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

BİYOİNFORMATİK VERİ SINIFLANDIRMA PROBLEMLERİ İÇİN BOYUT İNDİRGEMEYE DAYALI ÖZNİTELİK SEÇİMİ YAKLAŞIMLARININ GELİŞTİRİLMESİ

Hazırlayan
Umay Gülfem ELGÜN

Danışman
Dr. Öğr. Üyesi Özkan Ufuk NALBANTOĞLU

Yüksek Lisans Tezi

Ocak 2019
KAYSERİ

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

BİYOİNFORMATİK VERİ SINIFLANDIRMA PROBLEMLERİ İÇİN BOYUT İNDİRGEMEYE DAYALI ÖZNİTELİK SEÇİMİ YAKLAŞIMLARININ GELİŞTİRİLMESİ

(Yüksek Lisans Tezi)

Hazırlayan
Umay Gülfem ELGÜN

Danışman
Dr. Öğr. Üyesi Özkan Ufuk NALBANTOĞLU

Ocak 2019
KAYSERİ

BİLİMSEL ETİĞE UYGUNLUK

Bu çalışmadaki tüm bilgilerin, akademik ve etik kurallara uygun bir şekilde elde edildiğini beyan ederim. Aynı zamanda bu kural ve davranışların gerektirdiği gibi, bu çalışmanın özünde olmayan tüm materyal ve sonuçları tam olarak aktardığımı ve referans gösterdiğimi belirtirim.

Umay Gülfem ELGÜN

YÖNERGEYE UYGUNLUK

"Biyoinformatik Veri Sınıflandırma Problemleri için Boyut İndirgemeye Dayalı Öznitelik Seçimi Yaklaşımlarının Geliştirilmesi" adlı Yüksek Lisans tezi, Erciyes Üniversitesi Lisansüstü Tez Önerisi ve Tez Yazma Yönergesi'ne uygun olarak hazırlanmıştır.

Tezi Hazırlayan
Umay Gülfem ELGÜN

Danışman
Dr. Öğr. Üyesi Ö.Ufuk NALBANTOĞLU

Bilgisayar Mühendisliği ABD Başkanı
Prof. Dr. Veysel ASLANTAŞ

Dr. Öğr. Üyesi Özkan Ufuk NALBANTOĞLU danışmanlığında **Umay Gülfem ELGÜN** tarafından hazırlanan **"Biyoinformatik veri sınıflandırma problemleri için boyut indirgemeye dayalı öznitelik seçimi yaklaşımlarının geliştirilmesi"** adlı bu çalışma jürimiz tarafından Erciyes Üniversitesi Fen Bilimleri Enstitüsü **Bilgisayar Mühendisliği** Anabilim Dalında **yüksek lisans** tezi olarak kabul edilmiştir.

18 / 01 / 2019

JÜRİ:

Danışman : Dr. Öğr. Üyesi Ö. Ufuk NALBANTOĞLU .....

Üye : Doç. Dr. Celal ÖZTÜRK .....

Üye : Dr. Öğr. Üyesi Burcu GÜNGÖR .....

ONAY:

Bu tezin kabulü Enstitü Yönetim Kurulunun 29/01/2019. tarih ve 2019/07-15 sayılı kararı ile onaylanmıştır.

29 / 01 / 2019

Prof. Dr. Mehmet AKKURT

Enstitü Müdürü

ÖNSÖZ / TEŞEKKÜR

Araştırmalarımın her aşamasında bilgi, öneri ve yardımlarını esirgemeyerek akademik ortamda olduğu kadar insani ilişkilerde de sonsuz desteğiyle gelişmeme katkıda bulunan danışman hocam sayın Dr. Öğr. Üyesi Özkan Ufuk NALBANTOĞLU' na teşekkürü bir borç bilirim.

Çalışmalarım sırasında karşılaştığım zorlukları aşmamda yardımlarından dolayı Doç. Dr. Aycan GÜNDOĞDU' ya, Genom ve Kök Hücre Merkezi' nde aynı odayı paylaştığımız çalışma arkadaşlarıma ve yaşamımın her döneminde bana duydukları güven için aileme en derin duygularla teşekkür ederim.

Umay Gülfem ELGÜN

Kayseri, Ocak 2019

BİYOİNFORMATİK VERİ SINIFLANDIRMA PROBLEMLERİ İÇİN BOYUT İNDİRGEMEYE DAYALI ÖZNİTELİK SEÇİMİ YAKLAŞIMLARININ GELİŞTİRİLMESİ

Umay Gülfem ELGÜN

Erciyes Üniversitesi, Fen Bilimleri Enstitüsü

Yüksek Lisans Tezi, Ocak 2019

Danışman: Dr. Öğr. Üyesi Özkan Ufuk Nalbantoğlu

ÖZET

Öznitelik Seçimi (ÖS), makine öğrenme, veri madenciliği ve biyoinformatik gibi alanlarda, verilerin boyutsallığını azaltmak ve sınıflandırma algoritması gibi bir algoritmanın performansını artırmak için kullanılmaktadır. ÖS teknikleri, artan veri boyutlarından dolayı, birçok alandaki uygulamalarda belirgin bir ihtiyaç haline gelmiştir. Bu tezde, yüksek boyutluluğun getirdiği, boyutsallık laneti, aşırı öğrenme gibi sorunlardan; ÖS için sıklıkla kullanılan yaklaşımlardan ve her birinin avantaj ve dezavantajlarından bahsedilmektedir. Ayrıca, ÖS uygulamalarında sıkça rastlanan sorunlar için geliştirilecek olan teorik çözüm anlatılmaktadır. Asıl özellik uzayındaki verilerin daha düşük boyutlu bir uzaya taşınıp, ÖS'nin bu uzayda yapılması ve bu daha düşük boyutlu uzayda seçilen aşırı öğrenmeye dayanıklı özelliklerin, asıl uzaya geri yansıtılmasıyla oluşan sonuçlar incelenmiştir.

Anahtar Kelimeler: Öznitelik seçimi; veri sınıflandırma; boyut indirgeme; yüksek boyutlu veri.

DEVELOPING DIMENSIONALITY REDUCTION BASED FEATURE SELECTION APPROACHES FOR BIOINFORMATIC DATA CLASSIFICATION PROBLEMS

Umay Gülfem ELGÜN

Erciyes University, Graduate School of Natural and Applied Sciences

Master Thesis, January 2019

Supervisor: Asst. Prof. Özkan Ufuk NALBANTOĞLU

ABSTRACT

Feature selection (FS) is used in areas such as machine learning, data mining and bioinformatics to reduce the dimensionality of data and increase the performance of an algorithm such as a classification algorithm. FS techniques have become an apparent need in many applications because of the growing data size. In this paper, high-dimensionality problems, such as the curse of dimensionality and overfitting are considered. FS approaches that commonly used are explained and their advantages and disadvantages are discussed. It also proposes a theoretical solution for frequently encountered problems in FS applications. Datas harboring in a feature space are transformed to a lower dimensional space and features are selected in that space. This overfitting robust features are reconstructed to the primary space and are observed as the result of that process.

Keywords: Feature selection; data classification; dimensionality reduction; high-dimentional data.

İÇİNDEKİLER

BİYOİNFORMATİK VERİ SINIFLANDIRMA PROBLEMLERİ İÇİN BOYUT İNDİRGEMEYE DAYALI ÖZNİTELİK SEÇİMİ YAKLAŞIMLARININ GELİŞTİRİLMESİ

KISALTMALAR VE SİMGELER

Sembol	Anlamı	Birimi
ÖS	Özellik Seçimi	--
FS	Feature Selection	--
RFE	Recursive Feature Elimination	--
SVM	Support Vector Machine	--
LDA	Linear Discriminant Analysis	--
MSE	Mean Squares Error	--
CFS	Corelation-based Feature Selection	--
USC	Uncorrelated Shrunken Centroid	--
BSS/WSS	Between-group to Within-group Sum of Squares	--
TNoM	Threshold Number of Misclassification	--
FCBF	Fast Corelation-based Feature Selection	--
SFS	Sequential Forward Selection	--
SBE	Sequential Backward Elimination	--
YSA	Yapay Sinir Ağları	--
YBDÖS	Yüksek Boyut, Düşük Örnek Sayısı	--
DSA	Derin Sinir Ağları	--
DSK	Derin Sinirsel Kovalama	--
EAA	Eğri Altında kalan Alan	--
AUC	Area Under the Curve	--
ROC	Receiver Operating Characteristic	--
AUROC	Area Under the ROC Curve	--
AUPR	Area Under the PR Curve	--
RO	Rastgele Orman	--
DVM	Destek Vektör Makineleri	--
LR	Lojistik Regresyon	--
KEYK	K En Yakın Komşu	--
KA	Karar Ağaçları	--
GNB	Gaussian Naive Bayes	--
NOMA	Negatif Olmayan Matris Ayrıştırması	--
PNOMA	Projektif Negatif Olmayan Matris Ayrıştırması	--

HD	Hata Değeri	--
BHD	Bağıl Hata Değeri	--
TDA	Tekil Değer Ayrıştırması	--
TDÖS	Tek Değişkenli Özellik Seçimi	--
ÖÖS	Özyineli Öznitelik Eleme	--
İBH	İltihaplı Bağırsak Hastalığı	--
KK	Kolon Kanseri	--
T2D	Tip-2 Diyabet	--
T2Dk	Kadınlarda Tip-2 Diyabet	--

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

GİRİŞ

Günümüzde, genom teknolojilerindeki ilerlemeler ile transkriptom, proteom, metabolom ve sınıflandırma için öznitelik seçimi (ÖS), veri toplama ve depolama teknolojilerindeki hızlı gelişmelerle giderek daha önemli bir araştırma alanı haline gelmiştir. Bu gelişmeler, kuruluşların büyük, yüksek boyutlu ve karmaşık veri kümeleri oluşturmalarına olanak sağlamıştır [1-4]. Hesaplamalı biyolojinin temel problemlerinden biri, bu yüksek boyutlu veri kümeleri içerisinde biyolojik olarak anlamlı özniteliklerin belirlenerek mevcut hipotezlerin test edilebilmesi veya yeni hipotezler ortaya atılabilmesidir. Fakat, alışageldik düşük boyutlu ve bağıl olarak yüksek sayıdaki örneklerden oluşan veri kümelerine oranla, söz konusu veri kümeleri farklı problemleri de beraberinde getirmektedir. Yüksek boyutlu verilerde, genel olarak fazla öznitelik bulunması sebebiyle sınıflandırma performansının yüksek olması beklenmektedir. Ancak bunun yanında yüksek boyutlu veri setlerindeki bazı ilgisiz ve gereksiz özellikler ve bunlara ek olarak gürültü, sınıflandırmanın performansını kötü etkilemekte, hesaplama karmaşıklığını ve bellek depolama gereksinimlerini önemli ölçüde artırmaktadır [5]. Veri setlerini işlemeyi zorlaştıran bu sebeplerden dolayı boyutsallığın azaltılması gerekmektedir [1,5]. Boyut indirgeme, sınıflandırma veya kümeleme ön işlemlerinde hayati öneme sahiptir [6].

ÖS'de yüksek boyutlu asıl veriyi bir kere kullanıp, sonrasında sadece gerekli olan alt kümeyi kullanabilmek amaçlanır. Bu sayede gözlemlenen çıktı ile ilişkili öznitelikler ve imzalar ortaya çıkarılırken, belli durumlarda da gürültü ve ilgisiz özniteliklerin elenmesiyle veri analizinin performansının artırılabilmesi sağlanmaktadır. Öte yandan, seçilim ile genellikle işlemler hızlanır ve gereksiz bellek kullanımı engellenmiş olur. Öznitelik seçim teknikleri, birçok avantajının yanı sıra bir dezavantajı da beraberinde getirir. İlgili özniteliklerin bir alt kümesinin araştırılması, ek bir karmaşıklık katmanına sebep olur. Bu yüzden araştırma yapılırken, ilgili özniteliklerin yalnızca optimum alt kümesini bulmak gerekir [7]. Bu tezde, latent öznitelik seçimi ile gözlemlenen uzaydaki

ÖS yaklaşımlarını eşzamanlı olarak gerçekleştirmeye yönelik bir yaklaşım ve bununla bağıntılı algoritmalar geliştirilip test edilmiş ve bu amaçla boyut indirgeme ve indirgenmiş uzayda seçilen özelliklerin gözlenen uzaya yansıtılması fikri üzerinden yöntemler geliştirilmiştir.

1. BÖLÜM
GENEL BİLGİLER

20. yüzyıl sonu ve 21. yüzyıl başlarındaki döneme denk gelen bilimsel ve teknolojik ilerlemeler mikrobiyolojik kaynaklar için devrim niteliğindedir [8]. Mikrodizi ve kütle spektrometresi gibi kitle üretim teknolojileri kullanılarak birçok farklı mikrobiyal türleri içeren, çok sayıda biyolojik veri kümesi oluşturulmuştur [9]. Ayrıca, yeni nesil dizilemenin ortaya çıkışından bu yana bilim dünyası, daha önce görülmemiş miktarda genomik veri üretmiştir [10]. Dizileme metodolojilerindeki bu gelişmelerin yanında, DNA ve RNA dizilerinin yüksek çıktılı olarak dizilenebilmesiyle birlikte yeni biyoinformatik problemleri ortaya atılmış ve bu problemlerin çözümü üzerinde uğraşan yeni alt alanlar ortaya çıkmıştır [11]. Biyoinformatik bilimi, biyolojik sistemlerin yapısal özelliklerini inceleyen biyoloji biliminden yararlanması, biyolojik verileri elde etme amacıyla yaşam bilimleri ile ilişki kurması, verileri değerlendirme amacıyla istatistiksel yöntemlerden yararlanması ve büyük veri kümelerini anlamlandırma amacıyla bilgisayar algoritmalarını kullanmasıyla çok disiplinli bir özelliğe sahiptir [12].

Genellikle, biyolojik sistemlerde bulunan yapılara ve bunların dolaylı/dolaysız ürünlerinin niceliğine ait öznitelikler bize fenotip tespitine dair önemli bilgiler vermektedir. Fakat, yapılan çalışmalar incelendiğinde bazı özniteliklerin, fenotipe ait herhangi bir bilgi taşımadığı ve fenotip tespitinde yanıltıcı sonuçlar verebildiği görülmektedir. Bu yapılara ait özniteliklerin birçoğu sınıflandırma başarısı açısından olumsuz sonuçlara yol açmaktadır. Bu sebeple mevcut biyolojik ölçüm teknolojilerinde fenotip belirlemesi için öznitelik seçimi daha başarılı sonuçlar elde etmek için önemli bir adım olmaktadır.

Belli hastalıkların teşhisi için gerekli diagnostik süreçler doğası gereği (Örn: Tanı yönteminin invaziv operasyonlar gerektirmesi, hastalığa ait bilinen sinyallerin geç dönemlerde ortaya çıkması) erken tanıya olanak vermemektedir. Bununla birlikte, hastalığın patojenik süreçlerinin, klinik semptomlar görülmeden yıllar önce başlaması muhtemeldir [13]. Bu nedenle, ölçülebilir düzeyleri olan "Bir hastalık, fizyolojik

anormallik veya psikolojik durumun varlığını gösteren bir madde, fizyolojik özellik veya gen" olarak tanımlanan ve ölçülebilir seviyeleri klinik semptomlardan önce değişen biyolojik belirteçlere (biyobelirteçler) duyulan ihtiyaç büyük önem taşımaktadır [14]. Bu tür veri setlerinin kullanılabilirliği, hastalıkların teşhis ve tedavisi için çeşitli teknik ve ilaçların geliştirilmesine yardımcı olabilirken, aynı zamanda verilerin doğası ve oldukça yüksek hacmi, yararlı ve anlamlı bilgiler elde etmek için analizleri açısından veri madenciliğinin gücüne meydan okumaktadır.

Biyolojik/moleküler ölçümlerin öznitelik seçimindeki amaç en iyi alt öznitelik kümesini bulmaktır. En iyi alt özellik kümesinin belirlenmesi fenotiple ilişkili süreçleri (Örn: hastalıkların teşhis ve tedavi imkânını) iyileştirme potansiyeline sahiptir. Biyolojik/moleküler ölçümler normalde çok sayıda ilgisiz ya da benzer değerli öznitelikleri barındırmaktadır [15]. Öznitelik seçimindeki amaç, yüksek boyutlu veri kümesinde bulunan fenotiple ilişkili biyobelirteçleri belirlemektir. Böylece ayrım gücü yüksek olan öznitelikler belirlenebilir ve daha yüksek sınıflandırma doğruluğu ile veri kümesi sınıflandırılabilir. Diğer bir amaç ise, minimum sayıda öznitelik kullanarak en iyi sınıflandırma başarısını elde etmektir [16]. Buradaki biyolojik motivasyon, az sayıda biyobelirteç ile fenotip ayrıştırması yapılabilmesi durumunda bu biyobelirteçlerin doğrudan teşhis veya biyolojik süreç manipülasyonunda kullanılabilecek hedefler olarak belirlenebilecek olmasıdır. Örneğin, insan mikrobiyotasında bir hastalıkla ilişkili bakterilerin veri kümesi olarak kullanıldığı durumda öznitelik seçilimi ile az sayıda bakterinin biyobelirteç olarak belirlendiği durumu ele aldığımızı düşünelim. Tüm mikrobiyotanın hastalıkla toplu olarak ilişkilendirilmesi ancak hastalıkla ilişkili mikrobiyom bileşeni olduğu hipotezini destekleyecektir. Ancak öznitelik seçilimi ile ortaya konan az sayıdaki mikroorganizma bulgusunu sistem seviyesinden az sayıdaki doğrudan bileşen seviyesine taşıyabilmektedir. Bu da demektir ki, yalnızca bu mikroorganizmaların ölçümlerinden oluşan ufak çaplı tarama metodolojileri veya fiziksel süreçleri teşhis için pratikte uygulanabilecek teknolojilerin yolunu açabilir. Öte yandan hastalık patogenezi ve prognozu ile ilişkili olabilecek bileşen sayısı yine büyük bir ekosistemden tekil elemanlara kadar bu sayede indirilebilmektedir. Dolayısıyla ileri *in vivo* ve *in vitro* araştırmaların aday ajan sayısı azaltılmış olup bilimsel işleyiş süreçlerini zaman, maliyet ve iş gücü yönünden avantajlı hale getirmek mümkün olmaktadır.

Biyolojik belirteçlerin yüksek boyutlu verilerden elde edilmesindeki temel sorun, ilgili özellikler için sistematik bir araştırma ile çözülebilir. Bu çözüm veri setinin boyutsallığını

temsil eden, sınıflandırma doğruluğunu artıran ve hesaplama maliyetini azaltan küçük, ancak oldukça güvenilir ve ayırt edici bir alt kümeye indirgemektir [17]. Biyobelirteçleri bulmak için yaklaşımlardan biri, en alakalı özellik alt kümelerini seçmek için ÖS tekniklerinin kullanılmasıdır.

1.1. Literatür Çalışması

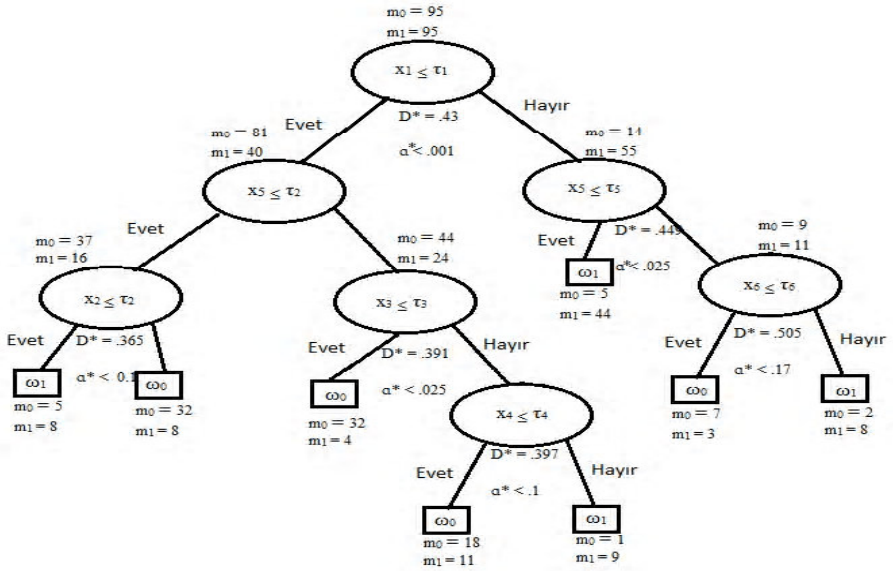
1.1.1. Öznitelik Seçimi

"Öznitelik seçimi, veri kümesini tanımlayan en az sayıda ilgili özellik kümesini bulma sorunu olarak tanımlanabilir [18]."

ÖS bir teknik kavram olarak incelendiğinde konuya odaklanan ilk çalışmaların 1980'li yıllarda başladığı görülmüştür.

Dr. E. M. Rours [19] çalışmasında özellik seçimi ve sınıflandırıcı tasarımının bütünleştirilmesi için bir çerçeve oluşturmayı amaçlamış ve özellik seçimi için Kolmogorov-Smirnov mesafesine ve Kolmogorov-Smirnov testine dayalı ardışık, hiyerarşik bir karar şeması kullanarak bunu başarmıştır. Bu karar kuralı, ikili karar ağacı yapısı için etkin bir otomatik prosedür sunmaktadır. Karar ağacı sınıflandırıcısı, mevcut eğitim verilerinden maksimum faydayı sağlar. Özellik seçimini sınıflandırıcının tasarımıyla birleştirerek, sınıflandırma için sadece en bilgilendirici özellikler korunur.

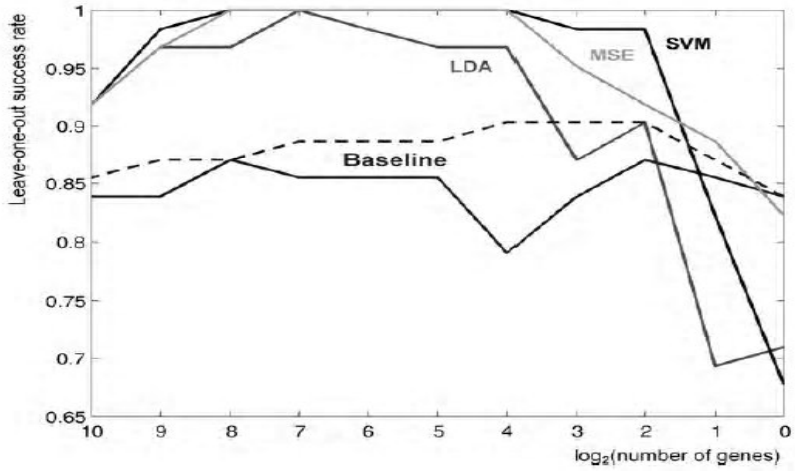
Karar ağacı oluşturma algoritması, iki sınıflı bir radar hedef tanıma problemine uygulanmıştır. Eğitim setinden türetilen ağaç Şekil 1.1 'de gösterilmiş, her bir düğümün alt popülasyon sayısı belirtilmiştir.



Şekil 1.1. Problemin Karar Ağacı Örneği

Eğitim veri seti ortalama % 79,5 oranında bir sınıflandırma oranına sahiptir. Bağımsız test seti ise % 73,5 oranında doğru bir sınıflandırma sağlamıştır. Bu sayılar, ağacın eğitim verilerine aşırı yüklenmediğini varsayabilmek için oldukça yakındır.

Isabelle Guyon vd. [20], kanserli hastalardan ve kontrol grubundan elde edilen DNA mikrodizi (microarray) teknolojisi ile elde edilmiş gen ifadesi verisinden küçük bir alt grup seçerek, hastalığın biyoinformatik teşhisi için uygun bir sınıflandırıcı keşfetmeyi amaçlamışlardır. Özyineli öznitelik eleme (İng: Recursive Feature Elimination (RFE)) temelli Destek Vektör Makinesi (İng: Support Vector Machine (SVM)) yöntemini kullanarak yeni bir gen seçimi yöntemi bulmuş ve bu teknik tarafından seçilen genlerin daha iyi sınıflandırma performansı sağladığını deneysel olarak göstermişlerdir.



Şekil 1.2. Özellik (gen) seçim yöntemlerinin karşılaştırılması (Kolon kanseri verileri)

Şekil 1.2'de görüldüğü gibi SVM (Support Vector Machine) metodu, Linear Discriminant Analysis (LDA), Mean Squared Error (MSE) ve Baseline metoduna göre daha iyi performans göstermiştir.

Yvan Saeys vd. [7], öznitelik seçimi tekniklerinin temel bir taksonomisini sunarak hem biyolojik uygulamalar hem de biyoinformatik uygulamaların birçoğunda bunların kullanımı, çeşitliliği ve potansiyelinin tartışıldığı bir tarama/değerlendirme makalesi hazırlamıştır.

Tablo 1.1. Özellik Seçimi Tekniklerinin Taksonomisi

Teknikler		**Avantajları**	**Dezavantajları**	**Örnekler**
Filtreleme	Tek Değişkenli	Ölçeklenebilir Hızlı Sınıflandırıcıdan bağımsız	Özellik bağımlılıklarını yoksayar Sınıflandırıcı ile etkileşimi yok sayar	x2 Öklid Mesafesi i-testi Bilgi kazancı Kazanç oranı
	Çok Değişkenli	Özellik bağımlılıkları olan modeller Sınıflandırıcıdan bağımsız Sarıcı yöntemlerden daha iyi hesaplama karmaşıklığı	Tek değişkenli tekniklerden daha yavaş Tek değişkenli tekniklerden daha az ölçeklenebilir Sınıflandırıcı ile etkileşimi yok sayar	Korelasyona dayalı özellik seçimi Markov blanket filtresi Hızlı korelasyona dayalı özellik seçimi
Sarmalayıcı	Deterministik	Basit Sınıflandırıcı ile etkileşime girer Özellik bağımlılıkları olan modeller Rasgele tekniklerden daha az hesaplama yoğunluğu	Aşırı öğrenme riski Yerel olarak optimum sonuçlara (açgözlü arama) takılmaya rastgele algoritmalardan daha eğilimli Sınıflandırıcıya bağlı seçim	Sıralı ileri seçim Sıralı geriye doğru eleme Artı q take-away r Beam arama
	Rastgele	Yerel optimal için daha az eğilimli Sınıflandırıcı ile etkileşime girer Özellik bağımlılıkları olan modeller	Hesaplamalı olarak yoğun Sınıflandırıcıya bağlı seçim Deterministik algoritmalardan daha fazla aşırı öğrenme riski	Benzetimli tavlama Rastgele tepe tırmanma Genetik algoritmalar Dağıtım algoritmalarının tahmini
Gömülü		Sınıflandırıcı ile etkileşime girer Özellik bağımlılıkları olan modeller Sarıcı yöntemlerinden daha iyi hesaplama karmaşıklığı	Sınıflandırıcıya bağlı seçim	Karar ağaçları Ağırlıklı Naive Bayes SVM'nin ağırlık vektörünü kullanarak özellik seçimi

Tablo 1.1'de her özellik seçimi türü için, ilgililerin hedeflerine ve kaynaklarına uygun bir teknik seçimi sağlayacak bilgiler sunulmuştur.

Tablo 1.2. Mikrodizi alanındaki her özellik seçimi tekniği için anahtar referanslar

Filtreleme			**Sarmalayıcı**	**Gömülü**
Tek Değişkenli		**Çok Değişkenli**	Sıralı Arama Genetik Algoritmalar Dağıtım algoritmalarının tahmini	Rastgele Orman Lojistik Ağırlıkları Regresyonu
Parametrik t-test ANOVA Bayesian Regresyon Gama	Model İçermeyen Wilcoxon derece toplamı BSS/WSS Derece Ürünleri Rastgele Permütasyonlar TNoM	Bivariate CFS MRMR USC Markov Blanket		

Tablo 1.3. Kütle spektrometresi alanındaki her özellik seçimi tekniği için anahtar referanslar

	Tek Değişkenli		**Çok Değişkenli**
	Parametrik	Model İçermeyen	
Filtreleme	t-test	Zirve olasılık kontrastı	CFS
	F-test	Kolmogorov-Smirnov test	Relied-F
Sarmalayıcı	Genetik Algoritmalar Doğa İlhamlı		
Gömülü	Rastgele Orman/Karar Ağaçları SVM ağırlık vektörü Sinir Ağı		

Tablo 1.2 ve Tablo 1.3'te, mikrodizi ve kütle spektrometresi alanında kullanılan minimum redundancy-maximum relevance (MRMR), corelation-based feature selection (CFS), uncorrelated shrunken centroid (USC), between-group to within-group sum of squares (BSS/WSS), Threshold Number of Misclassification (TNoM) gibi öznitelik seçimi algoritmaları verilmiştir.

Bir dizi tanınmış biyoinformatik uygulamada öznitelik seçimi teknikleri incelenirken, biyoinformatik alanındaki iki temel ortak sorun tespit edilmiştir: "Büyük girdi boyutları ve küçük örnek büyüklükleri." Bu problemlerle başa çıkabilmek için, biyoinformatik, makine öğrenmesi ve veri madenciliği konularında tasarlanan bir dizi ÖS tekniği bir sonraki bölümde temel hatlarıyla açıklanmaktadır.

1.1.1.1. Öznitelik Seçimi Teknikleri

Değerlendirme ölçütlerine dayalı olarak, öznitelik seçimi algoritmaları genel olarak üç kategoriye ayrılır [5,21]:

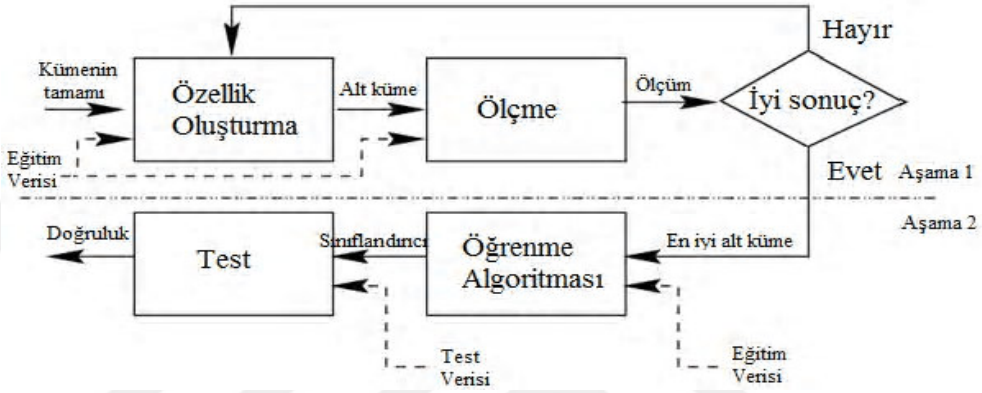
1) Filtreleme Teknikleri
2) Wrapper (Sarmalayıcı) Teknikler
3) Embedded (Gömülü) Teknikler

Filtreleme teknikleri, verilerin özniteliklerinin çeşitli istatistiki özelliklerini inceleyerek, özniteliklerin çözülmeye çalışılan probleme uygunluğunu değerlendirir. Çoğu durumda, bir öznitelik ilgililik skoru hesaplanır ve düşük puanlı öznitelikler çözüm kümesinden kaldırılır. Daha sonra, elenmiş olan öznitelik alt kümesi, sınıflandırma algoritmasına örüntüleri temsil eden girdi olarak sunulur. Bu yaklaşımda genellikle her öznitelik tekil olarak ele alınmakta, dolayısıyla doğrusal zamanda öznitelik seçilimi yapılabilmektedir. Filtreleme tekniklerinin avantajları, çok yüksek boyutlu veri kümelerine kolayca ölçeklendirilebilmeleri, hesaplama açısından basit ve hızlı olmaları ve sınıflandırma algoritmasından bağımsız çalışmalarıdır. Fakat sınıflandırıcı ile etkileşim kurulmaması, sınıflandırma performansının daha düşük olmasına sebep olabilir.

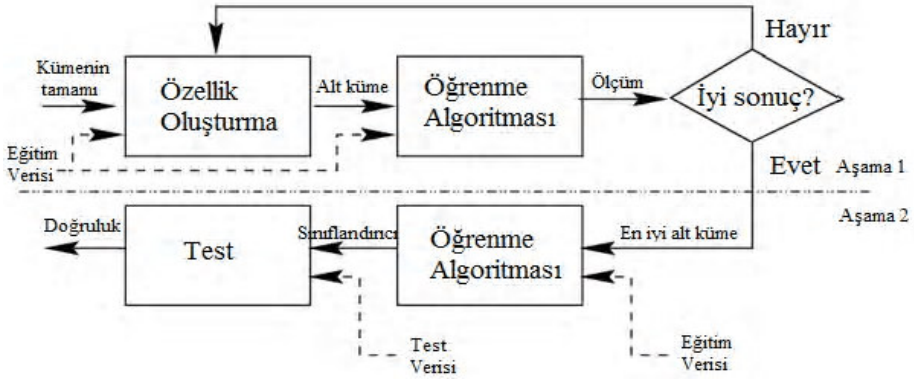
Filtreleme ve Sarmalayıcı yaklaşımların temel farkları, sarmalayıcı yaklaşımların, alt küme değerlendirme aşamasında bir sınıflandırma / öğrenme algoritmasına bağımlı olarak değerlendirme yapmasıdır. Filtreleme tekniklerini kullanan algoritmalara; Öklid Mesafesi, i-testi, korelasyona dayalı özellik seçimi (Correlation-based feature selection (CFS)), Markov blanket filtresi, Relief, hızlı korelasyona dayalı özellik seçimi (Fast correlation-based feature selection (FCBF)) ve INTERACT örnek gösterilebilir [7, 22].

Sarmalayıcı yaklaşımda çeşitli öznitelik altkümeleri oluşturulur ve bu altkümeyi içeren çözümün performansı öğrenme algoritması kriterlerine göre değerlendirilir. Olası öznitelik alt grupları uzayında bir arama prosedürü tanımlanır. Tüm öznitelik alt gruplarının alanını aramak için, bir arama algoritması, sınıflandırma modelinin etrafına "sarılır". Bununla birlikte, öznitelik alt kümeleri uzayı, öznitelik sayısıyla üstel olarak büyüdüğünden, tam kapsamlı arama yaklaşımı öznitelik kümelerinin eleman sayısı artışıyla hızla ölçeklenebilir olmaktan uzaklaşmakta ve polinom karmaşıklıkta bir çözüm sunamamaktadır. Bu nedenle, arama için en uygun alt kümeyi yönlendirmek için sezgisel arama yöntemleri kullanılır. Bu arama yöntemleri iki sınıfa ayrılabilir: deterministik ve rasgele arama algoritmaları.

Sarmalayıcı yaklaşım, sınıflandırıcı ile etkileşimde olduğundan, sınıflandırma performansının ampritik olarak genellikle yüksek olduğu gözlenmektedir. Buna karşılık, sınıflandırıcının oluşturulması için gerekli hesaplama maliyetini yükseltir. Sarmalayıcı tekniklere örnek olarak: Genetik algoritmalar, Sıralı ileri seçim (Sequential forward selection (SFS)), Sıralı geriye doğru eleme (Sequential backward elimination (SBE)) ve Greedy forward search verilebilir [7].



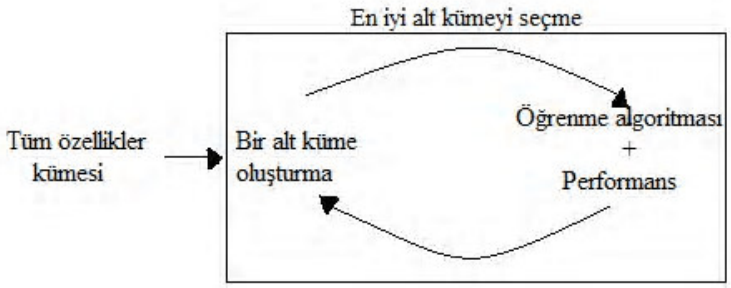
Şekil 1.3. Filtreleme Yöntemleri [23]



Şekil 1.4. Sarmalayıcı Yöntemler [23]

Gömülü tekniklerde ise, özniteliklerin optimal bir alt kümesini bulmak için, arama, sınıflayıcı yapısına yerleştirilir ve öznitelik alt kümeleri ile hipotezlerin kombine alanındaki bir arama olarak görülebilir. Sarıcı yaklaşımlardaki gibi, gömülü yaklaşımlar da belirli bir öğrenme algoritmasına özgüdür. Gömülü yöntemler, sınıflandırma modeliyle etkileşimi içerirken, aynı zamanda sarmalayıcı yöntemlerden çok daha az hesaplamaya tabi tutulması avantajına sahiptirler.

Gömülü yöntemler de öğrenme algoritmasının veya modelinin performansını optimize etmek amacıyla kullanıldıkları için, sarmalayıcı yöntemlere oldukça benzerdir. Sarmalayıcı yöntemlerle aralarındaki fark, öğrenme sırasında içsel bir model oluşturma yönteminin kullanılmasıdır. Yaygın olarak kullanılan gömülü yöntemler: Rastgele orman, karar ağaçları ve sinir ağlarıdır [7].



Şekil 1.5. Gömülü Yöntemler

1.1.1.2. Öznitelik Seçiminde Karşılaşılan Problemler

Pek çok örüntü tanıma tekniği, büyük oranda sınıflandırma problemiyle ilişkisiz özelliklerle baş etmek zorunda kalınca, bu örüntü tanıma tekniklerini ÖS teknikleriyle birleştirmek birçok uygulamada zorunluluk haline gelmiştir [24-25].

Modeli seçmek için kullanılan ölçütler, bir modelin uygunluğunu değerlendirmek için kullanılan ölçütler ile aynı değildir. Örneğin, bir model, eğitim verisinin performansını en üst düzeye çıkardığı için seçilebilir, fakat uygunluğu, görünmeyen verileri işleme becerisi ile belirlenebilir. Bir model eğilime göre genelleme yapmak için, eğitim verilerini "öğrenme" yerine, "ezberlemeye" başladığında aşırı öğrenme (overfitting) oluşur [26].

Diğer bir problem ise Richard E. Bellman tarafından ortaya atılan bir tabir olan boyutsallığın lanetidir (Curse of dimensionality). Boyutlar arttıkça, örneklemin dağıldığı uzayın determinantı çarpımsal olarak artmakta ve örneklem noktaları arasında metrik olarak büyük uzaklıklar oluşarak noktalar kümesi seyrekleşmektedir. Bu seyreklik, konvansiyonel istatistik güdümlü olan herhangi bir yöntem için sorun teşkil eder. İstatistiksel olarak sağlam ve güvenilir bir sonuç elde etmek için, veri miktarının boyusallıkla birlikte katlanarak büyümesi gerekir. Verileri organize etmek ve araştırmak çoğunlukla benzer özelliklere sahip grupları tespit etmeye dayanır. Dolayısıyla birçok öznitelik seçilim algoritması korelasyona sahip kümeler içerisinden temsilci seçimi prensibinden hareketle fonksiyon göstermektedir. Ancak, yüksek boyutlu verilerde, tüm

nesneler, birçok yönden birbiriyle benzeşmez şekilde gözlemlenir, ki bu söz konusu yöntemlerin seçilimde tasarlandığı şekilde çalışmaması ile sonuçlanabilir.

Sabit sayıda eğitim örneği ile, tahmin gücü, boyutluluk arttıkça azalır. Bu genel eğilim Hughes fenomeni olarak adlandırılır [21]. Tüm bunlardan hareketle boyutsallık lanetinin önüne geçmek, aşırı öğrenmeyi azaltmak ve model performansını artırmak, yani sınıflandırmada tahmin performansını artırmak, maliyetten tasarruf sağlama, modellerin, araştırmacı ya da kullanıcı tarafından yorumlanmasını kolaylaştırmak için sadeleştirilmesi ve verileri üreten süreçler hakkında derinlemesine bir kavrayış elde etmek amacı ile ÖS uygulanır [27].

Bahsi geçen yüksek boyutlu ve küçük örnekleme sahip veri kümeleri pek çok biyoinformatik problemi gibi mikrobiyota/mikrobiyom sınıflandırması problemlerinde de sıklıkla karşılaşılan bir problemdir. Buna karşılık bu uygulama sınıfına özgü geliştirilmiş kapsamlı bir yaklaşım sınıfı bulunmamaktadır. Söz konusu problemlerde yaygın uygulaması bulunan genel amaçlı öznitelik seçilim algoritmalarının yanında, bu problem sınıfına özgü tasarlanmış öznitelik seçimi yaklaşımlarının geliştirilmesi önemli bir araştırma sorusuna denk gelmektedir. Öte yandan, bu doğrultuda geliştirilecek olan potansiyel bir seçilim algoritması, doğrudan biyobelirteç keşfi problemine denk düşeceğinden yaşam bilimleri açısından büyük motivasyon taşımaktadır.

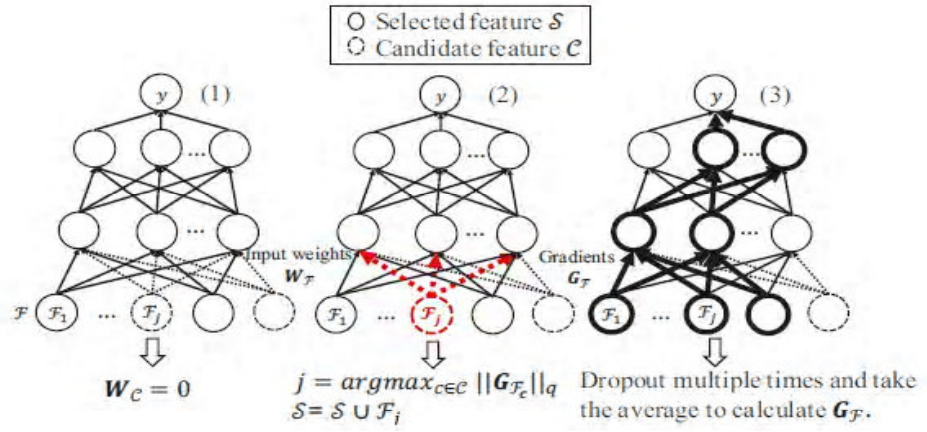
Öznitelik sayısının yüksek ve örnek sayısının nispeten düşük olduğu sınıflandırma problemlerinde ÖS, hesaplama karmaşıklığı yüksek ve optimal altı çözümlerin elde edilmesine açık bir kombinatorik problem haline dönüşmektedir. Bu tezde geliştirilen yöntemler ile seçilimin esasen düşük boyutlu uzayda yapılması ile söz konusu karmaşıklık ve optimumdan uzak çözümler elde etme riskleri azaltılmaya çalışılmıştır. Bu sayede dayanıklı ve hızlı öznitelik seçimi yaklaşımları ortaya konarak probleme yeni bir çözüm önerisi getirilmesi amaçlanmıştır.

1.1.2. Yapay Sinir Ağlarına Dayalı Öznitelik Seçimi

Profil oluşturan omik verilerin analizi, önemli biyolojik keşiflere yol açma potansiyeline sahiptir. Bu alanlara öncü olarak mikrodizileme veya RNA/DNA sekans verilerinin gen ifadelerinin sınıflandırılması ve kümelenmesi çokça çalışılmış konulardır. Farklı olarak ifade edilen genlerin tanımlanması konusundaki çalışmaların çoğu, en önemli değişikliklere odaklanmıştır ve verilerde daha ince örüntülerin tanınmasına izin vermeyebilir [28]. Bu verileri analiz etmek, gen düzenleyici hedeflerin, hastalık teşhisinin

ve ilaç geliştirmenin keşfedilmesi için muazzam bir potansiyel bulundurmaktadır. Yapay sinir ağları (YSA) çeşitli alanlarda en son teknoloji sonuçların elde edilmesine olanak sunarken, yorumlanma karmaşıklığından dolayı biyoloji ve sağlık hizmetleri gibi hipotez odaklı alanlarda göz korkutucu bulunmaktadır. Bu sebeple, kısıtlı kaynak ortamında, makul bütçede yüksek doğrulukta performansa yol açan daha çok bilgilendirici özelliği bulunan testlerin tasarlanması kritik öneme sahiptir [29-32]. YSA birçok farklı alanda etkin bir şekilde kullanılmıştır.

Liu vd. [33], biyoinformatikte genetik verileri kullanarak fenotip tahmin problemlerindeki gibi yüksek boyutlu, düşük örnek sayılı (YBDÖS) verilerle karşılaşıldığında, derin sinir ağları (DSA) benzeri yüksek kapasiteli yaklaşımların aşırı öğrenme ve yüksek varyanslı gradyanlardan dolayı sorun yaşayabileceğini öne sürmüştür. Çalışmalarında, Derin Sinirsel Kovalama (DSK) adlı YBDÖS verisine uyarlanmış bir DSA modeli önerilmiştir. DSK, aşırı öğrenmenin hafifletilmesi için yüksek boyutlu özelliklerin bir alt kümesini seçer ve düşük varyanslı gradyanları hesaplamak için birden fazla çıkış noktası üzerinden ortalama alır. YBDÖS verisine uygulanan ilk DSA yöntemi olarak DSK, yüksek boyutluluğa karşı direnç, az sayıda örnekten öğrenebilme yeteneği, öznitelik seçimindeki kararlılık ve uçtan uca eğitimin avantajlarından yararlanır. DSK'nin bu avantajları hem sentetik hem de gerçek biyolojik veri kümeleri üzerinde deneyler yapılarak gösterilmiştir.



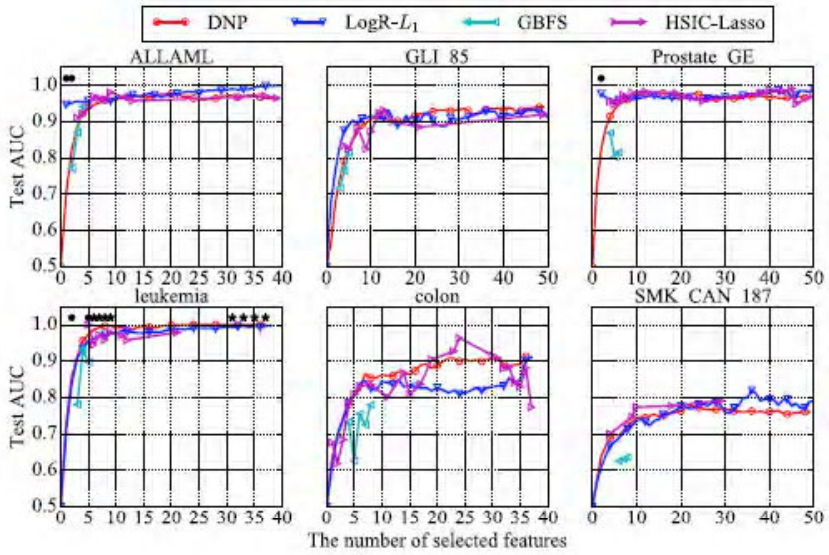
Şekil 1.6. Uçtan uca özellik seçilimi sağlayan yapay sinir ağı modeli

Şekil 1.6'da 1 numaralı figür, seçilen öznitelikler ve ilgili alt ağı, 2.şekil tek bir öznitelik seçimini, 3. Şekil ise birden fazla çıkış noktasıyla daha düşük varyanslı gradyanları hesaplamayı göstermektedir.

Tablo 1.4. Biyolojik Veri Setleri için İstatistikler

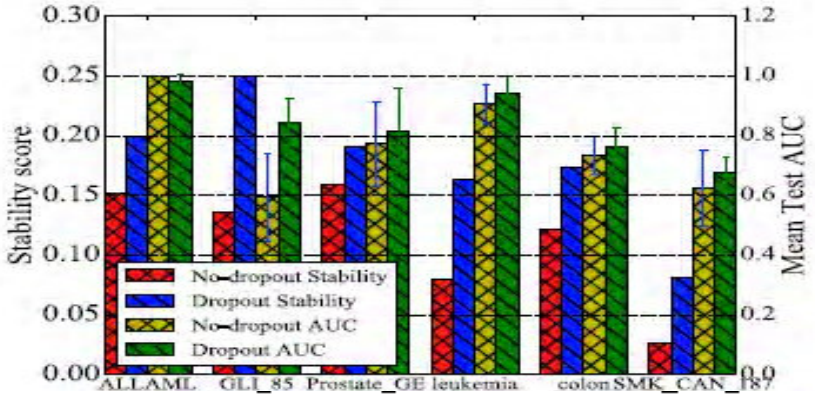
Veri	Kolon	Prostat	Lösemi
Örnek Sayısı	62	102	72
Boyut	2000	5966	7070
Veri	ALLAML	SMK_CAN_87	GLI_85
Örnek Sayısı	72	187	85
Boyut	7129	19993	22283

DSK'nin gerçek veri kümeleri üzerindeki performansını araştırmak için, hepsi YBDÖS probleminden muzdarip olan altı adet halka açık biyolojik veri kümesi kullanılmıştır. Bu veri setlerinin istatistikleri Tablo 1.4'te gösterilmektedir. Şekil 1.7'de, seçilen özelliklerin sayısına göre EAA (Eğri Altında Kalan Alan, İng: Area Under The Curve (AUC)) ortalama puanlarını verilmiştir. DSK, her yinelemede tek bir özellik seçer. Sonuç olarak, Şekil 1.7'de verilen DSK'nin performans eğrisi, EAA puanlarının iterasyonların sayısına göre nasıl değiştiğini gösterir. Altı veri kümesinde, DSK'nin EAA puanlarını test etmek, 10 iterasyondan daha az sürer.



Şekil 1.7. AUC ortalama puanları

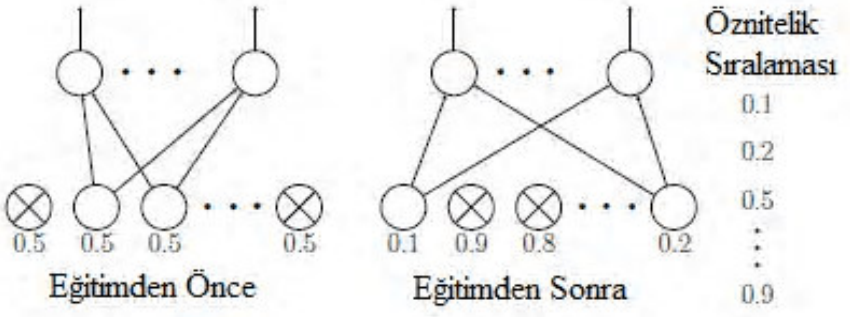
Farklı sayılarda öznitelikleri olan sentetik veri kümeleri üzerinde sınıflandırma ve öznitelik seçiminin performansı incelendiğinde DSK'nin genel olarak diğer yöntemlere göre (LogR, GBFS, HSIC-Lasso) daha iyi performans gösterdiği söylenebilir.



Şekil 1.8. Çoklu Çıkış Noktaları Olan ve Olmayan Yöntemler Arasında Kararlılık Kıyaslaması

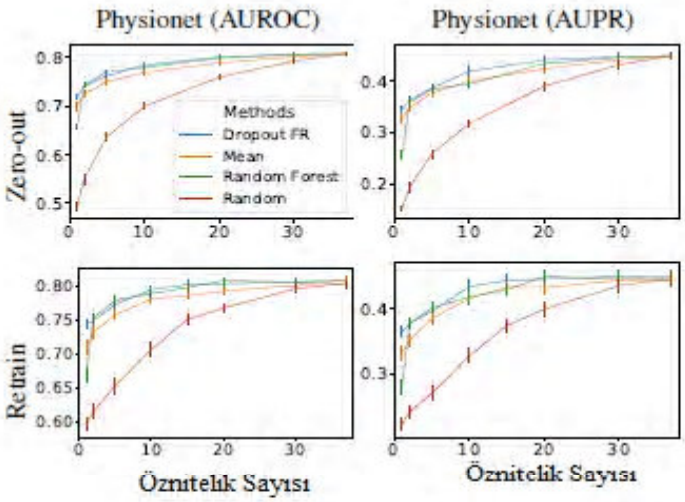
Şekil 1.8'de gösterildiği gibi, birden fazla çıkış noktası (dropout) olan DSK, altı veri kümesinin tamamı üzerinde diğerlerinden açıkça daha kararlıdır.

Chang vd. [32], derin öğrenme bazlı yeni bir genel özellik sıralama yöntemi önererek bir boşluğu kapatmayı hedeflemiştir. Bu basit ve etkili yöntemin hem statik hem de zaman serisi senaryolarında, iki farklı simülasyonda ve beş çok farklı veri kümesinde, sınıflandırmadan regresyona kadar birçok metotta, klasik yöntemlere göre eşit ya da daha iyi performans gösterdiği gözlemlenmiştir.



Şekil 1.9. Dropout Öznitelik Sıralama Şeması

Şekil 1.9'da eğitimden önce (Sol), her öznitelik için dropout oranı 0,5 olarak başlatılır. Eğitimden sonra (Sağ), her özellik farklı bir dropout oranı alır ve tüm öznitelikler bu dropout oranının büyüklüğüne göre sıralanır – düşük orandan yükseğe doğru.



Şekil 1.10. AUROC ve AUPR ile Physionet veri kümelerindeki yöntemlerin karşılaştırılması

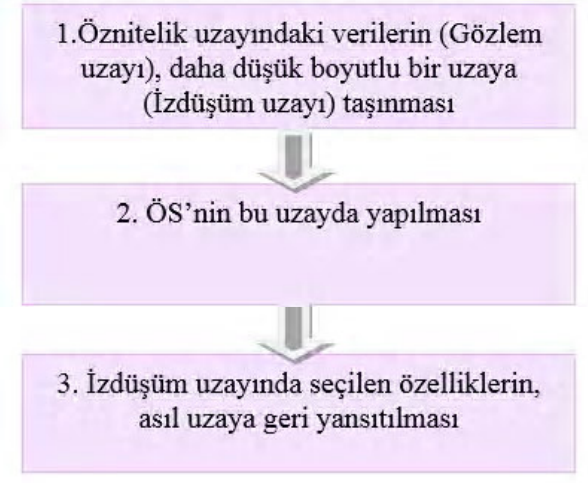
Şekil 1.10'da AUROC (Area under the ROC curve) ve AUPR (Area under the PR curve) ile Physionet veri kümelerindeki yöntemlerin karşılaştırılması gösterilmiştir. Etkileşimli ve etkileşimsiz 2 veri seti; tüm özellikler (40), sadece bilgilendirici öznitelik (20) ve en bilgilendirici 5 öznitelik için ayrı ayrı karşılaştırılmıştır. Etkileşimsiz veri kümesinde, Elastic Net ve LASSO dışındaki tüm yöntemlerin (Marginal, RF, Dropout ÖS, Mean, Shuffle, Derin ÖS) özellikleri karşılaştırırken mükemmel performans gösterdiği görülmektedir. Bu, yöntemlerin gürültüyü gerçek özelliklerden ayırt edemediğini, ancak bilgilendirici özelliklerin güçlü yönlerini sıralayabildiğini göstermektedir. Etkileşimli veri kümesinde ise Elastic Net, LASSO ve Marginal yönteminin çok daha kötü performans gösterdiğini, bu basit lineer katmanı ve tek özellikli istatistiksel testlerin ikinci dereceden etkileşim efektlerini yakalayamadığı çıkarılabilir.

Genel olarak, Dropout ÖS metodunun tekrarlayan sinir ağı mimarisinde zaman serisi veri kümelerindeki özellik önemlerini yakalayarak iyi performans gösterdiği görülmektedir.

2. BÖLÜM
GEREÇ VE YÖNTEM

Yapılan araştırmalar ve literatür taraması sonucunda, öznitelik seçimi konusunda kullanılan yöntemlerin büyük çoğunluğunun gözlem uzayında operasyon gerçekleştirdiği görülmektedir. Yüksek boyutlu problemlerde bu yaklaşım çok büyük bir uzayın içerisinde arama yapılacağı anlamı taşımaktadır. Bu durum sonucunda aramanın ne denli efektif olduğu tam olarak ortaya çıkarılamamakla birlikte birçok problemde boyutun yüksekliği aynı zamanda beraberinde büyük bir hesaplama yükü getirmekte, hatta çok büyük boyutlu problemler için ölçeklenebilir olmaktan çıkabilmektedir. Öte yandan, gözlem uzayı yüksek boyutlu olsa dahi, esasen örnek profillerinin bu uzayın küçük boyutlu bir alt uzayında dağılım göstermesi ve bir gizli uzay içerisinde yer alması oldukça olasıdır. Önerilen yöntemde, seçilimin öncelikle bu alt uzayda yapılması, daha sonra bu uzayın seçilmiş alt kümesini en iyi şekilde yansıtacak gözlenebilir öznitelik kümesini ortaya koyabilecek hiyerarşik bir yaklaşım ortaya konmaktadır.

Geliştirilen yöntemin aşamaları genel olarak aşağıdaki şekilde gibi özetlenmiştir:



Şekil 2.1. Geliştirilen Yöntemin Aşamaları

Şekil 2.1'deki geliştirilen yaklaşımın aşamaları sırasıyla aşağıda açıklanmaktadır.

1) Gözlem uzayındaki verilerin izdüşüm uzayına taşınması

Önerilen yaklaşımın genel amaçlı öznitelik seçimi algoritmalarından temel farkı, sadece sınıflandırma objektif fonksiyonunun en iyilenmesine yönelik kombinatoryal bir optimizasyonu değil, aynı zamanda veri kümesinin genel karaktertistiğini de en iyi şekilde yansıtan özniteliklerin seçilebilmesidir. Bu düzenlileştirme yaklaşımının ardında bulunan ana motivasyon, mikrobiyom üzerinden biyobelirteç seçilimi ve benzeri biyoinformatik uygulamalarda komünitenin fenotipe dayalı, sistem seviyesinde bir varyasyon göstermesi ve seçilen biyobelirtecin bu global varyasyonla ilişkilendirilebiliyor olmasıdır. Biyolojik olarak bu şekilde seçilen biyobelirteçlerin sistem dinamiklerinde kilit rol oynayan elemanlar olma olasılıkları daha yüksek olacaktır. Bu sebeple ilk aşamada gözlenebilir özniteliklerin, yani gözlem uzayının daha düşük boyutlu bir uzaya izdüşümü gerçekleştirilerek gizli öznitelik çıkarımı yapılması sağlanmaktadır. Bu aşama doğrudan veriyi temsil edebilen az sayıdaki içsel özniteliğin belirlendiği kısım olarak düşünülebilir. Bir veri kümesini en az farklılıkla geri döndürülebilecek şekilde daha az sayıda değişken ile dönüştürmenin kabul gören belli başlı yöntemleri vardır. Bunlar içerisinde TDA ve NOMA en yaygın olarak kullanılan iki doğrusal yöntemdir. İki yaklaşım da daha düşük boyutlu bir uzaya en az ortalama karesel hata ile geri döndürülebilir bir izdüşüm yapabilmektedir. Bu iki yaklaşım da ilk aşamadaki boyut indirgeme ve gizli özelliklerin ortaya konulabilmesi için geçerli olabilecekken, yapılan deneyler sonucunda NOMA yaklaşımının bu indirgeme için daha anlamlı olduğu görülmüştür. Bu sebeple ilk aşamadaki indirgeme NOMA ile gerçekleştirilmektedir.

2) İzdüşüm uzayında öznitelik seçilimi yapılması

NOMA indirgemesi sonucunda artık her bir örneği temsil edebilen bir gizli öznitelik oluşmaktadır. Bu gizli özniteliklerin her biri eldeki veri setinin farklı bir özelliğini temsil ediyorken bu özelliklerin tamamı, söz konusu sınıflandırma kategorileri ile ilişkili olmayabilir. Bu amaçla, öncelikle gizli uzayda bir öznitelik seçilimi yapılması ikinci aşamada uygulanan yaklaşımdır. Genel amaçlı olarak kullanılan birçok öznitelik seçimi algoritması bu iş için uygun iken, yapılan denemelerde kullanılan yöntemler içerisinde en tutarlı sonuçları ÖÖE sağladığı için bu yaklaşım varsayılan yöntem olarak önerilen yöntemin gizli öznitelik seçimi yöntemi olarak kabul edilmiştir. Bu aşamadaki seçilim, hem sınıflandırmaya yönelik özniteliklerin ortaya konulabilmesi, hem de bir sonraki

aşamadaki gözlenebilir öznitelik elemesindeki hataların azaltılması açısından önemlidir. Eğer bu aşamada seçilen öznitelikler sınıflandırmayı yeterli doğrulukta yapabiliyorlarsa, gözlenen uzaydan bu uzaya gerçekleştirilecek haritalama sonucunda da doğru sınıflandırma yapabilen düşük sayıda öznitelik ortaya çıkmış olacaktır.

3) İzdüşüm uzayında seçilen özniteliklerin gözlenen uzaya geri yansıtılması

Bu aşamadaki amaç, belirlenen (sınıflandırma ile ilişkili bulunan) gizli öznitelikleri bir haritalama/regresyon ile temsil edebilecek, gözlenebilen öznitelik seçiliminin yapılmasıdır. Bu amaçla, seçilen gizli öznitelik kümesi üzerinde en çok bileşene sahip olup, aynı zamanda seçilmeyen öznitelikler üzerindeki bileşeni en az olan gözlenebilen özniteliklerin seçilebilmesi sağlanmaktadır. Sütunları her bir örnek vektörünü temsil edecek şekilde d boyutlu n örnek için dxn'lik X matrisinin bir veri kümesini temsil ettiği varsayılırsa NOMA işlemi sonrasında $X \approx HW$ şeklinde m adet öznitelik vektörünü içeren dxm boyutunda H taban matrisi ve n adet m boyutlu kodlama vektörünü içeren mxn boyutunda W kodlama matrisine ayrıştığını varsayalım (m<d). İkinci adımda m gizli öznitelik içerisinden k adet ($k \leqslant m$) sınıflandırma ile ilişkili öznitelik indeks kümesi $S = \{s_1, s_2, .., s_k\}$, seçilmeyen özellikler ise $S' = \{s'_1, s'_2, .., s'_{m-k}\}$ indeks kümesi ile belirlensin. Bu durumda gözlenebilen öznitelikler içerisinde seçilmiş tabanlardaki içeriği, seçilmeyenlerden fazla olan öznitelik indeks skorlaması:

$$O(i) = \sum_{s_j \in S} H_{i,s_j}^2 - \sum_{s'_j \in S'} H_{i,s'_j}^2 , i \in \{1,2,3,..d\} \quad (2.1)$$

şeklinde karesel toplam farkı şeklinde alınabilir. Bu skorlama ile tüm gözlenebilen özniteliklere ait skorlar elde edildikten sonra en yüksek skoru alan öznitelikler seçilim sonucunu vermektedir. Son aşamada yeni bir sınıflandırma modeli, yalnızca bu öznitelikler kullanılarak işleme koyulur.

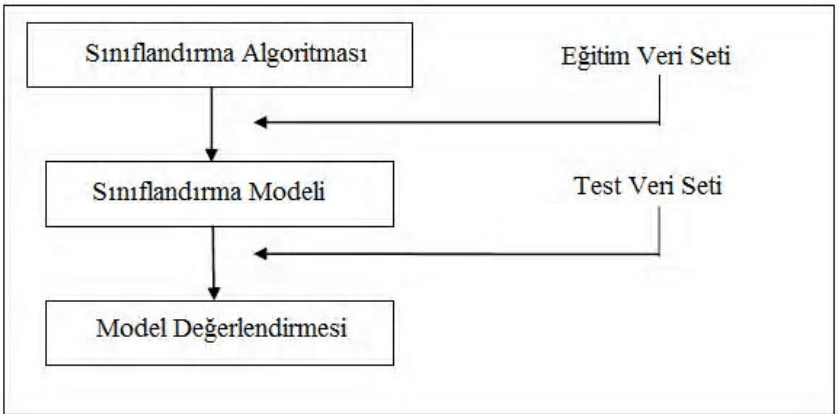
Bu temel yaklaşım ile verilerin gözlem uzayından daha düşük boyutlu izdüşüm uzayına taşınmasıyla, problemin gerçek boyutunun ortaya çıkarılması ve bu sayede boyutsallığın lanetinden kurtulmak amaçlanmıştır. Gözlem uzayında çok fazla veri bulunabilir fakat bu özelliklerin birçoğu birbiriyle çok ilişkili/benzerdir. Öznitelik seçilimi doğrudan bu öznitelik uzayında yapılırsa, çok fazla yerel çözümden bir tanesine ulaşılmış olur. Bu çözümün nitelikli bir çözüm olma olasılığı uzayın büyüklüğü ile azalmaktadır. İzdüşüm uzayında yapılan ÖS ise, aşırı öğrenmeye daha dayanıklı olması beklenmektedir. Bu yüzden izdüşüm uzayında oluşturulmuş aşırı öğrenmeye dayanıklı öznitelik seçiminin

gözlem uzayına yansıtılması ve bu uzayda özellik seçiminin yeniden oluşturulması hedeflenmiştir.

2.1. Kullanılan Yöntemler

2.1.1. Sınıflandırma Yöntemleri

Sınıflandırma problemleri, kategorik regresyon problemlerini içerisine alan problemlerdir. Bir örneklem kümesi içerisinde yer alan elemanları sonlu kardinalitesi olan başka bir kümeye içsel (injektif) olarak eşleyen bir fonksiyon olduğu varsayımından yola çıkarak, her bir örneğin fonksiyon değerinin denk geldiği elemana sınıf ismi verilmektedir. Bu tanıma göre, pratikte kategorilere ayrılabilen her türlü veri kümesinin kategorilerinin tahmin edilebilmesi yukarda bahsi geçen fonksiyonun ortaya çıkarılmasına denk gelmektedir. Bu problemin çözümü, kategorik bir regresyon çözümünü gerektirmektedir. Yapay öğrenme alanında bu çözüm, örneklem kümesinin fonksiyonun öğrenilmesi için kullanıldığı denetimli bir öğrenme yaklaşımına karşılık gelmektedir. Dolayısıyla sınıflandırma denetimli öğrenmenin bir türüdür. Sınıflandırıcının bir eğitim verisi ile eğitilmesi ve bilinmeyen bir veri üzerinde sınıflandırma için kullanılmadan önce test verileri ile değerlendirilmesi kabul gören mevcut konvansiyondur. Denetimli öğrenme süreci aşağıdaki şekilde gösterilebilir:



Şekil 2.2. Sınıflandırma için denetimli öğrenmenin genel süreci

Şekil 2.2' de görüldüğü gibi, denetimli öğrenme sürecinde, bir veri kümesi, eğitim seti ve test seti olarak ayrılır. Eğitim veri seti, sınıflandırıcıyı eğitmek ve bir sınıflandırma modeli oluşturmak için kullanılırken, test veri kümesi sınıflandırma modelini tahmin doğruluğu açısından değerlendirmek için kullanılır [34].

Sınıflandırma, kategori üyeliği bilinen gözlemleri içeren bir eğitim seti temelinde, yeni bir gözlemin, bir grup kategoriden hangisine ait olduğunu tespit etme problemidir. Birçok gerçek dünya problemi sınıflandırma problemi olarak modellenebilir. Örneğin belirli bir e-postanın "spam" veya "spam olmayan" sınıflara atanması, gelecek haber kategorilerinin (Örn: "Spor" ve "Eğlence") otomatik olarak atanması veya bir hastanın gözlemlenen özellikleri (cinsiyet, tansiyon, belirli semptomların varlığı veya yokluğu, vb.) ile o hastaya tanı koyabilmek gibi [35].

Eğitim modelinin performansının bir ölçüsü, tahmininin doğruluğudur. Doğruluk değeri en yaygın kullanılan değerlendirme ölçütlerinden biridir. Bu çalışmada da değerlendirme ölçütü olarak sınıflandırma doğruluğu kullanılmıştır. Sınıflandırma doğruluğunun başarısı doğru etiketlenmiş veri sayısının, etiketlenen toplam verilere oranı ile ölçülür. Tahminlerde hata oranının daha düşük olması, daha iyi ve daha güvenilir bir model olduğunu gösterir. Sınıflandırma doğruluğu ve hata oranları aşağıda gösterilen denklem (2.1) ve (2.2) kullanılarak hesaplanır.

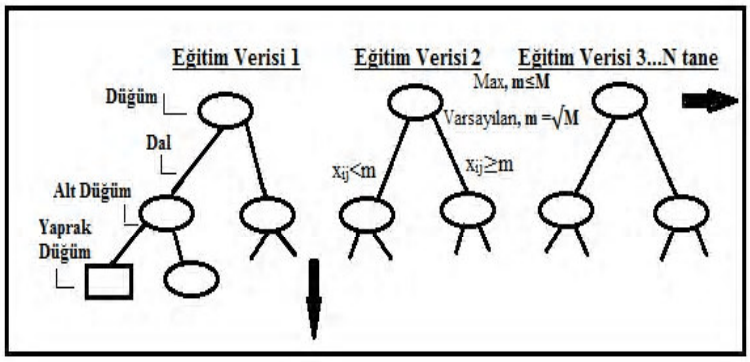
$$Doğruluk\ oranı = \frac{Doğru\ tahminlerin\ sayısı}{Toplam\ tahmin\ sayısı} \tag{2.2}$$

$$Hata\ oranı = \frac{Yanlış\ tahminlerin\ sayısı}{Toplam\ tahmin\ sayısı} \tag{2.3}$$

Tez çalışmasında kullanılan sınıflandırma algoritmaları: Rastgele Orman Algoritması, Destek Vektör Makineleri Algoritması, Lojistik Regresyon Algoritması, En Yakın Komşu Algoritması, Karar Ağaçları Algoritması ve Gaussian Naive Bayes Algoritmasıdır. Bu algoritmaların özellikleri aşağıda detaylandırılmış, algoritmalar uygulanırken Python programlama dili kullanılmış ve bu dil içerisindeki sklearn kütüphanesinden faydalanılmıştır.

2.1.1.1. Rastgele Orman Algoritması

Günümüzde Rastgele Orman (RO) algoritması, sınıflandırmada çok iyi performans sergilediği için toplu öğrenme yöntemlerine göre sıklıkla tercih edilmektedir. Son zamanlarda geliştirilen RO sınıflandırıcısı hem hızlı olması hem de yüksek doğruluk sağlaması yönünden diğer sınıflandırma yöntemlerine göre avantaj sağlamaktadır [36].



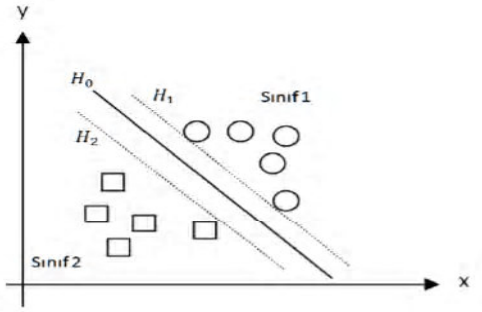
Şekil 2.3. RO yöntemine ait ağaç yapısı

RO sınıflandırıcısı ile bir ağaç üretmek için kullanıcı tarafından tanımlanması gereken 2 parametre vardır. Bu parametreler, en iyi bölünmeyi belirlemek için her bir düğüm için kullanılan değişkenlerin sayısı m ve geliştirilecek ağaçların sayısı N'dir [37]. Kullanıcı tarafından başlangıç m değeri rastgele seçilir sonraki m'ler genelleştirilmiş hatalara göre artırılır ya da azaltılır, bu şekilde en uygun m bulunur ve böylece deneysel olarak sınıflandırma duyarlığı artar, hata oranı azalır. m değişken değeri seçilirken, M değerinin (toplam değişken sayısı) karekköküne eşit olarak alınması genellikle optimum sonucu vermektedir [38].

Ağaç gelişiminden sonraki aşama, girdi verilerinin sınıflandırılması işlemidir. Bu işlemde, girdi verileri ormandaki her bir ağaca yerleştirilir. Belirlenen ağaçlar arasında oylama yapılır ve en çok oyu alan ağaç sınıfa atanır [39].

2.1.1.2. Destek Vektör Makineleri Algoritması

1995 yılında farklı sınıfların nesnelerine ait kümeler arasında, öznitelik uzayında en iyi bir hiper düzlemi oluşturan bir sınıflandırıcı geliştirilmiştir [40]. Bu en iyi hiper düzlemi oluşturmak için, hata fonksiyonunu minimize etmede iteratif bir eğitim algoritması kullanılmıştır [41].



Şekil 2.4. İki boyutlu uzayda doğrusal ayrılabilen verilerin görünümü [42]

Doğrusal bir Destek Vektör Makinesi (DVM) için formül:

$$u = \bar{w} \cdot \bar{x} - b \tag{2.4}$$

Bu denklemde w, hiper düzlemdeki normal vektördür ve x, giriş vektörüdür. En yakın noktalar u = ± 1 düzlemlerindedir. d mesafesi:

$$d = \frac{1}{\|w\|_2} \tag{2.5}$$

Maksimum d uzaklığı, optimizasyon problemi kullanılarak ifade edilebilir.

$$\min_{\bar{w},b} \frac{1}{2} \|\bar{w}\|^2 y_i(\bar{w} \cdot \bar{x}_i - b) \geq 1 \tag{2.6}$$

$\bar{x}_i$ ve y_i eğitilmiş örnekler için DVM 'nin doğru çıktısıdır. y_i, pozitif örnekler için +1, negatif örnekler için -1'dir.

Standart DVM iki sınıflı problemler için uygundur. Bunun üstesinden gelmek için bire karşı diğer sınıflar ya da çok sınıflı destek vektör makinesi yapıları kullanılabilir.

2.1.1.3. Lojistik Regresyon Algoritması

Lojistik regresyon (LR) modelinde yukarıdaki üç algoritmadan farklı olarak P (y | x) ifadesini açıklayabilmek için hem fonksiyonel bir form olan f , hem de parametre vektörü α bulunur.

$$P(y \mid x) = f(x, \alpha) \tag{2.7}$$

α parametreleri, genellikle maksimum olabilirlik kestirimi ile veri kümesine göre belirlenir. f fonksiyonel formu, lojistik regresyon için parametrik bir yöntem olarak bilinir. Genel olarak, bir lojistik regresyon modeli, veri kümesindeki iki kategoriden biri için sınıf üyelik olasılığını hesaplar.

Sadece orijinal ortak değişkenleri içeren bir lojistik regresyon modeline ana etki modeli denir. Bu model ürünler gibi etkileşim terimleri dahil olmak üzere, değişkenleri kovaryantlarda nonlineer yapar ve bu nedenle daha esnektir. Yüksek esneklik genel olarak daha çok istense de, daha önce görülmemiş vakalarda bir modelin doğruluğunu potansiyel olarak azaltabilir, bu yüzden modelin aşırı yüklenmesi için daha yüksek bir risk taşır (Eğitim durumlarını ezberlemek). Tahmini modellemede, eğitim durumlarının

belirlenmesi görevin sadece bir parçasıdır, en önemli hedef ise yeni vakaların doğru bir şekilde sınıflandırılmasıdır [43].

2.1.1.4. K En Yakın Komşu Algoritması

Noktalar arasındaki mesafelere dayalı bir sınıflandırıcı algoritmadır. Bu algoritmada her bir örneğe ait noktalar göz önüne alınır ve bu noktalara göre işlem gerçekleştirilir. K En Yakın Komşu (KEYK) algoritması diğer sınıflandırıcı algoritmalara göre eğitim maliyeti en düşük yaklaşımdır. Basit olmasına rağmen birçok çalışmada başarısını kanıtlamış ve en yaygın kullanılan sınıflandırıcı algoritmalarından biridir [44,45].

Hesaplama yapılırken, genellikle noktalar arası (öznitelikler arası) mesafe farkının kareleri toplamının karekköküne dayalı bir yöntem olan Öklid mesafesi kullanılır. Bu mesafe hesabında eğitim kümesindeki özniteliklere ait değerler ile test kümesindeki özniteliklere ait uzaklık hesabına bakılmaktadır [46]. Her bir sınıf için ayrı bir değer hesaplanmakta ve elde edilen sonuca göre örnek en büyük değere sahip sınıfın etiketiyle etiketlendirilmektedir. Burada diğer önemli bir husus en yakın kaç komşu değerine bakılacağıdır. Literatürde genellikle k katsayısı 3 alınırken, bu değer veri kümesine göre değişebilmektedir. Ancak n katsayısının tek sayılardan oluşması gerekmektedir.

Öklid Mesafe Hesabı:
i=öznitelik sayısı;
X=test kümesine ait öznitelikler kümesi;
Y=eğitim kümesine ait öznitelikler kümesi;

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_i - y\dot{i})^2} \tag{2.8}$$

2.1.1.5. Karar Ağaçları Algoritması

Karar Ağaçları (KA) öznitelikleri hiyerarşik olarak bir kural tablosuna göre değerlendirerek kural tablosundaki kriterlere göre sınıflandırma yapan bir sınıflandırma algoritmasıdır. Esasen kural tabanlı olan bu yaklaşımın öğrenilebilirliğini kural yapısının eğitim verisi kullanılarak en doğru sınıflandırmayı yapacak şekilde açgözlü bir algoritma ile öğrenilmesine dayalı teknikler sağlamaktadır. Öğrenmeyi sağlayan açgözlü algoritmalar ise bir ağaç yapısında konstrüktif olarak gerçeklenmektedir. KA yapısal

olarak, iki tip düğümden oluşur. Bunlar: ara ve terminal (yaprak) düğümlerdir. Ara düğümlerin seçimi, değişen k ve ölçme fonksiyonu için tüm olası değerler üzerinde en iyi değeri seçmeyi içerir. Terminal düğümleri ise bir karar / teşhis oluşturur. Tanı, her seferinde tek bir soruyu soran ve cevaba bağlı olarak başka bir soru kümesini içeren ağacın başka bir dalının izlendiği adım adım karar verme süreci ile sağlanır. Bir kümenin entropisi (Bölümlenmemiş set veya altkümelerinden herhangi biri.) aşağıdaki entropi formülü ile hesaplanabilir.

$$-\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n} \tag{2.9}$$

Burada k problem içindeki kategorilerin sayısını, n_i i. kategorideki sınıf sayısını ve n veri kümesindeki toplam sınıf sayısını gösterir [47].

Bu algoritma, veri kümesini, verilerin ayrılmasını en üst düzeye çıkaran ve ağaç benzeri bir yapıya yol açan bir ölçüte göre tekrar tekrar ayırır. Kullanılan en yaygın kriter bilgi kazanımıdır; her bölünmede, bu bölünmeye bağlı entropi azalmasının maksimize edildiği anlamına gelir. P (y | x) tahmini, y sınıf elemanlarının, veri düğümü x içeren yaprak düğümünün tüm elemanları üzerindeki oranıdır.

Diğer makine öğrenme yöntemleriyle karşılaştırıldığında, karar ağaçları kara kutu model olmama avantajına sahiptir ve kural olarak kolayca ifade edilebilirler. Birçok uygulama alanında, bu avantaj dezavantajlardan daha ağırdır, bu nedenle bu modeller tıpta açık kural modelleri şeklinde yaygın olarak kullanılmaktadır [43].

2.1.1.6. Gaussian Naive Bayes Algoritması

Bayes teoremine ve nitelikler arasındaki bağımsızlık varsayımına dayalı basit bir olasılıksal sınıflandırıcı algoritmadır [48]. Hesaplamaya alınan her bir parametrenin istatistiksel özelliklerine bakarak bir sınıflandırma işlemi gerçekleştirmektedir. Bu algoritmada her bir sınıfa ait olasılık değeri hesaplanarak sınıf olasılığı yüksek olan değere göre örnek etiketlendirme işlemine tabi tutulur. Bayes sınıflandırıcısının, gen ifade verilerinde diğer sınıflandırıcılara göre daha başarılı sonuçlar elde ettiğine dair gözlemler bulunmaktadır [41].

Bayes teoremi eğitim verilerine bakarak her bir test sınıfına ait en yüksek olasılık değerlerini inceleyerek etiketlendirme işlemini gerçekleştirmektedir. Bayes teoreminde

sürekli verileri tahmin etme amacıyla Gauss dağılımından yararlanılmaktadır. Gaussian Naive Bayes (GNB) sınıflandırıcısı için bir test sınıfına ait veriyi tahmin etmede aşağıdaki denklemlerden yararlanılmaktadır:

$$g(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(x-\mu)^2}{2\sigma}} \tag{2.10}$$

$$P(x|c_i) = g\left(x_k, \mu_{c_i}, \sigma_{c_i}\right) \tag{2.11}$$

Denlemlerde μ ortalamayı, σ standart sapmayı, x ise özniteliği ifade etmektedir. $P(x|c_i)$ ise X özniteliğinin c_i sınıfında olma olasılığını ifade etmektedir.

2.1.2. Negatif Olmayan Matris Ayrıştırması

Birçok veri seti, negatif olmayan değerler içermektedir (Örn: görüntü ve metin veri setleri). Bu yüzden kullanılan yöntemlerde negatif değerler oluşmakta ve bu negatif değerler verinin yorumlanmasında zorluklar oluşturmaktadır.

Negatif Olmayan Matris Ayrıştırması (NOMA), yakın zamanda gerçekleştirilen verinin negatif olmayan bir şekilde lineer olarak temsil edilmesini sağlayan bir tekniktir. Bu yöntem, indirgenen boyuttaki değerlerin negatiften farklı olmasıyla diğer metotlardan ayrılır. Yani yüksek boyuttan düşük boyuta indirgenen matriste negatif elemanlar bulunmaz. Bütün değerler pozitiftir [49].

2.1.2.1. NOMA Formülasyonu

$X = X_{ij}$ $m \times n$ boyutlarında bir negatif olmayan matris olsun, X matrisinin her bir elemanı 0'dan büyüktür ($X_{ij} \geq 0$) ve her bir sütunu analiz edilecek veri noktalarını temsil eder. İndirgenmek istenen yüksek boyutlu X matrisi, yaklaşık olarak, boyutları $m \times r$ olan $W = W_{ij}$ ve boyutları $r \times n$ olan $H = H_{ij}$ matrislerinin çarpımı şeklinde yazılabilmelidir [29,31-33]. NOMA, hata oranını en aza indirebilmek için;

$$\min_{\substack{\omega ij \geq 0 \\ hij \geq 0}} \|X - WH\|_F^2 \quad [50] \tag{2.12}$$

En uygun W ve H matrislerini arar. W matrisinin her bir satırı temel vektör olarak bilinirken, H matrisinin her bir sütunu kodlama katsayısı vektörü olarak bilinir. WH^T

çarpımı X matrisinin NOMA'sı olarak adlandırılır ve X matrisine uzaklığı en düşük normda olan matris olması hedeflenmektedir. Çoğu r değeri için, WH^T yaklaşık olarak ayrıştırılmaktadır. r değeri için optimal bir değer bulunması pratikte oldukça önemlidir, ancak r'nin seçimi sıklıkla problemlere bağımlıdır. Ancak, temel vektörlerin sayısı(r), yaklaşık veri matrisi rankının bir üst sınırını kadardır; $rank(WH^T) \leq min(rank(W),rank(H)) \leq r$. Ayrıca orijinal veri matrisi (X) rankının üst sınırı bilinmektedir; $rank(X) \leq min(m,n)$. Bu yüzden r genellikle m ve n'den küçük seçilir; $r \leq min(m,n)$. Böylece, temsili matris WH^T, orijinal veri matrisine (X) göre düşük boyutlu olup, gürültü ve veri artıklığını azaltma avantajı sağlar [50,51].

NOMA'nın bir başka önemli özelliği, W matrisindeki temel vektörler gibi öznitelikleri en aza indirgeyen sayısal yöntemlerin yeteneğidir. Bu öznitelikler, tanımlama ve sınıflandırma için daha sonradan kullanılabilir. NOMA W ve H matrislerinde negatif girişlere izin vermeyerek, parçaların bir bütün oluşturması için eksiği olmayan bir kombinasyonu mümkün kılar [52].

Negatif olmayan matris ayrıştırması (NOMA), son yıllarda giderek daha popüler bir veri işleme aracı haline gelmiş olup, veri madenciliğinde, gen ifadesi çalışmalarında, örüntü tanıma ve biyoinformatik dalları dahil olmak üzere çeşitli topluluklar tarafından yaygın olarak kullanılmaktadır. Veri setindeki her bir veri örneği, negatif olmayan ağırlıklarla ağırlıklandırılmış bir grup negatif olmayan vektörün doğrusal bir kombinasyonu ile tahmin edebilir [50-54]. NOMA, sayım veya ölçüm sonucunda ortaya çıkan doğal sayı veya negatif olmayan rasyonel sayılar şeklinde ortaya çıkan profilleri doğrudan benzer şekildeki profillerin ağırlıklı ortalaması şeklinde ifade edebilecek doğal bir yaklaşım ortaya sürebilmektedir.

Yang ve Oja'nın çalışmasında [54], bir negatif olmayan matris ayrıştırması çeşidi olan Projektif Negatif Olmayan Matris Ayrıştırması (PNOMA) yöntemi analiz edilmiştir. Yeni yöntem, bir yansıtma matrisini, rekonstrüksiyon hatasını en aza indirerek, düşük dereceli bir matris ve onun transpozu olarak ayırır. Orijinal veri matrisi ile bu yeni matris arasındaki benzerlik Frobenius matris normu veya modifiye edilmiş Kullback-Leibler sapması ile ölçülebilir. Her iki ölçüm de çarpımsal güncelleme kuralları ile en aza indirilmiştir. PNOMA formülasyonunun mevcut NOMA yöntemlerine ve kümeleme yaklaşımlarına bağıntılı olduğu gösterilmiştir. Ayrıca, Lagranj çarpanları kullanılarak yapılan türev işlemi, rekonstrüksiyon ve seyreklik arasındaki ilişkiyi ortaya koymaktadır. Üç ayrı veri tabanı üzerinde yapılan deneysel çalışma, PNOMA'nın kümelemede en iyi

ya da en iyiye yakın sonuçlara ulaşabileceğini göstermektedir. Önerilen algoritma, özellikle yüksek boyutlu veriler için, karşılaştırılan NOMA yöntemlerinden daha verimli çalışır.

Berry vd. [55]'nin çalışmasında, metin madenciliği ve spektral veri analizi alanlarında öznitelik çıkarımı ve tanımlaması için NOMA algoritmalarının geliştirilmesi ve kullanımı sunulmaktadır. Ortaya çıkan NOMA sonuçları için hem seyreklik hem de düzgünlük kısıtlamalarına dayanan melez yöntemlerin evrimi ve yakınsama özellikleri tartışılmıştır. NOMA çıktılarının, belirli bağlamlarda yorumlanabilirliği, büyük ölçekli ve zamanla değişen veri setleri için NOMA algoritmalarının modifikasyonuyla, bu tip veriler üzerinde kullanılmaya elverişli olduğu sonucu rapor edilmiştir.

Devarajan [56] araştırmasında, NOMA yöntemi, negatif olmayan bir matrisin (V), çarpımsal olmayan bir güncelleme algoritması vasıtasıyla, W ve H olmak üzere iki negatif olmayan matris halinde ayrışmasını içeren, denetimsiz, parça temelli bir öğrenme paradigması olarak tanıtmıştır. p x n boyutundaki bir profil matrisi (V), p kadar gen, n sayısınca örnek içerir. W'nun her bir sütunu bir meta-profili, H'nin her bir sütunu ise model bileşenini (metagene expression pattern) temsil eder. Bu yöntem, çeşitli uygulamalara vurgu yaparak, hesaplama biyolojisinde bir veri analiz ve yorumlama aracı olarak incelenmiştir. Araştırmacılar, bu yaklaşımı gen ifadesi çalışması uygulamalarına özelleştirerek V matrisini her bir örneğin gen ifadesi profillerinin popüle edildiği veri matrisi, W matrisini meta-gen profilleri kütüphanesinin barındırıldığı modelleme matrisi, H matrisini ise model bileşenlerinin ağırlıklı ortalama ile gen ifadesini oluşturacağı kodlama matrisi olarak tanımlamışlardır.

2.1.2.2. Bağıl Hata Değeri

X, m × n boyutlarında negatif olmayan bir matris olmak üzere, $X\text{-}WH^T$ farkının hesaplanmasıyla bulunan fark matrisinin her bir elemanının kareleri toplamının (Frobenius normu) eleman sayısına bölünmesi ile hata değeri (HD) bulunur.

$$HD = \text{Frobenius Norm} \; / (m \times n) \tag{2.13}$$

Başlangıçtaki X matrisinin tüm elemanları toplamına XT diyelim. X matrisinin enerjisini hesaplayıp X_enerji = XT / (m × n), HD ile bu değer oranlandığında ise bağıl hata değeri (BHD) elde edilir.

$$BHD = HD \,/\, X\_enerji \tag{2.14}$$

2.1.3. Tekil Değer Ayrışması

Bir matrisin Tekil Değer Ayrışması (TDA), öznitelik çıkarma ve boyut küçültme için sağlam, sayısal olarak güvenilir ve verimli bir teknik sağlar [57]. Tekil değer ayrışımına, birbirleriyle uyumlu üç bakış açısından bakılabilir. Bir yandan, korelasyonlu değişkenleri, orijinal veri öğeleri arasındaki çeşitli ilişkileri daha iyi ortaya koyan korelasyonsuz gruplara dönüştürmek için bir yöntem olarak görebiliriz. TDA aynı zamanda, verilerin en fazla çeşitlilik gösterdiği boyutları tanımlamak ve sıralamak için kullanılan bir yöntemdir. Bu bilgi üçüncü özellikle bağlantlıdır; en fazla varyasyonun nerede olduğunu belirlendiğinde, daha az boyut kullanılarak orijinal veri noktalarının en iyi yaklaşımını bulmak mümkündür. Bu nedenle TDA, veri indirgemek için bir yöntem olarak görülebilir.

Yüksek boyutlu, çok değişken bir veri setini alıp, orijinal verilerinin alt yapısını daha açık ve sıralı şekilde yansıtan daha düşük boyutlu bir alana indirgemek TDA'nın arkasındaki temel fikirlerdir. TDA'yı doğal dil işleme, biyoinformatik, veri madenciliği gibi birçok uygulama alanı için pratik yapan şey, verileri büyük ölçüde azaltmak için belirli bir eşik değeri altındaki varyasyonu basitçe görmezden gelmek, buna rağmen temel bağlantıların korunmuş olduğundan emin olmaktır.

TDA, lineer cebir temelli bir teoremden [58] yola çıkarak dikdörtgen bir matrisin (A) üç matrise (ortogonal bir matris (U), diyagonal bir matris (S) ve ortogonal bir matrisin (V) transpozisyonuna) parçalanabileceğini söyler. Teorem genellikle aşağıdaki gibidir;

$$A_{mn} = U_{mm}\ S_{mn}\ V_{nn} \tag{2.15}$$

$$U^T U = I, \qquad V^T V = I$$

U'nun sütunları AA^T'nin ortonormal özvektörleridir.V'nin sütunları, A^TA'nın ortonormal özvektörleridir ve S, U veya V'den özdeğerlerin kareköklerini azalan sırada içeren bir diyagonal matristir [58].

2.1.4. Öznitelik Seçimi

Son otuz yılda, makine öğrenimi ve veri madenciliği alanlarında üretilen verilerin boyutsallığı hızlı bir şekilde artmıştır. Son derece yüksek boyutsallığa sahip veriler, mevcut öğrenme yöntemlerine, ciddi zorluklar getirmiştir (Örn: Boyutsallık laneti). Çok

sayıda özelliğin bulunmasıyla, bir öğrenme modeli aşırı öğrenme eğilimi göstererek performans dejenerasyonlarına neden olabilir. Boyutsallık laneti problemini ele almak üzere, makine öğrenimi ve veri madenciliği araştırma alanında önemli bir dal olan boyut azaltma teknikleri çalışılmıştır. ÖS, uygulayıcılar arasında boyutsallığı azaltmak için yaygın olarak kullanılan bir tekniktir. İlgili özniteliklerin küçük bir alt kümesini, genellikle daha iyi öğrenme performansı sağlayan belirli bir geçerlilik değerlendirme kriteri kullanarak (Örn: sınıflandırma için daha yüksek öğrenme doğruluğu), düşük hesaplama maliyeti ve daha iyi model yorumlanabilirliiği ile seçmeyi amaçlamaktadır [35].

Aşağıda, kullanılan iki öznitelik seçim yönteminin işleyişi anlatılmıştır. Bu yöntemler uygulanırken python yazılımı kullanılmış ve sklearn kütüphanesinden faydalanılmıştır.

2.1.4.1. Tek Değişkenli Öznitelik Seçimi Yöntemi

Tek değişkenli öznitelik seçimi (TDÖS), tek değişkenli istatistiksel testlere dayanarak en iyi özellikleri seçmeye çalışır. Bu yöntem bir tahmin yöntemi kullanılmadan önce bir ön işlem adımı olarak görülebilir [59]. Tek değişkenli öznitelik seçimi, öznitelik ile karşı değişken arasındaki ilişki gücünü belirlemek için her bir özelliği tek tek inceler. Bu yöntemin anlaşılması ve çalıştırılması kolaydır ve genel olarak verilerin daha iyi anlaşılmasına yardımcı olur, ancak daha iyi bir genellemeyle öznitelik setini optimize etmek için önerilmez. Tek değişkenli özellik seçimi genellikle verileri, veri yapısını ve özelliklerini daha iyi anlamak için en iyi yöntemdir [60]. Bu yaklaşımda, bir özniteliğin farklı sınıflarda ortalamada birbirinden farklı değerler alıp almadığını test edecek istatistiki testlerin p değerlerine bakılarak en düşük p değerini veren özniteliklerin seçilmesi prensibi izlenir. Genellikle Student's t-testi ve Kluskal-Wallis testleri p değeri hesaplaması için kullanılan testlerdir.

2.1.4.2. Özyineli Öznitelik Eleme Yöntemi

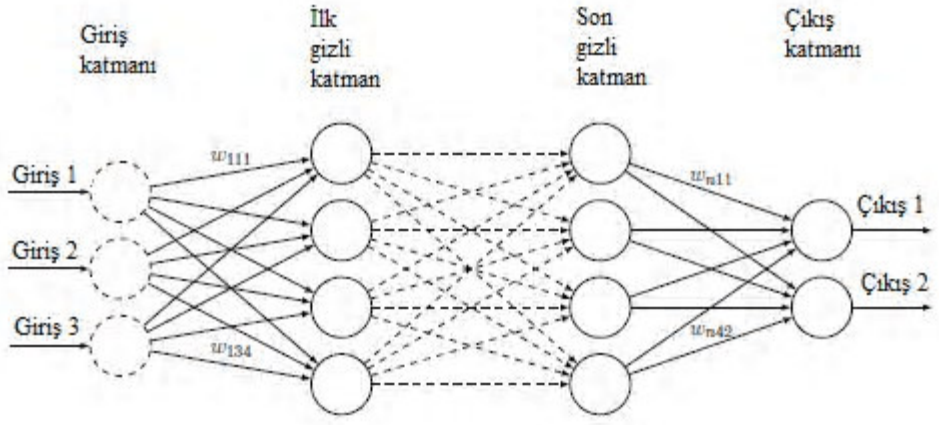
Özyineli Öznitelik Eleme (ÖÖE) yöntemi, temel olarak özellikleri önemlerinin bir ölçüsüne göre sıralayan bir özyineli bir süreç içerir. Her bir iterasyonda öznitelik önemi ölçülür ve en az ilgili olan çıkarılır. Bu ölçüm için sarmalayıcı yöntemler kullanılarak her özniteliğin sınıflandırmaya olan katkısı tek tek ölçülür. Başka bir olasılık ise, süreci hızlandırmak için her seferinde bir grup özniteliği kaldırmaktır. Bazı ölçümler için, her

bir özniteliğin nispi önemi, kademeli olarak eleme işlemi sırasında (özellikle yüksek korelasyonlu özellikler için) farklı öznitelikler alt kümesi üzerinde değerlendirildiğinde önemli ölçüde değişebilir. Bu yüzden yineleme işlemi oldukça önemlidir. Öznitelikler elendiği sıra, bir final sıralaması oluşturmak için kullanılır. Öznitelik seçim süreci, yalnızca bu sıralamadaki ilk n özelliklerin alınmasından oluşur [61].

2.1.5. İleri beslemeli Yapay Sinir Ağları

Sinir ağları, insan beynini modelleme çabasıyla oluşturulmuş, fakat seneler içerisinde ilk amaçlarından farklı olarak, birçok alanda güçlü makine öğrenimi araçları olarak kullanılmıştır.

Sinir ağlarının yapısını inceleyecek olursak, her biri çoklu nöron içeren çoklu katmanlara ayrılırlar. Her nöron, Şekil 2.5' de gösterildiği gibi, yandaki katmana ağırlıklarla (w) bağlanır. Her bir nöron, bir bias ve bir önceki katmanın çıktılarının ağırlıklı toplamını girdi olarak alan bir aktivasyon fonksiyonundan oluşur.

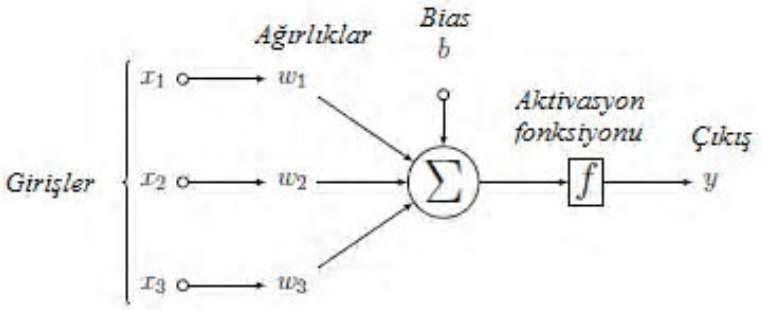


Şekil 2.5. İleri beslemeli Sinir Ağının Yapısı

- $n_1, \ldots, n_k$ i. Katmandaki nöronlar
- $m_1, \ldots, m_k$ i + 1. Katmandaki nöronlar
- f aktivasyon fonksiyonundaki nöronlar
- w_{ij} ni nöronu ve mj nöronu arasındaki bağlantı ağırlığı
- Bias(n_i) i. Nöronun biası
- Çıkış(n_i) i. nöronun çıkışı

$$\text{Çıkış(mi)} = \text{f}(\sum_{x=1}^{k} \text{Çıkış}(n_i) * \omega_{x_i} * Bias\ (m_i) \tag{2.16}$$

Bir tahminin nasıl elde edileceği denklem 2.14' te gösterilmiştir. Bir ağın bir girdi ile beslenmesi ve son çıktıyı elde etmek için yayılması gerekir. Giriş, sığ katmanlardan daha derinlere geçer ve bu ileriye doğru yayılma olarak adlandırılır.



Şekil 2.6. Nöron Tanımı

Sinir ağları birden çok parametreyle ve birçok yolla inşa edilebilir. Ayrıca, sinir ağları çok sığdan çok derine ve çok dardan çok genişe birçok form alabilir ve mimarisine birçok kısıtlama da eklenebilir. Evrişimli katman ve kodlayıcı katman bunlardan bazılarıdır. Tüm bu parametreler, çalışılan problemle ilgili olarak değiştirilebilir.

Kullanılan yöntemde, ilk model sadece sınıflandırma amaçlı oluşturulmuştur. İkinci modelde ise sinir ağına ilk katmanda girdi olarak verilen veri seti (X), ikinci katmanda aynı boyutta bir ağırlık vektörü (w) ile çarpılıp yine X boyutunda bir vektör elde edilmiştir. Öznitelikler ikinci katmanda 0-1 arasındaki rastgele değerlerle çarpıldıktan sonra, eklenen leaky (akışkan) RELUdan yalnızca katsayıları pozitif olan değerler doğrudan çıkarken, katsayıları negatif olanların ise 0 olarak çıkması ve bu sayede öznitelik seçiminin otomatik olarak yapılması amaçlanmıştır.

Şekil 2.7. Model 2 için Açıklayıcı Görsel

3.modelde ise akışkan RELUdan çıkan değerlerin L1 normu alınarak (mutlak değerlerinin toplamı) çıkışa ulaşan değerlerin çoğunun 0 olması amaçlanmıştır. Bu sayede öznitelik sayıları yüksek oranda azaltılmış olur.

Bu yöntem için python yazılımı kullanılmış ve tensorflow kütüphanesinden yararlanılmıştır.

Şekil 2.8. Model 3 için Açıklayıcı Görsel

3. BÖLÜM
BULGULAR

3.1. Yöntemin Testi için Kullanılan Veri Setleri

Yaklaşımı test etmek amacıyla, her biri farklı çalışmalarda kullanılmış, erişime açık 7 ayrı mikrobiyom veri seti kullanılmıştır. Her bir veri setinde insan mikrobiyom örnekleri elde edilerek içeriğinden total DNA izolasyonu yapılmış, elde edilen DNA'lar yeni nesil dizileme teknikleri ile dizilenerek ham DNA verisi elde edilmiştir. Daha sonra ham DNA verisinin MetaPlan2 [62] programı ile taksonomik atamaları yapılmış ve her veri setinin içerisinde hangi mikroorganizma türlerinin olduğu ve her bir türün hangi bağıl bollukta bulunduğu tespit edilmiştir.

Test için kullanılan veri setleri: İltihaplı Bağırsak Hastalığı (İBH) [63], Siroz [64], Kolon Kanseri (KK) [65], Obezite [66], Sedef Hastalığı (SH) [67], Tip-II-Diyabet (T2D) [68] ve Kadınlarda Tip-II-Diyabettir (T2Dk) [69]. Mikrobiyota verisinin oluşturulmasında kullanılan örnekler sedef hastalığı için deriden, geri kalan hastalıklar içinse bağırsaktan alınmıştır.

Veri setlerinde bulunan hasta ve sağlıklı insan sayıları Tablo 3.1'de verilmiştir. Çalışmada 808'i Sağlıklı kontrol grubu olmak üzere toplamda 1531 örnek kullanılmıştır. Bu 1531 örnekten, toplamda 1455 tür tespit edilmiş ve her bir örnek 1455 boyutlu vektörler ile temsil edilmiştir. Bu profil vektörlerinin her bileşeni, kendisine karşılık gelen türün örnek içerisindeki bağıl bolluğunu temsil etmektedir.

Tablo 3.1. Veri setlerindeki deney-kontrol grubu sayıları

İBH	127 Crohn Hastalığı, 21 Ülseratif Kolit, 234 Sağlıklı Kontrol
Siroz	114 Siroz, 118 Sağlıklı Kontrol
KK	60 Kanser, 26 Küçük Hücreli Karsinoma, 48 Sağlıklı Kontrol
Obezite	89 Obez, 25 Zayıflık Hastalığı, 164 Sağlıklı Kontrol
Sedef	34 Sedef Hastası, 36 Sağlıklı Kontrol
T2D	135 Diyabet, 155 Sağlıklı Kontrol
T2Dk	49 Tip-2 Diyabet, 43 İnsülin Direnci, 53 Sağlıklı Kontrol

Python programlama dili ve sklearn kütüphanesi kullanılarak 7 ayrı veri setinde test edilen yöntemlerin sonuçları bu bölümde verilmiştir. Her bir yöntem için veri setleri 0,2 oranında test ve 0,8 oranında eğitim seti olacak şekilde ayrılarak kullanılmıştır. Bunun için yine sklearn kütüphanesindeki "test_train_split" metodundan yararlanılmıştır.

3.2. Sınıflandırma Sonuçları

İltihaplı bağırsak hastalığı (İBH), siroz, kolon kanseri (KK), obezite, sedef hastalığı (SH), tip-II-diyabet (T2D) ve kadınlarda tip-II-diyabet (T2Dk) veri setlerine; rastgele orman (RO), destek vektör makinesi (DVM), lojistik regresyon (LR), k en yakın komşu (KEYK), karar ağaçları (KA) ve gaussian naive bayes (GNB) sınıflandırma algoritmaları uygulanmış, doğruluk değeri sonuçları Tablo 3.2'de verilmiştir.

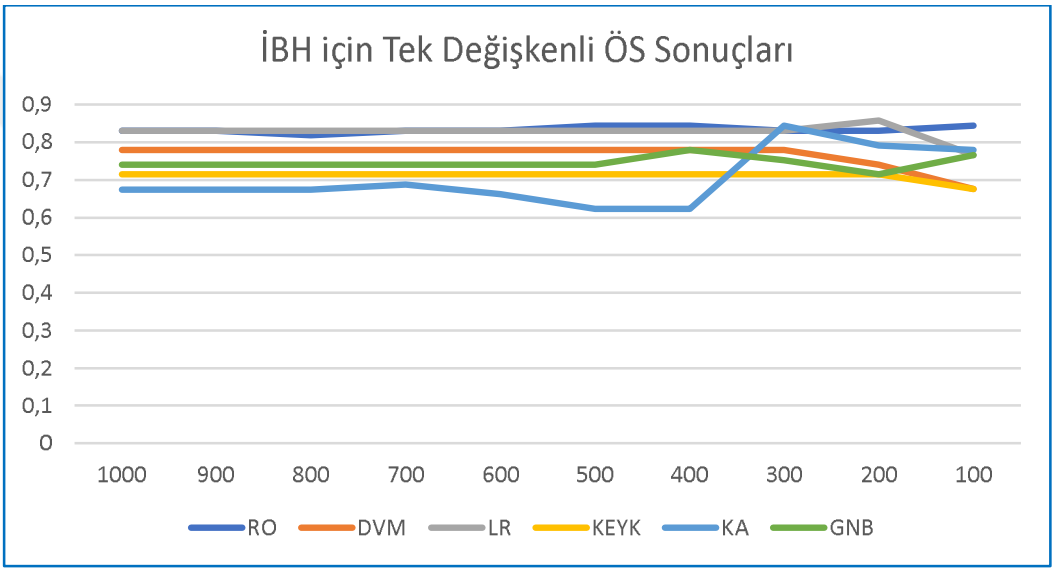
Tablo 3.2. Sınıflandırma Doğruluk Sonuçları

		Siroz	KK	Obezite	Sedef	T2D	T2Dk
RO	0.818	0.894	0.704	0.607	0.714	0.759	0.310
DVM	0.766	0.681	0.371	0.625	0.643	0.672	0.276
LR	0.831	0.681	0.519	0.536	0.571	0.672	0.380
KEYK	0.714	0.681	0.370	0.5	0.571	0.552	0.310
KA	0.662	0.766	0.481	0.518	0.714	0.603	0.207
GNB	0.740	0.702	0.629	0.25	0.643	0.741	0.241

3.3. Tek Değişkenli Öznitelik Seçimi Sonuçları

3.3.1. İltihaplı Bağırsak Hastalığı için TDÖS Sonuçları

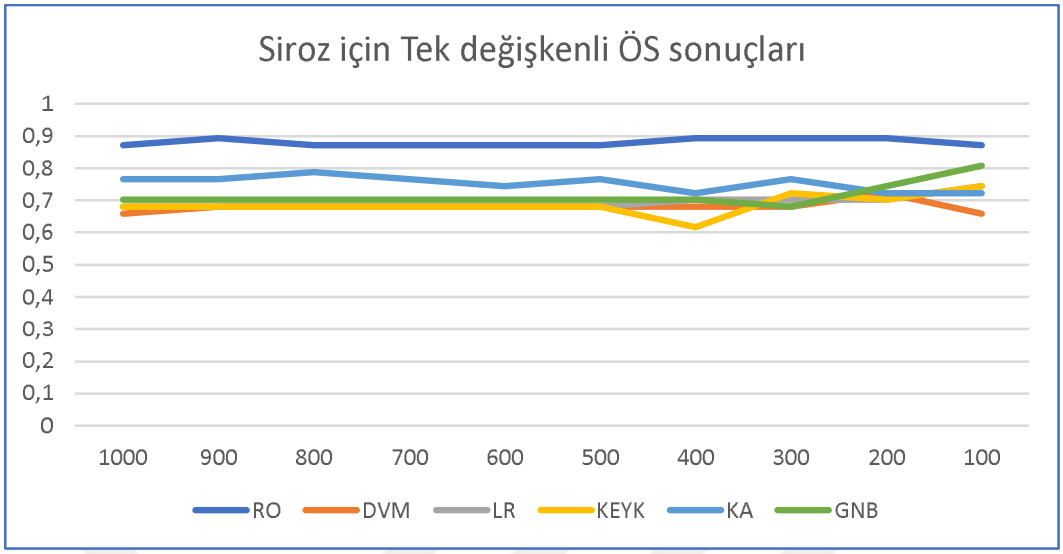
127 crohn hastası, 21 ülseratif kolit hastası ve 234 sağlıklı kontrol grubundan alınan bağırsak metagenomlarını içeren İltihaplı bağırsak hastalığı (İBH) veri setinde 3 ayrı sınıftan 382 örnek ve 1455 öznitelik bulunmaktadır. Öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.1'de verilmiştir.



Şekil 3.1. İBH için TDÖS Sonuçları

3.3.2. Siroz Hastalığı için TDÖS Sonuçları

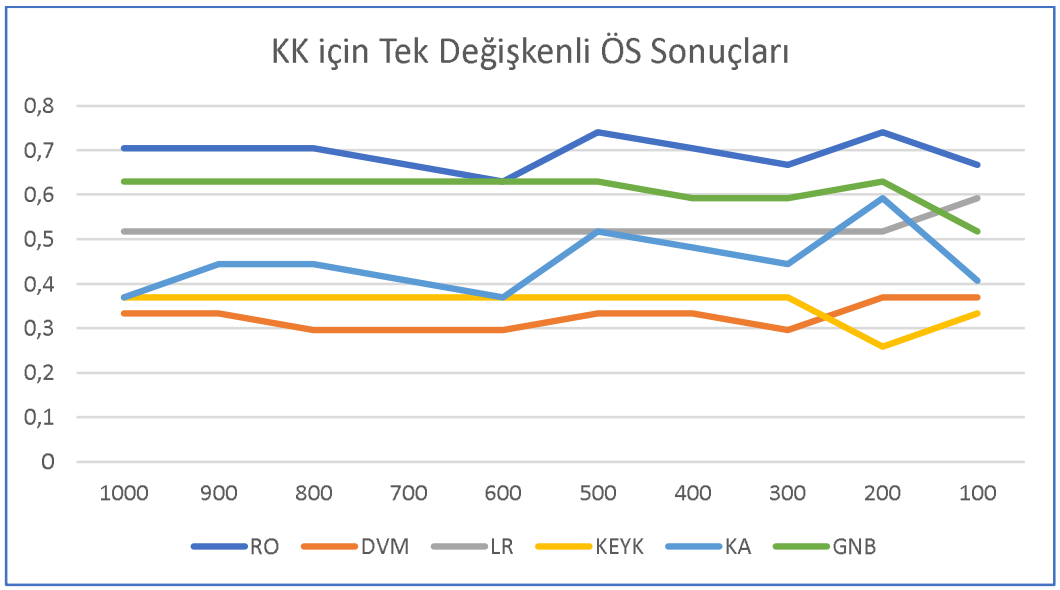
114 siroz hastası ve 118 sağlıklı kontrol grubundan alınan bağırsak metagenomlarını içeren Siroz veri setinde 2 ayrı sınıftan 232 örnek ve 1455 öznitelik bulunmaktadır. Öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.2'te verilmiştir.



Şekil 3.2. Siroz için TDÖS Sonuçları

3.3.3. Kolon Kanseri için TDÖS Sonuçları

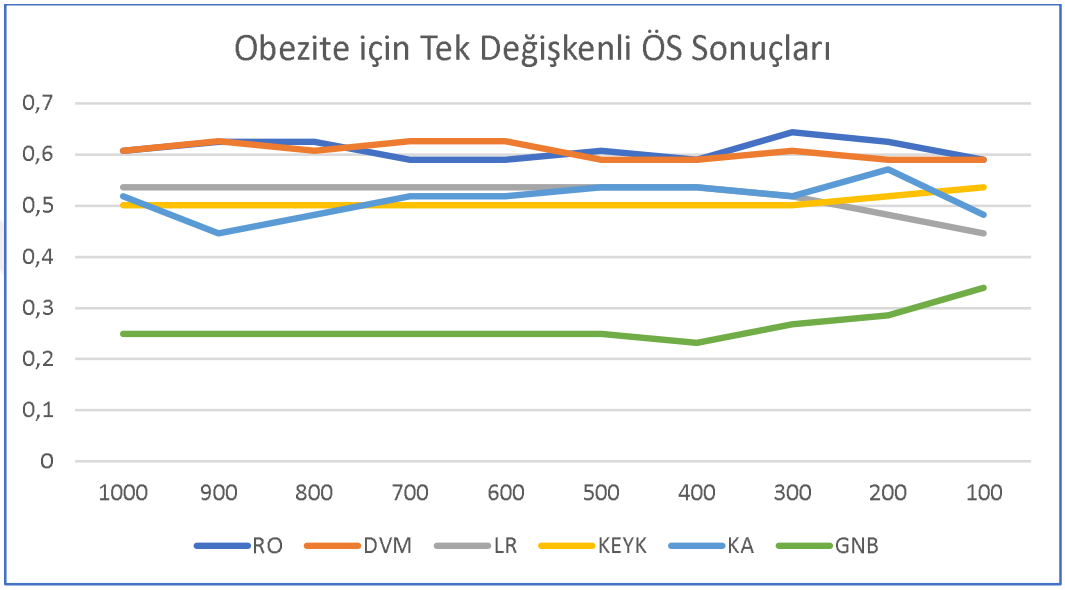
60 kanser hastası, 26 küçük hücreli karsinoma hastası ve 48 sağlıklı kontrol grubundan alınan bağırsak metagenomlarını içeren KK veri setinde 3 ayrı sınıftan 134 örnek ve 1455 öznitelik bulunmaktadır. Öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.3'te verilmiştir.



Şekil 3.3. KK için TDÖS Sonuçları

3.3.4. Obezite için TDÖS Sonuçları

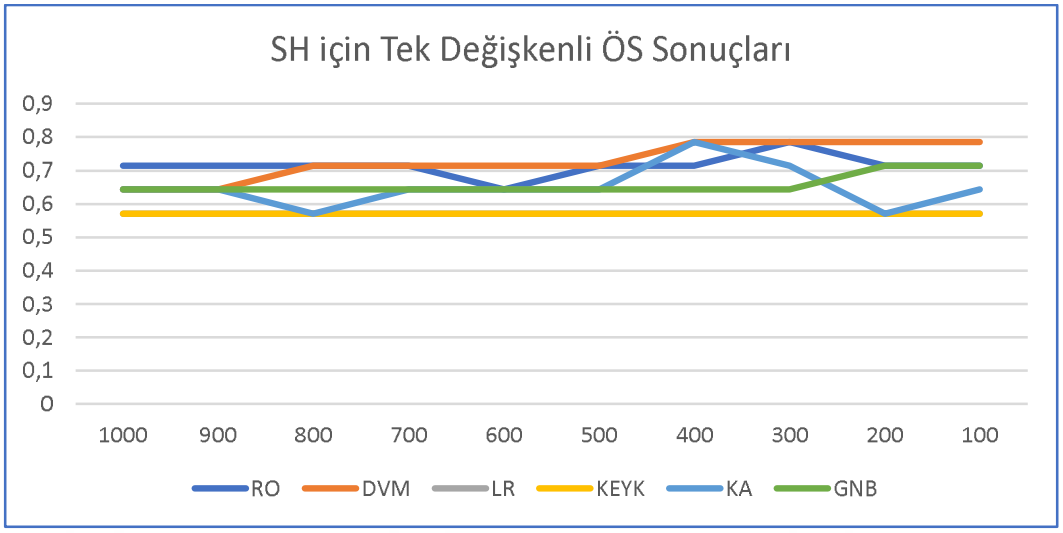
89 Obez, 25 zayıflık hastalığına sahip ve 164 sağlıklı kontrol grubundan alınan bağırsak metagenomlarını içeren obezite veri setinde 3 ayrı sınıftan 278 örnek ve 1455 öznitelik bulunmaktadır. Öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.4'te verilmiştir.



Şekil 3.4. Obezite için TDÖS Sonuçları

3.3.5. Sedef Hastalığı için TDÖS Sonuçları

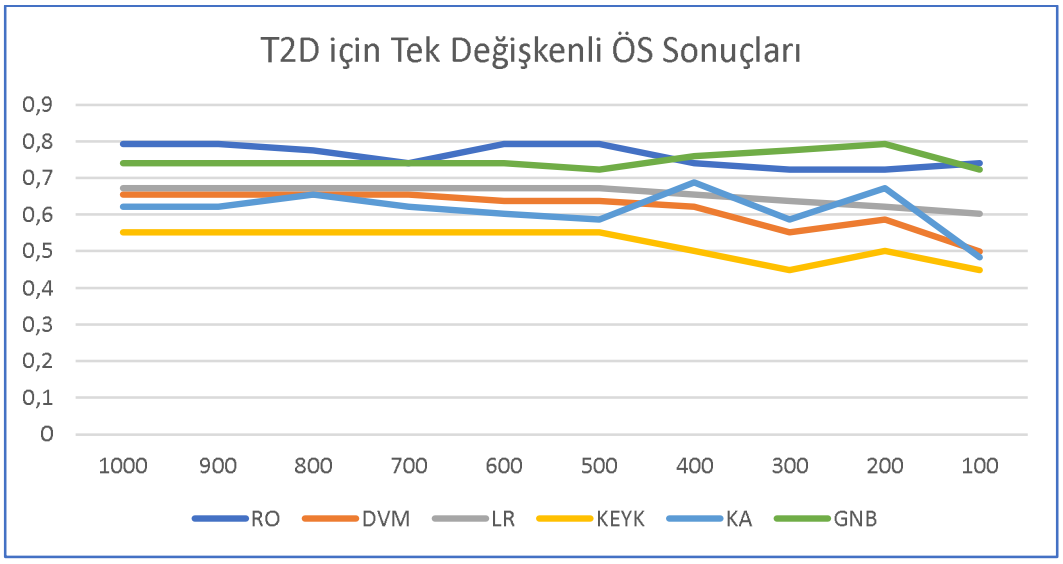
34 sedef hastası ve 36 sağlıklı kontrol grubundan alınan deri metagenomlarını içeren SH veri setinde 2 ayrı sınıftan 70 örnek ve 1455 öznitelik bulunmaktadır. Öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.5'te verilmiştir.



Şekil 3.5. SH için TDÖS Sonuçları

3.3.6. Tip-II-Diyabet için TDÖS Sonuçları

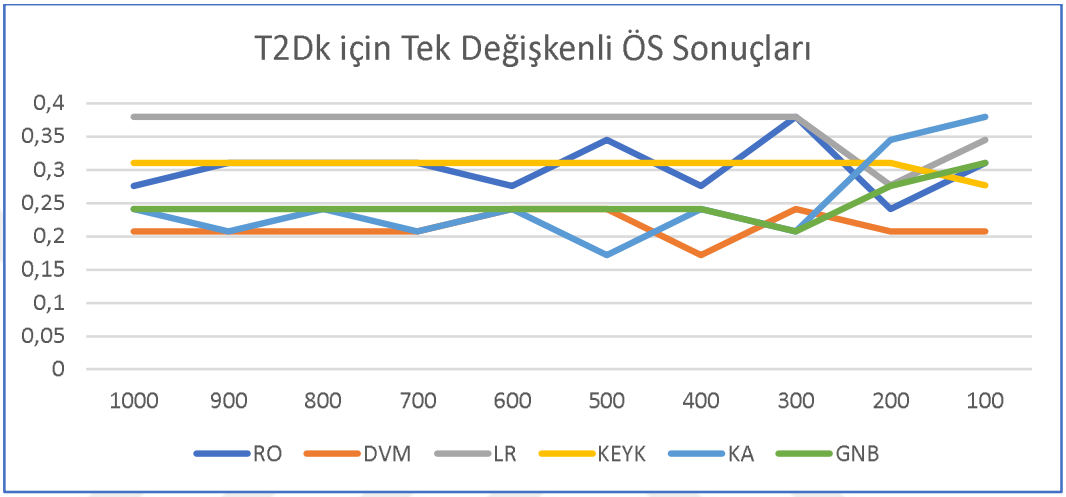
135 Tip-II diyabet hastası ve 155 sağlıklı kontrol grubundan alınan bağırsak metagenomlarını içeren SH veri setinde 2 ayrı sınıftan 290 örnek ve 1455 öznitelik bulunmaktadır. Öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.6'da verilmiştir.



Şekil 3.6. T2D için TDÖS Sonuçları

3.3.7. Kadınlarda Tip-II-Diyabet için TDÖS Sonuçları

49 Tip-II diyabet hastası, 43 insülin direncine sahip hasta ve 53 sağlıklı kontrol grubundan alınan bağırsak metagenomlarını içeren T2Dk veri setinde 3 ayrı sınıftan 145 örnek ve 1455 öznitelik bulunmaktadır. Öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.7'de verilmiştir.

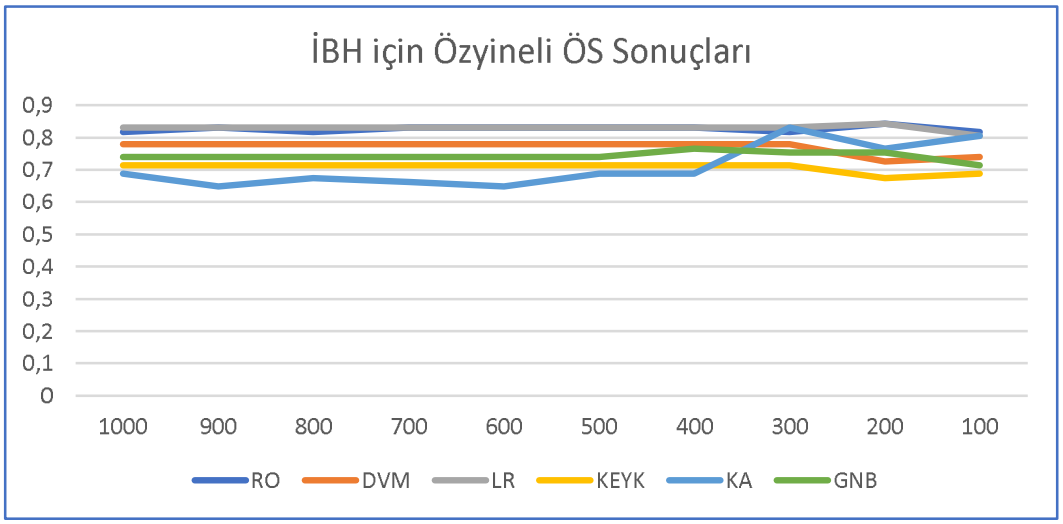


Şekil 3.7. T2Dk için TDÖS Sonuçları

3.4. Özyineli Öznitelik Eleme Sonuçları

3.4.1. İltihaplı Bağırsak Hastalığı için ÖÖE Sonuçları

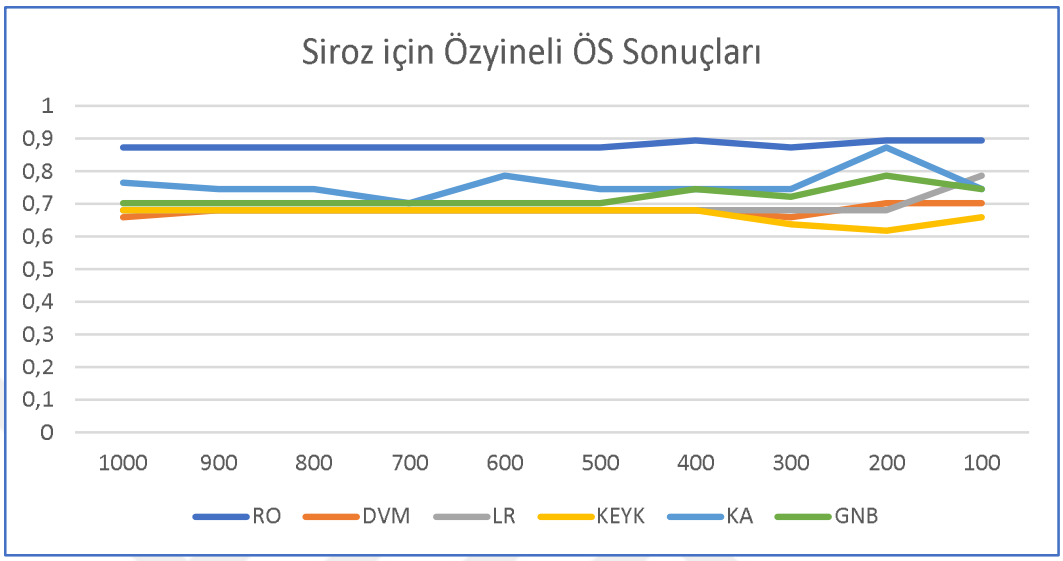
İltihaplı Bağırsak Hastalığı veri setinde, öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.8'de verilmiştir.



Şekil 3.8. İBH için ÖÖE Sonuçları

3.4.2. Siroz için ÖÖE Sonuçları

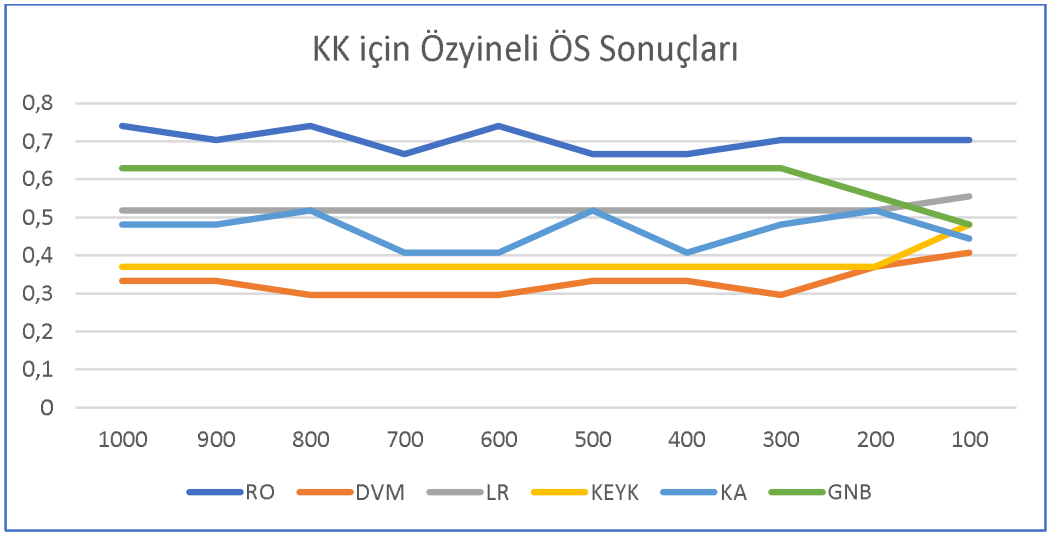
Siroz veri setinde öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.9'da verilmiştir.



Şekil 3.9. Siroz için ÖÖE Sonuçları

3.4.3. Kolon Kanseri için ÖÖE Sonuçları

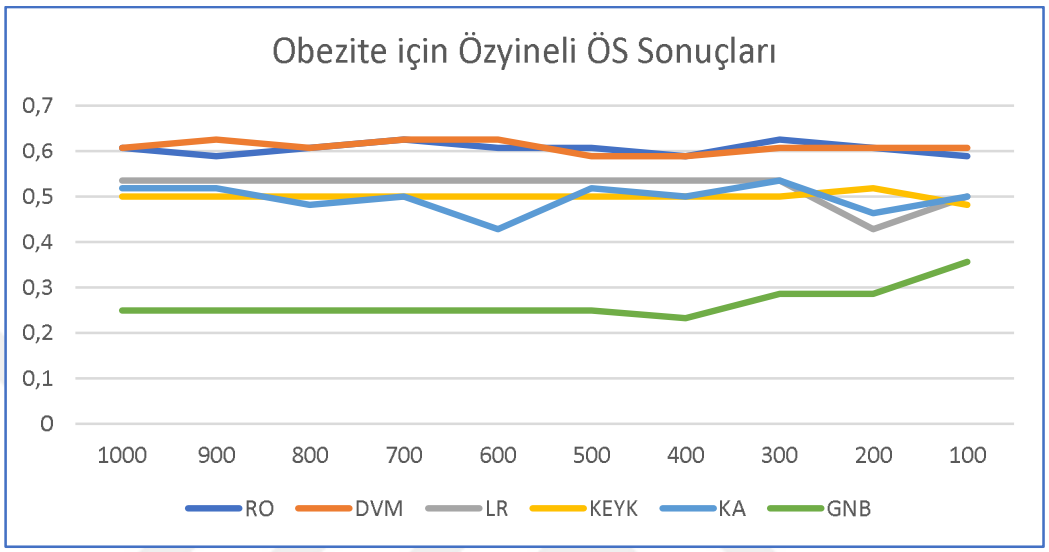
Kolon kanseri veri setinde öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.10'da verilmiştir.



Şekil 3.10. KK için ÖÖE Sonuçları

3.4.4. Obezite için ÖÖE Sonuçları

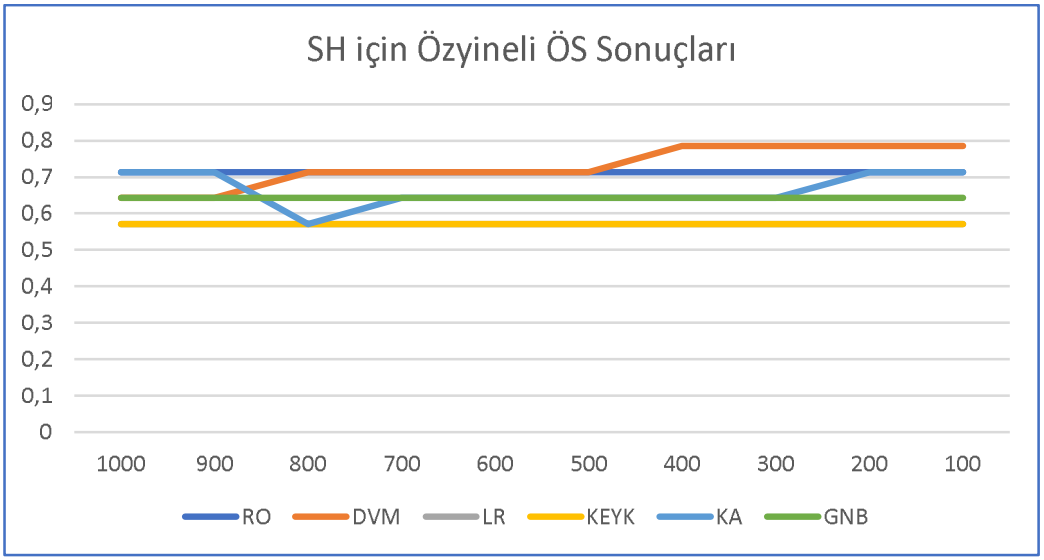
Obezite veri setinde öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.11'de verilmiştir.



Şekil 3.11. Obezite için ÖÖE Sonuçları

3.4.5. Sedef Hastalığı için ÖÖE Sonuçları

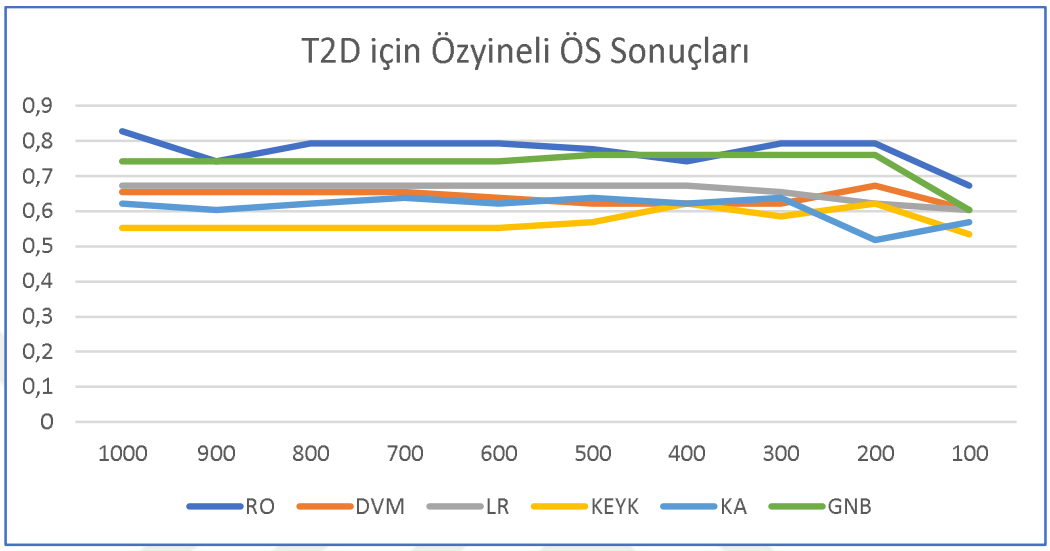
Sedef Hastalığı veri setinde öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.12'de verilmiştir.



Şekil 3.12. SH için ÖÖE Sonuçları

3.4.6. Tip-II-Diyabet için ÖÖE Sonuçları

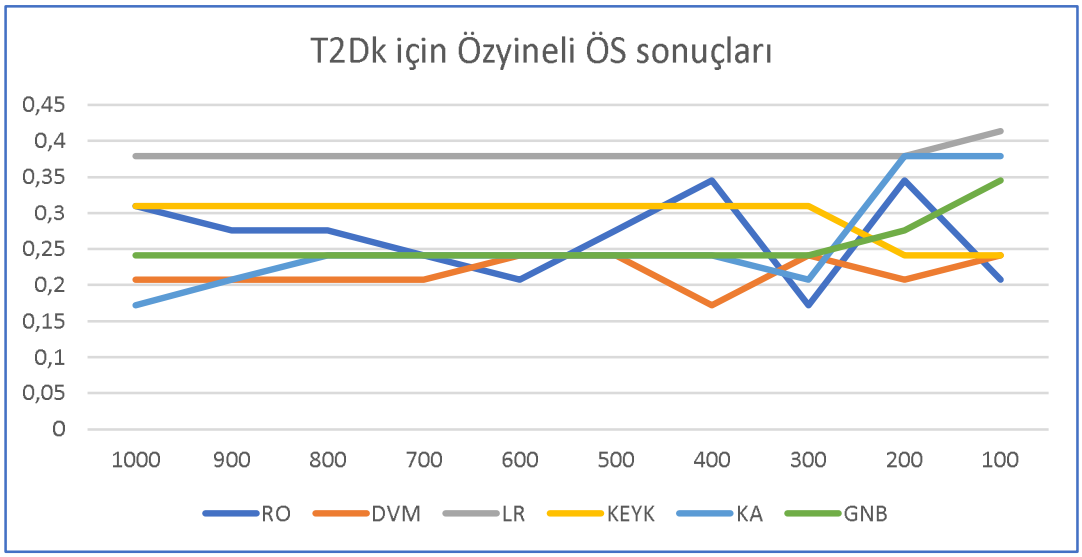
Tip-II diyabet veri setinde öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.13'te verilmiştir.



Şekil 3.13. T2D için ÖÖE Sonuçları

3.4.7. Kadınlarda Tip-II-Diyabet için ÖÖE Sonuçları

Kadınlarda tip-II diyabet veri setinde öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.14'te verilmiştir.

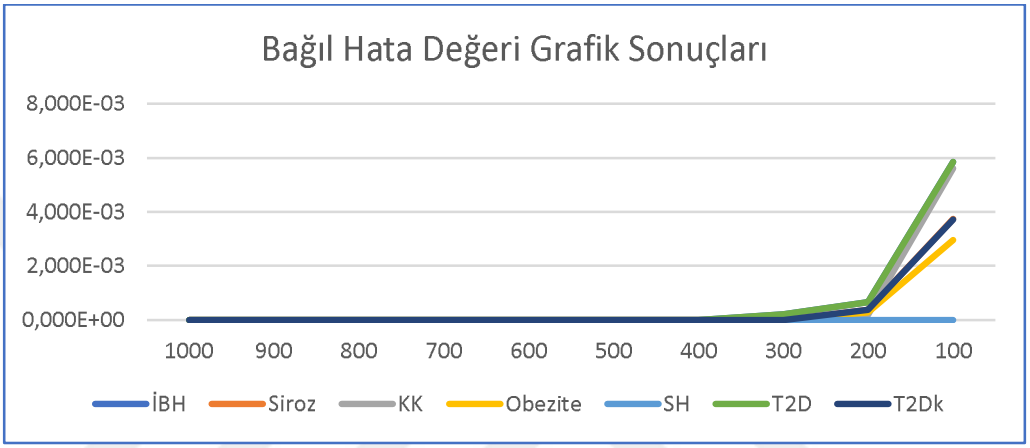


Şekil 3.14. T2Dk için ÖÖE Sonuçları

3.5. Negatif Olmayan Matris Ayrıştırması Sonuçları

3.5.1. Bağıl Hata Değeri Sonuçları

İltihaplı bağırsak hastalığı (İBH), siroz, kolon kanseri (KK), obezite, sedef hastalığı (SH), tip-II-diyabet (T2D) ve kadınlarda tip-II-diyabet (T2Dk) veri setleri için, öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen bağıl hata değeri sonuçları Şekil 3.15'de verilmiştir.



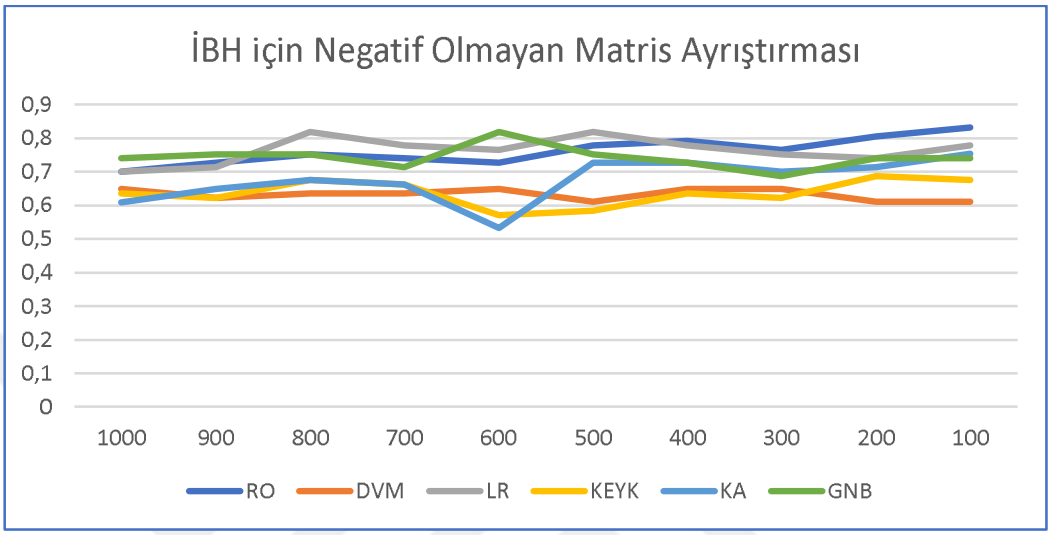
Şekil 3.15. Bağıl Hata Değeri Grafik Sonuçları

Tablo 3.3. . Bağıl Hata Değeri Sonuçları

		Siroz	KK	Obezite	SH	T2D	T2Dk
1000	4,673 E-07	7,389 E-06	2,544 E-07	1,589 E-06	3,331 E-08	4,673 E-07	9,289 E-07
900	5,283 E-07	1,458 E-06	2,867 E-07	1,742 E-06	3,757 E-08	5,283 E-07	9,001 E-07
800	5,908 E-07	2,877 E-06	5,374 E-07	5,026 E-06	4,186 E-08	5,908 E-07	1,033 E-06
700	8,138 E-07	3,135 E-06	2,633 E-07	1,825 E-06	2,771 E-08	8,138 E-07	9,454 E-06
600	1,195 E-06	2,912 E-06	1,185 E-06	2,982 E-06	7,309 E-08	1,195 E-06	8,461 E-06
500	4,961 E-06	3,362 E-06	4,508 E-07	1,291 E-06	3,911 E-08	4,961 E-06	3,551 E-06
400	4,511 E-05	1,803 E-05	8,094 E-07	3,602 E-06	4,724 E-08	4,511 E-05	1,074 E-05
300	2,190 E-04	5,372 E-05	3,101 E-06	2,983 E-05	2,422 E-08	2,190 E-04	7,232 E-05
200	6,710 E-04	3,660 E-04	1,890 E-04	2,920 E-04	2,590 E-08	6,720 E-04	3,660 E-04
100	5,840 E-03	3,740 E-03	5,610 E-03	2,950 E-03	1,208 E-07	5,840 E-03	3,700 E-03

3.5.2. İltihaplı Bağırsak Hastalığı için NOMA Sonuçları

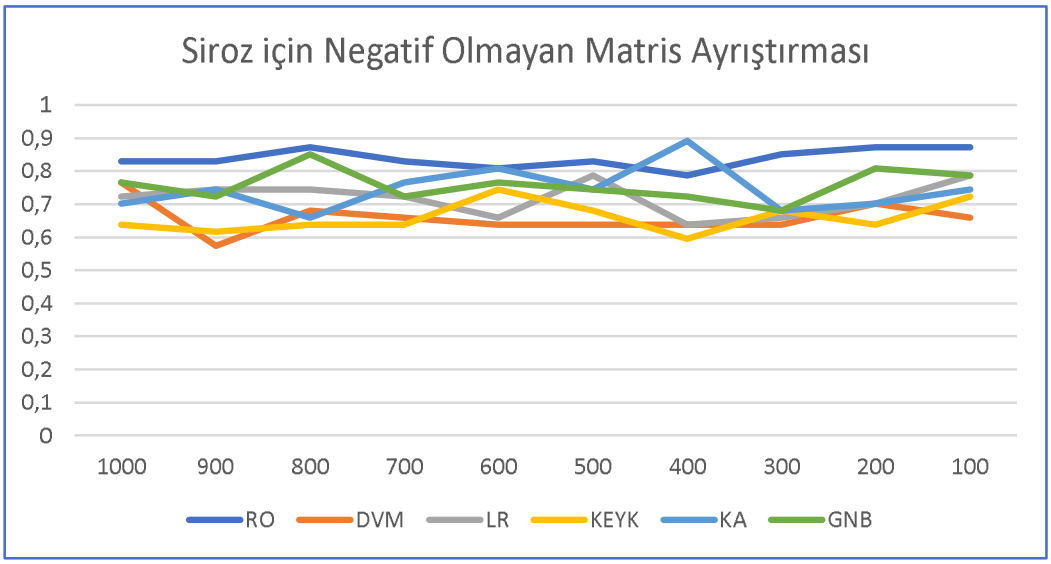
İBH veri setinde, öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.16'da verilmiştir.



Şekil 3.16. İBH için NOMA Sonuçları

3.5.3. Siroz için NOMA Sonuçları

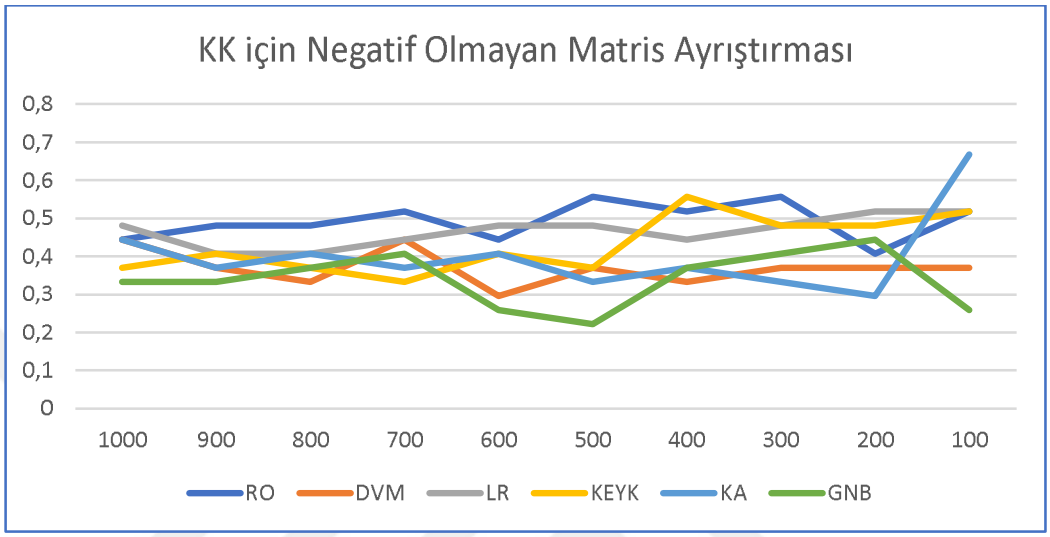
Siroz veri setinde öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.17'de verilmiştir.



Şekil 3.17. İBH için NOMA Sonuçları

3.5.4. Kolon Kanseri için NOMA Sonuçları

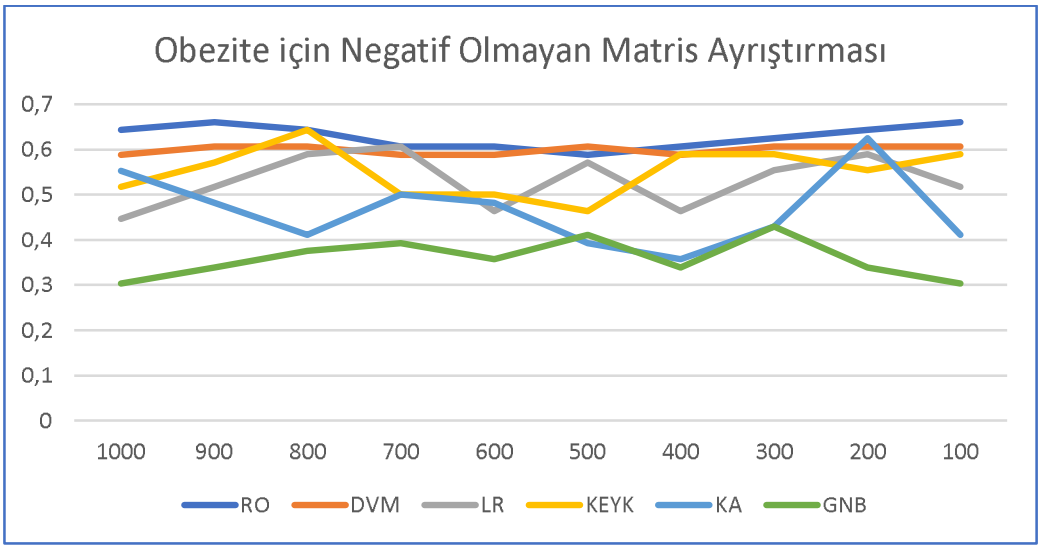
KK veri setinde öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.18'de verilmiştir.



Şekil 3.18. KK için NOMA Sonuçları

3.5.5. Obezite için NOMA Sonuçları

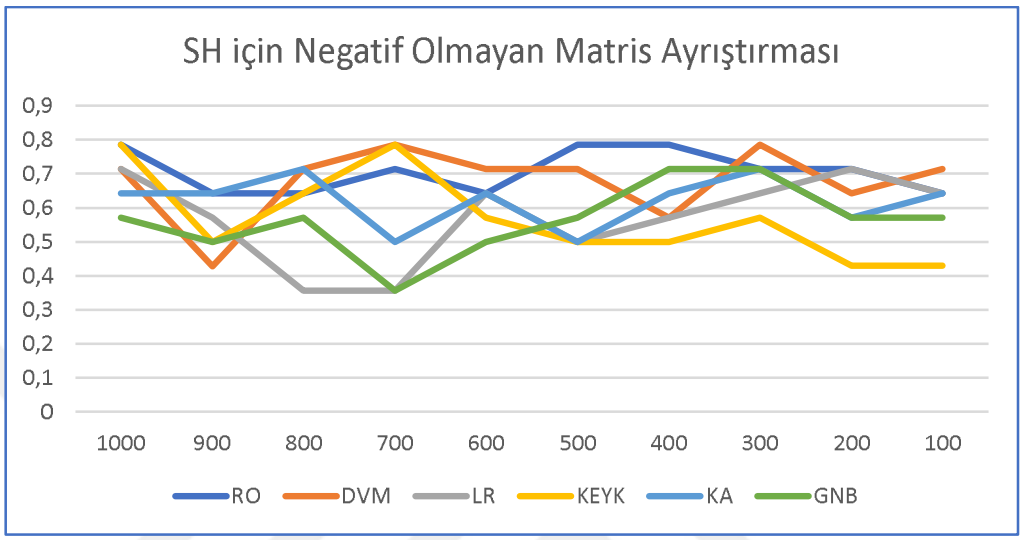
Obezite veri setinde öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.19'da verilmiştir.



Şekil 3.19. Obezite için NOMA Sonuçları

3.5.6. Sedef Hastalığı için NOMA Sonuçları

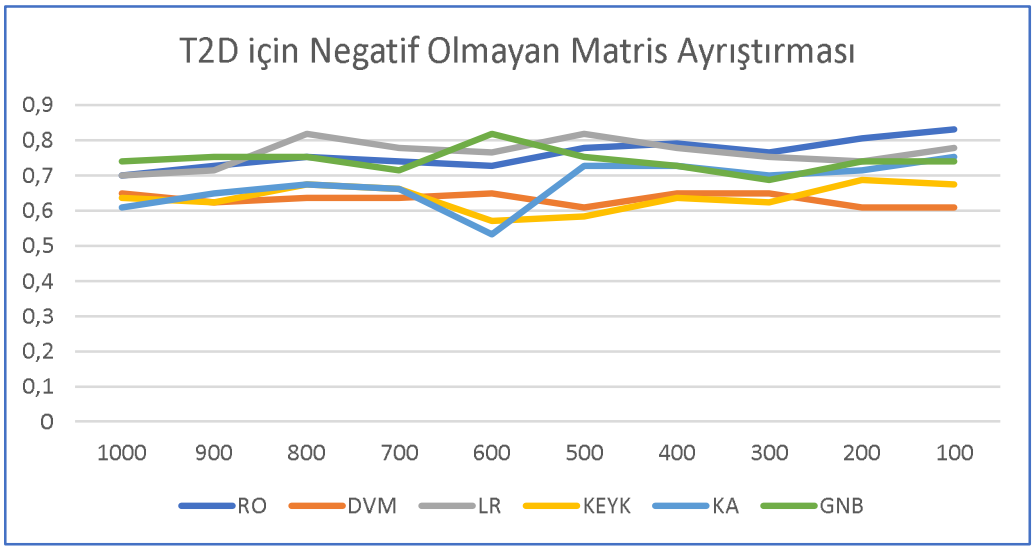
Sedef Hastalığı veri setinde öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.20'de verilmiştir.



Şekil 3.20. SH için NOMA Sonuçları

3.5.7. Tip-II-Diyabet için NOMA Sonuçları

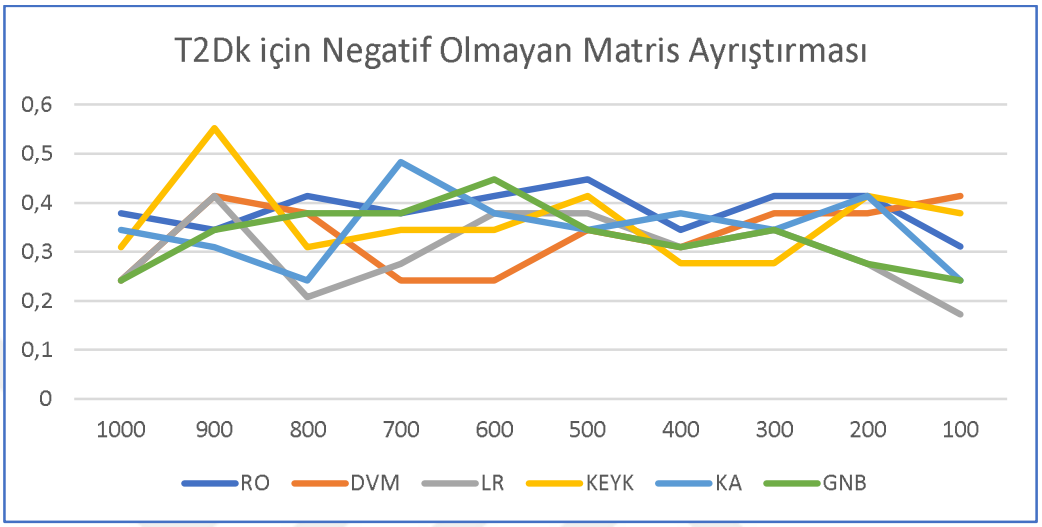
Tip-II diyabet veri setinde öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.21'de verilmiştir.



Şekil 3.21. T2D için NOMA Sonuçları

3.5.8. Kadınlarda Tip-II-Diyabet için NOMA Sonuçları

Kadınlarda tip-II diyabet veri setinde öznitelik sayısının 100-1000 arasında değerlere indirgenmesine göre değişen sınıflandırma sonuçları Şekil 3.22'de verilmiştir.



Şekil 3.22. T2Dk için NOMA Sonuçları

3.6. Tekil Değer Ayrışması Yöntemi için Sonuçlar

İltihaplı bağırsak hastalığı (İBH), siroz, kolon kanseri (KK), obezite, sedef hastalığı (SH), tip-II-diyabet (T2D) ve kadınlarda tip-II-diyabet (T2Dk) veri setlerine; TDA yöntemi uygulandığındaki sınıflandırma doğruluk değeri sonuçları ve indirgenmiş yeni matrislerin boyutları Tablo 3.4'te verilmiştir.

Tablo 3.4. TDA için Doğruluk Sonuçları

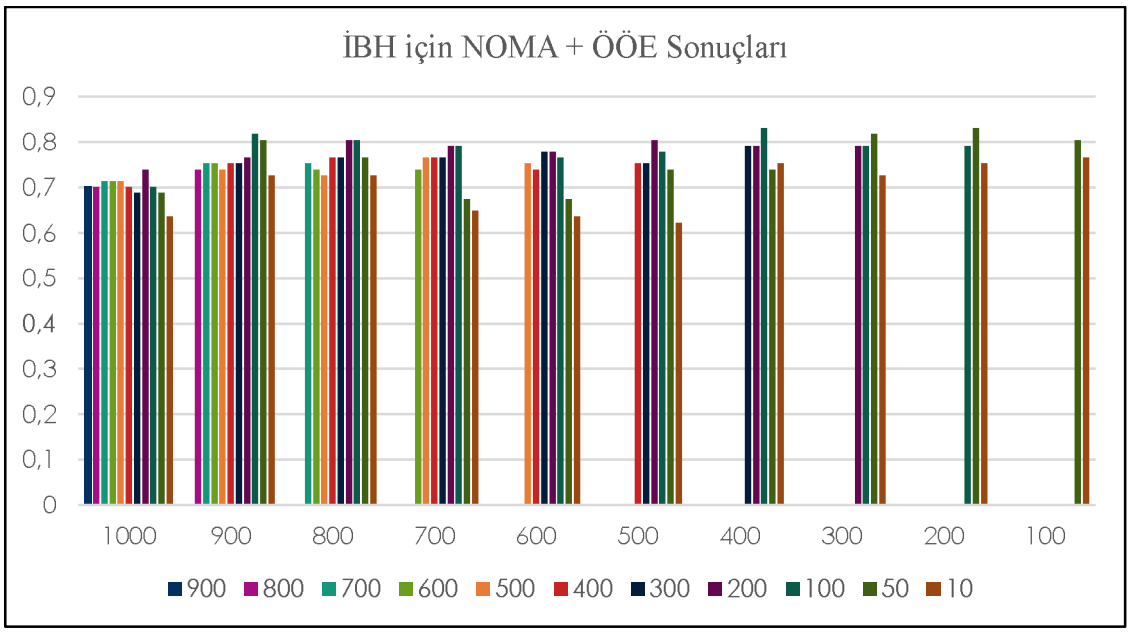
		Siroz	KK	Obezite	Sedef	T2D	T2Dk
	(382,382)	(232,232)	(134,134)	(278,278)	(70,70)	(290,290)	(145,145)
RO	0.636	0.702	0.444	0.607	0.571	0.759	0.311
DVM	0.610	0.425	0.371	0.607	0.429	0.569	0.414
LR	0.610	0.425	0.371	0.607	0.429	0.569	0.414
KEYK	0.415	0.532	0.333	0.571	0.571	0.483	0.276
KA	0.662	0.511	0.371	0.375	0.714	0.603	0.276
GNB	0.610	0.596	0.333	0.446	0.643	0.689	0.345

3.7. Negatif Olmayan Matris Ayrıştırması ve Özyineli Öznitelik Eleme Yöntemi için Sonuçlar

Bu bölümde, önerilen yöntem ile gerçekleştirilen öznitelik seçilimi ve sonucunda elde edilen Rastgele Orman (RO) sınıflandırma sonuçları sunulmaktadır. Sırasıyla önce NOMA yöntemi uygulanıp, öznitelik sayıları 100 ile 1000 arasında değerlere indirgenmiş veri setlerine, daha sonra ÖÖE yöntemi uygulanıp öznitelik sayıları 10 ile 900 arası değerlere indirgenmiştir. Bu şekilde ile NOMA farklı boyutlarda ara uzay boyutuna indirgeme yapılırken, ara uzayda ÖÖE ile ara öznitelik seçilimi yapılmış, daha sonra izdüşüm uzayında seçilen latent öznitelikler gözlemlenen uzaya yansıtılarak özellik seçilimi yapılmıştır.

3.7.1. İltihaplı Bağırsak Hastalığı için NOMA + ÖÖE Sonuçları

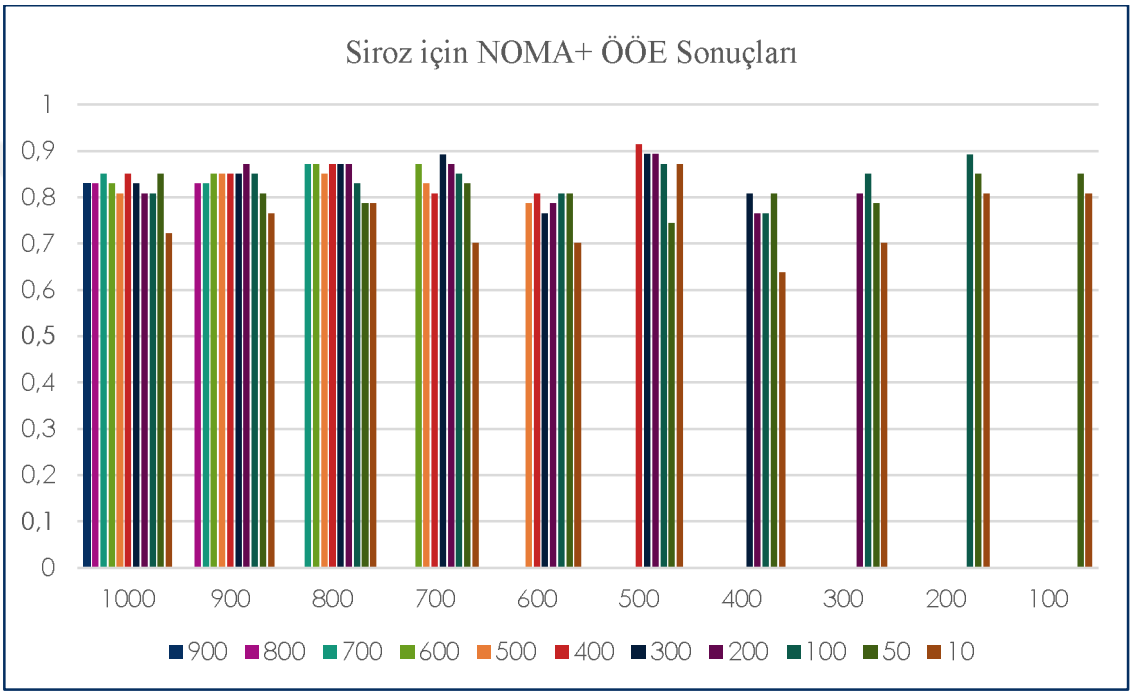
İltihaplı bağırsak hastalığı veri setinde çalışmada önerilen yöntem kullanılarak elde edilen sınıflandırma doğruluk değeri sonuçları Şekil 3.23'te verilmiştir. Her bir renk 100 ile 1000 arasında NOMA yöntemi ile indirgenmiş gizli uzay boyutlarını gösterirken aynı renkteki sonuçlar bu değerler üzerinden ÖÖE ile indirgenmiş ve geri yansıtılarak elde edilmiş öznitelik seçilimi sonuçları gösterir.



Şekil 3.23. İBH için Önerilen Yöntem Sonuçları

3.7.2. Siroz için NOMA + ÖÖE Sonuçları

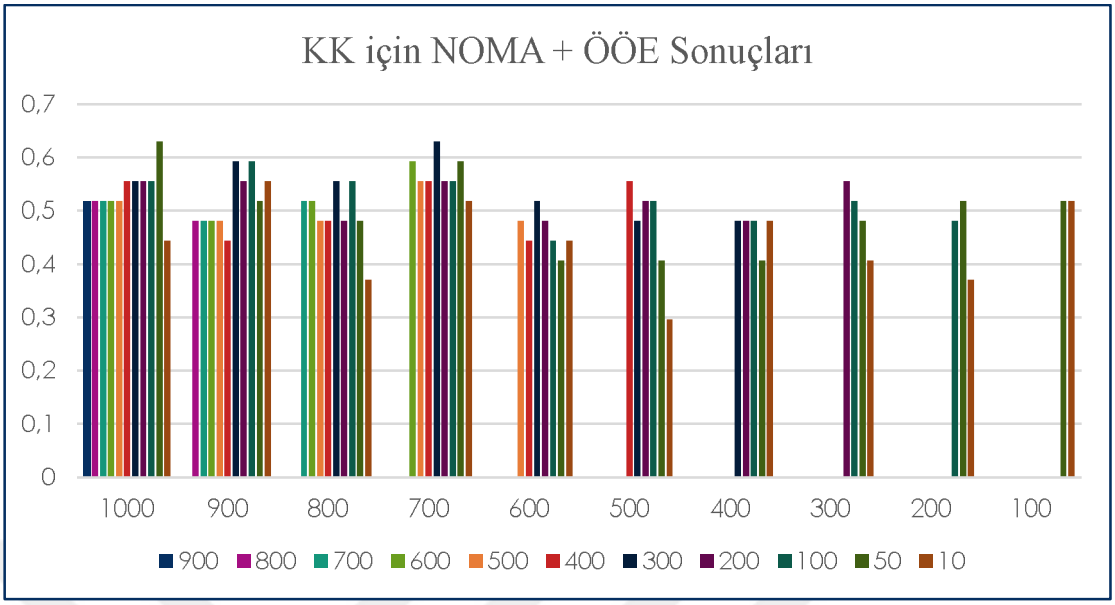
Siroz veri setinde önerilen yöntem kullanılarak elde edilen sınıflandırma doğruluk değeri sonuçları Şekil 3.24'te verilmiştir. Her bir renk 100 ile 1000 arasında NOMA yöntemi ile indirgenmiş gizli uzay boyutlarını gösterirken aynı renkteki sonuçlar bu değerler üzerinden ÖÖE ile indirgenmiş ve geri yansıtılarak elde edilmiş öznitelik seçilimi sonuçları gösterir.



Şekil 3.24. Siroz için Önerilen Yöntem Sonuçları

3.7.3. Kolon Kanseri için NOMA + ÖÖE Sonuçları

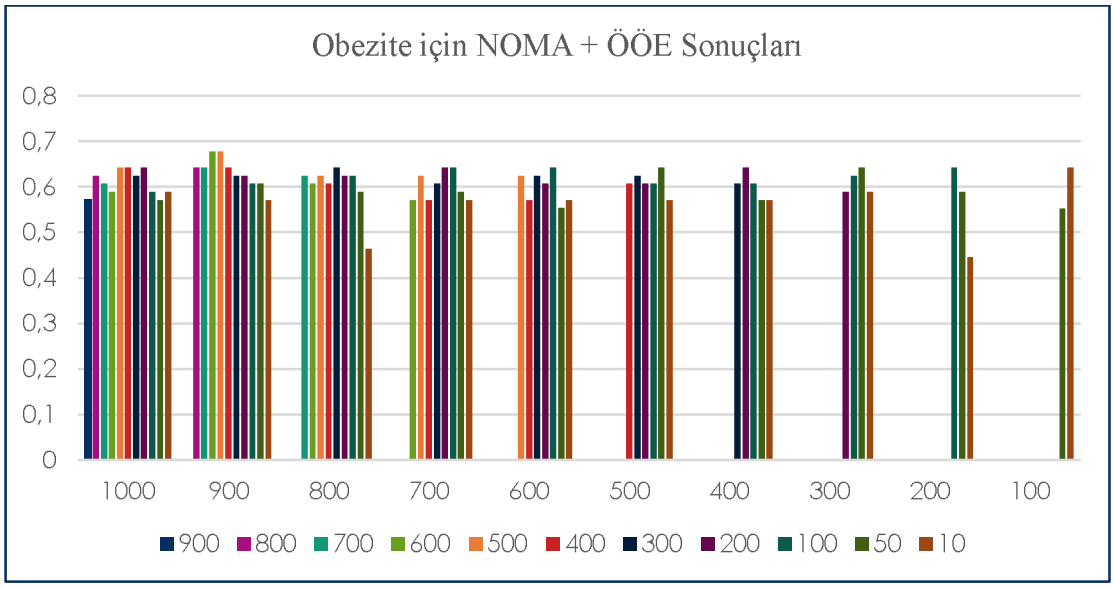
Kolon kanseri veri setinde 2 ayrı yöntem kullanılarak elde edilen sınıflandırma doğruluk değeri sonuçları Şekil 3.25'te verilmiştir. Her bir renk 100 ile 1000 arasında NOMA yöntemi ile indirgenmiş gizli uzay boyutlarını gösterirken aynı renkteki sonuçlar bu değerler üzerinden ÖÖE ile indirgenmiş ve geri yansıtılarak elde edilmiş öznitelik seçimi sonuçları gösterir.



Şekil 3.25. KK için Önerilen Yöntem Sonuçları

3.7.4. Obezite için NOMA + ÖÖE Sonuçları

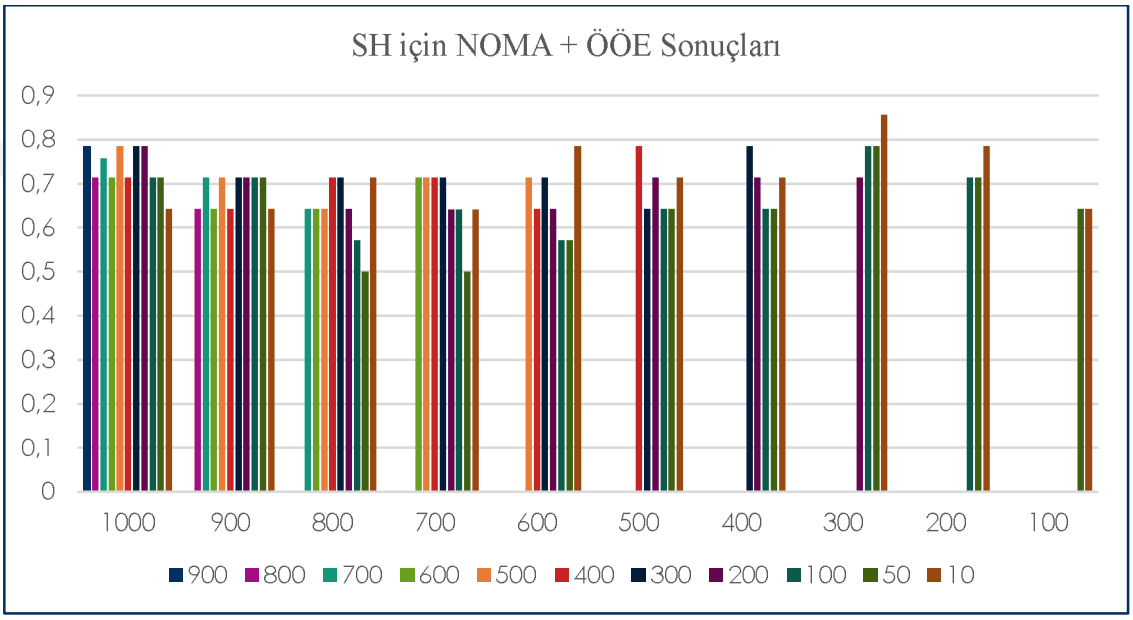
Obezite veri setinde 2 ayrı yöntem kullanılarak elde edilen sınıflandırma doğruluk değeri sonuçları Şekil 3.26'da verilmiştir. Her bir renk 100 ile 1000 arasında NOMA yöntemi ile indirgenmiş gizli uzay boyutlarını gösterirken aynı renkteki sonuçlar bu değerler üzerinden ÖÖE ile indirgenmiş ve geri yansıtılarak elde edilmiş öznitelik seçilimi sonuçları gösterir.



Şekil 3.26. Obezite için Önerilen Yöntem Sonuçları

3.7.5. Sedef Hastalığı için NOMA + ÖÖE Sonuçları

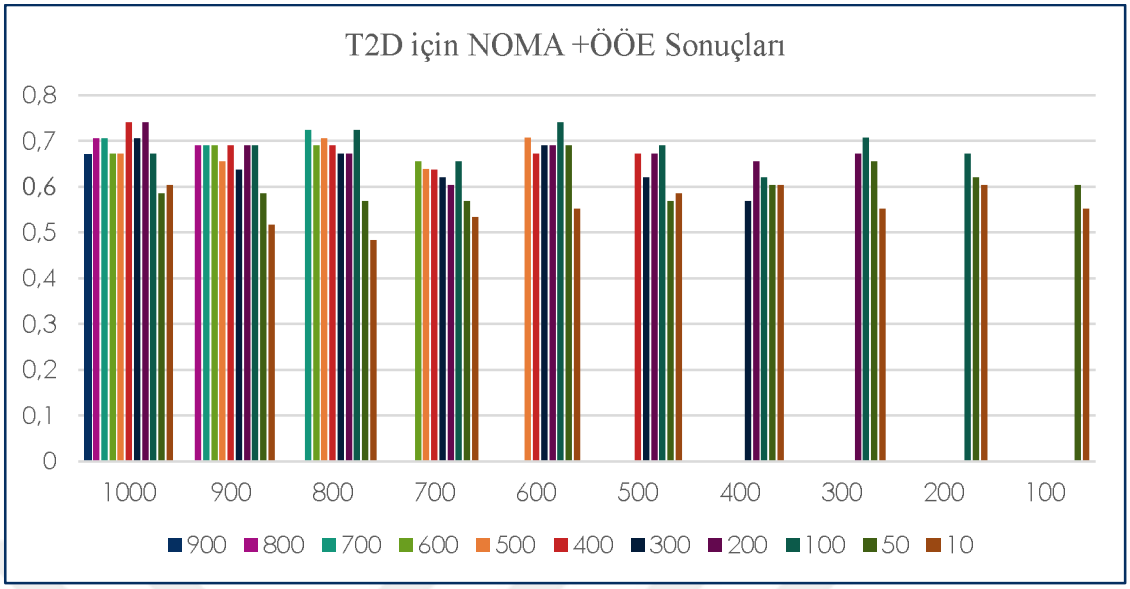
Sedef hastalığı veri setinde 2 ayrı yöntem kullanılarak elde edilen sınıflandırma doğruluk değeri sonuçları Şekil 3.27'de verilmiştir. Her bir renk 100 ile 1000 arasında NOMA yöntemi ile indirgenmiş gizli uzay boyutlarını gösterirken aynı renkteki sonuçlar bu değerler üzerinden ÖÖE ile indirgenmiş ve geri yansıtılarak elde edilmiş öznitelik seçilimi sonuçları gösterir.



Şekil 3.27. SH için Önerilen Yöntem Sonuçları

3.7.6. Tip-II-Diyabet için NOMA + ÖÖE Sonuçları

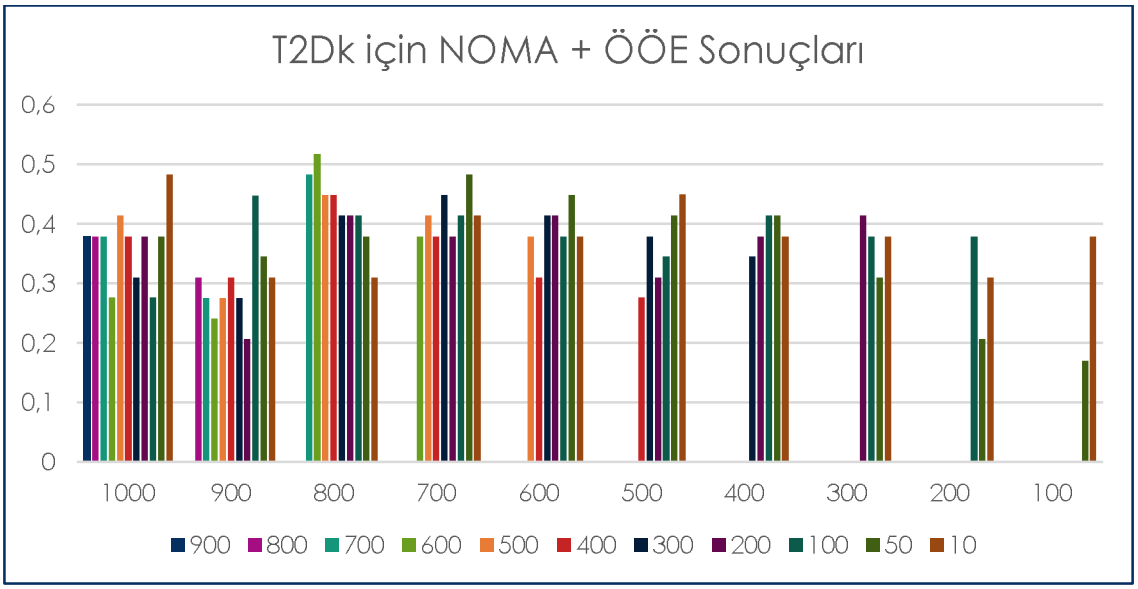
T2D veri setinde 2 ayrı yöntem kullanılarak elde edilen sınıflandırma doğruluk değeri sonuçları Şekil 3.28'de verilmiştir. Her bir renk 100 ile 1000 arasında NOMA yöntemi ile indirgenmiş gizli uzay boyutlarını gösterirken aynı renkteki sonuçlar bu değerler üzerinden ÖÖE ile indirgenmiş ve geri yansıtılarak elde edilmiş öznitelik seçilimi sonuçları gösterir.



Şekil 3.28. T2D için Önerilen Yöntem Sonuçları

3.7.7. Kadınlarda Tip-II-Diyabet için NOMA + ÖÖE Sonuçları

T2Dk veri setinde 2 ayrı yöntem kullanılarak elde edilen sınıflandırma doğruluk değeri sonuçları Şekil 3.29'da verilmiştir. Her bir renk 100 ile 1000 arasında NOMA yöntemi ile indirgenmiş gizli uzay boyutlarını gösterirken aynı renkteki sonuçlar bu değerler üzerinden ÖÖE ile indirgenmiş ve geri yansıtılarak elde edilmiş öznitelik seçilimi sonuçları gösterir.



Şekil 3.29. T2Dk için Önerilen Yöntem Sonuçları

3.8. Bulunan Biyobelirteç Adayları

Bölüm 3.6'da gösterilen NOMA + ÖÖE yöntemiyle; iltihaplı bağırsak hastalığı (İBH), siroz, kolon kanseri (KK), obezite, sedef hastalığı (SH), tip-II-diyabet (T2D) ve kadınlarda tip-II-diyabet (T2Dk) veri setleri içinden 10'ar biyobelirteç adayı tespit edilmiş ve bu biyobelirteçlerin tür isimleri Tablo 3.5'te verilmiştir.

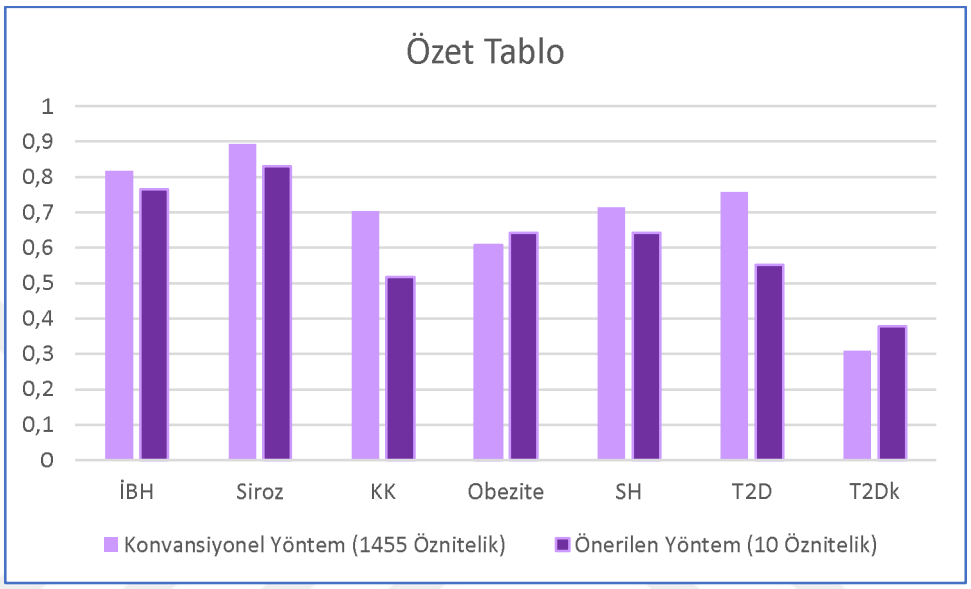
Tablo 3.5. Biyobelirteç Adaylarının Tür İsimleri

[redacted]	*Veillonella dispar*	*Alistipes indistinctus*	*Bacteroides coprocola*	*Bacteroides ovatus*	*Prevotella disiens*	*Ruminococcus torques*	*Lactobacillus rhamnosus*	*Veillonella parvula*	*Coprococcus eutactus*	*Veillonella sp3144*
KK	*Parabacteroides goldsteinii*	*Ruminococcus sp5139BFAA*	*Bacteroides coprophilus*	*Bifidobacterium longum*	*Bacteroides stercoris*	*Eubacterium eligens*	*Enterobacter cloacae*	*Bacteroides sp4347FAA*	*Neisseria unclassified*	*Ruminococcus gnavus*
[redacted]	*Anaerotruncus unclassified*	*Bacteroides uniformis*	*Bacteroides coprocola*	*Prevotella disiens*	*Ruminococcus sp5139BFAA*	*Enterobacter cloacae*	*Sutterella wadsworthensis*	*Eubacterium rectale*	*Paraprevotella clara*	*Ruminococcus torques*
[redacted]	*Clostridium perfringens*	*Bacteroides caccae*	*Eubacterium biforme*	*Bacteroides clarus*	*Odoribacter laneus*	*Eubacterium ramulus*	*Bacteroides uniformis*	*Bacteroides ovatus*	*Anaerotruncus unclassified*	*Bacteroides coprocola*
SEDEF	*Corynebacterium pyruvici produceens*	*Corynebacterium pseudodiphtheriticum*	*Escherichia coli*	*Staphylococcus capraecapitis*	*Corynebacterium striatum*	*Rothia mucilaginosa*	*Staphylococcus pseudintermedius*	*Lactococcus raffinolactis*	*Staphylococcus epidermidis*	*Acinetobacter radioresistens*
T2D	*Sutterella wadsworthensis*	*Parabacteroides goldsteinii*	*Atopobium rimae*	*Bifidobacterium angulatum*	*Clostridium citroniae*	*Ruminococcus sp5139BFAA*	*Clostridium asparagiforme*	*Coprococcus spART551*	*Blautia hydrogenotrophica*	*Oscillibacter unclassified*
T2Dk	*Streptococcus thermophllius*	*Faecalibacterium prausnitzii*	*Eubacterium eligens*	*Atopobium rimae*	*Ruminococcus obeum*	*Prevotella oralis*	*Alistipes onderdonkii*	*Streptococcus parasanguinis*	*Enterobacter cloacae*	*Succinivibrionaceae*

3.9. Önerilen Yöntemin Konvansiyonel Yöntemlerle Kıyaslanması

Şekil 3.30'da 1455 özniteliği olan İltihaplı bağırsak hastalığı (İBH), siroz, kolon kanseri (KK), obezite, sedef hastalığı (SH), tip-II-diyabet (T2D) ve kadınlarda tip-II-diyabet

(T2Dk) veri setlerinin rastgele orman sınıflandırma sonuçları ile önerilen yöntemle seçilmiş 10 öznitelik için rastgele orman sınıflandırma sonuçları karşılaştırılmıştır. Önerilen yaklaşım ile seçilen öznitelik sayısı oldukça düşük tutulsa da genel olarak başarılı sınıflandırma trendinin devam ettiği görülmüştür.



Şekil 3.30. Önerilen Yöntemin Konvansiyonel Yöntemlerle Karşılaştırılması

3.10. Yapay Sinir Ağları ile Elde Edilen Sonuçlar

Oluşturulan 3 model için de sinir ağlarında girdi olarak 7 ayrı veri seti kullanılmış (İltihaplı bağırsak hastalığı (İBH), siroz, kolon kanseri (KK), obezite, sedef hastalığı (SH), tip-II-diyabet (T2D) ve kadınlarda tip-II-diyabet (T2Dk)) ve test edilen doğruluk sonuçları Tablo 3.6'da verilmiştir. Bu sebeple modellerin giriş katmanında 1455 nöron bulunmaktadır. İlk model yalnızca sınıflandırma amaçlıdır. Model 2 için, giriş katmanından sonraki ikinci katmanda öznitelikler aynı boyutta bir ağırlık vektörü ile çarpılıp yine aynı boyutta bir vektör elde edilmiştir. Öznitelikler ikinci katmanda 0-1 arasındaki rastgele değerlerle çarpıldıktan sonra, eklenen akışkan RELUdan yalnızca katsayıları pozitif olan değerler doğrudan çıkarken, katsayıları negatif olanların ise 0 olarak çıkması ve bu sayede öznitelik seçiminin otomatik olarak yapılması amaçlanmıştır. 2.katmanda indirgenmiş özellik sayıları Tablo 3.6'da Model 2 kısmında verilmiştir.

Model 3'te ise Model 2'deki katmanlara ek olarak üçüncü bir katman bulunur. Bu katmanda akışkan RELUdan çıkan değerlerin L1 normu alınarak çıkışa ulaşan değerlerin

çoğunun 0 olması amaçlanmıştır. 3.katmandaki nöron sayıları Tablo 3.6'da Model 3 doğruluk sonuçlarının yanında verilmiştir.

Bu yaklaşım ile esasen bir önceki aşamada önerilen yöntemin türevlenebilir programlama ile otomatik olarak ve doğrusal olmayan bir yaklaşımla gerçekleştirilip gerçekleştirilemeyeceği sorusuna bir yanıt aranmıştır. YSA ilk katmanındaki ağırlıklı akışkan RELU katmanları öznitelik seçimini yaparken, ara katmanlardaki L1 normu çoğu sinir hücresini inaktif olmaya zorlayarak boyut indirgeme işlevini yerine getirmeyi otomatikleştirmeye çalışmaktadır.

Tablo 3.6. YSA ile Elde Edilen Sonuçlar

	Model 1	Model 2	Model 3
İBH	0.8051948	0.5974026 (721)	0.5194805 (726)
Siroz	0.80851066	0.44680852 (739)	0.46808519 (746)
KK	0.6296296	0.7037037 (750)	0.6296296 (762)
Obezite	0.53571427	0.58928573 (726)	0.58928573 (736)
SH	0.78571427	0.42857143 (735)	0.42857143 (752)
T2D	0.6034483	0.5689655 (717)	0.51724136 (709)
T2Dk	0.5862069	0.6551724 (683)	0.5862069 (717)

4. BÖLÜM
TARTIŞMA- SONUÇ VE ÖNERİLER

Yürütülen çalışma ile, yüksek boyutlu, görece düşük sayıda örnek noktasına sahip sınıflandırma problemlerinde öznitelik seçimini sağlayan yeni bir biyoinformatik yaklaşım önerilmiş ve bu yaklaşım hastalık mikrobiyotası verileri üzerinde test edilerek konvansiyonel yaklaşımlara oranla hangi başarımda olduğu incelenmiştir. Yaklaşımın temel rasyonel motivasyonu, gözlenebilir değişken sayısının yüksek sayıda olmasına rağmen, birçok veri sınıfında olduğu gibi bu değişkenlerin büyük varyasyonunun öznitelik uzayının içerisinde bir alt-uzayda yerleşik olduğu ve esas özniteliklerin bu alt uzaylarda gömülü olduğu gözleminden yola çıkmaktadır. Önerilen yöntem öncelikle gözlenen veriyi negatif olmayan matris ayrıştırması yaklaşımıyla bir alt uzayda temsil etmekte, daha sonra bu uzayda bir sarıcı ile sınıflandırarak özyineli öznitelik seçilimi sağlamaktadır. Gözlem uzayından indirgenmiş öznitelik uzayına en az kayıpla izdüşümü sağlayan öznitelik seçimi sonrasında ise gizli öznitelikleri destekleyen dayanıklı gözlemlenebilir özniteliklerin belirlenmesi sağlanmaktadır.

Yedi adet hastalık mikrobiyota veri seti üzerinde yürütülen deneylerde geliştirilen yaklaşımın konvansiyonel öznitelik seçilim algoritmaları ile rekabet edebilir seviyede sonuçlar ürettiği görülmüştür. Her ne kadar yöntem genel hatlarıyla gizli öznitelik uzayına izdüşüm için herhangi bir boyut indirgeme algoritması ve sınıflandırma için herhangi bir sınıflandırma algoritması ile yürütülebilecek jenerik bir çerçeve olarak belirlense de, yürütülen testlerde NOMA diğer boyut indirgeme yaklaşımlarına (Örn: Tekil değer ayrıştırması), RO ise bulgularda sonuçları verilen diğer sınıflandırma algoritmalarına oranla daha başarılı görülmüş ve önerilen algoritmanın varsayılan yöntemleri olarak kullanılmıştır. Bu yaklaşım ile seçilen öznitelik sayısı oldukça düşük değerlere taşınsa da (Örn: 10 öznitelik, 50 öznitelik) başarılı sınıflandırma trendinin devam ettiği görülmüştür. Bu sonuç, bir yönüyle pratikte düşük sayıda öznitelik ile başarılı sınıflandırma yapılabilmenin önemli olduğu durumlar için değerli bir çıktı

oluşturmaktadır. Örneğin, deneyleri yürütülen hastalık mikrobiyotası verisi içerisinden az sayıda takson ile başarılı sınıflandırma yapılabilmesi, biyobelirteç probleminin bir çözümüne denk gelmektedir. Buna göre, hastalık teşhis ve tedavisinde kullanılabilecek biyolojik ünitelerin keşfedilmesi, hastalık sınıflandırmada öznitelik olabilen türlerin belirlenmesine karşılık gelmektedir ki, önerilen yaklaşımın buna bir çözüm üretebildiği görülmektedir. Öte yandan, önerilen yaklaşım, konvansiyonel öznitelik seçiminden bağımsız olarak veriyi gizli uzayda da temsil edebilecek ana hatlardan oluşturulmaktadır. Bu yeteneğin biyolojik olarak karşılığı, fenotipin sebebiyet verdiği genel trend değişimini modelleyebiliyor olmasıdır. Örneğin, hastalık mikrobiyotasında görülen disbiyozis fenomeni, mikrobiyota komposizyonunun hastalık ile beraber homeostatik dengenin kaybedilmesi sonucu genel bir yapısal değişime maruz kalmasıdır. Bu ise genel hatlarıyla mikrobiyota profilinin değişime uğraması ve farklı bir rejime girmesi ile açıklanmaktadır. Bu, veri bilimi açısından gizli özniteliklerin değerlerinin değişimi olarak veriye yansıyacaktır. Önerilen yöntem, gizli özniteliklerin trend değişikliğini en az hatayla regrese edebilen özniteliklerin seçimine dayanmaktadır. Bu sebeple, seçilen az sayıda gözlemlenebilir özniteliğin ana trend değişikliğini yansıtan anahtar türler olduğu ve bu bulgunun biyobelirteç seçiminde de önemli elemanlara denk gelebileceği öne sürülebilir. Nitekim yürütülen deneylerde ortaya çıkarılan özniteliklerin önemli biyobelirteçlere denk geldiğini gözlemlemek mümkündür.

Önerilen öznitelik seçimi algoritması kullanılarak keşfedilen biyobelirteçler incelendiğinde her mikrobiyota ilişkili hastalık verisi için rapor edilen 10 biyobelirtecin hastalık için ne derece sağlıklı elemanlar olduğunu ortaya koymak amacıyla söz konusu taksonomik üniteler için bir literatür taraması yürütülmüştür. Bu çalışmada tespit edilen her biyobelirtecin ilişkilendirdiği hastalığın tespiti veya patogenezinde ilişkili olan taksonlar olduğu görülmüştür.

Ayrıntılı olarak Siroz hastalığıyla ilişkili olduğu tespit edilen biyobelirteçlerden *Veillonella,* siroz hastalığında bağırsak mikrobiyotasındaki bağıl bolluğunun arttığı daha önceden tespit edilmiş olan bir cinstir [70]. Öte yandan *Alistipes*, *Bacteroides* ve *Prevotella* cinslerinin mikrobiyota temelli siroz tedavisinde son yıllarda aday biyobelirteçler olarak öne çıktığı Horvath vd. tarafından rapor edilmiştir [71]. Bajaj vd. ise çalışmalarında *Ruminococcus* cinsinin sirozla ilişkili olduğunu gerçek zamanlı polimeraz zincir reaksiyonu yöntemiyle tespit etmiştir [72]. *Lactobacillus* üzerinde

yürütülen hayvan deneylerinde karaciğer hastalığına sahip farelerde bu cinsin hepatik fibrozise engel olarak hastalığı gerilettiği görülmektedir [73]. Bunun yanında üzerinde testlerin gerçekleştirildiği siroz mikrobiyomu çalışmasının sonucu olarak yayınlanan makalede önerilen teknikle keşfedilen *Coprococcus* cinsinin siroz ile ilişkili olduğu raporlanmıştır [74].

Kolon kanseri mikrobiyotası verisi üzerinden yapılan analizde elde edilen 10 biyobelirteç ile ilgili şu kanser bağıntıları görülmüştür. *Parabacteroides* ve *Bacteroides* cinslerinin IL-1 reseptörleri ile ilişkili kinaz proteinlerini etkileyerek kanser oluşumunda rol alabileceği fare deneyleri ile Klimesova vd. Tarafından gösterilmiştir [75]. Lucas vd. *Ruminnococcus* türlerinin enflamasyonla ve kanserle ilişkisini incelemiştir [76]. *Bifidobacterium* cinsine ait bakterilerin yokluklarında kalın bağırsak kanserinin daha yüksek olasılıkla görüldüğü ve bunun bütirik asit üretimi ile ilişkisi Rivière vd. tarafından raporlanmıştır [77]. Tespit edilen cinslerden bir diğeri olan *Eubacterium* kolon kanseri disbiyosizinde öne çıkan cinslerden biri olarak gözlemlenmiş olsa da [78], doğrudan kanser patogenezi veya prognozu ile doğrudan ilişkisini gösterir bir çalışma tarih itibariyle bilinmemektedir. Yurdakul vd. bir süre önce *Enterobacter* türlerinin kolon kanserini tetikleyebileceğini göstermiştir [79]. Önerilen yöntemin keşfettiği biyobelirteçlerden bir diğeri olan *Neisseria* türlerinin ise mukosal epitel hücrelerine hücum ederek hücreler arası tutunum molekülü (CAM-1) ifadesini artırdığı ve kolon kanseri oluşumuna elverişli bir ortam oluşturduğu henüz mikrobiyota çalışmalarının yaygınlaşmadığı bir dönemde Jarvis vd. tarafından gösterilmiştir [80]. Buradan yola çıkarak önerilen yöntemin kolon kanseri ile farklı yolaklardan pozitif ve negatif ilişkili bakteriyel türleri keşfedebilecek yetenekte olduğu sonucuna varılabilir.

Obez bireylere ait mikrobiyota profilleri incelenerek önerilen biyobelirteçler için obezite ile ilişkili mikrobiyom literatürüne bakıldığında yine ilginç keşifler göze çarpmaktadır. *Anaerotruncus massiliensis* türü son yıllarda bariatrik cerrahi müdahale geçirmiş hastaların bağırsaklarında tespit edilen ve obezite ile ilişkilendirilen bir tür iken [81], bu tür daha önceki çalışmalarda obezite ile ilişkili bir biyobelirteç olarak keşfedilmemiştir. Oysa ki bu tezde üzerinde çalışılmış olan obezite metagenomları İnsan Mikrobiyom Projesi ile elde edilmiş olan verilerden oluşmaktadır. Önerilen yöntem bu son yıllarda izole edilerek varlığı keşfedilen türü doğrudan öznitelik seçilimi ile keşfetmeyi başarmıştır. Yöntem tarafından obezite ilişkili diğer bir biyobelirteç olan *Bacteroides*

uniformis türünün yüksek yağlı diyete maruz kalarak obeziteye sürüklenmeye çalışılan fare modellerinde metabolik ve immünolojik aksaklıkları giderdiği ve bu türe ait hayvanlarda obezitenin ilerlemediği bir süre önce Olds [82] tarafından keşfedilmiştir. Öte yandan aynı cinsin bir başka türü olan *Bacteroides coprocola* türünün obezite ile ilişkili olduğu incelenen verinin raporlandığı ilk çalışmada dahi ortaya konmuş idi [83]. Önerilen yöntem aynı cinse ait bu iki türün de hastalıkla ilişkisini ortaya koyabilmiştir. Öte yandan diyet ve yaşam tarzı ile bağıl bolluğu ilişkilendirilebilen bir diğer cins *Prevotella* [84] da öne çıkan bir diğer öznitelik olarak görünmektedir. *Ruminnococcus* ise diğer çalışmalarda da obez ve obez olmayan bireyler arasında mikrobiyota kompozisyonu ayrımına yol açan cinsler arasındadır [85]. Fırsatçı patojenler olan *Enterobacter*lerin insanlardan izole edilmiş suşlarının steril farelerde obeziteye sebep olduğu bir süre önce raporlanmıştır [86]. Bu türler de önerilen algoritma tarafından yine bir obezite biyobelirteci olarak ortaya çıkmaktadır. Bu cins üzerinde yürütülen literatür çalışmasında ise metagenom çalışmalarında söz konusu cinslerin biyobelirteç olarak sıkça raporlanmadığı görülmektedir. Önerilen yaklaşımın, oldukça jenerik bir obezite veri seti (İnsan Mikrobiyom Projesi metagenom verisi) kullanmasına rağmen bu ajanı keşfedebilmiş olması, yöntemin biyobelirteç seçimi yönünden hedeflenen seçicilik özelliğine kavuşabildiğini gösteren bir bulgu olarak ele alınabilir. Keşfedilen diğer biyobelirteçlerden *Sutterella* cinsinin yine obezite ile ilişkili olduğu Hou vd. [87] tarafından kısa bir süre önce gösterilmiştir. Bunlarla birlikte yakın zamanda önerilen algoritma tarafından keşfedilen *Eubacterium rectale* [88] ve *Paraprevotella* [89] taksonlarının da obezite ile ilişkili oldukları raporlanmıştır. Bu bulgulardan yola çıkarak önerilen yaklaşımın geniş spektrumda obezite biyobelirteçlerini keşfedebildiği öne sürülebilir. Öte yandan obezite sınıflandırma başarısının yüksek olmamasına karşın keşfedilen her bir taksonun obezite sistematiği ile ilişkili olması, obezite disbiyozisinin karmaşık bir sistem olduğu ve az sayıda biyobelirteç ile değil, geniş spektrumlu bir sistem tarafından temsil edilebileceği görüşünü desteklemektedir.

İltihaplı bağırsak hastalığı, geniş bir disfonksiyon sınıfı olması ve birden fazla sınıfa giren hastalığı (Örn: Crohn Hastalığı, Ülseratif kolit) barındırmasına rağmen keşfedilen belirteçlerin filogenetik olarak benzer türlerde toplandığı görülmüştür. Bunların içinden *Clostridium* cinsinin bağırsak hastalıklarıyla doğrudan ilişkili olduğu uzun süredir bilinmektedir [90]. Bir süre önce Zhou ve Zhi tarafından yürütülen meta çalışmada *Bacteriodes* türlerine ait bakterilerin azlığının iltihaplı bağırsak hastalığı riskini artırdığı,

dolayısıyla bu türlerin bağırsağı koruyucu etkilerinin olabileceği öne sürülmüştür [91]. Keşfedilen biyobelirteçler arasında bulunan *Eubacterium* cinsinin ise yine kolit barındıran bağırsak hastalıkları ve enfeksiyon ile ilişkili olduğu metagenomik çalışmalarla gösterilmektedir [92]. Yine enflamasyon ve bağırsak dokusundaki baloncuklanma ile ilişkilendirilen *Anaerotruncus* türleri [93] keşfi yapılan bir diğer belirteç olarak ortaya çıkmaktadır. Önerilen algoritmanın hem bağırsağı koruma hem de bağırsakta enflamasyona sebebiyet verme potansiyeli olan elemanları keşfetmesi ve bunu bir cinse ait birden fazla türle ilişkilendirme yaparak gerçekleştirmesi yine yöntemin geniş metabolik bir bantta arama yapabildiğini göstermektedir.

Diğer metagenom verisi ile ilişkilendirilen hastalıklardan farklı olarak sedef hastalığı bağırsak değil deri üzerinden alınan mikrobiyom örnekleri ile elde edilmiştir. Dolayısı ile bulunan biyobelirteçler de deri florasına ait sedef ile ilişkili olabilecek taksonları ortaya koymaktadır. Drago vd. birinci derece kuzenlerden oluşan bir kohort üzerinde yürüttükleri mikrobiyal çalışmada *Corynebacterium* türlerinin sedef ile ilişkili olduğunu göstermiştir [94]. Bu çalışmada önerilen algoritma da *Corynebacterium* türlerini sedef ile ilişkili biyobelirteç olarak öne sürmektedir., Öte yandan deri florasının elemanları olarak görülebilen *E. coli*, *Staphylococcus* ve *Acinetobacter* türlerinin hastalık ile ilişkili olduğu algoritma tarafından öne sürülürken, bu türlerin sedef ile ilgili olabilecekleri literatürde de raporlanmaktadır [95]. Bir diğer biyobelirteç adayı olan *Rothia* türlerinin de diğer cilt hastalıklarına oranla sedef hastalığında spesifik olarak bağıl bolluğu artan bir tür olduğu yine bilinmektedir [96]. Diğer hastalık metagenomlarında görülen fenomene benzer şekilde sedef hastalığı için de önerilen biyobelirteçler içerisinde birbirlerine filogenetik olarak yakın türlerin olduğu görülmektedir. Konvansiyonel öznitelik seçimi algoritmalarından birçoğu korele türleri bir temsilci ile seçmeye meyletmektedir. Özellikle sınıflandırma başarısı yüksek olan sarmalama teknikleri ve bu çalışmada da kullanılan özyineli öznitelik eleme algoritması, birbiri ile korele elemanlardan birini seçmişken diğerini elemeye daha yatkındır. Ancak önerilen algoritmanın bu hassasiyeti koruyabildiği ve birbirine yakın türleri de biyobelirteç kümesinin içerisinde elemeden barındırabildiği görülmektedir.

Tip-II diyabet metagenomu sınıflandırma başarısı yüksek olmayan ve genellikle biyobelirteç seçiliminde problemlerle karşılaşılan kompleks bir hastalık metagenomu olarak bilinmektedir. Önerilen öznitelik seçimi yaklaşımının geniş bir etki kümesinden

farklı metabolizmalara ait biyobelirteçler seçebildiği görülmüştür. Spesifik olarak önerilen *Sutterella* fırsatçı patojen cinsinin insülin direnci ile ilişkili metabolik faaliyetlerde bulunduğu Moreno-Indias vd. tarafından morbid bir kohort üzerinde gösterilmiştir [97]. Geliştirilen algoritma tarafından önerilen *Parabacteriodes goldsteinii* türünün polisakkarit metabolizmasında dominant bir rol oynadığı geçtiğimiz birkaç yıllık periyotta keşfedilen bir bulgudur [98]. *Atopobium* cinsi hem oral florada hem de bağırsak mikrobiyomunda görülen komensal bir bakteri cinsidir. Bu cinsin oral floradaki bağıl bolluğu diyabet ile ilişkilendirilmişken [99], bağırsak florasındaki varlığı ile hastalık arasında bir ilişki bilinen literatürde bulunamamıştır. Önerilen algoritma, hastalıkla ilişkisi olan bu cinsin bağırsaktaki bolluğunun da diyabet ile ilişkili olduğunu iddia etmektedir. Bu bulgu, daha ileri çalışmaların tasarlanması açısından önemli bir çıktı olarak değerlendirilebilir. Probiyotik bir bakteri cinsi olan *Bifidobacterium* algoritma tarafından diyabet ile ilişkili olabileceği öne sürülen bir diğer elemandır. *Bifidobacterium* cinslerinin diyabet tedavisinde probiyotik ajanlar olarak kullanılabileceğine dair çalışmaların kısa bir süre öncesinde başladığı bilinmektedir [100]. Tip 2 diyabete sahip bireylerde bu türlerin daha az görülmesi, bu bakterilerin diyabet etkilerinden koruyucu metabolik faaliyetlerinden mahrum kalmalarıyla açıklanabilir. *Clostridium* ve *Coprococcus* cinsleri daha önceden diyabet ile ilişkisi olduğu bilinen bakteriler olmakla beraber önerilen algoritmanın da keşfedebildiği biyobelirteçlerdir [101]. Metabolik dengesizlik ve IFN-γ seviyeleri ile ilişkili olduğu hayvan deneyleri ile kontrollü olarak ispatlanan *Ruminnococcus* bakterileri biyobelirteç adayı olarak belirtilen bir diğer özniteliktir [102]. Bunun yanında *Blautia* literatürde doğrudan tip-2 diyabet ile ilişkilendirilmemiş olsa da glukoz toleransı ile ilgili olduğu Egshatyan vd. tarafından gösterilmiştir [103]. Önerilen yöntem, bu cinsi de tip-2 diyabetle ilişkili bir bakteri cinsi olarak öne sürmektedir. Benzer şekilde *Oscillibacter* cinsinin bağırsak bariyeri fonksiyonlarıyla ilişkili olduğu ve bu türlerin fazlalığının bağırsak geçirgenliğini artırarak tip-2 diyabeti olumsuz yönde etkileyebileceği belirtilmektedir. Nitekim diyabet modeli farelerde yürütülen çalışmalarda vildagliptin tedavisi gören farelerde bu türün birçok parametreyle korelasyon halinde olduğu ve hastalık tedavisinde etkin rol oynayabildiği öne sürülmüştür [104]. Tüm bu ilişkiler göz önüne alındığında, öznitelik seçimi algoritmasının geniş bir etki alanında ve farklı çalışmalarda keşfedilmiş olan pozitif ve negatif biyobelirteçleri toplu olarak keşfedebildiği görülmektedir.

Önerilen yaklaşımda, NOMA yöntemi ile boyut indirgendiğinde, geri kotarım hatasının gizli öznitelik sayısı azaldıkça arttığı gözlenmektedir. Buna göre daha düşük boyutlu bir uzaya indirgeme yapılıp eski uzaya geri dönüşüm yapıldığında, geri çağrılan mikrobiyota profilinin karesel hata cinsinden orijinal profille karşılaştırıldığında hatalı olduğu ve bu hatanın gizli boyut sayısı azaldıkça arttığı görülmüştür. Buradan, mikrobiyota bilgisini içeren alt uzayın çok düşük boyutlu olmadığı sonucu öne sürülebilir. Öte yandan NOMA, doğrusal bir boyut indirgeme paradigması sunduğundan asıl manifoldun doğrusal olarak haritalanamayan bir alt uzaya gömülmesi ve bunun doğrusal bir izdüşüm ile ortaya çıkarılamaması da olası bir senaryo olarak öne sürülebilir. Sonuç itibariyle, indirgenen alt uzay asıl profili temsil etmekten uzaklaşsa da aynı hata trendinin mikrobiyota sınıflandırmada ortaya çıkmadığı ve düşük boyutlarda da yüksek boyutlarla karşılaştırılabilir sınıflandırma başarılarına erişildiği görülmüştür. Bunun sebebinin denetimsiz öğrenmeye ait kalite metriklerinin denetimli öğrenme ile örtüşmemesi, dolayısıyla tüm profilin başarılı bir şekilde temsil edilememesinin fenotip sınıflandırması kalitesine doğrudan yansıyacak bir ölçüt olmaması olduğu düşünülebilir.

Geliştirilen yaklaşım doğrusal cebire dayalı boyut indirgeme yaklaşımları, açgözlü öznitelik seçimi algoritmaları ve klasik sınıflandırma algoritmalarından oluşan aşamalı bir iş akışı ile gerçeklenmektedir. Bu yaklaşımın yerine uçtan-uca türevsel programlama ile tüm aşamaların bütünleşik olarak gerçekleştirilebileceği bir derin öğrenme yaklaşımı denenmiştir. Öznitelik seçimi ve gizli uzaya indirgeme, seçilim ve 1-normu düzenlemesiyle bir ileri beslemeli yapay sinir ağına gömülmüştür. Ancak bu yaklaşım ile elde edilen ilk sonuçlar, önerilen el ile tasarlanmış algoritmaya oranla başarısız sonuçlar vermiştir. Literatürde oto kodlayıcılar kullanılarak gen ifadesi imzaları seçilimi için oluşturulmuş yaklaşımlar mevcuttur. Dolayısıyla her ne kadar doğrudan yapay sinir ağına implementasyon buna elvermemiş olsa da geliştirilen yaklaşımın uçtan-uca bir derin öğrenme yaklaşımı ile gerçeklenmesi olasıdır. İleride yürütülebilecek bu tip yaklaşımlar hem süreci bir boru hattı olmaktan çıkarıp otomatize ve genel-geçer bir yaklaşım haline getirebilecek, hem de mevcut hali ile doğrusal çözümlerin sunulduğu kısımlarda doğrusal-olmayan sonuçlara erişebilerek başarım gücünü potansiyel olarak yukarıya taşıma imkânı oluşturacaktır.

KAYNAKLAR

1. García-Torres M., Gómez-Vela F., Melián-Batista B., Moreno-Vega J.M., 2016. High-dimensional feature selection via feature grouping: A variable neighborhood search approach. **Information Sciences, 326**:102–118.
2. Lewis P., 1962. The characteristic selection problem in recognition systems. **IRE Transactions on Information Theory, 8**(2): 171–178.
3. Liu H., Motada H., 1998. Feature Selection for Knowledge Discovery and Data Mining, Kluwer Academic Publishers. Springer US, 214 pp.
4. Liu H., Motada H., 2002. On issues of instance selection. **Data Mining and Knowledge Discovery, 6**(2): 115–130.
5. Hu L., Gao W., Zhao K., Zhang P., 2018. Feature selection considering two types of feature relevancy and feature interdependency. **Expert Systems With Applications 93**: 423–434.
6. Zou Q., 2016. A novel features ranking metric with application to scalable visual and bioinformatics data classification. **Neurocomputing, 173**: 346–354.
7. Saeys Y., Inza I., Larrañaga P., 2007. A review of feature selection techniques in bioinformatics. **Bioinformatics, 23**(19): 2507–2517.
8. Cardenas E, Tiedje J.M., 2008. New tools for discovering and characterizing microbial diversity. **Current Opinion in Biotechnology, 19** (6): 544-549.
9. Ma S., and Huang J., 2008. Penalised feature selection and classification in bioinformatics. **Briefings in Bioinformatics, 9**(5): 392-403.
10. Dewhirst F.E., Chen T., Izard J, Paster B.J., Tanner A.C.R., Yu W.H., Lakshmanan A., Wade W.G., 2010. The Human Oral Microbiome. **J. Bacteriol, 192**(19): 5002-5017.
11. Bolouri H., 2014. Modeling Genomic Regulatory Networks With Big Data. **Trends In Genetics, 30**(5): 182-191.
12. Polat M., Karahan A. G., 2009. Multidisipliner yeni bir bilim dalı: biyoinformatik ve tıpta uygulamaları. **Süleyman Demirel Üniversitesi Tıp Fakültesi Dergisi, 16**(3): 41-50.
13. Mayo Clinic, 2013. Alzheimer's Disease: Tests and diagnosis. (Web Page: **https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/diagnosis-treatment/drc-20350453**), (Date Access: December 2018).

14. Lemos L. N., Morais D. N., Tsai S. M., Roesch L., Pylro V., 2017. Bioinformatics for Microbiome Research: Concepts, Strategies, and Advances. **The Brazilian Microbiome,** 111-123.
15. Huerta E., Duval B., Hao J., 2010. A Hybrid LDA and genetic algorithm for gene selection and classification of microarray data. **Neurocomputing, 73**: 2375-2383.
16. Kaya M., 2014. Gen İfade Verilerine Öznitelik Seçimi Ve Sınıflandırma. Gazi Üniversitesi, Yüksek Lisans, Ankara, 97 s.
17. Hanczar B., Courtine M., Benis A., Henegar C., Clément K., Zucker J.D., 2003. Improving classification of microarray data using prototype-based feature selection. **ACM SIGKDD Explorations Newsletter, 5**(2), 23-30.
18. Gilhan K., Yeonjoo K., Heuiseok L., Hyeoncheol K., 2010. An MLPbased feature subset selection for HIV-1 protease cleavage site analysis. **Artificial intelligence in medicine, 48**(2), 83-89.
19. Rours E. M., 1980. A combined nonparametric approach to Feature Selection and Binary Decision Tree design. **Pattern Recoqnition, 12**: 313-317.
20. Guyon I., Weston J., Barnhill S., Vapnik V., 2002. Gene Selection for cancer classification using support vector machines. **Machine Learning, 46**(1-3): 389–422.
21. Hughes, G.F. ,1968. On the mean accuracy of statistical pattern recognizers. **IEEE Transactions on Information Theory Archive, 14** (1): 55-63.
22. Sánchez-Maroño N, Alonso-Betanzos A., Tombilla-Sanromán M., 2007. Filter methods for feature selection- A comparative study, Intelligent Data Engineering and Automated Learning - IDEAL 2007: 8th International Conference, Birmingham, UK, 178-187.
23. Yu L., Ye J., 2007, Dimensionality Reduction for Data Mining-Techniques, Applications and Trends, (Web Page: **http://www.cs.binghamton.edu/~lyu/SDM07/DR-SDM07.pdf**), (Date accessed: November 2018).
24. Guyon I. and Elisseeff A., 2003. An Introduction to Variable and Feature Selection. **The Journal of Machine Learning Research, 3**: 1157-1182.
25. Liu H. and Yu L., 2005. Toward integrating feature selection algorithms for classification and clustering, **IEEE Transactions on Knowledge and Data Engineering, 17**: 491-502.

26. Hawkins D. M., 2004. The Problem of Overfitting. **Journal of Chemical Information and Computer Sciences, 44**: 1-12.

27. Xue B., Zhang M., Browne W. N., Yao X., 2016. A survey evolutionary computation approaches to feature selection, **IEEE Transaction on Evolutionary Computation, 20**(4): 606-626.

28. Danaee P., Ghaeini R, Hendrix D.A., 2016. A Deep Learning Approach for Cancer Detection and Relevant Gene Identification, **Biocomputing,** 219-229.

29. Maienschein-Cline M., Zhou J., Kevin P. W., Sciammas R., Dinner A.R., 2012. Discovering transcription factor regulatory targets using gene expression and binding data, **Bioinformatics, 28**(2):206-213.

30. Schadt E.E., Lamb J., Yang X., Zhu J., Edwards S., Guhathakurta D., Sieberts S.K., Monks S., Reitman M., Zhang C., Lum P.Y., Leonardson A., Thieringer R., Metzger J.M., Yang L., Castle J., Zhu H., Kash S.F., Drake T.A., Sachs A., Lusis A.J., 2005. An integrative genomics approach to infer causal associations between gene expression and disease, **Nature Genetics, 37**(7):710-7.

31. Shabana K. M., Nazeer K.A., Pradhan M., Palakal M., 2015. A computational method for drug repositioning using publicly available gene expression data, **BMC Bioinformatics, 16**(17): 1-5.

32. Chang C-H., Rampasek L., Goldenberg A., 2018, Dropout Feature Ranking for Deep Learning Models, Bioinformatics:1712.08645.

33. Liu B., Wei Y., Zhang Y., Yang Q., 2017, Deep neural networks for high dimension, low sample size data, IJCAI'17 Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2287-2293 pp.

34. Tan P.N., Steinbach M., Kumar V., 2006. Introduction to data mining (China ed.): Pearson Education Asia Ltd and Post & Telecom Press, 169 pp.

35. Tang J., Alelyani S., Liu H., 2014. Feature Selection for Classification: A Review, **Data Classification: Algorithms and Applications:** 37-64

36. Güngör O., Akar Ö. Ok A., 2011. Rastgele orman sınıflandırma yöntemi yardımıyla tarım alanlarındaki ürün çeşitliliğinin sınıflandırılması, Antalya TUFUAB 2011 VI. Teknik Sempozyumu.

37. Pal, M., 2005, Random Forest Classifier For Remote Sensing Classification. **International Journal Of Remote Sensing, 26**(1): 217-222.

38. Breiman, L., 2002. Manual On Setting Up, Using And Understanding Random Forests V3.1, (Web Pages: **http://oz.berkeley.edu/users/breiman/Using_random_forests_V3.1.pdf),** (Access Date : November 2018).
39. Liaw A., Wiener M., 2002. Classification And Regression By Random Forest. **R News, 2**(3):18-22.
40. Vapnik V., 1995. The Nature of Statistical Learning Theory, Second Edition, Springer-Verlag, New York, 317 pp.
41. Chandra B., Gupta M., 2011. An efficient statistical feature selection approach for classification of gene expression data. **Journal of Biomedical Informatics, 44**:529-535.
42. Yıldız O., Tez M., Bilge H. Ş., Akcayol M. A., Güler İ., 2012. Meme kanseri sınıflandırması için veri füzyonu ve genetik algoritma tabanlı gen seçimi, **Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 27**(3): 659-668.
43. Dreiseitl S., Ohno-Machado L., 2002. Logistic regression and artificial neural network classification models: a methodology review. **Journal of Biomedical Informatics, 35**(5–6):352-359.
44. Kuncheva L. I., 1995. Editing for the k-nearest neighbors rule by a genetic algorithm. **Pattern Recognition Letters, 16:**809-814.
45. Ho S. Y., Shu L., Chen H., 1995. Intelligent genetic algorithm with a new intelligent crossover using orthogonal arrays. Proceedings of the genetic and evolutionary computation conference, Florida, USA, 289-296.
46. Enas G. G., Choi S. S., 1986. Choice of smoothing parameter and efficiency of k-nearest neighbor classification. **Computer & Mathematics with Applications, 12**: 235-244.
47. Rudolfer S.M., Paliouras G., Peers I.S., 1999. A Comparison of Logistic Regression to Decision Tree Induction in the Diagnosis of Tunnel Syndrome. **Computers and Biomedical Research, 32**(5): 391-414.
48. Friedman N., Geiger D., Goldszmidt M., 1997. Bayesian Network Classifiers. **Machine Learning, 29:**131-163.
49. Yıldız K., Çamurcu Y, Doğan B., 2010. Veri Madenciliğinde Temel Bileşenler Analizi ve Negatifsiz Matris Çarpanlarına Ayırma Tekniklerinin Karşılaştırmalı Analizi. Akademik Bilişim'10-XII. Akademik Bilişim Konferansı Bildiriler.

50. Gong L., Mu T., Wang M., Liu H., Goulermase J.Y., 2017. Evolutionary nonnegative matrix factorization with adaptive control of cluster quality. **Neurocomputing, 272** (17): 237-249.

51. Shahnaza, F., Berry M.W., Pauca V.P., Plemmons R. J., 2006. Document clustering using nonnegative matrix factorization. **Information Processing and Management, 42**(2), 373–386.

52. Lee D. D., Seung H. S., 1999. Learning the parts of objects by non-negative matrix factorization. **Nature, 401**: 788–791.

53. Brunet J.P., Tamayo P., Golub T. R. and Mesirov J. P., 2004. Metagenes and molecular pattern discovery using matrix factorization. **Proceedings of the National Academy of Science, 101**(12) 4164–4169

54. Yang Z., Oja E., 2010. Linear and Nonlinear Projective Nonnegative Matrix Factorization. **IEEE Transactios On Neural Networks, 21**(5): 734 – 749.

55. Berry M. W. Et. Al., 2007. Algorithms and applications for approximate nonnegative matrix factorization, **Computational Statistics & Data Analysis, 52**:155 – 173.

56. Devarajan K., 2008. Nonnegative Matrix Factorization: An Analytical and Interpretive Tool in Computational Biology. **PLoS Computational Biology, 4**(7): e1000029.

57. Hong-Bo X., 2009. Of the mechanomyogram signal using a wavelet packet transform and singular value decomposition for multifunction prosthesis control. **Physiologial Measurement, 30** :441–457.

58. Baker K., 2005. Singular Value Decomposition Tutorial, 14-16, (Web Pages: **https://www.researchgate.net/publication/246546380_Singular_Value_Decomposition_Tutorial**), (Access Date: Aralık 2018).

59. Scikit Learn, Feature selection. (Web Pages: **https://scikit-learn.org/stable/modules/feature_selection.html**), (Access Date: Aralık 2018).

60. A blog on machine learning, data mining and visualization, 2014. Feature selection – Part I: univariate selection. (Web Pages: **https://blog.datadive.net/selecting-good-features-part-i-univariate-selection**), (Access Date: Aralık 2018).

61. Granitto P. M., Furlanello C., Biasiolia F., Gasperia F., 2006. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. **Chemometrics and Intelligent Laboratory Systems, 83**(2): 83-90.

62. Truong D.T., Franzosa E.A., Tickle T. L., Scholz M., Weingart G., Pasolli E., Tett A., Huttenhower C., Segata N., 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. **Nature Methods, 12**: 902-903.

63. Qin J., Li R., Raes J., Arumugam M., Burgdorf K.S., Manichanh C., et al., 2010. A human gut microbial gene catalogue established by metagenomic sequencing. **Nature, 464** (7285):59–65.

64. Qin N., Yang F., Li A., Prifti E., Chen Y., Shao L., et al., 2014. Alterations of the human gut microbiome in liver cirrhosis. **Nature, 513**(7516):59–64. pmid:25079328.

65. Zeller G., Tap J., Voigt A.Y., Sunagawa S., Kultima J.R., Costea P.I., et al., 2014. Potential of fecal microbiota for early-stage detection of colorectal cancer. **Moleculer System Biology, 10**(11): 766.

66. Le Chatelier E., Nielsen T., Qin J., Prifti E., Hildebrand F., Falony G., et al., 2013. Richness of human gut microbiome correlates with metabolic markers. **Nature, 500** (7464):541–546.

67. Tett A., Pasolli E., Farina S., Truong D.T., Asnicar F., et al., 2017. Unexplored diversity and strain-level structure of the skin microbiome associated with psoriasis, **NPJ Biofilms and Microbiomes, 3**(1): 14.

68. Qin J., Li Y., Cai Z., Li S., Zhu J., Zhang F., et al., 2012. A metagenome-wide association study of gut microbiota in type 2 diabete. **Nature, 490**(7418):55–60.

69. Karlsson F.H., Tremaroli V., Nookaew I., Bergström G., Behre C.J., Fagerberg B., et al., 2013. Gut metagenome in European women with normal impaired and diabetic glucose contro., **Nature, 498** (7452):99–103.

70. Horvath A., Bashir M., Schmerboeck B., Rainer F., Krones E., Baumann-Durchschein F., Douschan P., Spindelboeck W., Zollner G., Stauber R.E. and Fickert P., 2017. Intestinal colonisation by Veillonella spp. is predictive for mortality in stable cirrhosis and could be partially reduced by a multispecies probiotic in a randomized placebo controlled trial. **Journal of Hepatology, 66**(1): 128.

71. Fukui H., 2017. Gut Microbiome-based therapeutics in liver cirrhosis: basic consideration for the next step. **Journal of Clinical and Translational Hepatology, 5** (3): 249.

72. Bajaj J. S., Heuman D. M., Hylemon P. B., Sanyal A. J., White M. B., Monteith P., et al., 2014. Altered profile of human gut microbiome is associated with cirrhosis and its complications. **J Hepatol, 60**:940–947.

73. Hammes T.O., Leke R., Escobar, T.D.C., Fracasso L.B., Meyer F.S., Andrades M.É. and da Silveira T.R., 2017. *Lactobacillus rhamnosus* GG reduces hepatic fibrosis in a model of chronic liver disease in rats. **Nutricion Hospitalaria, 34**(3):702-709.

74. Qin N., Yang F., Li A., Prifti E., Chen Y., Shao L., Guo J., Le Chatelier E., Yao J., Wu L. and Zhou J., 2014. Alterations of the human gut microbiome in liver cirrhosis. **Nature, 513**(7516): 59.

75. Klimesova K., Kverka M., Zakostelska Z., Hudcovic T., Hrncir T., Stepankova R., Rossmann P., Ridl J., Kostovcik M., Mrazek J. and Kopecny J., 2013. Altered gut microbiota promotes colitis-associated cancer in IL-1 receptor–associated kinase M–deficient mice. **Inflammatory Bowel Diseases, 19**(6):1266-1277.

76. Lucas C., Barnich N. and Nguyen H.T.T., 2017. Microbiota, inflammation and colorectal cancer. **International Journal of Molecular Sciences, 18**(6): 1310.

77. Rivière A., Selak M., Lantin D., Leroy F. and De Vuyst L., 2016. Bifidobacteria and butyrate-producing colon bacteria: importance and strategies for their stimulation in the human gut. **Frontiers in microbiology, 7**: 979.

78. Gao Z., Guo B., Gao R., Zhu Q. and Qin H., 2015. Microbiota disbiosis is associated with colorectal cancer. **Frontiers in microbiology, 6**: 20.

79. Yurdakul D., Yazgan-Karataş A., Şahin F., 2015, Enterobacter strains might promote colon cancer, Current microbiology, 71(3): 403-411.

80. Jarvis G.A., Li J. and Swanson K.V., 1999. Invasion of human mucosal epithelial cells by Neisseria gonorrhoeae upregulates expression of intercellular adhesion molecule 1 (ICAM-1). **Infection and Immunity, 67**(3): 1149-1156.

81. Togo A.H., Valero R., Delerce J., Raoult D. and Million M., 2016. Anaerotruncus massiliensis, a new species identified from human stool from an obese patient after bariatric surgery. **New Microbes and New Infections, 14**: 56-57.

82. Olds W., 2014. Bacteroides uniformis CECT 7771 Ameliorates Metabolic and Immunological Dysfunction in Mice with High-Fat-Diet Induced Obesity. In Health and the Gut, Apple Academic Press, p: 97-132.

83. Gill S. R., Pop M., Deboy R. T., Eckburg P. B., Turnbaugh P. J., Samuel B. S., Gordon J. I., Relman D. A., Fraser-Liggett C. M. & Nelson K. E., 2006. Metagenomic analysis of the human distal gut microbiome. **Science, 312**:1355–1359.

84. Gorvitovskaia A., Holmes, S.P. and Huse S.M., 2016. Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. **Microbiome, 4**(1): 15.

85. Kasai C., Sugimoto K., Moritani I., Tanaka J., Oya Y., Inoue H., Tameda M., Shiraki K., Ito M., Takei Y. and Takase K., 2015. Comparison of the gut microbiota composition between obese and non-obese individuals in a Japanese population, as analyzed by terminal restriction fragment length polymorphism and next-generation sequencing, **BMC Gastroenterology, 15**(1): 100.

86. Fei N. and Zhao L., 2013. An opportunistic pathogen isolated from the gut of an obese human causes obesity in germfree mice. **The ISME Journal, 7**(4): 880.

87. Hou Y.P., He Q.Q., Ouyang H.M., Peng H.S., Wang Q., Li J., Lv X.F., Zheng Y.N., Li S.C., Liu H.L. and Yin A.H., 2017. Human gut microbiota associated with obesity in Chinese children and adolescents**, BioMed Research International, 2017**: 1-8

88. Gomes A.C., Hoffmann, C. and Mota J.F., 2018. The human gut microbiota: Metabolism and perspective in obesity. **Gut Microbes, 9**(4): 308-325.

89. Del Chierico F., Abbatini F., Russo A., Quagliariello A., Reddel S., Capoccia D., Caccamo R., Ginanni Corradini S., Nobili V., De Peppo F. and Dallapiccola B., 2018. Gut microbiota markers in obese adolescent and adult patients: age-dependent differential patterns. **Frontiers in Microbiology, 9**: 1210.

90. Tannock G.W., 2010. The bowel microbiota and inflammatory bowel disease. **International Journal of Inflammation, 2010**:1-9

91. Zhou Y. and Zhi F., 2016. Lower level of bacteroides in the gut microbiota is associated with inflammatory bowel disease: a meta-analysis. **BioMed Research International, 2016**(2): 1-9.

92. Perez-Muñoz M.E., Bergstrom K., Peng V., Schmaltz R., Jimenez-Cardona R., Marsteller N., McGee S., Clavel T., Ley R., Fu J. and Xia L., 2014. Discordance

between changes in the gut microbiota and pathogenicity in a mouse model of spontaneous colitis. **Gut Microbes, 5**(3): 286-485.

93. Hakansson A. and Molin G., 2011. Gut microbiota and inflammation. **Nutrients, 3**(6): 637-682.
94. Drago L., Grandi R., Altomare G., Pigatto,P., Rossi O. and Toscano M., 2016. Skin microbiota of first cousins affected by psoriasis and atopic dermatitis. **Clinical and Molecular Allergy, 14**(1): 2.
95. Noah P.W., 1990. The role of microorganisms in psoriasis. **In Seminars in dermatology, 9(**4): 269-276.
96. Manasson J., Reddy S.M., Neimann A.L., Segal L.N. and Scher J.U., 2016, Cutaneous Microbiota Features Distinguish Psoriasis from Psoriatic Arthritis, In Arthritis & Rheumatology, Vol. 68: 111, NJ USA: WILEY.
97. Moreno-Indias I., Sánchez-Alcoholado L., García-Fuentes E., Cardona F., Queipo-Ortuño M.I. and Tinahones F.J., 2016. Insulin resistance is associated with specific gut microbiota in appendix samples from morbidly obese patients. **American Journal of Translational Research, 8**(12): 5672.
98. Wu T.R., Lin C.S., Chang C.J., Lin T.L., Martel J., Ko Y.F., Ojcius D.M., Lu C.C., Young J.D. and Lai H.C., 2018. Gut commensal Parabacteroides goldsteinii plays a predominant role in the anti-obesity effects of polysaccharides isolated from Hirsutella sinensis, **BMJ Gut, 68**(2): 248-262.
99. Long J., Cai Q., Steinwandel M., Hargreaves M.K., Bordenstein S.R., Blot W.J., Zheng W. and Shu X.O., 2017. Association of oral microbiome with type 2 diabetes risk. **Journal of periodontal research, 52**(3): 636-643.
100. Bordalo Tonucci L., Dos Santos K.M.O., De Luces Fortes Ferreira C.L., Ribeiro S.M.R., De Oliveira L.L. and Martino H.S.D., 2017. Gut microbiota and probiotics: Focus on diabetes mellitus. **Critical reviews in food science and nutrition, 57**(11): 2296-2309.
101. Larsen N., Vogensen F.K., van den Berg F.W., Nielsen D.S., Andreasen A.S., Pedersen B.K., Al-Soud W.A., Sørensen S.J., Hansen L.H. and Jakobsen M., 2010. Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults. **PloS one, 5**(2): 9085.

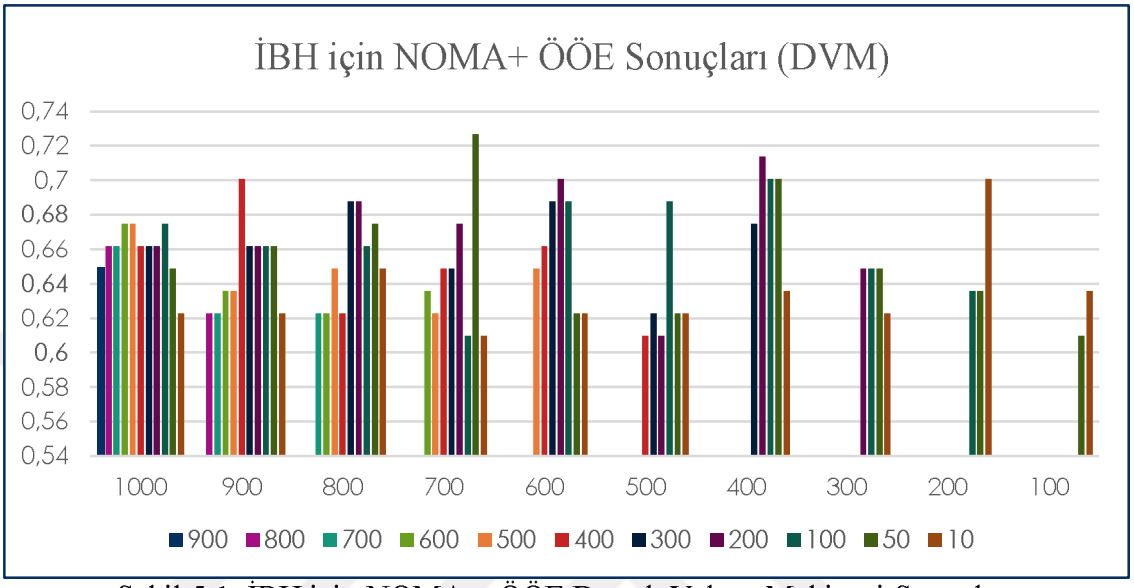
102. Krych Ł., Nielsen D.S., Hansen A.K. and Hansen C.H.F., 2015. Gut microbial markers are associated with diabetes onset, regulatory imbalance, and IFN-γ level in NOD mice. **Gut Microbes, 6**(2): 101-109.

103. Egshatyan L., Kashtanova D., Popenko A., Tkacheva O., Tyakht A., Alexeev D., Karamnova N., Kostryukova E., Babenko V., Vakhitova M. and Boytsov S., 2016. Gut microbiota and diet in patients with different glucose tolerance. **Endocrine Connections, 5**(1): 1-9.

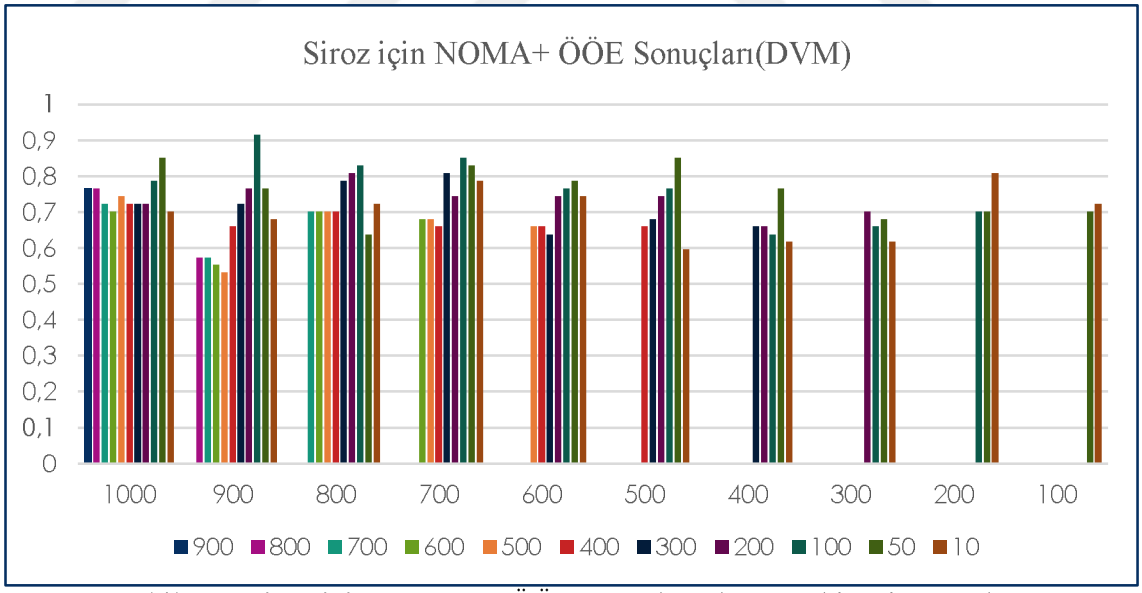
104. Zhang Q., Xiao X., Li M., Yu M., Ping F., Zheng J., Wang T. and Wang X., 2017. Vildagliptin increases butyrate-producing bacteria in the gut of diabetic rats. **PloS One, 12**(10):1-18

EKLER

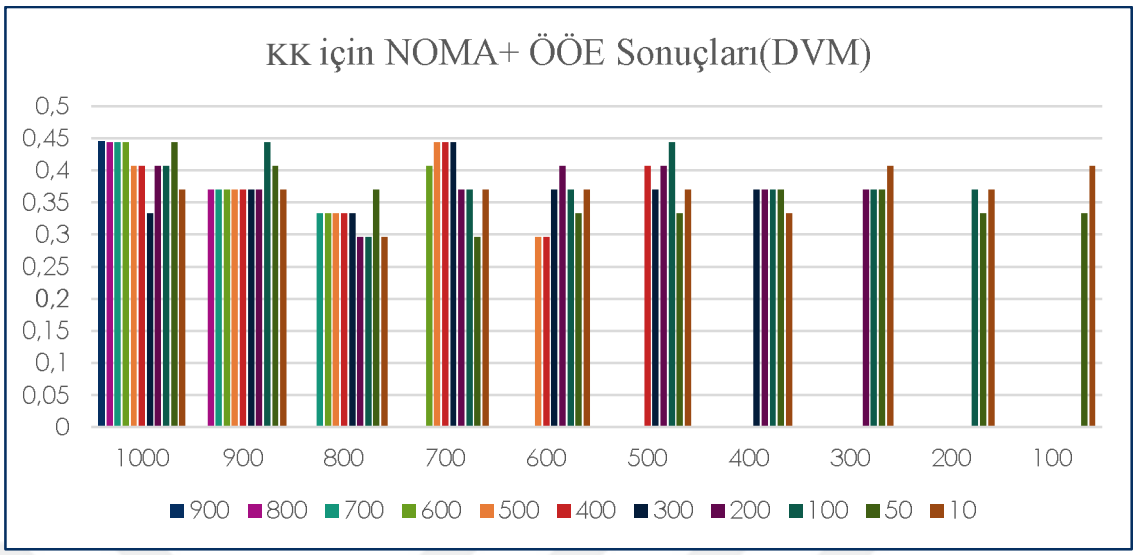
EK-1: Veri setleri için NOMA + ÖÖE Destek Vektör Makinesi (DVM) Sonuçları



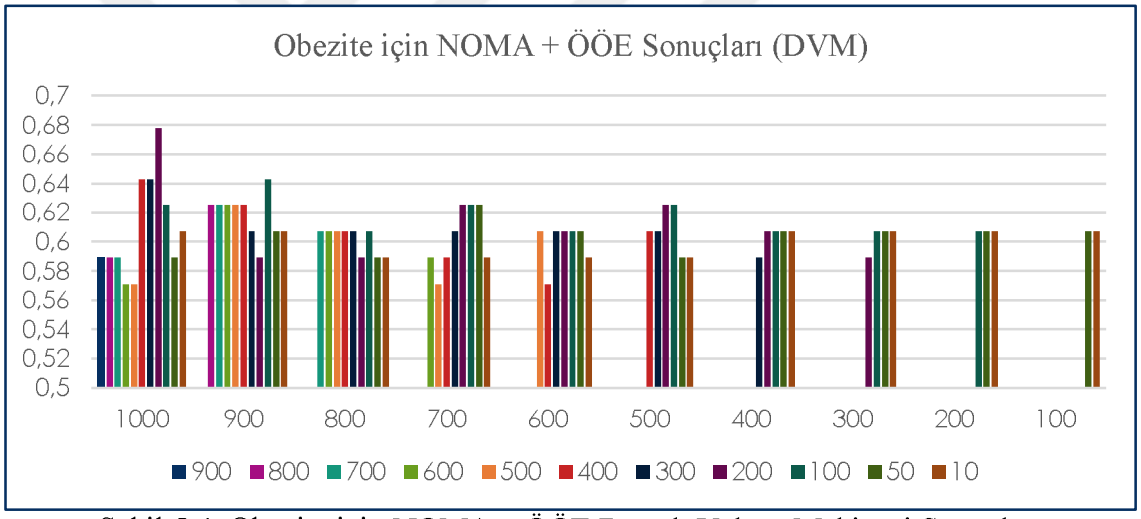
Şekil.5.1. İBH için NOMA + ÖÖE Destek Vektör Makinesi Sonuçları



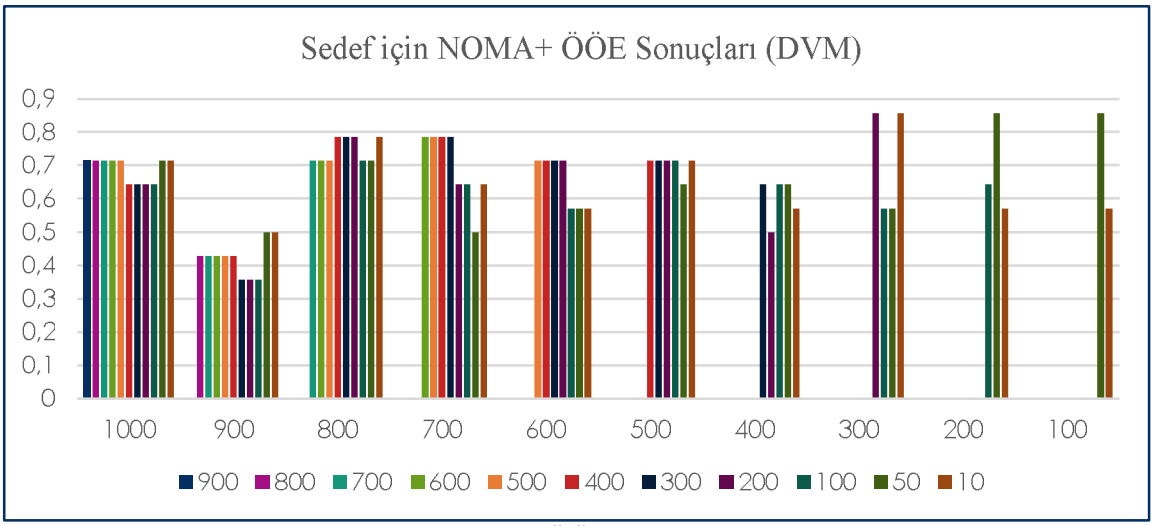
Şekil 5.2. Siroz için NOMA + ÖÖE Destek Vektör Makinesi Sonuçları



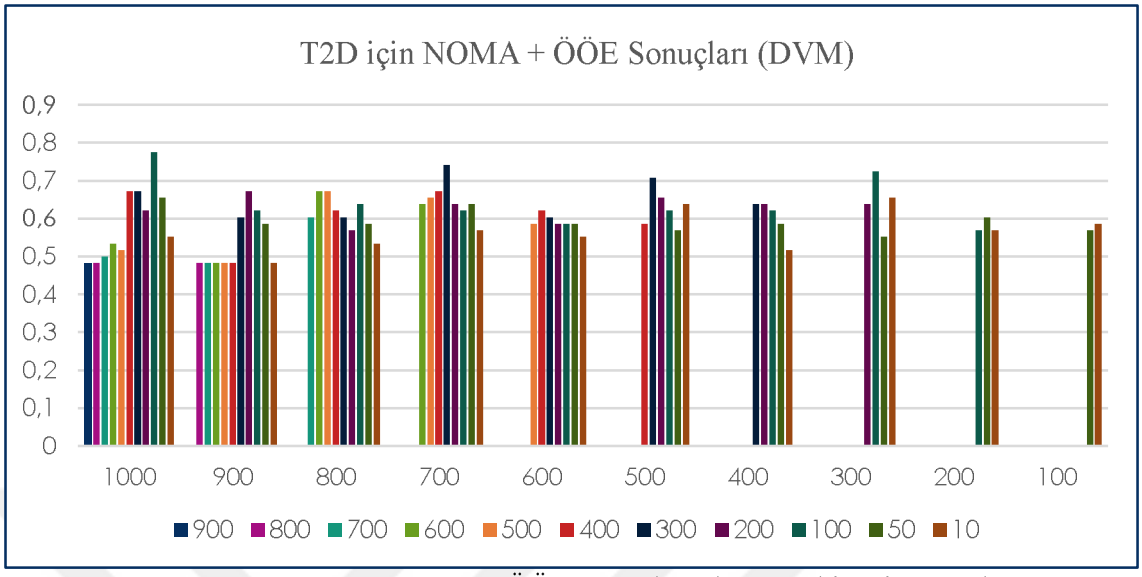
Şekil 5.3. KK için NOMA + ÖÖE Destek Vektör Makinesi Sonuçları



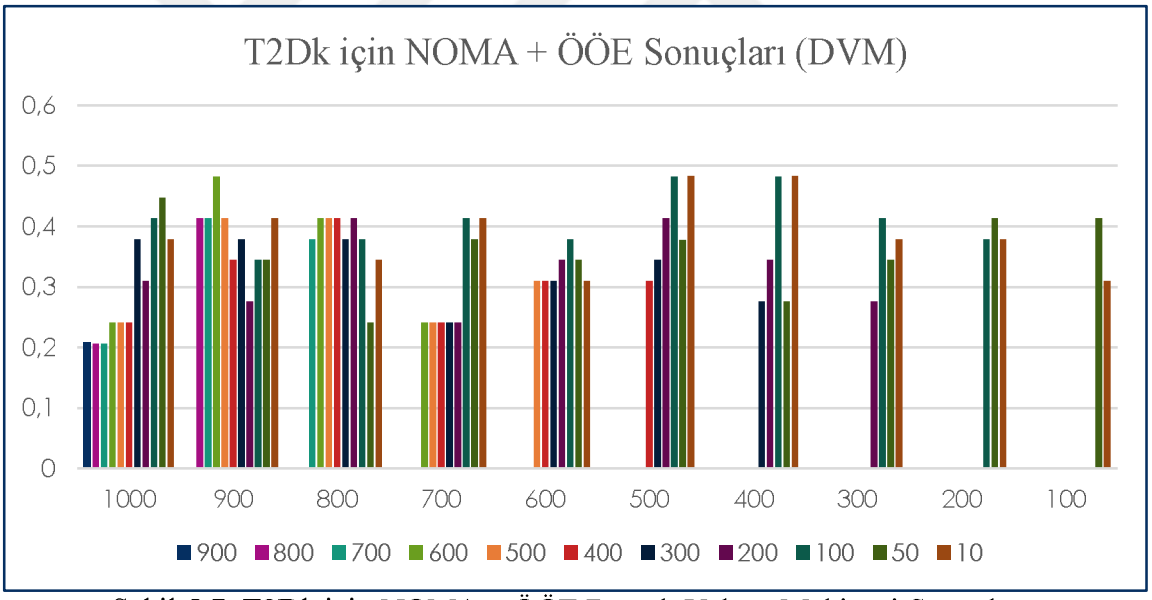
Şekil 5.4. Obezite için NOMA + ÖÖE Destek Vektör Makinesi Sonuçları



Şekil 5.5. Sedef Hastalığı NOMA + ÖÖE Destek Vektör Makinesi Sonuçları

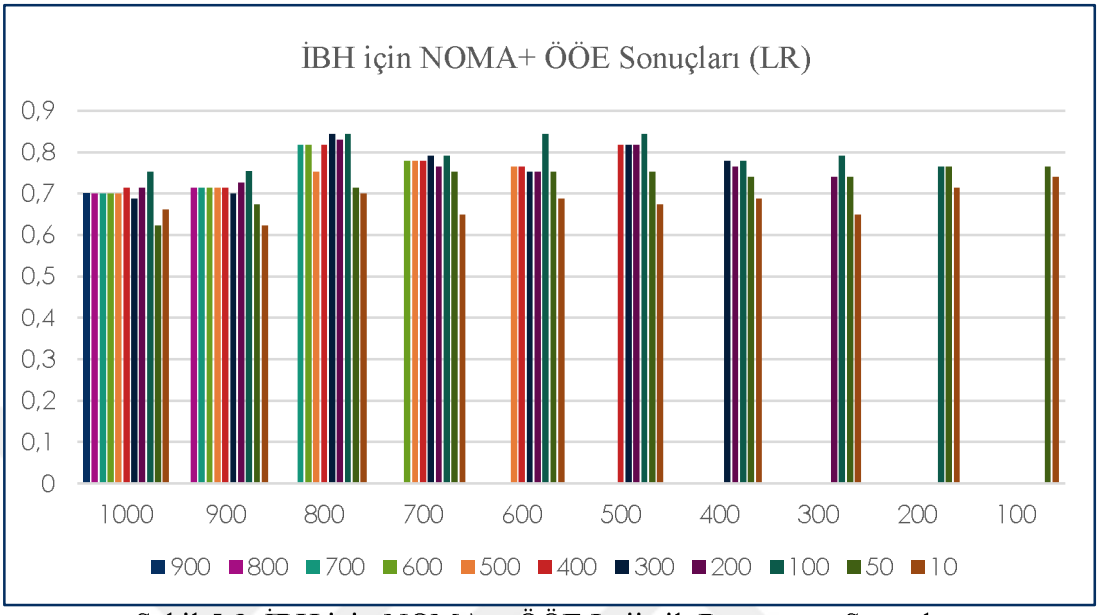


Şekil 5.6. T2D için NOMA + ÖÖE Destek Vektör Makinesi Sonuçları

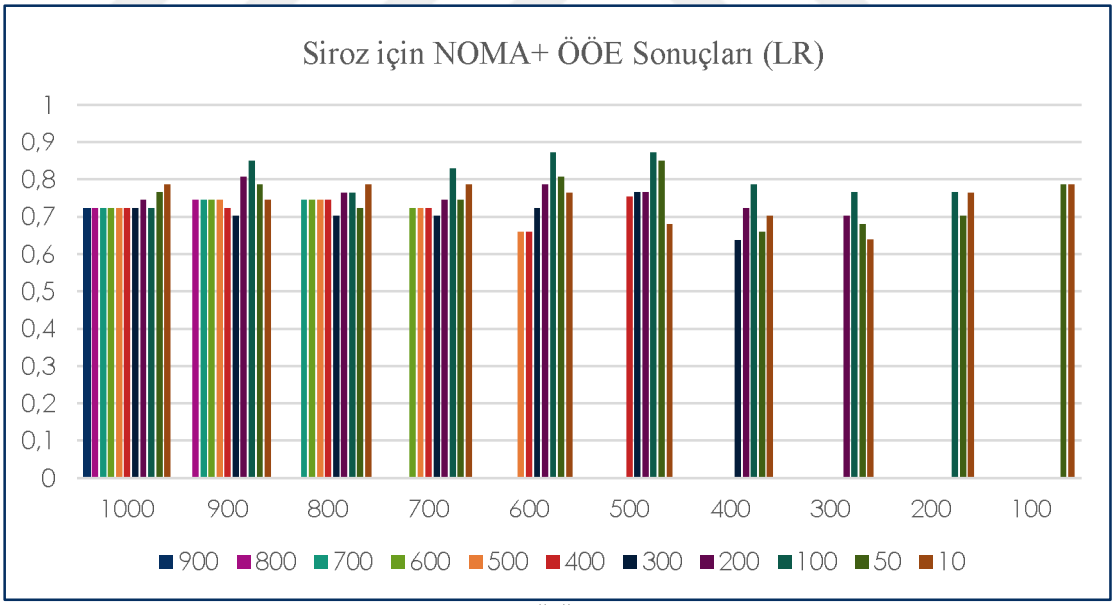


Şekil 5.7. T2Dk için NOMA + ÖÖE Destek Vektör Makinesi Sonuçları

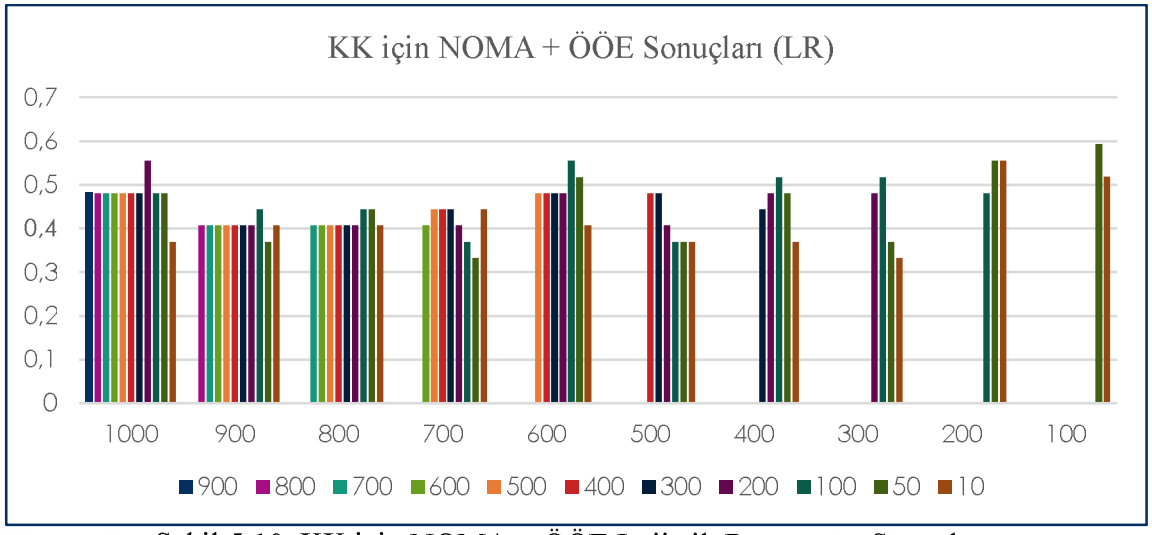
EK-2: Veri setleri için NOMA + ÖÖE Lojistik Regresyon (LR) Sonuçları



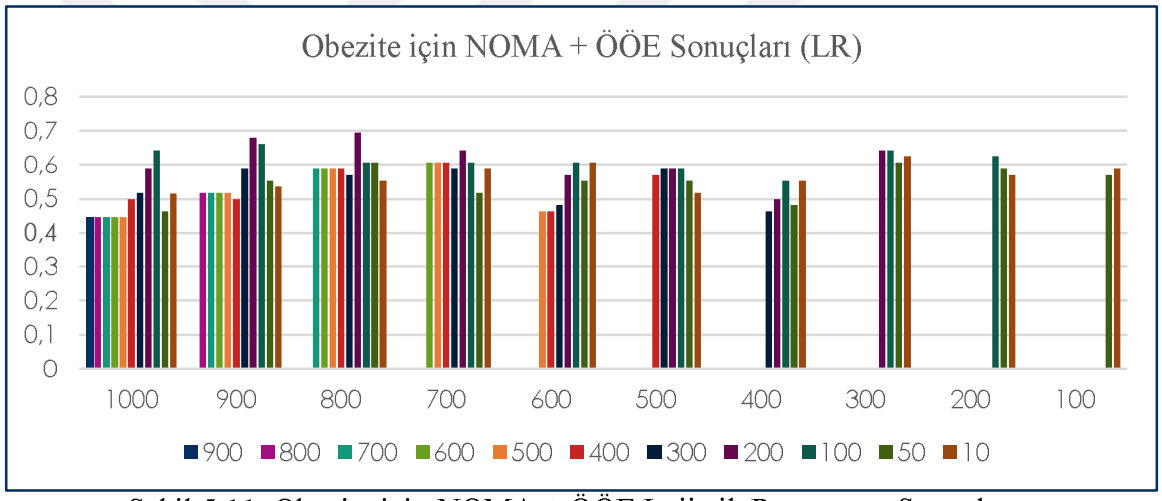
Şekil 5.8. İBH için NOMA + ÖÖE Lojistik Regresyon Sonuçları



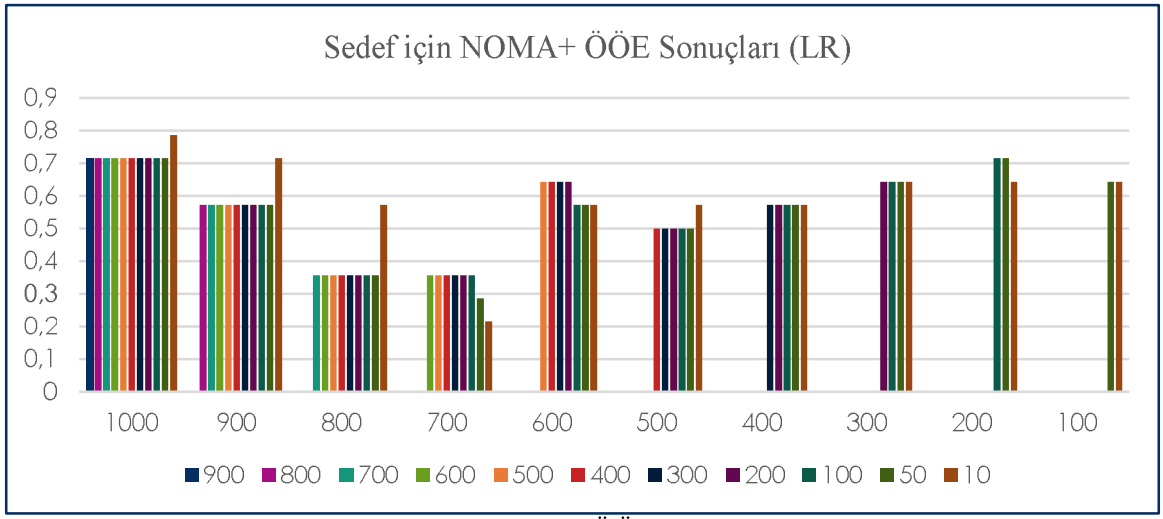
Şekil 5.9. Siroz için NOMA + ÖÖE Lojistik Regresyon Sonuçları



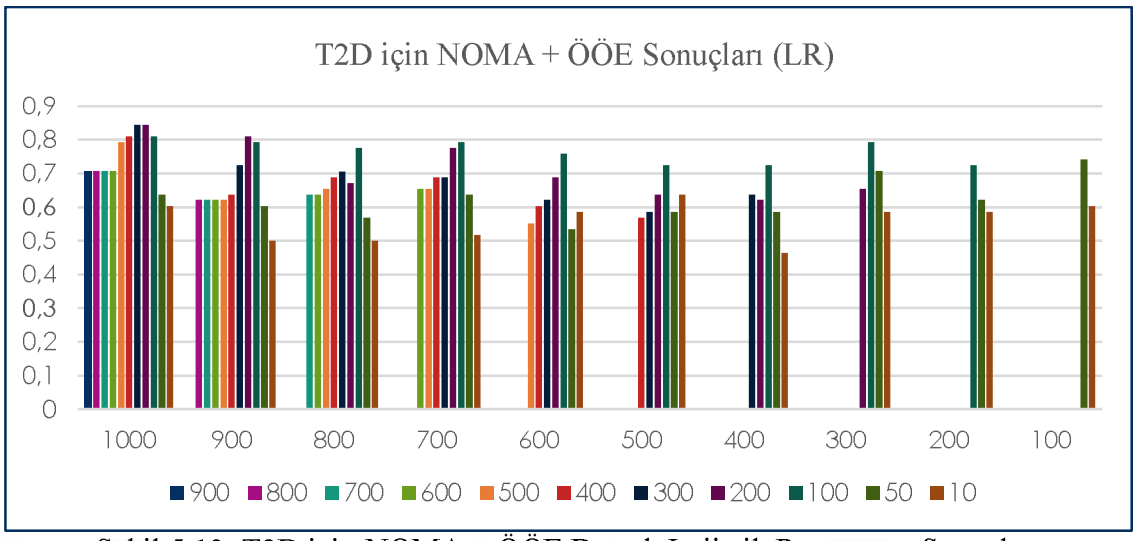
Şekil 5.10. KK için NOMA + ÖÖE Lojistik Regresyon Sonuçları



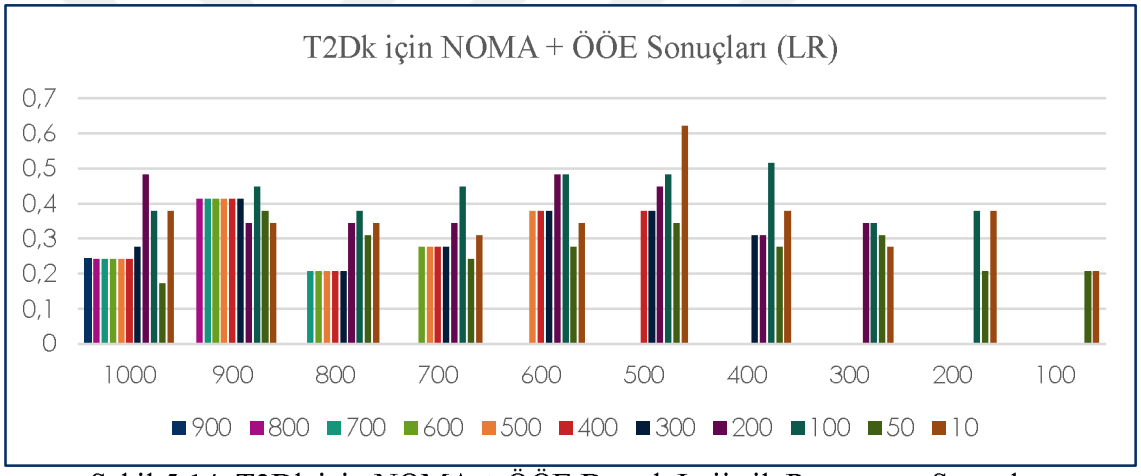
Şekil 5.11. Obezite için NOMA + ÖÖE Lojistik Regresyon Sonuçları



Şekil 5.12. Sedef Hastalığı NOMA + ÖÖE Lojistik Regresyon Sonuçları

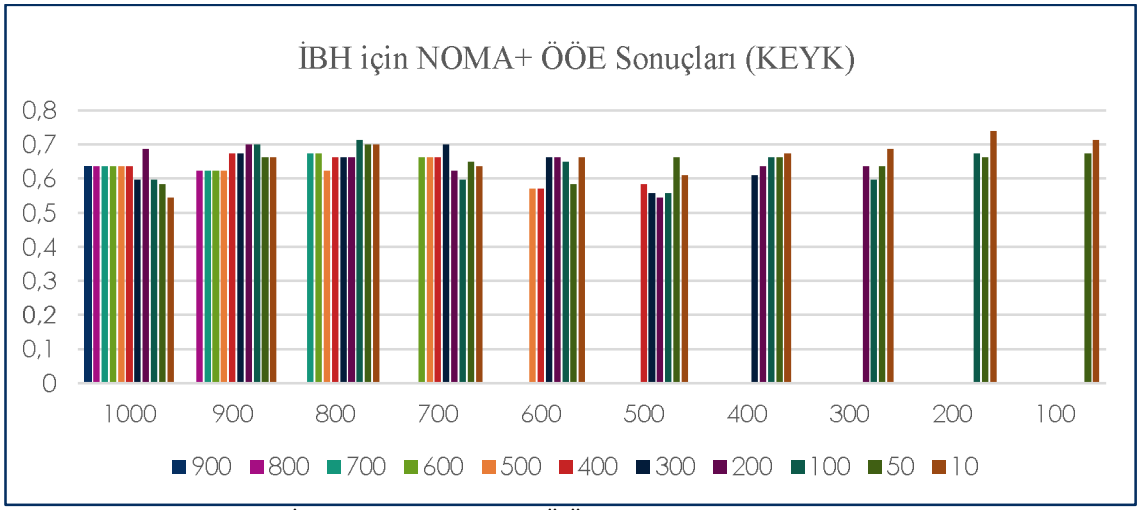


Şekil 5.13. T2D için NOMA + ÖÖE Destek Lojistik Regresyon Sonuçları

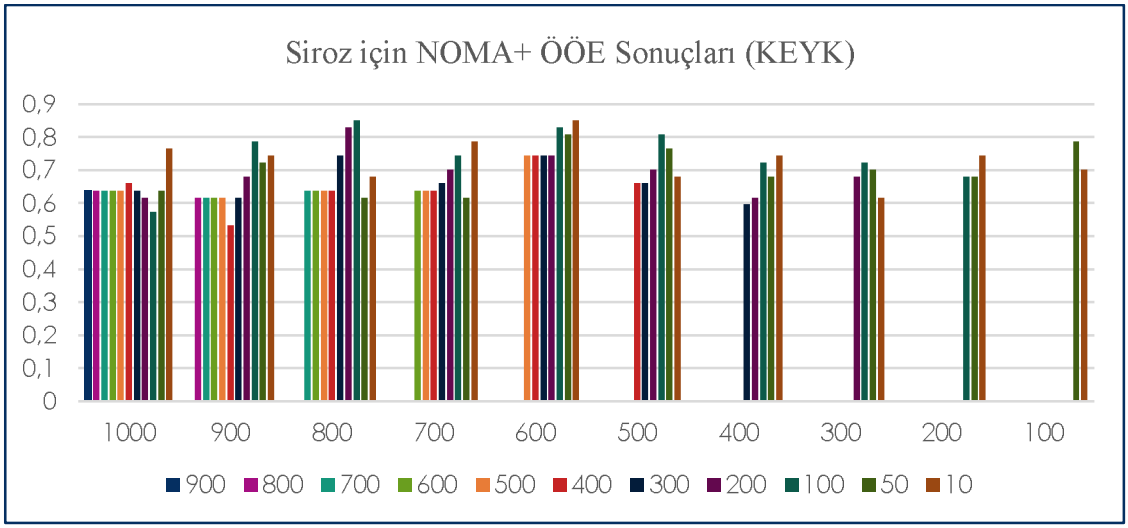


Şekil 5.14. T2Dk için NOMA + ÖÖE Destek Lojistik Regresyon Sonuçları

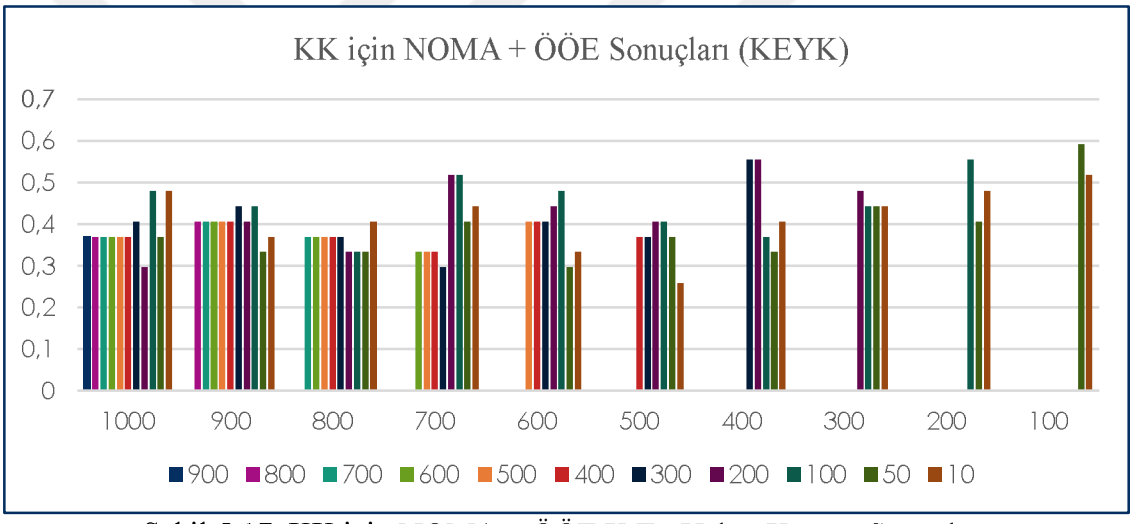
EK-3: Veri setleri için NOMA + ÖÖE K En Yakın Komşu (KEYK) Sonuçları



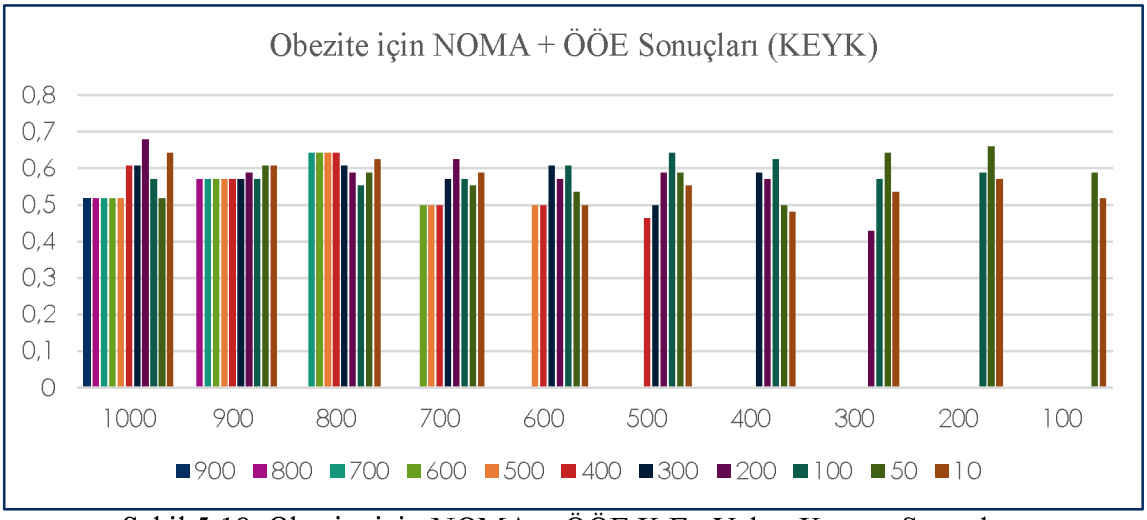
Şekil 5.15. İBH için NOMA + ÖÖE K En Yakın Komşu Sonuçları



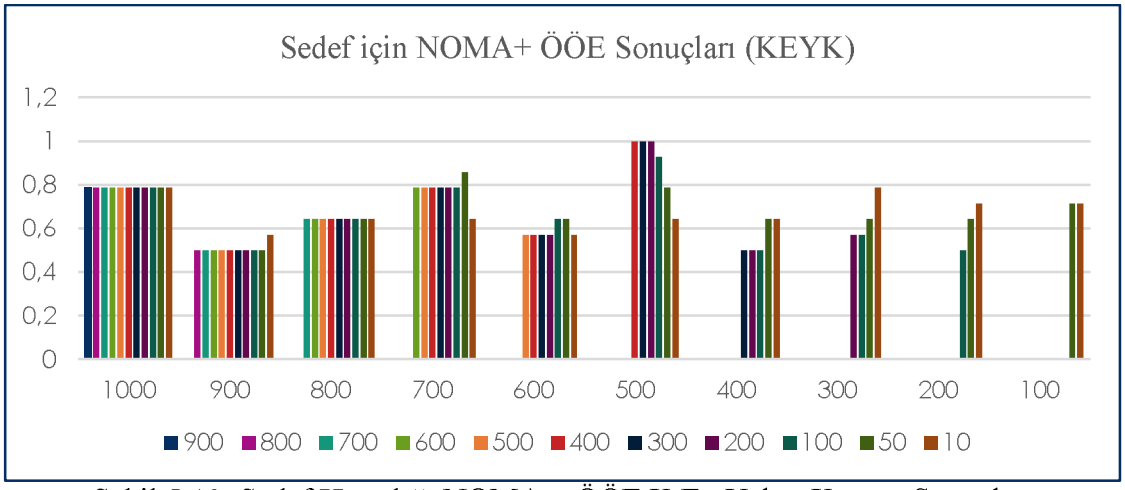
Şekil 5.16. Siroz için NOMA + ÖÖE K En Yakın Komşu Sonuçları



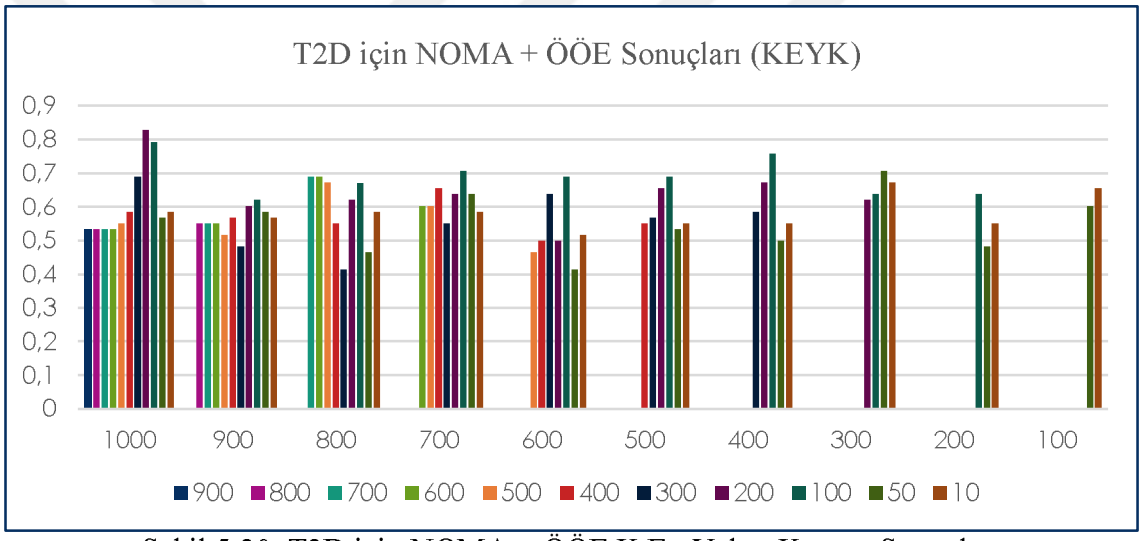
Şekil 5.17. KK için NOMA + ÖÖE K En Yakın Komşu Sonuçları



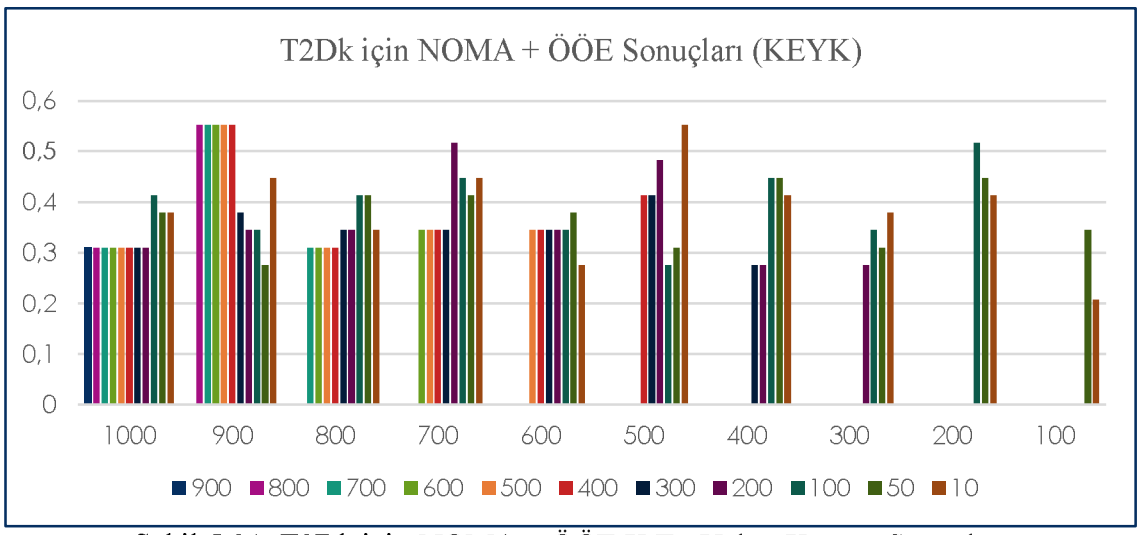
Şekil 5.18. Obezite için NOMA + ÖÖE K En Yakın Komşu Sonuçları



Şekil 5.19. Sedef Hastalığı NOMA + ÖÖE K En Yakın Komşu Sonuçları

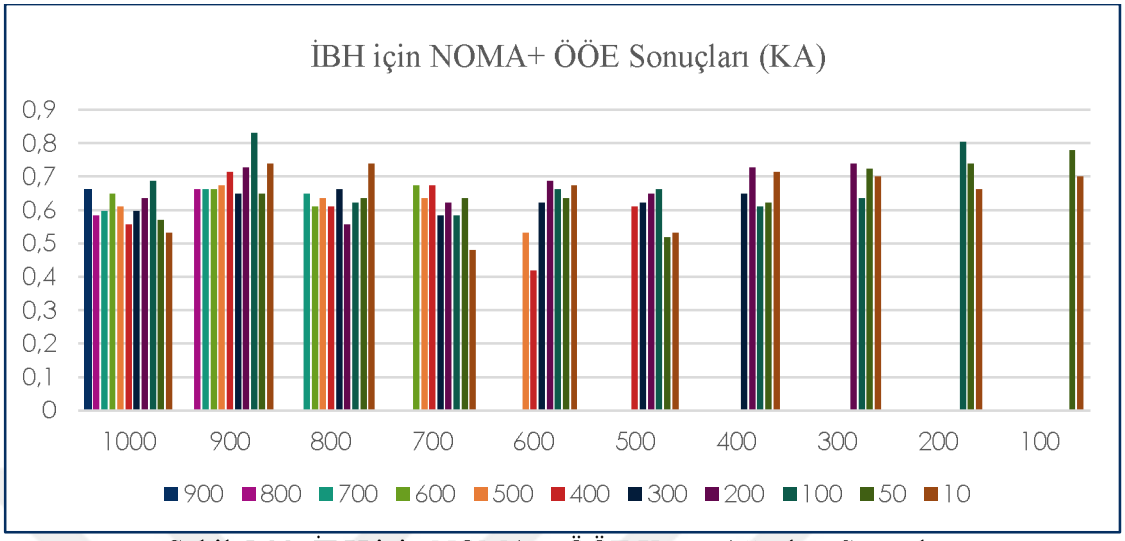


Şekil 5.20. T2D için NOMA + ÖÖE K En Yakın Komşu Sonuçları

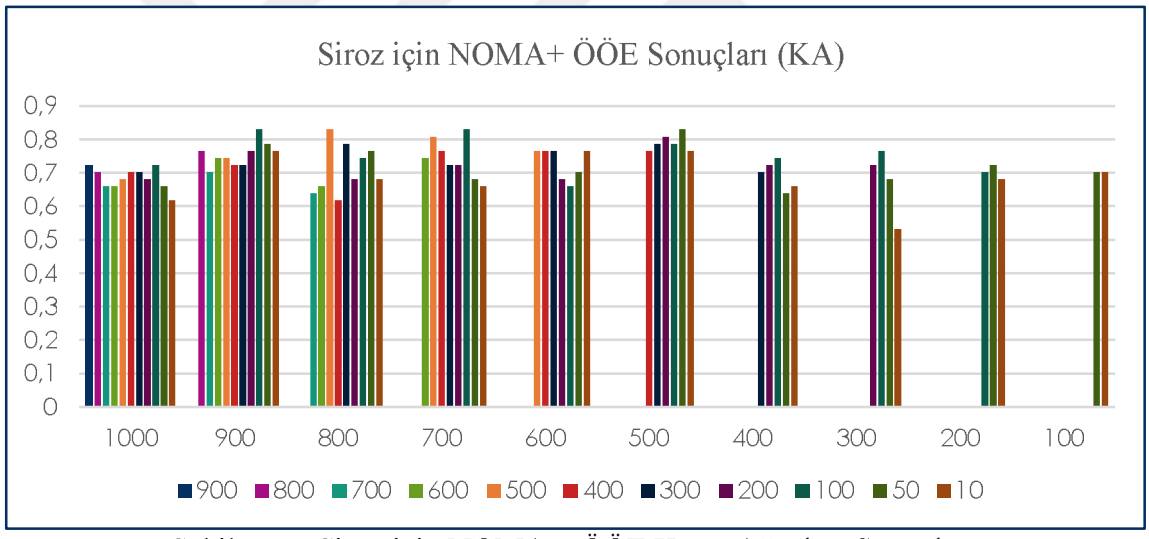


Şekil 5.21. T2Dk için NOMA + ÖÖE K En Yakın Komşu Sonuçları

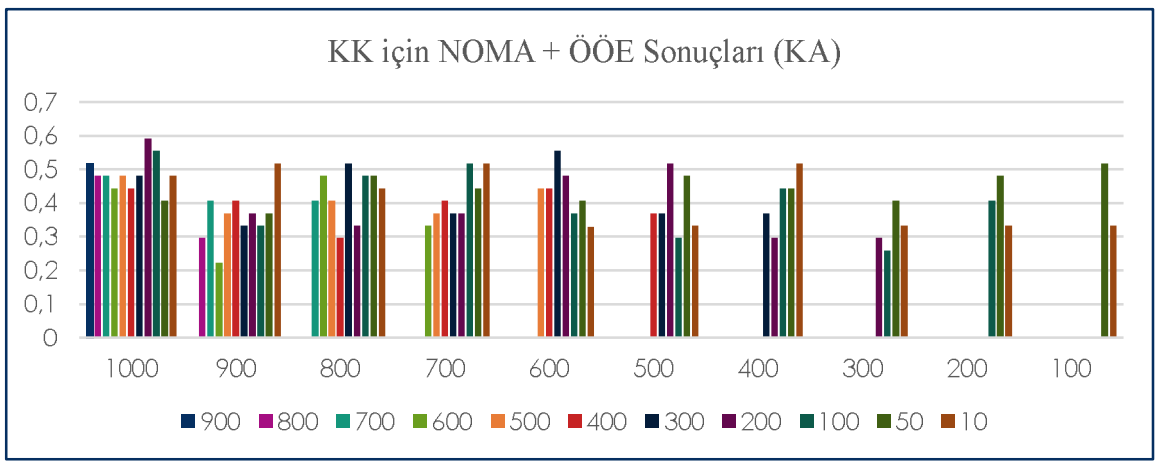
EK-4: Veri setleri için NOMA + ÖÖE Karar Ağaçları (KA) Sonuçları



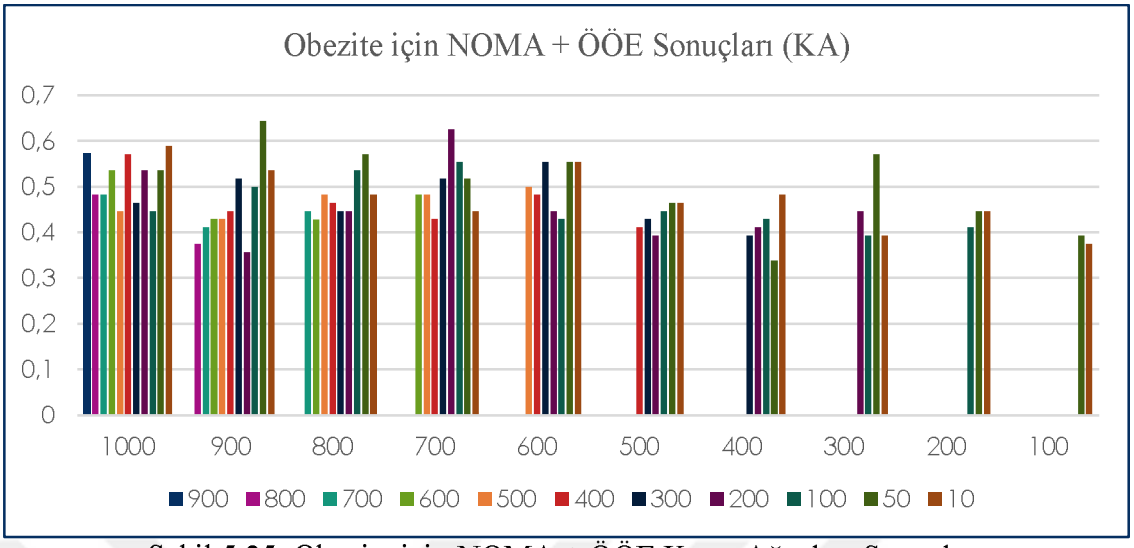
Şekil 5.22. İBH için NOMA + ÖÖE Karar Ağaçları Sonuçları



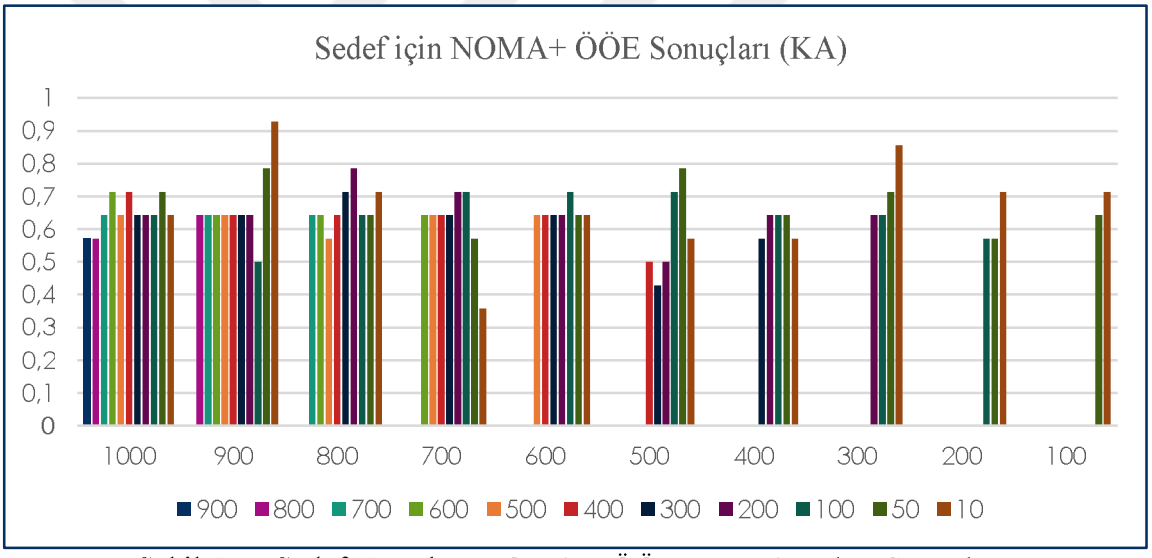
Şekil 5.23. Siroz için NOMA + ÖÖE Karar Ağaçları Sonuçları



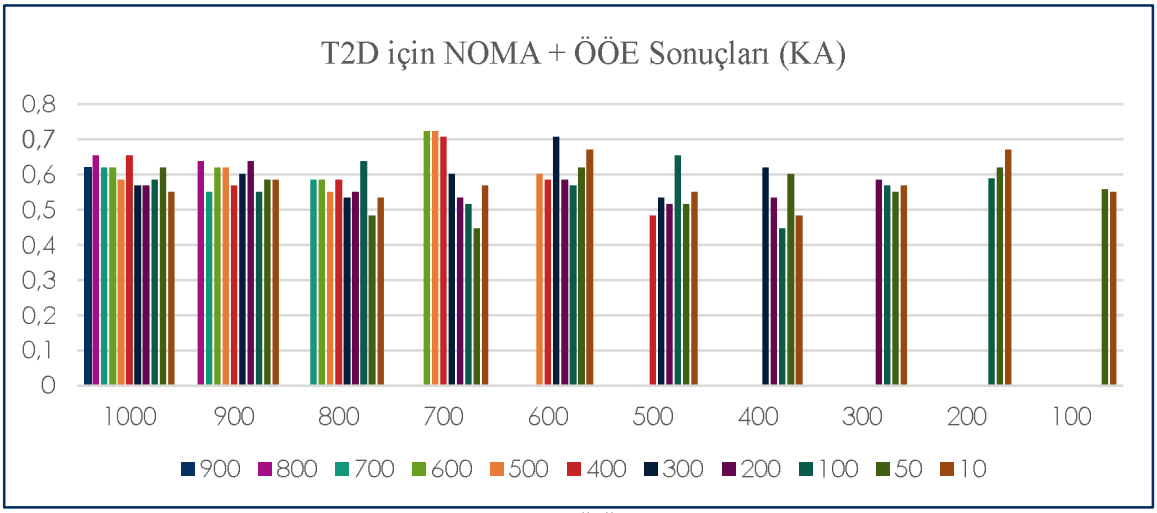
Şekil 5.24. KK için NOMA + ÖÖE Karar Ağaçları Sonuçları



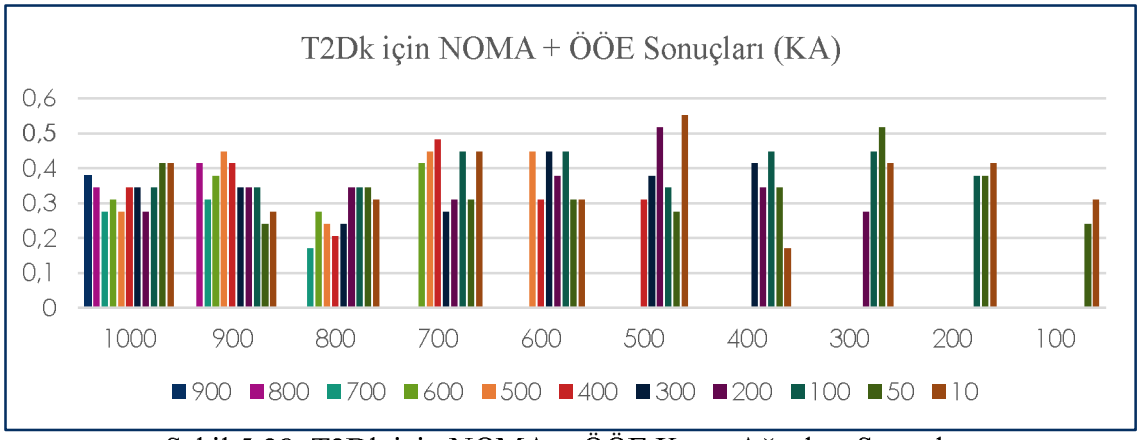
Şekil 5.25. Obezite için NOMA + ÖÖE Karar Ağaçları Sonuçları



Şekil 5.26 Sedef Hastalığı NOMA + ÖÖE Karar Ağaçları Sonuçları

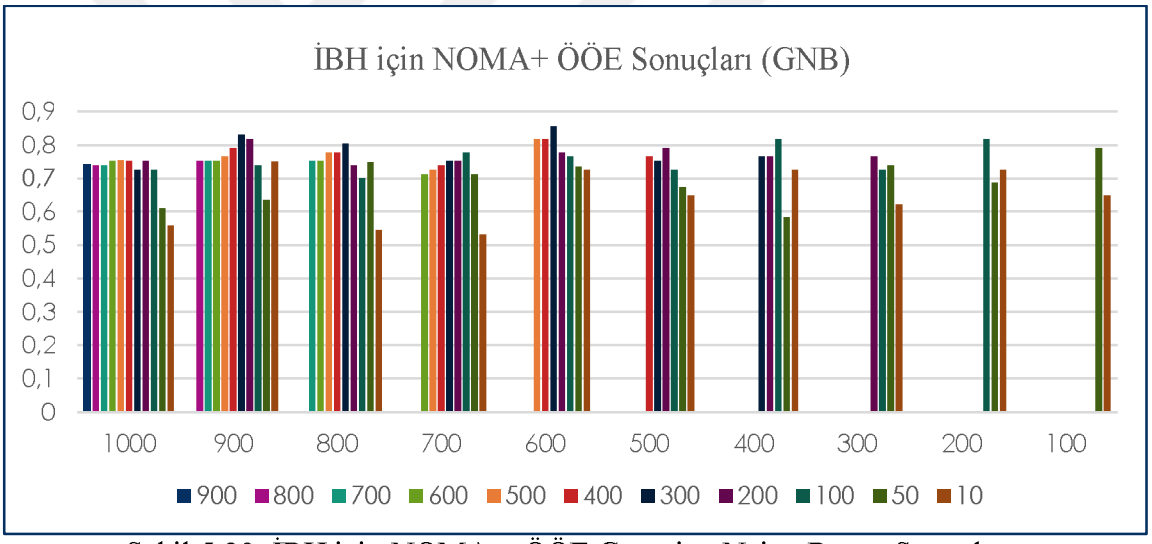


Şekil 5.27. T2D için NOMA + ÖÖE Karar Ağaçları Sonuçları

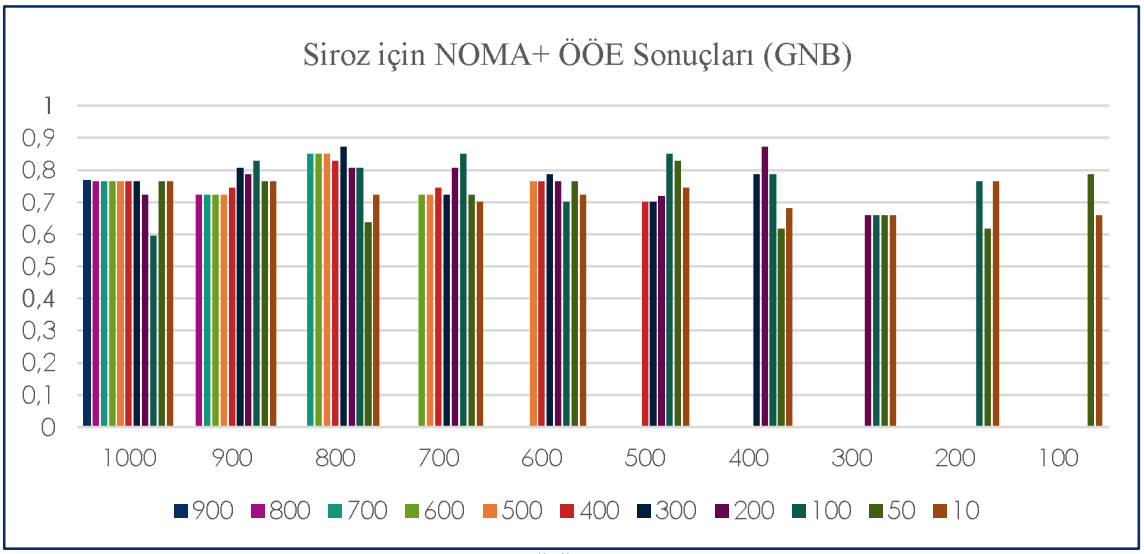


Şekil 5.28. T2Dk için NOMA + ÖÖE Karar Ağaçları Sonuçları

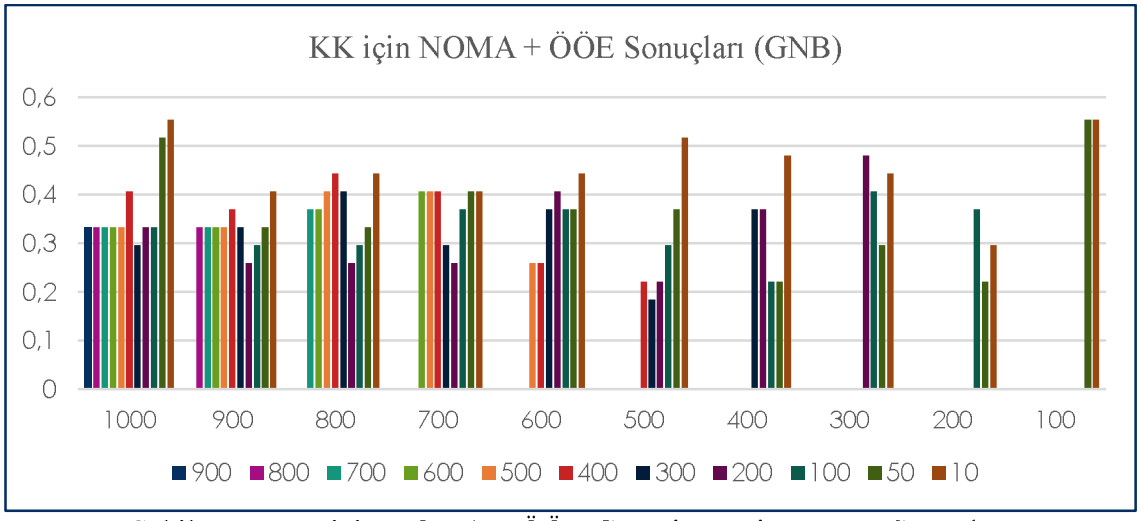
EK-5: Veri setleri için NOMA + ÖÖE Gaussian Naive Bayes (GNB) Sonuçları



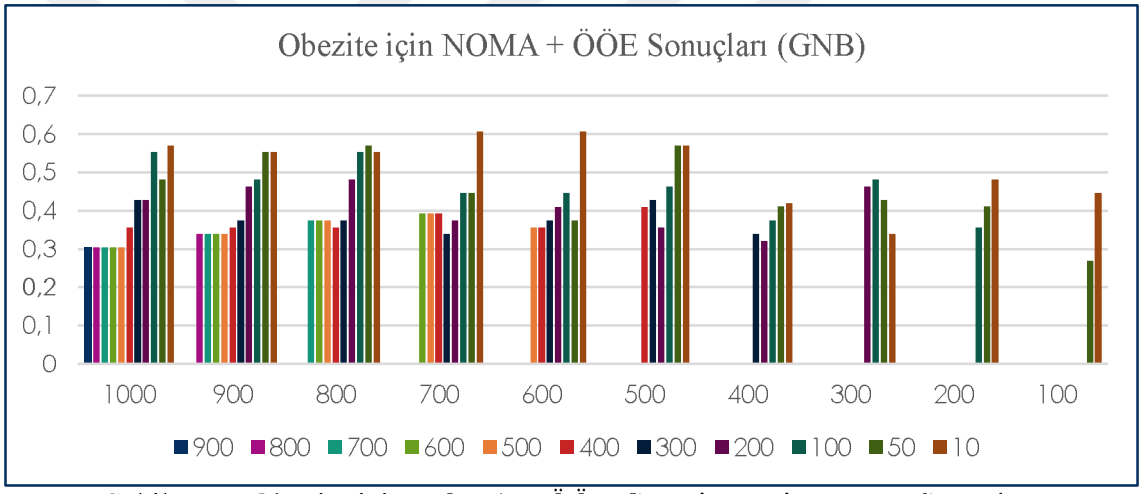
Şekil 5.29. İBH için NOMA + ÖÖE Gaussian Naive Bayes Sonuçları



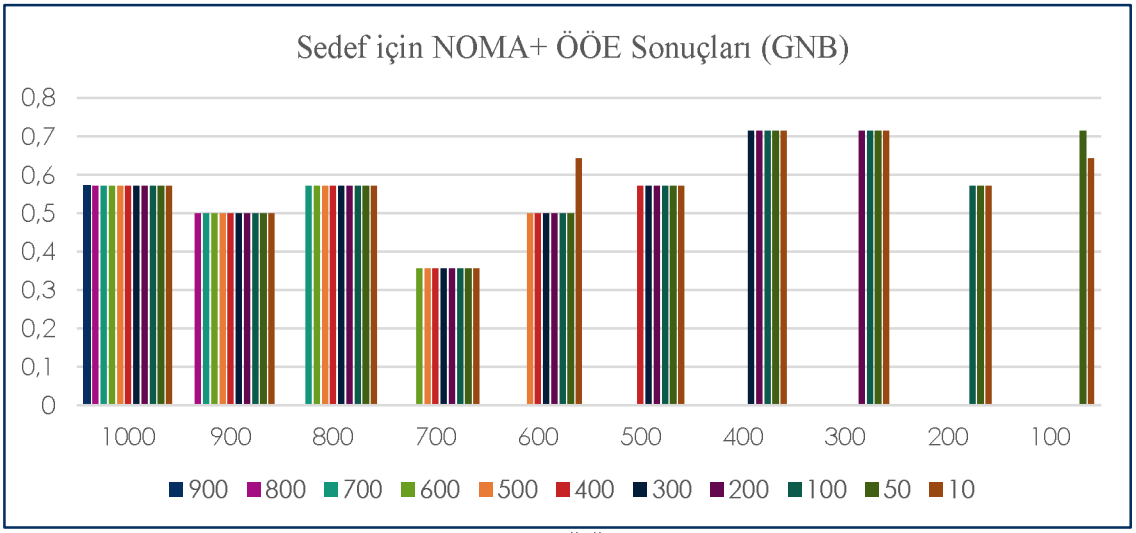
Şekil 5.30. Siroz için NOMA + ÖÖE Gaussian Naive Bayes Sonuçları



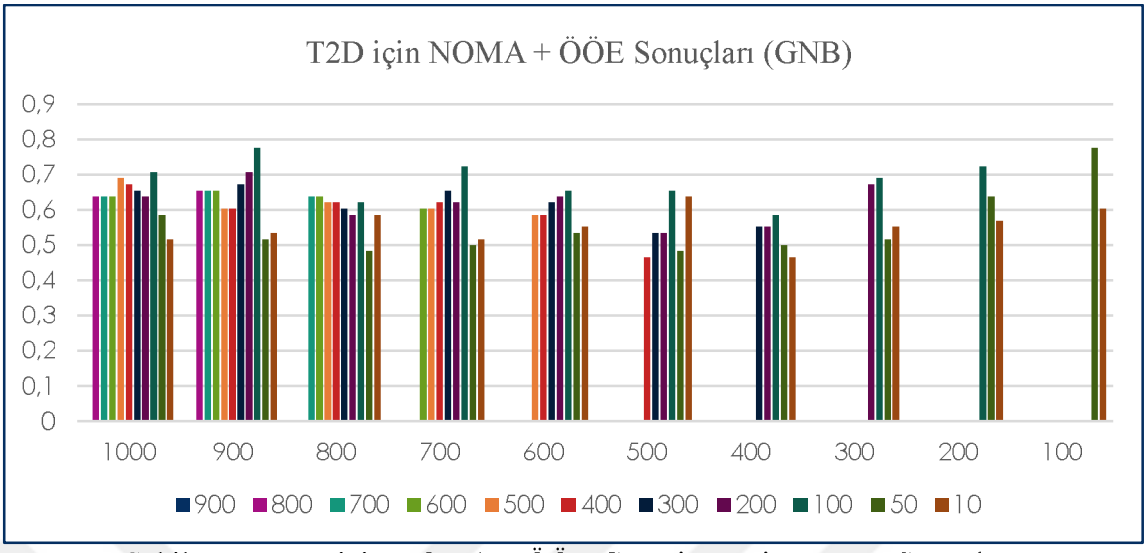
Şekil 5.31. KK için NOMA + ÖÖE Gaussian Naive Bayes Sonuçları



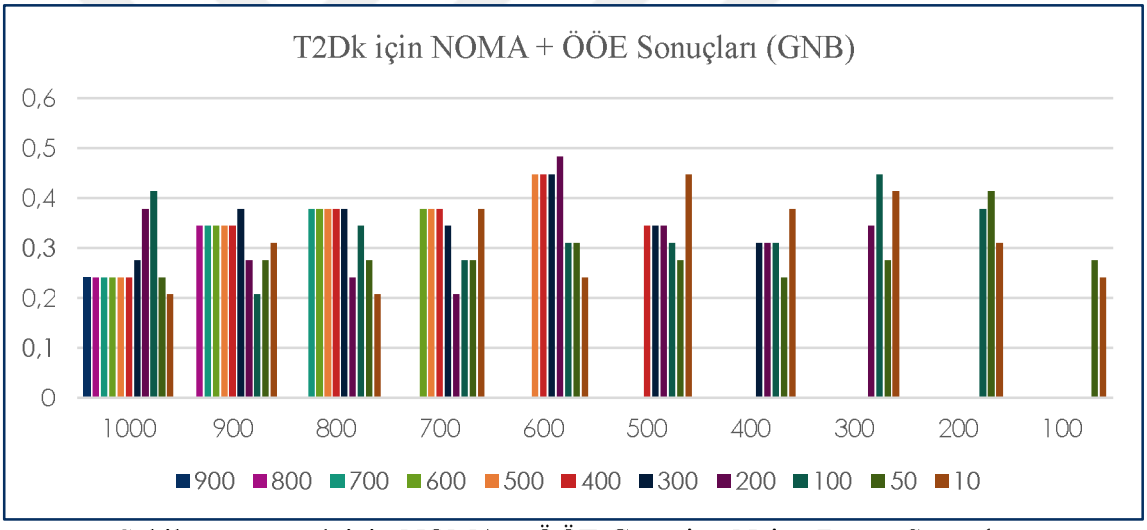
Şekil 5.32. Obezite için NOMA + ÖÖE Gaussian Naive Bayes Sonuçları



Şekil 5.33. Sedef Hastalığı NOMA + ÖÖE Gaussian Naive Bayes Sonuçları



Şekil 5.34. T2D için NOMA + ÖÖE Gaussian Naive Bayes Sonuçları



Şekil 5.35. T2Dk için NOMA + ÖÖE Gaussian Naive Bayes Sonuçları

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı, Soyadı: Umay Gülfem ELGÜN
Uyruğu: Türkiye (TC)
Doğum Tarihi ve Yeri: 1 Ocak 1994, Kayseri
Medeni Durumu: Bekâr
Tel: 0530 552 51 02
e-mail: gulfemelgun@gmail.com
Yazışma Adresi: Betül Ziya EREN Genom ve Kök Hücre Merkezi, Erciyes Üniversitesi 38039 Talas/KAYSERİ

EĞİTİM

Derece	Kurum	Mezuniyet Tarihi
Yüksek Lisans	EÜ Bilgisayar Mühendisliği	2019
Lisans	EÜ Bilgisayar Mühendisliği	2007
Lise	Nuh Mehmet Küçükçalık A.L, Kayseri	2011

YABANCI DİL

İngilizce