**T.C.**
**HACETTEPE ÜNİVERSİTESİ**
**SAĞLIK BİLİMLERİ ENSTİTÜSÜ**

# MİKRODİZİ GEN İFADE VERİLERİNDE FARKLI ÖZNİTELİK SEÇİM YÖNTEMLERİ İLE SINIFLAMA YÖNTEMLERİNİN PERFORMANSLARININ DEĞERLENDİRİLMESİ

**Özlem ARIK**

**Biyoistatistik Programı**
**DOKTORA TEZİ**

**ANKARA**
**2020**

# TEŞEKKÜR

Doktora eğitimim ve tez dönemim süresince çalışmalarımda tecrübeleri ve bilgisiyle her zaman yol gösteren, yardımı ve desteğiyle güç veren Saygıdeğer Danışman Hocam Prof. Dr. Erdem KARABULUT'a,

Tez izleme komitesi üyeleri olarak sağladığı değerli katkı ve eleştirileri için Sayın Prof. Dr. Meriç YAVUZ ÇOLAK, Sayın Dr. Öğr. Üyesi. Sevilay KARAHAN ve tezin değerlendirilmesindeki katkılarından dolayı diğer Sayın Jüri üyelerine,

Bilgilerini benimle paylaşan Biyoistatistik Anabilim Dalı öğretim üyeleri Sayın Prof. Dr. C. Reha ALPAR, Sayın Prof.Dr. A. Ergun KARAAĞAOĞLU, Sayın Prof. Dr. Pınar ÖZDEMİR, Sayın Doç. Dr. Jale KARAKAYA, Sayın Dr. Öğr. Üyesi Osman DAĞ ve araştırma görevlilerine,

Ara ara telefonla arayıp fikir alışverişi yaptığım Arş. Gör. Merve KAŞIKÇI`ya ve doktora yeterlilik sınavı kader ortağım Arş. Gör. Merve BAŞOL`a

Biyoistatistik Ana Bilim Dalı Sekreteri Şef Menekşe TARLA`ya,

Eğitim-Öğretim hayatımda bu günlere gelmemi benden daha çok isteyen Babam ve Anneme,

Doktora eğitiminin tez döneminde hayatımıza dâhil olan, yardım istediğim her an ve duasıyla her zaman bizim yanımızda olan Nazmiye ÖZEN`e,

Doktora sürecinde ve hayatımın her anında bana her zaman destek olan, yardımını hiçbir zaman esirgemeyen yol arkadaşım sevgili eşim Dr. İbrahim ARIK`a,

Saçlarının bir telini dünyaya değişemeyeceğim biricik çocuklarım canım oğlum Ahmet Erdem`e ve canım kızım Ayşenur`a,

Çok teşekkür ederim.

Kanser hastalığından vefat eden Anneannem ve tüm kanser hastalarına…

# ÖZET

**Arık, Ö., Mikrodizi Gen İfade Verilerinde Farklı Öznitelik Seçim Yöntemleri ile Sınıflama Yöntemlerinin Performanslarının Değerlendirilmesi, Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü Biyoistatistik Programı Doktora Tezi, Ankara, 2020.** İstatistik, biyoloji, bilgisayar, matematik ve genetik bilimlerini bir arada kullanan disiplinler arası bir bilim dalı olan biyoinformatik sayesinde, hangi anormalliklerin hangi hastalığa neden olduğu gösterilebilmektedir. Kanser hastalığında mikrodizi gen ifade verileri ile yapılan teşhis, sınıflama işlemleri, kanserin yapısında etkili olan genlerin belirlenmesi erken teşhiste önemlidir. Bu tez çalışmasında da akciğer, lenfoma, rahim ağzı, prostat, meme ve lösemi kanser türlerine ait mikrodizi gen ifade verileri üzerinde çalışılmıştır. Verilerin öznitelik sayısı fazla olduğu için daha az sayıda öznitelik ile çalışmak amacıyla varFilter, nsFilter, rf, lasso, rfe ve limma öznitelik seçim yöntemleri ele alınmıştır. Öznitelik seçimi yapılmış veri setlerinde Naive Bayes, Destek Vektör Makineleri, k-En Yakın Komşu ve Yapay Sinir Ağları sınıflama yöntemleri ile son yıllarda popülerlik kazanan Derin Öğrenme yöntemi ile sınıflama modelleri oluşturulmuştur. Veri setlerinde, ele alınan öznitelik seçim yöntemlerinin hangi sınıflama yöntemlerinde daha iyi olduğunu göstermek ve oluşturulan sınıflama modellerinin performanslarını karşılaştırmak için doğruluk, duyarlılık, seçicilik ve ROC eğrisi altında kalan alan değerleri elde edilmiştir. Genellikle lasso ve limma öznitelik seçim yöntemlerinde oluşturulan sınıflama modelleri diğer öznitelik seçim yöntemlerinde oluşturulan modellere göre daha başarılıdır. Derin Öğrenme yöntemi de klasik veri madenciliği sınıflama yöntemlerine göre çoğunlukla daha iyi performans göstermiştir. Veri setleri üzerinde öznitelik seçim yöntemi uygulamadan Derin Öğrenme sınıflama modelleri de elde edilmiştir. Öznitelik seçim yöntemlerini uygulayarak ve uygulamadan elde edilen Derin Öğrenme modellerinin performansları da karşılaştırılmıştır. Ayrıca benzetim çalışması yapılmıştır ve gerçek veri setlerine benzer sonuçlar elde edilmiştir.

**Anahtar Kelimeler**: Veri Madenciliği, Biyoinformatik, Öznitelik Seçimi, Mikrodizi, Gen, Kanser.

# ABSTRACT

**Arık, Ö., Evaluation of The Performance of Classification Methods with Different Feature Selection Methods in Microarray Gene Expression Data, Hacettepe University, Graduate School of Health Sciences, Biostatistics Program, PhD thesis, Ankara, 2020.** Bioinformatics is an interdisciplinary branch of science that combines statistics, biology, computing, mathematics, and genetics, and thanks to the analysis in bioinformatics, it can be shown which abnormalities causes which disease. In cancer disease, diagnosis with microarray gene expression data, classification procedures and identification of genes that are effective in the structure of cancer are of great importance for early diagnosis of the disease. In the thesis, microarray gene expression data of lung, kidney, lymphoma, cervical, prostate, breast and leukemia cancer types were studied. Since the number of features of the data is high, varFilter, nsFilter, rf, lasso, rfe and limma feature selection methods have been discussed. In filtered data sets, classification models were constructed with Naive Bayes, Support Vector Machines, k-Nearest Neighbor, Artificial Neural Networks and Deep Learning method, which has gained popularity in recent years. Accuracy, sensitivity, specificity and AUC were obtained to demonstrate which classification methods are better in the subject feature selection methods and to compare the performance and success of the generated classification models. Generally, classification models obtained in lasso and limma feature selection methods are more successful than models obtained in other feature selection methods. Deep Learning method is also generally more successful than classical data mining classification methods. Deep learning classification models were also obtained without applying the feature selection method on the datasets. It was compared whether there is a difference between the performances of deep learning models obtained by applying and without applying feature selection methods. In addition, implementation steps were carried out in four different simulation data. Similar results were obtained on real and simulation datasets.

**Keywords:** Data Mining, Bioinformatics, Feature Selection, Microarray, Gene, Cancer.

# İÇİNDEKİLER

## SİMGELER ve KISALTMALAR

| | |
|---|---|
| **A** | Adenin |
| **ALL** | Akut Lenfoblastik Lösemi |
| **AML** | Akut Miyeloid Lösemi |
| **BLAST** | Basic Local Alingment Search Tool |
| **C** | Sitozin |
| **CART** | Classification and Regression Trees |
| **DDBJ** | DNA Japonya Veritabanı (DNA Data Bank of Japan) |
| **DN** | Doğru Negatif |
| **DNA** | Deoksiribo Nükleik Asit |
| **DÖ** | Derin Öğrenme |
| **DP** | Doğru Pozitif |
| **DVM** | Destek Vektör Makineleri |
| **EAKA** | Eğri Altında Kalan Alan |
| **EBI** | The European Bioinformatics Institute (Avrupa Biyoinformatik Enstitüsü) |
| **EMBL** | The European Molecular Biology Laboratory (Avrupa Moleküler Biyoloji Laboratuvarı) |
| **FGED** | Functional Genomics Data Society (İşlevsel Genomik Veri Toplumu) |
| **G** | Guanin |
| **GDS** | Veri Seti |
| **GenBank** | Gen Bankası |
| **GEO** | Gene Expression Omnibus |
| **GPL** | Platform Kaydı |
| **GSE** | Veri Seri Kaydı |
| **GSM** | Örnek Kaydı |
| **kNN** | k-En Yakın Komşu |
| **lasso** | Least Absolute Shrinkage and Selection Operator |
| **limma** | Linear Model for Microarray Data |
| **MIAME** | Minimum İnformation About a Microarray Experiment |
| **NB** | Naive Bayes (Saf Bayes) |
| **NCBI** | National Center for Biotechnology Information (Ulusal Biyoteknoloji Bilgi Merkezi) |

| | |
|---|---|
| **NLM** | National Library of Medicine (Ulusal Tıp Kütüphanesi) |
| **OMIM** | Online Mendelian Inheritance in Man |
| **rf** | Random Forest (Rastgele Orman) |
| **rfe** | Recursive Feature Elimination (Özyinelemeli Öznitelik Eleme) |
| **RNA** | Ribo Nükleik Asit |
| **SVM-RFE** | Support Vector Machines-Recursive Feature Elimination (Destek Vektör Makineleri- Özyinelemeli Öznitelik Eleme) |
| **T** | Timin |
| **U** | Urasil |
| **YN** | Yanlış Negatif |
| **YP** | Yanlış Pozitif |
| **YSA** | Yapay Sinir Ağları |

# ŞEKİLLER

# TABLOLAR

# 1. GİRİŞ

Sosyolojik araştırmalar, üretim sektörü, devlet yönetimi gibi günlük hayatımızın neredeyse her alanında kullanılmakta olan istatistik bilimi ile verilerin analizine en çok ihtiyaç duyan tıp bilimi her zaman bir arada olmuştur (1). Sağlık alanında yapılan çalışmaların genel amacı sağlıklı bireylerin fiziki, ruhi ve sosyal açıdan iyi olarak hayatlarını devam ettirecek öneriler sunmak, hastalık durumlarında hastalığın nedeni, seyri, teşhis ve tanısı, risk etkenleri ve uygulanacak tedavi yöntemleri ile ilgili önerilerde bulunmaktır. Bu amaç doğrultusunda yapılan araştırmalarda ilgili verilerin toplanması, analiz edilmesi ve çıkan sonuçlar aracılığıyla doğru kararların verilmesinde ise biyoistatistik bilim dalından yararlanılmaktadır (2,3).

Biyoistatistiğin yanı sıra önemi her geçen gün artan biyoinformatik ise; biyoloji, bilgisayar, matematik, istatistik ve genetik alanlarını içermektedir. En karmaşık ve en önemli veri tipi olan genetik temelli verilerin anlaşılabilmesi için gelişen, disiplinler arası bilim dalı olan biyoinformatik; biyoloji dizi verilerini, gen içeriklerini ve sıralamalarını analiz etmeyi ve bu sayede makro moleküler yapıları ve fonksiyonları tahmin etmeyi amaçlamaktadır (4). Biyoinformatiğin en önemli araştırma konularından birisi gen analizidir. Bu alanda kullanılan DNA mikrodizi teknolojisi sayesinde genlerin bilinen ve bilinmeyen fonksiyonları tespit edilmektedir. Böylece hasta ile sağlıklı dokulardaki gen farklılıklarını ve benzerliklerini ortaya çıkarmak için tüm genlerin eşzamanlı ifadeleri belirlenir.

Hücrelerin yapısındaki genomun dinamik değişiklikleri ile ilerleyen kanserde genetik bozukluklar her bir kanser tipine özgün bir şekilde gelişim göstermektedir (5,6). Genom dizileme ve biyoinformatik alanlarındaki gelişmeler sayesinde kanserli hücrelerin genom yapısı ve kanserin iç dinamiklerinde yaşanan değişikliklerin anlaşılması ile daha iyi tanı, tedavi ve önleme çalışmaları yapılmaktadır. Günümüzde kanser hücrelerinde keşfedilen genetik değişiklikler sayesinde ilaç geliştirilmesi, sağlıklı hücrelerin korunmasına yardımcı olan, hedefe yönelik kanser tedavisi planlaması ve birçok kanserin gelişme riskine karşılık önlem alınması yapılabilmektedir (7).

Kanser gibi hastalıkların teşhis ve sınıflamasında, mikrodizi gen ifade verileri ile hastalıkla direkt ilişkili genleri bulmak büyük önem kazanmaktadır. Genetik

verilerin incelenmesinde ise verinin çok büyük boyutlarda olması sebebiyle klasik istatistiksel yöntemler ile anlamlı sonuçların elde edilmesi zor olduğu için çeşitli veri madenciliği yöntemleri ve bilgisayar programcılığı ile analiz yapılabilmektedir (8,9).

Tez çalışmasında satırda bireyleri, sütunda ise öznitelikleri (genleri) ve yanıt değişkenini (tümörün yapısı) içeren büyük boyuttaki mikrodizi gen ifade verileri kullanılmıştır. Öncelikle öznitelik seçim yöntemleriyle önemli ve anlamlı genler seçilmiştir. Daha sonra seçilen veriler ile hasta-sağlıklı sınıflamasının yapılmasıyla tümörün yapısında etkili olan genler belirlenmiştir. Kullanılan veri setlerinde ilk olarak veri madenciliğinde ön işleme adımı gerçekleştirilmiş olup, daha iyi başarıya sahip modeller elde etmek amacıyla az sayıda öznitelik ile çalışmak için bazı öznitelik seçim yöntemlerinden yararlanılmıştır. Sınıflama modellerini oluşturmak için derin öğrenme ile birlikte veri madenciliğinde sık kullanılan sınıflama yöntemleri tercih edilmiştir. Öznitelik seçim yöntemlerinin kullanılan sınıflama yöntemlerinin hangisinde daha iyi performans verdiğini göstermek amacıyla model performans ölçüleri gibi yöntemlerden yararlanılmıştır (10,11). Bu yöntemlerin, bazı kanser verileri ile Benzetim verileri üzerinde uygun bilgisayar programları ile uygulaması yapılarak yorumlanması hedeflenmiştir.

"Mikrodizi Gen İfade Verilerinde Öznitelik Seçim Yöntemlerinin Sınıflama Yöntemleri Başarısına Etkisi" başlıklı tez çalışması altı bölümden oluşmaktadır. Giriş bölümünde çalışma konusu hakkında kısa bilgiler verilerek çalışmanın amaçlarından bahsedilmiştir. Ayrıca tezde yer alan diğer bölümlerin de içeriği ile ilgili kısa açıklamalar yapılmıştır. Genel Bilgiler bölümünde; tez çalışmasının konusu ile bağlantılı olan veri, veri tabanı, veri madenciliği kavramı ve tarihçesi, biyoinformatik, biyoinformatikte sık kullanılan veri tabanları ve programları, temel genetik kavramlar, gen ve gen ifadesi, mikrodizi teknolojisi ve gen ifade verileri ile kanser sınıflandırması konuları ile ilgili bilgiler verilmiştir. Gereç ve Yöntem bölümünde; çalışmada kullanılan öznitelik seçim yöntemleri, sınıflama yöntemleri, model performans ölçüleri ile gerçek ve Benzetim veri setlerinde hakkında açıklamalar yapılmıştır. Bulgular bölümünde de; R ve Matlab programları aracılığıyla ilk üç bölümde bahsedilen konular ile ilgili uygulama çalışması yapılmış olup, gerçek ve Benzetim veri setlerine ait sonuçlar tablo ve şekiller aracılığıyla verilmiştir. Tartışma bölümünde; çalışmanın başında belirlenen amaçlara paralel

olarak çalışmanın sonunda elde edilen yorumlara yer verilmiştir ve literatürde yer alan diğer çalışmaların sonuçları ile tez çalışmasının sonuçları karşılaştırılmıştır. Son bölüm olan Sonuç ve Öneriler'de ise çalışma ile elde edilen sonuçlar kısa ve net olarak açıklanmış olup daha sonraki çalışmalara tavsiye olacak şekilde önerilerde bulunulmuştur.

# 2. GENEL BİLGİLER

## 2.1. Veri Madenciliği Kavramı ve Tarihçesi

Son yıllarda teknolojinin ve bilgi sistemlerinin gelişmesiyle kamu ve özel kuruluşlara ait veri tabanlarında çok hızlı bir biçimde veriler depolanmaktadır (12). Büyük miktarda biriken ham veri setlerinden anlamlı, değerli ve faydalı bilgilerin ortaya çıkarılmasına veri madenciliği denir. İlk kez 1980`li yıllarda literatüre giren veri madenciliği günümüzde yaygın bir şekilde kullanılmaktadır (13,14). Klasik istatistiksel yöntemlerin büyük miktardaki veriler için geçerli ve güvenilir sonuçlar verememesiyle birlikte Tukey tarafından 1977 yılında ortaya atılan Keşfedici Veri Analizi yöntemleri ile daha iyi sonuçlar elde edilmeye başlanır ve veri madenciliği kavramının temeli atılmış olur (15,16). Öncelikle istatistik, makine öğrenme ve yapay zeka gibi kavramların bir araya gelmesiyle disiplinler arası bir alan olan veri madenciliği gün geçtikçe daha çok kabul görmektedir (17,18). Verilerin ve veriler arasındaki bağlantı ve ilişkilerin incelenmesinde kullanılan klasik istatistiksel yöntemler veri madenciliği yöntemlerinin temelini oluşturur ve araç olarak kullanılır (18).

İnsan gibi düşünebilme ve sezgisel yaklaşımı temel alan ve yüksek kapasitede güçlü bilgisayarların kullanılmasını gerektiren yapay zeka sayesinde farklı sorunlara çözümler getirilir (13,18). İstatistik ve yapay zekanın birlikte kullanıldığı makine öğrenmesi ile ileri seviyede çalışmalar yapılarak veriler değerlendirilir, verilerden anlamlı sonuçlar çıkarılır ve bu sonuçlara ait kararlar alınır. Genel olarak büyük veri yığınları içerisinden yararlı bilgiye ulaşmak için istatistik, yapay zeka ve makine öğrenme gibi disiplinlerden yararlanarak gerçekleştirilen işlemlerin tümüne veri madenciliği denilebilir (19-21).

### 2.1.1. Veri Madenciliği Süreci

Büyük veri setleri üzerinden önemli ve faydalı bilgileri çıkararak kullanıcının başarılı sonuçlar elde etmesine yardımcı olan veri madenciliği belli bir süreç akışı içermektedir (21). Veri madenciliği sürecine ait işlemler Şekil 2.1.`de gösterilmiştir (22) .


**Şekil 2.1.** Veri madenciliği süreci.

**Problemin Belirlenmesi**

İyi bir performansa sahip sonuç elde etmek için ayrıntılı ve net olarak problem tanımı yapılmalıdır. Problem tanımı ile hedefler, gereklilikler, kısıtlamalar ve stratejiler belirlenmelidir (13,23).

**Veriyi Anlama**

Ulaşılmak istenen sonuçlar ile kullanılması planlanan veriler arasında bir ilişki oluşturulur. Veriler ile amaçlar kesinleşir, amaçlar ile veriler tekrar gözden geçirilir. Böylece verinin kalitesi ve yeterliliği değerlendirilir (24).

**Veriyi Hazırlama**

Modelin kurulması amacıyla eldeki ham veri seti üzerinde uygulama yapabilmek için veriyi kullanılabilir hale getirmektir. Veriyi hazırlama aşamasında; veri temizleme, veri dönüştürme, veri birleştirme, veri indirgeme gibi veri ön işleme adımları kullanılır. İlerleyen aşamalarda bir sorunla karşılaşmamak için veri hazırlama aşamasına dikkat edilmeli ve gereken önem verilmelidir (12,25).

**Modelin Kurulması**

Belirlenen problemin çözülmesi için kullanılan veri üzerinde uygun veri madenciliği algoritmalarını uygulayarak en iyi ve en doğru sonucu veren modeli oluşturmaktır (26,27).

**Modelin Değerlendirilmesi**

Modelin kurulması ile elde edilen sonuçların yüksek performansta belirli özelliklere sahip olması beklenir. Faydalı, yeni, anlamlı ve güvenilir sonuçlar bir modelde olması istenilen belli başlı özelliklerdir. Çeşitli yöntemler aracılığıyla model değerlendirilerek modelin performansı ortaya çıkarılır (21,28,29).

**Modelin Kullanılması-Uygulama**

Kurulan ve değerlendirilen model, hedefe yönelik olarak başka bir uygulamanın aracı olarak ya da tek başına bir uygulama olarak da kullanılabilir (21).

**Modelin İzlenmesi**

Kurulan, değerlendirilen ve kullanılmakta olan modellerin zamanla gerçekleşen değişiklikler sebebiyle yeniden düzenlenmesi gerekebilir. Dolayısıyla model takibi ve izlenmesi yapılarak güncellenmelidir (21).

### 2.1.2. Veri Madenciliği Uygulama Alanları

Büyük veri setlerinin yer aldığı bankacılık, iletişim, sigorta, yüzey analizi ve coğrafi bilgi sistemleri, görüntü tanıma ve robot görüş sistemleri, uzay bilimleri ve teknolojisi, meteoroloji ve atmosfer bilimleri, sosyal bilimler ve davranış bilimleri, metin madenciliği, internet madenciliği, kimya, biyoloji, sağlık, tıp ve genetik gibi birçok alanda veri madenciliği uygulanmaktadır (18). Tıp ve genetik alanında ise gen haritasının çözümlenmesi ve genetik hastalıkların belirlenmesi, kanserli hücrelerin tespiti ve kanserli hücreyi etkileyen genlere karar verilerek tedavi sürecinin planlanmasında veri madenciliğinden yararlanılmaktadır (13).

### 2.1.3. Veri Madenciliği Yöntemleri

Veri madenciliğinde kullanılan yöntemler genel olarak Şekil 2.2.`de gösterildiği gibi tahmin edici ve tanımlayıcı olmak üzere ikiye ayrılmaktadır (30,31). Verinin sahip olduğu özellikleri kullanarak ve veriler arasındaki benzerlik, ilişki gibi ölçülerden yararlanarak karar verme sürecinde belirleyici olabilecek özelliklerin oluşturulmasında kullanılan yöntemler tanımlayıcı yöntemlerdir. Bu yöntemlere örnek olarak kümeleme ve birliktelik kuralları yöntemleri verilebilir (21).


**Şekil 2.2.** Veri madenciliği yöntemleri.

Herhangi bir olaya ait sonuçları bilinen veriler aracılığıyla oluşturulan model sayesinde problem ile ilgili karşılaşılan yeni durumların sonuçlarının elde edilmesinde tahmin edici yöntemler kullanılmaktadır. Regresyon ve sınıflama; tahmin edici yöntemler arasındadır. Sınıflama yöntemleri sayesinde hangi sınıfa ait

olduğu bilinen veriler ile bir sınıflama modeli elde edilir ve yeni eklenen verilerin hangi sınıfa dâhil olacağına karar verilir (18).

### 2.2. Biyoinformatik

20. yüzyılın ikinci yarısında biyolojik bilginin çok fazla artmasıyla oluşan karmaşık bilginin işe yarar hale gelmesi için güçlü araçlara ihtiyaç duyulmuştur (5). Uygulamalı matematik, bilgisayar bilimleri, istatistik, biyoloji ve genetik alanlarını kapsayan disiplinler arası bir bilim dalı olan biyoinformatik sayesinde de büyük boyuttaki biyolojik veriler düzenlenir, analiz edilir ve daha anlaşılır hale getirilir (32,33).

Bilgisayarla moleküler grafiklerin çizimine ait ilk makalenin 1966 yılında Scientific American dergisinde yayınlanması biyoinformatik için gerçek anlamda başlangıç sayılabilir (5). Biyoinformatik terimi 1980`li yılların ortalarından sonra kullanılmaya başlamış ve İnsan Genom Projesi sonucu ortaya çıkan genetik bilginin işlenmesi için biyoinformatiğe olan ihtiyaç artmıştır (5,34). Ekim 1990`da başlayan ve 13 yıllık uluslararası bir emeği içeren proje, 30-35 bin insan geninin ortaya çıkarılmasını ve biyolojik çalışmalarda kullanılabilecek şekilde hizmete sunulmasını temel amaç edinmiştir (5,35).

Genom projeleri, yapısal ve fonksiyonel genomik, kıyaslamalı genomik, proteomik, hesaplamalı biyoloji ve mikrodiziler biyoinformatiğin ana konularıdır (33,36). Dolayısıyla biyoinformatik biliminin çoğunlukla üzerinde çalıştığı veri türü genetik veridir ve buna bağlı olarak gen ifade verisidir. Biyoinformatiğin en önemli uygulamalarından biri, farklı yapılardaki biyolojik bilgilerin yönetilmesi ve etkili kullanımı için yeni araçlar geliştirmek; diğeri ise biyolojik verileri, kullanıcıların ulaşabileceği ortamlarda saklayabilecek ve yeni gelen verileri var olanların yanına ekleyerek düzenleyecek algoritma ve yazılımlar oluşturmaktır (32,33).

#### 2.2.1. Biyoinformatikte Sık Kullanılan Veri Tabanları ve Programları

Biyoinformatikte araştırmacıların yararlanmasına açık olan ve nükleotid dizi bilgilerinin saklanması, düzenlenmesi ve kullanılması için işbirliği ile çalışan GenBank (Gen Bankası; ABD-Maryland), EMBL (Avrupa Moleküler Biyoloji Laboratuvarı; İngiltere-Hinxton) ve DDBJ (DNA Japonya Veritabanı (DNA Data

Bank of Japan); Japonya-Mishima) olmak üzere üç kurum vardır. 1988`de Maryland`da kurulan ve NLM (National Library of Medicine)`nin bir kolu olan NCBI (National Center for Biotechnology Information) ise web`e dayalı en önemli biyolojik veritabanıdır. NCBI içerisinde Pubmed makalelerini, Genetics ve Biochemistry, EMolecular Biology of the Cell gibi kitaplar bulunmaktadır. Genler ve genetik hastalıklarla ilgili ayrıntılı biyoteknolojik ve tıbbi bilgilerin bulunduğu bir servis olan OMIM (Online Mendelian Inheritance in Man) ise NCBI`ın alt hizmetlerinden biridir (5,37).

EBI (The European Bioinformatics Institute) gibi enstitülerle iyi projelerde yer alarak biyoinformatiğin gelişimine önemli katkıları olan veri tabanlarından biri de Ensemle`dır. Özellikle ökaryot genomları üzerine çalışmaktadır (5,38). BLAST (Basic Local Alingment Search Tool) programı ise biyoinformatikte önemli bir yeri olan sık kullanılan bilgisayar programıdır. BLAST`ın amacı, bilgisayar aracılığıyla genom verilerini analiz etmek için bilgisayar programları geliştirmek ve bir kaynakta moleküler biyoloji ile ilgili bilgileri toplamak olmuştur. BLAST dizi eşleştirme programı ile eldeki DNA dizisi, ayrıntılı analiz edilebilir (37,38).

### 2.3. Temel Genetik Kavramlar: DNA, RNA, Gen

1800`lü yıllarda Mendel`in çalışmalarıyla başlamış olan genetik çalışmalarının asıl materyali ise tüm hücreli canlılarda bulunan, canlının gelişimi için ihtiyacı olan biyolojik bilgiyi taşıyan ve bu bilginin kendinden sonrakilere aktarılmasında görev alan DNA (Deoksiribo Nükleik Asit)`dır (39-41). Kalıtım molekülü olarak da bilinen DNA`nın yapısı James Watson ve Francis Crick tarafından ilk kez 1953 yılında keşfedilmiştir (33,42,43).

Şekil 2.3.`te gösterildiği gibi sarmal bir yapısı olan DNA; Guanin (G), Sitozin (C), Timin (T) ve Adenin (A) olmak üzere dört çeşit nükleotidden oluşmaktadır. Nükleotidlerin her biri; bir fosfat grubu, bir organik baz ve beş-karbonlu şekerden meydana gelmektedir. Sadece T ile A birbirlerine bağlanırken, G ile de C birbirlerine bağlanabilmektedir (44). Birbirlerine bağlanan bu nükleotidlere, nükleotid çifti (baz çifti) adı verilmektedir (40).


**Şekil 2.3.** DNA`nın yapısı.

DNA`da taşınan genetik bilginin proteine dönüştürülmesi sürecinde önemli rolü olan RNA (Ribo Nükleik Asit)`nın ise üç türü vardır. Bunlar; taşıyıcı RNA, ribozomal RNA ve mesajcı RNA`dır. Her birinin farklı işlevi vardır. Ancak genel olarak RNA sayesinde DNA'dan elde edilen bilgi taşınır (transkripsiyon) ve protein sentezi gerçekleştirilir (translasyon) (41,45). DNA'daki T nükleotidi yerine RNA'da Urasil (U) bulunur ve RNA sarmal yapıda değildir, tek ipliktir (33,41).

RNA aracılığı ile hücrenin tüm aktivitelerinin gerçekleşmesinden sorumlu moleküllerin yani proteinin yapısını belirleyen gen ise DNA`nın bir parçası olmakla birlikte bir canlının her türlü özelliklerini belirleyen en temel kalıtsal birimdir (33,41). İnsanda bulunan yirmi üç kromozomda yer alan genlerin her birinin fizyolojik ve morfolojik görevleri bulunmaktadır. Genotip ise bir organizmanın genetik yapısıdır (33,43,45).

### 2.3.1. Gen ve Gen İfadesi

1800`lü yıllarda Gregor Mendel`in bitkisel özelliklerin nesilden nesile aktarılması ile ilgili bilimsel çalışmalarına ait yaptığı yayınlar, kalıtım ve genlerle ilgili olan genetik biliminin temelini oluşturmaktadır. Mendel`in çalışmaları ile birlikte yeni bir dönem başlamış ve genetik bilginin sırasıyla DNA, RNA ve proteine aktarıldığı gösterilmiştir (46-48). Genetik kelimesinin kökünü oluşturan gen ise hücre çekirdeğindeki kromozomlarda yer alan, fiziksel özellikleri tanımlamak gibi genetik görevleri olan, başlangıç ve bitiş noktaları bulunan DNA bölgeleridir. İnsanların her bir hücresinde yaklaşık olarak 25000 gen bölgesi bulunmaktadır ve her

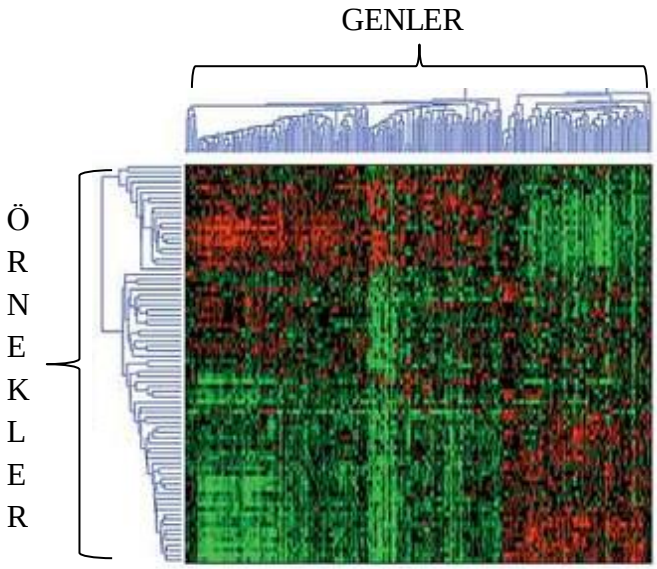
birinin özellikleri farklıdır (49). Kromozom üzerinde yer alan genlerin konum bilgisine de lokus denir ve her bir genin lokusu farklıdır (41).

Üzerinde çalışılan bir durum için genlerin aktif olup olmadıklarını aktif ise ne kadar aktif olduklarını gösteren gen ifadesi (gen ekspresyonu) ise genlerin DNA`dan RNA yapılarına ve proteine dönüşmesi aşamasıdır. Proteinin fazla üretilmesi ile gen ifade düzeyinin yüksek olması arasında pozitif yönde doğrusal bir ilişki vardır (48). Organlarımızın tümü aynı genetik materyali içerir. Ancak farklı hücrelerde genlerin farklı ifade edilmeleri sebebiyle meme, akciğer, beyin gibi hücreler birbiriyle aynı fonksiyonlara sahip değildirler (10,50).

### 2.3.2. Mikrodizi Teknolojisi ve Gen İfade Verileri: Veri formatı, Veri işleme, Gen ifade veri matrisi

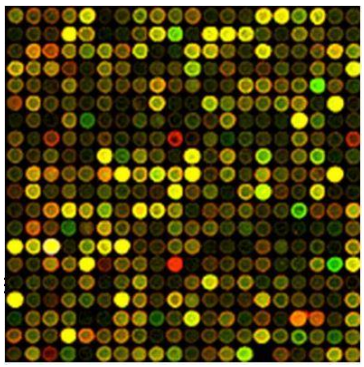
Altmış yedi yıl önce James Watson ve Francis Crick tarafından 21 Şubat 1953`te resmen keşfedilen DNA ile birlikte canlıların yaşamı üzerinde genetik kodun etkilerini araştırmak önemli bir konu haline gelmiştir. İlerleyen teknoloji ile birlikte atılan adımlardan biri olan mikrodizi teknolojisi sayesinde de bir organizmaya ait genomun ifadeleri tek bir seferde incelenebilmektedir (42, 51). Hızlı olmak, genlerin hasta ve sağlıklı hücrelerdeki etkinliğini incelemek, hastalıkları kategorilere ayırabilmek gibi özellikler mikrodizi teknolojisinin avantajları iken; pahalı olması, aynı anda çok fazla veri elde edilmiştirği için tüm verilerin analizinin yapılmasının uzun sürmesi ve yorumlamasının karmaşık olabilmesi ise mikrodizi teknolojisinin dezavantajlarıdır. Elde edilen verilerin büyüklüğü ve karmaşıklığından dolayı analiz ve yorumlama için hesaplamaya dayalı genomik yaklaşımlara ihtiyaç artmıştır. Biyoistatistiksel analizlerin ve biyoinformatiğin bu ihtiyaçların giderilmesinde büyük bir yeri vardır (6,52).

Mikrodizi teknolojisi sayesinde yapılan gen ifade analizinde DNA çipleri, başta insanlar olmak üzere canlılardaki farklı genlerin ifade düzeylerinin belirlenmesi için Şekil 2.4.`te de gösterildiği gibi kullanılmaktadır (6,9). Gen ifade analizi ile elde edilen gen ifade verileri kullanılarak genlerin nerede ve ne zaman aktif oldukları böylece kendilerini ne kadar ifade ettikleri gösterilir (6,53).




**Şekil 2.4.** Gen ifade verisinin yansıyan görüntüsü.

Mikrodiziler, binlerce farklı DNA parçalarının sentezlendiği ya da yerleştirildiği binlerce noktadan (spot) oluşan çipler olarak da ifade edilmektedir. Cam, plastik ya da silikondan yapılan katı yüzey çip olarak adlandırılır. Prob ise çipin yüzeyindeki her bir noktadır (51,54). Çalışılacak olan hedef organizmanın tüm genlerini çok küçük bir alanda incelemek ve binlerce genin ifade düzeylerini aynı anda çalışmak için çipler üzerindeki küçük alanlara yerleştirilen problardan yararlanılır. Kullanılan yöntem ise bir genomda yer alan bilgilerin baz eşleşmesi kuralına bağlı hibridizasyon temellidir (51,54,55). Çipler üzerine yerleştirilen genler belli işlemlerden geçtikten sonra spotlarda ifade seviyelerini yansıtırlar. Şekil 2.5.`de gösterildiği gibi spotlar üzerindeki renkler genlerin her birinin kendini ifade etme düzeyidir (56).



**Şekil 2.5.** DNA mikrodizisi.

Gen ifade edilememiş ya da okunamamış ise prob siyah renkte gözükecektir. Yeşil, sağlıklı bireyleri; kırmızı ise hasta bireyleri işaret etmektedir. Hasta ya da sağlıklı olma durumu birbirine yakın ise sarı ile gösterilmektedir. Bilgisayar çözümlemesiyle bu renkler sayısal değerlere dönüştürülerek analiz için uygun hale getirilir (11,56). Bu tez çalışmasında kullanılan gerçek veri setleri de bu şekilde elde edilen verilerden oluşmaktadır.

Deneysel dizayn, deney platformu ve yöntemlerdeki çeşitlilikler nedeniyle mikrodizi deneylerine ait verilerin gösteriminde uluslararası geçerli bir prosedür olmadığı için mikrodizi deneylerinin yapılış şekillerinin ve verilerinin belli bir standartta ulaşılabilir olması amacıyla MIAME (Minimum İnformation About a Microarray Experiment) kriterlerine bağlı olarak mikrodizi veri tabanları, veri ve bilgi paylaşımı yapmaktadır. 2017 yılında FGED (Functional Genomics Data Society) tarafından ortaya atılan MIAME kriterlerine göre bir mikrodizi deneyi açıklanırken; işlenmiş veri, ham veri gibi deney ile ilgili temel bilgiler, deneysel tasarım, genomik koordinatlar, laboratuvar ve veri işleme adımları gibi mikrodizi deneylerine ait özellikler yani deney hakkında olması gereken minimum bilgi mikrodizi veri tabanlarında ulaşılabilir olmalıdır (11,53,55).

Halka açık mikrodizi veri tabanlarından en önemlileri; Amerika kökenli ve dünyanın en kapsamlı biyolojik veri tabanı olan NCBI'nın altında bulunan GEO (Gene Expression Omnibus) ile Avrupa kökenli büyük ve kapsamlı bir biyolojik veri tabanı olan EBI'nın altındaki ArrayExpress`dir (55). Veri analizinin yapılabilmesi için mikrodizi deneyi sonucunda elde edilen özellikle sarı, kırmızı ve yeşil renklerin yer aldığı mikrodizi resim verisi görüntü işleme, arka plan düzeltme, normalleştirme, özetleme gibi ön işleme adımlarından geçirilerek sayısal değerler elde edilir (57-59).

İlk olarak renklerle ifade edilen gen ifade verilerinin sayısal değerlerinin elde edilmesi ile gen ifade veri matrisi oluşturulur. Hasta ve sağlıklı olmak üzere iki adet sınıfa ait örneklem büyüklükleri $n_1$ ve $n_2$ ile m adet genin bulunduğu gen ifade veri matrisi Şekil 2.6.`da verildiği gibidir. $H_1,H_2,\ldots,H_n$ hasta bireyleri; $S_1,S_2,\ldots,S_n$ sağlıklı bireyleri; $GEN_1,GEN_2,\ GEN_3,\ldots,GEN_m$ ifadeleri ölçülen genleri ve $Y_{ij}$'ler ise genlerin ifade düzeylerini göstermektedir. Satırda genler, sütunda örnekler olmak üzere mxn boyutlu gen ifade veri matrisi satır veri yapısı biçimindedir (9,60).

Öznitelik seçimi ve sınıflama gibi analizleri uygulayabilmek için gen ifade veri kümesinin transpozu alınarak satırlara örnekler, sütunlara genler yerleştirilir.

| | Hasta Grup | | | | Sağlıklı Grup | | | |
|---|---|---|---|---|---|---|---|---|
| | **$H_1$** | **$H_2$** | **. . .** | **$H_n$** | **$S_1$** | **$S_2$** | **. . .** | **$S_n$** |
| **$GEN_1$** | $Y_{11}$ | $Y_{12}$ | . . . | $Y_{1n}$ | $Y_{1n+1}$ | . . . | . . . | . . . |
| **$GEN_2$** | $Y_{21}$ | $Y_{22}$ | . . . | $Y_{2n}$ | . . . | . . . | . . . | . . . |
| **$GEN_3$** | $Y_{31}$ | $Y_{32}$ | . . . | $Y_{3n}$ | . . . | . . . | . . . | . . . |
| . . . | . . . | . . . | . . . | . . . | . . . | . . . | . . . | . . . |
| **$GEN_m$** | $Y_{m1}$ | $Y_{m2}$ | . . . | $Y_{mn}$ | . . . | . . . | . . . | $Y_{mn+n}$ |

**Şekil 2.6.** Gen ifade veri matrisi yapısı.

Sınıf sayısı ikiden fazla ve her bir sınıftaki örnek yani birey sayıları birbirinden farklı olabilir. Farklı senaryolar olduğunda analiz için kullanılacak olan yöntemlerde değişiklik gösterecektir. Elde edilen gen ifade veri matrisi üzerinde veri madenciliği ve istatistiksel yöntemlerin uygulanması ile kanser gibi hastalıklar üzerinde etkili genler belirlenebilir, ortak işleve sahip genler kümelenebilir, bireylere ait hasta-sağlıklı sınıflaması yapılabilir (48).

### 2.3.3. Gen İfade Verileri ile Kanser Sınıflandırması

Kanser, hücresel düzeyde genetik bir hastalıktır. Özellikle hücrelerin nasıl büyüdüklerini ve bölündüklerini yani işleyiş şeklini kontrol eden genlerde meydana gelen bazı değişiklikler kansere neden olmaktadır. Hücrelerdeki işlerin çoğunu yerine getiren proteinlerin oluşturulma talimatlarını taşıyan genlerin bazılarında meydana gelen farklılıklar hücrelerin normalden farklı büyümesine böylece kansere sebep olmaktadır. Kanser hücrelerinin genetik değişimi normal hücrelere göre daha fazladır ve çok hızlı mutasyon geçirme özelliklerine sahiptir. Kanserin genetik yapısı herkeste farklı olmakla birlikte aynı tümör içinde bile farklı türlerde mutasyonlara sahip hücre çeşitleri bulunmaktadır (41). Genel olarak kanserin dört evresi vardır ve ilk evrelerde fark edilirse tedavi şansı daha yüksek olur. Dolayısıyla erken teşhis çok önemlidir. Kanser teşhis ve sınıflama aşamasında da gen ifade verileri ile çalışmak büyük önem kazanmaktadır (53).

2000`li yılların başında gen ifade verileri ile yapılan kanser çalışmalarında, kanser alt sınıflarının bulunması ve bilinen sınıflara bireylerin atanması konuları incelenmiştir (61). Literatürde yer alan çalışmalarda genel olarak; yumurtalık kanseri, lenfoma, kolon kanseri, ALL (Akut Lenfoblastik Lösemi), AML (Akut Miyeloid Lösemi), mide kanseri, akciğer kanseri, merkezi sinir sistemi kanseri ve pankreas kanseri gibi kanser türlerine ait veri setleri ile çalışılmıştır (8).

# 3. GEREÇ VE YÖNTEM

## 3.1. Öznitelik Seçim Yöntemleri

Özniteliğin; istatistikte sık kullanılan, örnekten örneğe farklı değerler alan özellik ya da durumları ifade eden değişken kavramı ile benzer bir tanımı vardır. Herhangi bir veri setini oluşturan örneklemdeki örneklerin niteliklerinin, özelliklerinin her birisi özniteliktir. Genellikle yüksek boyutlu veriler arasında olan mikrodizi gen ifade verilerinde öznitelik (gen) sayısı çok, örnek sayısı ise oldukça azdır. Ancak teoride çıkarım yapılacak bir veri setinde örnek sayısının öznitelik sayısına göre üstel olarak artıyor olması gerekir. Dolayısıyla istenmeyen veri yapısına sahip olan mikrodizi gen ifade verileri ile çalışmak bir sorundur (62).

Mikrodizi gen ifade verileri gibi büyük veri seti üzerinde yapılması planlanan uygulamanın amacına yönelik olarak, özniteliklerin tümünü kullanmak yerine gereksiz özniteliklerin çıkarılarak, en faydalı ve en önemli özniteliklerin seçilmesiyle orjinal veri setini temsil edebilecek en iyi öznitelik alt kümesinin belirlenmesi işlemine öznitelik seçimi denir. Öznitelik seçimi sayesinde hız ve başarı performansı açısından da daha iyi modeller elde edilir (63,64).

Öznitelik seçimini gerçekleştirmek için kullanılan farklı yöntemler vardır. Genel olarak bu yöntemler istatistiksel yöntemler, sarmal yöntemler ve gömülü yöntemler olmak üzere üçe ayrılmaktadır (65). Sadece istatistiksel bilgiyi kullanarak seçim yapan istatistiksel yöntemlere filtreleme yöntemleri de denir. Fisher skor, bilgi kazancı, gini katsayısı, ki-kare, kazanç oranı gibi yöntemler filtreleme yöntemlerine örnek olarak verilebilir. Bu yöntemlerde önce öznitelik seçimi yapılır, daha sonra veri madenciliği yöntemleri uygulanır. Sezgisel arama, genetik algoritma, parçacık sürü optimizasyonu gibi yöntemler ise sarmal yöntemler arasındadır. Bu yöntemlerde öznitelik seçimi için veri madenciliği yöntemleri bir araç olarak kullanılmaktadır. Öznitelik seçim yöntemi ve veri madenciliği yönteminin aynı anda uygulandığı yöntemlere ise gömülü yöntemler denir. En yaygın kullanılanları ise karar ağaçları ve svm-rfe (support vector machines-recursive feature elimination)`dır (63,66).

Genel olarak üç başlıkta toplanan öznitelik seçim yöntemlerinin her birinin avantajları ve dezavantajları bulunmaktadır. Sınıflama başarısı, hız ve veri madenciliği yöntemine bağlı olma açısından yöntemler karşılaştırıldıklarında

filtreleme yöntemleri herhangi bir veri madenciliği yöntemine bağlı olmadan hızlı bir şekilde çalışır ve sınıflama başarısı değişkenlik göstermektedir. Sarmal yöntemlerin ise veri madenciliği yöntemlerine bağlılığı vardır. Sınıflama başarısı yüksektir fakat hesaplama karmaşıklığı olduğu için daha yavaş sonuç vermektedir. Sarmal yöntemler gibi gömülü yöntemler de veri madenciliği yöntemlerine bağlı olarak çalışırlar. Sınıflama performansları ise değişkenlik göstermektedir ve sarmal yöntemlere göre daha hızlı çalışmaktadırlar (65). Dolayısıyla gömülü yöntemler sarmal yöntemler ile filtreleme yöntemlerin bir birleşimidir denilebilir. Özellikle mikrodizi gen ifade verilerinde gömülü yöntemlerin kullanımı daha idealdir (67).

Bu tez kapsamında da kullanılan veri setlerinde çok sayıda öznitelik olduğu için sınıflama sonucunu daha fazla etkileyecek olan önemli ve anlamlı özniteliklerin bulunması amacıyla R programı içinde yer alan altı farklı öznitelik seçim yöntemi kullanılmıştır. Ele alınan yöntemlerden bazıları gömülü yöntemler arasındadır. Kullanılan yöntemlere ait açıklamalar bir sonraki bölümlerde yer almaktadır.

### 3.1.1. ExpressionSet Nesnesine Uygulanabilen Öznitelik Seçim Yöntemleri

Özellikle mikrodizi gen ifade verileri gibi yüksek boyutlu veri setlerinde öznitelik seçim yöntemleri uygulanırken algoritmaların bazılarının da yapısı sebebiyle bilgisayar bellek sorunlarıyla karşılaşılabilir. Bilgisayar belleğinin analiz yapmaya izin vermediği durumlarda mikrodizi gen ifade verileri üzerinde öznitelik seçimi, sınıflama, kümeleme gibi uygulamaların yapılabilmesi için oluşturulmuş, kullanılabilir R paketleri bulunmaktadır. Çok sayıda mikrodizi veri setleri başta olmak üzere genomik veri kaynaklarını ve açık kaynak kodlu analiz araçlarını içeren Bioconductor ortamı ile R programı entegre edilmeye uygundur. www.bioconductor.org adresinden Bioconductor uygulamasına ve gerekli R paketlerine ulaşılabilir (68). Mikrodizi çalışmalarına ait gen ifade verileri Bioconductor ExpressionSet nesnesi içerisinde yer almaktadır. ExpressionSet nesnesi, bir çip üzerindeki gen sayısı G, örnek sayısı N ile gösterildiğinde GxN boyutunda bir veriyi depolamaktadır. Aynı zamanda ExpressionSet nesnesi Şekil 3.1.`de gösterildiği gibi veri setine ilişkin veri setinin konusunu oluşturan deneyin

açıklaması, öznitelik sayısı, deneydeki örneklere ait bilgiler gibi çeşitli bilgilerin olduğu metadata bilgilerini içermektedir (68).

**channel_count (kanal sayısı):** "1"

**dataset_id (veri kümesi tanımlayıcısı):** "GDS3837"

**description (deneyin tanımı):** "Analysis of paired tumor and adjacent normal lung tissue specimens obtained from nonsmoking female non-small cell lung carcinoma (NSCLC) patients in Taiwan. Results provide insight into potential prognostic biomarkers and therapeutic targets for NSCLC."

"lung cancer" "control" "37-80 years" "Tumor stage: normal,1A,1B,2A,2B,3A,3B,4,2,1"

**email (GEO veri deposu e-posta adresi):** "geo@ncbi.nlm.nih.gov"

**feature_count (Öznitelik sayısı):** "54675"

**institute (kurum):** "NCBI NLM NIH"

**name (veri deposunun adı):** "Gene Expression Omnibus (GEO)"

**platform (platform):** "GPL570"

**platform_organism (platform organizma):** "Homo sapiens"

**platform_technology_type (platform teknolojisi):** "in situ oligonucleotide"

**pubmed_id (pubmed tanımlayıcısı):** "20802022"

**ref (referans):** "Nucleic Acids Res. 2005 Jan 1;33 Database Issue:D562-6"

**reference_series (referans serisi):** "GSE19804"

**sample_count (örnek sayısı):** "120"

**sample_id (örnek tanımlayıcıları):** "GSM494556,GSM494557,...

**sample_organism (örnek organizması):** "Homo sapiens"

**sample_type (örnek türü):** "RNA"

**title (başlık):** "Non-small cell lung carcinoma in female nonsmokers"

**type (tür):** "Expression profiling by array" "disease state" "age" "other" "individual"

**update_date (güncelleme tarihi):** "Sep 03 2011"

**value_type (değer türü):** "transformed count"

**web_link (web linki):** http://www.ncbi.nlm.nih.gov/geo

**Şekil 3.1.** GDS3837(Akciğer Kanseri) veri seti ile ilgili metadata bilgileri.

Farklı bilgi kaynaklarının tek bir yapıya dönüştürülerek daha kullanışlı hale gelmeleri için ExpressionSet nesnesi oluşturulmuştur (69). Bu tez çalışmasında da NCBI GEO veri tabanından alınan kanser türlerine ait mikrodizi gen ifade verileri Bioconductor aracılığıyla ExpressionSet nesnesine dönüştürülerek uygulama için hazır hale getirilmiştir. ExpressionSet nesnesi kullanılarak, mikrodizi gen ifade

verilerinde öznitelik seçim işlemini gerçekleştirmek için ***geneFilter*** ve ***CMA*** paketlerinden yararlanılmaktadır (70). Bioconductor bileşeni olan ***geneFilter*** paketi içerisinde yer alan `varFilter()` ve `nsFilter()` fonksiyonları aracılığıyla öznitelik seçimi yapılmaktadır. İstatistiksel yöntemler olan varFilter ve nsFilter yöntemlerine ilişkin açıklamalar aşağıda yer almaktadır.

**varFilter**

Yüksek boyutlu veri setleri üzerinde öznitelik seçimi yapmak için hız ve bellek gibi bilgisayar kaynaklı sorunları ortadan kaldırmak amacıyla oluşturulan R paketlerinden biri olan ***genefilter*** paketi içerisinde yer alan `varFilter()` fonksiyonu kullanılmaktadır. Veri seti öncelikle ExpressionSet nesnesine dönüştürülerek öznitelik seçimi için hazır hale getirilir. varFilter yönteminde; veri setindeki özniteliklerin her biri için varyans değerleri elde edilir. Büyükten küçüğe doğru sıralanan varyans değerleri içinde belli bir sınırdan önce gelenleri belirlenir. Belirlenen varyans değerlerine sahip öznitelikler daha sonraki aşamalarda kullanılmak için seçilir. Diğer bir deyişle varFilter ile seçim neticesinde örnekler arasında çok değişiklik gösteren öznitelikler seçilirken az değişiklik gösteren öznitelikler atılır. `varFilter()` fonksiyonu içerisinde yer alan `var.cutoff` değeri ise veri setindeki toplam özniteliklerin ne kadarı ile çalışılmak isteniyorsa onu ifade etmek için kullanılmaktadır. Örneğin özniteliklerin %20`sini seçmek için `var.cutoff=0.80`, %10`unu seçmek için ise `var.cutoff=0.90` olarak belirlenir. Tez çalışmasında kullanılan mikrodizi gen ifade verilerinden biri olan akciğer kanseri verisinde varFilter ile yapılan öznitelik seçiminin R programındaki işlem adımları aşağıda verilmiştir. `var.cutoff` değeri de 0.90 alınmıştır (68).

```
R >eset_akciğer<-GDS2eSet(gdsakciğer,do.log2=TRUE)
R >dim(eset_akciğer)
 Features  Samples
   54675      120
R >seçim_akciğer1<- varFilter(eset_akciğer,var.cutoff=0.90)
R >dim(seçim_akciğer1)
 Features  Samples
    5468      120
```

Yukarıda verilen örnekten de anlaşılacağı üzere ExpressionSet nesnesi biçiminde olan ve toplamda 54675 özniteliğin bulunduğu akciğer kanseri veri setinde varyans değerlerine göre seçim yapan varFilter yöntemi ile belirlenen özniteliklerin %10`u seçilmiştir. Bundan sonraki işlemlerde kullanılmak üzere anlamlı ve önemli 5468 öznitelik seçilmiştir (70). Çalışmada kullanılan diğer veri setleri üzerinde de varFilter`a ait aynı işlemler gerçekleştirilmiştir.

**nsFilter**

varFilter`da olduğu gibi öznitelik seçimi için ExpressionSet nesnesine dönüştürülen veri setine ait anotasyon paketindeki bilgiler yani açıklama bilgileri de kullanılarak seçim yapılacak ise nsFilter yönteminden yararlanılır. Eğer ExpressionSet nesnesinin açıklama bilgisi yok ise Bioconductor`ın ***hgu133plus2.db*** paketi kurulur ve nesneye atanır. Aşağıda verilen adımlarda yer alan `eset` nesnesinde `Annotation:GPL570` şeklinde açıklama paketi bulunur. Bioconductor`da GPL570, `hgu133plus2.db` ile ifade edilir. GPL ise GEO`da verinin düzenlenmesi için kullanılan öğelerden biridir. Platform üretici kurumu, platform tanımı, teknoloji çeşidi gibi platform bilgilerini kapsayan platform kayıt dosyasının GPL ile başlayan bir ismi vardır. GPL570 platform kaydı ise Affymetrix Human Genome U133 Plus 2.0 dizileri ile ilgilidir. ***hgu133plus2.db*** paketinin eset nesnesine atanması ile açıklama bilgilerinin bulunduğu nesne ile öznitelik seçim işlemine başlanır. `nsFilter()` fonksiyonunun kullanılması ile devamlı düşük sinyal gösteren öznitelikler ve varyans hesabı ile örnekler arasında çok değişiklik göstermeyen öznitelikler seçilmemektedir (68,71). varFilter yönteminde yer alan `var.cutoff` değeri gibi bir ölçüt kullanarak özniteliklerin ne kadarı ile çalışılacağı da önceden belirlenmemektedir. Tez çalışmasında kullanılan mikrodizi gen ifade verilerinden biri olan akciğer kanseri verisinde nsFilter ile yapılan öznitelik seçiminin R programındaki işlem adımları aşağıda verilmiştir.

```
R >eset_akciğer<-GDS2eSet(gdsakciğer,do.log2=TRUE)
R >eset<-eset_akciğer
R >annotation(eset)<-"hgu133plus2.db"
R >seçim_akciğer2<-nsFilter(eset)
R >dim(seçim_akciğer2)
```

```
Features  Samples
   10091      120
```

Yukarıda verilen örnekten de anlaşılacağı üzere ExpressionSet nesnesi biçiminde olan ve toplamda 54675 özniteliğin bulunduğu akciğer kanseri veri setinde anotasyon paketindeki bilgileri de kullanan nsFilter yöntemi ile bundan sonraki adımlarda kullanılmak üzere anlamlı ve önemli 10091 öznitelik seçilmiştir (70). Çalışmada kullanılan diğer veri setleri üzerinde de nsFilter`a ait aynı işlemler gerçekleştirilmiştir.

### 3.1.2. CMA Paketi ile Öznitelik Seçimi

Mikrodizi gen ifade verilerinde öznitelik seçimini gerçekleştirmek için mikrodizi deneyleri ile elde edilen gen ifade verilerine ait çeşitli bilgileri ve veri setini içeren ExpressionSet nesnesi kullanılarak R programında bulunan ***CMA*** paketinden de yararlanılmaktadır. Paket içerisinde öznitelik seçimi için Welch, t, F, Kruskal, Wilcox, rastgele orman, lasso ve boosting gibi yöntemler mevcuttur. Tez çalışmasında ise ***CMA*** paketi içerisinde yer alan rastgele orman, lasso, özyinelemeli öznitelik eleme ve limma yöntemleri kullanılmıştır. Bu yöntemler ***CMA*** paketinde bulunan `geneSelection()` fonksiyonu aracılığıyla uygulanmaktadır (68,72).

#### Rastgele Orman (Random Forest-rf)

Veri madenciliğinde yaygın olarak kullanılan ve başarı oranı yüksek bir sınıflama yöntemi olan rastgele orman (random forest-rf**)**, öznitelik seçimi için de kullanılmaktadır (63). Temeli karar ağaçlarına dayanan bu yöntem, Breiman tarafından 2001 yılında önerilmiş olup veri setinin farklı alt kümelerinde uygulanan birçok karar ağacını içermektedir (73). Karar ağaçları yukarıdan aşağıya doğru kök, dal, yaprak şeklinde bir ağaç yapısına sahiptirler. Bu yapı oluşturulurken kullanılan algoritma önemlidir. Genellikle ortak bilgi, bilgi kazancı, gini katsayısı, F test, t test ve $\chi^2$ gibi ölçütler karar ağacının yapısında kullanılmaktadır (65). rf yönteminde ağaçlar biraraya gelerek ormanı oluşturmaktadır. Orman oluşumunda meydana gelen sonuçlar bir arada değerlendirilerek nihai sonuca ulaşılır. Bu yöntemde karar ağacı oluşturmak amacıyla CART (Classification and Regression Trees) algoritması uygulandığı için gini katsayısı ölçüt olarak kullanılmaktadır. Gini katsayısının en az

olduğu öznitelik en iyi bölümlemenin yapılacağı özniteliktir. Eşitlik 3.1.`deki gibi Gini katsayısı hesaplanmaktadır.

$$Gini = 1 - \sum_{i=1}^{n}(p_i)^2 \quad (3.1.)$$

Burada n seçilen veriyi, $p_i$ ise veri satırındaki her bir verinin, o satırdaki tüm değerlerin bölümünden gelen kareleri toplamını ifade etmektedir. Öznitelik seçim işleminde önemli olan, etkili ve anlamlı öznitelikleri belirlemektir. Tez çalışmasında olduğu gibi daha sonra sınıflama yöntemleri gibi yöntemler uygulanmaktadır. İlk olarak hangi özniteliklerin seçileceğine karar vermek önemlidir. Çünkü öznitelikler ne kadar iyi seçilirse sınıflama performansı da o kadar yüksek olacaktır (63). Gömülü öznitelik seçim yöntemlerinden olan rf yöntemi belli bir işlem akışı ile çalışmaktadır. İlk olarak, tüm öznitelik durumları kontrol edilir ve kullanılacak karar ağacı ölçütüne göre her bir öznitelik için önem değerleri hesaplanır. Daha sonra hesaplanan önem değerine göre öznitelikler sıralanır ve en yüksek önem değeri olan öznitelik kök değeri olarak belirlenir. Son olarak belirlenen başarı oranı ve iterasyon sayısına kadar ağaç yapısı bir önceki adıma dönerek genişletilir ve kriter sağlandığında algoritma sonlandırılır (65). Öznitelik seçimi için kullanılan gömülü yöntemler içerisinde sınıflama algoritması da olduğu için biraz yavaş çalışmaktadır ve hesaplama maliyeti daha fazladır.

Tez çalışmasında kullanılan mikrodizi gen ifade verilerinden biri olan akciğer kanseri verisinde rf yöntemi ile yapılan öznitelik seçiminin R programındaki işlem adımları verilmiştir (68).

```
R >takciğer<-t(exprs(eset_akciğer))
R >durumakciğer<-pData(eset_akciğer)$disease.state
R >öğrenme_akciğer<-GenerateLearningsets(y=durumakciğer,
method="CV", fold=5, strat=TRUE)
R >seçim_akciğer3<-GeneSelection(takciğer, durumakciğer,
learningsets=öğrenme_akciğer, method="rf")
```

Çalışmada kullanılan diğer veri setleri üzerinde de rf yöntemine ait aynı işlemler gerçekleştirilmiştir. rf yönteminin uygulaması için R`ın ***randomForest*** paketi de kullanılmıştır.

**Lasso (Least Absolute Shrinkage and Selection Operator)**

Veri setine ait performansı iyi olan bir model oluşturmak için öznitelikler içerisinden modele en çok etki edecek öznitelikleri seçerken, bir arama algoritmasından yararlanan gömülü yöntemlerden biri de lasso (least absolute shrinkage and selection operator)`dur (66). Lasso, regresyon analizinde katsayı tahmini ve değişken seçimini aynı anda yapabilen yöntem olarak ilk kez 1996 yılında Tibshirani tarafından geliştirilmiştir (74,75). Regresyon analizi ile bağımsız değişken(ler)in değerinden yararlanarak bağımlı (yanıt) değişken değeri tahmin edilir. Bağımsız değişken(ler) ile yanıt değişkeni arasında doğrusal ilişki olduğu durumda kullanılan doğrusal regresyon modelinin oluşturulması için en küçük kareler yönteminden yararlanılır. Yöntemde, bağımsız değişkenlere ilişkin katsayıların yani parametrelerin tahmini yapılır. Ancak bağımsız değişken sayısı çok fazla olduğu zaman, değişkenler arasında doğrusal ya da doğrusala yakın ilişkinin gözlendiği çoklu bağlantı gibi birtakım sorunlar ortaya çıkar (76). Çoklu bağlantı olduğunda katsayı tahminleri belirsiz olur ve tahminlerin varyansları, standart hataları büyür, $R^2$ olması gerekenden büyük çıkar. Yanıt değişkeni değerini tahmin etmek için oluşturulacak regresyon modelinde birtakım değişiklikler yapılarak farklı yöntemlerden yararlanılır. Lasso da bu yöntemlerden biridir (66).

Gen ifade verilerinde bağımsız değişkenlerin karşılığı özniteliklerdir ve çok sayıda öznitelik bulunmaktadır. Lasso yöntemi ile veri setinde yer alan öznitelikler kullanılarak oluşturulan modelde, hem aşırı uyum ve çoklu bağlantı gibi sorunlar ortadan kalkmış olur hem de daha az önemli özniteliklerin katsayıları sıfır olarak hesaplanır. Böylece lasso ile otomatik olarak öznitelik seçimi yapılmış olur (77).

Lasso yöntemi ile özniteliklerin katsayı hesabı için Eşitlik 3.2. ve Eşitlik 3.3.`ten yararlanılır. L1 ceza fonksiyonu ile cezalı en küçük karelerin özel bir durumu olan lasso tahmini Eşitlik 3.2.`deki gibi hesaplanmaktadır.

$$\hat{\beta}^{lasso} = arg_{\beta} min \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\, \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \qquad (3.2.)$$

$\sum_{j=1}^{p} |\beta_j| \leq t$ kısıtı altında lasso tahmini ;

$$\hat{\beta}^{lasso} = arg_{\beta} min \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\, \beta_j\right)^2 \qquad (3.3.)$$

Eşitlik 3.3. ile elde edilir. $t \geq 0$ parametresi ayar parametresi olup, tahminlere uygulanan büzülmenin miktarını kontrol eder (77,78).

Yanıt değişkeni Y normal, binom, poisson gibi dağılımlara sahip olabilir ve iki kategorili olduğu zaman genelleştirilmiş doğrusal modellerden yararlanılır. Genelleştirilmiş doğrusal modellerde rastgele bileşen, sistematik bileşen ve bağ fonksiyonu olmak üzere üç ana bileşen vardır. Model lojistik regresyon ve dağılım binom olduğu zaman bağ fonksiyonu $log[\mu_i/(1-\mu_i]$ olur ve model ile dağılıma göre bağ fonksiyonu değişir. Lojistik regresyon modeli için kullanılan lasso tahmini;

$$\hat{\beta}(\lambda) = arg_\beta \min\left(n^{-1}\sum_{i=1}^{n} \rho_{(\beta)}(X_i, Y_i\right) + \lambda\|\beta\|_1 \qquad (3.4.)$$

Eşitlik 3.4. ile hesaplanır. Eşitlikte yer alan $\rho$ fonksiyonu ise;

$$\rho_{(\beta)}(x,y) = -y(\sum_{j=0}^{p}\beta_j x^{(j)}) + \log(1+\exp(\sum_{j=0}^{p}\beta_j x^{(j)})) \qquad (3.5.)$$

Eşitlik 3.5`deki gibi elde edilir (75). Tez çalışmasında kullanılan mikrodizi gen ifade verilerinden biri olan akciğer kanseri verisinde lasso yöntemi ile yapılan öznitelik seçiminin R programındaki işlem adımları ise aşağıda verilmiştir (68).

```
R >takciğer<-t(exprs(eset_akciğer))
R >durumakciğer<-pData(eset_akciğer)$disease.state
R >öğrenme_akciğer<-GenerateLearningsets(y=durumakciğer,
method="CV", fold=5, strat=TRUE)
R >seçim_akciğer4<-GeneSelection(takciğer, durumakciğer,
learningsets=öğrenme_akciğer, method="lasso")
```

Çalışmada kullanılan diğer veri setleri üzerinde de lasso yöntemine ait aynı işlemler gerçekleştirilmiştir. Lasso yönteminin uygulaması için R`ın ***glmnet*** paketi de kullanılmıştır.

**Özyinelemeli Öznitelik Eleme (Recursive Feature Elimination-rfe)**

2000`li yıllarda kanser verileri üzerinde yapılan çalışmalarda, sınıflama yöntemi olan destek vektör makineleri ile gömülü öznitelik seçim yöntemlerinden olan özyinelemeli öznitelik eleme (recursive feature elimination-rfe) bir araya getirilerek oluşturulan yöntem ile seçilen öznitelikler aracılığıyla sınıflama performansı yüksek bir model elde etmek amaçlanmıştır. Literatürde rf gibi farklı

sınıflama yöntemlerinin öznitelik seçim yöntemleri ile bir arada kullanılmasıyla elde edilen öznitelik seçim yöntemleri de olmuştur. rfe için en sık tercih edilen yöntem ise genelleme yeteneği iyi ve doğruluk değeri yüksek olan destek vektör makineleridir. rfe, sınıf ayrımının en büyük sınırını yani marjini sağlayan öznitelikleri seçerek bir destek vektör makineleri sınıflayıcısının kullanımı gibi çalışmaktadır (65,67).

rfe`da, ilk olarak özniteliklerin hepsi kullanılarak destek vektör makineleri ya da rastgele orman gibi sınıflama yöntemleri aracılığıyla bir model elde edilir ve özniteliklerin her birine ait önem puanları hesaplandıktan sonra en düşük önem puanı olan öznitelik çıkarılarak yeniden model oluşturulur ve tekrar önem puanları hesaplanır. İstenilen sayıda öznitelik kalana kadar işleme devam edilir. Sınıflama yönteminin uygulanması ile hesaplanan ağırlık vektörlerine göre özniteliklerin önem puanları belirlenir ve sıralama yapılır. En yüksek önem puanına sahip özniteliğin sınıflama işlemi üzerindeki etkisi en fazla olacağı için öznitelik kümesinde yer almaya devam eder. Ancak sınıflamadaki en az etkiye sahip olan öznitelik, en düşük önem puanına sahip olduğu için veri setinden çıkarılarak bir sonraki sınıflama modelinde yer almaz. Mikrodizi gen ifade verileri gibi büyük veri setlerinde öznitelikleri teker teker çıkarmak fazla zaman alacağı için her tekrarda önem puanı düşük olan birden fazla öznitelik çıkarılabilir. İstenilen performans düzeyine ulaşılana kadar işlemlere devam edilir (79).

Tez çalışmasında kullanılan mikrodizi gen ifade verilerinden biri olan akciğer kanseri verisinde rfe yöntemi ile yapılan öznitelik seçiminin R programındaki işlem adımları aşağıda verilmiştir (68).

```
R >takciğer<-t(exprs(eset_akciğer))
R >durumakciğer<-pData(eset_akciğer)$disease.state
R >öğrenme_akciğer<-GenerateLearningsets(y=durumakciğer,
method="CV", fold=5, strat=TRUE)
R >seçim_akciğer5<-GeneSelection(takciğer, durumakciğer,
learningsets=öğrenme_akciğer, method="rfe")
```

Çalışmada kullanılan diğer veri setleri üzerinde de rfe yöntemine ait aynı işlemler gerçekleştirilmiştir. Destek vektör makineleri kullanıldığı için rfe yönteminin uygulamasında R`ın ***e1071*** paketi de kullanılmıştır.

**Limma (Linear Models for Microarray Data)**

R programında Bioconductor`a ait ***CMA*** paketinde yer alan `geneSelection()` fonksiyonundaki metodlardan biri olan limma (linear models for microarray data), ilk kez 2003 yılında Smyth tarafından ortaya atılmıştır (80). RNA sekansı ya da mikrodizi teknolojileri sayesinde elde edilen gen ifade verilerinin analizinde farklı olan gen ifadelerini belirlemek ve deney tasarımlarını analiz etmek için limma kullanılır. Bağımlı yani yanıt değişkeninin çok sınıflı olduğu durumlarda F istatistiğinden, iki sınıflı olduğu durumlarda ise t istatistiğinden yararlanılır (72,81). Ancak test istatistiklerinin formülünde serbestlik derecelerini de işin içine katarak, modifiye edilmiş t istatistiği gibi sonuçlar üzerinden yorumlar yapılır (80).

İki grup arasında incelenen öznitelik bakımından fark olup olmadığına karar vermek için t-istatistiği kullanılır. Mikrodizi verileri ile çalışıldığı zaman hasta-sağlıklı gibi iki grup arasında bir genin ortalama ifadesinin farklı olup olmadığını belirlemek için de t-istatistiğinden yararlanılır. Ancak mikrodizi gen ifade verilerinde olduğu gibi küçük örneklem büyüklüğü olduğunda hata varyansını tahmin etmede güçlükleri hesaba katacak şekilde değiştirilmiş bir t istatistiği yaklaşımı kullanılmalıdır. t istatistiğinde yapılan değişiklik ile küçük değişimi olan genlerin seçilmesi önlenir. Bu çalışmada öznitelikler ile hasta-sağlıklı şeklinde iki sınıfın olduğu veri setlerinde limma yönteminin uygulanmasıyla, öznitelikler (genler) açısından hasta-sağlıklı şeklinde iki grup arasında fark olup olmadığına bakılır. İki grup arasıında farklı ifade edilmiş genler belirlenerek öznitelik seçimi gerçekleştirilmiş olur (82).

Tez çalışmasında kullanılan mikrodizi gen ifade verilerinden biri olan akciğer kanseri verisinde limma yöntemi ile yapılan öznitelik seçiminin R programındaki işlem adımları aşağıda verilmiştir (68).

```
R >takciğer<-t(exprs(eset_akciğer))
R >durumakciğer<-pData(eset_akciğer)$disease.state
R >öğrenme_akciğer<-GenerateLearningsets(y=durumakciğer,
method="CV", fold=5, strat=TRUE)
R >seçim_akciğer6<-GeneSelection(takciğer, durumakciğer,
learningsets=öğrenme_akciğer, method="limma")
```

Çalışmada kullanılan diğer veri setleri üzerinde de limma yöntemine ait aynı işlemler gerçekleştirilmiştir. Limma yönteminin uygulaması için aynı zamanda R`ın ***limma*** paketi de kullanılmıştır (72).

### 3.2. Sınıflama Yöntemleri

Bağımsız değişkenler ve kategorik yanıt değişkeninin yer aldığı veriler kullanılarak geleceğe yönelik tahmin yapmak amacıyla anlamlı bir model elde etmek için sınıflama yöntemlerinden yararlanılır (12). Tez çalışmasında, öznitelikler bağımsız değişkenleri ifade ederken, yanıt değişkeni hasta-sağlıklı şeklinde iki sınıftan oluşmaktadır. Örnekler, öznitelikler ve kategorik yanıt değişkeninin yer aldığı mikrodizi gen ifade verilerinde örneklerin hangi sınıfta olduğu bilinmektedir. Bu bilgiler de kullanılarak oluşturulan sınıflama modelleri ile yeni bir örnek geldiğinde bu örneğin hangi sınıfta olacağı tahmin edilir. Kısacası yeni ortaya çıkan bir verinin hangi sınıfa dâhil olacağına karar verilir (18,83).

Sınıflama işlemi için takip edilen bir süreç vardır. Veri önişleme adımından sonra eldeki veri setinin hepsi model oluşturmada kullanılmaz. Bir kısmı eğitim veri kümesi, diğer kısmı test veri kümesi olmak üzere iki bölüme ayrılır. Ayırma işleminde genel olarak %80-%20 ya da %75-%25 gibi oranlar dikkate alınır (30). Eğitim veri kümesinde sınıfı bilinen örnekler ile sınıflama kurallarının yer aldığı bir model oluşturulur. Test veri kümesinde ise elde edilen model test edilerek doğruluğu ölçülür (84). Eğitim veri kümesinde kurulan model ile elde edilen sonuçların çapraz geçerlilik, bootstrap gibi yöntemlerle genelleştirilmesi sağlanır. Test veri kümesi aracılığıyla da modelin performansı belirlenir. Bu çalışmada veri setinin %75`lik kısmı eğitim, %25`lik kısmı ise test veri kümesi olarak kullanılmıştır ve model genelleştirilmesi için 5-kat çapraz geçerlilik yapılmıştır. Çapraz geçerlilik yönteminde, veri kümesi k tane alt kümeye bölünür. k-1 tane küme eğitimde, bir tanesi testte kullanılır. k kere işlem tekrarlanır ve her seferinde elde edilen doğruluk değerlerinin ortalaması alınır ve modelin doğruluk performansı hesaplanmış olur. Bootstrap yönteminde ise veri kümesinden belli sayıda yerine koyarak seçim yapılması ile eğitim veri kümesi oluşturulur. Test veri kümesini ise eğitim veri kümesine girmeyen örnekler oluşturmaktadır. Eğitim ve test veri kümeleri için ayrı ayrı hesaplanan doğruluk değerlerinin toplamı ile modelin doğruluk performansı elde

edilir (17,68). Bu çalışmada ise veri kümesinin beş parçaya bölünmesi ile birinci parça test, diğerleri ise eğitim veri kümesini oluşturur. Eğitim veri kümesinde sınıflama yönteminin uygulanması ile elde edilen sınıflama modeli ile bir öngörüde bulunulur. Bulunulan öngörünün doğruluk değeri gibi performans değerleri elde edilir. İkinci parçanın test verisi olması ile bu işlem devam eder. Sırasıyla diğer parçalar için süreç tekrarlanır. Beş defa işlemlerin tekrarlanması ile elde edilen doğruluk değerleri gibi performans değerlerinin ortalaması alınarak sınıflama modelinin performansı belirlenir (68). Veri setlerinin hepsinde çok iyi sonuç veren tek bir yöntem olmadığı için birçok farklı sınıflama yöntemleri geliştirilmiştir (30).

Mikrodizi gen ifade verileri gibi büyük veri setleri ile çalışıldığı zaman, hem hesaplama süresini azaltmak hem de daha iyi performansa sahip modeller elde etmek için öznitelik seçim işleminden sonra veri madenciliğinin en sık kullanılan yöntemlerinden biri olan sınıflama yöntemleri uygulanmaktadır. Gen ifade verilerinin yer aldığı çalışmalara bakıldığında en çok tercih edilen sınıflama yöntemlerinin naive bayes, destek vektör makineleri ve k-en yakın komşu olduğu görülmüştür (65). Tez çalışmasında bu üç sınıflama yöntemine ilave olarak yapay sinir ağları ve derin öğrenme yöntemleri de kullanılmıştır. Kullanılan sınıflama yöntemlerine ilişkin açıklamalar ilerleyen bölümlerde yer almaktadır.

### 3.2.1. Naive Bayes

Büyük veri setlerinde uygulaması yapıldığında hız ve doğruluk açısından yüksek performans gösteren Naive Bayes (NB) sınıflama yöntemi istatistiksel bir yöntemdir (83,85,86). Temeli 1760`lı yıllarda ortaya çıkan bayes teoremine bağlı olan ve İngiliz matematikçi Thomas Bayes`ten ismini alan sınıflama yöntemi, her bir sınıftaki özniteliklerin birbirinden bağımsız olduğu ve eşit öneme sahip olduğu varsayımına dayalıdır. Olasılık temelli, uygulanabilirliği ile basit algoritma yapısı olan etkili bir yöntemdir (18,87,88). Genellikle bağımsızlık varsayımı sağlanamıyor olsa da, birçok alanda uygulaması olan NB sınıflama yönteminin gen ifade verilerinde de diğer sınıflama yöntemlerine göre daha iyi sonuçlar verdiği ile ilgili görüşler vardır (65,87). Önceden sınıflanmış eldeki verileri kullanarak yeni gelen verinin hangi sınıfa ait olduğunu tahmin etmek için önsel olasılıklar ve koşullu olasılıkların yer aldığı Bayes teoreminden yararlanılır (16,62).

Örneklemde yer alan verilerin genelde %75 ya da %80`lik kısmı eğitim veri kümesi olarak kullanılarak öğrenme yapılır. Geriye kalan %25 ya da %20`lik kısmı oluşturan test veri kümesi ile de bazı özellikleri verilen örneklerin hangi sınıfa ait olduğu belirlenir (87,89). NB sınıflama yöntemi ile test veri kümesine ait verinin sınıfını tahmin etmek için Eşitlik 3.6., Eşitlik 3.7. ve Eşitlik 3.8. gibi eşitliklerden yararlanılmaktadır (16,30).

$X = \{x_1, x_2, \dots, x_n\}$ : hangi sınıfa ait olduğu bilinmeyen veri kümesi ve

$C = \{C_1, C_2, \dots, C_m\}$ : veri kümesinde m tane sınıf olması durumunda Bayes teoremine göre $P(C_j|X)$ olasılığı Eşitlik 3.7. ile elde edilir.

$$P(X) = \sum_{i=1}^{k} P(x_i|C_j).P(C_j) \tag{3.6.}$$

ile

$$P(C_j|X) = \frac{P(x_i|C_j)P(C_j)}{P(x_i)} \tag{3.7.}$$

$P(x_i|C_j)$ olasılığı basitleştirilerek hesaplamalardaki işlem yükü azaltılır. Bunun için örneğe ait $x_i$ değerlerinin birbirinden bağımsız olduğu varsayılarak Eşitlik 3.8.`den yararlanılır.

$$P(X|C_j) = \prod_{k=1}^{n} P(X_k|\, C_j) \tag{3.8.}$$

Sınıfı belirlenecek olan örnek X`in sınıfını tahmin etmek amacıyla Eşitlik 3.7.`de yer alan payda değerleri aynı olduğu için pay değerleri karşılaştırılarak en büyük olanı seçilir ve sınıfı bilinmeyen örneğin bu sınıfta olduğuna karar verilir (16,30).

Çalışmada NB sınıflama modelini elde etmek için R`ın ***caret*** paketinin `train()` fonksiyonu kullanılarak `method="nb"` seçilmiştir.

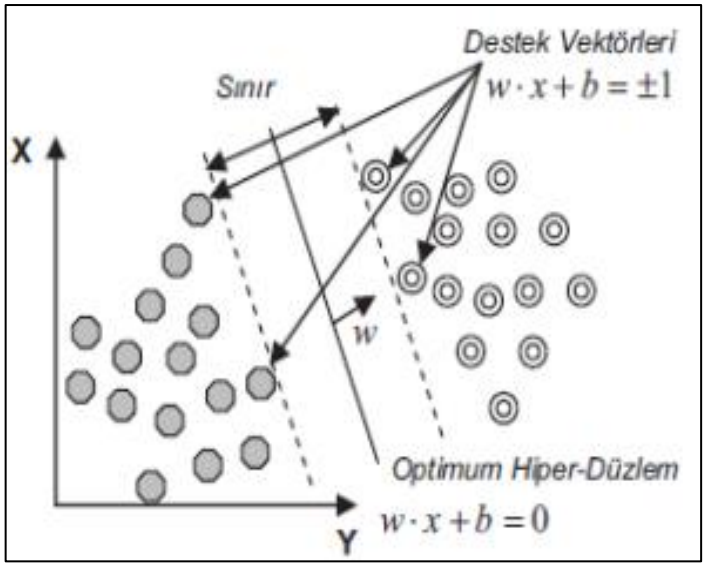
### 3.2.2. Destek Vektör Makineleri

1960`lı yıllarda Vladimir Vapnik ve Alexey Chervoenkis tarafından ilk kez ortaya çıkarılan Destek Vektör Makineleri (DVM) sınıflama yönteminin kullanımı 1990`lı yıllarda ilk başarılı uygulamaların yapılmasıyla yaygınlaşmıştır (90,91). Sınıf sayısının iki olduğu durumlar için geliştirilmiş olmasına rağmen zamanla sınıf

sayısının ikiden fazla olduğu ve doğrusal olarak ayrılamayan veri setleri için de genişletilerek uygulanabilir şekle gelmiştir (92).

Öznitelik sayısının çok olduğu gen ifade verileri gibi büyük veri setlerinde ve daha birçok alanda genellikle performans düzeyi yüksek olan sınıflama sonuçları verdiği için çok tercih edilen bir yöntemdir (91,92). Ayrıca aykırı değerler ve gürültülü veriler içeren veri setlerinde de sağlamdır (86).

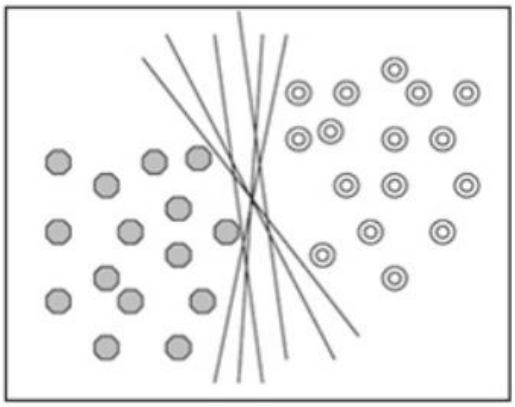
DVM sınıflama yönteminde, farklı sınıflara ait öznitelik kümeleri kullanılarak sınıfların örnekleri arasındaki ayrımı yaparken doğru sınıflamayı en iyi yapacak olan düzlem belirlenir ve hiperdüzlem olarak adlandırılır. Sınır genişliğini sınırlı bir duruma getiren noktalara da destek vektörleri denir. Şekil 3.2.`de gösterildiği gibi sınırlar birbirinden ne kadar uzak olursa o kadar uygundur (65,83,93).



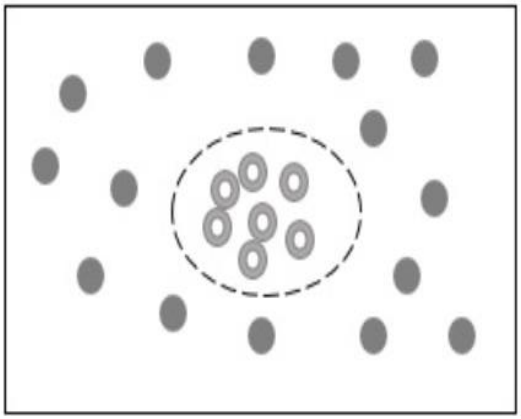

**Şekil 3.2.** Destek vektörleri.

Uygulamada hiperdüzlemin belirlenmesinde iki durum karşımıza çıkabilir. Bunlardan biri verilerin doğrusal biçimde ayrılabilecekleri bir yapıda olması diğeri ise verilerin doğrusal biçimde ayrılamayan bir yapıda olmasıdır (94). Dolayısıyla hiperdüzlemin belirlenmesinde doğrusal ya da doğrusal olmayan fonksiyonlardan yararlanılarak sınıflama yapılmaktadır (83).

Doğrusal olarak ayrılabilen ve doğrusal olarak ayrılamayan veriler Şekil 3.3. ile Şekil 3.4.`te gösterilmiştir (91).



**Şekil 3.3.** Doğrusal olarak ayrılabilen veriler.



**Şekil 3.4.** Doğrusal olarak ayrılamayan veriler.

Doğrusal olarak ayrılamayan veriler olduğunda polinomial fonksiyon, sigmoid fonksiyon, doğrusal fonksiyon ve radyal tabanlı fonksiyon gibi yaygın kullanılan çekirdek (kernel) fonksiyonları aracılığıyla başka bir boyutta verilerin doğrusal olarak ayrılmasına olanak sağlanır (92).

Diğer sınıflama yöntemlerinde olduğu gibi DVM`de de eğitim veri kümesi üzerinde çalışılır. Doğrusal olarak ayrılabilen verilerin olduğu ve y ile gösterilen sınıf etiketinin -1 ve +1 şeklinde belirtildiği iki sınıflı bir sınıflama probleminde DVM`nin eğitimi için n tane örnekten oluşan eğitim veri kümesinde $\{x_i, y_i\}$; $i = 1, \ldots, n$ kadar olduğu düşünüldüğünde en uygun hiperdüzleme ilişkin hesaplamalar Eşitlik 3.9. ve Eşitlik 3.10.`daki gibidir.

$$y_i = 1 \text{ için; } wx_i + b \geq 1 \tag{3.9.}$$

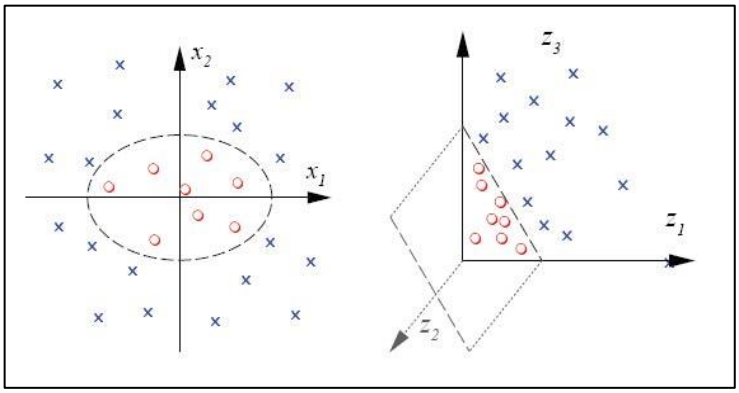
$$y_i = -1 \text{ için; } wx_i + b < 1 \tag{3.10.}$$

Eşitliklerde yer alan $w$, ağırlık vektörünü; $x$, N boyutlu uzayı ve $b$, hata terimi olarak da isimlendirilen yanlılığı ifade etmektedir (83, 94). Doğrusal olarak ayrılamayan verilerin varlığında ise hiperdüzleme ilişkin hesaplamalar genel olarak Eşitlik 3.11.`deki gibidir. Burada $K(x_i, x_j)$ çekirdek fonksiyonu iken $a$ Lagrange çarpanıdır (92).

$$f(x) = a_i y_i K(x_i, x_j) + b \tag{3.11.}$$

Her iki durumda da verilerin hiperdüzlemin doğru tarafında doğru sınıfta olmalarının yanı sıra verilerin hiperdüzlemden belli bir uzaklıkta bulunmasıyla en uygun hiperdüzlem belirlenmiş olur, böylece sınıflamanın başarısı olumlu yönde etkilenir (83).

Şekil 3.5.`te gösterildiği gibi doğrusal olarak ayrılamayan veri olduğunda çekirdek fonksiyonu aracılığıyla veri ayrılabilir duruma gelmektedir (94).




**Şekil 3.5.** Doğrusal olarak ayrılamayan ve çekirdek fonksiyonu ile farklı bir boyuta dönüştürülerek ayrılabilir şekle gelen veriler.

Bu çalışmada DVM sınıflama modelini elde etmek için R`ın ***caret*** paketinin `train()` fonksiyonu kullanılarak `method="svmLinear"` seçilmiştir.

### 3.2.3. k-En Yakın Komşu

k-En Yakın Komşu (kNN) sınıflama yöntemi, Fix ve Hodges tarafından ilk kez 1950`li yılların başında ortaya atılmış ve 1960`lı yılların sonlarına doğru Cover ve Hart tarafından geliştirilerek popüler hale gelmiştir (95-97). Örnek ya da bellek tabanlı yöntemlerin en temeli olan kNN sınıflama yönteminde, eğitim veri kümesi

aracılığıyla önceden sınıflama yapılması ile hangi sınıfa dâhil olduğu bilinmeyen örneğin sınıfı belirlenir (88,98,99).

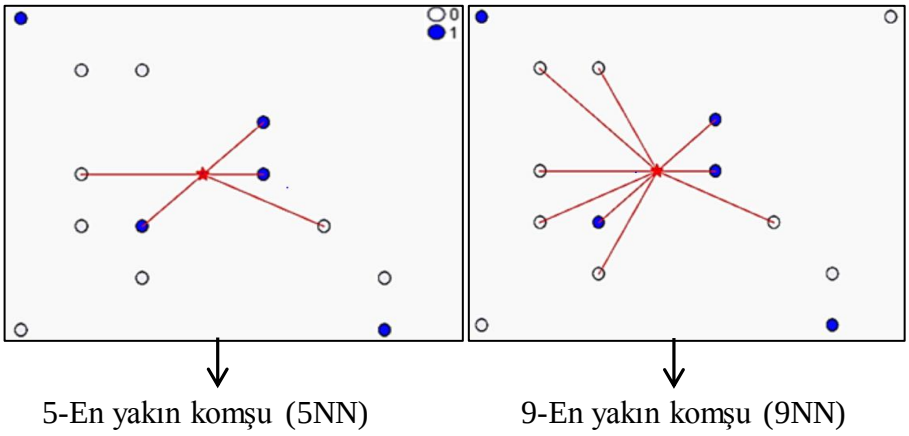
Sınıflaması yapılacak olan örneğin eğitim kümesinde en yakın uzaklıkta olduğu k tane komşusu belirlenir ve bu komşuların çoğu hangi sınıfta ise yeni örnek te o sınıfa dâhil edilir. Sınıflama işlemi için uzaklık ölçülerinden yararlanılarak yeni gelen örneğin eğitim veri kümesindeki örnekler arasındaki benzerliğine bakılır ve en yakın eğitim veri kümesi örneği belirlenmesiyle örnek için sınıf tahmini yapılmaktadır (88,97). Bunun için de Minkowski, Öklit, Manhattan gibi uzaklık ölçülerinden yararlanılır. Daha önce yapılan çalışmalarda ise genelde Öklit uzaklık ölçüsü kullanılmıştır (100). Öklit uzaklık ölçüsünü elde etmek için öznitelikler arası mesafe farkının kareleri toplamının karekökü alınır (65,101). Eşitlik 3.10.`da verildiği gibi eğitim veri kümesindeki özniteliklerin değerleri ile yeni gelen yani test edilecek örneğe ait veri kümesindeki özniteliklerin değerleri arasındaki mesafeler hesaplanarak Öklit uzaklık ölçüsü bulunur.

Test edilecek örneğe ait öznitelikler kümesi X, eğitim veri kümesine ait öznitelikler kümesi Y ve öznitelik sayısı i ile ifade edilmiştirğinde Öklit uzaklık ölçüsü;

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_i - y_i)^2} \qquad (3.12.)$$

Eşitlik 3.12. ile hesaplanmaktadır (65).

Sınıfların her biri için farklı değerler bulunur ve en büyük değere sahip sınıfa, yeni gelen yani test edilen örnek atanır (65,101). Dolayısıyla sınıfların özniteliklerinin belirlenmiş olması önemlidir (100). Veri setlerine göre değişen k katsayısı, genelde çalışmalarda üç alınmıştır (65). Veri setine yeni gelen bir örneğin k değerinin beş ve dokuz olduğu iki durum için, sıfır ve bir olmak üzere iki sınıftan hangi sınıfa dâhil olacağı Şekil 3.6.`da gösterilmektedir. En yakın komşuların çoğunluğu hangi sınfta ise yeni örnekte o sınıfta yer alır. k=5 olduğunda en yakın komşuların çoğunluğu bir sınıfında olduğu için yeni örnek o sınıfta yer alır, k=9 olduğunda ise en yakın komşuların çoğunluğu sıfır sınıfında olduğu için yeni örnekte sıfır sınıfına dâhil olur (97).



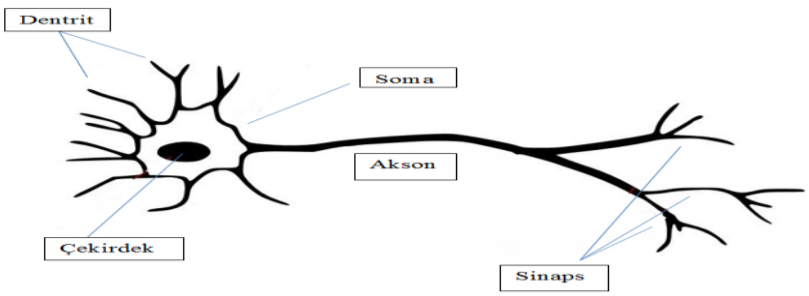

**Şekil 3.6.** k-En yakın komşu.

kNN yapısı itibariyle basit olduğu için pek çok alanda yararlanılan ve en yaygın kullanılan sınıflama yöntemleri arasındadır. Performansı birçok çalışmada ispatlanan, etkili ve anlaşılır sonuçlar veren kNN sınıflama yöntemi, özellikle büyük veri setlerinde kullanılmaktadır (65,86,97,102).

Çalışmada kNN sınıflama modelini elde etmek için R`ın ***CMA*** paketinin `knnCMA()` fonksiyonu kullanılmıştır.

### 3.2.4. Yapay Sinir Ağları

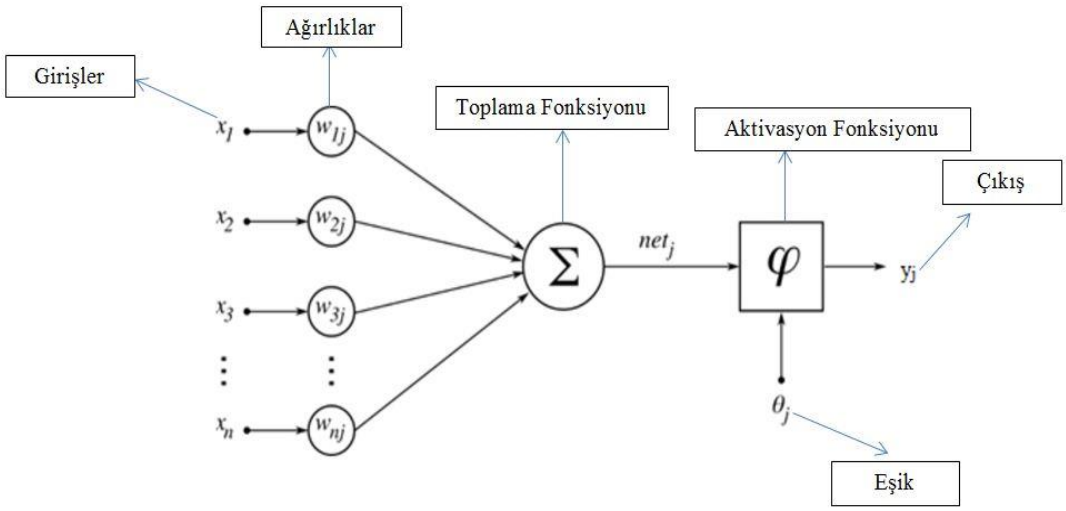
Yapay Sinir Ağları (YSA) kullanımı ilk kez McCulloch ve Pitts tarafından 1943 yılında ortaya atılmıştır ve 1980`lerden sonra yaygınlaşmaya başlamıştır (16,99,103). YSA, insan beynindeki sinir hücrelerinin işleyişini bilgisayar ortamında matematiksel olarak modelleyen bir yapıdır. Dolayısıyla ilk olarak biyolojik sinir hücresinin yapısını incelemek gerekir. Biyolojik sinir hücresinin yapısı Şekil 3.7.`de gösterilmiştir (94). Biyolojik sinir hücresinde bulunan yapılara karşılık sırasıyla YSA`da yer alan kavramlar bulunmaktadır.




**Şekil 3.7.** Biyolojik sinir hücresinin yapısı.

Biyolojik sinir hücresinde yer alan çekirdek (nöron), dentrit, soma (hücre gövdesi), akson ve sinaps yapılarına karşılık YSA`da sırasıyla algılayıcı, toplama işlevi, etkinleştirme işlevi, algılayıcı çıktısı ve ağırlıklar gelmektedir (83).

Biyolojik sinir hücresinin yapısından esinlenerek oluşturulan yapay sinir hücresinin yapısı ise Şekil 3.8.`de gösterildiği gibidir (94).




**Şekil 3.8.** Yapay sinir hücresinin yapısı.

YSA`nın en küçük ve temel öğesi nörondur, diğer bir deyişle sinir hücreleridir. Sinir hücreleri rastgele bir araya gelmezler. Genel olarak giriş, gizli ve çıkış olmak üzere üç tabakanın birbirine bağlanmasıyla oluşan sinir ağları; hafizaya alma, öğrenme gibi özelliklere sahiptirler. Dışarıdan veri alan sinir hücrelerini içeren tabaka giriş tabakası iken çıktıları dışarı ileten sinir hücrelerinin olduğu tabaka da çıkış tabakasıdır. Giriş ve çıkış tabakaları arasında bulunan gizli tabaka ise çok sayıda sinir hücresine sahiptir (18,83,94). Kısacası YSA`nın genel yapısı girdi verisinin bulunduğu giriş tabakasını, çıkış verilerinden oluşan çıkış tabakasını ve bu iki tabaka arasında yer alan gizli tabakayı içermektedir. Girdiler $x_i$, ağırlıklar $w_{ij}$, eşik değeri $\theta_j$ ve çıkış $y_j$ ile gösterildiğinde, çıkış değerinin hesaplanması için aşağıda verilen Eşitlik 3.13.`ten yararlanılır (94).

$$y_j = f(\sum_{i=1}^{n} x_i w_{ij} + \theta_j) \qquad (3.13.)$$

Eşitlik 3.13.`ten anlaşılacağı üzere girdi değerleri ağırlık katsayıları ile çarpılır ve eşik değeri ile toplanır. Aktivasyon diğer adıyla etkinleştirme fonksiyonu $f$ aracılığıyla $y_j$ çıktı değeri elde edilir. Eşik değer, biyolojik sinir hücresinin iletime geçebilmesi yani etkin olabilmesi için gerekli olan değerdir. İstenen sonuca ulaşmak için eşik ve ağırlık değerleri ayarlanarak sinir ağı eğitilmelidir. Doğrusal fonksiyon ($f(x) = x$), Lojistik Sigmoid ($f(x) = 1/(1 + \exp(-x))$, Hiperbolik Tanjant ($f(x) = \tanh(x)$) gibi fonksiyonlar, aktivasyon fonksiyonları içerisinde sık kullanılanlarıdır (83).

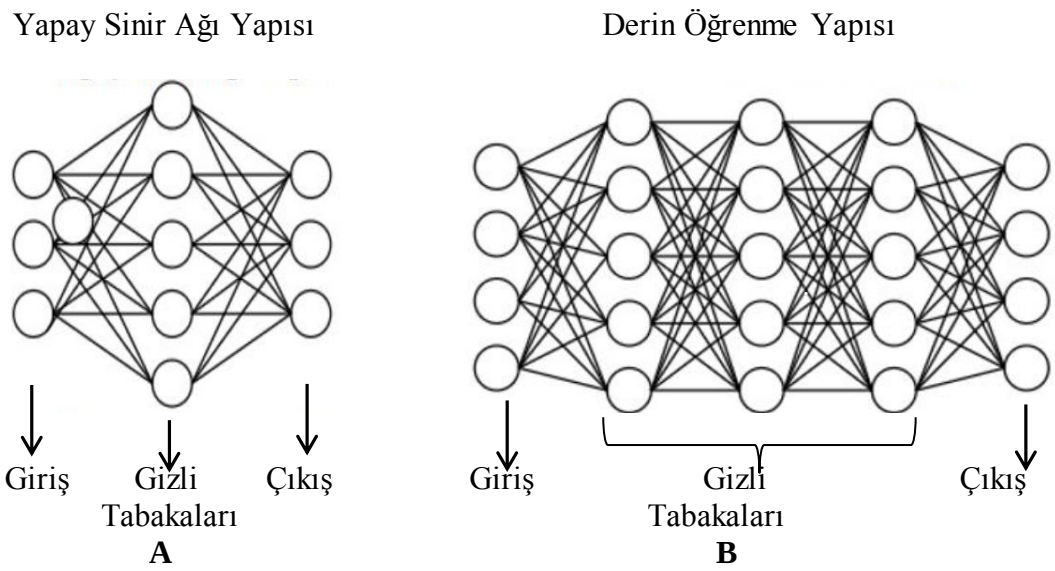
İnsan, hayatının her aşamasında yaşadıklarından yeni şeyler öğrenir ve bu öğrendiklerini daha sonra karşılaştığı herhangi bir durumda kullanarak karar verebilir. YSA`da insanın öğrenme yeteneğinden ilham almaktadır. Veri madenciliğinde kullanılan sınıflama yöntemlerinin neredeyse hepsinde olduğu gibi YSA`da da önce eğitim veri kümesinde öğrenme yani eğitim gerçekleştirilir. Devamında ise test veri kümesinde kullanma, diğer bir deyişle test aşaması gerçekleştirilir (18,83). YSA`dan yararlanarak sağlık, iletişim, üretim, askeri ve savunma sanayi gibi birçok farklı alanda tahmin, sınıflama, teşhis gibi uygulamalar yapılmaktadır (16). YSA`nın farklı düzenlere uygulanmasının zor olabilmesi, yapı içerisinde ne olduğunun bilinememesi, model oluşturma aşaması ve yorumlanmasının karar ağaçları gibi yöntemlere göre nispeten kolay olmaması, modelin doğru kurulması için ağın eğitimindeki dengenin sağlanabilmesi gibi dezavantajları bulunmaktadır. Ağ fazla eğitildiğinde önceden gözlenmemiş bir örneğe ilişkin tahmin yapamazken, az eğitildiğinde de yanlış tahmin yapabilir (16,18).

Çalışmada YSA sınıflama modelini elde etmek için R`ın ***nnet*** paketinin `nnet()` fonksiyonu kullanılmıştır.

### 3.2.5. Derin Öğrenme

2000`li yıllarda Geoffrey Hinton ve Ruslan Salakhutdinas`ın yapmış oldukları makale ile çok tabakalı YSA`nın nasıl çalıştığı gösterilmiştir. 2006 yılında Deep Belief Network çalışmasında da çok tabakalı derin yapıların çalışma şekli ve eksik özelliklerin kendi kendini tamamlama şekli ifade edilmiş ve buna Derin Öğrenme (DÖ) adı verilmiştir (104). Daha sonra pek çok yöntemleri ortaya atılan

DÖ ilk kez 2012 yılında bilim insanları tarafından kullanılarak zamanla yaygın bir kullanım alanına sahip olmuştur (105). 3.2.4 bölümünde anlatılan giriş, gizli ve çıkış olmak üzere üç tabakadan oluşan YSA`nın ileri düzeyli genişletilmiş bir yaklaşımı olan DÖ yönteminde gizli tabaka birden fazladır, böylece veriler kapsamlı bir şekilde temsil edilebilir (104). YSA ve birçok gizli tabakadan oluşan DÖ yapısı Şekil 3.9.`da verilmiştir (92).




**Şekil 3.9.** Yapay sinir ağı (A) ve derin öğrenme (B) yapısı.

Şekil 3.9.`dan da anlaşılacağı üzere, gizli tabakalar sayesinde çıkış tabakasındaki sonuç elde edilir. DÖ yöntemleri genel olarak ardışık tabakaların derin yapılarını kapsamaktadır. Ardışık tabaka ile anlatılmak istenen, her tabakanın çıkışı bir sonraki tabakanın girişini oluşturur. Böylece hiyerarşik bir şekilde öğrenme sağlanır. Giriş tabakasındaki nöronlar kullanılarak $y = f(x, w)$ doğrusal fonksiyonu ile gizli tabakalardaki nöronlar hesaplanır. Hesaplanan nöronlar üzerinde aktivasyon fonksiyonları uygulanarak çıkış tabakasındaki nöronlar elde edilir. İşlemler sonucunda elde edilen çıkış tabakası giriş tabakasının doğrusal olmayan bir biçimidir. DÖ yöntemleriyle genellikle doğrusal olmayan problemler çözülmeye çalışılır çünkü bu tip problemlerde diğer yöntemlere göre daha iyi sonuçlar vermiştir (92).

Veri madenciliğinde kullanılan sınıflama yöntemlerinde uygulanan ön işleme, öznitelik seçimi gibi adımların DÖ`de yapılmasına gerek olmadığı ile ilgili bilgi vardır. Ancak DÖ modellerinin kara kutu (black box) olduğu da bilinmektedir. DÖ

çok sayıda gizli tabakadan oluşması sebebiyle kendi içinde ne tür işlemlerin gerçekleştiği ile ilgili bilinmeyenler mevcuttur. Yani modellerin iç işleyişi tam olarak anlaşılıp açıklanamamaktadır. DÖ`nün bu özelliği dezavantajlarından biridir. Çünkü birçok uygulama alanında zorluk çıkarmaktadır. RNA dizileme ve mikrodizi gen ifade verilerinde, klasik veri madenciliği sınıflama yöntemlerinin kullanıldığı çalışmalara göre DÖ yönteminin sınıflama amaçlı kullanıldığı çalışmalar daha azdır. Mikrodizi gen ifade verileri gibi büyük veri setlerinde de verilerin tamamı ile analiz yapıldığında hız ve bellek açısından maliyet gibi sıkıntılar meydana gelmektedir (89,92,105).

Diğer sınıflama yöntemlerinde olduğu gibi DÖ`de de veri seti eğitim ve test veri kümeleri olmak üzere ikiye bölünür. Daha sonra sınıflama modelini oluşturmak için H2O.ai ekibi (2017) tarafından geliştirilen ***h2o*** paketi kullanılır. `h2o.init()` fonksiyonu ile bu paket çalıştırılarak veri kümeleri h2o nesnelerine dönüştürülür. `h2o.deeplearning()` fonksiyonu aracılığıyla da h2o veri nesnesinden DÖ modeli elde edilir (68). Bu tez çalışmasında, öznitelik seçim yöntemlerinin sınıflama yöntemleri performansına etkisini göstermek ve karşılaştırma yapabilmek için sonuçların belli bir standartta olması amacıyla diğer sınıflama yöntemlerinde olduğu gibi öznitelik seçim yöntemlerinin uygulanması ile DÖ sınıflama modelleri elde edilmiştir. Ayrıca öznitelik seçim yöntemi uygulamadan da DÖ modelleri oluşturulmuştur. Elde edilen sonuçlara ait ayrıntılı açıklamalar Bulgular bölümünde yer almaktadır.

Günümüzde DÖ; görüntü tanıma, ses analizi, doğal dil öğrenme, video işleme gibi birçok alanda kullanılmaktadır (89). Özellikle sağlık alanında, kanser teşhisi ve sınıflamada, ilaç geliştirmede, medikal görüntü ve sinyal verilerinde DÖ yöntemleri sıklıkla uygulanmaktadır. Soruların çözümünde doğruluk açısından göstermiş olduğu yüksek performans neticesinde giderek daha çok tercih edilmektedir (106).

Derin sinir ağlarından farklı olarak değişik veri türleri ve amaçlar için oluşturulmuş Yığınlı Otomatik Kodlayıcılar, Derin İnanç Ağları, Evrişimsel Sinir Ağları, Tekrarlayan Sinir Ağları gibi DÖ yapıları da vardır (92).

### 3.3. Model Performans Ölçüleri

Sınıflama yöntemlerinin ne kadar doğru sınıflama yaptığı ile ilgili performansını değerlendirmek ve yorumlamak için çeşitli model performans ölçülerinden yararlanılmaktadır. Oluşturulan sınıflama modeli iki sınıflı olduğunda duyarlılık, seçicilik, F-ölçütü, dengeli doğruluk, Kappa, Matthews korelasyon katsayısı, doğruluk ve ROC eğrisi altında kalan alan gibi performans ölçüleri aracılığıyla sınıflama modelinin başarısı belirlenerek, veriyi ne kadar doğru sınıflandırdığı gösterilir. Bu tez kapsamında ele alınan veri setlerinde sınıf sayılarının dağılımında dengesizlik durumu da olmadığı için model performans ölçülerinden doğruluk, duyarlılık, seçicilik ve ROC eğrisi altında kalan alan gibi çok tercih edilen ölçüler kullanılmıştır. Tablo 3.1. göz önünde bulundurularak, kullanılan ölçüler hakkında açıklamalar verilmiştir. Sınıf sayısının iki olduğu bir veri setinde dört olası sonuç vardır. Bu sonuçlar Tablo 3.1.`de verilen sınıflama tablosundaki gibidir (92,107).

**Tablo 3.1.** Gerçek ve tahmin sonuçlarına ait sınıflama tablosu.

| | **Gerçek Sınıf** | |
|---|---|---|
| **Tahmin Edilen Sınıf** | Pozitif | Negatif |
| Pozitif | A<br>(Doğru Pozitif - DP) | B<br>(Yanlış Pozitif - YP) |
| Negatif | C<br>(Yanlış Negatif - YN) | D<br>(Doğru Negatif - DN) |

Çalışmada kullanılan performans ölçüleri de sınıflama tablosunda yer alan değerler ile hesaplanabilmektedir.

#### Doğruluk (Accuracy)

Sınıflama yönteminin performansının belirlenmesinde çok tercih edilen basit bir yöntemdir ve genel başarı ölçüsüdür. Doğru sınıflandırılmış örnek sayısının (DP+DN) toplam örnek sayısına (DP+YP+YN+DN) oranıdır. Sınıf sayılarının dağılımında dengesizlik durumu olduğu zaman yanlış yorumlara sebep olacağı için kullanılmaması gerekir.

Modelin doğruluk değeri Eşitlik 3.14.`deki gibi hesaplanır ve hata oranı ise 1-Doğruluk değerine karşılık gelmektedir (107,108).

$$Doğruluk = \frac{DP+DN}{DP+YP+YN+DN} \quad (3.14.)$$

**Duyarlılık (Sensitivity)**

Gerçekte pozitif sınıfta yer alan örnekler içerisinde tahmin edilen sınıfı pozitif olan örneklerin oranıdır. Sınıflama yönteminin pozitif değere sahip olan örnekleri belirlemedeki performansıdır ve Eşitlik 3.15. ile elde edilir (108).

$$Duyarlılık = \frac{DP}{DP+YN} \quad (3.15.)$$

**Seçicilik (Specificity)**

Gerçekte negatif sınıfta yer alan örnekler içerisinde tahmin edilen sınıfı negatif olan örneklerin oranıdır. Sınıflama yönteminin negatif değerlere sahip örnekleri belirlemedeki performansıdır ve Eşitlik 3.16. ile hesaplanır (108).

$$Seçicilik = \frac{DN}{YP+DN} \quad (3.16)$$

**ROC Eğrisi Altında Kalan Alan (Area Under the ROC Curve)**

İlk kez 1950`li yıllarda sinyal algılamada kullanılan ROC eğrisi biyomedikal çalışmalarda çok kullanılmaktadır. ROC eğrisinin oluşturulmasındaki asıl amaç, sınıflama yönteminden elde edilen sonucu doğruluk değerleri bakımından incelemektir. Dolayısıyla ilk olarak duyarlılık ve seçicilik değerleri hesaplanır. Yatay ekseni 1-seçicilik, dikey ekseni ise duyarlılık değerlerinden oluşan ROC grafiği üzerinde elde edilen eğri altında kalan alan (EAKA) 0,5 ile 1 arasında değişen değerler almaktadır. EAKA 1`e ne kadar yakınsa kullanılan yöntemin sınıflama performansı o kadar iyi iken; EAKA değeri 0,5 olan yöntem sınıflamada oldukça başarısızdır (92,108).

Tez çalışmasında kullanılan verilerin yanıt değişkeni, hasta-sağlıklı şeklinde iki sınıftan oluşmaktadır. Hastalığı gerçekten var olduğu bilinen kişilerden yüzde kaçının sınıflama yöntemi ile hasta olarak sınıflandığını gösteren ölçü duyarlılık değeri iken; sağlıklı kişilerin yüzde kaçının sınıflama yöntemi ile sağlıklı olarak sınıflandığını gösteren ölçü seçiciliktir. Hasta ve sağlıklı kişilerin yüzde kaçının sınıflama yöntemi ile doğru bir şekilde hasta veya sağlıklı olarak sınıflandığını

gösteren değer doğruluk değeridir. EAKA da kullanılan sınıflama yönteminin hasta ve sağlıklıları ayırt etme gücünü göstermektedir (109).

### 3.4. Çalışmada Kullanılan Veri Setleri

Mikrodizi gen ifade verilerinde öznitelik seçim yöntemlerinin sınıflama yöntemleri başarısına etkisini göstermek için kullanılan veri setleri ile ilgili ayrıntılı açıklamalar bu bölümde yer almaktadır.

#### 3.4.1. Gerçek Veri Setleri

Çalışmada kullanılan gerçek veri setleri akciğer, lenfoma, rahim ağzı, meme, prostat ve lösemi olmak üzere altı farklı kanser türüne ait, mikrodizi deneylerinden elde edilen gen ifade verilerinden oluşmaktadır. Veri setlerine NCBI tarafından hizmete sunulan GEO veri deposu üzerinden ulaşılabilmektedir. GEO veri deposuna ise www.ncbi.nlm.nih.gov/geo/ adresinden erişilebilir. Biyoinformatikte verilerin analizinde sık kullanılan R, WEKA, Orange gibi programlar ile GEO veri deposu koordineli çalışabilmektedir. GPL platform kaydı, GSM örnek kaydı, GSE veri seri kaydı ve GDS veri seti olmak üzere GEO`da verinin düzenlenmesi dört bileşenden oluşmaktadır ve yaklaşık 600000 örneğe ait binlerce veri seti bulunmaktadır (48,68). GDS veri seti, platform ve örnek veri kayıtlarına ait bilgileri içeren bir kümedir. Normalleştirme, arka plan işlemleri gibi mikrodizilerde ön işleme adımları veri setinde yapılmış olup, ölçümlerin eşdeğer biçimde yapıldığı kabul edilmektedir. Her bir veri setinin GDS ile başlayan kodu vardır (68).

Bu tez çalışmasında kullanılan mikrodizi gen ifade verileri ile ilgili bilgiler Tablo 3.2.`de yer almaktadır. Tabloda GEO veri deposunda yer alan veri setlerine ait GDS ile başlayan GEO kodu,veri kaynağı, başlığı ve ilgili web adresi gibi ayrıntılı bilgiler bulunmaktadır. Tablo 3.3.`te ise gerçek veri setlerine ilişkin başlıca özellikler verilmiştir. Akciğer ve rahim ağzı kanseri veri setlerinde 54675 öznitelik var iken; meme, prostat kanseri, lenfoma ve lösemi veri setlerinde ise 22283 öznitelik vardır. Ele alınan veri setlerinde hasta ve sağlıklı şeklinde iki sınıf değeri içeren yanıt değişkeni bulunmaktadır. Toplam örnek sayısına bakıldığında akciğer kanseri yüz yirmi örnek ile en çok örnek sayısına sahip veri setidir. On dört örnek sayısı ile

lenfoma veri seti en az örnek sayısına sahiptir. Her bir veri setinde yer alan hasta-sağlıklı sınıflarına ait örnek sayıları birbirine eşit veya çok yakındır.

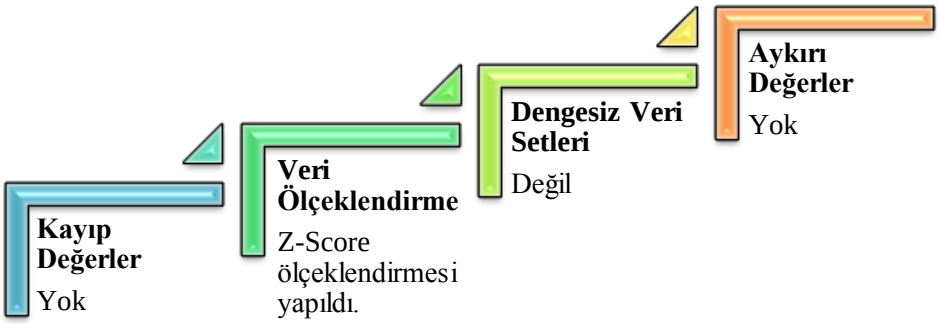
**Tablo 3.2.** Çalışmada kullanılan mikrodizi gen ifade verileri ile ilgili bilgiler.

| Veri Seti | Kaynak | GEO-başlık | GEO-kod | Web adresi |
|---|---|---|---|---|
| **Akciğer kanseri** | Lu ve ark, 2010 | Non-small cell lung carcinoma in female nonsmokers | GDS3837 | https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3837 |
| **Lenfoma** | Dürig ve ark, 2007 | T-cell prolymphocytc leukemia with inv(14) (q11q32) | GDS2908 | https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS2908 |
| **Rahim ağzı kanseri** | Noordhuis ve ark, 2011 | Lymph node-positive, early stage cervical cancer | GDS4664 | https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4664 |
| **Meme kanseri** | Graham ve ark, 2010 | Breast cancer: histologically normal breast epithelium | GDS3716 | https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3716 |
| **Prostat kanseri** | Sun ve ark, 2009 | Recurrent and non-recurrent prostate cancer primarytumors | GDS4109 | https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4109 |
| **Lösemi** | Stirewalt ve ark, 2008 | Acute myeloid leukemia | GDS3057 | https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3057 |

**Tablo 3.3.** Gerçek veri setlerinin başlıca özellikleri.

| Veri Seti | Örnek Sayısı n | | | Öznitelik Sayısı p |
|---|---|---|---|---|
| | Hasta | Sağlıklı | Toplam | |
| **Akciğer kanseri** | 60 | 60 | 120 | 54675 |
| **Lenfoma** | 6 | 8 | 14 | 22283 |
| **Rahim ağzı kanseri** | 19 | 20 | 39 | 54675 |
| **Meme kanseri** | 24 | 18 | 42 | 22283 |
| **Prostat kanseri** | 40 | 39 | 79 | 22283 |
| **Lösemi** | 38 | 26 | 64 | 22283 |

Genomik ya da klinik veri setlerinin veri önişleme aşamasından geçirilmesi veri analizlerinde önemlidir. Özellikle büyük veri setlerinde verinin analizine başlamadan önce veri önişleme adımlarına dikkat edilirse sorunsuz şekilde analiz gerçekleştirilir (68). Tez çalışmasında da altı farklı kanser türüne ait mikrodizi gen ifade verilerinde kayıp değer incelemesi, verinin ölçeklenmesi, sınıf sayılarının dağılımında dengesizlik durumu incelemesi ve aykırı değer analizi yapıldı. Yapılan veri önişleme adımları ve sonuçları Şekil 3.10.`da gösterilmiştir (68).




**Şekil 3.10.** Çalışmada kullanılan temel önişleme yöntemleri.

Ele alınan veri setlerinde kayıp ve aykırı değerler olmadığı, sınıf sayılarının dağılımında dengesiz veri setleri durumunun gözlenmediği sonucuna ulaşılmıştır. Veriye ait ortalama ve standart sapmayı dikkate alan Z-score ölçeklendirmesi kullanılarak özniteliklerin birbirine olan üstünlük durumu önlenmiştir.

Ayrıca öznitelik seçim ve sınıflama yöntemleri uygulanırken; veri setinin satırlar örnekleri, sütunlar öznitelikleri içerecek şekilde matris biçiminde olması gerektiği için satır veri yapısı biçiminde olan gen ifade veri matrisinin transpozu alınarak işlemlere devam edilmiştir.

Bu tez çalışmasında kullanılan akciğer kanseri, lenfoma, rahim ağzı kanseri, meme kanseri, prostat kanseri ve lösemi kanser türlerine ait mikrodizi gen ifade verilerinin uygulama için hazır olan matris yapısında beş örnek ve beş özniteliğin yer aldığı matris gösterimleri de Tablo 3.4.`te yer almaktadır.

**Tablo 3.4.** Kanser türlerine ait mikrodizi gen ifade verilerinin 5x5`lik matris gösterimi.

| | | **1007_s_at** | **1053_at** | **117_at** | **121_at** | **1255_g_at** |
|---|---|---|---|---|---|---|
| **Akciğer kanseri** | **GSM494565** | 3,522370 | 3,076800 | 3,398528 | 3,023763 | 2,021298 |
| | **GSM494594** | 3,503374 | 2,877318 | 3,112263 | 3,137363 | 1,863292 |
| | **GSM494604** | 3,363255 | 3,011691 | 2,971893 | 3,077472 | 1,961931 |
| | **GSM494564** | 3,518862 | 2,964107 | 2,874331 | 3,125130 | 2,131362 |
| | **GSM494591** | 3,662490 | 3,143335 | 3,053122 | 3,127609 | 1,911726 |
| | | **1007_s_at** | **1053_at** | **117_at** | **121_at** | **1255_g_at** |
| **Lenfoma** | **GSM135264** | 10,61314 | 8,650334 | 7,372430 | 11,25709 | 6,752213 |
| | **GSM135265** | 11,49526 | 7,694880 | 8,783980 | 11,96011 | 7,571373 |
| | **GSM135266** | 11,79969 | 8,155324 | 8,377644 | 11,09625 | 6,488644 |
| | **GSM135267** | 10,27181 | 7,360189 | 8,208966 | 12,08547 | 6,444601 |
| | **GSM135268** | 11,42527 | 8,589464 | 8,510171 | 11,55742 | 6,145677 |
| | | **1007_s_at** | **1053_at** | **117_at** | **121_at** | **1255_g_at** |
| **Rahim Ağzı kanseri** | **GSM651831** | 3,445435 | 3,192743 | 2,887991 | 3,230809 | 2,035768 |
| | **GSM651832** | 3,474722 | 3,118611 | 2,720178 | 3,243599 | 2,011095 |
| | **GSM651833** | 3,254458 | 3,079707 | 2,884485 | 3,268322 | 2,045279 |
| | **GSM651834** | 3,494838 | 3,169465 | 2,863827 | 3,210272 | 2,012436 |
| | **GSM651835** | 3,435562 | 3,137199 | 2,724030 | 3,188539 | 2,042137 |
| | | **1007_s_at** | **1053_at** | **117_at** | **121_at** | **1255_g_at** |
| **Meme kanseri** | **GSM512539** | 11,26526 | 4,738768 | 6,368070 | 9,880043 | 6,165912 |
| | **GSM512540** | 11,74639 | 7,312883 | 7,927185 | 9,809768 | 6,457791 |
| | **GSM512541** | 10,91625 | 4,963474 | 7,230741 | 9,715619 | 6,236493 |
| | **GSM512542** | 11,21535 | 7,136479 | 6,571373 | 9,766363 | 5,860466 |
| | **GSM512543** | 11,57710 | 6,127221 | 7,709429 | 9,420802 | 4,990955 |
| | | **1007_s_at** | **1053_at** | **117_at** | **121_at** | **1255_g_at** |
| **Prostat kanseri** | **GSM617581** | 9,639883 | 11,09315 | 9,561288 | 10,78962 | 10,52484 |
| | **GSM617582** | 9,578184 | 10,87912 | 9,288405 | 10,68141 | 10,59572 |
| | **GSM617588** | 10,465158 | 11,75926 | 10,708222 | 11,50109 | 11,45373 |
| | **GSM617590** | 10,094869 | 11,23697 | 9,524150 | 11,00267 | 10,84674 |
| | **GSM617592** | 10,853076 | 11,54332 | 10,537995 | 11,95870 | 11,44305 |
| | | **1007_s_at** | **1053_at** | **117_at** | **121_at** | **1255_g_at** |
| **Lösemi** | **GSM239371** | 1,711680 | 2,743269 | 2,141096 | 2,930001 | 1,183925 |
| | **GSM239487** | 1,569530 | 2,837544 | 2,773157 | 2,827697 | 1,123660 |
| | **GSM239489** | 1,752260 | 2,801486 | 2,181720 | 2,989673 | 1,196670 |
| | **GSM239492** | 1,583509 | 2,795430 | 2,169207 | 2,927432 | 1,156526 |
| | **GSM239497** | 1,742571 | 2,646976 | 2,490039 | 2,983045 | 1,234317 |

### 3.4.2. Benzetim Çalışması ile Elde Edilen Veri Setleri

Mikrodizi gen ifade verilerine ait benzetim çalışmasının planlanmasında Doulaye Dembele`nin yazarı olduğu 2013 yılında yayınlanmış "A Flexible Microarray Data Simulation Model" başlıklı makaleden yararlanılmıştır. Veri setleri R programının ***madsim*** (Microarray Data Simulation) paketi aracılığıyla türetilmiştir. Mikrodizi gen ifade verisi üretmek için kullanılan model ile mevcut platformlar tarafından yaygın olarak oluşturulan verilerle benzer özelliklere sahip veriler türetilmektedir (110).

Mikrodizi çalışmaları sonucu oluşan veri üzerinde direkt analiz yapılamadığı için veri bazı işlemlerden geçirilir. Belli bir dosya formatında saklanarak analizlerin yapılması için hazır hale getirilir. Platforma göre dosyaların yapısı değişmektedir.

Affymetrix, Agilent ve Illumina olmak üzere üç değişik platform, yani veri dosyaları vardır. Bu platformların birbirinden farklı olan özellikleri, kullanılan teknolojilerdir. Affymetrix, Agilent ve Illumina platformlarındaki veriler ile R ve Bioconductor üzerinde çalışılmaktadır (68). Veriyi üretmek için uygulanan fonksiyonda kullanıcının farklı veriler oluşturmasını sağlayan parametreler, diğer bir deyişle değişkenlik gösteren özellikler bulunmaktadır. Bu parametrelere ait açıklamalar veri setlerinin türetilmesi için kullanılan fonksiyondan sonra yer almaktadır. R programında yer alan ***madsim*** paketindeki `madsim()` fonksiyonu aracılığıyla veri setlerinin türetilmesi için kullanılan fonksiyonlardan biri aşağıdaki gibidir (111).

```
R >bnz-1<-madsim(mdata=NULL,n=1000,ratio=0,
fparams=data.frame(m1=40,m2=40,shape2=4,lb=4,ub=14,pde=0.02,
sym=0.5,dparams=data.frame(lambda1=0.13,lambda2=2,
muminde=1,sdde=0.5), sdn=0.4, rseed=50)
```

Fonksiyonda yer alan parametrelerden ilki `mdata`; başlangıç olarak kullanılacak sayısal değerlerin olduğu, uzunluğunun yüzden büyük olması önerilen bir veri çerçevesidir, NULL (boş) olarak ayarlandığında üretilen veriler bütünüyle yapaydır. `n`; oluşturulan verideki öznitelik (gen) sayısını belirten bir tamsayıdır. Gen ifade verilerinin genellikle log2 tabanındaki değerleri kullanılır ve bu şekilde veri türetmek için varsayılan ayarlarda `ratio=0` alınır. `fparams` ise `m1,m2,shape2,lb,ub,pde` ve `sym` olmak üzere yedi bileşenden oluşmaktadır. `m1` ve `m2`; hasta ve sağlıklı örnek sayılarıdır. Mikrodizi gen ifade verilerine ait yapılan benzetim çalışmasında yüksek değerlerden daha küçük değerler elde etmek için beta dağılımı kullanılır ve beta dağılımı şekil parametreleri için varsayılan değerler `shape1=2, shape2=4` olarak ayarlanmıştır. `lb` ve `ub` parametreleri ise log2 yoğunlukları değişim aralığında alt ve üst sınırı ifade etmektedir. Gerçek Affymetrix GeneChip© dizi verileri için gen ifade profili oluşturmak yani mikrodizi görüntülerini sayısallaştırmak amacıyla `lb` ve `ub` için sırasıyla [2,6] ve [8,16] aralıklarındaki değerler kullanılır. Fonksiyonun varsayılan ayarlarında `lb=4`, `ub=14` alınmıştır. `pde`; veri setinde farklı şekilde ifade edilmiş genlerin yüzdesidir, varsayılan ayarlarda `pde=0.02`dir. `sym=0,5` olduğu için yukarı ve aşağı düzenlenmiş genlerin sayısı neredeyse aynıdır. `dparams` ise `lambda1, lambda2, muminde` ve `sdde` olmak üzere dört bileşen

içermektedir. Üstel dağılım parametresi olan `lambda1`; ortalama gen ifade seviyelerinin değişim aralığını ifade etmektedir. Kuvvetli ve zayıf olarak ifade edilen genlerin düşük ve yüksek değişkenlikte nasıl dağıtılacağı belirlenir ve 0.13 alınmıştır.`lambda2`, `muminde` ve `sdde` parametreleri ise farklı ifade edilmiş genler için değişiklikler oluşturmak amacıyla kullanılır, yani varyasyon parametreleridir. Bu parametreler ile hasta ve sağlıklı örnekler arasında ortalama gen ifade seviyelerindeki değişiklikler meydana gelir. Varsayılan değeri 0.4 olan `sdn` ise ilave gürültü için standart sapma olarak kullanılır. Son olarak `rseed`; bilgisayar tarafından rastgele sayı üretmek için kullanılan başlangıç tamsayı değeridir, varsayılan değer ellidir (111).

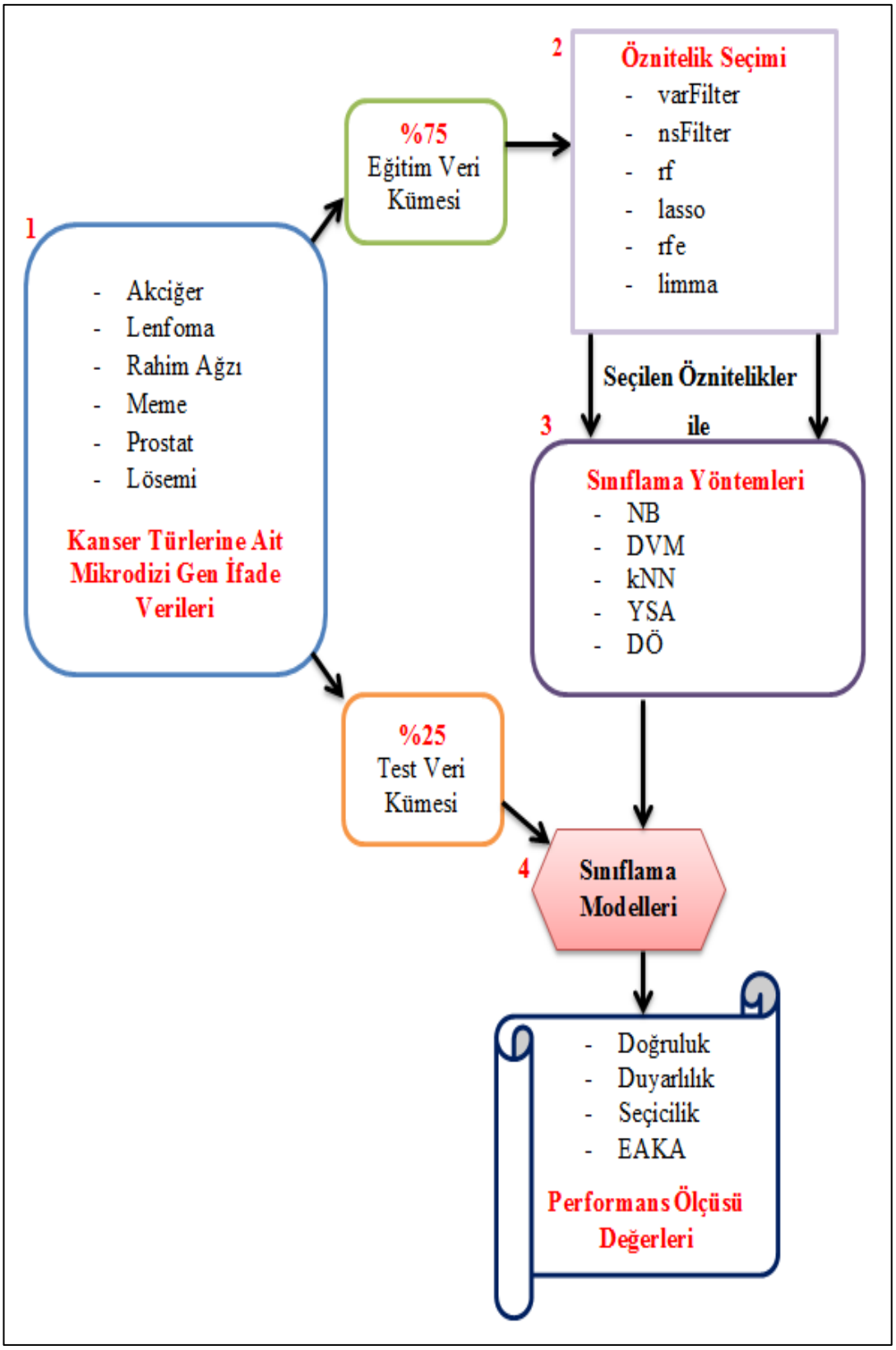
Benzetim çalışması ile elde edilen benzetim-1 (bnz-1), benzetim-2 (bnz-2), benzetim-3 (bnz-3) ve benzetim-4 (bnz-4) veri setlerine ait özellikler Tablo 3.5`te gösterilmiştir.

**Tablo 3.5.** Benzetim çalışması ile elde edilen veri setlerinin başlıca özellikleri.

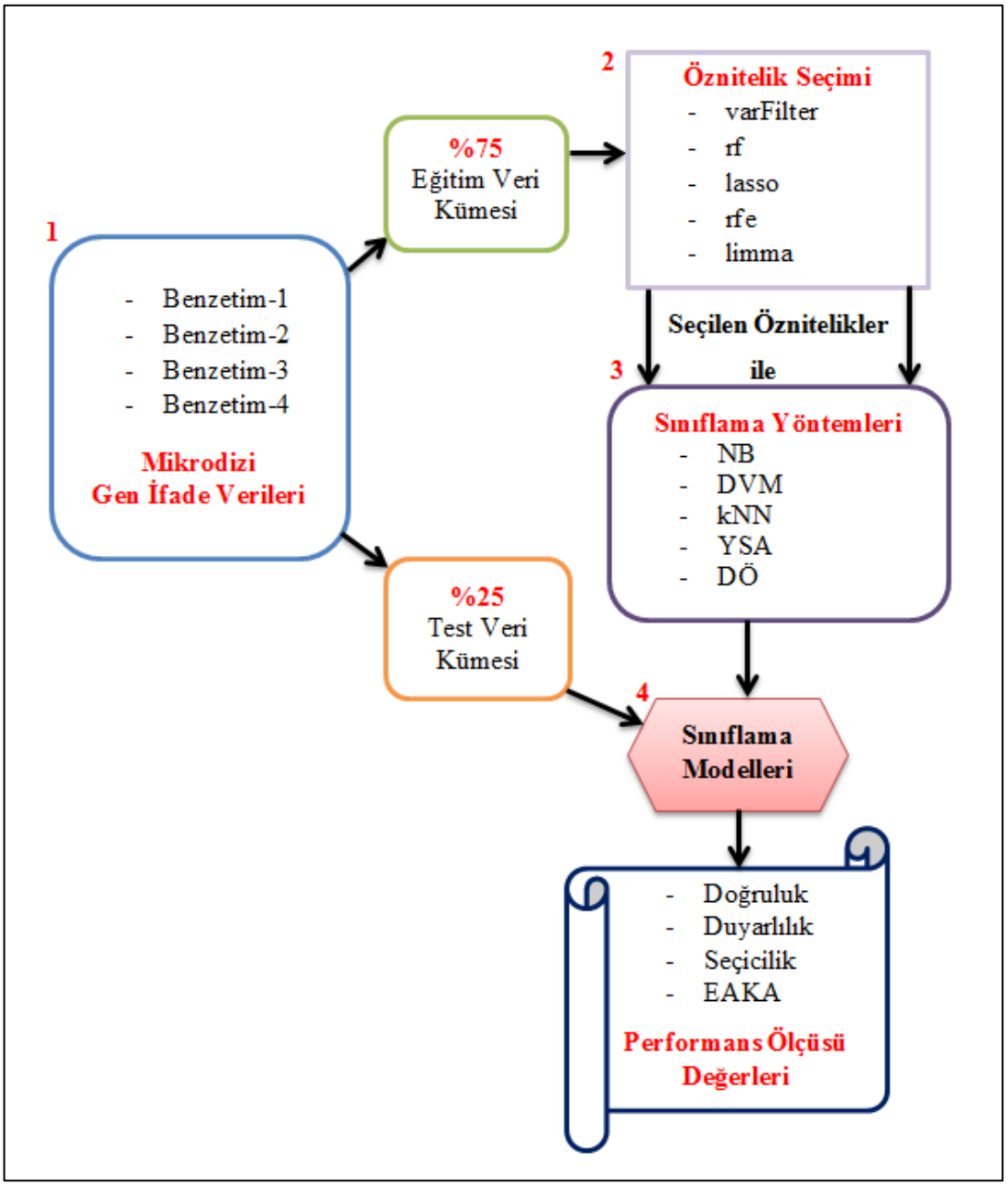
| Veri Seti | Öznitelik Sayısı p |
|---|---|
| **Bnz-1** | 250 |
| **Bnz-2** | 500 |
| **Bnz-3** | 750 |
| **Bnz-4** | 1000 |

Bnz-1, bnz-2, bnz-3 ve bnz-4 veri setlerinin her birinde kırk hasta ve kırk sağlıklı olmak üzere toplam seksen örnek vardır. Bnz-1 veri setinde 250, bnz-2 veri setinde 500, bnz-3 veri setinde 750 ve bnz-4 veri setinde ise 1000 öznitelik bulunmaktadır. Dört veri setinde örnek sayıları aynı iken; öznitelik sayıları değişmektedir. Mikrodizi gen ifade verilerinin genel özelliği örnek sayısı az, öznitelik sayısının çok olmasıdır. Bu tez çalışmasındaki benzetim çalışması ile öznitelik sayısında artışın olduğu veri setleri oluşturulmuştur. Bin tekrar yapılarak, öznitelik sayısındaki değişim ile öznitelik seçim yöntemlerinin ve sınıflama yöntemlerinin performansları değerlendirilmiştir.

Gereç ve Yöntem bölümünde bahsedilen yöntemlerin gerçek veri setleri ve benzetim çalışması ile elde edilen veri setleri üzerinde uygulama adımları sırasıyla Şekil 3.11. ve Şekil 3.12.`de gösterilmiştir.




**Şekil 3.11.** Gerçek veri setlerinde kullanılan yöntemlerin temel uygulama adımları.




**Şekil 3.12.** Benzetim çalışmasından elde edilen veri setlerinde kullanılan yöntemlerin temel uygulama adımları.

Gerçek ve benzetim çalışması ile elde edilen veri setlerinde Şekil 3.11. ve Şekil 3.12`de özet olarak verilen işlem akışının gerçekleştirilmesiyle, tez çalışmasına ait uygulama sonuçları elde edilmiştir. Elde edilen sonuçlar, tablo ve şekiller aracılığıyla Bulgular bölümünde yer almaktadır.

# 4. BULGULAR

Akciğer, lenfoma, rahim ağzı, meme, prostat ve lösemi kanser türlerine ait mikrodizi gen ifade verilerinden oluşan gerçek veri setlerine ve benzetim çalışmasına ait bulgular tablo ve şekil olmak üzere iki farklı şekilde verilmiştir.

## 4.1. Gerçek Veri Setlerine Ait Bulgular

İlk olarak akciğer kanserine ait sonuçlar Tablo 4.1.`de yer almaktadır.

**Tablo 4.1.** Akciğer kanseri veri setinde öznitelik seçim yöntemleriyle belirlenen öznitelikler kullanılarak oluşturulan sınıflama modellerinin performanslarının karşılaştırılması.

| Öznitelik Seçim Yöntemi | Sınıflama Yöntemi | Doğruluk | Duyarlılık | Seçicilik | EAKA |
|---|---|---|---|---|---|
| varFilter | NB | 0,833 | 0,928 | 0,751 | 0,839 |
| | DVM | 0,833 | 0,928 | 0,751 | 0,839 |
| | kNN | 0,733 | 0,562 | 0,929 | 0,753 |
| | YSA | 0,799 | 0,849 | 0,750 | 0,800 |
| | DÖ | 0,951 | 1,000 | 0,923 | 0,945 |
| nsFilter | NB | 0,766 | 0,764 | 0,769 | 0,776 |
| | DVM | 0,900 | 0,882 | 0,923 | 0,889 |
| | kNN | 0,767 | 0,562 | 1,000 | 0,751 |
| | YSA | 0,833 | 0,919 | 0,750 | 0,776 |
| | DÖ | 0,933 | 1,000 | 0,926 | 0,950 |
| rf | NB | 0,965 | 0,967 | 0,960 | 0,965 |
| | DVM | 0,908 | 0,900 | 0,917 | 0,910 |
| | kNN | 0,950 | 0,967 | 0,933 | 0,941 |
| | YSA | 0,858 | 0,767 | 0,950 | 0,875 |
| | DÖ | 0,970 | 1,000 | 0,965 | 0,976 |
| lasso | NB | 0,975 | 0,970 | 0,990 | 0,982 |
| | DVM | 0,958 | 0,967 | 0,950 | 0,960 |
| | kNN | 0,960 | 0,970 | 0,950 | 0,960 |
| | YSA | 0,925 | 0,900 | 0,950 | 0,939 |
| | DÖ | 0,965 | 0,955 | 0,970 | 0,961 |
| rfe | NB | 0,975 | 0,967 | 0,983 | 0,976 |
| | DVM | 0,917 | 0,950 | 0,883 | 0,930 |
| | kNN | 0,975 | 0,967 | 0,983 | 0,976 |
| | YSA | 0,933 | 0,883 | 0,983 | 0,940 |
| | DÖ | 0,733 | 1,000 | 0,700 | 0,753 |
| limma | NB | 0,970 | 0,970 | 0,970 | 0,962 |
| | DVM | 0,958 | 0,983 | 0,933 | 0,950 |
| | kNN | 0,970 | 0,970 | 0,970 | 0,962 |
| | YSA | 0,858 | 0,767 | 0,950 | 0,785 |
| | DÖ | 0,986 | 1,000 | 0,975 | 0,988 |

Akciğer kanseri veri setine ait sonuçlar incelendiğinde, varFilter öznitelik seçim yöntemi uygulandıktan sonra DÖ sınıflama yöntemi ile elde edilen modelin diğer yöntemlere göre performansı daha iyidir. NB ve DVM`nin de performansları aynıdır ve en iyi ikinci sıradaki modellerdir. YSA sınıflama yöntemi ile elde edilen modelin başarısı da NB ve DVM`den sonra gelmektedir. kNN ile elde edilen sınıflama modelinin performansında ise seçicilik değeri dışında performans ölçüsü değerleri diğer modellere göre oldukça düşüktür.

nsFilter öznitelik seçim yönteminin kullanılması ile oluşturulan sınıflama modelleri arasında en iyi performans DÖ yöntemi ile elde edilmiştir. DÖ modelinin başarısını ise sırasıyla DVM ve YSA takip etmektedir. Seçicilik performansı açısından en iyi model kNN`dir. Ancak diğer performans ölçüsü değerlerinde kNN ve NB`nin performansları birbirine yakındır ve diğer yöntemlere göre genel olarak daha düşüktür.

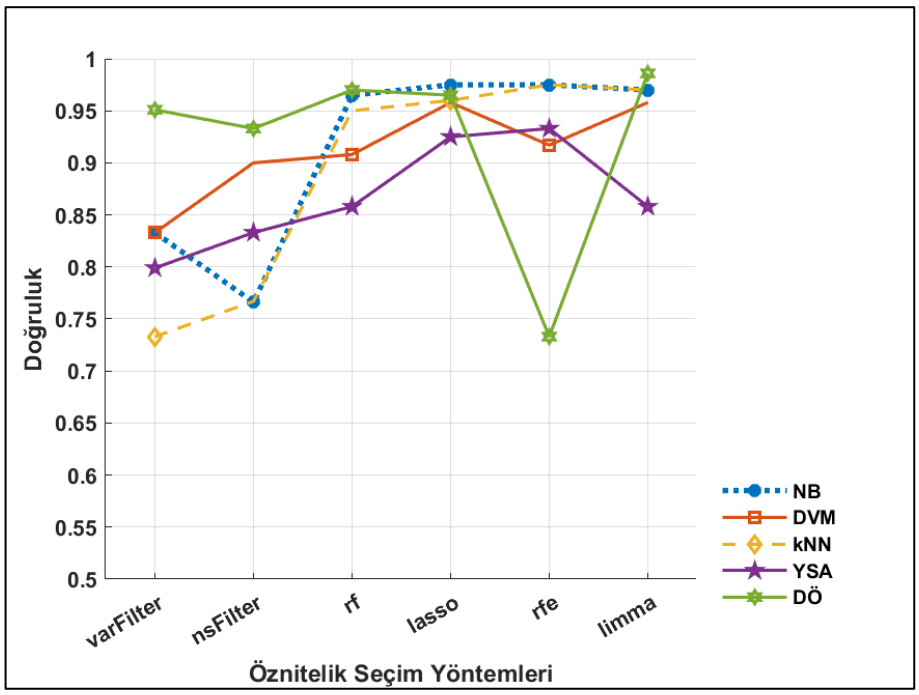
rf ile seçilen öznitelikler aracılığıyla oluşturulan sınıflama modelleri içerisinde en iyi performansa varFilter ve nsFilter`da olduğu gibi DÖ yöntemi ile ulaşıldı. DÖ sınıflama yöntemini sırasıyla NB, kNN, DVM yöntemlerinin performansları takip etmektedir. YSA`nın ise diğer sınıflama yöntemlerine göre performansı düşüktür.

lasso öznitelik seçim yönteminde ise NB yöntemi ile elde edilen sınıflama modeli en iyi performansa sahiptir. DVM, kNN ve DÖ`nün performans değerleri de yaklaşık olarak birbirlerine yakındır ve oldukça iyidir. Sınıflama yöntemleri ile kıyaslandığında YSA`nın başarısı çok az miktarda düşüktür.

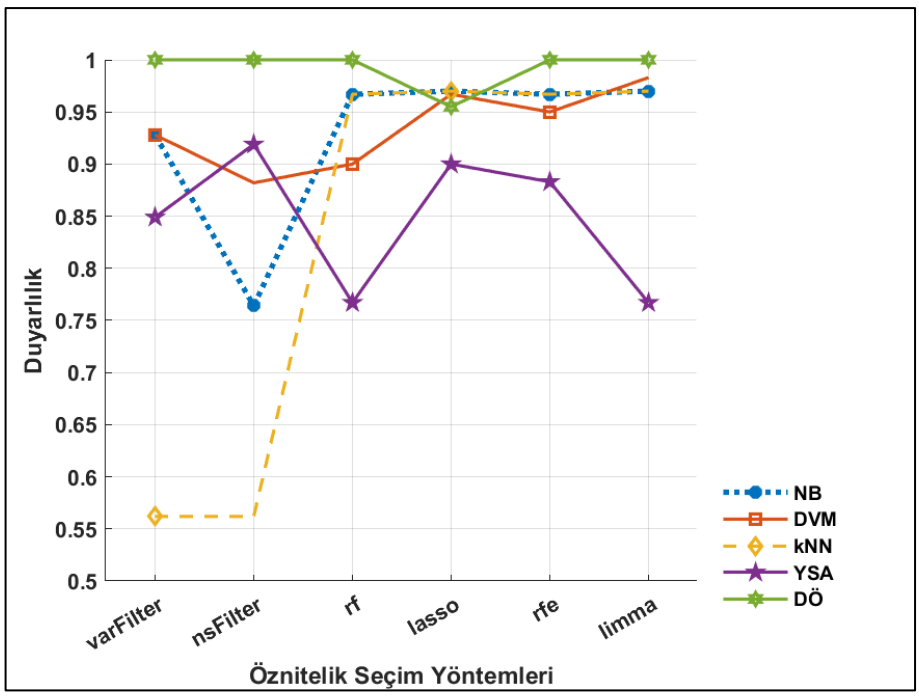
rfe öznitelik seçim yöntemi sonrasında uygulanan sınıflama yöntemlerinin performans sıralamasında ise daha önce ortaya çıkan sıralamalara göre farklılıklar vardır. Duyarlılık değeri hariç DÖ modelinin performansı diğer sınıflama yöntemleri ile elde edilen modellere göre daha düşüktür. NB ve kNN aynı ve en iyi performansa sahiptir, bu sırayı YSA ve DVM takip etmektedir. YSA yöntemi ile oluşturulan sınıflama modelleri arasında en iyi performansa rfe öznitelik seçim yönteminin kullanılması ile elde edilen sınıflamada ulaşılmıştır.

Son olarak, limma öznitelik seçim yöntemi kullanılarak elde edilen sınıflama modelleri arasında en iyi performans DÖ yöntemi ile elde edilmiştir. NB ve kNN`nin de performansları aynı olup DÖ`den sonra gelmektedir. DVM ile elde edilen

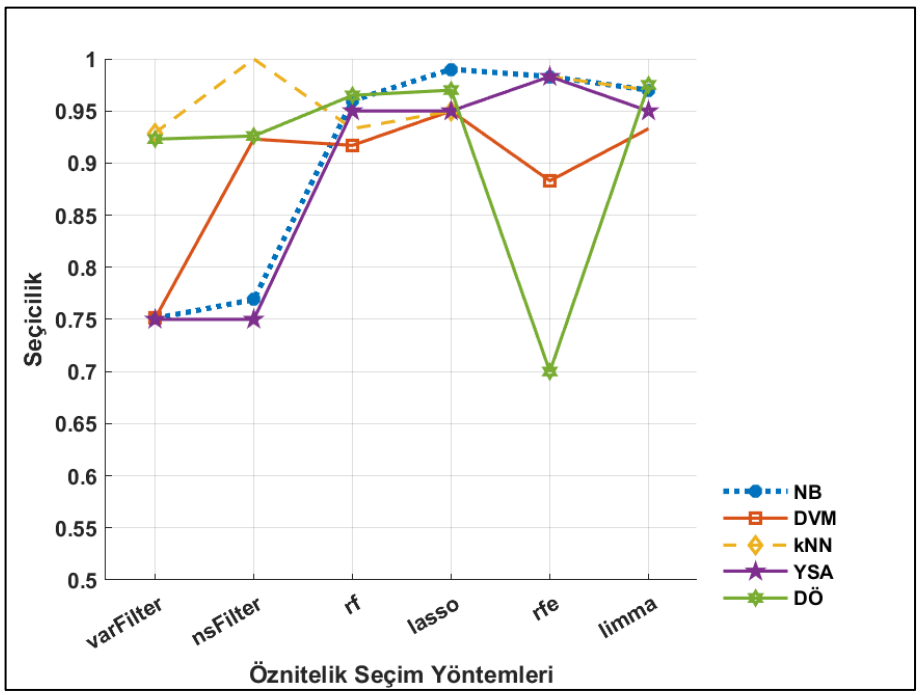
sınıflama modelinin başarısı da NB ve kNN`ye yakındır ve YSA`dan daha iyidir. Elde edilen sonuçlar grafikler aracılığıyla da Şekil 4.1.`de verilmiştir.



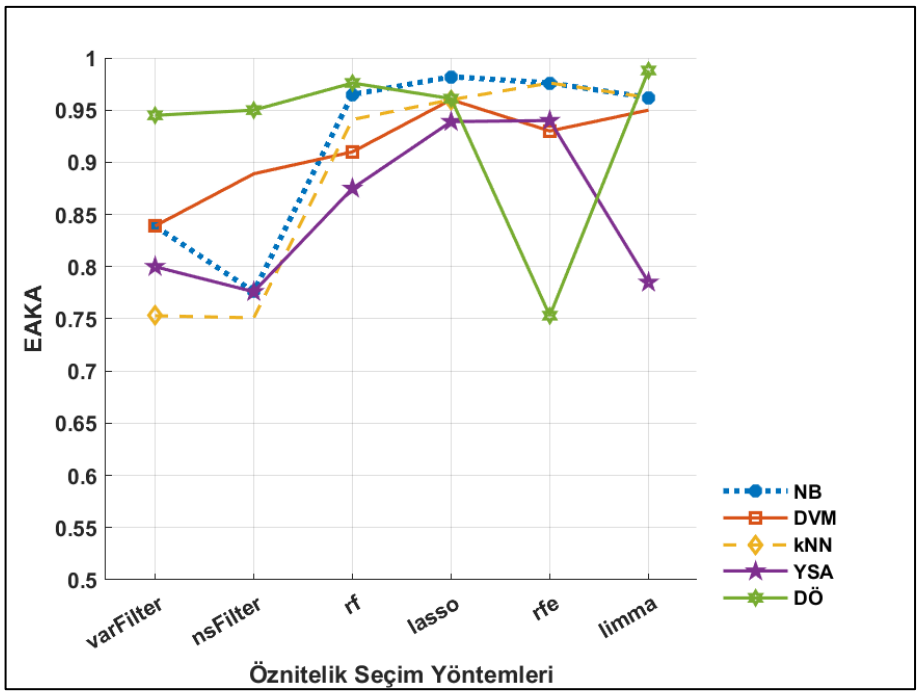

**A**




**B**




**C**




**D**

**Şekil 4.1.** Akciğer kanseri veri seti için farklı öznitelik seçim yöntemlerinde sınıflama yöntemlerinin doğruluk (A), duyarlılık (B), seçicilik (C) ve EAKA (D) performanslarının karşılaştırılması.

Akciğer kanseri veri setinde sınıflama yöntemleri ile elde edilen modellerin genel olarak performansı iyidir. lasso ve limma öznitelik seçim yöntemleri ile elde edilen sınıflama modellerinin başarısı daha yüksek iken; varFilter ve nsFilter öznitelik seçim yöntemlerinde ise sınıflama yöntemleri daha düşük performansa sahiptir. Sınıflama yöntemleri içerisinde genel olarak DÖ daha iyi, YSA ise daha düşük performans göstermiştir.

**Tablo 4.2.** Lenfoma veri setinde öznitelik seçim yöntemleriyle belirlenen öznitelikler kullanılarak oluşturulan sınıflama modellerinin performanslarının karşılaştırılması.

| Öznitelik Seçim Yöntemi | Sınıflama Yöntemi | Doğruluk | Duyarlılık | Seçicilik | EAKA |
|---|---|---|---|---|---|
| varFilter | NB | 0,750 | 0,500 | 1,000 | 0,750 |
| | DVM | 0,750 | 0,500 | 1,000 | 0,750 |
| | kNN | 0,750 | 0,667 | 1,000 | 0,800 |
| | YSA | 0,750 | 0,500 | 1,000 | 0,750 |
| | DÖ | 0,975 | 0,970 | 0,980 | 0,985 |
| nsFilter | NB | 0,750 | 1,000 | 0,667 | 0,800 |
| | DVM | 0,750 | 0,667 | 1,000 | 0,800 |
| | kNN | 0,750 | 0,500 | 1,000 | 0,750 |
| | YSA | 0,750 | 1,000 | 0,500 | 0,700 |
| | DÖ | 0,988 | 0,985 | 0,992 | 0,990 |
| rf | NB | 1,000 | 1,000 | 1,000 | 1,000 |
| | DVM | 0,933 | 0,900 | 1,000 | 0,980 |
| | kNN | 0,900 | 0,800 | 1,000 | 0,960 |
| | YSA | 0,633 | 0,660 | 0,606 | 0,650 |
| | DÖ | 0,949 | 1,000 | 0,900 | 0,960 |
| lasso | NB | 0,933 | 0,900 | 1,000 | 0,980 |
| | DVM | 0,933 | 0,900 | 1,000 | 0,980 |
| | kNN | 0,870 | 0,750 | 0,950 | 0,930 |
| | YSA | 0,700 | 0,750 | 0,650 | 0,700 |
| | DÖ | 1,000 | 1,000 | 1,000 | 1,000 |
| rfe | NB | 0,833 | 0,600 | 1,000 | 0,800 |
| | DVM | 0,867 | 0,700 | 0,950 | 0,900 |
| | kNN | 0,867 | 0,700 | 0,950 | 0,900 |
| | YSA | 0,567 | 0,500 | 0,600 | 0,550 |
| | DÖ | 0,900 | 1,000 | 0,850 | 0,950 |
| limma | NB | 0,950 | 0,800 | 1,000 | 0,960 |
| | DVM | 0,950 | 0,800 | 1,000 | 0,960 |
| | kNN | 0,950 | 0,800 | 1,000 | 0,960 |
| | YSA | 0,740 | 0,500 | 1,000 | 0,800 |
| | DÖ | 1,000 | 1,000 | 1,000 | 1,000 |

Lenfoma veri setine ait sonuçları içeren Tablo 4.2. incelendiğinde, varFilter öznitelik seçim yöntemi uygulandıktan sonra DÖ sınıflama yöntemi ile elde edilen modelin

performansı diğerlerine göre oldukça yüksektir. kNN sınıflama yöntemi ile oluşturulan modelin performansı ise NB, DVM ve YSA`ya göre çok az bir farkla daha iyidir. NB, DVM ve YSA yöntemleri ile elde edilen modellerin ise performans değerleri aynıdır.

nsFilter öznitelik seçim yönteminin kullanılması ile oluşturulan modeller arasında en iyi performans varFilter`da olduğu gibi DÖ sınıflama yöntemi ile elde edilmiştir. DÖ`den sonra benzer performans ölçüleri olan NB ile DVM ve kNN ile YSA yöntemleri ile oluşturulan sınıflama modelleri gelmektedir.

rf ile seçilen öznitelikler aracılığıyla oluşturulan sınıflama modelleri içerisinde en iyi performansa NB yöntemi ile ulaşıldı. NB sınıflama yöntemini sırasıyla DÖ, DVM ve kNN yöntemlerinin performansları takip etmektedir. Performans değeri en düşük olan sınıflama modeli ise YSA yöntemi ile elde edilmiştir.

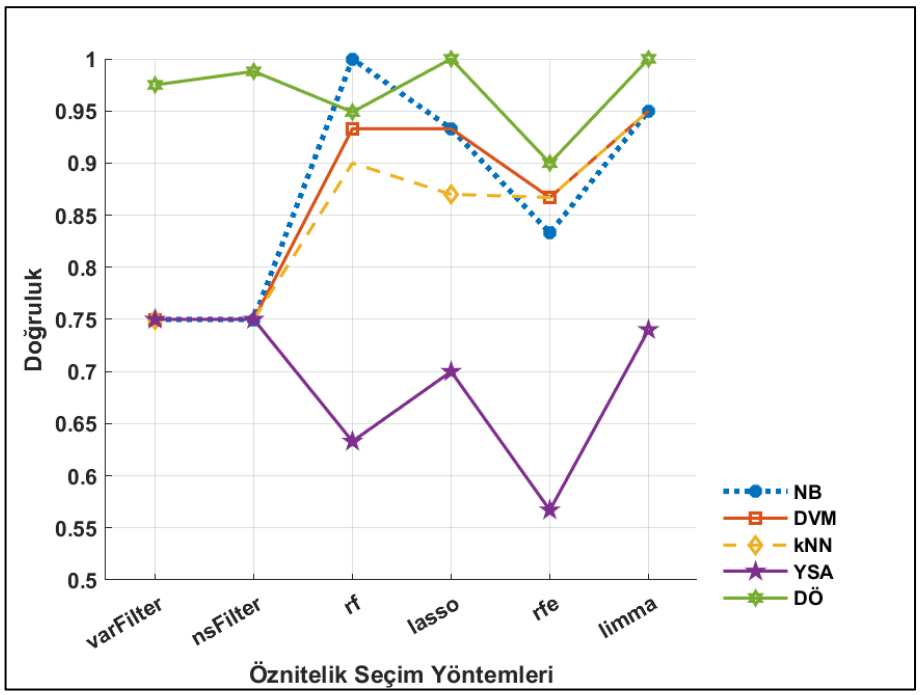
lasso öznitelik seçim yönteminde ise DÖ yöntemi ile elde edilen sınıflama modeli en iyi performansa sahiptir. NB ve DVM`nin performans değerleri aynıdır ve DÖ`den sonra gelmektedir. kNN`nin ise performans değerleri DÖ, NB ve DVM sınıflama yöntemlerine göre daha düşük iken; YSA`ya göre daha iyidir. YSA`nın başarısı ise diğer sınıflama yöntemlerine göre daha düşüktür.

rfe öznitelik seçim yöntemi sonrasında elde edilen sınıflama modelleri arasında DÖ en iyi performansa sahiptir. DVM ve kNN sınıflama yöntemleri aynı performans değerlerine sahiptir ve DÖ`nün performans değerlerine yakındır. NB`nin performans değerleri ise DVM ve kNN`den sonra gelmektedir. YSA ise en düşük performans değerlerine sahip sınıflama yöntemidir.

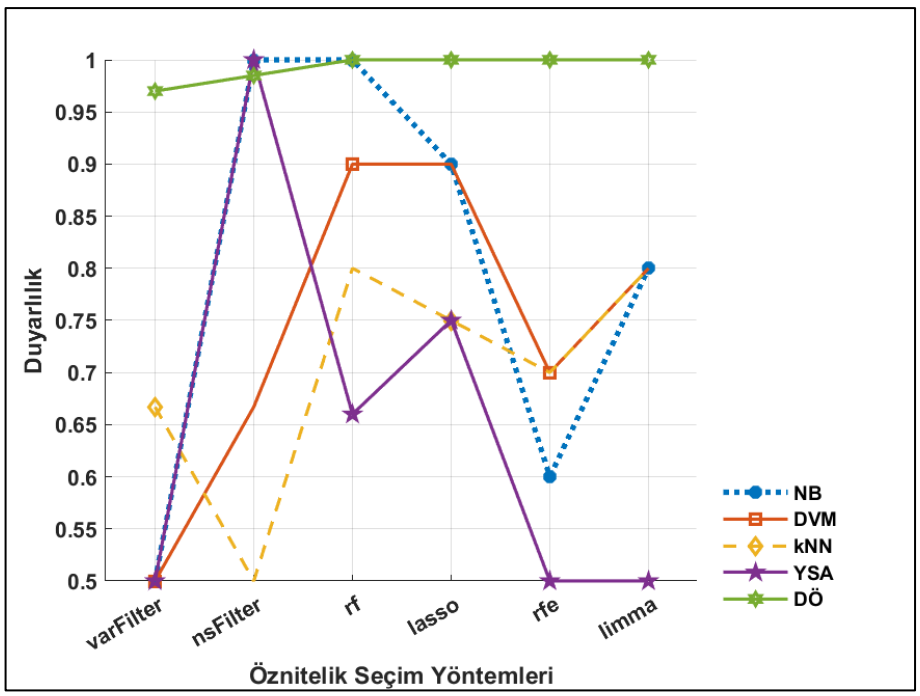
Son olarak, limma öznitelik seçim yöntemi kullanılarak elde edilen sınıflama modelleri arasında en iyi performans DÖ yöntemi ile elde edilmiştir. NB, DVM ve kNN yöntemleri ile oluşturulan sınıflama modellerinin performans değerleri aynı olup DÖ`den sonra gelmektedir. Sınıflama yöntemleri içerisinde YSA`ya ait modelin seçicilik performansı hariç diğer ölçüleri en düşük performansa sahiptir.

Akciğer kanseri veri setinde olduğu gibi lenfoma veri setinde de genel olarak DÖ yöntemi ile oluşturulan sınıflama modellerinin performans ölçüsü değerleri diğer modellere göre daha iyi çıkmıştır. Öznitelik seçim yöntemleri içinde de lasso

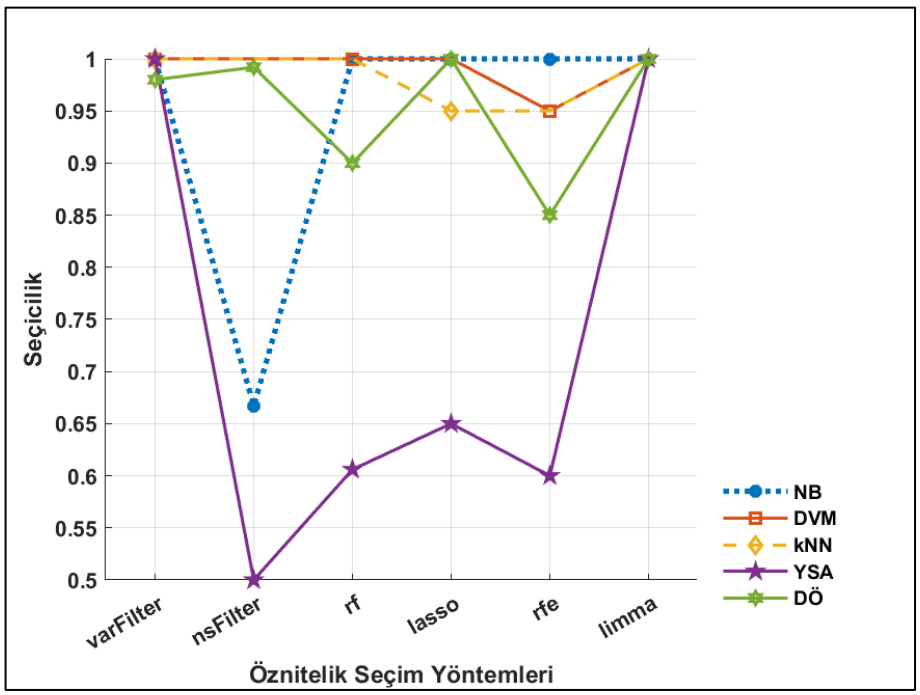
ve limma yöntemlerinin kullanılması ile elde edilen sınıflama modellerinin başarısı daha yüksektir. Elde edilen sonuçlar grafikler aracılığıyla da Şekil 4.2.`de verilmiştir.



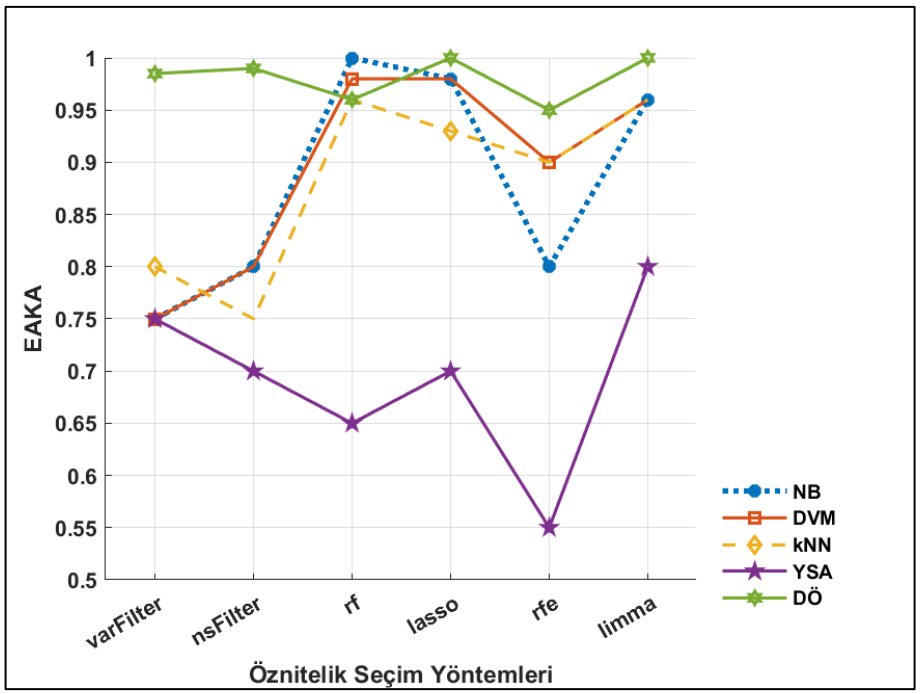

**A**




**B**




**C**




**D**

**Şekil 4.2.** Lenfoma veri seti için farklı öznitelik seçim yöntemlerinde sınıflama yöntemlerinin doğruluk (A), duyarlılık (B), seçicilik (C) ve EAKA (D) performanslarının karşılaştırılması.

Lenfoma veri setinde genel olarak DÖ yönteminin performansı en iyidir. nsFilter ve varFilter öznitelik seçim yöntemlerinde YSA`nın performans değerleri NB, DVM ve kNN` ye yakındır. Ancak diğer öznitelik seçim yöntemlerinde en düşük performansa sahip sınıflama yöntemi YSA`dır. rf öznitelik seçim yönteminde NB, lasso ve limma öznitelik seçim yöntemlerinde ise DÖ sınıflama modelleri doğruluk, duyarlılık, seçicilik ve EAKA açısından neredeyse %100 performans göstermiştir. Genel olarak rf, lasso ve limma öznitelik seçim yöntemlerinin kullanılması ile elde edilen sınıflama modellerinin başarısı daha yüksektir.

**Tablo 4.3.** Rahim ağzı kanseri veri setinde öznitelik seçim yöntemleriyle belirlenen öznitelikler kullanılarak oluşturulan sınıflama modellerinin performanslarının karşılaştırılması.

| Öznitelik Seçim Yöntemi | Sınıflama Yöntemi | Doğruluk | Duyarlılık | Seçicilik | EAKA |
|---|---|---|---|---|---|
| varFilter | NB | 0,658 | 0,600 | 0,714 | 0,705 |
| | DVM | 0,867 | 1,000 | 0,714 | 0,850 |
| | kNN | 0,920 | 1,000 | 0,833 | 0,950 |
| | YSA | 0,600 | 0,600 | 0,600 | 0,600 |
| | DÖ | 0,885 | 0,845 | 0,985 | 0,880 |
| nsFilter | NB | 0,700 | 0,666 | 0,734 | 0,725 |
| | DVM | 0,867 | 1,000 | 0,714 | 0,850 |
| | kNN | 0,700 | 0,571 | 0,833 | 0,782 |
| | YSA | 0,550 | 0,500 | 0,600 | 0,500 |
| | DÖ | 0,910 | 0,902 | 0,928 | 0,915 |
| rf | NB | 0,639 | 0,617 | 0,650 | 0,671 |
| | DVM | 0,689 | 0,567 | 0,800 | 0,724 |
| | kNN | 0,721 | 0,683 | 0,750 | 0,758 |
| | YSA | 0,639 | 0,567 | 0,700 | 0,581 |
| | DÖ | 0,862 | 1,000 | 0,823 | 0,836 |
| lasso | NB | 0,671 | 0,633 | 0,700 | 0,679 |
| | DVM | 0,593 | 0,533 | 0,650 | 0,623 |
| | kNN | 0,614 | 0,617 | 0,600 | 0,628 |
| | YSA | 0,651 | 0,733 | 0,600 | 0,656 |
| | DÖ | 0,925 | 1,000 | 0,923 | 0,952 |
| rfe | NB | 0,593 | 0,567 | 0,600 | 0,608 |
| | DVM | 0,693 | 0,783 | 0,600 | 0,716 |
| | kNN | 0,618 | 0,705 | 0,500 | 0,607 |
| | YSA | 0,568 | 0,400 | 0,750 | 0,625 |
| | DÖ | 0,865 | 0,950 | 0,850 | 0,865 |
| limma | NB | 0,693 | 0,633 | 0,750 | 0,729 |
| | DVM | 0,639 | 0,683 | 0,600 | 0,674 |
| | kNN | 0,696 | 0,700 | 0,690 | 0,757 |
| | YSA | 0,721 | 0,467 | 0,950 | 0,779 |
| | DÖ | 0,965 | 1,000 | 0,929 | 0,985 |

Rahim ağzı kanseri veri setine ait sonuçları içeren Tablo 4.3. incelendiğinde, varFilter öznitelik seçim yöntemi uygulandıktan sonra kNN sınıflama yöntemi ile elde edilen modelin performansı diğerlerine göre daha iyidir. NB, DVM ve YSA`ya göre daha iyi performansı olan DÖ`nün performans değerleri kNN`den sonra gelmektedir. Sınıflama yöntemleri içerisinde en düşük performans YSA`nındır.

nsFilter öznitelik seçim yönteminin kullanılması ile oluşturulan sınıflama modelleri arasında en iyi performans DÖ yöntemi ile elde edilmiştir. DÖ`den sonra birbirine yakın performans ölçüleri olan DVM, kNN ve NB gelmektedir. YSA`nın ise performans ölçüsü değerleri diğer sınıflama yöntemlerine göre daha düşüktür.

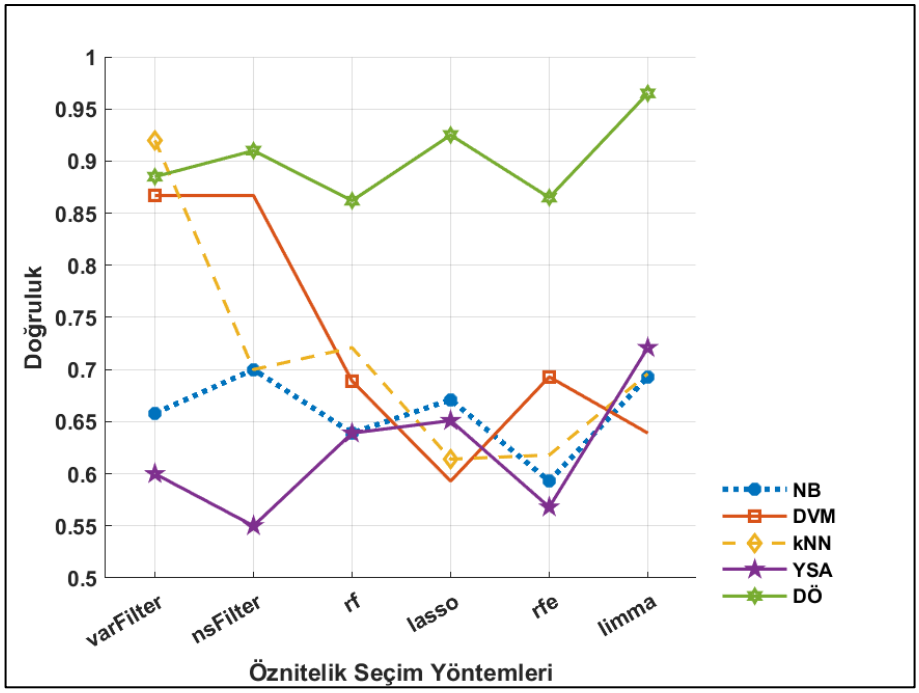
rf ile seçilen öznitelikler aracılığıyla oluşturulan sınıflama modelleri içerisinde en iyi performansa nsFilter`da olduğu gibi DÖ yöntemi ile ulaşıldı. DÖ sınıflama yöntemini sırasıyla kNN, DVM ve NB yöntemlerinin performansları takip etmektedir. Performans değeri en düşük olan sınıflama modeli ise YSA yöntemi ile elde edilmiştir.

lasso öznitelik seçim yönteminde ise DÖ sınıflama yöntemi ile oluşturulan modelin performansı diğer sınıflama yöntemlerine göre oldukça yüksektir. DÖ`den sonra NB`nin performansı gelmektedir. Şimdiye kadar elde edilen sonuçlardan farklı olarak, DVM yöntemi ile en düşük performansa sahip sınıflama modeli elde edilmiştir. YSA`nın performansı ise kNN`den daha iyi olup NB`ye yakındır.

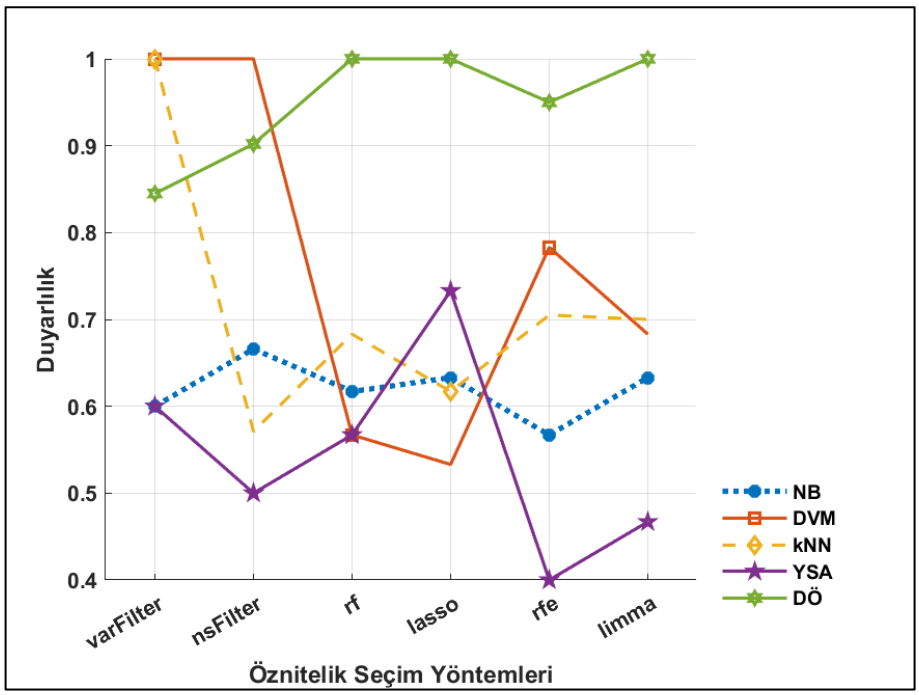
rfe öznitelik seçim yöntemi sonrasında elde edilen sınıflama modelleri arasında DÖ daha iyi performansa sahiptir. DÖ`den sonra DVM gelmektedir. NB ile kNN sınıflama yöntemlerinin performans değerleri ise birbirine yakındır. varFilter, nsFilter ve rf öznitelik seçim yöntemlerinde olduğu gibi rfe öznitelik seçim yönteminde de YSA en düşük performansa sahip sınıflama yöntemidir.

Son olarak, limma öznitelik seçim yöntemi kullanılarak elde edilen sınıflama modelleri arasında en iyi performans DÖ yöntemi ile elde edilmiştir. YSA yöntemi ile oluşturulan sınıflama modeli doğruluk, seçicilik ve EAKA performans ölçüleriyle ikinci sıradaki en iyi modeldir. kNN ile NB sınıflama yöntemlerinin ise performans değerleri birbirine yakındır. lasso öznitelik seçim yönteminde olduğu gibi DVM sınıflama yöntemi ile elde edilen modelin performans ölçüsü değerleri diğerlerine göre daha düşüktür.

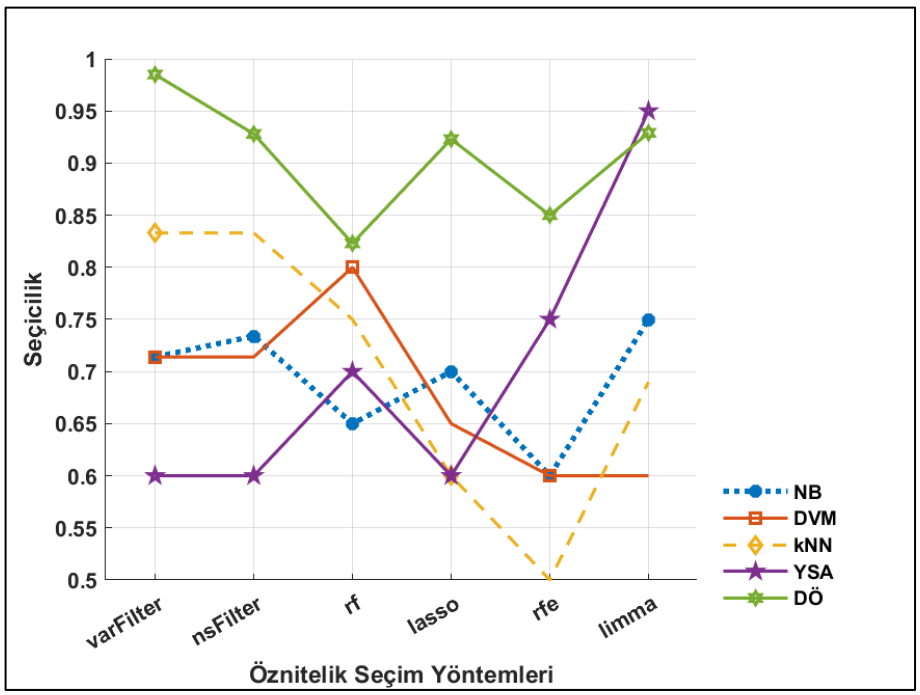
Elde edilen sonuçlar grafikler aracılığıyla da Şekil 4.3.`te verilmiştir.



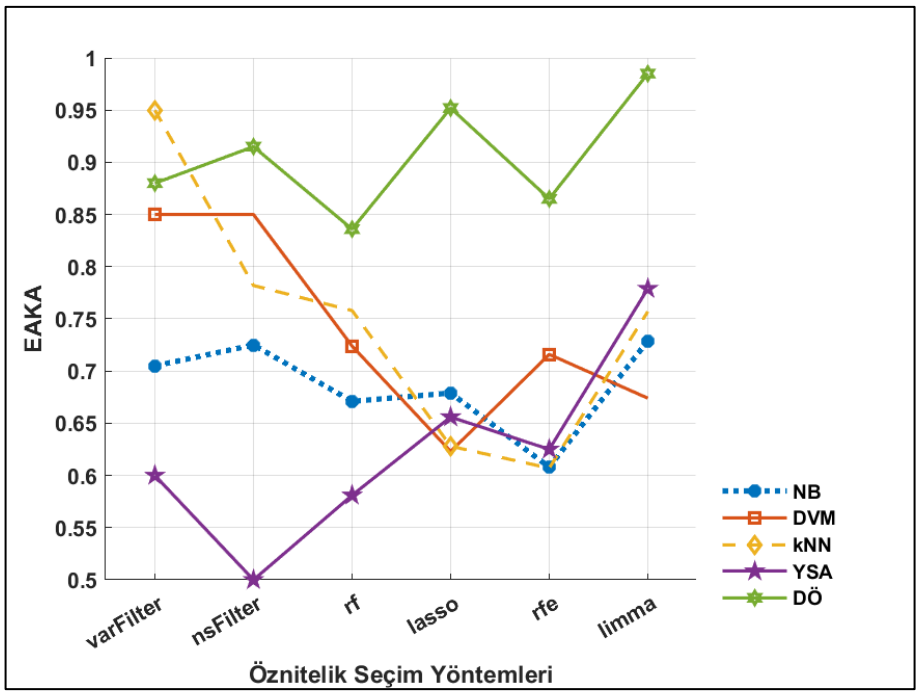

**A**




**B**




**C**




**D**

**Şekil 4.3.** Rahim ağzı veri seti için farklı öznitelik seçim yöntemlerinde sınıflama yöntemlerinin doğruluk (A), duyarlılık (B), seçicilik (C) ve EAKA (D) performanslarının karşılaştırılması.

Rahim ağzı kanseri veri setinde sınıflama yöntemleri ile elde edilen sınıflama modelleri içerisinde genel olarak DÖ yönteminin performansı en iyidir. varFilter öznitelik seçim yönteminde ise kNN`nin performansı daha iyidir. lasso ve limma öznitelik seçim yöntemleri hariç diğer yöntemlerde YSA`nın performans değerleri diğer sınıflama yöntemlerine göre daha düşüktür. lasso ve limma`da ise DVM ile elde edilen sınıflama modeli en düşük performansa sahiptir. Genel olarak öznitelik seçim yöntemleri içerisinde rf, lasso ve limma öznitelik seçim yöntemlerinin kullanılması ile elde edilen sınıflama modellerinin başarısı daha yüksektir.

**Tablo 4.4.** Meme kanseri veri setinde öznitelik seçim yöntemleriyle belirlenen öznitelikler kullanılarak oluşturulan sınıflama modellerinin performanslarının karşılaştırılması.

| Öznitelik Seçim Yöntemi | Sınıflama Yöntemi | Doğruluk | Duyarlılık | Seçicilik | EAKA |
|---|---|---|---|---|---|
| varFilter | NB | 0,636 | 0,666 | 0,625 | 0,645 |
| | DVM | 0,540 | 0,333 | 0,750 | 0,541 |
| | kNN | 0,727 | 0,875 | 0,673 | 0,691 |
| | YSA | 0,732 | 0,747 | 0,696 | 0,652 |
| | DÖ | 0,797 | 0,840 | 0,790 | 0,800 |
| nsFilter | NB | 0,540 | 0,333 | 0,750 | 0,541 |
| | DVM | 0,540 | 0,333 | 0,750 | 0,541 |
| | kNN | 0,818 | 0,963 | 0,667 | 0,798 |
| | YSA | 0,732 | 0,747 | 0,714 | 0,745 |
| | DÖ | 0,818 | 0,960 | 0,800 | 0,810 |
| rf | NB | 0,752 | 0,590 | 0,850 | 0,803 |
| | DVM | 0,714 | 0,740 | 0,683 | 0,710 |
| | kNN | 0,689 | 0,740 | 0,633 | 0,685 |
| | YSA | 0,573 | 0,880 | 0,150 | 0,630 |
| | DÖ | 0,900 | 1,000 | 0,890 | 0,910 |
| lasso | NB | 0,693 | 0,750 | 0,617 | 0,780 |
| | DVM | 0,648 | 0,790 | 0,450 | 0,640 |
| | kNN | 0,639 | 0,740 | 0,500 | 0,632 |
| | YSA | 0,566 | 0,770 | 0,383 | 0,543 |
| | DÖ | 0,909 | 1,000 | 0,850 | 0,933 |
| rfe | NB | 0,620 | 0,700 | 0,517 | 0,637 |
| | DVM | 0,664 | 0,700 | 0,617 | 0,660 |
| | kNN | 0,636 | 0,730 | 0,517 | 0,653 |
| | YSA | 0,575 | 0,660 | 0,467 | 0,508 |
| | DÖ | 0,775 | 1,000 | 0,750 | 0,753 |
| limma | NB | 0,744 | 0,750 | 0,733 | 0,757 |
| | DVM | 0,671 | 0,710 | 0,633 | 0,667 |
| | kNN | 0,633 | 0,690 | 0,567 | 0,630 |
| | YSA | 0,523 | 0,720 | 0,300 | 0,530 |
| | DÖ | 1,000 | 1,000 | 1,000 | 1,000 |

Meme kanseri veri setine ait sonuçları içeren Tablo 4.4. incelendiğinde, varFilter öznitelik seçim yöntemi uygulandıktan sonra DÖ sınıflama yöntemi ile elde edilen modelin performansı diğer yöntemlere göre daha iyidir. Performans değerleri birbirine yakın olan kNN ile YSA ise DÖ`den sonra gelmektedir. NB ile elde edilen sınıflama modelinin performansı kNN ve YSA ile oluşturulan sınıflama modellerinin performanslarına göre daha düşüktür. Duyarlılık değerinin düşük olduğu DVM ile elde edilen sınıflama modelinin performans değerleri ise diğer yöntemler ile elde edilen sınıflama modelleri içerisinde genel olarak en düşüktür.

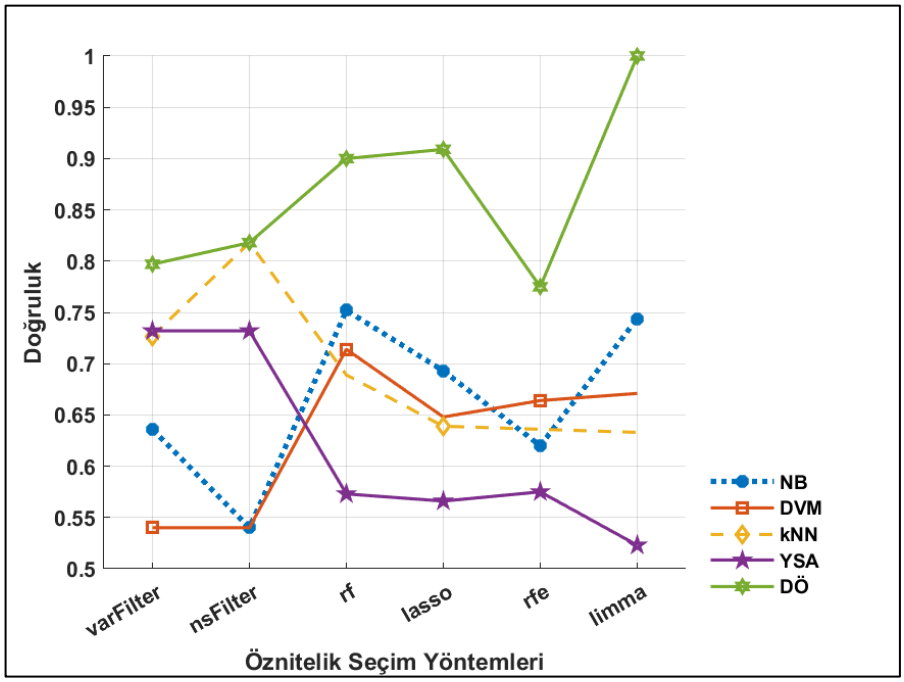
nsFilter öznitelik seçim yönteminin kullanılması ile oluşturulan sınıflama modellerinin performanslarına ait sonuçlar varFilter ile benzerdir. En iyi performans DÖ yöntemi ile elde edilmiştir. DÖ`den sonra sırasıyla kNN ve YSA yöntemleri ile elde edilen sınıflama modellerinin performansları gelir. Diğer sınıflama yöntemlerine göre daha düşük performans değerlerine sahip olan NB ve DVM ile oluşturulan modellerin performans değerleri aynıdır.

rf ile seçilen öznitelikler ile oluşturulan sınıflama modelleri içerisinde en iyi performansa varFilter ve nsFilter`da olduğu gibi DÖ yöntemi ile ulaşıldı. DÖ sınıflama yöntemini sırasıyla NB, DVM ve kNN yöntemlerinin performansları takip etmektedir. YSA yöntemi ile elde edilen modelin özellikle doğruluk, seçicilik ve EAKA ölçüleri açısından performansı diğer sınıflama yöntemlerine göre daha düşüktür.

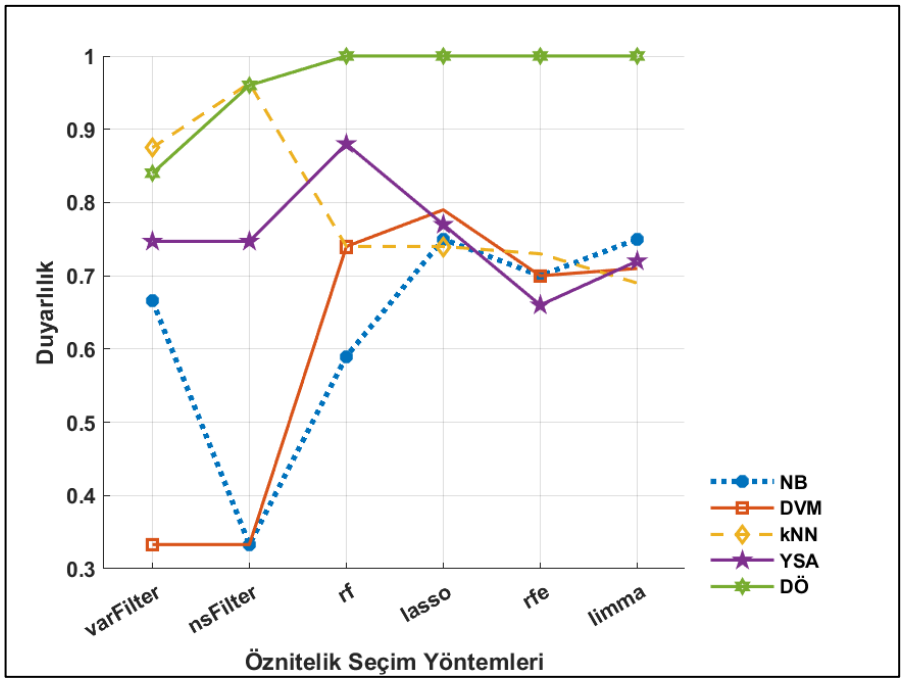
lasso öznitelik seçim yönteminde de DÖ yöntemi ile elde edilen sınıflama modeli en iyi performansa sahiptir. NB`nin performans değerleri ise DÖ`ye göre oldukça düşüktür. Ancak diğer sınıflama yöntemlerine göre daha iyidir. DVM ve kNN yöntemleri ile elde edilen sınıflama modellerinin performans değerleri birbirine yakındır ve YSA`dan daha iyi performansa sahiptirler. Diğer sınıflama yöntemlerine göre genel olarak en düşük performansı olan yöntem YSA`dır.

rfe öznitelik seçim yöntemi sonrasında elde edilen sınıflama modelleri arasında DÖ daha iyi performansa sahiptir. NB, DVM ve kNN sınıflama yöntemlerinin performans değerleri yaklaşık olarak birbirine yakındır. lasso ve rf öznitelik seçim yöntemlerinde olduğu gibi rfe öznitelik seçim yönteminde de YSA en düşük performansa sahip sınıflama yöntemidir. Son olarak, limma öznitelik seçim yöntemi kullanılarak elde edilen sınıflama modelleri arasında en iyi performans DÖ

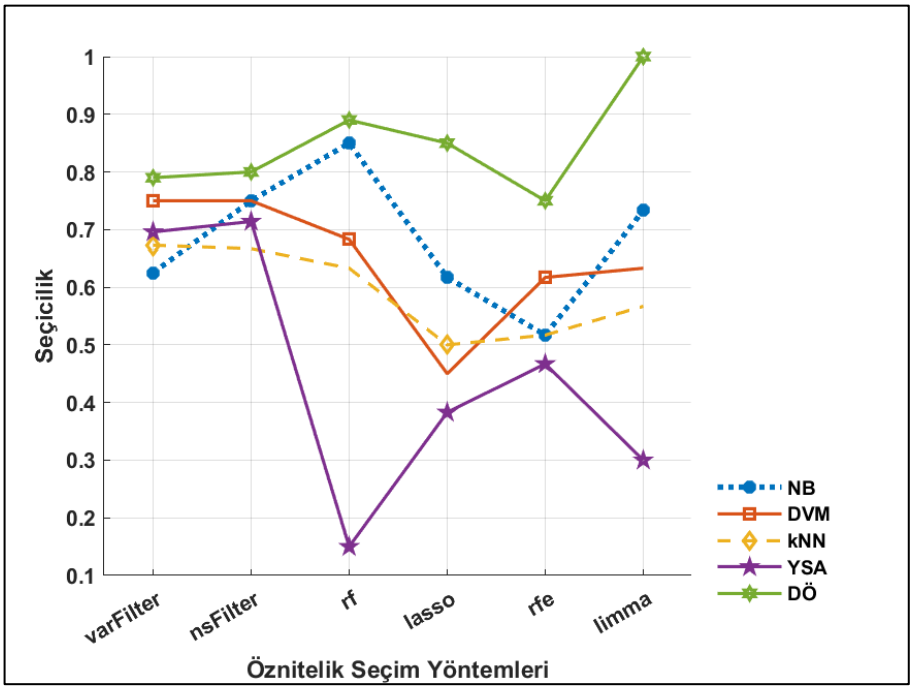
yöntemi ile elde edilmiştir. DÖ`den sonra NB`nin performansı gelmektedir. DVM ile kNN yöntemlerinin performans değerleri birbirine yakındır ve YSA`dan daha iyidir. Sınıflama yöntemleri ile elde edilen modeller arasında YSA en düşük performansa sahiptir. Elde edilen sonuçlar grafikler aracılığıyla da Şekil 4.4.`te verilmiştir.



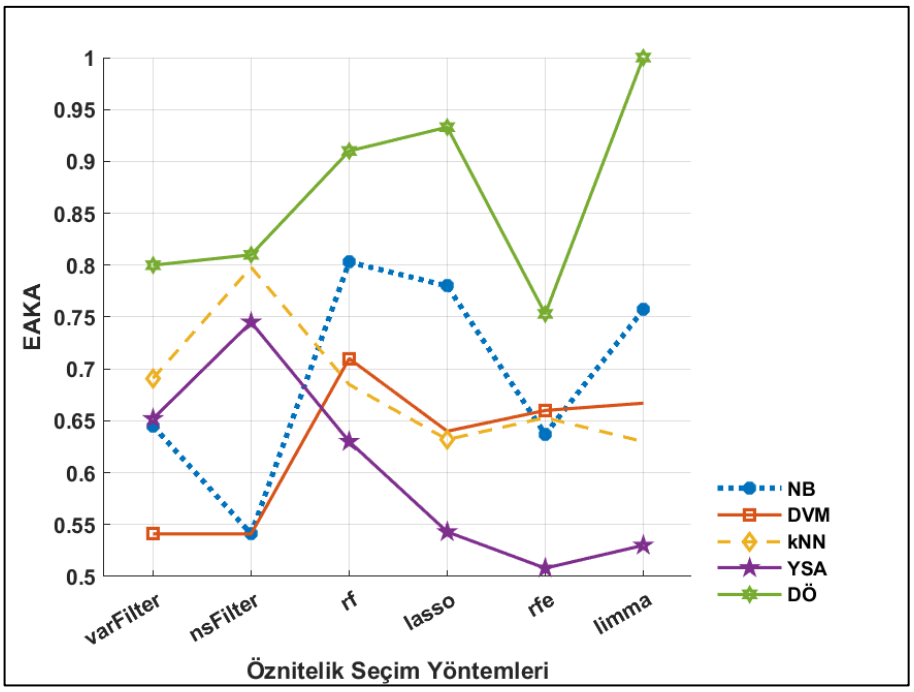

**A**




**B**




**C**




**D**

**Şekil 4.4.** Meme kanseri veri seti için farklı öznitelik seçim yöntemlerinde sınıflama yöntemlerinin doğruluk (A), duyarlılık (B), seçicilik (C) ve EAKA (D) performanslarının karşılaştırılması.

Meme kanseri veri setinde sınıflama yöntemleri ile elde edilen modeller içerisinde genel olarak YSA daha düşük, DÖ daha yüksek performansa sahiptir. Çoğunlukla lasso, limma ve rf öznitelik seçim yöntemlerinin kullanılması ile elde edilen sınıflama modellerinin başarısı daha yüksektir.

**Tablo 4.5.** Prostat kanseri veri setinde öznitelik seçim yöntemleriyle belirlenen özmitelikler kullanılarak oluşturulan sınıflama modellerinin performanslarının karşılaştırılması.

| Öznitelik Seçim Yöntemi | Sınıflama Yöntemi | Doğruluk | Duyarlılık | Seçicilik | EAKA |
|---|---|---|---|---|---|
| varFilter | NB | 0,700 | 0,666 | 0,727 | 0,697 |
| | DVM | 0,600 | 0,666 | 0,545 | 0,606 |
| | kNN | 0,600 | 0,666 | 0,545 | 0,606 |
| | YSA | 0,500 | 0,555 | 0,454 | 0,505 |
| | DÖ | 0,845 | 0,810 | 0,990 | 0,858 |
| nsFilter | NB | 0,600 | 0,666 | 0,545 | 0,606 |
| | DVM | 0,800 | 0,666 | 0,909 | 0,809 |
| | kNN | 0,650 | 0,540 | 0,860 | 0,660 |
| | YSA | 0,650 | 0,777 | 0,545 | 0,650 |
| | DÖ | 0,900 | 1,000 | 0,870 | 0,888 |
| rf | NB | 0,644 | 0,636 | 0,650 | 0,646 |
| | DVM | 0,596 | 0,561 | 0,625 | 0,605 |
| | kNN | 0,583 | 0,482 | 0,675 | 0,580 |
| | YSA | 0,493 | 0,231 | 0,801 | 0,550 |
| | DÖ | 0,835 | 0,850 | 0,820 | 0,840 |
| lasso | NB | 0,618 | 0,586 | 0,650 | 0,625 |
| | DVM | 0,578 | 0,511 | 0,625 | 0,598 |
| | kNN | 0,581 | 0,614 | 0,550 | 0,604 |
| | YSA | 0,506 | 0,275 | 0,750 | 0,500 |
| | DÖ | 0,870 | 1,000 | 0,850 | 0,875 |
| rfe | NB | 0,454 | 0,325 | 0,583 | 0,483 |
| | DVM | 0,504 | 0,557 | 0,450 | 0,532 |
| | kNN | 0,568 | 0,454 | 0,675 | 0,628 |
| | YSA | 0,543 | 0,350 | 0,750 | 0,502 |
| | DÖ | 0,800 | 1,000 | 0,620 | 0,820 |
| limma | NB | 0,682 | 0,611 | 0,750 | 0,673 |
| | DVM | 0,581 | 0,561 | 0,600 | 0,585 |
| | kNN | 0,580 | 0,664 | 0,500 | 0,650 |
| | YSA | 0,569 | 0,307 | 0,825 | 0,570 |
| | DÖ | 0,950 | 1,000 | 0,900 | 0,960 |

Prostat kanseri veri setine ait sonuçları içeren Tablo 4.5. incelendiğinde, varFilter öznitelik seçim yöntemi uygulandıktan sonra DÖ sınıflama yöntemi ile elde edilen modelin performansı diğer yöntemlere göre çok daha iyidir. NB yöntemi ile

oluşturulan modelin başarısı ise DÖ yönteminden sonra gelmektedir; fakat diğer yöntemlere göre daha iyidir. DVM ve kNN yöntemlerinin kullanılmasıyla elde edilen sınıflama modellerinin ise performans değerleri aynıdır. Bu performans değerleri NB yöntemi ile elde edilen sınıflama modelinin performans değerlerine göre daha düşük iken; YSA yöntemi ile oluşturulan sınıflama modelinin performans değerlerine göre daha iyidir.

nsFilter öznitelik seçim yönteminin kullanılması ile oluşturulan sınıflama modelleri arasında en iyi performansa DÖ yöntemi ile ulaşıldı. DÖ modelinin başarısını DVM takip etmektedir. Performans değerlerinin genelde iyi olduğu NB ise prostat kanseri veri setinde nsFilter öznitelik seçim yönteminde en düşük performansa sahip sınıflama yöntemidir.

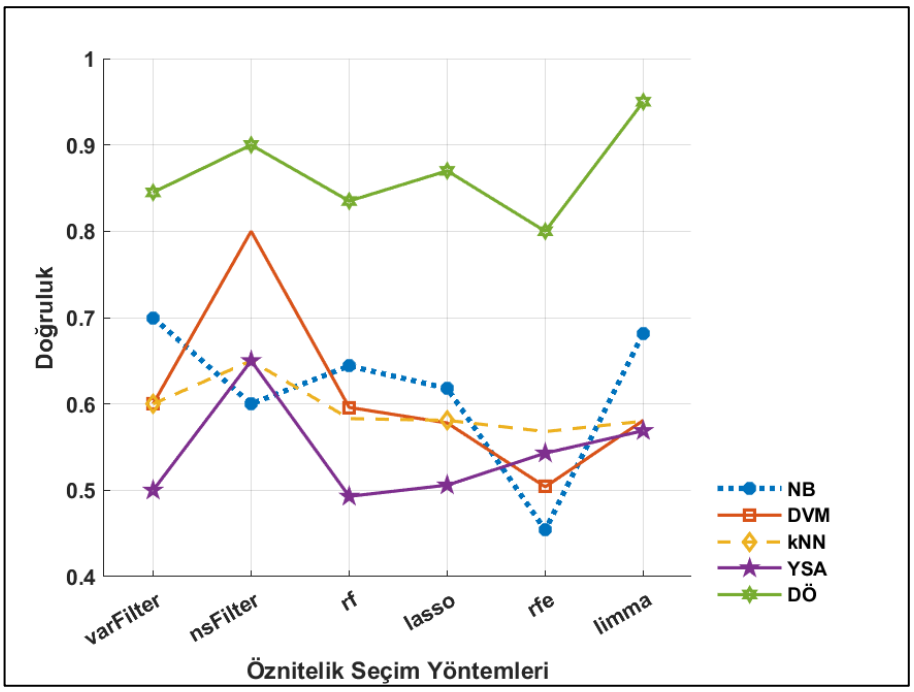
rf ile seçilen öznitelikler aracılığıyla oluşturulan sınıflama modelleri içerisinde en iyi performansa varFilter ve nsFilter`da olduğu gibi DÖ yöntemi ile ulaşıldı. DÖ sınıflama yöntemini ise performans değerleri birbirine yakın olan NB, DVM ve kNN yöntemleri takip etmektedir. Sınıflama yöntemleri içerisinde YSA`ya ait modelin seçicilik performansı DÖ`ye yakındır. Ancak duyarlılık performans ölçüsü değeri başta olmak üzere YSA`nın performansı diğer sınıflama yöntemlerine göre daha düşüktür.

lasso öznitelik seçim yönteminde en iyi performans ölçüsü değerleri DÖ sınıflama yöntemi ile elde edilmiştir. NB, kNN ve DVM yöntemleri ile oluşturulan sınıflama modellerinin performans ölçüsü değerleri birbirine yakındır ve DÖ`den sonra gelmektedir. rf`da olduğu gibi YSA yönteminin seçicilik değeri hariç performans ölçüsü değerleri diğer yöntemlere göre düşüktür.

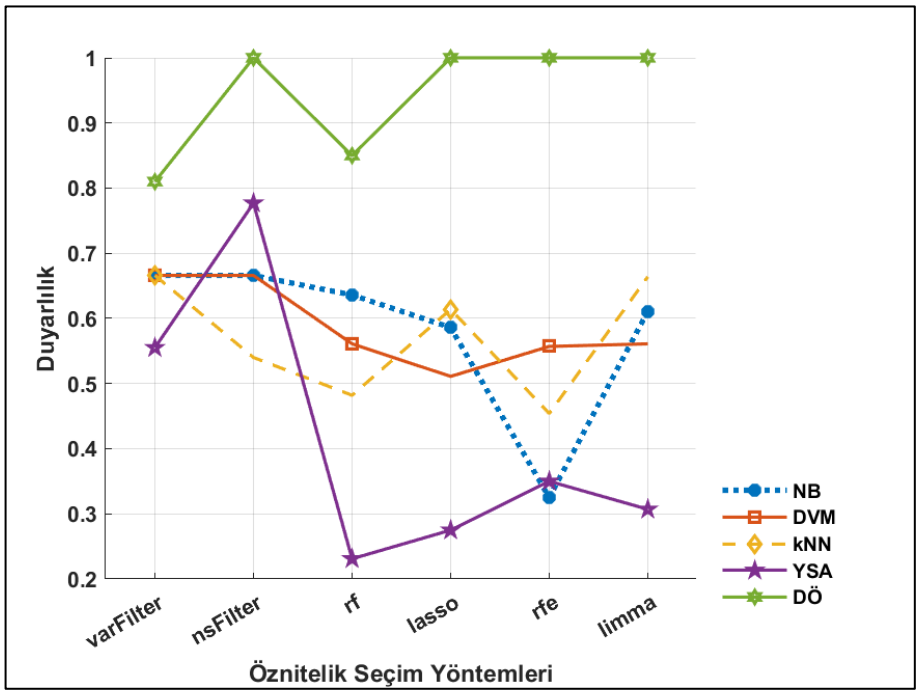
rfe öznitelik seçim yönteminin kullanılması ile oluşturulan sınıflama modelleri sıralamasında en iyi model DÖ`nündür. DÖ`den sonra kNN ve DVM sınıflama yöntemleri ile oluşturulan modellerin performansı gelmektedir. En düşük performansa sahip olan sınıflama modeli nsFilter`da olduğu gibi NB yöntemi ile elde edilmiştir.

Son olarak, limma öznitelik seçim yöntemi kullanılarak elde edilen sınıflama modelleri arasında en iyi performans DÖ yöntemi ile elde edilmiştir. NB yöntemi ile oluşturulan sınıflama modelinin performans değerleri DÖ`den sonra gelir. Performans ölçüsü değerleri birbirine yakın olan DVM ve kNN yöntemleri ile NB`ye

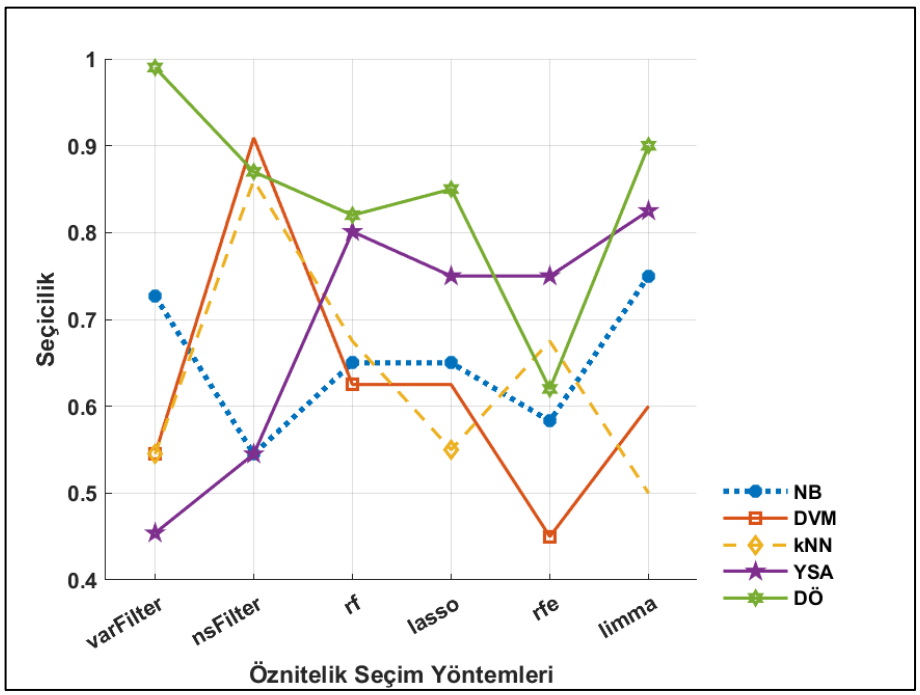
göre daha düşük, YSA`ya göre daha yüksek performansa sahip sınıflama modelleri elde edilmiştir. Elde edilen sonuçlar grafikler aracılığıyla da Şekil 4.5.`te verilmiştir.



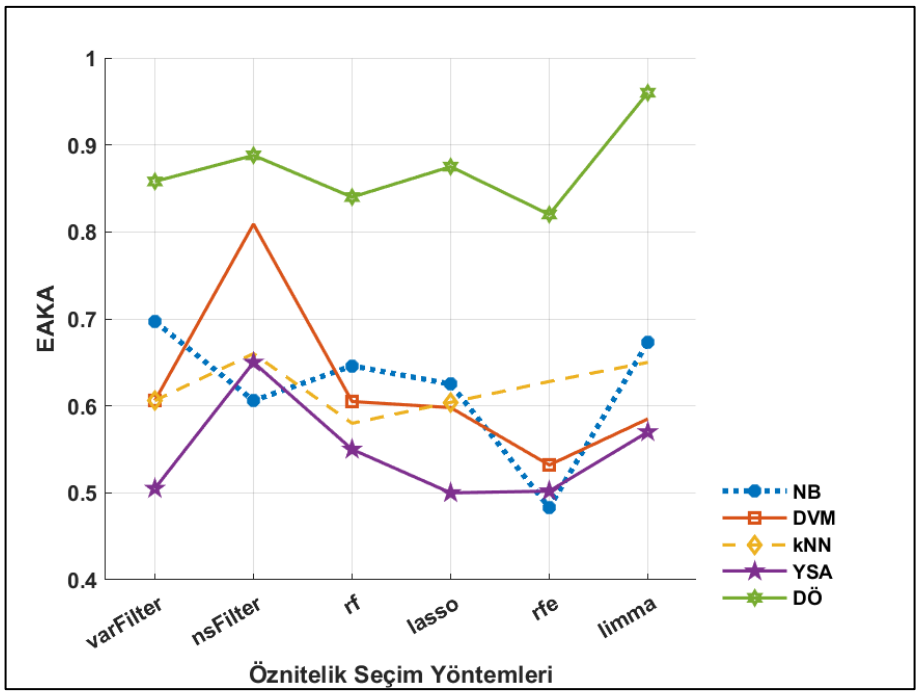

**A**




**B**




**C**




**D**

**Şekil 4.5.** Prostat kanseri veri seti için farklı öznitelik seçim yöntemlerinde sınıflama yöntemlerinin doğruluk (A), duyarlılık (B), seçicilik (C) ve EAKA (D) performanslarının karşılaştırılması.

Prostat kanseri veri setinde sınıflama yöntemleri ile elde edilen modeller içerisinde genel olarak DÖ daha yüksek, YSA daha düşük performansa sahiptir. Genel olarak rfe öznitelik seçim yöntemlerinin kullanılması ile elde edilen sınıflama modellerinin başarısı daha düşüktür. Özellikle limma ve nsFilter öznitelik seçim yöntemlerinde DÖ`nün başarısı daha iyidir.

**Tablo 4.6.** Lösemi veri setinde öznitelik seçim yöntemleriyle belirlenen öznitelikler kullanılarak oluşturulan sınıflama modellerinin performanslarının karşılaştırılması.

| Öznitelik Seçim Yöntemi | Sınıflama Yöntemi | Doğruluk | Duyarlılık | Seçicilik | EAKA |
|---|---|---|---|---|---|
| varFilter | NB | 1,000 | 1,000 | 1,000 | 1,000 |
| | DVM | 0,937 | 0,800 | 1,000 | 0,900 |
| | kNN | 1,000 | 1,000 | 1,000 | 1,000 |
| | YSA | 0,750 | 1,000 | 0,640 | 0,701 |
| | DÖ | 0,967 | 0,950 | 1,000 | 0,975 |
| nsFilter | NB | 0,801 | 0,714 | 0,888 | 0,815 |
| | DVM | 0,895 | 0,860 | 0,930 | 0,900 |
| | kNN | 0,937 | 0,833 | 1,000 | 0,900 |
| | YSA | 0,900 | 0,800 | 1,000 | 0,900 |
| | DÖ | 0,938 | 0,900 | 1,000 | 0,919 |
| rf | NB | 0,891 | 0,971 | 0,773 | 0,954 |
| | DVM | 0,937 | 1,000 | 0,867 | 0,971 |
| | kNN | 0,922 | 0,985 | 0,807 | 0,915 |
| | YSA | 0,840 | 0,946 | 0,680 | 0,870 |
| | DÖ | 0,960 | 1,000 | 0,950 | 0,965 |
| lasso | NB | 0,906 | 1,000 | 0,853 | 0,981 |
| | DVM | 0,971 | 0,980 | 0,932 | 0,975 |
| | kNN | 0,937 | 1,000 | 0,867 | 0,945 |
| | YSA | 0,971 | 0,980 | 0,932 | 0,975 |
| | DÖ | 0,985 | 1,000 | 0,980 | 0,985 |
| rfe | NB | 0,816 | 0,811 | 0,827 | 0,815 |
| | DVM | 0,948 | 0,960 | 0,920 | 0,955 |
| | kNN | 0,985 | 1,000 | 0,960 | 0,985 |
| | YSA | 0,708 | 0,850 | 0,547 | 0,685 |
| | DÖ | 0,780 | 0,650 | 0,923 | 0,793 |
| limma | NB | 0,937 | 1,000 | 0,867 | 0,950 |
| | DVM | 0,940 | 0,962 | 0,930 | 0,955 |
| | kNN | 0,953 | 1,000 | 0,887 | 0,968 |
| | YSA | 0,985 | 0,975 | 1,000 | 0,990 |
| | DÖ | 1,000 | 1,000 | 1,000 | 1,000 |

Lösemi veri setine ait sonuçları içeren Tablo 4.6. incelendiğinde, varFilter öznitelik seçim yöntemi uygulandıktan sonra NB ve kNN sınıflama yöntemleri ile elde edilen modellerin performansı diğerlerine göre daha iyidir. Sırasıyla DÖ ve DVM

yöntemleri ile elde edilen sınıflama modellerinin performansları birbirine yakındır ve NB ile kNN`den sonra gelmektedir. YSA ile oluşturulan sınıflama modelinin performans ölçüsü değerleri ise diğerlerine göre daha düşüktür.

nsFilter öznitelik seçim yönteminin kullanılması ile oluşturulan sınıflama modelleri arasında en iyi başarı DÖ yöntemi ile elde edilmiştir. DÖ`den sonra sırasıyla kNN, YSA, DVM ve NB yöntemleri ile oluşturulan sınıflama modellerinin performansları gelmektedir. YSA yöntemi ile elde edilen sınıflama modelinin performansı varFlter`a göre oldukça iyidir. nsFilter`da NB yöntemi ile oluşturulan sınıflama modelinin performans ölçüsü değerleri biraz daha düşüktür.

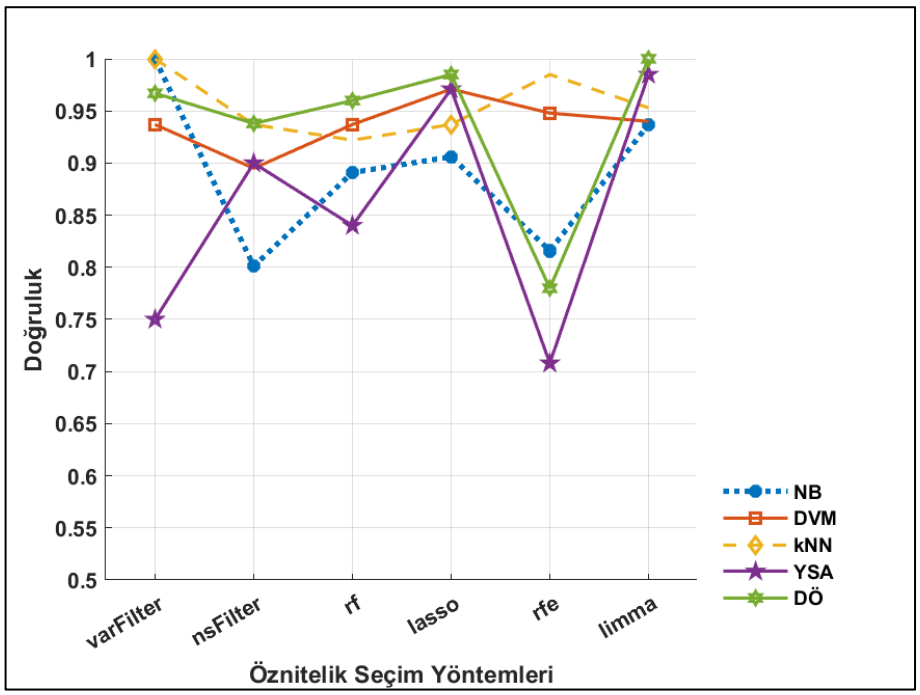
rf ile seçilen öznitelikler aracılığıyla oluşturulan sınıflama modelleri içerisinde DÖ ve DVM yöntemlerinin performans değerleri birbirine yakındır ve en iyi performansa sahiptirler. Sonrasında kNN ve NB ile elde edilen sınıflama modellerinin performansları gelmektedir. YSA`ya ait sınıflama modeli ise daha düşük performans ölçüsü değerlerine sahiptir.

lasso öznitelik seçim yönteminde ise DÖ yöntemi ile elde edilen sınıflama modeli en iyi performansa sahiptir. YSA ve DVM`nin performans değerleri aynıdır ve DÖ`den sonra gelmektedir. kNN ve NB yöntemleri ile oluşturulan sınıflama modellerinin performans değerleri ise birbirine yakındır. Ancak kNN`nin daha iyidir. NB yöntemi ile daha düşük performans ölçüsü değerlerine sahip sınıflama modeli elde edilmiştir.

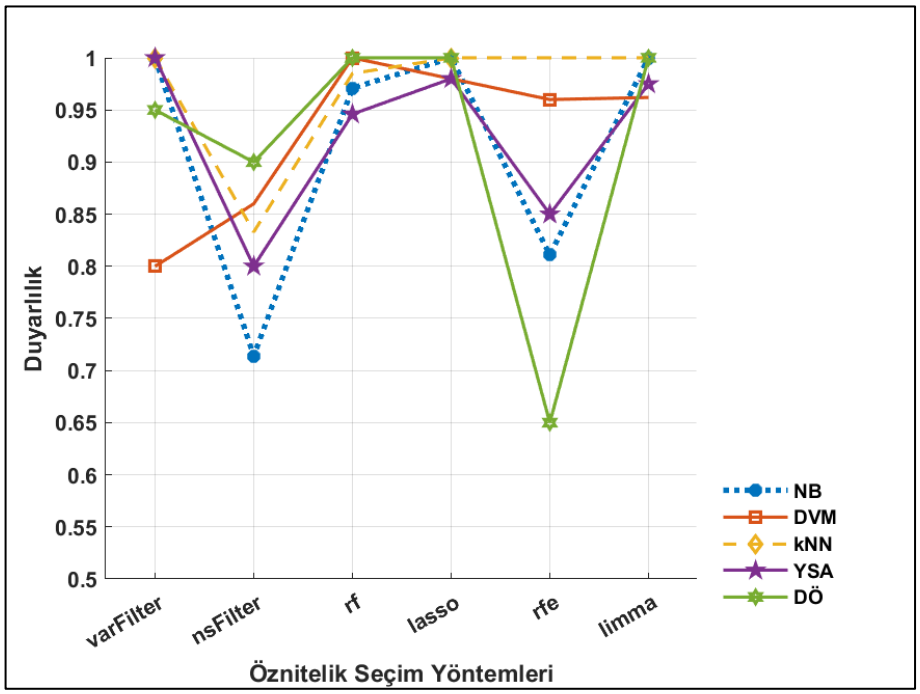
rfe öznitelik seçim yöntemi sonrasında elde edilen sınıflama modelleri arasında kNN en iyi performansa sahiptir. Sırasıyla DVM ve NB yöntemleri ile elde edilen sınıflama modellerinin performansları kNN`den sonra gelmektedir. YSA ise en düşük performans ölçüsü değerlerine sahip sınıflama yöntemidir. İlk kez DÖ yöntemi ile oluşturulan sınıflama modelinin performansı kNN, DVM ve NB`den sonra, YSA`dan önce yer aldı.

Son olarak, limma öznitelik seçim yöntemi kullanılarak elde edilen sınıflama modelleri arasında en iyi performans DÖ yöntemi ile elde edilmiştir. DÖ`den sonra sırasıyla YSA, kNN, DVM ve NB yöntemleri ile oluşturulan sınıflama modellerinin performansları gelmektedir. YSA yöntemi ile elde edilen sınıflama modelinin performansının iyi olduğu nadir durumlardan biridir. nsFilter ve lasso`da olduğu gibi NB yöntemi ile daha düşük performans ölçüsü değerlerinin olduğu sınıflama modeli

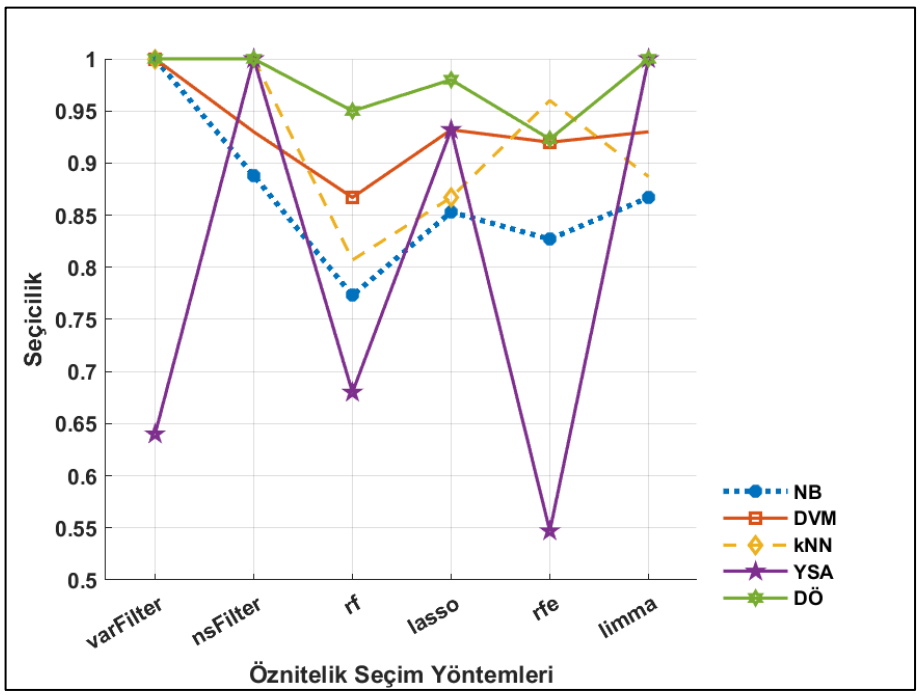
oluşturulmuştur. Elde edilen sonuçlar grafikler aracılığıyla da Şekil 4.6.`da verilmiştir.



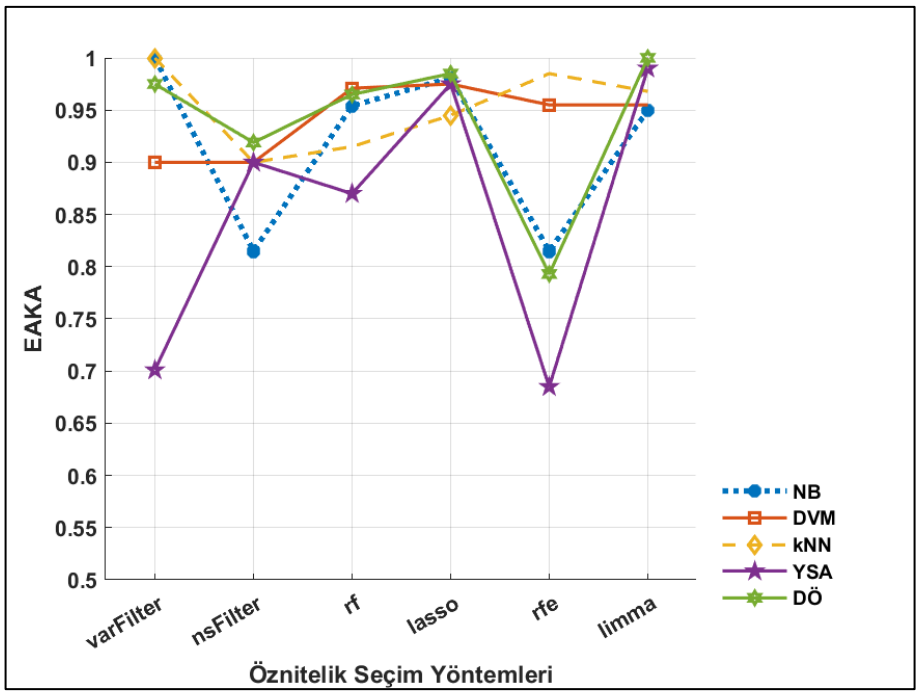

**A**




**B**




**C**




**D**

**Şekil 4.6.** Lösemi veri seti için farklı öznitelik seçim yöntemlerinde sınıflama yöntemlerinin doğruluk (A), duyarlılık (B), seçicilik (C) ve EAKA (D) performanslarının karşılaştırılması.

Lösemi veri setinde sınıflama yöntemleri ile elde edilen modeller içerisinde DÖ yöntemi ile oluşturulan sınıflama modellerinin başarısı çoğunlukla daha iyidir. YSA ve NB sınıflama yöntemleri ile elde edilen modellerin performans değerleri diğer sınıflama yöntemlerine göre daha düşüktür. Diğer sınıflama yöntemlerinin de genel olarak performansları iyidir. Kullanılan öznitelik seçim yöntemleri içerisinde sınıflama yöntemlerinin en iyi performansa sahip olduğu öznitelik seçim yöntemleri ise lasso ve limma`dır.

Çalışmada ele alınan mikrodizi gen ifade verilerine ait Tablo 4.1. ile Tablo 4.6. aralığındaki tablolarda yer alan DÖ sınıflama modeli sonuçları, öznitelik seçim yöntemlerinin kullanılması ile oluşturulan DÖ modellerine aittir. DÖ öznitelik seçim işlemini kendi içinde yaptığı için tez çalışmasında kullanılan veri setlerinde öznitelik seçim yöntemi uygulamadan DÖ yönteminin kullanılmasıyla oluşturulan sınıflama modellerinin performansları da elde edilmiştir. DÖ yapısı içerisinde öznitelik seçim işleminin ne şekilde gerçekleştiği başta olmak üzere, gizli tabakalar içerisinde işleyişin tam olarak nasıl olduğu ile ilgili açıklayıcı tam bilgi henüz yoktur. Ayrıca mikrodizi gen ifade verileri gibi büyük veri setleri üzerinde çalışırken hız ve bellek gibi bilgisayar kaynaklı sorunlar ile karşılaşma olasılığı da yüksektir. Dolayısıyla öznitelik seçimi gibi veri setinin boyutunu azaltan işlemlerin yapılmadan analize geçilmesi ayrı bir dezavantaj da olabilir. Bu tez çalışmasında kullanılan her bir veri seti üzerinde öznitelik seçim yöntemlerini uygulamadan ve öznitelik seçim yöntemlerinin uygulanması ile DÖ sınıflama modelleri elde edilmiştir. Böylece veri setleri üzerinde, DÖ açısından iki durum arasında ne kadar farklılık olduğu Tablo 4.7.`de yer alan sonuçlar ile gösterilmeye çalışılmıştır. Tablo incelendiğinde, akciğer kanseri veri setinde DÖ yöntemi ile elde edilen sınıflama modelleri içerisinde limma öznitelik seçim yöntemi uygulandığında en iyi performansta; rfe öznitelik seçim yöntemi uygulandığında ise en düşük performansta DÖ sınıflama modelleri elde edilmiştir. Öznitelik seçim yöntemi uygulamadan oluşturulan DÖ sınıflama modelinin performansı ise rfe öznitelik seçim yönteminin uygulanması ile oluşturulan DÖ sınıflama modelinin performans değerlerinden genel olarak daha iyi iken; varFilter, nsFilter, rf, lasso ve özellikle limma öznitelik seçim yöntemlerinde elde edilen DÖ sınıflama modellerinin performans ölçüsü değerlerine göre daha düşüktür.

**Tablo 4.7.** Gerçek veri setlerinde öznitelik seçim yöntemi uygulamadan ve öznitelik seçim yöntemlerini uygulayarak DÖ yöntemi kullanılması ile oluşturulan sınıflama modellerinin performanslarının karşılaştırılması.

| Veri Seti | Öznitelik Seçim | Doğruluk | Duyarlılık | Seçicilik | EAKA |
|---|---|---|---|---|---|
| Akciğer | - | 0,923 | 1,000 | 0,850 | 0,909 |
| | varFilter | 0,951 | 1,000 | 0,923 | 0,945 |
| | nsFilter | 0,933 | 1,000 | 0,926 | 0,950 |
| | rf | 0,970 | 1,000 | 0,965 | 0,976 |
| | lasso | 0,965 | 0,955 | 0,970 | 0,961 |
| | rfe | 0,733 | 1,000 | 0,700 | 0,753 |
| | limma | 0,986 | 1,000 | 0,975 | 0,988 |
| Lenfoma | - | 0,975 | 0,970 | 0,990 | 0,975 |
| | varFilter | 0,975 | 0,970 | 0,980 | 0,985 |
| | nsFilter | 0,988 | 0,985 | 0,992 | 0,990 |
| | rf | 0,949 | 1,000 | 0,900 | 0,960 |
| | lasso | 1,000 | 1,000 | 1,000 | 1,000 |
| | rfe | 0,900 | 1,000 | 0,850 | 0,950 |
| | limma | 1,000 | 1,000 | 1,000 | 1,000 |
| RahimAğzı | - | 0,870 | 1,000 | 0,850 | 0,865 |
| | varFilter | 0,885 | 0,845 | 0,985 | 0,880 |
| | nsFilter | 0,910 | 0,902 | 0,928 | 0,915 |
| | rf | 0,862 | 1,000 | 0,823 | 0,836 |
| | lasso | 0,925 | 1,000 | 0,923 | 0,952 |
| | rfe | 0,865 | 0,950 | 0,850 | 0,865 |
| | limma | 0,965 | 1,000 | 0,929 | 0,985 |
| Meme | - | 0,815 | 0,980 | 0,810 | 0,830 |
| | varFilter | 0,797 | 0,840 | 0,790 | 0,800 |
| | nsFilter | 0,818 | 0,960 | 0,800 | 0,810 |
| | rf | 0,900 | 1,000 | 0,890 | 0,910 |
| | lasso | 0,909 | 1,000 | 0,850 | 0,933 |
| | rfe | 0,775 | 1,000 | 0,750 | 0,753 |
| | limma | 1,000 | 1,000 | 1,000 | 1,000 |
| Prostat | - | 0,750 | 1,000 | 0,700 | 0,785 |
| | varFilter | 0,845 | 0,810 | 0,990 | 0,858 |
| | nsFilter | 0,900 | 1,000 | 0,870 | 0,888 |
| | rf | 0,835 | 0,850 | 0,820 | 0,840 |
| | lasso | 0,870 | 1,000 | 0,850 | 0,875 |
| | rfe | 0,800 | 1,000 | 0,620 | 0,820 |
| | limma | 0,950 | 1,000 | 0,900 | 0,960 |
| Lösemi | - | 0,932 | 0,990 | 0,875 | 0,910 |
| | varFilter | 0,967 | 0,950 | 1,000 | 0,975 |
| | nsFilter | 0,938 | 0,900 | 1,000 | 0,919 |
| | rf | 0,960 | 1,000 | 0,950 | 0,965 |
| | lasso | 0,985 | 1,000 | 0,980 | 0,985 |
| | rfe | 0,780 | 0,650 | 0,923 | 0,793 |
| | limma | 1,000 | 1,000 | 1,000 | 1,000 |

DÖ modellerine ait sonuçları içeren Tablo 4.7. incelendiğinde, lenfoma veri setinde DÖ modellerine ait sonuçlar incelendiğinde, DÖ yöntemi ile elde edilen sınıflama modelleri içerisinde lasso ve limma öznitelik seçim yöntemleri uygulandığında en iyi performansta; rfe öznitelik seçim yöntemi uygulandığında ise en düşük performansta DÖ sınıflama modelleri elde edilmiştir. Öznitelik seçim yöntemi uygulamadan oluşturulan DÖ sınıflama modelinin performansı ise rf ve rfe öznitelik seçim yöntemlerinin uygulanması ile oluşturulan DÖ sınıflama modellerinin performans değerlerinden genel olarak daha iyi iken; lasso ve limma öznitelik seçim yöntemlerinde elde edilen DÖ sınıflama modellerinin performans değerlerine göre daha düşüktür. varFilter ve nsFilter öznitelik seçim yöntemlerinde elde edilen DÖ sınıflama modellerinin performans değerlerine yakın performansa sahiptir.

Rahim ağzı kanseri veri setinde DÖ modellerine ait sonuçlar incelendiğinde, DÖ yöntemi ile elde edilen sınıflama modelleri içerisinde limma öznitelik seçim yöntemi uygulandığında en iyi performansta; rfe öznitelik seçim yöntemi uygulandığında ise en düşük performansta DÖ sınıflama modelleri elde edilmiştir. Öznitelik seçim yöntemi uygulamadan oluşturulan DÖ sınıflama modelinin performansı ise rf ve rfe öznitelik seçim yöntemlerinin uygulanması ile oluşturulan DÖ sınıflama modellerinin performans ölçüsü değerlerinden genel olarak daha iyi iken; varFilter, nsFilter, lasso ve özellikle limma öznitelik seçim yöntemlerinde elde edilen DÖ sınıflama modellerinin performans ölçüsü değerlerine göre daha düşüktür.

Meme kanseri veri setinde DÖ modellerine ait sonuçlar incelendiğinde, DÖ yöntemi ile elde edilen sınıflama modelleri içerisinde limma öznitelik seçim yöntemi uygulandığında en iyi performansta; rfe öznitelik seçim yöntemi uygulandığında ise en düşük performansta DÖ sınıflama modelleri elde edilmiştir. Öznitelik seçim yöntemi uygulamadan oluşturulan DÖ sınıflama modelinin performansı ise varFilter, nsFilter ve rfe öznitelik seçim yöntemlerinin uygulanması ile oluşturulan DÖ sınıflama modellerinin performans değerlerinden genel olarak daha iyi iken; rf, lasso ve özellikle limma öznitelik seçim yöntemlerinde elde edilen DÖ sınıflama modellerinin performans değerlerine göre daha düşüktür.

Prostat kanseri veri setinde DÖ modellerine ait sonuçlar incelendiğinde, DÖ yöntemi ile elde edilen sınıflama modelleri içerisinde nsFilter ve limma öznitelik

seçim yöntemleri uygulandığında en iyi performansta; rfe öznitelik seçim yöntemi uygulandığında ise en düşük performansta DÖ sınıflama modelleri elde edilmiştir. Genellikle öznitelik seçim yöntemi uygulamadan oluşturulan DÖ sınıflama modelinin performans ölçüsü değerleri ise öznitelik seçim yöntemleri uygulayarak elde edilen DÖ sınıflama modellerinin performans ölçüsü değerlerine göre daha düşüktür.

Lösemi veri setinde DÖ modellerine ait sonuçlar incelendiğinde, DÖ yöntemi ile elde edilen sınıflama modelleri içerisinde limma öznitelik seçim yöntemi uygulandığında en iyi performansta; rfe öznitelik seçim yöntemi uygulandığında ise en düşük performansta DÖ sınıflama modelleri elde edilmiştir. Öznitelik seçim yöntemi uygulamadan oluşturulan DÖ sınıflama modelinin performansı ise rfe öznitelik seçim yönteminin uygulanması ile oluşturulan DÖ sınıflama modelinin performans ölçüsü değerlerinden genel olarak daha iyi iken; varFilter, rf, lasso ve özellikle limma öznitelik seçim yöntemlerinde elde edilen DÖ sınıflama modellerinin performans ölçüsü değerlerine göre daha düşüktür. nsFilter öznitelik seçim yönteminde elde edilen DÖ sınıflama modelinin performans ölçüsü değerleri ile öznitelik seçim yöntemi uygulamadan oluşturulan DÖ sınıflama modelinin performans ölçüsü değerleri birbirine yakındır.

### 4.2. Benzetim Çalışmasına Ait Bulgular

Mikrodizi gen ifade verilerine ait benzetim çalışması ile elde edilen bnz-1, bnz-2, bnz-3 ve bnz-4 veri setleri üzerinde bin tekrar ile varFilter, rf, lasso, rfe ve limma öznitelik seçim yöntemleri ile önemli öznitelikler seçilmiştir. Seçilen öznitelikler ile NB, DVM, kNN, YSA ve DÖ yöntemleriyle oluşturulan sınıflama modellerinin doğruluk, duyarlılık, seçicilik ve EAKA şeklinde performans ölçüleri elde edilmiştir. Her bir veri setine ilişkin elde edilen değerler tablolar ile verilmiştir.

Öznitelik seçim yöntemleri ile seçilen öznitelikler kullanılarak sınıflama yöntemleri ile oluşturulan sınıflama modelleri sayesinde elde edilen tahminler üzerinden model performans ölçüleri hesaplanmıştır. Farklı veri setlerinde, farklı öznitelik seçim yöntemlerinde ve farklı sınıflama yöntemlerinde oluşturulan sınıflama modellerine ait elde edilen performans ölçüsü değerleri karşılaştırılmıştır.

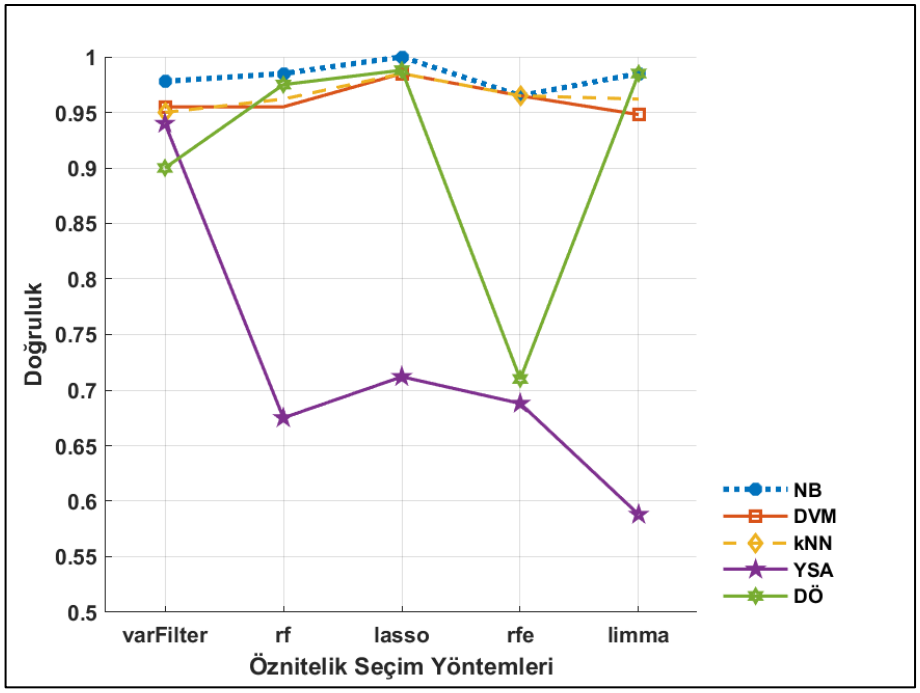
İlk olarak bnz-1 veri setine ait elde edilen sonuçlar Tablo 4.8.`de verilmiştir.

**Tablo 4.8.** Bnz-1 veri setinde öznitelik seçim yöntemleriyle belirlenen öznitelikler kullanılarak oluşturulan sınıflama modellerinin performanslarının karşılaştırılması.

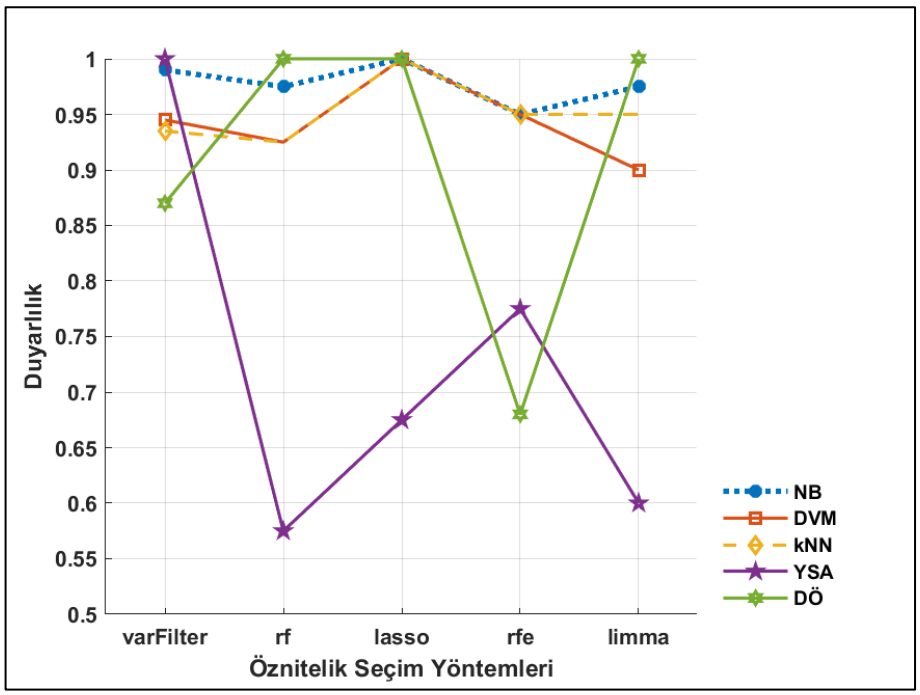
| Öznitelik Seçim Yöntemi | Sınıflama Yöntemi | Doğruluk | Duyarlılık | Seçicilik | EAKA |
|---|---|---|---|---|---|
| varFilter | NB | 0,978 | 0,990 | 0,970 | 0,985 |
| | DVM | 0,955 | 0,945 | 0,970 | 0,960 |
| | kNN | 0,950 | 0,935 | 0,980 | 0,970 |
| | YSA | 0,940 | 1,000 | 0,930 | 0,955 |
| | DÖ | 0,900 | 0,870 | 0,963 | 0,933 |
| rf | NB | 0,985 | 0,975 | 1,000 | 0,990 |
| | DVM | 0,955 | 0,925 | 0,975 | 0,965 |
| | kNN | 0,962 | 0,925 | 1,000 | 0,975 |
| | YSA | 0,675 | 0,575 | 0,775 | 0,650 |
| | DÖ | 0,975 | 1,000 | 0,950 | 0,985 |
| lasso | NB | 1,000 | 1,000 | 1,000 | 1,000 |
| | DVM | 0,985 | 1,000 | 0,975 | 0,990 |
| | kNN | 0,985 | 1,000 | 0,975 | 0,990 |
| | YSA | 0,712 | 0,675 | 0,750 | 0,709 |
| | DÖ | 0,988 | 1,000 | 0,980 | 0,995 |
| rfe | NB | 0,965 | 0,950 | 1,000 | 0,965 |
| | DVM | 0,965 | 0,950 | 1,000 | 0,965 |
| | kNN | 0,965 | 0,950 | 1,000 | 0,965 |
| | YSA | 0,688 | 0,775 | 0,600 | 0,605 |
| | DÖ | 0,710 | 0,680 | 0,750 | 0,700 |
| limma | NB | 0,985 | 0,975 | 1,000 | 0,990 |
| | DVM | 0,948 | 0,900 | 0,975 | 0,955 |
| | kNN | 0,962 | 0,950 | 0,975 | 0,960 |
| | YSA | 0,588 | 0,600 | 0,575 | 0,672 |
| | DÖ | 0,985 | 1,000 | 0,975 | 0,990 |

Bnz-1 veri setine ait sonuçlar incelendiğinde, varFilter hariç öznitelik seçim yöntemlerinde YSA yöntemi ile elde edilen sınıflama modellerinin başarısı diğer modellerin başarısına göre oldukça düşüktür. NB, DVM, kNN ve DÖ yöntemlerinin performans ölçüsü değerleri ise daha yüksektir. Diğer öznitelik seçim yöntemlerinden farklı olarak rfe öznitelik seçim yönteminde, DÖ ile elde edilen sınıflama modelinin performansı YSA yöntemi ile elde edilen sınıflama modelinin performansı gibi düşüktür. Çoğunlukla NB, DVM ve kNN sınıflama yöntemleri birbirine yakın performans ölçüsü değerlerine sahiptir. DÖ sınıflama yöntemi ile elde edilen modellerin diğer modellere göre performans ölçüsü değerlerinin daha iyi olduğu durumlar vardır. lasso ile seçilen öznitelikler kullanılarak sınıflama yöntemleri ile oluşturulan sınıflama modelleri daha iyi performans göstermiştir.

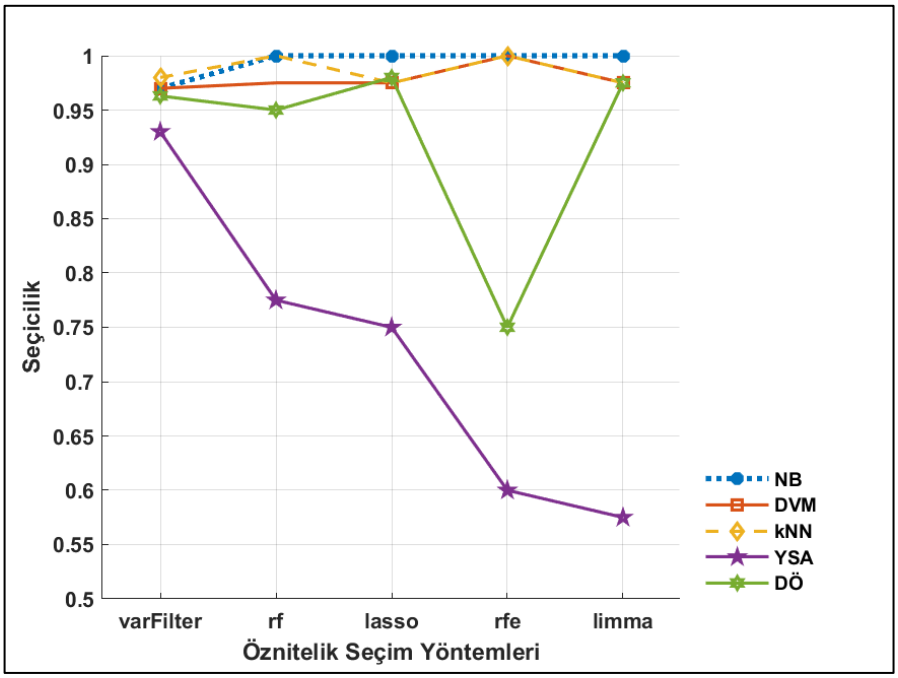
Elde edilen sonuçlar grafikler aracılığıyla da Şekil 4.7.`de verilmiştir.



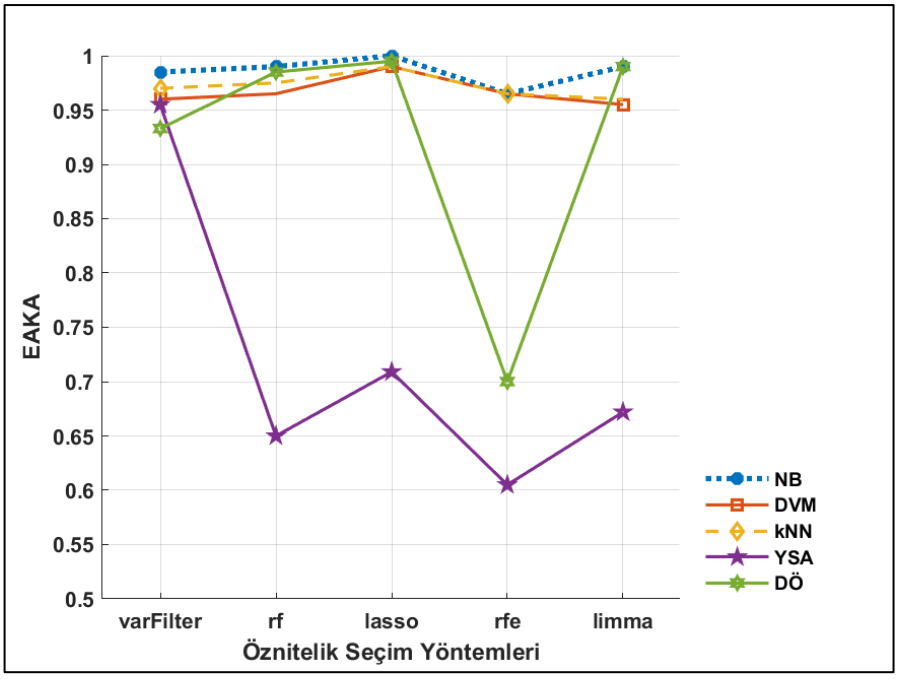

**A**




**B**




**C**




**D**

**Şekil 4.7.** Bnz-1 veri seti için farklı öznitelik seçim yöntemlerinde sınıflama yöntemlerinin doğruluk (A), duyarlılık (B), seçicilik (C) ve EAKA (D) performanslarının karşılaştırılması.

Şekil 4.7. incelendiğinde, rf, lasso, rfe ve limma öznitelik seçim yöntemlerinde oluşturulan YSA sınıflama modelleri düşük performans değerlerine sahiptir. DÖ

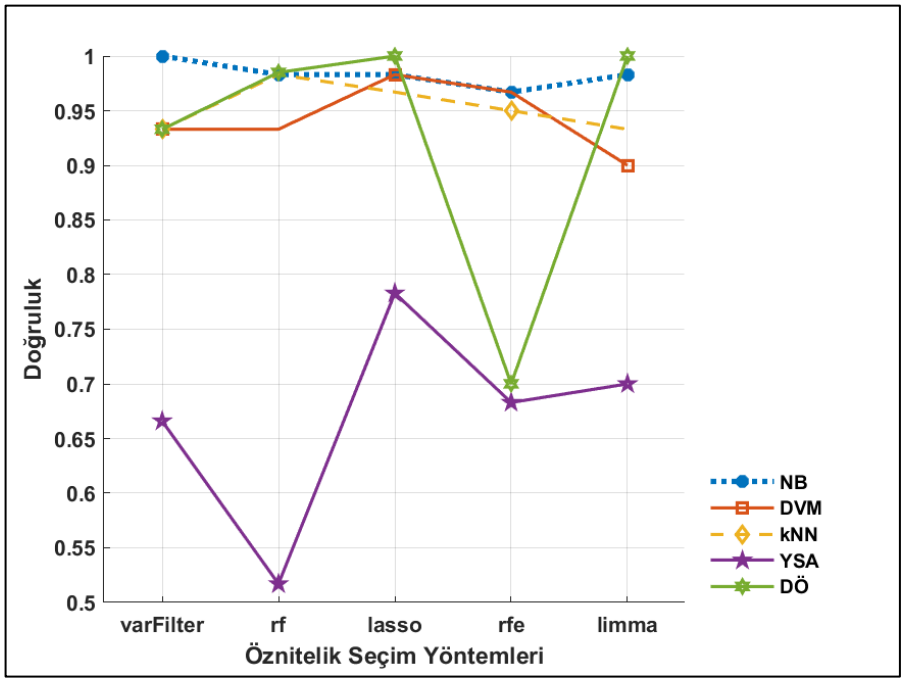
yöntemi ile elde edilen sınıflama modellerinin içerisinde, rfe öznitelik seçim yönteminde oluşturulan DÖ modelinin performansı diğerlerine göre daha düşüktür. Diğer durumlarda oluşturulan sınıflama modellerinin performans ölçüsü değerleri ise birbirine yakındır ve oldukça yüksektir.

**Tablo 4.9.** Bnz-2 veri setinde öznitelik seçim yöntemleriyle belirlenen öznitelikler kullanılarak oluşturulan sınıflama modellerinin performanslarının karşılaştırılması.

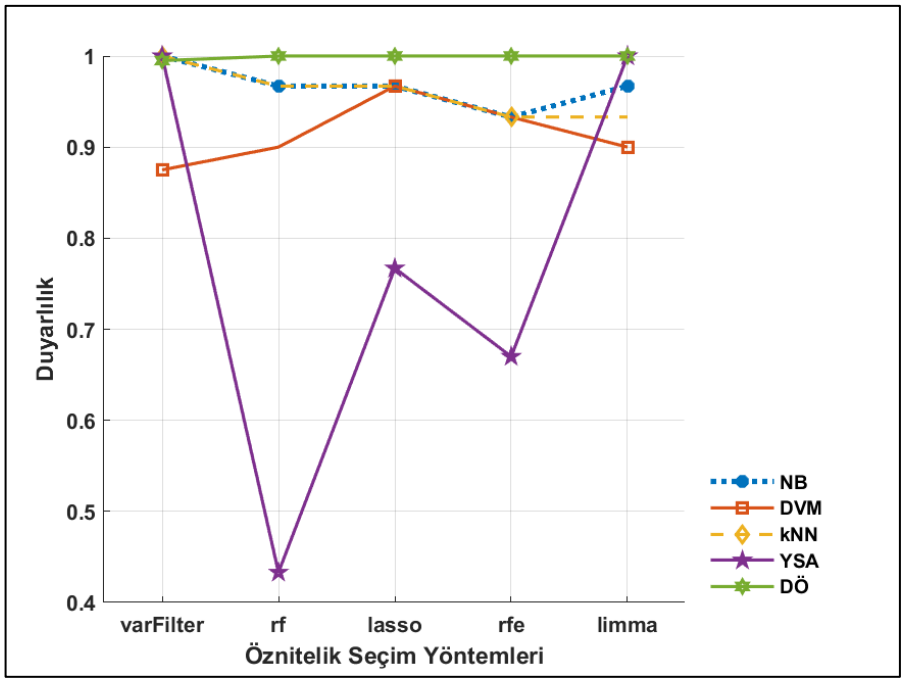
| Öznitelik Seçim Yöntemi | Sınıflama Yöntemi | Doğruluk | Duyarlılık | Seçicilik | EAKA |
|---|---|---|---|---|---|
| varFilter | NB | 1,000 | 1,000 | 1,000 | 1,000 |
| | DVM | 0,933 | 0,875 | 1,000 | 0,940 |
| | kNN | 0,933 | 1,000 | 0,889 | 0,945 |
| | YSA | 0,666 | 1,000 | 0,285 | 0,650 |
| | DÖ | 0,933 | 0,995 | 0,873 | 0,931 |
| rf | NB | 0,983 | 0,967 | 1,000 | 0,980 |
| | DVM | 0,933 | 0,900 | 0,967 | 0,950 |
| | kNN | 0,983 | 0,967 | 1,000 | 0,980 |
| | YSA | 0,517 | 0,433 | 0,600 | 0,550 |
| | DÖ | 0,985 | 1,000 | 0,980 | 0,995 |
| lasso | NB | 0,983 | 0,967 | 1,000 | 0,980 |
| | DVM | 0,983 | 0,967 | 1,000 | 0,980 |
| | kNN | 0,967 | 0,967 | 0,967 | 0,970 |
| | YSA | 0,783 | 0,767 | 0,800 | 0,800 |
| | DÖ | 1,000 | 1,000 | 1,000 | 1,000 |
| rfe | NB | 0,967 | 0,933 | 1,000 | 0,970 |
| | DVM | 0,967 | 0,933 | 1,000 | 0,970 |
| | kNN | 0,950 | 0,933 | 0,967 | 0,960 |
| | YSA | 0,683 | 0,670 | 0,700 | 0,682 |
| | DÖ | 0,820 | 1,000 | 0,700 | 0,820 |
| limma | NB | 0,983 | 0,967 | 1,000 | 0,980 |
| | DVM | 0,900 | 0,900 | 0,900 | 0,890 |
| | kNN | 0,933 | 0,933 | 0,933 | 0,930 |
| | YSA | 0,700 | 1,000 | 0,400 | 0,800 |
| | DÖ | 1,000 | 1,000 | 1,000 | 1,000 |

Bnz-2 veri setine ait sonuçlar incelendiğinde, öznitelik seçim yöntemlerinin her birinde en düşük performans ölçüsü değerleri YSA yöntemi ile elde edilen sınıflama modellerinde elde edilmiştir. rfe öznitelik seçim yönteminde DÖ ile elde edilen sınıflama modelinin performansı diğer DÖ modellerine göre daha düşüktür. Genel olarak NB, DVM ve kNN sınıflama yöntemleri ile elde edilen modeller birbirine yakın ve oldukça iyi performans ölçüsü değerlerine sahiptir. lasso ve limma öznitelik seçim yöntemlerinde DÖ yöntemi ile elde edilen modellerin diğer modellere göre

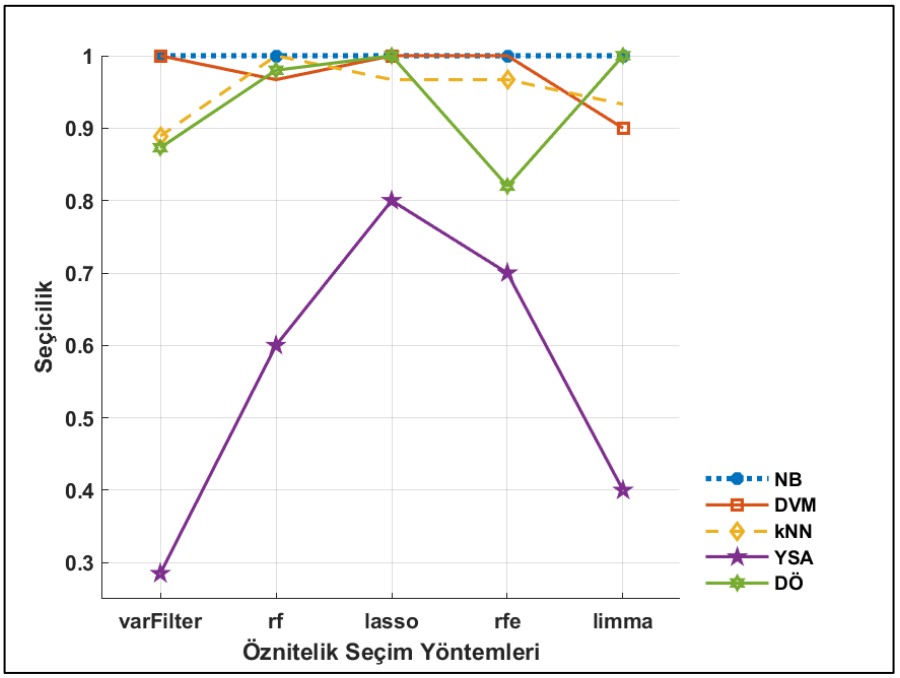
performans ölçüsü değerleri daha yüksektir. varFilter, lasso ve limma ile seçilen öznitelikler kullanılarak sınıflama yöntemleri ile oluşturulan sınıflama modelleri daha iyi performans göstermişlerdir. Elde edilen sonuçlar grafikler aracılığıyla da Şekil 4.8.`de verilmiştir.



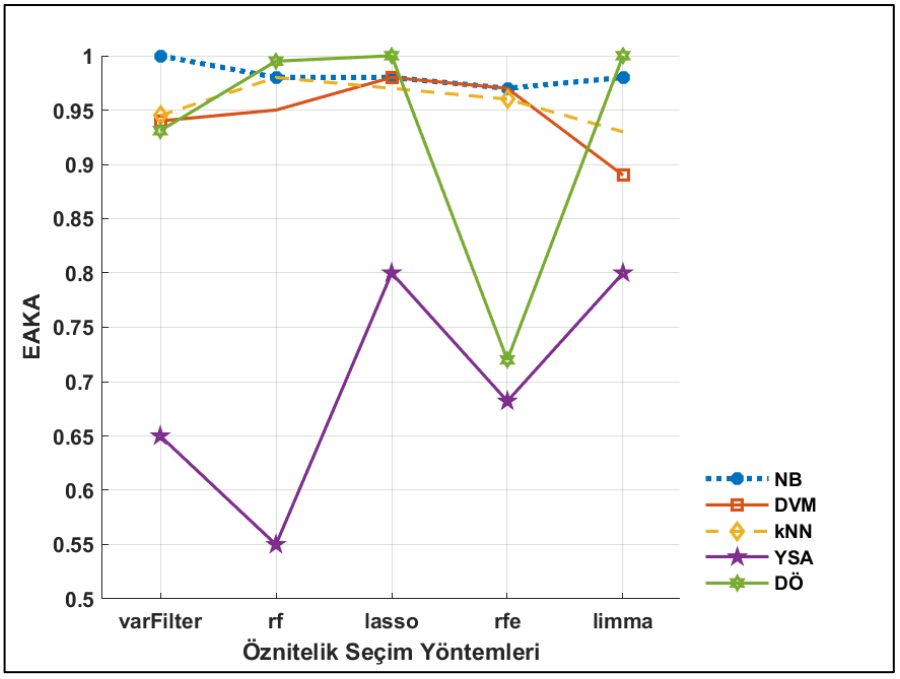

**A**




**B**




**C**




**D**

**Şekil 4.8.** Bnz-2 veri seti için farklı öznitelik seçim yöntemlerinde sınıflama yöntemlerinin doğruluk (A), duyarlılık (B), seçicilik (C) ve EAKA (D) performanslarının karşılaştırılması.

Şekil 4.8. incelendiğinde, genel olarak öznitelik seçim yöntemlerinde oluşturulan YSA sınıflama modelleri düşük performans ölçüsü değerlerine sahiptir. NB, DVM

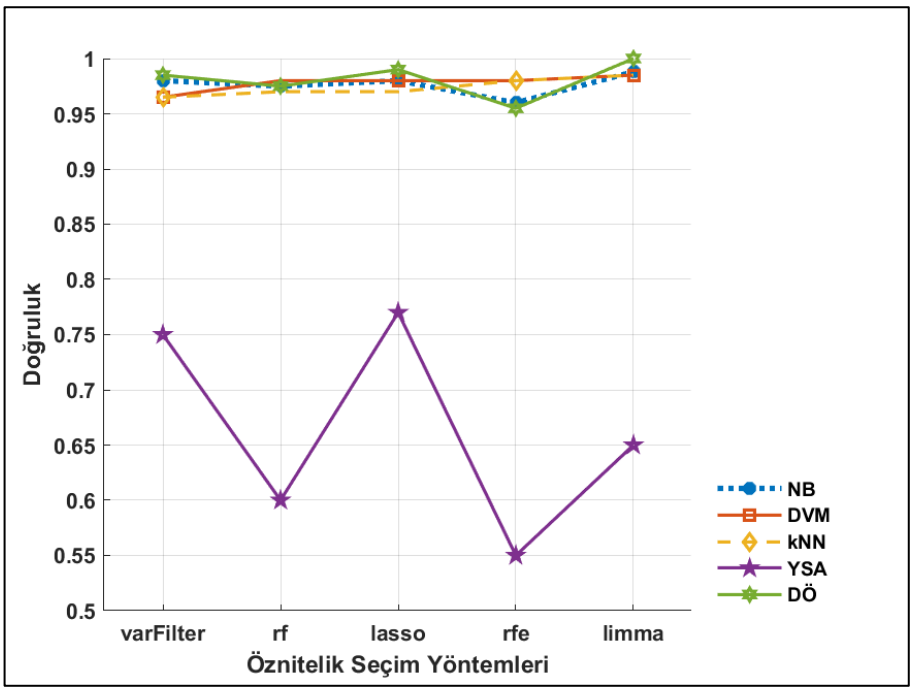
ve kNN sınıflama yöntemlerinin performans ölçüsü değerleri birbirine yakındır ve yaklaşık olarak %80`in üzerindedir. DÖ yöntemi ile elde edilen sınıflama modellerinin içerisinde ise rfe öznitelik seçim yönteminde oluşturulan DÖ modelinin performansı diğerlerine göre oldukça düşüktür.

**Tablo 4.10.** Bnz-3 veri setinde öznitelik seçim yöntemleriyle belirlenen öznitelikler kullanılarak oluşturulan sınıflama modellerinin performanslarının karşılaştırılması.

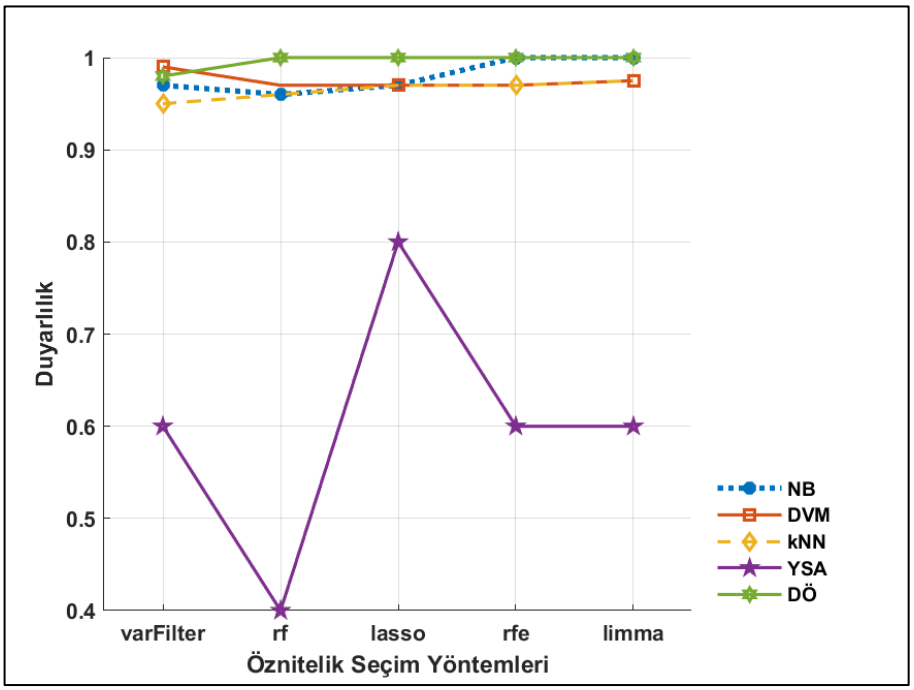
| Öznitelik Seçim Yöntemi | Sınıflama Yöntemi | Doğruluk | Duyarlılık | Seçicilik | EAKA |
|---|---|---|---|---|---|
| varFilter | NB | 0,980 | 0,970 | 0,990 | 0,980 |
| | DVM | 0,965 | 0,990 | 0,960 | 0,970 |
| | kNN | 0,965 | 0,950 | 1,000 | 0,965 |
| | YSA | 0,750 | 0,600 | 0,800 | 0,800 |
| | DÖ | 0,985 | 0,980 | 0,990 | 0,990 |
| rf | NB | 0,975 | 0,960 | 0,990 | 0,975 |
| | DVM | 0,980 | 0,970 | 1,000 | 0,985 |
| | kNN | 0,970 | 0,960 | 0,980 | 0,965 |
| | YSA | 0,600 | 0,400 | 0,800 | 0,575 |
| | DÖ | 0,975 | 1,000 | 0,875 | 0,978 |
| lasso | NB | 0,980 | 0,970 | 1,000 | 0,985 |
| | DVM | 0,980 | 0,970 | 1,000 | 0,985 |
| | kNN | 0,970 | 0,970 | 0,970 | 0,970 |
| | YSA | 0,770 | 0,800 | 0,750 | 0,750 |
| | DÖ | 0,990 | 1,000 | 0,990 | 0,998 |
| rfe | NB | 0,960 | 1,000 | 0,920 | 0,970 |
| | DVM | 0,980 | 0,970 | 1,000 | 0,985 |
| | kNN | 0,980 | 0,970 | 1,000 | 0,985 |
| | YSA | 0,550 | 0,600 | 0,500 | 0,555 |
| | DÖ | 0,955 | 1,000 | 0,910 | 0,965 |
| limma | NB | 0,988 | 1,000 | 0,980 | 0,995 |
| | DVM | 0,985 | 0,975 | 1,000 | 0,990 |
| | kNN | 0,985 | 0,975 | 1,000 | 0,990 |
| | YSA | 0,650 | 0,600 | 0,670 | 0,660 |
| | DÖ | 1,000 | 1,000 | 1,000 | 1,000 |

Bnz-3 veri setine ait sonuçlar incelendiğinde, genel olarak YSA sınıflama yöntemi ile oluşturulan modellerin performans ölçüsü değerleri en düşüktür. NB, DVM ve kNN sınıflama yöntemleri ile elde edilen modellerin performans düzeyleri birbirine yakındır ve %90`ın üzerindedir. Bnz-1 ve bnz-2 veri setlerinden farklı olarak bnz-3 verisinde rfe öznitelik seçim yöntemi ile DÖ sınıflama yönteminin kullanılmasıyla elde edilen modelin performans ölçüsü değerleri daha iyi çıkmıştır. varFilter, lasso ve limma öznitelik seçim yöntemleri uygulandıktan sonra elde edilen sınıflama

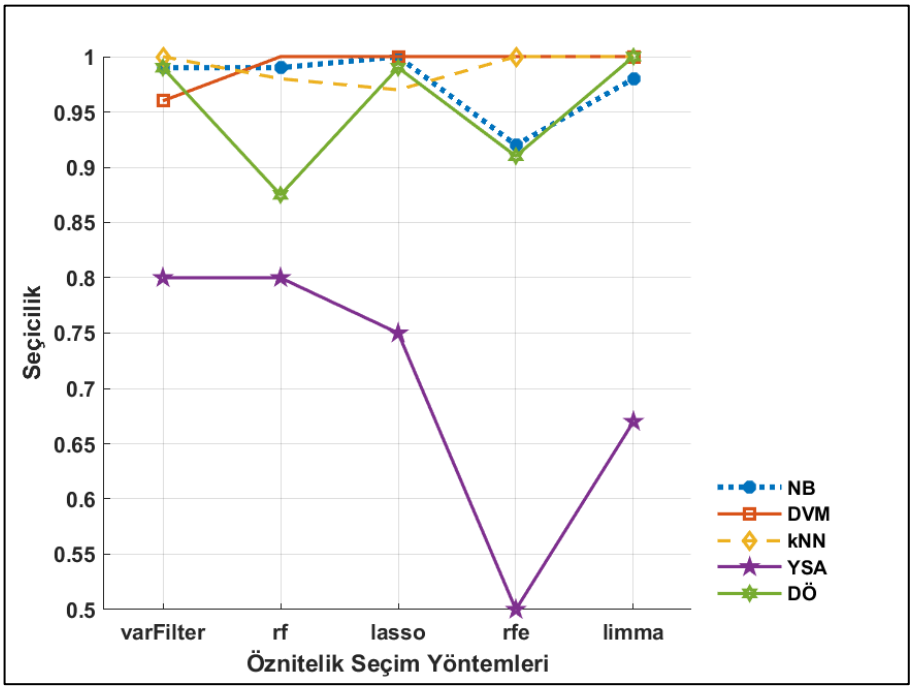
modelleri içerisinde DÖ yöntemi en iyi performansı göstermiştir. Elde edilen sonuçlar grafikler aracılığıyla da Şekil 4.9.`da verilmiştir.



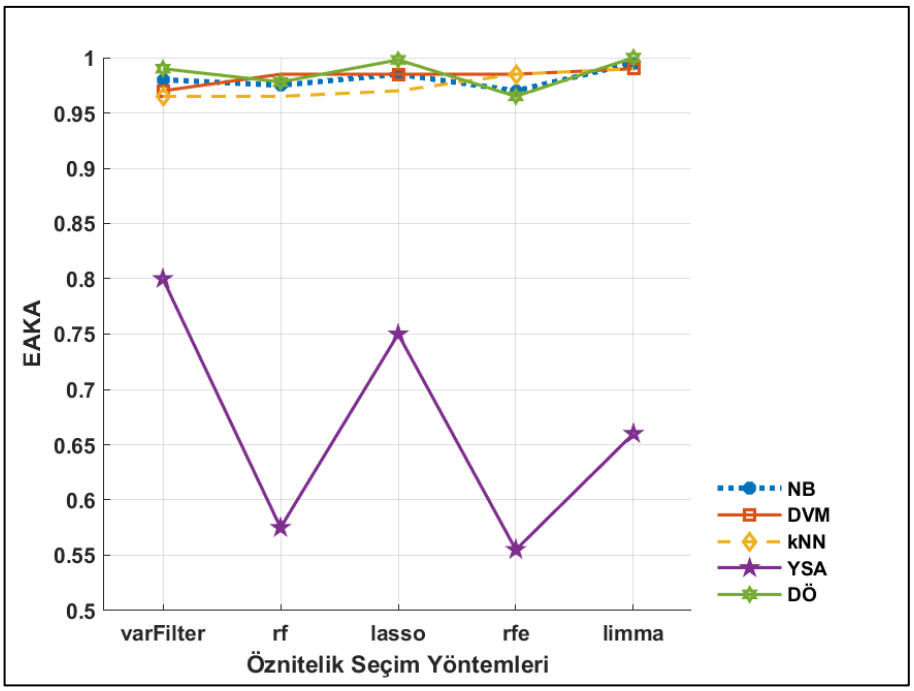

**A**




**B**




**C**




**D**

**Şekil 4.9.** Bnz-3 veri seti için farklı öznitelik seçim yöntemlerinde sınıflama yöntemlerinin doğruluk (A), duyarlılık (B), seçicilik (C) ve EAKA (D) performanslarının karşılaştırılması.

Şekil 4.9. incelendiğinde, çoğunlukla NB, DVM ve kNN sınıflama yöntemlerinin performans ölçüsü değerleri birbirine yakındır ve DÖ`den sonra geldikleri durumlar vardır. rf ve rfe öznitelik seçim yöntemlerini uygulayarak YSA ve DÖ sınıflama yöntemlerinin kullanılması ile elde edilen modellerin performans ölçüsü değerleri diğer öznitelik seçim yöntemlerinde oluşturulan modellerin değerlerine göre daha düşüktür. Öznitelik seçim yöntemlerinin her birinde YSA sınıflama yöntemi ile elde edilen modellerin performans ölçüsü değerleri diğer sınıflama yöntemlerininki kadar yüksek değildir.

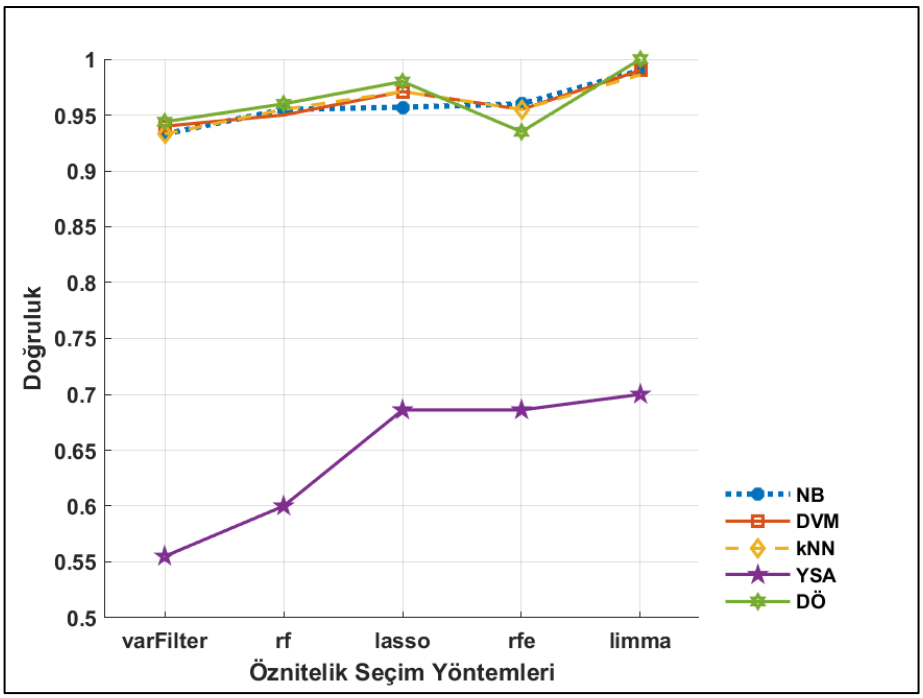
**Tablo 4.11.** Bnz-4 veri setinde öznitelik seçim yöntemleriyle belirlenen öznitelikler kullanılarak oluşturulan sınıflama modellerinin performanslarının karşılaştırılması.

| Öznitelik Seçim Yöntemi | Sınıflama Yöntemi | Doğruluk | Duyarlılık | Seçicilik | EAKA |
|---|---|---|---|---|---|
| varFilter | NB | 0,933 | 0,875 | 0,995 | 0,940 |
| | DVM | 0,940 | 0,880 | 1,000 | 0,950 |
| | kNN | 0,933 | 0,875 | 0,995 | 0,940 |
| | YSA | 0,555 | 0,600 | 0,520 | 0,550 |
| | DÖ | 0,944 | 0,920 | 0,972 | 0,950 |
| rf | NB | 0,955 | 0,930 | 1,000 | 0,960 |
| | DVM | 0,950 | 0,915 | 0,985 | 0,950 |
| | kNN | 0,955 | 0,930 | 1,000 | 0,960 |
| | YSA | 0,600 | 0,610 | 0,560 | 0,576 |
| | DÖ | 0,960 | 1,000 | 0,920 | 0,970 |
| lasso | NB | 0,957 | 0,914 | 1,000 | 0,960 |
| | DVM | 0,971 | 0,943 | 1,000 | 0,975 |
| | kNN | 0,971 | 0,943 | 1,000 | 0,975 |
| | YSA | 0,686 | 0,571 | 0,800 | 0,700 |
| | DÖ | 0,980 | 0,960 | 1,000 | 0,985 |
| rfe | NB | 0,960 | 1,000 | 0,920 | 0,970 |
| | DVM | 0,955 | 0,940 | 1,000 | 0,960 |
| | kNN | 0,955 | 0,940 | 1,000 | 0,960 |
| | YSA | 0,686 | 0,400 | 0,971 | 0,665 |
| | DÖ | 0,935 | 1,000 | 0,900 | 0,945 |
| limma | NB | 0,990 | 1,000 | 0,980 | 0,995 |
| | DVM | 0,990 | 1,000 | 0,980 | 0,995 |
| | kNN | 0,986 | 0,971 | 1,000 | 0,990 |
| | YSA | 0,700 | 0,800 | 0,600 | 0,710 |
| | DÖ | 1,000 | 1,000 | 1,000 | 1,000 |

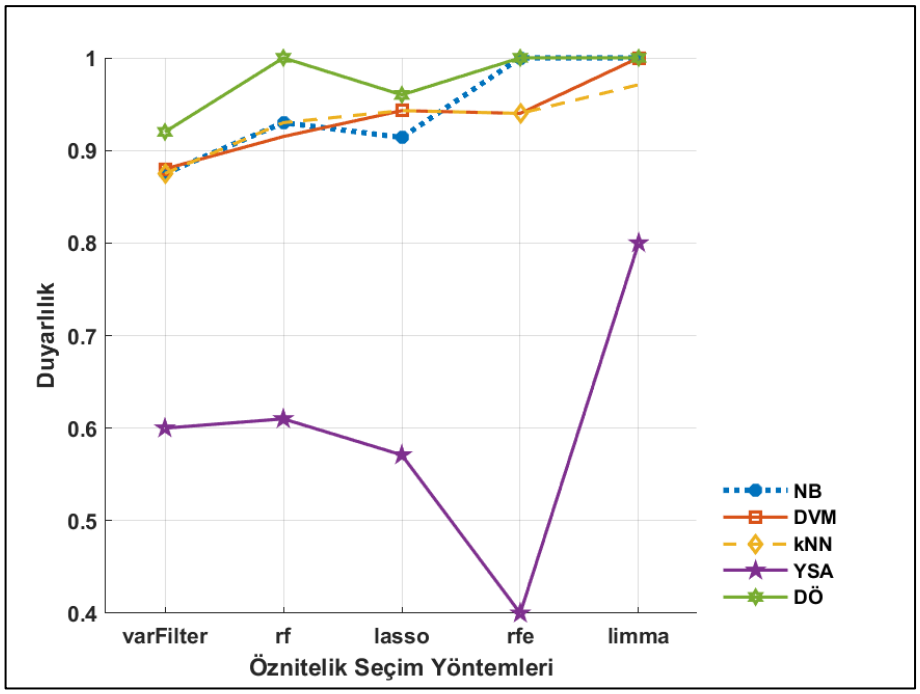
Tablo 4.11.`de verilen bnz-4 veri setine ait sonuçlar incelendiğinde, bnz-3 verisinde olduğu gibi bnz-4 verisinde de rfe öznitelik seçim yöntemi ile DÖ sınıflama

yönteminin kullanılmasıyla elde edilen modelin performans ölçüsü değerleri bnz-1 ve bnz-2 verisinde elde edilen DÖ sınıflama modeline göre daha iyi çıkmıştır. Öznitelik seçim yöntemleri içerisinde özellikle limma ve lasso öznitelik seçim yöntemleri uygulandıktan sonra sınıflama yöntemlerinin kullanılması ile elde edilen modellerin performans ölçüsü değerleri oldukça yüksektir. Genel olarak YSA sınıflama yöntemi ile oluşturulan modellerin performans ölçüsü değerleri diğer modellere göre oldukça düşüktür. NB, DVM ve kNN sınıflama yöntemleri ile elde edilen modeller ise birbirine yakın performans ölçüsü değerlerine sahiptir.

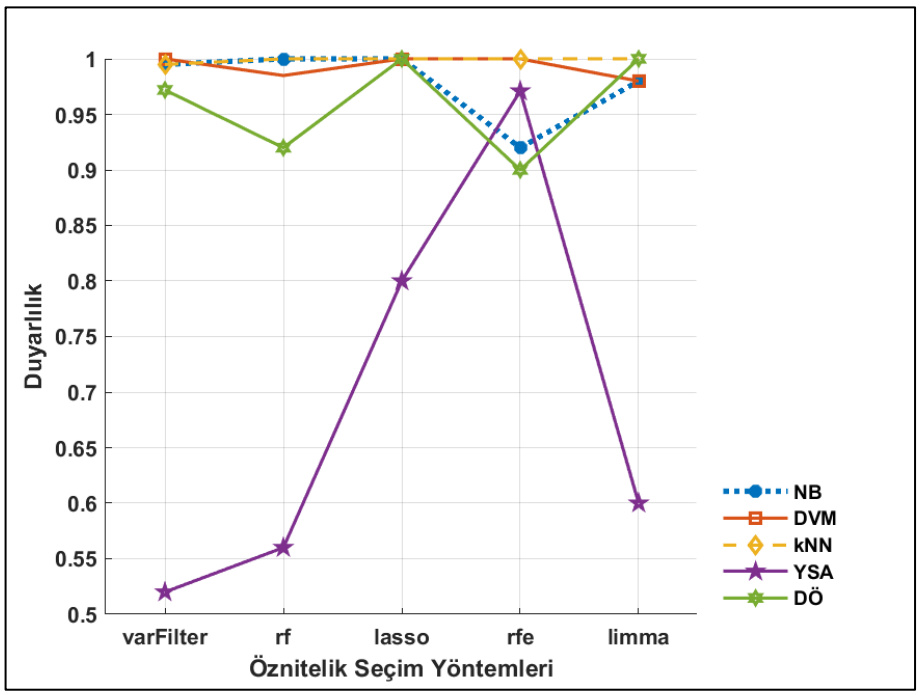
Son olarak bnz-4 veri setine ait elde edilen sonuçlar grafikler aracılığıyla da Şekil 4.10.`da verilmiştir. Şekil incelendiğinde, öznitelik seçim yöntemlerinin her birinde genel olarak DÖ sınıflama yöntemi ile elde edilen modellerin başarısı daha iyi iken; YSA sınıflama yöntemi ile elde edilen modellerin başarısı daha düşüktür. limma ve lasso öznitelik seçim yöntemlerini uygulayarak sınıflama yöntemlerinin kullanılması ile elde edilen sınıflama modellerinin performans ölçüsü değerleri diğer öznitelik seçim yöntemlerinde oluşturulan sınıflama modellerinin performans ölçüsü değerlerine göre daha yüksektir.



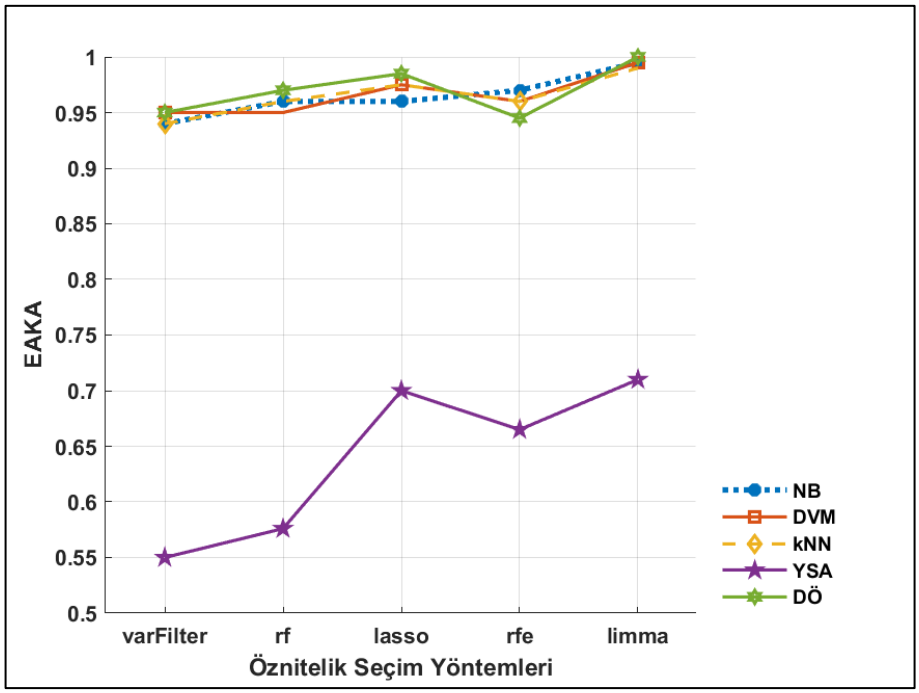

**A**


Duyarlılık
1
0.9
0.8
0.7
0.6
0.5
0.4
varFilter
rf
lasso
rfe
limma
Öznitelik Seçim Yöntemleri
NB
DVM
kNN
YSA
DÖ

**B**


Duyarlılık
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
varFilter
rf
lasso
rfe
limma
Öznitelik Seçim Yöntemleri
NB
DVM
kNN
YSA
DÖ

**C**




**D**

**Şekil 4.10.** Bnz-4 veri seti için farklı öznitelik seçim yöntemlerinde sınıflama yöntemlerinin doğruluk (A), duyarlılık (B), seçicilik (C) ve EAKA (D) performanslarının karşılaştırılması.

YSA`nın genişletilmiş biçimi olan DÖ yapısında çok sayıda gizli tabaka olduğu için öznitelik seçim işleminin nasıl yapıldığının tam olarak açıklanamaması ve mikrodizi gen ifade verileri gibi büyük veri setleri üzerinde çalışırken bilgisayar kaynaklı sorunlar ile karşılaşma olasılığının da yüksek olması gibi durumlar, öznitelik seçim yöntemi uygulamadan DÖ yöntemini uygulamanın dezavantajları sayılabilir.

Sık kullanılan klasik veri madenciliği sınıflama yöntemleri ile birlikte giderek daha çok tercih edilen DÖ yönteminin de yer aldığı bu tez çalışmasında, her bir veri seti üzerinde öznitelik seçim yöntemlerinin uygulanması ile DÖ sınıflama modelleri ve öznitelik seçim yöntemlerini uygulamadan DÖ sınıflama modelleri elde edilmiştir. Böylece veri setleri üzerinde, DÖ açısından iki durum arasında ne kadar farklılık olduğu gösterilmeye çalışılmıştır.

Elde edilen sonuçlar Tablo 4.12.`de yer almaktadır. Tabloda yer alan - işareti öznitelik seçiminin uygulanmadığı durumu ifade etmektedir.

**Tablo 4.12.** Benzetim çalışmasından elde edilmiş veri setlerinde öznitelik seçim yöntemi uygulamadan ve öznitelik seçim yöntemlerini uygulayarak DÖ yöntemi kullanılması ile oluşturulan sınıflama modellerinin performanslarının karşılaştırılması.

| Veri Setleri | Öznitelik Seçim Yöntemi | Doğruluk | Duyarlılık | Seçicilik | EAKA |
|---|---|---|---|---|---|
| Bnz-1 | - | 0,786 | 0,750 | 0,976 | 0,827 |
| | varFilter | 0,900 | 0,870 | 0,963 | 0,933 |
| | rf | 0,975 | 1,000 | 0,950 | 0,985 |
| | lasso | 0,988 | 1,000 | 0,980 | 0,995 |
| | rfe | 0,710 | 0,680 | 0,750 | 0,700 |
| | limma | 0,985 | 1,000 | 0,975 | 0,990 |
| Bnz-2 | - | 0,736 | 0,707 | 0,927 | 0,750 |
| | varFilter | 0,933 | 0,995 | 0,873 | 0,931 |
| | rf | 0,985 | 1,000 | 0,980 | 0,995 |
| | lasso | 1,000 | 1,000 | 1,000 | 1,000 |
| | rfe | 0,820 | 1,000 | 0,700 | 0,820 |
| | limma | 1,000 | 1,000 | 1,000 | 1,000 |
| Bnz-3 | - | 0,845 | 0,815 | 0,985 | 0,855 |
| | varFilter | 0,985 | 0,980 | 0,990 | 0,990 |
| | rf | 0,975 | 1,000 | 0,875 | 0,978 |
| | lasso | 0,990 | 1,000 | 0,990 | 0,998 |
| | rfe | 0,955 | 1,000 | 0,910 | 0,965 |
| | limma | 1,000 | 1,000 | 1,000 | 1,000 |
| Bnz-4 | - | 0,889 | 0,808 | 0,977 | 0,862 |
| | varFilter | 0,944 | 0,920 | 0,972 | 0,950 |
| | rf | 0,960 | 1,000 | 0,920 | 0,970 |
| | lasso | 0,980 | 0,960 | 1,000 | 0,985 |
| | rfe | 0,935 | 1,000 | 0,900 | 0,945 |
| | limma | 1,000 | 1,000 | 1,000 | 1,000 |

DÖ modellerine ait sonuçları içeren Tablo 4.12. incelendiğinde, bnz-1 veri setinde DÖ yöntemi ile elde edilen sınıflama modelleri içerisinde lasso öznitelik seçim yöntemi uygulandığında en iyi performansta DÖ sınıflama modeli; rfe öznitelik seçim yöntemi uygulandığında ise en düşük performansta DÖ sınıflama modeli elde edilmiştir. Öznitelik seçim yöntemi uygulamadan oluşturulan DÖ sınıflama modelinin performans ölçüsü değerleri rfe öznitelik seçim yönteminin uygulanması ile oluşturulan DÖ sınıflama modelinin performans değerlerinden daha iyidir. varFilter, rf, lasso ve limma öznitelik seçim yöntemlerinde elde edilen DÖ sınıflama modellerinin performans ölçüsü değerlerine göre öznitelik seçim yöntemi uygulamadan oluşturulan DÖ sınıflama modelinin performans ölçüsü değerleri daha düşüktür.

Bnz-2 veri setinde DÖ modellerine ait sonuçlar incelendiğinde, limma ve özellikle lasso öznitelik seçim yöntemleri uygulandığında en iyi performansta DÖ sınıflama modelleri; rfe öznitelik seçim yöntemi uygulandığında ise en düşük performansta DÖ sınıflama modeli elde edilmiştir. Öznitelik seçim yöntemi uygulamadan oluşturulan DÖ sınıflama modelinin performansı ise rfe öznitelik seçim yönteminin uygulanması ile oluşturulan DÖ sınıflama modelinin performans değerlerine daha yakın iken; diğer öznitelik seçim yöntemlerinde elde edilen DÖ sınıflama modellerinin performans değerlerine göre daha düşüktür.

Bnz-3 veri setinde DÖ modellerine ait sonuçlar incelendiğinde, lasso ve özellikle limma öznitelik seçim yöntemleri uygulandığında en iyi performansta DÖ sınıflama modelleri elde edilmiştir. Öznitelik seçim yöntemi uygulamadan oluşturulan DÖ sınıflama modelinin performans ölçüsü değerleri öznitelik seçim yöntemlerinin uygulanması ile oluşturulan DÖ sınıflama modellerinin performans ölçüsü değerlerinden daha düşüktür. Diğer koşullarda ise elde edilen performans değerleri birbirine yakındır.

Bnz-4 veri setinde DÖ modellerine ait sonuçlar incelendiğinde, bnz-3 veri setine benzer sonuçlar elde edilmiştir. Özellikle limma ve lasso öznitelik seçim yöntemleri uygulandığında en iyi performansta DÖ sınıflama modelleri elde edilmiştir. Öznitelik seçim yöntemi uygulamadan oluşturulan DÖ sınıflama modelinin performans ölçüsü değerleri öznitelik seçim yöntemlerinin uygulanması ile oluşturulan DÖ sınıflama modellerinin performans ölçü değerlerine göre düşüktür.

Bnz-1 ve bnz-2 veri setlerinde rfe öznitelik seçim yönteminin kullanılması ile oluşturulan DÖ sınıflama modelinin performans ölçü değerleri oldukça düşüktür. Benzetim çalışması ile elde edilen dört veri setinde de lasso ve limma öznitelik seçim yöntemlerinde elde edilen DÖ sınıflama modellerinin performans ölçü değerleri diğer durumlara göre daha yüksektir. Genel olarak veri setlerinde öznitelik seçim yöntemi kullanmadan DÖ yönteminin uygulanması ile oluşturulan sınıflama modelinin performans ölçü değerleri öznitelik seçim yöntemlerinin kullanılması ile DÖ yönteminin uygulanmasıyla elde edilen sınıflama modellerine göre daha düşük çıkmıştır.

# 5. TARTIŞMA

Bu çalışmada altısı gerçek dördü benzetim çalışması ile elde edilen toplam on farklı mikrodizi gen ifade verisi kullanılmıştır. Akciğer, lenfoma, rahim ağzı, meme, prostat ve lösemi kanser türlerine ait mikrodizi gen ifade verileri gerçek veri setlerini, bnz-1, bnz-2, bnz-3 ve bnz-4 veri setleri de benzetim çalışmasından elde edilen veri setlerini oluşturmaktadır. Veri setlerinin her biri, hasta ve sağlıklı olmak üzere iki sınıf içermektedir. Ayrıca sınıflardaki örnek sayıları bakımından da dengeli bir dağılıma sahiptir. Akciğer ve rahim ağzı kanseri veri setlerinde 54675 öznitelik vardır. Meme, prostat kanseri, lenfoma ve lösemi veri setlerinde ise 22283 öznitelik bulunmaktadır. Bnz-1 veri setinde iki yüz elli, bnz-2 veri setinde ise beş yüz, bnz-3 veri setinde yedi yüz elli, bnz-4 veri setinde ise bin öznitelik vardır. Öznitelik sayısını azaltmak için altı farklı öznitelik seçim yöntemi kullanılmıştır. Bunlar; varFilter, nsFilter, rf, lasso, rfe ve limma`dır. Sadece nsFilter yöntemi benzetim çalışmasından elde edilen veri setlerinde uygulanmamıştır. Veri setleri üzerinde seçilen önemli ve faydalı öznitelikler kullanılarak NB, DVM, kNN, YSA ve DÖ şeklinde beş farklı sınıflama yöntemi aracılığıyla sınıflama modelleri oluşturulmuştur. Sınıflama yöntemleri uygulanırken verilerin %75`i eğitim, %25`i test için kullanılmıştır ve 5-kat çapraz geçerlilik yapılmıştır. Elde edilen modellerin hasta-sağlıklı sınıflamasını ne kadar doğru yaptığını gösteren doğruluk, duyarlılık, seçicilik ve EAKA olmak üzere dört farklı model performans ölçüsü hesaplanmştır. Hesaplanan bu değerler ile her bir veri seti için kullanılan öznitelik seçim yöntemlerinin sınıflama yöntemleri başarısına etkisi incelenmiştir, yöntemler arasında karşılaştırmalar yapılmıştır.

Mikrodizi gen ifade verilerinde ele alınan öznitelik seçim yöntemlerinin kullanılan sınıflama yöntemlerinin performansına etkisi incelendiğinde; genel olarak lasso ve limma öznitelik seçim yöntemlerinin kullanılması ile elde edilen veri setleri üzerinde sınıflama yöntemlerinin uygulanması ile oluşturulan sınıflama modellerinin performans ölçü değerleri daha yüksek hesaplanmıştır. Sınıflama modelleri içerisinde de çoğunlukla DÖ sınıflama yöntemi ile oluşturulan modellerin performansı daha iyidir. Fakat öznitelik seçim yöntemleri içerisinde rfe yönteminin kullanılması ile oluşturulan DÖ sınıflama modellerinin performansları daha düşük çıkmıştır. NB, DVM ve kNN yöntemleri kullanılarak elde edilen sınıflama modellerinin

performansları ise birbirine oldukça yakındır ve DÖ`den sonra gelmektedir. Düşük performans ölçüsü değerlerine sahip modeller çoğunlukla sınıflama yöntemi YSA kullanıldığında elde edilmiştir. Bazı durumlarda ise sırasıyla NB, DVM ve kNN sınıflama yöntemleri ile elde edilen modellerin başarısı daha düşük çıkmıştır.

Tez kapsamında mikrodizi gen ifade verileri ile ilgili yapılan benzetim çalışmasında, mikrodizi gen ifade verilerinin genel özelliğini yansıtacak şekilde örnek sayısı az, öznitelik sayısı daha fazla olacak şekilde veriler türetilerek öznitelik seçim ve sınıflama yöntemlerinin performansları incelenmiştir. Veri setlerinde örnek sayıları aynıdır fakat öznitelik sayısı artmaktadır. Genellikle, en iyi sınıflama performansına DÖ sınıflama yöntemi ile ulaşılmıştır. Özellikle bnz-1 ve bnz-2 veri setlerinde rfe öznitelik seçim yönteminin kullanılması ile oluşturulan DÖ sınıflama modellerinin performansları daha düşük çıkmıştır. kNN, DVM ve NB sınıflama yöntemlerinin performans ölçüsü değerleri ise birbirine yakındır ve DÖ yönteminden sonra gelmektedir. YSA sınıflama yöntemi ile performans ölçüsü değerleri daha düşük olan modeller elde edilmiştir. lasso ve limma öznitelik seçim yöntemleri kullanıldığında sınıflama yöntemlerinin uygulanması ile elde edilen modellerin performans ölçüsü değerleri daha iyidir. Kısacası, gerçek veri setleri sonuçlarına benzer sonuçlar elde edilmiştir.

DÖ yönteminin uygulanması ile ilgili, öznitelik seçim işleminin yapılmasına gerek olmadığı ile ilgili var olan bilgi doğrultusunda, veri setleri üzerinde öznitelik seçimi yapılmadan DÖ sınıflama yöntemini uygulayarak da DÖ sınıflama modelleri elde edilmiştir. Çalışmanın amacı öznitelik seçim yöntemlerinin sınıflama yöntemleri başarısına etkisini göstermek olduğu için yöntemler arasında karşılaştırmaların yapılması için belli bir standartın olması açısından kullanılan sınıflama yöntemlerinin hepsi aynı işlem basamaklarından geçirilmiştir. Böylece öznitelik seçim yöntemlerinin uygulanması ile ve öznitelik seçim yöntemini uygulamadan DÖ sınıflama modelleri oluşturulmuştur. Öznitelik seçimi uygulamadan elde edilen DÖ sınıflama modellerinin performansları özellikle rf, lasso ve limma öznitelik seçim yöntemlerinin uygulanması ile oluşturulan DÖ sınıflama modellerinin performanslarından daha düşük iken; rfe öznitelik seçim yöntemi uygulandığında elde edilen DÖ sınıflama modellerinin performanslarından daha iyidir. Dolayısıyla çalışmada ele alınan veri setleri ve öznitelik seçim

yöntemleri kapsamında rfe hariç diğer öznitelik seçim yöntemlerinin kullanılarak DÖ sınıflama modellerinin elde edilmesinin performansı artırdığı söylenebilir. Mikrodizi gen ifade verileri üzerinde DÖ yönteminin bu şekilde bir uygulamasına daha önceki çalışmalarda rastlanmamıştır.

Literatüre bakıldığında kanser verileri üzerinde öznitelik seçiminin yapılarak sınıflama yöntemlerinin uygulanması ile elde edilen sınıflama modellerinin performans ölçüleri aracılığıyla karşılaştırmalarının yapıldığı çalışmalar vardır. Ancak üzerinde çalışılan veri setleri, uygulanan öznitelik seçim ile sınıflama yöntemleri ve kullanılan performans ölçüleri her birinde farklıdır. Ayrıca DÖ yönteminin hem mikrodizi gen ifade verileri üzerinde hem de sınıflama amaçlı uygulamasının bulunduğu çalışmaya literatürde rastlanmamış olup yok denecek kadar azdır.

Tez çalışmasının konusunu oluşturan biyoinformatik, veri madenciliğinde sınıflama yöntemleri, öznitelik seçim yöntemleri ve model performans ölçülerinin bir arada yer aldığı son yıllarda yapılan çalışmalardan bazıları incelendiğinde ise; Demircioğlu ve ark., 253 örnek ve 15154 tane genin bulunduğu halka açık yumurtalık kanseri gen veri kümesini kullanmışlardır. Veri setinde 253 tane örneğin 91 tanesi sağlıklı, 162 tanesi hastadır. Sınıflama yapılırken %40'ı eğitim, %60'ı test verisi olacak şekilde rastgele örnekler seçilmiştir. Veri kümesinde gen sayısı fazla olduğu için, Welch t testi ve Fisher korelasyon skorlama olmak üzere iki farklı öznitelik seçim yöntemi ile kNN ve DVM sınıflama yöntemlerini kullanarak modellerin doğruluk oranlarını karşılaştırmışlardır. %100 gibi yüksek başarı oranlarına ulaşılmıştır (8). Devi ve ark., lenfoma ve kolon kanseri veri setleri üzerinde çalışmışlardır. Karşılıklı bilgi tabanlı öznitelik seçimi ile seçilen öznitelikleri kullanarak kNN, DVM ve YSA sınıflama yöntemleri aracılığıyla sınıflama modellerini oluşturmuşlardır. Modellerin doğruluk değerleri elde edilmiştir. Kolon kanseri verisi için karşılıklı bilgi ile bulunan genler kullanılarak sınıflama yöntemlerinin doğruluk değerleri kNN ile %61,29, YSA ile %61,29, DVM (Radyal) ile %64,51, DVM (doğrusal) ile %74,19, DVM (pol) ile %64,51, DVM (quad) ile %38,70 bulunmuştur. Lenfoma veri setinde karşılıklı bilgi ile bulunan genler ile oluşturulan sınıflama yöntemlerinin doğruluk değerleri YSA ile %100, kNN ile %90,9, DVM (Radyal) ile %90,9, DVM (doğrusal) ile %100, DVM (pol) ile

%90,9, DVM (quad) ile %86,36 bulunmuştur (112). Sina ve ark., küçük yuvarlak mavi hücre tümörü, kolon, lösemi, akciğer ve prostat kanseri olmak üzere beş halka açık mikrodizi gen ifade verisinde çalışmışlardır. Yedi yöntem ile öznitelik seçimi yapmışlardır. Her bir veri seti için DVM, NB ve karar ağacı olmak üzere üç sınıflama yöntemi ile oluşturulan modellerin hata sonuçları bulunmuştur. Beş veri setinde yedi öznitelik seçim yöntemi ile DVM, NB ve karar ağacı sınıflama yöntemleri kullanılarak elde edilen modellerin hata oranları karşılaştırılmıştır. NB için %2 ile, DVM için %1,4 ile, karar ağacı için %1,5 ile en düşük hata oranları, genler arasındaki ilişkiyi maksimum, artıklığı minimum tutarak bir filtreleme yaklaşımı kullanan karınca koloni algoritması içeren denetimsiz bir öznitelik seçim yönteminde elde edilmiştir (113). Jin ve ark., yaptıkları çalışma ile çoklu destek vektör veri açıklama tabanlı öznitelik seçim yöntemini beş genel mikrodizi veri seti üzerinde uygulayarak hızlı ve efektif olduğunu göstermeye çalışmışlardır. Kolon, lösemi, tümör ve novartis veri setleri üzerinde ortalama %90`ın üstünde başarı yakalanmıştır. Akciğer kanseri verisinde ise istenilen başarı yakalanamamıştır (114). H. Banka ve S. Dara, Hamming uzaklığı yöntemi ile hesaplanan ve yaklaşık olarak Hamming uzaklıklarını kullanan yöntem ile gen ifade verilerindeki önemli özniteliklerin daha iyi performans ile bulunabileceğini göstermişlerdir. Lösemi, lenfoma ve kolon veri setleri üzerinde öznitelik seçim yöntemi uygulanarak LibLinear, çok katmanlı algılama, DVM, rf ve karar ağacı sınıflandırıcıları ile bu yöntemin başarısı ölçülmüştür. Diğer öznitelik seçim yöntemlerinin başarıları ile karşılaştırılmıştır. %50 eğitim, %50 test olarak veri seti ikiye bölünmüştür ve 10-kat çapraz geçerlilik yapılmıştır. Her bir veri seti için farklı sınıflama yöntemlerinin performansı daha yüksek çıkmıştır (115). Chen ve ark., binlerce öznitelikten daha az sayıda önemli olam öznitelikleri seçmek için karar ağacı sınıflandırıcı ile birlikte parçacık sürüsü optimizasyonunu on bir farklı kanser veri kümesinde uygulamışlardır. Sonrasında da DVM, geriye yayılım sinir ağı, kendi kendini düzenleyen harita, karar ağacı gibi yöntemlerin sınıflama başarıları hesaplanmıştır. DVM ile %72,46, geriye yayılım sinir ağı ile %42,58, kendi kendini düzenleyen harita ile %52,60, karar ağacı ile %93,14 sınıflama başarısı hesaplanmıştır (116). Shilaskar ve ark., üç farklı veri kümesinde öznitelik seçimi yapmışlardır. Mesafe ölçüsüyle sıralanan öznitelikler seçim işlemi için ileriye doğru seçme, ileriye doğru ekleme ve geriye doğru elimine

etme yöntemleri kullanılarak karşılaştırılmıştır. DVM`nin sınıflama amacıyla kullanıldığı yöntemde ileriye doğru ekleme yönteminin diğer iki yönteme göre sonuçlarının daha iyi olduğu görülmüştür (117). Lorena ve ark., öznitelik seçimi için yirmi üç farklı gen ifadesi verisinde Fisher'ın doğrusal ayırma analizine bağlı olan BW oranı yöntemini kullanmışlardır. DVM, sınıflama yöntemi olarak tercih edilmiştir (118). Kulkarni ve ark., ortak genetik programlama ve genetik evrimleştirilmiş karar ağaçları sınıflandırıcılarını sınıflama için kullanmışlardır. Öznitelik seçimi için ise t test ve ortak bilgi yöntemlerinden yararlanmışlardır. Bağırsak kanseri verisinde yapılan testlerde ortak bilginin ve genetik programlamanın kullanıldığı yapı ile %100 başarı oranı elde edilmiştir. Algoritmanın geçerliliğini test etmek için sınıflama sonucunda duyarlılık, doğruluk, seçicilik ve ROC eğrileri ile sonuçları karşılaştırmışlardır. Gömülü bir yapı olduğu için performans açısından sarmal yöntemlere göre daha iyi sonuçlar vermiştir (119). Peraz ve ark., diffüz büyük b-hücreli lenf kanseri ve prostat kanseri veri setleri üzerinde student-t testi, ROC eğrisi analizi, wilcoxon testi ve bulanık gen filtreleme yöntemleri ile öznitelik seçimi; kNN, NB, DVM ve YSA sınıflama yöntemleri ile sınıflama yapılmıştır. Yani; iki veri seti üzerinde dört farklı öznitelik seçim yöntemi ve dört farklı sınıflama yöntemi uygulanmıştır. Sonuç olarak prostat verisinde bulanık gen filtreleme yöntemi ile belirlenen genler ile oluşturulan sınıflama modellerinin daha iyi performans gösterdiğini gözlemlemişlerdir. Bulanık gen filtreleme yöntemiyle oluşturulan verilerin sınıflamasında ise NB %94 sınıflama performansıyla sonuncu sırada yer almıştır. Lenf veri setinde de bulanık gen filtreleme ile seçilen genlerin sınıflama performansları yüksektir. Bulanık gen filtreleme, oluşturulan veri setlerinin sınıflandırılmasında NB %97 ile sonuncu sıradadır (120). Hu ve ark., meme, lösemi, akciğer, barsak, lenf, prostat ve yumurtalık olmak üzere yedi farklı kanser verisinde C4.5, bagging C4.5, rf, adaboost C4.5 ve LibSVMs şeklinde beş farklı sınıflama yöntemini karşılaştırmışlardır. Hu ve arkadaşları veri ön işleme sonrasında her veri seti için kesikli değere sahip elli gen seçmişlerdir. 10-kat çapraz geçerlilik uygulanan yedi kanser verisine ait beş farklı sınıflama yönteminin doğruluk oranları karşılaştırıldığında; en yüksek doğru sınıflama ortalamasına sahip olan yöntem %94,8 ile rf, ardından %94,1 ile adaboost C4.5 olmuştur. En sonda %88,3 ile LibSVMs gelmektedir (121). Ling ve ark., beyin tümörü, akciğer kanseri ve lösemi

mikrodizi verilerinde sınıflama yöntemlerinin karşılaştırılmasını yapmışlardır. Veri setlerinde öznitelik seçimi yapılmıştır. Daha sonra 10-kat çapraz geçerlilik uygulamışlardır. Verilerin %90'ını eğitim, %10'unu test için kullanmışlardır. Karar ağacı, NB, DVM, kNN ve YSA olmak üzere beş farklı sınıflama yöntemi uygulamışlardır ve yöntemleri doğru sınıflama oranları ile karşılaştırmışlardır. Sonuç olarak veri setlerinde en iyi performans kNN yöntemi ile elde edilmiştir. kNN`den sonra en iyi performansı gösteren yöntem NB olmuştur (122). Statnikov ve ark., yirmi iki veri setinde hem veri setinden gen seçimi yapılarak hem de veri setinin tamamını kullanarak rf ve DVM sınıflama yöntemlerinin performanslarını incelemişlerdir. Gen veri setlerinin tamamı kullanıldığı zaman on beş veri setinde DVM'nin rf'a göre daha iyi performans gösterdiği, dört veri setinde ise rf'ın daha iyi performans gösterdiği ve üç veri setinde de iki sınıflama yönteminin aynı performansa sahip olduğu gözlenmiştir. Veri setinden gen seçimi yapıldığında, on yedi veri setinde rf`a göre DVM'nin daha üstün performans, üç veri setinde rf'ın DVM'ye göre daha üstün performans ve iki veri setinde ise aynı performansı gösterdikleri gözlenmiştir (123).

Bahsedilen çalışmalardan anlaşılacağı üzere hangi öznitelik seçim yönteminde hangi sınıflama yönteminin kullanılacağı ile ilgili meta bilgi yoktur. Kullanılan veri seti, öznitelik seçim yöntemi ve değerlendirilen performans ölçüsü gibi değişkenlik gösteren koşullara göre sınıflama yöntemlerinin başarısı da değişmektedir. Ancak kNN, DVM ve NB sınıflama yöntemleri mikrodizi gen ifade verilerinde çok tercih edilen yöntemlerdir. Çalışmada da bu yöntemlere yer verilmiştir. Genel olarak kNN, DVM ve NB sınıflama yöntemleri ile birbirine yakın ve iyi performans ölçüsü değerlerine sahip sınıflama modelleri elde edilmiştir. DÖ yöntemi de sahip olduğu yüksek performans ölçüsü değerleri ile son zamanların popüler uygulama konusu olmuştur. Bu çalışmada da en iyiler arasındadır. Öznitelik seçim yöntemleri içerisinde limma, mikrodizi verileri için kullanılan bir yöntemdir ve ele alınan veri setleri üzerinde limma ile öznitelik seçimi yapıldığında sınıflama modellerinin performans ölçüsü değerleri yüksektir. Bir diğeri, gömülü öznitelik seçim yöntemlerinden olan lasso ile seçim yapıldığında genellikle performans ölçüsü değerleri daha iyi sınıflama modelleri elde edilmiştir.

Literatürde mikrodizi gen ifade verilerinin NB, DVM, kNN ve YSA gibi klasik veri madenciliği yöntemleri ile sınıflandırıldığı çalışmalar bulunmaktadır. Ancak DÖ yönteminin sınıflamada kullanıldığı çalışma sayısı oldukça azdır. DÖ yöntemine bakıldığında ise sağlık alanında daha çok biyomedikal görüntü verileri ve sinyal verilerinin olduğu çalışmalarda kullanıldığı görülmektedir (92).

# 6. SONUÇ VE ÖNERİLER

Tez çalışmasında; öznitelik sayısının fazla, örnek sayısının az olduğu veri tipine sahip olan mikrodizi gen ifade verilerinde öznitelik sayısını azaltmak için yararlanılan öznitelik seçim yöntemlerinin, veri madenciliğinde sık kullanılan sınıflama yöntemlerinin performans düzeyine etkisini göstermek amaçlanmıştır. Bu amaç doğrultusunda altısı gerçek ve dördü benzetim çalışmasından elde edilen toplam on farklı veri seti üzerinde çalışılmıştır.

Altı farklı kanser türüne ilişkin mikrodizi gen ifade verileri üzerinde öncelikle varFilter, nsFilter, rf, lasso, rfe ve limma öznitelik seçim yöntemlerinin uygulanmasıyla elde edilen önemli özniteliklerin bulunduğu veri setlerine NB, DVM, kNN, YSA ve DÖ sınıflama yöntemleri uygulanarak hasta ve sağlıklı şeklinde sınıflamanın yapıldığı sınıflama modelleri oluşturulmuştur. Bu modellere ilişkin doğruluk, duyarlılık, seçicilik ve EAKA şeklinde model performans ölçüsü değerleri elde edilmiştir. Her bir veri setine ait sonuçlar incelendiğinde;

Akciğer kanseri veri setinde;

- rfe öznitelik seçim yöntemi hariç öznitelik seçim yöntemlerinde performans düzeyi en yüksek olan sınıflama yöntemi DÖ`dür.
- Öznitelik seçim yöntemlerinin her birinde farklı performans değerlerine sahip olan NB, DVM ve kNN sınıflama yöntemlerinin performans ölçüsü değerleri birbirine yakındır.
- lasso ve limma öznitelik seçim yöntemlerinde elde edilen sınıflama yöntemlerinin genel olarak performansları diğer öznitelik seçim yöntemlerinde hesaplanan performans değerlerine göre daha iyidir.
- varFilter ve nsFilter öznitelik seçim yöntemlerinde kNN yönteminin, rfe öznitelik seçim yönteminde DÖ yönteminin başarısı daha düşüktür.
- Diğer öznitelik seçim yöntemlerinde ise en düşük performansı gösteren sınıflama yöntemi YSA`dır. YSA ise en iyi sınıflama performansına, lasso öznitelik seçim yönteminde ulaşmıştır.
- lasso dışında diğer öznitelik seçim yöntemlerinde DÖ sınıflama yönteminin kullanılması ile elde edilen modellerin duyarlılık değerinin çok yüksek olması

hasta bireyleri belirlemedeki performanslarının çok iyi olduğunu göstermektedir.

- nsFilter öznitelik seçim yönteminde kNN, limma öznitelik seçim yönteminde ise DÖ sınıflama yöntemlerinin seçicilik değerlerinin yüksek olması sağlıklı bireylerin belirlenmesinde sınıflama modellerinin performanslarının çok iyi olduğunu ifade etmektedir.
- Hasta ve sağlıklı bireylerin doğru olarak sınıflandırıldığını ifade eden ve birbirine yakın değerleri olan doğruluk ve EAKA ölçülerine bakıldığında limma öznitelik seçim yönteminde DÖ sınıflama modeli çok başarılıdır.

Lenfoma veri setinde;

- rf hariç diğer öznitelik seçim yöntemlerinde performans düzeyi en yüksek olan sınıflama yöntemi DÖ`dür. rf öznitelik seçim yönteminde ise NB sınıflama yönteminin performansı diğer sınıflama yöntemlerine göre daha iyidir.
- varFilter öznitelik seçim yönteminde NB, DVM ve YSA sınıflama yöntemlerinin performansları aynıdır.
- lasso öznitelik seçim yönteminde ise NB ve DVM yöntemleri ile elde edilen sınıflama modellerinin performansları aynıdır.
- limma öznitelik seçim yönteminde ise NB, DVM ve kNN olmak üzere üç sınıflama yönteminin performansı aynıdır.
- Özellikle rf, lasso, rfe ve limma öznitelik seçim yöntemlerinde, YSA diğer sınıflama yöntemlerine göre genel olarak daha düşük performans ölçüsü değerlerine sahip sınıflama yöntemidir.
- Genel olarak sınıflama yöntemlerinin performanslarının daha iyi olduğu öznitelik seçim yöntemleri sırasıyla limma, lasso ve rf`tır.
- Özellikle lasso, limma, rf ve rfe öznitelik seçim yöntemlerinde DÖ, nsFilter ve rf öznitelik seçim yöntemlerinde ise NB sınıflama yöntemi ile elde edilen modellerin duyarlılık değerinin çok yüksek olması hasta bireyleri belirlemedeki performanslarının çok iyi olduğunu göstermektedir.
- varFilter öznitelik seçim yönteminde NB, DVM, kNN ve YSA, nsFilter öznitelik seçim yönteminde DVM ve kNN, rf öznitelik seçim yönteminde ise NB, DVM ve kNN, lasso öznitelik seçim yönteminde NB, DVM ve DÖ, rfe

öznitelik seçim yönteminde NB, limma öznitelik seçim yönteminde ise NB, DVM, kNN, YSA ve DÖ sınıflama yöntemlerinin seçicilik değerlerinin yüksek olması sınıflama modellerinin sağlıklı bireyleri belirlemedeki performanslarının çok iyi olduğunu ifade etmektedir.

- Hasta ve sağlıklı bireylerin doğru olarak sınıflandırıldığını ifade eden doğruluk ve EAKA ölçülerine bakıldığında DÖ sınıflama yöntemi, lasso ve limma öznitelik seçim yöntemlerinde çok başarılıdır.

Rahim ağzı kanseri veri setinde;

- varFilter dışında diğer öznitelik seçim yöntemlerinde performans düzeyi en yüksek olan sınıflama yöntemi DÖ`dür. varFilter öznitelik seçim yönteminde ise kNN sınıflama yönteminin performansı diğer sınıflama yöntemlerine göre daha iyidir.
- lasso ve limma hariç diğer öznitelik seçim yöntemlerinde YSA en düşük performansa sahip sınıflama yöntemidir. lasso ve limma öznitelik seçim yöntemlerinde ise DVM sınıflama yönteminin performansı diğer sınıflama yöntemlerine göre daha düşüktür.
- varFilter öznitelik seçim yönteminde DVM ve kNN; nsFilter öznitelik seçim yönteminde DVM ve DÖ; rf, lasso, rfe ve limma öznitelik seçim yöntemlerinde DÖ sınıflama yöntemleri ile oluşturulan modellerin duyarlılık değerlerine bakıldığında hasta bireyleri belirlemedeki performanslarının çok iyi olduğunu göstermektedir.
- Sağlıklı bireyleri belirlemenin en doğru yapıldığı durumlar ise varFilter, limma, nsFilter ve lasso öznitelik seçim yöntemlerinde DÖ sınıflama yöntemi ile oluşturulan modeller ile limma öznitelik seçim yönteminde YSA sınıflama yöntemi ile oluşturulan modeldir.
- Hasta ve sağlıklı bireylerin sınıflamasının doğru olarak yapıldığını ifade eden doğruluk ve EAKA ölçülerine bakıldığında limma ve lasso öznitelik seçim yöntemlerinde DÖ sınıflama yöntemi çok başarılıdır.

Meme kanseri veri setinde;

- Öznitelik seçim yöntemlerinde performans ölçüsü değerleri genel olarak en yüksek olan sınıflama yöntemi DÖ`dür. NB, DVM ve kNN sınıflama

yöntemlerinin performans ölçüsü değerleri ise çoğunlukla birbirine yakın olup DÖ`den sonra gelmektedir.

- limma öznitelik seçim yönteminde sınıflama yöntemlerinin performans ölçüsü değerleri genellikle daha iyidir.
- rfe öznitelik seçim yönteminde ise DÖ sınıflama yönteminin başarısı diğer öznitelik seçim yöntemlerinde elde edilen DÖ modellerine göre daha düşüktür. Ancak rfe`daki diğer sınıflama yöntemlerine göre daha başarılıdır.
- rf, lasso, rfe ve limma öznitelik seçim yöntemlerinde çoğunlukla en düşük performans ölçüsü değerlerine sahip olan sınıflama yöntemi YSA`dır. varFilter ve nsFilter öznitelik seçim yöntemlerinde YSA sınıflama yönteminin performans ölçüsü değerleri birbirine yakın olup, diğer öznitelik seçim yöntemlerinde elde edilen YSA modellerine göre daha iyi performans ölçüsü değerleri vardır.
- rf, lasso, rfe ve limma öznitelik seçim yöntemlerinde DÖ sınıflama yöntemi ile oluşturulan modellerin duyarlılık değerlerine bakıldığında hasta bireyleri belirlemedeki performanslarının çok iyi olduğu anlaşılmaktadır.
- Sağlıklı bireyleri belirlemenin en doğru yapıldığı durum ise limma öznitelik seçim yönteminde DÖ sınıflama yöntemi ile oluşturulan modeldir.
- Hasta ve sağlıklı bireylerin doğru olarak sınıflandırıldığını ifade eden doğruluk ve EAKA ölçülerine bakıldığında DÖ sınıflama yöntemi, lasso ve limma öznitelik seçim yöntemlerinde çok başarılıdır.

Prostat kanseri veri setinde;

- Öznitelik seçim yöntemlerinde performans ölçüsü değerleri en iyi olan sınıflama yöntemi DÖ`dür. DÖ sınıflama yöntemi, en iyi performans ölçüsü değerlerine limma ve nsFilter öznitelik seçim yöntemlerinde sahip iken; rfe`da daha düşük performans ölçüsü değerlerine sahiptir.
- nsFilter ve rfe öznitelik seçim yöntemleri dışında diğer öznitelik seçim yöntemlerinde başarı oranı en düşük olan sınıflama yöntemi YSA`dır.
- NB ise nsFilter ve rfe öznitelik seçim yöntemlerinde performansı en düşük olan, sınıflama yöntemidir.

- Öznitelik seçim yöntemlerinin her birine bakıldığında NB, DVM ve kNN sınıflama yöntemlerinin performans ölçüsü değerleri genellikle birbirine yakındır.
- nsFilter, lasso, rfe ve limma öznitelik seçim yöntemlerinde DÖ sınıflama yöntemi ile oluşturulan modellerin duyarlılık değerleri oldukça yüksek olup hasta bireyleri belirlemedeki performanslarının çok iyi olduğunu göstermektedir.
- varFilter ve limma öznitelik seçim yöntemlerinde DÖ, nsFilter öznitelik seçim yönteminde ise DVM sınıflama yönteleri ile oluşturulan modellerin seçicilik değerlerine bakıldığında sağlıklı bireyleri belirlemedeki performanslarının çok iyi olduğunu ifade etmektedir.
- Hasta ve sağlıklı bireylerin sınıflamasının doğru olarak yapıldığını ifade eden doğruluk ve EAKA ölçülerine bakıldığında nsFilter ve limma öznitelik seçim yöntemlerinde DÖ modeli çok başarılıdır.

Lösemi veri setinde;

- rfe ve varFilter hariç diğer öznitelik seçim yöntemlerinde en iyi performans ölçüsü değerlerine sahip sınıflama yöntemi DÖ`dür.
- varFilter öznitelik seçim yönteminde NB ve kNN sınıflama yöntemleri, rfe öznitelik seçim yönteminde ise kNN sınıflama yöntemi en iyi performans ölçüsü değerlerine sahiptir.
- Özellikle rfe, varFilter ve rf öznitelik seçim yöntemlerinde sınıflama performans ölçüsü değerleri en düşük olan yöntem YSA`dır. Genelde performans ölçüsü değerleri daha düşük olan YSA sınıflama yönteminin lasso ve limma öznitelik seçim yöntemlerinde performans ölçüsü değerleri oldukça iyidir.
- lasso, limma ve varFilter öznitelik seçim yöntemlerinde sınıflama yöntemlerinin genel olarak performans ölçüsü değerleri diğer öznitelik seçim yöntemlerindeki performans ölçüsü değerlerine göre daha yüksektir.
- rf, lasso ve limma öznitelik seçim yöntemlerinde DÖ, varFilter öznitelik seçim yönteminde NB, kNN ve YSA, rf öznitelik seçim yönteminde DVM, lasso ve limma öznitelik seçim yöntemlerinde NB ve kNN, rfe öznitelik seçim yönteminde ise kNN sınıflama yöntemleri ile oluşturulan modellerin

duyarlılık değerlerinin oldukça yüksek olması hasta bireyleri belirlemedeki performanslarının çok iyi olduğunu göstermektedir.

- varFilter öznitelik seçim yönteminde NB, DVM, kNN ve DÖ sınıflama yöntemlerinin, nsFilter öznitelik seçim yönteminde de kNN, YSA ve DÖ, limma öznitelik seçim yöntemlerinde YSA ve DÖ sınıflama yöntemlerinin seçicilik değerlerinin yüksek olması sınıflama modellerinin sağlıklı bireyleri belirlemedeki performanslarının çok iyi olduğunu göstermektedir.
- Hasta ve sağlıklı bireylerin doğru olarak sınıflandırıldığını ifade eden doğruluk ve EAKA ölçülerine bakıldığında limma öznitelik seçim yönteminde DÖ, varFilter öznitelik seçim yönteminde NB ve kNN sınıflama yöntemleri çok başarılıdır.

Kanser türlerine ait veri setleri gibi mikrodizi gen ifade verisi olarak benzetim çalışması ile elde edilen veri setlerinde varFilter, rf, lasso, rfe ve limma öznitelik seçim yöntemlerinin uygulanmasıyla elde edilen önemli özniteliklerin yer aldığı veri setlerine NB, DVM, kNN, YSA ve DÖ sınıflama yöntemleri uygulanarak hasta ve sağlıklı şeklinde sınıflamanın yapıldığı sınıflama modelleri elde edilmiştir. Bu modellere ilişkin doğruluk, duyarlılık, seçicilik ve EAKA olmak üzere performans ölçüsü değerleri incelenmiştir.

Sonuçlara bakıldığında bnz-1 veri setinde;

- Öznitelik seçim yöntemlerinde genel olarak NB sınıflama yönteminin performans ölçüsü değerleri en iyi olup ilk sıralardadır.
- varFilter öznitelik seçim yönteminde DVM ve kNN sınıflama yöntemlerinin performansları birbirine benzer olup NB sınıflama yönteminin performansından sonra gelmektedir.
- rf, lasso ve limma öznitelik seçim yöntemlerinde NB, DVM, kNN ve DÖ sınıflama yöntemlerinin performans ölçüsü değerleri yüksek olup, birbirine yakındır.
- rfe öznitelik seçim yönteminde ise performans ölçüsü değerleri aynı olan NB, DVM ve kNN sınıflama yöntemlerinin performansları DÖ yönteminden daha iyidir ve DÖ`nün performans ölçüsü değerleri de oldukça düşüktür.

- varFilter hariç diğer öznitelik seçim yöntemlerinde sınıflama performansı en düşük olan yöntem YSA`dır. Genelde başarı oranı düşük olan YSA sınıflama yönteminin varFilter öznitelik seçim yönteminde ise performansı oldukça iyidir.
- rf, lasso ve limma öznitelik seçim yöntemlerinde DÖ, varFilter öznitelik seçim yönteminde YSA, lasso öznitelik seçim yönteminde NB, DVM ve kNN sınıflama yöntemleri ile oluşturulan modellerin duyarlılık değerlerinin oldukça yüksek olması modellerin hasta bireyleri belirlemedeki performanslarının çok iyi olduğunu ifade etmektedir.
- rfe öznitelik seçim yönteminde NB, DVM ve kNN, lasso ve limma öznitelik seçim yöntemlerinde NB, rf öznitelik seçim yönteminde NB ve kNN sınıflama yöntemlerinin seçicilik değerlerinin yüksek olması sınıflama modellerinin sağlıklı bireyleri belirlemede iyi performansa sahip olduklarını göstermektedir.
- Hasta ve sağlıklı bireylerin doğru olarak sınıflandırıldığını ifade eden doğruluk ve EAKA ölçülerine bakıldığında varFilter, rf, lasso ve limma öznitelik seçim yöntemlerinde NB sınıflama yöntemi, rf, lasso ve limma öznitelik seçim yöntemlerinde ayrıca DÖ sınıflama yöntemi çok başarılıdır.

Bnz-2 veri setinde;

- rfe hariç diğer öznitelik seçim yöntemlerinde DÖ sınıflama yönteminin performans ölçüsü değerleri yüksek olup ilk sıralardadır. varFilter öznitelik seçim yönteminde de NB sınıflama yönteminin performans ölçüsü değerleri en iyidir.
- NB, DVM ve kNN sınıflama yöntemlerinin performans ölçüsü değerleri lasso ve rfe öznitelik seçim yöntemlerinde birbirine yakın olup oldukça iyidir.
- Öznitelik seçim yöntemlerinde performans ölçüsü değerleri en düşük olan sınıflama yöntemi YSA`dır. Genel olarak lasso ve limma öznitelik seçim yöntemlerinde YSA sınıflama yönteminin performans ölçüsü değerleri diğer öznitelik seçim yöntemlerindeki YSA sınıflama yönteminin performans ölçüsü değerlerine göre daha iyidir.
- rf, lasso, rfe ve limma öznitelik seçim yöntemlerinde DÖ, varFilter öznitelik seçim yönteminde NB, kNN ve YSA, limma öznitelik seçim yönteminde ise

YSA sınıflama yöntemleri ile oluşturulan modellerin duyarlılık değerleri oldukça yüksektir dolayısıyla hasta bireyleri belirlemede sınıflama modellerinin performanslarının çok iyi olduğunu göstermektedir.

- varFilter öznitelik seçim yönteminde NB ve DVM, rf öznitelik seçim yönteminde NB ve kNN, lasso öznitelik seçim yönteminde NB, DVM ve DÖ, rfe öznitelik seçim yönteminde NB ile DVM ve limma öznitelik seçim yönteminde ise NB ve DÖ sınıflama yöntemlerinin seçicilik değerlerinin yüksek olması sınıflama modellerinin sağlıklı bireyleri belirlemedeki performanslarının çok iyi olduğunu ifade etmektedir.
- Hasta ve sağlıklı bireylerin sınıflamasının doğru olarak yapıldığını ifade eden doğruluk ve EAKA ölçülerine bakıldığında varFilter öznitelik seçim yönteminde NB, lasso ve limma öznitelik seçim yöntemlerinde ise DÖ yöntemi çok başarılıdır.

Bnz-3 veri setinde;

- DÖ sınıflama yönteminin başta lasso ve limma öznitelik seçim yöntemlerinde olmak üzere genel olarak performans ölçüsü değerleri oldukça yüksektir.
- Öznitelik seçim yöntemlerinde performans ölçüsü değerleri en düşük olan sınıflama yöntemi YSA`dır. YSA sınıflama yönteminin performans ölçüsü değerleri öznitelik seçim yöntemleri içinde rfe`da daha düşüktür. lasso ve varFilter öznitelik seçim yöntemlerinde YSA sınıflama yönteminin performans ölçüsü değerleri diğer öznitelik seçim yöntemlerindeki YSA sınıflama yöntemlerinin performans ölçüsü değerlerine göre daha iyidir.
- Genel olarak öznitelik seçim yöntemlerinde NB, DVM ve kNN sınıflama yöntemlerinin performans ölçüsü değerleri iyi olup birbirine yakındır. Bazen DÖ`den daha iyi, bazen de DÖ`den sonra gelmektedir.
- rf ve lasso öznitelik seçim yöntemlerinde DÖ, rfe ve limma öznitelik seçim yöntemlerinde NB ve DÖ sınıflama yöntemleri ile oluşturulan modellerin duyarlılık değerinin oldukça yüksek olması hasta bireyleri belirlemedeki performanslarının çok iyi olduğunu göstermektedir.
- varFilter öznitelik seçim yönteminde kNN, rf öznitelik seçim yönteminde DVM, lasso öznitelik seçim yönteminde NB, DVM ve DÖ, rfe öznitelik seçim yönteminde DVM ve kNN, limma öznitelik seçim yönteminde ise

DVM, kNN ve DÖ sınıflama yöntemlerinin seçicilik değerinin yüksek olması sınıflama modellerinin sağlıklı bireyleri belirlemedeki performanslarının çok iyi olduğunu ifade etmektedir.

- Hasta ve sağlıklı bireylerin doğru olarak sınıflandırıldığını ifade eden doğruluk ve EAKA ölçülerine bakıldığında varFilter, lasso ve limma öznitelik seçim yöntemlerinde DÖ; ayrıca limma`da NB, DVM ve kNN sınıflama yöntemleri ile elde edilen modellerin daha başarılı olduğu görülmektedir.

Bnz-4 veri setinde;

- rfe hariç diğer öznitelik seçim yöntemlerinde genel olarak DÖ sınıflama yöntemi, performans ölçüsü değerleri en iyi olan yöntemdir.
- varFilter öznitelik seçim yönteminde DVM, rf öznitelik seçim yönteminde NB ve kNN, lasso öznitelik seçim yönteminde DVM ve kNN, limma öznitelik seçim yönteminde NB ve DVM sınıflama yöntemlerinin performans ölçüsü değerleri DÖ`den sonra gelen ikinci en iyi sınıflama yöntemleridir.
- NB, DVM ve kNN sınıflama yöntemlerinin performans ölçüsü değerleri genel olarak öznitelik seçim yöntemlerinin her birinde birbirine yakındır.
- Öznitelik seçim yöntemlerinde uygulanan sınıflama yöntemleri içerisinde YSA sınıflama yöntemi performans ölçüsü değerleri en düşük olan sınıflama yöntemidir. limma öznitelik seçim yönteminde YSA sınıflama yönteminin performans ölçüsü değerleri diğer öznitelik seçim yöntemlerindeki YSA`lara göre daha iyidir.
- rf öznitelik seçim yönteminde DÖ, rfe öznitelik seçim yönteminde NB ve DÖ, limma öznitelik seçim yönteminde ise NB, DVM ve DÖ sınıflama yöntemleri ile oluşturulan modellerin duyarlılık değerlerinin oldukça yüksek olması sınıflama modellerinin hasta bireyleri belirlemedeki performanslarının çok iyi olduğunu göstermektedir.
- varFilter öznitelik seçim yöntemlerinde DVM, rf öznitelik seçim yöntemlerinde NB ve kNN, lasso öznitelik seçim yönteminde NB, DVM, kNN ve DÖ, rfe öznitelik seçim yönteminde DVM ve kNN, limma öznitelik seçim yönteminde ise kNN ve DÖ sınıflama yöntemlerinin seçicilik

değerlerinin yüksek olması sınıflama modellerinin sağlıklı bireyleri belirlemedeki performanslarının çok iyi olduğunu ifade etmektedir.

- Hasta ve sağlıklı bireylerin doğru olarak sınıflandırıldığını ifade eden doğruluk ve EAKA ölçülerine bakıldığında lasso ve limma öznitelik seçim yöntemlerinde DÖ; ayrıca limma`da NB ve DVM sınıflama yöntemleri ile elde edilen modellerin daha başarılı olduğu görülmektedir.

Genel olarak; kullanılan veri setlerinde veri madenciliğinin ilk aşaması olan ön işlemeden sonra öznitelik sayısı fazla olduğu için varFilter, nsFilter, rf, lasso, rfe ve limma isimli altı farklı öznitelik seçim yöntemi sayesinde bundan sonraki adımlarda kullanılacak olan anlamlı ve önemli öznitelikler belirlenmiştir. Öznitelik seçimi yapılmış veri setleri ile NB, DVM, kNN, YSA ve DÖ sınıflama yöntemleri aracılığıyla sınıflama modelleri elde edilmiştir. Elde edilen modeller sayesinde hasta-sağlıklı sınıflaması yapılmıştır. Ne kadar doğru sınıflama yapıldığı ile ilgili doğruluk, duyarlılık, seçicilik ve EAKA olmak üzere model performans ölçüleri elde edilmiştir. Böylece mikrodizi gen ifade verilerinde farklı öznitelik seçim yöntemleri ile sınıflama yöntemlerinin performansların değerlendirilmiştir.

Ayrıca veri setleri üzerinde öznitelik seçim yöntemi uygulamadan DÖ yöntemi uygulanarak DÖ sınıflama modelleri oluşturulmuştur. Genel olarak öznitelik seçim yöntemi kullanılmadan elde edilen DÖ sınıflama modellerinin performansları rfe hariç diğer öznitelik seçim yöntemlerinin uygulanması ile oluşturulan DÖ sınıflama modellerinin performanslarından düşük çıkmıştır.

Son söz olarak; akciğer kanseri, lösemi ve bnz-3 veri setlerinde oluşturulan modellerin performans ölçüsü değerleri genel olarak daha yüksektir. Rahim ağzı, meme prostat kanseri ve bnz-4 veri setlerinde ise performans ölçüsü değerleri daha düşüktür. Genel olarak limma ve lasso öznitelik seçim yöntemleri uygulanan veri setlerinde sınıflama yöntemlerinin kullanılması ile performans ölçüsü değerleri daha yüksek sınıflama modelleri elde edilmiştir. rfe öznitelik seçim yöntemi kullanılarak elde edilen veriler üzerinde DÖ yöntemi uygulandığında oluşturulan sınıflama modelleri çoğunlukla düşük performans ölçüsü değerlerine sahip olmuştur. Ancak DÖ ile genellikle daha yüksek, YSA ile daha düşük performans ölçüsü değerlerinin olduğu sınıflama modelleri elde edilmiştir. NB, DVM ve kNN sınıflama yöntemleri ile oluşturulan sınıflama modellerinin performans ölçüsü değerleri de çoğunlukla

birbirine yakın çıkmıştır. Birkaç durum dışında, DÖ sınıflama yöntemi ile elde edilen modellerin performans ölçüsü değerlerini takip etmişlerdir. Özellikle NB ve DVM sınıflama yöntemleri bazı durumlarda YSA sınıflama yönteminden daha düşük performans ölçüsü değerlerine sahip olmuştur.

Genel olarak bakıldığında, öznitelik sayısının çok olduğu mikrodizi gen ifade verilerinde lasso ve limma öznitelik seçim yöntemlerinin kullanılması tercih edilebilir. DÖ yöntemi mikrodizi gen ifade verileri gibi büyük boyutlu verilerin sınıflandırılmasında klasik veri madenciliği yöntemlerine göre daha başarılı sonuçlar vermiştir ve kullanılması önerilmektedir. Büyük boyutlu veri setlerinde DÖ yöntemi uygulamasından önce rfe öznitelik seçim yönteminin kullanılmaması tavsiye edilmektedir. Gelecekteki çalışmalar için başta genetik alanı olmak üzere, tıbbi görüntüleme cihazları ile elde edilen verilerin tanınması gibi farklı alanlarda, DÖ yönteminin kullanılması planlanmaktadır. Ayrıca fazla özniteliğin bulunduğu veri kümelerinde farklı öznitelik seçim yöntemlerini de dâhil ederek lasso, limma, rf ve rfe gibi öznitelik seçim yöntemleri arasında hız ve sınıflama performansı açısından karşılaştırmalar yapılması hedeflenmektedir. Sık kullanılan diğer veri madenciliği yöntemlerinden olan kümeleme ve birliktelik kurallarının da mikrodizi gen ifade verileri üzerinde uygulaması yapılabilir.

# 7. KAYNAKLAR

1. Boyacıoğlu H, Güneri P. Sağlık Araştırmalarında Kullanılan Temel İstatistik Yöntemler. Hacettepe Dişhekimliği Fakültesi Dergisi. 2006;30(3):33-39.

2. Öğüş E. To Be Together Medicine And Biostatistics İn History: Review. Turkiye Klinikleri J Biostat. 2017;9(1):74-83.

3. Öner TÖ, Can Ş. Sağlıkta Biyoistatistiksel Uygulamalar. İzmir Kâtip Çelebi Üniversitesi Sağlık Bilimleri Fakültesi Dergisi. 2018;3(1):39-45.

4. Karabulut E, Karaağaoğlu E. Biyoinformatik ve Biyoistatistik. Hacettepe Tıp Dergisi. 2010;41:162-170.

5. Polat M, Karahan AG. Multidisipliner Yeni Bir Bilim Dalı: Biyoinformatik ve Tıpta Uygulamaları. S.D.Ü. Tıp Fak. Derg. 2009;16(3):41-50.

6. Yoldaş A, Karaboz İ. DNA Mikroarray Teknolojisi ve Uygulama Alanları. Elektronik Mikrobiyoloji Dergisi TR. 2010;8(1):1-19.

7. Baykara O. Kanser Tedavisinde Güncel Yaklaşımlar. Balıkesir Sağlık Bilimleri Dergisi. 2016;5(3):154-165.

8. Demircioğlu HZ, Bilge HŞ. Yumurtalık Kanseri Veri Kümesindeki Gen İfadelerinin Veri Madenciliği İle Analizi. Marmara Fen Bilimleri Dergisi. 2015; 4:125-134.

9. Coşkun E, Karaağaoğlu E. Veri Madenciliği Yöntemleri ile Mikrodizilim Gen İfade Analizi. Hacettepe Tıp Dergisi. 2011;42:180-189.

10. Haznedar B, Arslan MT, Kalınlı A. Karaciğer Mikrodizi Kanser Verisinin Sınıflandırılması için Genetik Algoritma Kullanarak ANFIS'in Eğitilmesi. Sakarya Üniversitesi Fen Bilimleri Enstitüsü Dergisi. 2017;21(1):54-62.

11. Özkan Y, Selçukcan Erol Ç. Biyoenformatik DNA Mikrodizi Veri Madenciliği. 2. Baskı. İstanbul: Papatya Yayıncılık; 2017.

12. Dolgun MÖ. Veri Madenciliği Sınıflama Yöntemlerinin Başarılarının; Bağımlı Değişken Prevelansı, Örneklem Büyüklüğü ve Bağımsız Değişkenler Arası İlişki Yapısına Göre Karşılaştırılması [Doktora tezi]. Ankara: Hacettepe Üniversitesi; 2014.

13. Çelik N. Amiyotrofik Lateral Skleroz (ALS) Hastalığının Genetik ve Klinik Veri İlişkisinin Veri Madenciliği Yöntemleri ile İncelenmesi [Yüksek Lisans Tezi]. Antalya: Akdeniz Üniversitesi; 2017.

14. Çataloluk H. Gerçek Tıbbi Veriler Üzerinde Veri Madenciliği Yöntemlerini Kullanarak Hastalık Teşhisi [Yüksek Lisans Tezi]. Bilecik: Bilecik Üniversitesi; 2012.

15. Tukey JW. Exploratory Data Analysis, 2.cilt,18.Baskı. ABD: Addison-Wesley Publishing Company; 1977.

16. Karaibrahimoğlu A. Veri Madenciliğinden Birliktelik Kuralı ile Onkoloji Verilerinin Analiz Edilmesi: Meram Tıp Fakültesi Onkoloji Örneği [ Doktora Tezi]. Konya: Selçuk Üniversitesi; 2014.

17. Han J, Kamber M, Pei J. Data Mining Concepts and Techniques. 3.Baskı. ABD: Elsevier; 2012.

18. Poyraz O. Tıp`Da Veri Madenciliği Uygulamaları: Meme Kanseri Veri Seti Analizi [Yüksek Lisans Tezi]. Edirne: Trakya Üniversitesi; 2012.

19. Bircan H, Çam S. Veri Madenciliğinde Kümeleme Analizi ve Sağlık Sektöründe Bir Uygulaması. C.Ü. İktisadi ve İdari Bilimler Dergisi. 2016;17(2):85-96.

20. Gündoğdu ÖE. Veri Madenciliğinde Genetik Algoritmalar [Yüksek Lisans Tezi]. Kocaeli: Kocaeli Üniversitesi; 2007.

21. Toprak U. Karsinogenezde Mutasyonlar Arası İlişkilerin Veri Madenciliği Metotları İle Tespiti [Yüksek Lisans Tezi]. Trabzon: Karadeniz Teknik Üniversitesi; 2015.

22. Swift RS. Accelerating Customer Relationships: Using CRM and Relationship Technologies. ABD: Prentice Hall; 2001.

23. Larose DT. Discovering Knowledge in Data An Introduction to Data Mining. New Jersey: Wiley; 2005.

24. Alaybeg F. Veri Madenciliği Giriş, Yöntemleri ve Metodolojileri. [İnternet]. Mart,2019. Erişim adresi: https://medium.com/@furkanalaybeg/veri-madencili%C4%9Fi-ve-y%C3%B6ntemleri-d0e2fd238e44.

25. Hand DJ. Classifier Technology and the Illusion of Progress. Statist. Sci. 2006;21(1):1-14.

26. Akküçük U. Veri Madenciliği: Kümeleme ve Sınıflama Algoritmaları. İstanbul: Yalın Yayıncılık; 2011.

27. Edelstein HA. Introduction to Data Mining and Knowledge Discovery. 3. Baskı. ABD: Two Crows Corporation; 1999.

28. Sever H, Oğuz B. Veri Tabanlarında Bilgi Keşfine Formel Bir Yaklaşım Kısım I: Eşleştirme Sorguları ve Algoritmalar. Bilgi Dünyası. 2002;3(2):173-204.

29. Haberal İ. Veri Madenciliği Algoritmaları Kullanılarak Web Günlük Erişimlerinin Analizi [Yüksek Lisans Tezi]. Ankara: Başkent Üniversitesi; 2007.

30. Göker H. Üniversite Giriş Sınavında Öğrencilerin Başarılarının Veri Madenciliği Yöntemleri İle Tahmin Edilmesi. [Yüksek Lisans Tezi]. Ankara: Gazi Üniversitesi; 2012.

31. Ergün K. Veri Madenciliğine Giriş. Balıkesir Üniversitesi MF Endüstri Mühendisliği Bölümü Veri Madenciliği Ders Notu.

32. Luscombe NM, Greenbaum D, Gerstein M. What İs Bioinformatics? An İntroduction And Overview. Yearbook Of Medical Informatics. 2001;10(01):83-100.

33. Tanır D. Genomik Veri Tabanlarında İndeksleme ve Arama Yöntemleri Üzerine [Doktora Tezi]. İzmir: Ege Üniversitesi; 2017.

34. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, ve ark. Bioconductor: Open Software Development For Computational Biology And Bioinformatics. Genome Biol. 2004;5(10):R80.

35. Collins FS, Morgan M, Patrinos A. The Human Genome Project Lessons from Large-Scale Biyology. Science. 2003;300:286-290.

36.Telefoncu A. Biyoinformatik I. Biyoinformatik Lisansüstü Yaz Okulu. İzmir:2003;s.3-20.

37. Hopkins MM, Ibarreta D, Gaisser S, Enzing CM, Ryan J, Martin PA, ve ark. Putting Pharmacogenetics into Practice. Nature Biotechnology. 2006; 24(4): 403-410.

38. Sayitoğlu M. Kanser Tedavisine Farmakogenetik Yaklaşım. Turkiye Klinikleri J Med Sci. 2007;27:434-441.

39. Roberts HF. Plant Hybridization before Mendel. Princeton: Princeton University Press, 1929.

40. Tisdall J. Beginning Perl for Bioinformatics. ABD: O'Reilly; 2001.

41. Özdoğan A. Gen Kümeleme İşleminin Özdüzenleyici Haritalar Kullanılarak Gen Ekspresyonu, Motif Sıklık Ve Gen Konum Verilerinden Faydalanılarak Gerçekleştirimi. [Yüksek Lisans Tezi]. İstanbul: Yıldız Teknik Üniversitesi; 2009.

42. Watson JD, Crick FH. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature. 1953;171:737-738.

43. Lesk AM. Introduction to Bionformatics. New York: Oxford University Press Inc.; 2002.

44.DNA'nın Keşfi [İnternet]. Ocak, 2019. Erişim adresi: https://www.biyologlar.com/dnanin-kesfi.

45. Setubal J, Meidanis J. Introduction to Computational Molecular Biology. ABD: PWS Publishing Company; 1997.

46.Mendel G. Versuche über Plflanzenhybriden. Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr. 1865; 3–47.

47. Klug WS, Cummings MR, Spencer CA. Genetik Kavramlar. 8. Baskı. Ankara: Palme Yayıncılık; 2009.

48.Zararsız G. Gen Ekspresyon Verilerinde Kümelemeye Dayalı Yeni Bir Sınıflandırma Yaklaşımı. [Yüksek Lisans Tezi]. Kayseri: Erciyes Üniversitesi; 2012.

49. Bakırcı ÇM. Temel Genetik Kavramlar: Nükleotit, DNA, Gen, Kromozom Nedir?. [İnternet]. 2019. Erişim adresi: https://evrimagaci.org/temel-genetik-kavramlar-nukleotit-dna-gen-kromozom-nedir-7566.

50. Lüleyap HÜ. Moleküler Genetiğin Esasları. İzmir: Nobel Kitabevi; 2008.

51. Sarıkaş A, Odabaşıoğlu N, Altay G. Gen İfade Verilerinde Eksik Değerleri Düzelten Kestirim Yöntemlerinin Karşılaştırılması Comparison of Estimation Methods for Missing Value Imputation of Gene Expression Data. Tıp Teknolojileri Kongresi; 27-29- Ekim 2016; Antalya. s.114-117.

52. Sassanfar S, Walker G. DNA Microarray Technology. What Is It and How Is It Useful, MIT, Biology Science Outreach. 2003.

53. İdil NB. Gen İfade Verileri ile İşlemsel Kanser Sınıflandırılması [Yüksek Lisans Tezi]. Ankara: Başkent Üniversitesi; 2009.

54. Gershon D. Microarray Technology: An Array of Opportunities. Nature. 2002: 885-891.

55. Öztemur Y, Aydos A, Gür-Dedeoğlu B. Meme Kanseri Mikrodizin Verilerinin Biyoinformatik Yöntemler ile Bir Araya Getirilmesi - Meta-Analiz Yaklaşımları. Turk Hij Den Biyol Derg, 2015;72(2):155-162.

56. George GVS and Raj VC. Review on Feature Selection Techniques and The İmpact Of SVM For Cancer Classification Using Gene Expression Profile. International Journal Of Computer Science & Engineering Survey. 2011;16-27.

57. Babu MM. An İntroduction to Microarray Data Analysis. Grant RP. Computational Genomics: Theory and Application. Oxford: Horizon Bioscience; 2004.

58. Quackenbush J. Computational Analysis of Microarray Data. Nat Rev Genet. 2001;2(6): 418-27.

59. Gibson G. Microarray Analysis. PLoS Biol. 2003;1(1):e15.

60. Jagota A. Microarray Data Analysis and Visualization. Bioinformatics by the Bay Press: Santa Cruz; 2001.

61. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP ve ark. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. 1999;286(5439):531-537.

62. Kaya A. Bilgisayar Destekli Tanı Sistemi ile Akciğer Nodüllerinin Nitelendirilmesi [Doktora Tezi]. Ankara: Hacettepe Üniversitesi; 2015.

63. Budak H. Özellik Seçim Yöntemleri ve Yeni Bir Yaklaşım. Süleyman Demirel University Journal of Natural and Applied Sciences. 2018;22 (Special Issue):21-31.

64. Yazıcı B, Yaslı F, Yıldız Gürleyik H, Turgut UO, Aktas MS, Kalıpsız O. Veri Madenciliğinde Özellik Seçim Tekniklerinin Bankacılık Verisine Uygulanması Üzerine Araştırma ve Karşılaştırmalı Uygulama. 9. Ulusal Yazılım Mühendisliği Sempozyumu; 15 - 17 Eylül 2015; İzmir. UYMS-15. s.72-83.

65. Kaya M. Gen İfade Verilerinde Öznitelik Seçimi ve Sınıflandırma [Yüksek Lisans Tezi].Ankara: Gazi Üniversitesi; 2014.

66. Var E, İnan A. Sınıflandırma İçin Diferansiyel Mahremiyete Dayalı Öznitelik Seçimi. Journal of the Faculty of Engineering and Architecture of Gazi University. 2018;33(1):323-336.

67. Zengin HY. Sosyal Ağ Analizinin Hastalık Biyobelirteçlerinin Belirlenmesinde Kullanımı [Doktora Tezi]. Ankara: Hacettepe Üniversitesi; 2018.

68. Özkan Y, Selçukcan Erol Ç. Kanser Biyoenformatiğinde Yapay Zeka. 2. Baskı. İstanbul: Papatya Yayıncılık; 2019.

69. Falcon S, Morgan M, Gentleman R. An Introduction to Bioconductor's ExpressionSet Class [İnternet]. Temmuz, 2019. Erişim adresi:

https://www.bioconductor.org/packages/release/bioc/vignettes/Biobase/inst/doc/ExpressionSetIntroduction.pdf.

70. Gentleman R, Carey V, Huber W, Hahne F. Genefilter: Genefilter: Methods For Filtering Genes From High-Throughput Experiments [İnternet]. 2019. Erişim adresi: https://bioconductor.org/packages/devel/bioc/manuals/genefilter/man/genefilter.pdf.

71. Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. PNAS. 2010; 107(21):9546-9551.

72. Slawski M, Boulesteix AL, Bernau C. CMA: Synthesis of microarray-based classification [İnternet]. Haziran, 2019. Erişim adresi: https://www.bioconductor.org/packages/release/bioc/manuals/CMA/man/CMA.pdf.

73. Daş B, Türkoğlu İ. DNA Dizilimlerinin Sınıflandırılmasında Karar Ağacı Algoritmalarının Karşılaştırılması. Elektrik – Elektronik – Bilgisayar ve Biyomedikal Mühendisliği Sempozyumu; 27 – 29 Kasım 2014; Bursa. Eleco 2014. s.381-383.

74. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological). 1996;58(1):267-288.

75. Fontı V. Feature Selection using LASSO. Vrije Universiteit Amsterdam; 2017.s.26.

76. Ludwig N, Feuerriegel S, Neumann D. Putting Big Data Analytics to Work: Feature Selection for Forecasting Electricity Prices Using the LASSO and Random Forests. Journal Of Decision Systems. 2015;24(1):19-36.

77. Muthukrishnan R, Rohini R. LASSO: A Feature Selection Technique in Predictive Modeling for Machine Learning. IEEE International Conference on Advances in Computer Applications; 2016; Coimbatore. ICACA 2016. s.18-20.

78.Küçük A. Doğrusal Regresyonda Rıdge, Lıu ve Lasso Tahmin Edicileri Üzerine Bir Çalışma [Yüksek Lisans Tezi]. Ankara: Hacettepe Üniversitesi; 2019.

79. Wang C, Xiao Z, Wang B, Wu J. Identification of Autism Based on SVM-RFE and Stacked Sparse Auto-Encoder. IEEE Access. 2019;7:118030-118036.

80. Smyth GK. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Statistical Applications in Genetics and Molecular Biology. 2004;3(1):1-26.

81. Smyth GK, Hu Y, Ritchie M, Silver J, Wettenhall J, McCarthy ve ark. limma: Linear Models for Microarray Data [İnternet]. Eylül, 2019. Erişim adresi: https://bioconductor.org/packages/release/bioc/manuals/limma/man/limma.pdf.

82. Smyth GK, Ritchie M, Thorne N, Wettenhall J, Shi W, Hu Y. limma: Linear Models for Microarray and RNA-Seq Data User's Guide [İnternet]. Kasım, 2019.Erişimadresi:https://www.bioconductor.org/packages/devel/bioc/vignettes/limma/inst/doc/usersguide.pdf.

83. Pala T. Tıbbi Karar Destek Sisteminin Veri Madenciliği Yöntemleriyle Gerçekleştirilmesi [Yüksek Lisans Tezi]. İstanbul: Marmara Üniversitesi; 2013.

84. Akman M, Genç Y, Ankaralı H. Random Forests Yöntemi ve Sağlık Alanında Bir Uygulama. Turkiye Klinikleri J Biostat.2011;3(1):36-48.

85. Wibawa AP, Kurniawan AC, Murti DM, Adiperkasa RP, Putra SM, Kurniawan SA ve ark. Nugraha YR. Naive Bayes Classifier for Journal Quartile Classification. iJES. 2019;7(2):91-99.

86. Onan A, Korukoğlu S. Metin Sınıflandırmada Öznitelik Seçim Yöntemlerinin Değerlendirilmesi. XVIII. Akademik Bilişim Konferansı; 30 Ocak-5 Şubat 2016; Adnan Menderes Üniversitesi, Aydın.

87. Korkem E. Mikroarray Gen Ekspresyon Veri Setlerinde Random Forest ve Naive Bayes Sınıflama Yöntemleri Yaklaşımı [Yüksek Lisans Tezi]. Ankara: Hacettepe Üniversitesi; 2013.

88. Kayaalp F, Başarslan MS, Polat K. Kronik Böbrek Hastalığını Tanımlamada Bir Hibrit Sınıflandırma Örneği. Electric Electronics, Computer Science, Biomedical Engineerings Meeting;18-19 Nisan 2018; İstanbul Arel Üniversitesi, İstanbul.

89. Çiftçi F, Kaleli C, Günal S. Öznitelik Seçme ve Makine Öğrenmesi Yöntemleriyle Eğitmen Performansının Tahmin Edilmesi. Anadolu Journal of Educational Sciences International. 2018; 8(2): 419-440.

90. Vapnik VN. An Overview of Statistical Learning Theory. IEEE Transactıons on Neural Networks.1999;10(5):988-999.

91. Aydın Haklı D. Sınıf Dengesizliği Sorununu Çözmek için Kullanılan Algoritmaların Farklı Sınıflandırma Yöntemlerinde Performanslarının Karşılaştırılması. Ankara: Hacettepe Üniversitesi; 2018.

92. Kaşıkçı M. Transkriptom Veri Seti Üzerinde Derin Öğrenme Yöntemi ile Klasik Veri Madenciliği Yöntemlerinin Sınıflama Performanslarının Karşılaştırılması [Yüksek Lisans Tezi]. Ankara: Hacettepe Üniversitesi; 2019.

93. Gümüş E. Makina Öğrenme Yöntemleriyle Genom Dizilim Verilerinin Analizi [Doktora Tezi]. İstanbul: İstanbul Üniversitesi; 2013.

94. Çakmak I. Makine Öğrenmesi Yöntemleriyle Tümör Kontrol Olasılığının Hesaplanması [Yüksek Lisans Tezi]. Trabzon: Karadeniz Teknik Üniversitesi; 2017.

95. Fix E, Hodges JL. Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties. Technical Report 4. USAF School of Aviation Medicine, Randolph Field, Texas;1951. Report No:4.

96. Cover MT, Hart P. Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory. 1967;13(1):21-27.

97. Elasan S. Veri Madenciliğinde Farklı Karar Ağaçları ve K-En Yakın Komşuluk Yöntemlerinin İncelenmesi: Kadın Hastalıkları ve Doğum Verisinde Bir Uygulama [Doktora Tezi]. Van: Van Yüzüncü Yıl Üniversitesi; 2019.

98.Taşcı E, Onan E. K-En Yakın Komşu Algoritması Parametrelerinin Sınıflandırma Performansı Üzerine Etkisinin İncelenmesi. XVIII. Akademik Bilişim Konferansı; 30 Ocak-5 Şubat 2016; Adnan Menderes Üniversitesi, Aydın.

99. Öğüş E, Can MB, Çamur E, Koru M, Özkan Ö, Rzayeva Z. Veri Kümelerinden Bilgi Keşfi: Veri Madenciliği; 15. Ulusal Biyoistatistik Kongresi, Uluslararası Katılımlı; 20-23 Ağustos 2013; Aydın.

100. Daş B, Türkoğlu İ. DNA Dizilimlerindeki Nükleotit Çiftlerinin Frekans Değerlerine Göre Farklı Sınıflandırma Yöntemleri ile Karşılaştırılması. Tıp Teknolojileri Ulusal Kongresi; 25 – 27 Eylül 2014; Kapadokya. TıpTekno`14. s.191 -194.

101. Demircioğlu HZ. Biyoinformatikte Çok Boyutlu Verilerin Boyut İndirgenerek Sınıflandırılması [Yüksek Lisans Tezi]. Ankara: Gazi Üniversitesi; 2015.

102. Koyuncugil As, Özgülbaş N. Veri Madenciliği: Tıp ve Sağlık Hizmetlerinde Kullanımı ve Uygulamaları. Bilişim Teknolojileri Dergisi. 2009;2(2):21-32.

103. McCulloch WS, Pitts WH. A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics.1943;5:115-133.

104. Hinton GE, Salakhutdinov RR. Reducing the Dimensionality of Data with Neural Networks. Science.2006;313(5786):504-507.

105. Türkçetin AÖ. Akciğer Kanserinin Tespit Edilmesinde Derin Öğrenme Algoritmalarının Kullanılması [Yüksek Lisans Tezi]. Isparta: Isparta Uygulamalı Bilimler Üniversitesi; 2019.

106. Doğan F, Türkoğlu İ. Derin Öğrenme Algoritmalarının Yaprak Sınıflandırma Başarımlarının Karşılaştırılması. Sakarya University Journal of Computer and Information Sciences.2018;1:10-21.

107. Çoşkun C, Baykal A. Veri Madenciliğinde Sınıflandırma Algoritmalarının Bir Örnek Üzerinde Karşılaştırılması. XIII. Akademik Bilişim Konferansı. 2-4 Şubat 2011.

108. Kılıçkap M. Bilgi Kuramı Yaklaşımı ile Bilgisayarlı Tomografik Koroner Anjiyografinin Tanısal Değerinin Değerlendirilmesi [Yüksek Lisans Tezi]. Ankara: Hacettepe Üniversitesi; 2012.

109. Dişçi R. Tanı Testlerinin Değerlendirilmesi ROC Analizi. İstanbul Üniversitesi, Onkoloji Enstitüsü; 2013.

110. Dembele D. A Flexible Microarray Data Simulation Model. Microarrays. 2013;2:115-130.

111. Dembele D.madsim: A Flexible Microarray Data Simulation Model [İnternet]. Mayıs, 2019. Erişim adresi: https://cran.r-project.org/web/packages/madsim/madsim.pdf.

112. Devi A, Vanitha C, Devaraj D, Venkatesulu M. Gene Expression Data Classification Using Support Vector Machine and Mutual İnformation-Based Gene Selection. Procedia Computer Science. 2015;47:13-21.

113. Sina T, Ali N, Reza R, Parham M. Gene selection for microarray data classifcation using a novel ant colony optimization. Neurocomputting.2015; 168:1024–1036.

114. Jin C, Li Z, Bangjun W, Fanzhang L, Jiwen Y. A Fast Gene Selection Method for Multi Cancer Classification Using Multiple Support Vector Data Description. Journal of Biomedical Informatics.2015; 53:381–389.

115. Banka H, Dara S. A Hamming Distance Based Binary Particle Swarm Optimization (HDBPSO) Algorithm for High Dimensional Feature Selection, Classification and Validation. Pattern Recognition Letters. 2015;52:94-100.

116. Chen KH, Wang KJ, Wang KM, Angelia MA. Applying Particle Swarm Optimizationbased Decision Tree Classifier for Cancer Classification on Gene Expression Data. Applied Soft Computing. 2014;24:773-780.

117. Shilaskar S, Ghatol A. Feature Selection for Medical Diagnosis: Evalution for Cardiovascular Diseases. Expert Systems with Applications. 2013;40:4146-4153.

118. Lorena A, Costa I, Spolaor N, Souto M. Analysis of Complexity İndices for Classification Problems: Cancer Gene Expression Data. Neurocomputing. 2012: 75:33-42.

119. Kulkarni A, Kumar N, Ravi V, Murthy US. Colon Cancer Prediction with Genetics Profiles Using Evolutionary Techniques. Expert Systems With Applications.2011;38:2752-2757.

120. Peraz M, Marwala T. The Fuzzy Gene Filter: A Classifier Performance Assesment [İnternet]. Nisan, 2020. Erişim adresi: https://arxiv.org/abs/1108.4545.

121. Hu H, Li J, Plank A, Wang H, Daggard G. A Comparative Study of Classification Methods for Microarray Data Analysis Conference: Data Mining and Analytics 2006, Proceedings of the Fifth Australasian Data Mining Conference (AusDM2006); 29-30 Kasım 2006; Sydney, NSW, Australia. Proceedings. s.33-37.

122. Ling NE, Hasan YA. Classification on Microarray Data. Regional Conference on Mathematics, Statistics and Applications (IRCMSA); 2006; University of Science Malaysia. s.1-8.

123. Statnikov A, Wang L, Aliferis CF. A Comprehensive Comparison Of Random Forests And Support Vector Machines For Microarray-Based Cancer Classification. BMC Bioinformatics. 2008;9:319-329.

# 8. EKLER

**EK-1:** Tez Çalışması Orijinallik Raporu

Mikrodizi Gen İfade Verilerinde Farklı Öznitelik Seçim Yöntemleri İle Sınıflama Yöntemlerinin Performanslarının Değerlendirilmesi

ORIJINALLIK RAPORU

| %10 | %8 | %2 | %7 |
|---|---|---|---|
| BENZERLIK ENDEKSI | İNTERNET KAYNAKLARI | YAYINLAR | ÖĞRENCI ÖDEVLERI |

BIRINCIL KAYNAKLAR

| | | |
|---|---|---|
| 1 | Submitted to Hacettepe University<br>Öğrenci Ödevi | %2 |
| 2 | www.openaccess.hacettepe.edu.tr:8080<br>İnternet Kaynağı | %1 |
| 3 | openaccess.hacettepe.edu.tr:8080<br>İnternet Kaynağı | %1 |
| 4 | avesis.marmara.edu.tr<br>İnternet Kaynağı | <%1 |
| 5 | dpu.edu.tr<br>İnternet Kaynağı | <%1 |
| 6 | e-dergi.marmara.edu.tr<br>İnternet Kaynağı | <%1 |
| 7 | es.scribd.com<br>İnternet Kaynağı | <%1 |
| 8 | busqueda.bvsalud.org<br>İnternet Kaynağı | <%1 |