PLT	Sodyum
56	PLT	PLT	RDW-CV	HbA1c(IFCC)
57	RDW-CV	RDW-CV	MCHC	Ürik Asit
58	CRP	CRP	CRP	Cinsiyet

Ek 54: Kazanım Oranı Nitelik Seçimi Sonrası Yapay Sinir Ağları Çıktısı

```

Correctly Classified Instances      240          86.3309 %
Kappa statistic                    0.7283
Mean absolute error                0.1885
Root mean squared error            0.3418
Relative absolute error             37.8927 %
Root relative squared error        68.5283 %
Total Number of Instances         278

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
      0.799   0.062   0.937     0.799   0.862     0.737   0.898   0.921    0
      0.938   0.201   0.801     0.938   0.864     0.737   0.898   0.840    1
Weighted Avg.   0.863   0.127   0.874     0.863   0.863     0.737   0.898   0.883

=== Confusion Matrix ===

  a  b  <-- classified as
119 30 | a = 0
 8 121 | b = 1

```

Ek 55: Relief-f Nitelik Seçimi Sonrası Destek Vektör Makineleri Çıktısı

```
Correctly Classified Instances      234          84.1727 %
Kappa statistic                    0.6825
Mean absolute error                 0.1583
Root mean squared error             0.3978
Relative absolute error             31.8171 %
Root relative squared error         79.7715 %
Total Number of Instances          278

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,839    0,155    0,862     0,839    0,850     0,683    0,842     0,810     0
      0,845    0,161    0,820     0,845    0,832     0,683    0,842     0,764     1
Weighted Avg.   0,842    0,158    0,842     0,842    0,842     0,683    0,842     0,789

=== Confusion Matrix ===

  a  b  <-- classified as
125 24 |  a = 0
 20 109 |  b = 1
```

Ek 56: Relief-f Nitelik Seçimi Sonrası Karar Ağaçları Çıktısı

```
Correctly Classified Instances      221          79.4964 %
Kappa statistic                    0.5837
Mean absolute error                 0.5017
Root mean squared error             0.523
Relative absolute error             100.8612 %
Root relative squared error         104.8605 %
Total Number of Instances          278

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,872    0,295    0,774     0,872    0,820     0,589    0,722     0,662     0
      0,705    0,128    0,827     0,705    0,762     0,589    0,720     0,776     1
Weighted Avg.   0,795    0,217    0,799     0,795    0,793     0,589    0,721     0,715

=== Confusion Matrix ===

  a  b  <-- classified as
130 19 |  a = 0
 38 91 |  b = 1
```

Ek 57: Bilgi Kazanımı Nitelik Seçimi Sonrası Naive Bayes Çıktısı

```
Correctly Classified Instances      218          78.4173 %
Kappa statistic                    0.5629
Mean absolute error                 0.312
Root mean squared error             0.4196
Relative absolute error             62.7259 %
Root relative squared error         84.1346 %
Total Number of Instances          278

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,846    0,287    0,773     0,846    0,808     0,566    0,793     0,776     0
      0,713    0,154    0,800     0,713    0,754     0,566    0,793     0,738     1
Weighted Avg.   0,784    0,225    0,786     0,784    0,783     0,566    0,793     0,758

=== Confusion Matrix ===

  a  b  <-- classified as
126 23 |  a = 0
 37 92 |  b = 1
```

Ek 58: Bilgi Kazanımı Nitelik Seçimi Sonrası Rastgele Orman Çıktısı

```
Correctly Classified Instances      217          78.0576 %
Kappa statistic                    0.5554
Mean absolute error                0.3716
Root mean squared error            0.4178
Relative absolute error            74.6914 %
Root relative squared error        83.7809 %
Total Number of Instances         278

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0.846    0.295    0.768    0.846    0.805    0.559    0.807    0.799    0
      0.705    0.154    0.798    0.705    0.749    0.559    0.807    0.768    1
Weighted Avg.    0.781    0.230    0.782    0.781    0.779    0.559    0.807    0.785

=== Confusion Matrix ===
      a  b  <-- classified as
    126  23 |  a = 0
     38  91 |  b = 1
```

Ek 59: Bilgi Kazanımı Nitelik Seçimi Sonrası K-En Yakın Komşuluk Algoritması Çıktısı

```
Correctly Classified Instances      215          77.3381 %
Kappa statistic                    0.5413
Mean absolute error                0.3518
Root mean squared error            0.4209
Relative absolute error            70.7158 %
Root relative squared error        84.401 %
Total Number of Instances         278

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0.832    0.295    0.765    0.832    0.797    0.544    0.718    0.688    0
      0.705    0.168    0.784    0.705    0.743    0.544    0.718    0.657    1
Weighted Avg.    0.773    0.236    0.774    0.773    0.772    0.544    0.718    0.674

=== Confusion Matrix ===
      a  b  <-- classified as
    124  25 |  a = 0
     38  91 |  b = 1
```

Ek 60: Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Kümeleme Sonuçları

	Küme 1	Küme 2	Küme 3	Küme 4	Küme 5	Min	Max
Vitamin D	0	1	1	0	0		
Cinsiyet	Y	Y	X	X	Y		
Yaş	19,6667	36,4474	36,4835	46,16	50,4909		
Glukoz	94,0938	95,9291	86,985	89,3099	100,5303	70	110
HbA1c	5,4629	5,597	5,542	5,4546	5,8368	4	6,2
HbA1c(IFCC)	36,553	37,0655	37,2756	36,9742	39,7109	20	42
Kreatinin	0,7277	0,8312	0,7425	0,7226	0,8861	0,84	1,25

Üre	25,9836	30,3689	26,977	27,6852	29,1374	17	43
Ürik Asit	5,1852	5,4268	5,0638	4,7087	5,3428	3,5	7,2
ALT	20,7358	19,5871	18,4242	22,4899	26,0481	7	40
AST	23,7339	22,8541	20,7991	19,6866	23,009	9	45
ALP	65,7045	65,677	64,982	67,5618	66,0748	30	120
LDH	167,1739	163,7208	166,613	169,4991	169,4308	120	250
GGT	16,4524	16,24	16,0351	16,501	17,2675	0	55
Albümin	9,9693	11,3039	9,6587	9,2894	9,8702	3,5	5,2
Bilirübin	0,4184	0,4075	0,4133	0,4133	0,4111	0	0,5
Sodyum	138,7592	139,4534	138,8166	139,0585	138,8769	136	146
Potasyum	4,4888	4,4444	4,4746	4,5168	4,4739	3,5	5,1
Total Kolesterol	181,6918	187,169	191,9939	189,7876	202,3216	120	200
HDL	44,9653	43,5644	48,8044	48,5224	42,5147	40	60
LDL	113,7492	114,5742	121,6602	133,5612	124,5942	0	130
Trigliserid	161,3315	172,9707	158,2979	160,2068	244,6382	40	160
Demir	82,2492	81,454	75,57	71,5978	89,7107	40	160
Demir Bağlama	268,7256	256,283	267,7406	285,1731	257,865	155	355
Kalsiyum	177,4913	136,9954	254,6669	92,6421	141,6147	8,8	10,6
Fosfor	4,0571	4,0397	4,0728	4,0409	4,0506	2,3	4,7
Magnezyum	2,1008	2,0451	2,0611	2,0729	2,0663	1,3	2,7
CK	74,0568	79,0723	72,6562	67,985	75,6687	30	294
ASO	185,7275	185,6	187,4615	182,376	183,8255	0	200

Ek 60: Diyabet Hastalığı Olmayan Hastaların Veri Seti Üzerinde Çalışılan Kümeleme Sonuçları (devamı)

CRP	1,5379	1,8249	1,5404	0,8616	1,1476	0	0,8
WBC	6,8942	7,4647	6,7373	6,6396	6,8754	4	11

RBC	4,6704	5,0174	4,5547	4,4593	4,7182	3,8	5,8
HGB	12,9986	13,9025	12,5845	12,5356	13,2696	11,5	16,5
HCT	39,845	42,2284	38,6473	38,7406	40,6819	37	47
MCV	85,69	84,7937	85,1298	87,3052	86,502	78	100
MCH	27,945	27,86	27,7407	28,2286	28,2128	27	32
MCHC	32,6179	32,9124	32,5743	32,313	32,625	30	35
RDW-CV	14,085	14,08	14,126	14,1358	13,9975	11,5	14,5
PLT	267,2071	252,8902	273,1853	276,218	263,1118	150	400
MPV	9,4171	9,2992	9,5325	9,2088	9,4025	8	12
PCT	0,2476	0,2311	0,2557	0,2517	0,2437	0,12	0,36
PDW	49,29	51,2226	49,7638	44,4892	49,3529	25	65
LY1	2,1834	2,4833	2,0732	2,1858	2,1576	1,5	4
MO1	0,3903	0,4529	0,3686	0,3665	0,3937	0	0,8
NE1	3,9775	4,1221	3,9691	3,7673	3,987	2	7,5
EO1	0,1851	0,2082	0,1826	0,1694	0,1806	0	0,4
BA1	0,0309	0,0335	0,0298	0,0315	0,0308	0	0,1
MO2	5,7364	6,0854	5,5827	5,6102	5,807	0	10
NE2	57,0793	54,2557	58,3045	56,0502	57,4706	40	75
EO2	2,7221	2,805	2,7472	2,6102	2,6743	0	6
BA2	0,46	0,4695	0,4455	0,4928	0,4625	0	2
LY2	32,1429	34,1861	31,1834	33,396	31,7491	20	50
Sedimentasyon	10,9783	9,615	11,2888	13,5939	10,2174	0	20
İnsülin	14,3954	12,5526	13,3212	13,1824	13,0671	2,6	37,6
T3	3,6845	3,4435	3,1686	3,1046	3,1436	1,71	3,71
T4	1,1955	1,2017	1,1588	1,0772	1,1484	0,7	1,48
TSH	2,1039	1,9596	2,0283	4,0038	1,8839	0,35	4,94
Ferritin	48,863	75,5281	43,6618	35,6782	75,3071	10	291

VB12	308,1594	309,034	432,424	385,0375	296,9208	187	883
------	----------	---------	---------	----------	----------	-----	-----

Ek 61: K-Ortalama Algoritması Kümeleme Sonrası Yapay Sinir Ağları Çıktısı

```

Correctly Classified Instances      154          55.3957 %
Kappa statistic                    0.1014
Mean absolute error                 0.4899
Root mean squared error             0.5094
Relative absolute error             98.4825 %
Root relative squared error         102.1498 %
Total Number of Instances          278

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0.597    0.496    0.582    0.597    0.589      0.101    0.548    0.564    0
      0.504    0.403    0.520    0.504    0.512      0.101    0.548    0.494    1
Weighted Avg.   0.554    0.453    0.553    0.554    0.553      0.101    0.548    0.532

=== Confusion Matrix ===

  a  b  <-- classified as
89 60 | a = 0
64 65 | b = 1

```

Ek 62: K-Ortalama Algoritması Kümeleme Sonrası Naive Bayes Çıktısı

```

Correctly Classified Instances      149          53.5971 %
Kappa statistic                     0
Mean absolute error                 0.4974
Root mean squared error             0.4987
Relative absolute error             99.998 %
Root relative squared error         100 %
Total Number of Instances          278

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      1,000    1,000    0,536    1,000    0,698      ?      0,493    0,533    0
      0,000    0,000      ?      0,000      ?      ?      0,493    0,461    1
Weighted Avg.   0,536    0,536      ?      0,536      ?      ?      0,493    0,499

=== Confusion Matrix ===

  a  b  <-- classified as
149  0 | a = 0
129  0 | b = 1

```

Ek 63: K-Ortalama Algoritması Kümeleme Sonrası Destek Vektör Makineleri

Çıktısı

```
Correctly Classified Instances      146          52.518 %
Kappa statistic                    0.0494
Mean absolute error                0.4748
Root mean squared error            0.6891
Relative absolute error            95.4514 %
Root relative squared error        138.1682 %
Total Number of Instances         278

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
      0.530    0.481    0.560    0.530    0.545      0.049    0.525    0.549      0
      0.519    0.470    0.489    0.519    0.504      0.049    0.525    0.477      1
Weighted Avg.   0.525    0.476    0.527    0.525    0.526      0.049    0.525    0.516

=== Confusion Matrix ===

  a  b  <-- classified as
79 70 |  a = 0
62 67 |  b = 1
```

Ek 64: K-Ortalama Algoritması Kümeleme Sonrası Rastgele Orman Çıktısı

```
Correctly Classified Instances      141          50.7194 %
Kappa statistic                    -0.0276
Mean absolute error                0.4883
Root mean squared error            0.497
Relative absolute error            98.164 %
Root relative squared error        99.6621 %
Total Number of Instances         278

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
      0.772    0.798    0.528    0.772    0.627     -0.032    0.566    0.638      0
      0.202    0.228    0.433    0.202    0.275     -0.032    0.566    0.500      1
Weighted Avg.   0.507    0.534    0.484    0.507    0.464     -0.032    0.566    0.574

=== Confusion Matrix ===

  a  b  <-- classified as
115 34 |  a = 0
103 26 |  b = 1
```

Ek 65: K-Ortalama Algoritması Kümeleme Sonrası K-En Yakın Komşuluk Algoritması Çıktısı

```
Correctly Classified Instances      141          50.7194 %
Kappa statistic                    0.0436
Mean absolute error                0.4929
Root mean squared error            0.6992
Relative absolute error            99.0785 %
Root relative squared error        140.2038 %
Total Number of Instances         278

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
      0.309    0.264    0.575    0.309    0.402      0.050    0.514    0.543      0
      0.736    0.691    0.480    0.736    0.581      0.050    0.514    0.471      1
Weighted Avg.   0.507    0.462    0.531    0.507    0.485      0.050    0.514    0.510

=== Confusion Matrix ===

  a  b  <-- classified as
46 103 |  a = 0
34 95 |  b = 1
```

Ek 66: K-Ortalama Algoritması Kümeleme Sonrası Karar Ağaçları Çıktısı

```
Correctly Classified Instances      140          50.3597 %
Kappa statistic                    -0.0379
Mean absolute error                  0.4893
Root mean squared error              0.5014
Relative absolute error              98.3621 %
Root relative squared error          100.54 %
Total Number of Instances           278

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
      0,785    0,822    0,525    0,785    0,629    -0,046    0,574    0,614    0
      0,178    0,215    0,418    0,178    0,250    -0,046    0,576    0,501    1
Weighted Avg.    0,504    0,540    0,475    0,504    0,453    -0,046    0,575    0,562

=== Confusion Matrix ===

  a  b  <-- classified as
117 32 | a = 0
106 23 | b = 1
```

Ek 67: Etik Kurul Kararı

T.C. ÇUKUROVA ÜNİVERSİTESİ TIP FAKÜLTESİ GİRİŞİMSEL OLMAYAN KLİNİK ARAŞTIRMALAR ETİK KURULU

Toplantı Sayısı	Tarih
89	14 Haziran 2019

KARAR NO 62- Çukurova Üniversitesi Mühendislik Mimarlık Fakültesi Endüstri Mühendisliği Bölümü'nde Doç. Dr. Oya Hacire Yüreğir yönetiminde, Doktor Öğretim Üyesi Birsen İrem Selamoğlu'nun katkılarıyla, Uğur Engin Eşsiz, tarafından yürütülmesi öngörülen, "Laboratuvar Tetkik Sonuçları Üzerinde Makine Öğrenmesi Uygulamaları" başlıklı doktora tez projesi araştırma etiği yönünden değerlendirildi. Toplantıya katılan üyelerin oybirliğiyle uygun olduğuna karar verildi.

BAŞKAN	Prof Dr Selim Kadioğlu Tıp Tarihi ve Etik Anabilim Dalı	
ÜYELER	Prof Dr Davut Alptekin Tıbbi Biyoloji Anabilim Dalı	
	Prof Dr Dinçer Yıldızdaş Çocuk Sağlığı ve Hastalıkları Anabilim Dalı	
	Prof Dr Gülşah Seydaoğlu Biyoistatistik Anabilim Dalı	
	Prof Dr Gürhan Sakman Genel Cerrahi Anabilim Dalı	
	Prof Dr Murat Gündüz Anesteziyoloji ve Reanimasyon Anabilim Dalı	
	Doç Dr Ezgi Özyılmaz Saraç Göğüs Hastalıkları Anabilim Dalı	
	Av. Zehra Bulut Hukukçu Üye	
	Dr Neşe Kayrın Kurum Dışı Üye	

Çukurova Üniversitesi Tıp Fakültesi Dekanlık Binası, Balcalı 01330 Adana
Telefon: 0322 338 60 60 dahili 3465, Faks: 0322 338 67 22