

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

**YÜKSEK BOYUT DÜŞÜK ÖRNEKLEM GENİŞLİĞİ
DURUMUNDA SINIFLAMA ALGORİTMALARININ
PERFORMANSLARININ KARŞILAŞTIRILMASI**

Ülger AYDOĞAN CULHA

**Biyoistatistik Programı
DOKTORA TEZİ**

**ANKARA
2015**

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

**YÜKSEK BOYUT DÜŞÜK ÖRNEKLEM GENİŞLİĞİ
DURUMUNDA SINIFLAMA ALGORİTMALARININ
PERFORMANSLARININ KARŞILAŞTIRILMASI**

Ülger AYDOĞAN CULHA

**Biyoistatistik Programı
DOKTORA TEZİ**

**TEZ DANIŞMANI
Doç. Dr. Erdem KARABULUT**

**ANKARA
2015**

Anabilim Dalı :BİYOİSTATİSTİK
 Program :BİYOİSTATİSTİK
 Tez Başlığı :YÜKSEK BOYUT DÜŞÜK ÖRNEKLEM GENİŞLİĞİ DURUMUNDA
 SINIFLAMA ALGORİTMALARININ PERFORMANSLARININ
 KARŞILAŞTIRILMASI
 Öğrenci Adı-Soyadı :ÜLGER AYDOĞAN CULHA
 Savunma Sınavı Tarihi :28.12.2015

Bu çalışma jürimiz tarafından yüksek lisans/doktora tezi olarak kabul edilmiştir.

Jüri Başkanı: Prof. Dr. OSMAN SARAÇBAŞI
 (Hacettepe Üniversitesi)
 Tez danışmanı: Doç. Dr. ERDEM KARABULUT
 (Hacettepe Üniversitesi)
 Üye: Prof. Dr. ATILLA HALİL ELHAN
 (Ankara Üniversitesi)
 Üye: Prof. Dr. MEHTAP AKÇİL OK
 (Baskent Üniversitesi)
 Üye: Yrd. Doç. Dr. NİMET ANIL DOLGUN
 (Hacettepe Üniversitesi)

(imza)
 (imza)
 (imza)
 (imza)
 (imza)

ONAY

Bu tez Hacettepe Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca yukarıdaki jüri tarafından uygun görülmüş ve Sağlık Bilimleri Enstitüsü Yönetim Kurulu kararıyla kabul edilmiştir.

(imza)
 Prof.Dr. Ersin FADILLIOĞLU
 Müdür

TEŞEKKÜR

Tez çalışmamın her aşamasında desteğini ve bilgisini esirgemeyen sevgili danışmanım Doç. Dr. Erdem Karabulut'a, tezin ilerleyişinde çok değerli katkıları olan tez izlemesi komitesi üyeleri hocalarım Prof. Dr. Osman Saraçbaşı ve Prof. Dr. Atilla H. Elhan'a teşekkür ederim.

Doktora eğitimim süresince üzerimde emeği ve katkısı olan tüm hocalarıma ve arkadaşlarıma teşekkür ederim.

Son olarak, sevgisi ve desteği ile her zaman yanımda olan annem Çiçek Aydoğan, babam Kenan Aydoğan ve eşim Ali Rıza Culha'ya teşekkür ederim.



ÖZET

Aydoğan Culha, Ü. Yüksek Boyut Düşük Örneklem Genişliği Durumunda Sınıflama Algoritmalarının Performanslarının Karşılaştırılması. Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü, Biyoistatistik Programı Doktora Tezi, Ankara, 2015. Günümüzde özellikle genetik, biyoinformatik, tıbbi görüntüleme gibi alanlarda her bir gözleme ait binlerce özelliğin/niteliğin aynı anda ölçülmesine imkan veren teknolojiler oldukça popülerdir. Bu ise “Yüksek Boyut Düşük Örneklem Genişlikli (YBDÖG) Düzenler” olarak adlandırılan farklı bir veri yapısını karşımıza çıkarmakta olup, klasik istatistiksel yöntemlerle bu verilerin analizinde bir takım zorluklar yaşanmaktadır. Yüksek boyutlu verilerin sınıflandırılmasında farklı üstünlüklere sahip farklı makine öğrenmesi yöntemlerinin düşük örneklem genişliği söz konusu olduğunda performanslarındaki değişim ya da yanlılık eğilimleri ise dikkate alınmamaktadır. Bu çalışma ile YBDÖG veri setlerinde Destek Vektör Makineleri (DVM), Uzaklık Ağırlıklı Ayırma (UAA), Random Forest (RForest), Random Ferns (RFerns) ve Naive Bayes (NB) algoritmalarının sınıflama performanslarının karşılaştırılması amaçlanmıştır. Benzetim teknikleri kullanılarak boyut, örneklem genişliği, korelasyon yapısı, gürültü oranı ve prevalans açısından farklı karakteristiklere sahip yapay veri setleri üretilmiştir. Her bir senaryonun 1000 kez tekrarı ile sınıflama performanslarına ilişkin ortalama değerler hesaplanmıştır. Çalışma sonucunda dengeli dağılım durumunda UAA yönteminin diğerlerinden daha iyi performansa sahip olduğu görülmüştür. Diğer taraftan RFerns yöntemi düşük gözlem sayısında diğer yöntemlere göre daha başarılı sonuçlar ortaya koymuştur. Tüm yöntemlerin dengesiz dağılım durumunda klasik veri yapılarında olduğu kadar yüksek başarı sağlamadığı da görülmüştür.

Anahtar kelimeler: Yüksek boyut düşük örneklem genişlikli düzenler, büyük p küçük n, sınıflama, uzaklık ağırlıklı ayırma, random ferns.

ABSTRACT

Aydoğan Culha, Ü. A Comparison of Performances of Classification Algorithms in High Dimension Low Sample Size Settings. Hacettepe University Institute of Health Sciences, Ph.D. Thesis in Biostatistics, ANKARA, 2015.

Nowadays, technologies that enable measuring thousands of features for each observation are very popular in many areas such as genetic, bioinformatics, medical imaging etc. This leads to different data settings which are called “High Dimension Low Sample Size (HDLSS) Settings”. There are some problems to analyze these settings with classical statistical methods. Performances and bias tendencies of various machine learning algorithms which have different advantages on high dimension are often ignored in case of small sample size. With this thesis, it is aimed to compare classification performances of Support Vector Machine (SVM), Distance Weighted Discrimination (DWD), Random Forest (RForest), Random Ferns (RFerns) and Naive Bayes (NB) algorithms in HDLSS settings. By simulation studies, artificial data sets that have different characteristics in terms of dimension, sample size, correlation structure, noise ratio and prevalence were generated. Each scenario was iterated for 1000 times and mean values for classification performances were calculated. In case of balanced distribution it was seen that DWD show better performance rather than other methods. On the other hand, for low sample size, RFerns revealed good performance. In case of imbalanced distribution, and for all methods, it was also seen that high success as was seen in classical data structure cannot be achieved.

Key Words: High dimension low sample size settings, large p small n, classification, distance weighted discrimination, random ferns.

İÇİNDEKİLER

ONAY SAYFASI	iii
TEŞEKKÜR	iv
ÖZET	v
ABSTRACT	vi
İÇİNDEKİLER	vii
SİMGELER-KISALTMALAR	viii
ŞEKİLLER	ix
TABLOLAR	xi
1. GİRİŞ	1
1.1. Çalışmanın Amacı ve Kapsamı	2
2. GENEL BİLGİLER	4
2.1. Yüksek Boyut Düşük Örneklem Genişlikli (YBDÖG) Düzenler	4
2.1.1. YBDÖG Düzenlerin Geometrik Gösterimi	8
2.2. Sınıflama ve Sınıflama Algoritmaları	10
2.2.1. Destek Vektör Makineleri	12
2.2.2. Uzaklık Ağırlıklı Ayırma	15
2.2.3. Naive Bayes	18
2.2.4. Random Forest	20
2.2.5. Random Ferns	22
2.3. Sınıflama Performanslarının Değerlendirilmesi	25
3. GEREÇ VE YÖNTEM	27
3.1. Benzetim Çalışması	27
3.1.1. Senaryolar	27
3.1.2. Veri Türetme Süreçleri	29
4. BULGULAR	32
5. TARTIŞMA	61
6. SONUÇ VE ÖNERİLER	63
KAYNAKLAR	65
EKLER	69
EK 1: Benzetim Çalışması Kod Örneği	
EK 2: Senaryolara İlişkin Duyarlılık, Seçicilik ve Matthews Korelasyon Katsayısı Değerleri	

SİMGELER-KISALTMALAR

AUAA	Ağırlıklı Uzaklık Ağırlıklı Ayırma
CRD	Centroid Rule Discriminator
DVM	Destek Vektör Makinesi
k-NN	k-En Yakın Komşuluğu
MLP	Çok Katmanlı Algılayıcı (Multi Layer Perceptron)
NB	Naive Bayes
NCI	Ulusal Kanser Ajansı (National Cancer Institute)
PCA	Temel Bileşenler Analizi (Principal Component Analysis)
RFerns	Random Ferns
RForest	Random Forest
SOM	Öz Örgütlemeli Ağ (Self Organizing Map)
UAA	Uzaklık Ağırlıklı Ayırma
YBDÖG	Yüksek Boyut Düşük Örneklem Genişlikli

ŞEKİLLER

	Sayfa
2.1. $n=3, d=2(a), d=20(b), d=200(c), d=20000(d)$ olan 10 örnek için saçılım grafiği	9
2.2. DVM'lerin Basit Gösterimi	13
2.3. UAA'nın Basit Gösterimi	16
2.4. Karar Ağacı Formunda Fern Yapısı ve Genel Karar Ağacı	23
3.1. Çalışma Kapsamında Kullanılan Senaryolar	29
4.1. Dengeli Dağılım ve Yüksek Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre DVM'nin Sınıflama Performansı	40
4.2. Dengeli Dağılım ve Yüksek Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre UAA'nın Sınıflama Performansı	41
4.3. Dengeli Dağılım ve Yüksek Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre RForest'in Sınıflama Performansı	42
4.4. Dengeli Dağılım ve Yüksek Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre RFerns'in Sınıflama Performansı	43
4.5. Dengeli Dağılım ve Yüksek Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre NB'nin Sınıflama Performansı	44
4.6. Dengeli Dağılım ve Orta Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre DVM'nin Sınıflama Performansı	44
4.7. Dengeli Dağılım ve Orta Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre UAA'nın Sınıflama Performansı	45
4.8. Dengeli Dağılım ve Orta Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre RForest'in Sınıflama Performansı	46
4.9. Dengeli Dağılım ve Orta Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre RFerns'in Sınıflama Performansı	46
4.10. Dengeli Dağılım ve Orta Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre NB'nin Sınıflama Performansı	47
4.11. Dengesiz Dağılım ve Yüksek Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre DVM'nin Sınıflama Performansı	54

4.12.	Dengesiz Dağılım ve Yüksek Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre RForest'in Sınıflama Performansı	55
4.13.	Dengesiz Dağılım ve Yüksek Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre RFerns'in Sınıflama Performansı	56
4.14.	Dengesiz Dağılım ve Yüksek Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre NB'nin Sınıflama Performansı	57
4.15.	Dengesiz Dağılım ve Orta Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre DVM'nin Sınıflama Performansı	58
4.16.	Dengesiz Dağılım ve Orta Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre RForest'in Sınıflama Performansı	58
4.17.	Dengesiz Dağılım ve Orta Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre RFerns'in Sınıflama Performansı	59
4.18.	Dengesiz Dağılım ve Orta Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre NB'nin Sınıflama Performansı	60

TABLÖLAR

	Sayfa
2.1. Olumsuzluk Tablosu	25
4.1. Dengeli Dağılım Durumunda Diğer Senaryolara Göre Doğruluk Değerleri	32
4.2. Dengeli Dağılım Durumunda Diğer Senaryolara Göre F Değerleri	36
4.3. Dengesiz Dağılım Durumunda Diğer Senaryolara Göre Doğruluk Değerleri	48
4.4. Dengesiz Dağılım Durumunda Diğer Senaryolara Göre F Değerleri	52



1. GİRİŞ

Teknolojinin her geçen gün büyük bir hızla ilerleyişine paralel olarak farklı alanlardaki farklı sorunların çözümleri için yeni araçlar geliştirilmektedir. Bu araçlarla elde edilen veriler ve bilgiler ise farklılaşmaktadır. Donoho'nun (1) 2000'li yılların başında ifade ettiği gibi bu yüzyıl tam anlamıyla veri yüzyılı olarak nitelendirilebilir. Örneğin bilgisayar teknolojilerinin gelişimine paralel olarak veri saklama teknolojilerinin/kapasitelerinin gelişmesi ile çok yüksek boyutlara ulaşan verilerin toplanması ve saklanması olanaklı hale gelmiştir. İnternet ortamı, elektronik kayıt sistemleri gibi ortamlarda her saniye yeni bilgiler kayıt altına alınmakta ve devasa boyutlarda veri havuzları oluşmaktadır. Bu ise araştırmacıları çok fazla bilgi içeren ancak bilginin yüksek boyut içinde gizlendiği işlenmemiş veri yığınları ile karşı karşıya bırakmaktadır. Bununla baş edebilmek için veri madenciliği algoritmaları gündeme gelmiş ve veri içindeki gizli bilgi ve örüntüleri ortaya çıkaracak algoritmalar geliştirilmiştir (2,3).

Benzer gelişmeler durmaksızın devam ederken son yıllarda çok farklı alanlarda binlerce niteliğin eş zamanlı ölçümünü yapabilen farklı cihazlar geliştirilmiştir. Bu konuda en bilinen örnek ise mikro dizilim gen ifade çalışmaları ve bu çalışmalarda kullanılan araçlardır. Bu çalışmalarda bir bireye ait binlerce hatta on binlerce genin ifade düzeyleri aynı anda ölçülmekte ve hastalıkla ilişkili genler saptanmaktadır. Bu ise birçok kompleks hastalık için tedaviler geliştirmeye olanak sağlamaktadır. Ancak genetik, tıbbi görüntüleme, proteomik, spektroskopi gibi birçok alanda karşımıza çıkan bu araçlarla analizlerin maliyetlerinin yüksek olması araştırmacıları gözlem sayısı konusunda sınırlandırmaktadır. Bu da gözlem sayısının değişken (boyut) sayısından çok daha düşük olduğu veri yapılarının ortaya çıkmasına sebep olmaktadır (4-6).

Yüksek Boyut Düşük Örneklem Genişlikli (YBDÖG) düzenler olarak adlandırılan ve önemi giderek artan bu yapıların analizinde ise bir takım sorunlarla karşılaşilmektedir. Bilindiği üzere klasik çok değişkenli istatistiksel yöntemler varsayımların sağlanamaması ve boyut sorunu (curse of dimensionality) nedeniyle değişken sayısının çok yüksek olduğu ya da gözlem sayısından daha yüksek olduğu durumlarda tercih edilmemektedir. Bununla birlikte, esasında yüksek boyut durumu için geliştirilmiş olan veri madenciliği ve makine öğrenmesi algoritmalarının

performanslarının, gözlem sayısının düşük olduğu durumlarda detaylı olarak incelenmediği gözlenmiştir. YBDÖG düzenlerin geometrik yapılarına ilişkin farklı yaklaşımlarla incelemeler yapılmış olsa da, özellikle sınıflama problemleri açısından bu veri yapılarının karakterlerindeki değişimin mevcut makine öğrenmesi algoritmalarının performansı üzerindeki etkilerini gösteren kapsamlı çalışmalara da rastlanamamıştır.

1.1. Çalışmanın Amacı ve Kapsamı

Bu tez çalışması ile YBDÖG düzenler detaylı olarak ele alınmakta ve bu veri yapılarında sınıflama problemlerine değinilmektedir. Makine öğrenmesi yöntemleri arasında yüksek performansları ile dikkat çeken Destek Vektör Makineleri (DVM), Random Forest (RForest) ve Naive Bayes (NB) sınıflayıcıları ile görece yeni ve güçlü sınıflayıcılar olan Uzaklık Ağırlıklı Ayırma (UAA) ve Random Ferns (RFerns) yöntemlerinin söz konusu düzenlerde sınıflama performanslarının karşılaştırılması amaçlanmıştır. Böylece farklı niteliklerdeki YBDÖG veri setleri için sınıflama performansı açısından en kullanışlı yöntemler belirlenmeye çalışılmıştır.

Tez çalışması altı bölümden oluşmaktadır. Birinci bölümde; YBDÖG düzenler özetlenmiş olup çalışmanın amacı ve kapsamı vurgulanmıştır.

İkinci bölümde; YBDÖG düzenler detaylı olarak incelenmiş, çalışma kapsamında daha önce yapılan çalışmalara ilişkin genel bilgiler verilmiş ve çalışmaya iten sebepler açıklanmıştır. Ayrıca çalışmada kullanılan sınıflama algoritmalarına ve sınıflama başarılarının değerlendirilmesinde kullanılan performans ölçütlerine ilişkin bilgiler bu bölümde özetlenmiştir.

Üçüncü bölümde; çalışmanın yöntemi detayları ile sunulmuştur. Amaç doğrultusunda oluşturulan senaryolara göre R yazılımından faydalanılarak benzetim teknikleri ile türetilen yapay veri setleri üzerinde söz konusu algoritmaların sınıflama performansları incelenmiştir.

Dördüncü bölümde; bulgular, tablo ve grafiklerle sunulmuştur. Yapay veri setleri üzerinde sınıflama algoritmalarının performansları raporlanmıştır.

Beşinci bölümde; elde edilen bulgular tartışılmış olup bulgular ışığında araştırmacılara hangi durumlarda hangi yöntemlerin daha tercih edilebilir olduğuna dair bilgiler verilmeye çalışılmıştır.

Son bölümde ise çalışma sonucunda elde edilen bilgiler özetlenmiş, çalışmayı ileriye taşıyabilecek ve literatüre katkı sağlayabileceği düşünülen öneriler getirilmiştir.



2. GENEL BİLGİLER

2.1. Yüksek Boyut Düşük Örneklem Genişlikli (YBDÖG) Düzenler

Hayatı kolaylaştırmak ve mevcut sorunları çözebilmek adına bilim ve teknoloji alanında her geçen gün yeni gelişmeler yaşanmaktadır. Özellikle ölümcül hastalıklar için ilaç ya da tedaviler geliştirilmesi ve yaşam kalitesini artıracak çözümler üretilmesi bilim insanlarının öncelikleri arasında yer almaktadır. Bununla birlikte teknoloji alanındaki birçok yeni gelişme yeni soruları ve yeni gelişmeleri de beraberinde getirmektedir.

Ulusal Kanser Ajansı'nın (NCI) tahminleri, 2012 yılında Dünya'da yeni gelişen kanser olgularının sayısının 14,1 milyon, kansere bağlı ölümlerin sayısının ise 8,2 milyon olduğunu göstermektedir (7). Bu sayılar, kanserin insanlar için ne derece büyük bir tehdit oluşturduğunu göstermek açısından oldukça önemlidir. Dolayısıyla çağın hastalığı olarak da adlandırılan kanserin nedenlerinin belirlenmesi ve tedavilerin geliştirilmesi son dönemlerde öncelikli alanlar arasına girmektedir. Kanserın çözümüne yönelik yeni tedavilerin ve ilaçların geliştirilmesi için bilim insanları her geçen gün farklı bakış açıları ortaya koymakta ve öneriler getirmektedirler. Örneğin daha önce genel olarak hastalığa sebep olan çevresel faktörler araştırılırken genetik alanındaki gelişmeler ile birlikte hastalığı etkileyebilecek genetik faktörler üzerinde artık daha çok durulduğu görülmektedir. Golub ve diğerlerinin (8) mikro dizilim gen ifade düzeylerinin, akut miyeloid lösemi ile akut lenfositik löseminin sınıflandırılmasında kullanılabileceğini gösterdiği çalışma ise bu açıdan önemli bir adım sayılabilir. Bu ve benzeri çalışmalar, mikro dizilim verileri ile geleneksel yöntemlere göre daha doğru tanıların elde edilebileceğini göstermiştir. Ardından kanserin tedavisi için gen ifade düzeylerine dayanarak tümörlü ve sağlam sınıfların belirlenmesi genetik alanında büyük önem kazanmıştır. Böylece bir kişiye ait binlerce genin ifade düzeylerinin eş zamanlı olarak ölçülebildiği mikro dizilim çalışmalarının popülerliği artmıştır. Ancak mikro dizilim çalışmalarının maliyetlerinin yüksek olması araştırmacıyı denek/gözlem sayısı konusunda sınırlandırmaktadır (9,10). Dolayısıyla mikro dizilim çalışmaları sonucunda değişken sayısının gözlem sayısından çok daha yüksek olduğu yüksek boyut düşük örneklem genişlikli veri yapıları ortaya çıkmaktadır. Bu da yukarıda değinildiği gibi farklı sorunları beraberinde getirmektedir.

Söz konusu veri yapıları sadece mikro dizilim çalışmalarına özgü bir durum olmaktan öte biyoinformatik, spektroskopi, proteomik, tıbbi görüntüleme gibi alanlarda da sıklıkla karşılaşılan ortak bir problem haline dönüşmüştür (11).

Tıbbi görüntüleme analizleri için örnek verilecek olursa; çeşitli organlara ya da dokulara ait görüntüleme sonuçlarının sayısal çıktıları genel olarak pikseller ya da vokseller (3 boyutlu görüntülemelerde 1 piksel karşılığı) ile ifade edildiğinden her bir görüntüden yüzlerce hatta binlerce nitelik ölçülebilmektedir (12). Ancak hem görüntülerin sayısallaştırılmasındaki zorluklar hem de yüksek maliyetli görüntüleme cihazları bu araştırmalarda da örnek sayısının sınırlandırılmasına neden olmaktadır.

Değişken sayısının çok yüksek, gözlem sayısının ise oldukça düşük olduğu bu veri yapıları literatürde “*high dimension low sample size data/settings*”, “*large p, small n*”, “*high dimensional, undersampled data*” gibi farklı şekillerde isimlendirilmektedir. Tez çalışması kapsamında bu veri yapıları “Yüksek Boyut Düşük Örneklem Genişlikli” (YBDÖG) düzenler olarak ifade edilmektedir.

Örnekleri giderek artan ve birçok disiplini ilgilendiren YBDÖG veri yapılarına özel yöntemler geliştirilmesi ya da var olan yöntemler ile problemlere çözümler aranması özellikle istatistik alanı açısından önem arz etmektedir. Çünkü klasik çok değişkenli istatistiksel yöntemler, genel olarak, iyi belirlenmiş az sayıda değişken ve değişken sayısından çok daha fazla gözlemin bulunduğu veri setlerinde çıkarsama yapmak için geliştirilmiş yöntemlerdir. Dolayısıyla YBDÖG veri setlerinin klasik yöntemler ile analizlerinde bir takım sorunlar bulunmaktadır. Örneğin çoğu klasik yaklaşım örnek kovaryans matrisini kullanmaktadır. Ancak YBDÖG düzenlerde yeterli veri olmaması nedeniyle kovaryans matrisinin güvenilir kestirimi elde edilememektedir. Diğer taraftan klasik çok değişkenli istatistiksel analizlerde temel varsayımlardan birisi küreselliktir. Küresellik, veri matrisi ile varyans kovaryans matrisinin tersinin karekökünün çarpımını gerektirmektedir. Ancak YBDÖG düzenlerde varyans kovaryans matrisi tam ranklı olmadığı için tersi alınamamaktadır. Dolayısıyla varsayımın sağlanamaması klasik yaklaşımları bu yapıların analizinde yetersiz kılmaktadır (13,14).

Klasik yaklaşımların yetersizliği araştırmacıları boyut sorunu (curse dimensionality) ile baş etme konusunda daha başarılı olan makine öğrenmesi algoritmalarını kullanmaya yönlendirmektedir. Daha önce de bahsedildiği gibi son

dönemlerde veri madenciliği ve makine öğrenmesi kavramları veri analizinde büyük önem kazanmıştır. Veri madenciliği yöntemleri özellikle artan boyut sorununun çözümünde önemli katkılar sağlamıştır. Bu yöntemler eğitim verisi üzerinde öğrenme yaparak yeni gözlemler için genelleme yapmaktadırlar. Diğer bir deyişle yüksek boyutlu verilerde gizli kalmış bilgileri ya da örüntüleri açığa çıkarmakla ilgilenmektedirler. Ancak bu yöntemler de genel olarak gözlem sayısının değişken sayısından daha büyük olduğu durumlar için geliştirilmiş ve bu tip veriler üzerinde yoğun olarak kullanılmıştır. YBDÖG düzenler ile daha sık karşılaşmaya başladıkça bu yöntemlerin YBDÖG veri setlerinde uygulanabilirliği ve performansları da merak konusu olmaya başlamış ve önem kazanmıştır.

Literatürde, özellikle mikro dizilim çalışmalarından elde edilen veriler üzerinde sınıflama ve kümeleme gibi analizler yapmak amacıyla makine öğrenmesi ve veri madenciliği algoritmalarının sıklıkla uygulandığı gözlenmektedir. Özellikle ortak kullanıma açık gerçek veri setleri üzerinde farklı boyut indirgeme teknikleri ile değişken seçimi yaparak ya da tüm veri setini kullanarak farklı sınıflayıcıların performanslarının karşılaştırıldığı çalışmalar bulunmaktadır.

Dudoit ve diğ. (15) gen ifade profillerine dayanarak tümör sınıflama çalışmalarında sınıflama yöntemlerini karşılaştırmışlardır. Çalışmada üç gerçek veri seti kullanılmış olup lenfoma verisinde 4682 gen ve 81 örnek, lösemi verisinde 6817 gen ve 72 örnek, NCI (National Cancer Institute) 60 verisinde 9703 gen ve 61 örnek için gen ifade düzeyleri ile çalışılmıştır. Bu çalışma sonucunda en yakın komşuluk ve diyagonal doğrusal ayırma analizi gibi yöntemlerin, karar ağaçları gibi daha karmaşık sınıflayıcılardan daha iyi performans gösterdiği ifade edilmiştir.

Cho ve Won (16) ise çalışmalarında, lösemi, kolon kanseri ve lenfoma veri setlerinde farklı değişken seçim yöntemleri kullanarak çok katmanlı algılayıcı (*multi layer perceptron-MLP*), k-en yakın komşuluğu (k-NN), DVM, öz örgütlemeli ağ (*Self Organization Map-SOM*) yöntemlerini karşılaştırmışlardır. Çalışma sonucunda bilgi kazancı (*Information Gain*) ve Pearson korelasyon katsayısı en başarılı değişken seçim yöntemleri, MLP ve k-NN ise en başarılı sınıflayıcılar olarak bulunmuştur.

Bu çalışmaların örneklerini artırmak mümkündür. Örneğin Furey ve diğ. (17) 5 farklı mikro dizilim veri setini kullanarak DVM'lerin kanser olgularını sınıflamada

başarılı sonuçlar elde ettiğini göstermişlerdir. Pirooznia ve diğ. (18) ise 8 serbest erişimli mikro dizilim verisi üzerinde hem değişken seçimi hem de sınıflama algoritmalarının birlikte performanslarını değerlendirmişlerdir. DVM, yapay sinir ağları, RForest, karar ağaçları ve bayesçi sınıflayıcıları kullandıkları çalışmada en iyi performansa sahip değişken seçim yöntemi ve sınıflama yöntemi kombinasyonunun her veri seti için farklılaştığını gözlemişlerdir. Bununla birlikte değişken seçiminin yeni gözlemleri sınıflama başarısını olumlu etkilediğini öne sürmüşlerdir. Wang ve arkadaşlarının (19) makine öğrenmesi yaklaşımları ile kanser sınıflama çalışmasında ise; karar ağaçları, DVM ve NB sınıflayıcılarının seçilen az sayıda gen ile yüksek doğrulukta sınıflama yaptıkları gösterilmiştir. Bununla birlikte araştırmacılar mikro dizilim veri setlerinin doğası gereği fazla gürültüye sahip olması nedeniyle, seçilen genlerin iyi sınıflayıcı oluşturmak için önerilmesinde tereddütleri olduğunu da ifade etmişlerdir.

Genel kullanıma açık verilerin sınırlı sayıda olması nedeniyle -yukarıda birkaç örneğine değinilen- bu çalışmaların çoğunda aynı veri setlerinin kullanıldığı görülmektedir. Bununla birlikte, farklı çalışmalardan elde edilmiş söz konusu veri setlerinin farklı karakterlerde olduğu göz önüne alındığında, bu çalışmalar ile YBDÖG veri setlerinde makine öğrenmesi algoritmalarının sınıflama performansları ile ilgili genel bir kanıya ulaşılamayacağı düşünülmektedir.

Söz konusu çalışmalarda dikkat çeken bir diğer nokta ise bahsi geçen gerçek veri uygulamalarının çoğunda verilerin önce boyut indirgeme işlemine tabii tutulmuş olmalarıdır. Kimi çalışmalarda korelasyon yapılarına dayanarak kimi çalışmalarda ise temel bileşenler analizi (PCA) gibi yöntemler kullanılarak indirgenmiş veriler üzerinde sınıflama yapılmaktadır. Literatürde bu yöntemlerin çok kullanıldığı görülmekle birlikte YBDÖG veri setlerinde boyut indirgeme işlemleri için temel bileşenler analizi (PCA) gibi yöntemlerin doğrudan uygulanamayacağını belirten çalışmalar da bulunmaktadır. Özellikle bu verilerin geometrik yapıları üzerinde yapılan ve aşağıda detaylarına da değinilecek olan çalışmalar sonucunda YBDÖG veri yapılarının klasik yapılardan farklı olmasının dikkate alınmadığı belirtilmektedir (13,20). Buna karşın uygulamada sağladığı kolaylık ve başarılar nedeniyle de kullanılmaya devam ettiği görülmektedir. Xing ve diğ. (10), amaç küçük değişken setleri için hesaplama ve yorumlama kolaylığı sağlamak olduğunda değişken seçimi

yöntemlerinin; amaç daha iyi bir kestirici elde etmek olduğunda ise kovaryans matrisine ilişkin düzeltme tekniklerinin kullanılması gerektiğini ifade etmektedirler.

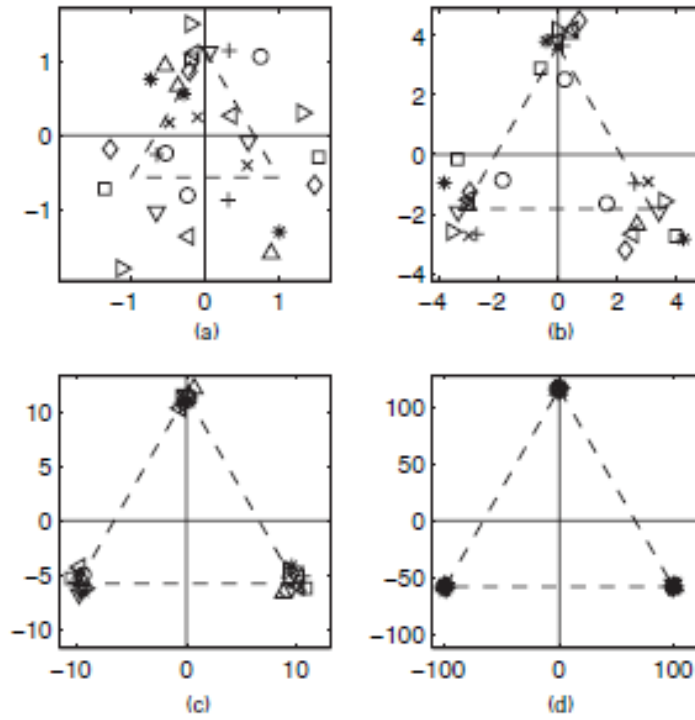
Kovaryans matrisinden kaynaklanan sorunların çözümü konusunda da farklı ayırma yöntemleri ya da düzeltme teknikleri üzerinde sıklıkla durulmaktadır. Dudoit ve diğ. (15) “diyagonal doğrusal ayırma” analizi adını verdikleri örnek kovaryans matrisinin yalnızca köşegen elemanlarını kullanan teknik ile sınıflama yapmayı önermişlerdir. Bickel ve Levina (21) da “Naive Bayes” yaklaşımı ile değişkenlerin bağımsız olduğu varsayılarak, kovaryans matrisinin tüm elemanları kullanılmaksızın, başarılı sonuçlar elde edildiğini göstermişlerdir.

2.1.1. YBDÖG Düzenlerin Geometrik Gösterimi

Verilerin görselleştirilmesi ve yöntemlerin geçerliliklerinin değerlendirilmesi için literatürde genellikle farklı asimptotik yaklaşımların kullanıldığı görülmektedir. Bu çalışmaların çoğunu gözlem sayısı sonsuza giderken ($n \rightarrow \infty$) değişken sayısının sabit (d sabit) tutulduğu çalışmalar oluşturmaktadır. Bununla birlikte hem gözlem hem de değişken sayısının birlikte arttığı yaklaşımlar da mevcuttur.

Son dönemlerde ise YBDÖG verilerin analizinde farklı alternatifler geliştirebilmek adına bu verilerin geometrik yapılarını incelemek üzere tasarlanan çalışmaların sayısı giderek artmaktadır. Bu çalışmaların çoğunu da YBDÖG düzenlerde örnek kovaryans matrisinin özellikleri ile ilgilenen farklı asimptotik çalışmalar oluşturmaktadır.

Hall’un (13) gözlem sayısı (n) sabit iken değişken sayısının (boyut, d) sonsuza gitmesi yaklaşımını kullanarak YBDÖG veri setlerine farklı bir bakış açısı getiren çalışması, hem bu verilerin görselleştirilmesi anlamında hem de sonuçları açısından dikkat çekmektedir. ‘ d -asimptotik’ olarak adlandırılan bu yaklaşımla söz konusu veri yapılarında geometrik yapının beklenenden farklı olduğu bildirilmektedir (13,14). $d=2$, $d=20$, $d=200$ ve $d=20000$ boyutları için 10 farklı örneğe ait $n=3$ gözlemin dağılım yapılarının yer aldığı Şekil 2.1’de bu durum açıkça görülmektedir. Görüldüğü üzere boyut arttıkça gözlemler arasındaki uzaklıklar benzer olmakta ve her bir örneğe ilişkin gözlemler düzgün simpleksin köşelerinde yığılma eğilimi göstermektedirler. Ayrıca çok yüksek boyutlarda sıfır ortalamalı Gauss dağılımına sahip rasgele örneklemelerin popülasyon ortalamasına nadiren yaklaştığı da ifade edilmektedir (22).



Şekil 2.1. $n=3$, $d=2$ (a), 20(b), 200(c) ve 20000(d) olan 10 örnek için saçılım grafiği. Hall (13)'dan alınmıştır.

Marron ve diğ. (23) ise özel olarak YBDÖG düzenlerde DVM'lerin genelleme performanslarını olumsuz etkileyen problemlerin varlığını bu asimptotik yaklaşımlardan faydalanarak göstermişlerdir. Bu sorunu çözmek adına önerdikleri yeni sınıflama yöntemi olan UAA yöntemini ise yapay veri setleri üzerinde DVM'ler ile kıyaslayarak başarısını ortaya koymuşlardır. Söz konusu çalışmada verilerin yalnızca gözlem sayısı ve boyut sayısı dikkate alınarak oluşturulması ve veri yapısını etkileyebilecek diğer faktörlerin olmaması ise bir eksiklik olarak düşünülmektedir. Dolayısıyla tez çalışması kapsamında özel olarak YBDÖG düzenler için geliştirilmiş UAA sınıflayıcısının diğer makine öğrenmesi algoritmalarına göre sınıflama performanslarının çok yönlü olarak ele alınması ve değerlendirilmesi de önem arz etmektedir.

Diğer taraftan Hall (13) da benzetim teknikleri ile üretilen farklı veri setleri üzerinde DVM, UAA, en yakın komşuluk ve ortalama farkları olarak da adlandırılan Centroid Rule Discriminator (CRD) sınıflayıcılarının performanslarını karşılaştırmıştır. Belirlenen ortalama ve standart sapma değerleri için Gauss dağılımından türetilen veri setlerinde dağılım yapısı, gözlem ve değişken sayılarına

göre senaryolar oluşturulmuş ve ortalama hata değerleri elde edilmiştir. Sıfır ortalama ve özdeş kovaryans matris için Gauss dağılımından türetilen dengeli dağılıma sahip veri setinde boyut arttıkça DVM, UAA ve CRD yöntemlerinin hata oranlarının birbirine yakınsadığı gözlenmiş, en yakın komşuluk algoritmasının performansının ise çok geride kaldığı görülmüştür. Ortalama ve standart sapmaların değiştirildiği diğer bir senaryoda ise boyut arttıkça ve yüksek ortalama değerleri için DVM'lerin hatalarının sıfıra yakınsadığı bildirilmiştir. Aynı senaryolar için UAA'nın da benzer performansı gösterdiği, düşük boyutlarda ise daha başarılı olduğu belirtilmiştir. Sınıf dağılımının dengesizliği dikkate alındığında ise UAA yönteminin başarısızlığı ifade edilmiştir (13).

Kısaca özetlemek gerekirse serbest erişimli gerçek veri setlerinin kısıtlı olması ve genellikle benzetim teknikleri uygulanan çalışmalarda yalnızca boyut, gözlem sayısı ve prevalansın dikkate alınması literatürde bu konuda bir boşluk olduğunu düşündürmektedir. Aynı zamanda boyut indirgeme ya da değişken seçimi ile az sayıda değişkenle çalışılmasında literatürde ifade edilen teorik olumsuzluklar ve bilgi kaybı endişesi, çalışmaların tüm değişkenleri içerecek şekilde planlanmasının bu aşamada daha uygun olduğu kanaatini oluşturmaktadır. Bu nedenlerle bu tez çalışması ile boyut, gözlem sayısı ve prevalans değerlerinin yanı sıra özellikle YBDÖG verilerde önemli olan gürültü değişkenlerinin oranları, hem bağımlı değişken ile bağımsız değişkenler arasındaki hem de bağımsız değişkenlerin kendi aralarındaki korelasyon yapıları da dikkate alınarak karşılaşılabilecek veri yapılarına ilişkin bir çok senaryo geliştirilmiştir. Aynı zamanda bu tez çalışması kapsamında doğrudan veri seti kullanılarak -değişken seçimi gibi herhangi bir ön işlem olmaksızın- farklı karakterlerdeki YBDÖG veri setleri üzerinde en etkili sınıflama yöntemleri belirlenmeye çalışılmıştır. Böylece, araştırmacılara, mevcut tüm bilgileri kullanarak pratikte en hızlı ve en doğru şekilde tahminler elde etmelerinde fayda sağlayacak öneriler sunulması planlanmıştır.

2.2. Sınıflama ve Sınıflama Algoritmaları

Sınıflama, benzer nesnelerin gruplandırılarak verilerin sadeleştirilmesi ve yeni gözlemler geldiğinde bunların benzer nesnelerin bulunduğu gruplara atanması için kestirim yapmakta kullanılan oldukça faydalı istatistiksel bir araçtır. Sınıflama işleminde, belirli nitelikleri olan ve belirli sınıflarda yer alan gözlemlere dayanarak

bir karar kuralı oluşturulmaktadır. Bu kurala dayanarak da belirli nitelikleri olan ancak sınıfı belli olmayan yeni nesnelerin hangi sınıfta yer alacağı tahmin edilebilmektedir.

Literatürde çok sayıda sınıflama yöntemi bulunmaktadır. Bunların bazıları klasik istatistiksel yaklaşımları içermekte bazıları ise makine öğrenmesi temeline dayanmaktadır. Klasik istatistiksel yaklaşımlar dağılıma ilişkin varsayımlar içermekteyken birçok makine öğrenmesi algoritması ile sınıflama yaparken bu gereklilik ortadan kalkmaktadır. Özellikle YBDÖG düzenlerde varsayımsal sorunlardan kaynaklanan problemler klasik yaklaşımların doğrudan uygulanmasını olanaksız kılmaktadır. Makine öğrenmesi algoritmalarının ise bu düzenlerde doğrudan uygulanabileceği belirtilmekte ve çok sayıda uygulaması ile karşılaşılmaktadır (24). Ancak daha önce bahsedildiği gibi bu yöntemlerin YBDÖG düzenlerde sınıflama performansları kapsamlı olarak ele alınmamıştır.

Tez kapsamında beş farklı sınıflama yöntemi dikkate alınmıştır. Bunlar Destek Vektör Makineleri (DVM), Uzaklık Ağırlıklı Ayırma (UAA), Random Forest (RForest), Random Ferns (RFerns) ve Naive Bayes (NB) algoritmalarıdır. Çalışma kapsamında bu yöntemlerin dikkate alınmasının bazı gerekçeleri bulunmaktadır. DVM ve RForest yöntemleri birçok alanda uygulaması bulunan güçlü sınıflayıcılardır ancak YBDÖG düzenlerde karşılaştıkları sorunlar ve farklı karakterlere sahip YBDÖG düzenlerdeki performansları yeteri kadar incelenmemiştir. NB sınıflayıcılar, özellikle düşük örneklem genişliğinde yüksek performansları ile bilinmektedir. Ancak gözlem sayısı az iken boyut arttıkça performansındaki değişime ilişkin açıklayıcı çalışmalara rastlanamamıştır. UAA ise özellikle YBDÖG düzenler için geliştirilmiş olmakla birlikte veri yapısından ne şekilde etkilendiği konusu çok net değildir. Ayrıca DVM'lerin bir uyarlaması olduğu için genellikle performans karşılaştırması bu iki yöntem arasında yapılmaktadır. RFerns ise diğer yöntemlere göre oldukça yeni olduğundan ve genel anlamda mühendislik çalışmaları için tasarlandığından sınıflama problemleri için makine öğrenmesi yöntemleri arasında henüz çok bilinmemektedir. YBDÖG düzenlerde RFerns yönteminin sınıflama performansına ilişkin herhangi bir çalışma da bulunmamaktadır.

Aşağıda tez çalışmasında kullanılacak olan algoritmalara ilişkin detaylı bilgiler yer almaktadır.

2.2.1. Destek Vektör Makineleri

Destek vektör makineleri (DVM) ilk olarak Vapnik tarafından 1960'larda ortaya atılmış ancak 90'lı yıllara kadar çok dikkat çekememiştir. Son dönemlerde ise yüksek performansı ve doğrusal olmayan veri yapıları ile baş edebilme özelliğinden dolayı sıklıkla kullanılan güçlü bir sınıflama yöntemidir. Başlangıçta doğrusal ayrılabilen ve iki sınıflı sınıflama problemlerinin çözümü için önerilmiş olsa da doğrusal olarak ayrılamayan veri yapılarında ve çok sınıflı sınıflama problemlerinde performansı ile öne çıkmaktadır (25-27).

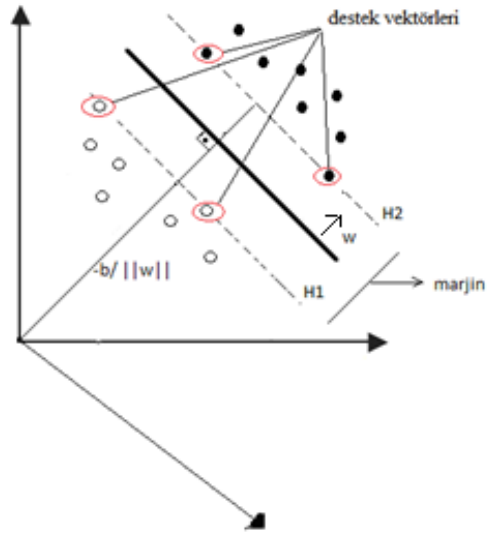
Yöntemin asıl amacı yeni veri geldiğinde sınıflama hatasını minimuma indirmektir. Doğrusal olarak ayrılabilen verilerde, farklı sınıflardaki birbirine en yakın noktaların birbirine olan uzaklığını maksimize edecek ayırıcı doğruyu bulmaya çalışmaktadır. Söz konusu noktalar destek vektörleri olarak adlandırılmaktadır. Doğrusal olarak ayrılmayan yapılarda ise girdi verisini (input data) haritalama (mapping) teknikleriyle daha yüksek boyuta taşımakta ve verinin taşındığı yeni boyutta sınıflar arası uzaklığı maksimum yapacak ayırıcı düzlemi araştırmaktadır (28).

DVM'ler sınıflama yapabilmek için karar fonksiyonları oluşturmaktadır. Doğrusal olarak ayrılabilen iki sınıflı bir sınıflama problemi için, eğitim seti $\{x_i, y_i\}$ $i=1, \dots, n$ ve sınıf etiketleri $y_i \in \{-1, +1\}$ olsun. Buna göre ayırıcı doğru,

$$\langle w, x \rangle + b = 0 \quad (2.1)$$

şeklinde gösterilir. Eşitlikte yer alan w ayırıcı doğrunun normal vektörü, b ise yanlılık terimidir.

Şekil 2.2'de görüldüğü gibi ayırıcı doğruya eşit uzaklıkta ve paralel olan doğrular karar sınırlarını oluşturmaktadır. Destek vektörleri bu doğrular üzerinde yer almaktadır ve bu doğrular marjin sınırları olarak ifade edilmektedir. Yöntem optimizasyon teknikleri kullanarak tekrarlı denemelerle bu iki doğru arasındaki uzaklığı en büyük yapacak ayırıcı doğruyu bulmaya çalışmaktadır. Bu uzaklığa "marjin" adı verilmektedir.



Şekil 2.2. DVM'lerin basit gösterimi.

Aydoğan (28)'dan alınmıştır.

Marjin sınırlarına ilişkin eşitlikler aşağıdaki gibidir:

$$\langle w, x_1 \rangle + b = +1, \quad y_i = +1 \text{ için} \quad (2.2)$$

$$\langle w, x_2 \rangle + b = -1 \quad y_i = -1 \text{ için} \quad (2.3)$$

Destek vektörlerinin ayırıcı doğruya olan uzaklığı $1/\|w\|$ 'dir. Marjin değeri ise her iki sınıftaki destek vektörlerinin ayırıcı doğruya olan uzaklıkları toplamı olacağından $2/\|w\|$ değerine eşittir. $\|w\|$, w 'nin Öklid normudur. En iyi ayırıcı düzlemi bulmak için optimizasyon problemi,

$$\min \tau(w) = \frac{1}{2} \|w\|^2 \quad (2.4)$$

$$y_i(\langle x_i, w \rangle + b) - 1 \geq 0, \quad \forall i \text{ için} \quad (2.5)$$

şeklinde kurulur. Yukarıdaki problemlerin çözümü Lagrange çarpanları yöntemini gerektirmektedir. Optimizasyon problemi lagrange çarpanları ile çözüldükten sonra sınıflama hatasını en aza indirecek ayırıcı düzlemi bulacak karar fonksiyonu aşağıdaki forma dönüşür:

$$f(x) = \sum(\alpha_i y_i x_i \cdot x) + b = 0 \quad (2.6)$$

Eşitlikte yer alan α_i 'ler lagrange çarpanlarıdır.

Yukarıda sözü edilen yaklaşım sert marjın yaklaşımı olarak adlandırılır ve verilerin tamamının doğrusal olarak ayrılabilirdiği durumlar için önerilmiştir. Gerçek hayatta verilerin tamamının doğrusal olarak ayrılabilirdiği durumlar nadir görülmektedir. Özellikle gürültü verilerinin varlığı nedeniyle bu durum bozulabilmektedir. Bu gibi durumlarda DVM'ler soft marjın yaklaşımını kullanmakta ve deneme hatalarını tolere edebilecek karar fonksiyonları oluşturmaktadır. Soft marjın yaklaşımında marjın sınırlarına gevşek (slack) değişken ($\xi_i > 0$) eklenir ve nesne fonksiyonuna hata için bir maliyet parametresi (C) eklenir. Bu parametre marjın maksimizasyonu ile deneme hatası arasındaki ödünleşimi ifade eder. Bu durumda marjın sınırları ve nesne fonksiyonu aşağıdaki forma dönüşür:

$$\langle w, x_1 \rangle + b \geq +1 - \xi_i, \quad y_i = +1 \text{ için} \quad (2.7)$$

$$\langle w, x_2 \rangle + b \leq -1 - \xi_i, \quad y_i = -1 \text{ için} \quad (2.8)$$

$$\min \tau(w, \xi) = \frac{1}{2} \|w\|^2 + C (\sum_i \xi_i)^k \quad (2.9)$$

Verilerin doğrusal olarak ayrılamadığı durumlarda da DVM'lerin uygun çözümü bulunmaktadır. Söz konusu durumlarda DVM'ler doğrusal olmayan haritalama yaparak verileri daha yüksek boyuta sahip olay uzayına taşımaktadır.

$$x \in R^n \rightarrow \Phi(x) \in R^f \quad (2.10)$$

Dönüştürülmüş yeni olay uzayında karar fonksiyonu aşağıdaki gibidir.

$$\langle w, \Phi(x) \rangle + b = 0 \quad (2.11)$$

Eşitlikte görüldüğü üzere doğrusal ayrılabilir durumlarda olduğu gibi doğrudan girdi verisi kullanmak yerine burada haritalama fonksiyonu ($\Phi(x)$) işleme dahil edilmektedir. Buna göre marjın sınırları ve nesne fonksiyonu da aşağıdaki forma dönüşmektedir.

$$\langle w, \Phi(x) \rangle + b \geq +1 \quad (2.12)$$

$$\langle w, \Phi(x) \rangle + b \leq -1 \quad (2.13)$$

$$\min_{w \in H, b \in R} \tau(w) = \frac{1}{2} \|w\|^2 \quad (2.14)$$

$$y_i(\langle \Phi(x), w \rangle + b) - 1 \geq 0, \quad \forall i \text{ için} \quad (2.15)$$

Burada kullanılacak haritalama fonksiyonunun belli olmaması ve yüksek boyutta işlem yapmanın zor olması nedeni ile çekirdek düzenlemesi (kernel trick) adı verilen düzenleme uygulanmaktadır. Böylelikle dönüştürülmüş uzaydaki haritalama fonksiyonu yerine doğrudan girdi uzayındaki verilerin kullanımına imkan veren çekirdek fonksiyonlar işleme dahil edilmektedir. Literatürde çok sayıda çekirdek fonksiyon bulunmakla birlikte en yaygın kullanılan fonksiyonlar; doğrusal fonksiyon, polinomial fonksiyon, sigmoid fonksiyon ve radyal tabanlı fonksiyonlardır.

Yukarıda genel özellikleri anlatılan DVM'lerin uygulamada çok sayıda avantajı bulunmaktadır. Bunlardan en önemlileri,

- Yüksek sınıflama performansı
- Doğrusal olarak ayıramayan veri yapılarında kullanılabilme
- Yüksek boyutta çalışabilme

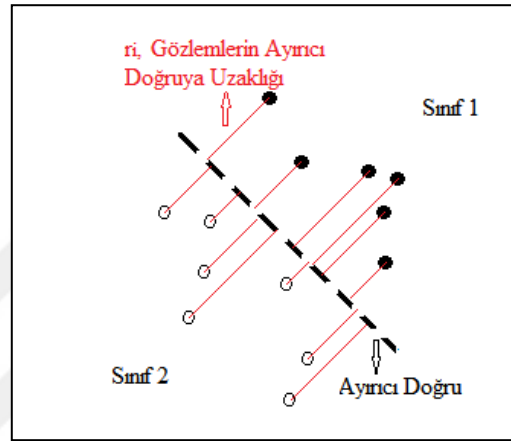
şeklinde sıralanabilir. Yöntem mikro dizilim verilerinin analizi, el yazısı tanıma, nesne tanıma ve görüntü tanıma gibi çalışmalarda kullanılmaktadır.

2.2.2. Uzaklık Ağırlıklı Ayırma

Uzaklık Ağırlıklı Ayırma (UAA) Marron ve diğ. tarafından özellikle YBDÖG düzenler dikkate alınarak geliştirilmiş görece yeni bir sınıflama yöntemidir (23). Yöntemin temelleri 2002 yılında atılmış olsa da 2007 yılında orijinal halini almıştır.

YBDÖG düzenlerde sınıflama problemlerine farklı bir açıdan bakmayı amaçlayan Marron ve diğ. DVM'lerde ayırıcı doğrunun normal vektöründe projeksiyon yapmışlardır. Bunun sonucunda ise projeksiyonların aynı olduğu ve veri noktalarının marjın sınırlarında yığılma eğiliminde olduğu gözlenmiştir. Veri yığılması (data piling) adını verdikleri bu özellik ise genelleme performansında

kayıplara neden olmaktadır. Çünkü destek vektörlerinin etkisi ile bulunan optimum ayırıcı düzlem bu yığılmadan etkilenmektedir ve aşırı uyum nedeniyle yeni verilerin yanlış sınıflandırılmasına yol açabilmektedir. Diğer taraftan yüksek boyutta destek vektörlerinin sayısının artma eğilimi de dikkate alındığında, marjin üzerinde yığılan bu noktaların ayırıcı düzlem belirlemede yaratacağı etki öngörülebilmektedir. UAA yönteminin geliştirilmesindeki asıl motivasyon da bahsi geçen genelleme performansındaki kayıpları önlemektir (14,29).



Şekil 2.3. UAA'nın basit gösterimi.

UAA, DVM'ler gibi geniş marjin tabanlıdır ve sınıflama işlemini karar fonksiyonu oluşturarak yapmaktadır. Ancak karar fonksiyonu oluştururken DVM'lerden farklı olarak, yalnızca destek vektörlerini dikkate almak yerine tüm veri noktalarını kullanmaktadır (29). Bunun için her bir veri noktasının ayırıcı doğruya olan uzaklıklarının terslerinin toplamını minimum yapacak ayırıcı doğruyu araştırır. Yani sınıflar arasındaki minimum uzaklıklar yerine ortalama uzaklığı büyültmeye çalışmaktadır. Böylece ayırıcı doğrunun belirlenmesinde her bir noktanın etkisi dikkate alınmaktadır (Şekil 2.3.).

UAA'nın DVM'den diğer bir farkı ise kullandığı optimizasyon problemidir. DVM'lerde kuadratik programlama probleminin çözümü gerekirken, UAA ikinci dereceden koni programlama probleminin çözümünü gerektirmektedir (30). Bu ise DVM'ye göre hesaplama anlamında zorluk içermektedir.

$\{x_i, y_i\}$ 'lerden oluşan ve sınıf etiketleri $y_i \in \{-1, +1\}$ olan eğitim seti dikkate alındığında, i. veri vektörünün ayırıcı düzleme olan uzaklığı:

$$r_i = y_i(x_i^T + \beta) + \xi_i \quad (2.16)$$

şeklinde ifade edilir. Tüm noktaların doğrusal olarak ayrılması her zaman mümkün olmayacağı için burada da DVM'ye benzer şekilde gevşek (slack) değişken (ξ) işleme dahil edilmektedir.

UAA probleminin çözümü, bu uzaklıkların tersleri toplamını minimize edecek, diğer bir deyişle noktaların ayırıcı düzleme uzaklıklarının harmonik ortalamasını maksimize edecek ayırıcı düzlemi bulmakla ilgilenmektedir.

$$\min \sum_i \frac{1}{r_i} + C e' \xi \quad (2.17)$$

$$r = YX'w + \beta y + \xi \geq 0 \quad (2.18)$$

$$\|w\|^2 \leq 1, \quad \xi \geq 0 \quad (2.19)$$

w ayırıcı doğrunun normal vektörüdür ve $\beta \in \mathbb{R}$ ise bu doğrunun pozisyonunu belirlemektedir. Sınıflar arasındaki açıklık DVM'lerde $2/\|w\|$ iken UAA'da eşitlikten de anlaşılacağı üzere

$$\sum_i^n \frac{1}{r_i} \quad (2.20)$$

formuna dönüşmektedir.

Yöntem başlangıçta doğrusal ve iki sınıflı sınıflama problemleri için geliştirilmiş olmakla birlikte çok sınıflı sınıflama problemleri için genişletilmiş formları da mevcuttur (29). UAA, YBDÖG düzenlerde sınıflama amacı ile kullanımı dışında özellikle mikro dizilim çalışmalarda biyolojik olmayan deneysel bazı sistematik yanlışlıkların düzeltilmesi için de önerilmektedir. Örnek toplama süreçlerinin uzunluğu, analizlerin farklı merkezlerde yapılması, farklı mikro dizilim baskı tablalarının (print batches) kullanılması ya da farklı hibridleme protokollerinin uygulanması gibi çeşitli sistematik yanlışlık kaynakları bulunmaktadır. Benito ve diğ.

(2004) çalışmalarında UAA algoritmasının söz konusu hataların düzeltilmesinde de etkili olduğunu göstermişlerdir (31,32).

UAA yönteminin en önemli avantajı daha önce de belirtildiği gibi YBDÖG düzenlerde sağladığı başarıdır. Ancak sınıfların örneklem genişliklerinde dengesizlik söz konusu olduğunda UAA'nın düşük performans gösterdiği belirtilmektedir (33). Sınıflardan herhangi biri diğerinden daha büyük olduğunda standart UAA yaklaşımında baskın (büyük örnekleme sahip) olan sınıf karar sınırını küçük sınıfa doğru kaydırmakta, bu ise yeni gelen verilerin genellikle baskın sınıfa atanmasına neden olmaktadır (34). Qiao ve diğ. (2010) bu sorunu çözmek üzere "Ağırlıklı Uzaklık Ağırlıklı Ayırma" (AUAA) yöntemini önermişlerdir. AUAA'da büyük olan sınıfa küçük, küçük olan sınıfa büyük ağırlık verilerek en uygun çözüme ulaşmak amaçlanmaktadır (34,35).

2.2.3. Naive Bayes

Naive Bayes (NB), Bayes teoremine dayanan ve veri madenciliği uygulamaları içinde etkili bir sınıflama yöntemidir. Bayes teoremindeki hesaplama zorlukları, değişkenlerin bağımsızlığı varsayımı ile sadeleştirilmekte böylece NB kolay uygulanabilir bir forma dönüşmektedir. Bu varsayım her ne kadar gerçek uygulamalarda karşılaşılan bir durum olmasa da, yöntemin uygulamalarda gösterdiği yüksek performans dikkat çekmektedir (36). NB ile her değişkenin önemi ayrı ayrı dikkate alınmakta ve sonuç üzerindeki etkileri olasılıksal olarak hesaplanmaktadır.

NB sınıflayıcılar değişkenlerin bağımsızlığı varsayımına dayandığı için kovaryans matrisine ihtiyaç duymamakta, değişken varyansları yeterli olmaktadır. Bu da küçük eğitim setleri ile çalışabilmesine olanak sağlamaktadır.

NB sınıflayıcıların çalışma prensiplerini incelemek için $X = (X_1, X_2, \dots, X_n)$ seti ele alınsın ve hedef değişken Y olsun. Bu durumda NB sınıflayıcı,

$$P(Y = y_i | X = x) \quad (2.21)$$

koşullu olasılığını hesaplamaya çalışmakta ve bu olasılığı yüksek olan sınıfa atama yapmayı hedeflemektedir. Bayes teoremine dayanarak söz konusu olasılık,

$$P(y_i | x) = \frac{P(x|y_i)P(y_i)}{P(x)} \quad (2.22)$$

formuna dönüşmektedir. NB sınıflayıcıların temel varsayımı olan değişkenlerin bağımsızlığı bu formüle eklendiğinde ise aşağıdaki eşitlik elde edilmektedir.

$$P(y_i|x) = \frac{P(x_1|y_i)P(x_2|y_i)...P(x_n|y_i)P(y_i)}{P(x_1,x_2,...,x_n)} \quad (2.23)$$

Burada paydadaki terim sabit olduğundan ihmal edilmekte ve nihai sonsal olasılık aşağıdaki şekli almaktadır.

$$P(y_i|x) \propto P(x_1|y_i)P(x_2|y_i) ... P(x_n|y_i)P(y_i) \quad (2.24)$$

Sınıflama problemleri için en yüksek olasılığa sahip sınıfa atama yapmak hedeflendiğinden NB sınıflama kuralı için karar kuralı:

$$\operatorname{argmax} P(Y = y_i) \prod_i P(X_i|Y = y) \quad (2.25)$$

şeklinde oluşturulmaktadır. Buradan anlaşılacağı üzere bağımsızlık koşulu olmaksızın hesaplama yapıldığında $2(2^n-1)$ yerine $2n$ tane parametreye ihtiyaç duyulur. Bu da yöntemin basit olması avantajını sağlamaktadır.

Kategorik değişkenler ile sınıflama yapmaya elverişli şekilde tasarlanan NB algoritması sürekli ve kesikli sayısal verilerle sınıflama yapmak üzere uyarlanmıştır. Burada iki yaklaşım esas alınmaktadır. İlki sürekli ya da kesikli sayısal verilerin kategorik değişkene dönüştürülmesi ve ardından standart aşamaların uygulanarak önsel olasılıkların hesaplanmasını gerektirmektedir. Ancak keyfi olabilecek bu işlem ciddi bilgi kayıplarına neden olabileceği için özellikle yüksek boyutlu yapılar için çok kullanışlı değildir. Diğer yaklaşım ise değişkenlerin dağılımlarına ilişkin olasılık yoğunluk fonksiyonları yardımıyla önsel olasılıkların hesaplanmasını gerektirmektedir. Normal dağılıma ilişkin $P(x_i|y_j)$ önsel olasılığının hesaplanması:

$$P(x_i|y_j) : \frac{1}{\sigma_{ij}\sqrt{2\pi}} \exp\left(\frac{-(x-\mu_{ij})^2}{2\sigma_{ij}^2}\right), -\infty < x < \infty, -\infty < \mu_{ij}, \sigma_{ij} > 0 \quad (2.26)$$

şeklinde yapılmaktadır. Burada normal dağılım örneği verilmiş olmakla birlikte diğer dağılımlar içinde ilgili olasılık yoğunluk fonksiyonları yardımı ile önsel olasılıklar hesaplanabilmektedir.

NB sınıflayıcılar ile ilgili önemli bir diğer nokta da “Sıfır Olasılık Sorunu” adı verilen problemle karşılaşılması durumudur. Yöntemde sınıflama yapmak amacıyla dikkate alınan yeni veri setinde eğitim setinde olmayan herhangi bir değişken ya da değişken kategorisi olması durumunda $P(x_i|y_j)$ olasılığı 0 olmaktadır. Bu da diğer sonsal olasılıkları etkilemektedir. Bu sorunu çözmek için Laplace düzeltmesinden faydalanılabilmekte ve olasılığı hesaplanacak değere “1” değeri eklenmektedir (2,37). Bu düzeltmede, eğitim setinin yeteri kadar büyük olduğu ve eklenen değer ile hesaplanan olasılığın ise ihmal edilebilir olduğu varsayılmaktadır. Böylece diğer değişkenlerin sonsal olasılıkları etkilenmemektedir.

NB yönteminin avantajlarına değinilecek olursa, küçük miktardaki veri ile parametre kestirimi yapılabilmesi yöntemin en önemli yanlarından biridir. Ayrıca değişkenlerin bağımsızlığı varsayımından dolayı kovaryans matrisi yerine yalnızca değişkenlerin varyanslarının belirlenmesinin yeterli olması yüksek boyut durumunda önemli avantajlar sağlamaktadır.

2.2.4. Random Forest

Random Forest (RForest) 2001 yılında Breiman (38) tarafından önerilmiş, karar ağaçlarının birleştirilmesine dayanan, popüler ve etkili bir sınıflama algoritmasıdır. Karar ağaçları; karar düğümü, dallar ve yapraklardan oluşmaktadır. Karar düğümleri yapılacak testi belirtmekte, yapraklar ise sınıfları belirtmektedir. Sınıflama işlemini bitirmeye aday her bir dal için dalın ucunda sınıflama gerçekleşmiyorsa yeni bir karar düğümü oluşmakta ve dallar sonlanana kadar bu işlem devam etmektedir. Oluşan yaprak ile nesnenin/bireyin sınıfı belirlenmiş olmaktadır. RForest ise esas olarak birçok karar ağacını bir araya getirerek sınıflama performansında iyileşme sağlamayı amaçlamaktadır. Bunun için veri setinden “bootstrap” tekniği ile örnekler seçerek karar ağaçlarında öğrenme süreçlerini gerçekleştirmektedir. Sınıflama yapan her bir karar ağacı için oylama yapılarak en yüksek oyu alan karar ağacının yaptığı sınıflama seçilmektedir. Her ağaç için dallara ayırıcı değişkenlerin seçimi rasgele yapılmaktadır.

RForest’ta sınıflama adımları basitçe aşağıdaki gibidir (39):

- İlk aşama örnek seçimi ile başlamaktadır. Eğitim verisi oluşturulurken veri setinin üçte biri hata oranını hesaplamak üzere test veri seti (out-of-bag) olarak

ayrılmaktadır. Eğer orijinal veri setinde test seti varsa hata oranı bu veri seti ile de hesaplanabilmektedir.

- Daha sonra karar ağaçları bootstrap ile oluşturulan her bir örnek setine uygulanmaktadır. Bu aşamada her bir düğümde rasgele seçilen az sayıda değişken ikili ayırma için kullanılmaktadır. Breiman tarafından sınıflama problemlerinde değişken sayısının $p/3$ olarak kullanılması önerilmiş olsa da gereğinden az ya da daha fazla değişken seçiminin sınıflama performansı üzerinde olumsuz etki oluşturabileceği belirtilmektedir. Bu nedenle farklı değişken sayıları belirlenerek bu sayılar için test verisinde hata oranlarının hesaplanması ve hata oranının en küçük olduğu değişken sayısının kullanılması daha çok tercih edilmektedir.

- En çok bilgi kazancı sağlayan değişken dalları ayırmada kullanılmakta ve gini indeksine dayanarak dalları ayırma kriteri belirlemektedir. Bütün değişkenler kullanılıp dallara ayırmada kullanılacak başka değişken kalmadığında ya da tüm gözlemler aynı sınıfta kaldığında ağaç büyümesi tamamlanmakta ve sınıfı belirleyen yapraklar oluşmaktadır.

- Ağaçların büyümesi tamamlandığında test setindeki gözlemler için tahmin yapılmaktadır.

- Her gözlem için test setindeki (out-of-bag) tahminler kullanılarak doğruluk ve hata değerleri hesaplanmakta, daha sonra da tüm gözlemler için bu değerlerin ortalamaları hesaplanmaktadır. Bu değerlere göre en çok oyu alan sınıf yeni verinin atanacağı sınıfı belirtmektedir.

RForest yöntemi ile sınıflama yapabilmenin yanı sıra değişken önemlilikleri belirlenerek değişken seçimi de yapılabilmektedir. Bunun için gini indeksi kullanılabildiği gibi permutasyona dayalı hesaplamadan da yararlanılabilmektedir. Gini indeksinin değişken seçiminde yanlılık eğilimi olduğu ifade edilmekte ve özellikle son dönemlerde yapılan çalışmalarda dengesiz sınıf dağılımı durumlarında değişken önemliliği konusunda yarattığı sakıncalar ifade edilmektedir (40). Permutasyona dayalı hesaplama ise değişken önemliliğini belirleme açısından daha popüler ve kullanışlı bir ölçüdür. Burada değişkenin önemini belirlemek için diğer değişkenler sabitken ilgili değişkene ait gözlemler karıştırılarak test setindeki (out-of-bag) tahmin hatalarındaki değişime bakılmaktadır. Tüm ağaçlar için elde edilen bu farklar z skoruna dönüştürülerek değişkenlerin önemlilik sıralaması belirlenmektedir.

RForest eksik ve kayıp verilerin analizinde de sıklıkla kullanılmaktadır. Kayıp gözlemlerin kestirimi eğitim ve test setinde farklılık göstermektedir. Eğitim setinde iki farklı şekilde eksik değer kestirimi yapılabilmektedir. İlkinde, ilgili değişken sürekli veri tipinde ise eksik olan gözlem için ilgili sınıftaki ilgili değişkenin tüm değerlerinin ortancası hesaplanmakta ve eksik gözlem yerine bu değer atanmaktadır. Eğer ilgili değişken kategorik ise bu kez ilgili sınıfta o değişken için en çok görülen değer eksik gözlem yerine atanmaktadır. İkinci yöntemde ise kayıp gözlem yerine kabataslak ve keyfi atama yapılmakta daha sonra ağaç çalıştırılarak yakınlık değerleri hesaplanmaktadır. Sürekli veri tipindeki gözlem için ilgili değişkenin kayıp gözlem ile kayıp olmayan gözlemler arasındaki yakınlık değeri ile ağırlıklandırılan değişkenin kayıp olmayan gözlemleri üzerinden ortalama hesaplanmakta ve bu değer atanmaktadır. Kategorik veri tipi söz konusu olduğunda ise yakınlık değeri ile ağırlıklandırılmış değişkenin kayıp olmayan gözlemleri arasında en sık karşılaşılan değer kayıp gözlem yerine atanmaktadır. Test setlerinde ise kayıp gözlemin sınıfı belli olduğu durumlarda eğitim setinden atama yapılabilmektedir. Sınıf belli olmadığında ise her bir gözlem sınıf sayısı kadar çoğaltılmakta, ilk kopyalar birinci sınıf olarak varsayılmakta ve birinci sınıf kayıp gözlemleri hesaplanmada kullanılmaktadır. Sınıf sayısı kadar bu işlem tekrar etmekte ve her kopya seti içinde en yüksek oyu alan sınıf hangisi ise gözlem o sınıfa atanarak değeri hesaplanmaktadır.

RForest'in bunlar dışında yüksek boyut ile çalışabilme, aşırı uyum sorununun birçok yöneme göre daha az olması, hızlı öğrenme süreci gibi avantajları bulunmaktadır. Sınıflama problemlerinde yüksek doğruluk sağlayan RForest gözlemler arası yakınlık hesaplayarak kümeleme problemlerinde de başarıyla uygulanmaktadır.

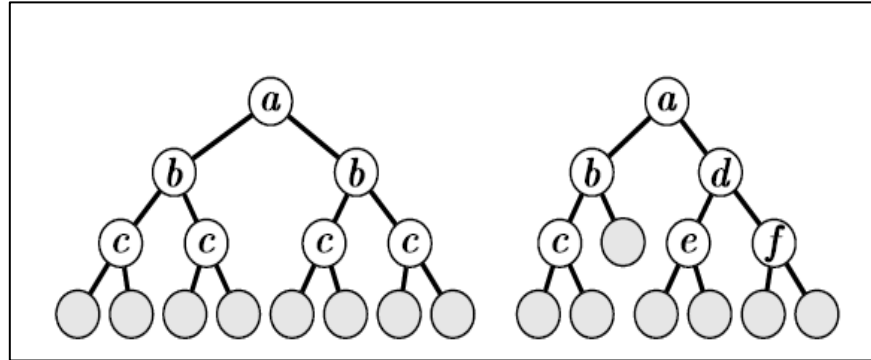
Yöntemin dezavantajları ise karar ağaçları gibi görsel sunumunun bulunmaması ve güven aralığı değerleri vermemesi olarak sıralanabilir (41).

2.2.5. Random Ferns

Random Ferns (RFerns) görece daha yeni bir makine öğrenmesi algoritmasıdır. Özuysal ve diğ. (2007) tarafından iki ekran üzerindeki görüntülerde eş noktaların belirlenmesi amacıyla mühendislik uygulamalarına yönelik geliştirilmiş basit ve etkili bir algoritmadır (42). Özellikle görüntü tanıma analizlerinde parçaların

sınıflandırılması için ağaçların yerine kullanılmak üzere tasarlanmıştır. Yöntemin ağaçlara göre daha basit ve hızlı uygulandığı bununla birlikte onlar kadar da güvenilir olduğu belirtilmektedir (43). İsim ve form/şekil olarak Random Forest’a benzemesine karşın genel anlamda iki açıdan farklılık göstermektedir. Ana fikrini Semi-Naive Bayes algoritmasına dayandırmaktadır ve ağaçlar yerine “*fern*” adı verilen hiyerarşik olmayan yapılar ile nesneler sınıflandırılmaktadır. Burada sözü edilen hiyerarşi, ağaçlarda dallanma aşamasında en iyi ayrımı yapacak şekilde ayırıcı değişkenlerin sıralanmasıdır. RFerns’de dallanma RForest’a benzer şekilde yukarıdan aşağıya olmakla birlikte dallanmalarda kullanılan değişkenler arasında hiyerarşik bir düzen bulunmamaktadır.

Hiyerarşik olmayan yapıya sahip bu yöntemde her bir fern ikili (binary) test setleri içermekte ve sınıflayıcı olarak kullanılan ikili testler tamamen rasgele seçilmektedir. Bu testler ilgilenilen parçaların/görüntülerin hangi sınıfa atanacağını belirlemek için olasılıklara dönüştürülmekte ve Naive Bayes yaklaşımı ile sonuçlar birleştirilmektedir. Öğrenme aşamasında bu sıralı test setleri RForest’tan farklı olarak tüm gözlemlere uygulanmaktadır. RForest’a benzer şekilde yapraklar sonsal olasılıkları içermektedir (Şekil 2.4.).



Şekil 2.4. Karar ağacı formunda fern yapısı (solda), genel karar ağacı (sağda) Kursu (45)’dan alınmıştır.

Bayes yaklaşımı dikkate alındığında problem:

$$\operatorname{argmax} P(x_1, x_2, \dots, x_d | y = Y_i) P(y = Y_i) \quad (2.27)$$

şeklinde tanımlanmaktayken NB için değişkenlerin bağımsızlığı varsayımı, problemi:

$$P(x_1, x_2, \dots, x_d | y = Y_i) = \prod_{j=1}^d P(x_j | Y_i) \quad (2.28)$$

formuna dönüştürür. Burada tüm değişkenler ilişkisiz olarak varsayılmakta değişkenler arasındaki tüm ilişkiler ihmal edilmektedir.

RFerns'de ise d tane değişken s büyüklüğünde M tane küçük sete bölünmekte ve fern adı verilen bu setlerin birbirinden tamamen bağımsız olduğu ancak set içindeki değişkenlerin ilişkili oldukları varsayılmaktadır. Bu durumda koşullu olasılık $X_m = \{X_{m,1}, X_{m,2}, \dots, X_{m,s}\}$, $m=1, 2, \dots, M$ için:

$$P(x_1, x_2, \dots, x_d | y = Y_i) = \prod_{m=1}^M P(X_m | Y_i) \quad (2.29)$$

olarak hesaplanmaktadır. Semi-Naive Bayes yaklaşımı ile her bir fern içindeki ilişkiler dikkate alınıp fernler arasında ilişki olmadığı varsayıldığından, hem standart NB'deki basitlik korunmakta hem de korelasyon dengesi sağlanmaktadır (44). Değişkenlerin fernler halinde gruplanması da tamamen rasgele gerçekleştirilmektedir (42).

RFerns orijinalinde yalnızca ikili (binary) değişken içeren veri setleri ve dengeli dağılım durumu için geliştirilmiş olsa da, Kursu (45) tarafından hem sürekli ve kategorik değişkenlerle çalışabilecek forma getirilmiş hem de dengesiz dağılımlara uyarlanmıştır. Bayes yaklaşımlarda sürekli değişkenlerle çalışabilmek için genellikle bir olasılık dağılımı varsayımı ve uyarlaması yapılmaktadır. Ancak bunun gerektirdiği varsayımsal sorunlar ve güvenilirlik sorunu nedeniyle ağaç yaklaşımlarında genellikle bilgi kazancına dayanarak sürekli değişkenler için eşik değerler, kategorik değişkenler için ise alt kategori grupları belirlenmektedir. RFerns'de ise yöntemin stokastik doğasını korumak için bu eşik değerler de rasgele belirlenmektedir. Sürekli değişkenler için ilgili değişkene ait rasgele iki değer ortalaması kullanılmaktadır. Kategorik değişkenlerde ise ilgili değişkenin tüm kategorileri içinde gözlemlerin tamamını ya da hiçbirini içermeyen kategori haricinde bir kategori rasgele seçilmektedir. Bu şekilde ayırmanın aşırı uyumdan koruduğu ve kolay anlaşılamayan etkileşimleri ortaya çıkardığı belirtilmektedir.

Görüntü tanıma, hareket tanıma çalışmalarında RFerns uygulamalarına literatürde rastlanmaktadır. Ayrıca uygulamalarda kullanıcıya kolaylık olması açısından R yazılımında RFerns paketi geliştirilmiştir. Böylece yöntemin ilerleyen zamanda kendisine daha geniş bir uygulama alanı bulacağı düşünülmektedir.

2.3. Sınıflama Performanslarının Değerlendirilmesi

Performans ölçütleri öğrenme modellerinin başarısının değerlendirilmesinde büyük önem arz etmektedir. Bu ölçütler ile yöntemlerin nesneleri doğru sınıflara atama becerileri karşılaştırılabilmekte ve yeni nesneleri en doğru şekilde sınıflayacak yöntemi belirlemek için bir öngörü sağlanmaktadır.

Literatürde çok farklı ölçütler bulunmakla birlikte bu tez çalışması kapsamında doğruluk, duyarlılık, seçicilik, F ölçütü ve Matthews korelasyon katsayısı dikkate alınmıştır. Söz konusu ölçütlerin daha kolay anlaşılabilmesi için öncelikle Tablo 2.1.'de yer alan 2x2 olumsuzluk tablosuna bakmak fayda sağlamaktadır.

Tablo 2.1. Olumsuzluk Tablosu

Gerçek Durum				
Tahmin (Test Sonucu)		Pozitif (1)	Negatif (0)	Toplam
	Pozitif (1)	Doğru Pozitif (a)	Yanlış Pozitif (b)	a+b
	Negatif (0)	Yanlış Negatif (c)	Doğru Negatif (d)	c+d
	Toplam	a+c	b+d	a+b+c+d

İki sınıflı sınıflama problemleri için test sonucu ile gerçek durumdaki sonuçların yer aldığı çapraz tablodan yararlanarak ölçütler aşağıdaki şekilde hesaplanmaktadır.

Doğruluk, test sonucunda doğru olarak belirlenen pozitif ve negatif olguların/nesnelerin tüm olguların/nesnelerin içindeki payı olarak ifade edilebilir. En yaygın ve basit performans ölçülerindendir. Tablo 2.1. yardımıyla doğruluk değeri aşağıdaki eşitlik ile elde edilir.

$$Doğruluk = \frac{a+d}{a+b+c+d} \quad (2.30)$$

Duyarlılık, testin gerçek durumda pozitif olan olguları/nesneleri doğru belirleme başarısıdır.

$$Duyarlılık = \frac{a}{a+c} \quad (2.31)$$

Seçicilik, testin gerçek durumda negatif olan olguları/nesneleri doğru belirleme başarısıdır.

$$Seçicilik = \frac{d}{b+d} \quad (2.32)$$

F Ölçütü: Özellikle sınıf dağılımlarının dengesiz olduğu durumlarda tercih edilen bu ölçüt Recall ve precision değerlerinin harmonik ortalamasıdır. Recall aynı zamanda duyarlılığa eşittir, precision ise pozitif kestirim oranına eşittir. Pozitif kestirim oranı ise testin pozitif bulduğu olguların gerçekte de pozitif olma oranını ifade etmektedir.

$$F \text{ Ölçütü} = 2x \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2a}{2a+b+c} \quad (2.33)$$

Matthews Korelasyon Katsayısı: Dengesiz dağılım durumunda tercih edilen diğer ölçütlerden biridir ve phi katsayısı da olarak bilinmektedir. Aşağıdaki şekilde hesaplanmaktadır.

$$Matthew KK = \frac{(axd)-(bxc)}{\sqrt{(a+b)(a+c)(d+b)(d+c)}} \quad (2.34)$$

3. GEREÇ VE YÖNTEM

Çalışma kapsamında sınıflama yöntemlerinin performanslarını değerlendirmek için farklı senaryolara göre benzetim teknikleri ile türetilmiş yapay veri setleri kullanılmıştır. Benzetim çalışmaları ve analizler için R yazılımından faydalanılmıştır.

Sınıflama aşamasında DVM için e1071 paketindeki 'ksvm' fonksiyonu, UAA için DWD paketindeki 'kdwd' fonksiyonu, RForest için randomForest paketindeki 'randomForest' fonksiyonu, RFerns için RFerns paketinde 'rFerns' fonksiyonu ve NB için e1071 paketindeki 'naiveBayes' fonksiyonu kullanılmıştır. Benzetim uygulamaları için türetilen yapay veri setlerinde gözlemlerin %75'i eğitim seti, %25'i test seti olarak ayrılmıştır. Her bir senaryo 1000 defa tekrar edilerek DVM, UAA, RForest, RFerns ve NB sınıflayıcılarının test setleri üzerindeki performanslarına ilişkin ölçütler ortalama olarak raporlanmıştır. Tez çalışması kapsamında doğruluk, duyarlılık, seçicilik, F ölçütü ve Matthews korelasyon katsayısı olmak üzere 5 farklı performans ölçütü dikkate alınmıştır. Benzetim uygulamalarına ilişkin örnek komutlar EK 1'de verilmiştir.

3.1. Benzetim Çalışması

3.1.1. Senaryolar

YBDÖG düzenler dikkate alınarak örneklem büyüklüğü, değişken sayısı (boyut), bağımlı değişkenler ile bağımsız değişkenler arasındaki ilişki, bağımsız değişkenlerin kendi aralarındaki ilişki, dengeli ve dengesiz dağılım durumu ile gürültü (noise) oranlarının farklılaştırıldığı veri yapısı senaryoları oluşturulmuştur. Çalışmada yalnızca iki sınıflı sınıflama problemleri dikkate alınmış olup senaryoların genel çerçevesi aşağıdaki gibi belirlenmiştir:

Örneklem büyüklüğünün sınıflayıcıların performansı üzerindeki etkisini incelemek üzere öncelikle YBDÖG düzenlerde genel olarak karşılaşılan örneklem büyüklükleri araştırılmıştır. Literatür göz önüne alınarak örneklem büyüklükleri $n=20$, $n=50$ ve $n=100$ için senaryolar hazırlanmıştır.

Değişken sayılarının (boyut) etkisini değerlendirmek üzere $d=100$, $d=500$ ve $d=1000$ için senaryolar oluşturulmuştur.

Bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiler dikkate alınarak bunların sınıflama performansları üzerindeki etkileri incelenmiştir. İlişki düzeyleri düşük, orta ve yüksek şeklinde gruplandırılmıştır. Korelasyon matrisinde her bir eleman için sabit bir değer atamak yerine her bir tekrarda (iterasyon) belirlenen korelasyon düzeyi için ilgili aralıktan rasgele seçim yapılarak değer atanması sağlanmıştır. Düşük korelasyon senaryosu için her bir tekrarda korelasyon matrisinin elemanları 0,10 ile 0,30 arasından rasgele seçim yapılarak belirlenmiştir. Bu aralık orta düzey ilişki için 0,40 ile 0,60, yüksek düzey ilişki için 0,70 ile 0,90 olarak belirlenmiştir.

Bağımsız değişkenlerin kendi aralarındaki ilişkiler de düşük (0,10-0,30), orta (0,40-0,60) ve yüksek (0,70-0,90) olarak belirlenmiştir. Korelasyon değerlerinin belirlenen aralıklardan yine rasgele olarak seçilmesi sağlanmıştır.

Bir diğer senaryo sınıf dağılımlarının dengeli olup olmaması durumunu dikkate almaktadır. Dengeli dağılım için bağımlı değişkenin prevalansı 0,50 olarak dikkate alınmıştır. Dengesiz dağılım durumunu izlemek için ise prevalans 0,35 olarak ele alınmıştır.

Gürültü oranları orta ve yüksek düzeylerde belirlenmiştir. Orta düzey gürültü varlığı için bağımsız değişkenlerin %50'si bağımlı değişken ile ilişkisiz ve yüksek düzey gürültü varlığı için bağımsız değişkenlerin %90'ı bağımlı değişken ile ilişkisiz olarak belirlenmiştir. İlişkisizlik durumu için korelasyon aralığı 0,00-0,10 olarak dikkate alınmıştır. Burada da, bağımlı değişken ile bağımsız değişkenler arasındaki korelasyonların belirlenmesinde olduğu gibi rasgele süreçler işletilmiştir.

Yapay veri setleri türetmeye yönelik senaryolara ilişkin bilgiler Şekil 3.1'de özetlenmiştir. Gürültü oranının %50 (orta) olarak belirlendiği durumlarda, bağımlı değişken ile bağımsız değişkenler arasında yüksek korelasyon olması halinde değişkenlerinin yarısı ilişkisiz iken diğer yarısının yüksek ilişkili olması bağımsız değişkenler arasındaki korelasyon düzeyleri açısından uyumsuzluk yaratmakta ve varyans kovaryans matrisi pozitif tanımlı olmamaktadır. Bundan dolayı söz konusu senaryolar çalışma dışında tutulmuştur. Çalışma kapsamında toplam 270 senaryoya ilişkin performans karşılaştırması yapılmış ve sonuçları raporlanmıştır.

Örneklem Büyüklüğü	• $n=20$ • $n=50$ • $n=100$
Değişken Sayısı (Boyut)	• $d=100$ • $d=500$ • $d=1000$
Bağımlı Değişken ile Bağımsız Değişkenler Arasındaki İlişki Düzeyi	• Yüksek (0,70-0,90) • Orta (0,40-0,60) • Düşük (0,10-0,30)
Bağımsız Değişkenlerin Kendi Aralarındaki İlişki Düzeyi	• Yüksek (0,70-0,90) • Orta (0,40-0,60) • Düşük (0,10-0,30)
Gürültü Oranı (Noise)	• Yüksek (%90) • Orta (%50)
Prevalans	• Dengeli Dağılım $P(X=1) = 0,50$ • Dengesiz Dağılım $P(X=1) = 0,35$

Şekil 3.1. Çalışma Kapsamında Kullanılan Senaryolar

3.1.2. Veri Türetme Süreçleri

Veri türetme aşamasında genel olarak Demirtaş ve diğ. tarafından (46, 47) önerilen eş zamanlı ikili ve sürekli değişken üretimi prosedürlerinden yararlanılmıştır. Bu amaçla R yazılımında geliştirilmiş BinNor (48) paketindeki ilgili fonksiyonlar kullanılmıştır.

Bu süreçte önemli aşamalardan birisi korelasyon matrisinin oluşturulmasıdır. Tüm veriler standart normal dağılımdan türetildiği için ve bağımlı değişken gerçek kesikli (dikotom) bir veri olmadığından korelasyon matrisinin oluşturulmasında “BinNor” paketindeki *compute.sigma.star* fonksiyonundan yararlanılmıştır. Bu

fonksiyon ile bağımsız sürekli değişken ile bağımlı ikili değişken arasındaki ilişkiler için biserial korelasyon, sürekli değişkenlerin kendi aralarındaki ilişkiler için Pearson korelasyon katsayısı hesaplanmaktadır. Çalışma kapsamında sadece sürekli bağımsız değişkenler dikkate alınmış olmakla birlikte söz konusu fonksiyon içinde ikili değişkenlerin kendi aralarındaki ilişkiyi ifade eden tetrakorik korelasyon değerleri de hesaplanabilmektedir. Ayrıca oluşturulan korelasyon matrisinin pozitif tanımlı olmaması durumunda söz konusu fonksiyon en yakın pozitif tanımlı matrisi vermektedir.

Gürültü oranlarının senaryolara dahil edilmesi de korelasyon matrislerinin oluşturulması sürecinde dikkate alınmıştır.

Veri setindeki toplam değişken sayısı (boyut) d olsun. Bu durumda oluşturulacak korelasyon matrisinde köşegen altında kalan üçgendeki elemanların sayısı aşağıdaki gibidir:

$$(d \times (d - 1))/2 \quad (3.1)$$

Belirlenen gürültü oranı için bağımlı değişkenle ilişkisiz değişken (gürültü) sayısı:

$$(d - 1) \times \text{Gürültü Oranı} \quad (3.2)$$

kadardır. Bağımlı değişken ile ilişkili bağımsız değişkenlerin sayısı ise Eşitlik 3.3'te ifade edildiği gibi bağımsız değişken sayısından gürültü değişkenlerinin sayısı çıkarılarak aşağıdaki şekilde:

$$(d - 1) - ((d - 1) \times \text{Gürültü Oranı}) \quad (3.3)$$

hesaplanır. Verideki toplam değişken sayısı " d " olduğu için bağımlı değişken çıkarılarak bu işlem $d-1$ tane bağımsız değişken üzerinden yapılmıştır. Ondalıklı çıkan değerler yuvarlanarak tam sayıya dönüştürülmektedir. Korelasyon düzeyleri belirlenecek bağımsız değişkenlerin sayısı ise matrisin alt üçgeninde yer alan toplam eleman sayısından bağımsız değişken sayısı çıkarılarak hesaplanır ve aşağıdaki gibi olur:

$$(d \times (d - 1))/2 - (d - 1) \quad (3.4)$$

Hem gürültü değişkenleri hem de ilişkili değişkenler için belirlenen bu sayılara dayanarak tekdüze (uniform) dağılımdan, belirlenen korelasyon düzeyinde, rasgele değerler üretilerek korelasyon matrisine yerleştirilir.

Korelasyon matrisi üretme sürecini bir örnek üzerinden ifade etmek gerekirse 1 bağımlı değişken ve 29 bağımsız değişkene sahip bir veri seti dikkate alınsın. Bağımlı değişken ikili (binary), bağımsız değişkenler ise sürekli veri tipinde olsun. Bu durumda boyut $d=30$ olur ve korelasyon matrisinin köşegeni altında kalan elemanlarının sayısı $(30 \times (30-1))/2 = 435$ olur.

Gürültü oranının %90 olarak belirlendiği bir senaryo için bağımlı değişkenle ilişkisiz (gürültü) değişkenlerin sayısı $29 \times 0,90 = 26,1 \approx 26$ olur. Belirli bir ilişki düzeyine sahip olanların sayısı ise $29 - 26 = 3$ olur. 26 gürültü değişkeninin bağımlı değişkenle aralarındaki korelasyon 0 ile 0,10 arasından rasgele seçilerek belirlenir. Bunun için gürültü değişkeni sayısı kadar tekdüze dağılımdan 0 ile 0,10 arasında sayı türetilir.

Bu örnek için bağımlı değişken ile bağımsız değişkenler arasındaki korelasyonun yüksek, bağımsız değişkenlerin kendi aralarındaki korelasyonun ise düşük düzeyde olduğu varsayılabilir. Bu durumda bağımlı değişken ile ilişkili 3 bağımsız değişkene ait korelasyon değerleri tekdüze dağılım kullanılarak 0,70 ile 0,90 arasından rasgele olarak belirlenir. Bağımsız değişkenlerin kendi aralarındaki korelasyon için ise geriye kalan $435 - 29 = 406$ korelasyon değeri 0,10-0,30 arasından rasgele olarak belirlenir. Böylece korelasyon matrisine ait alt üçgen elde edilmiş olur.

Korelasyon matrisi oluşturulduktan sonra bu matrise dayanarak belirlenen gözlem sayısı ve değişken sayısı için “rmvnorm” fonksiyonu yardımıyla çok değişkenli standart normal dağılımdan veriler türetilmiştir. Bağımlı değişken belirlenen prevalans düzeyine göre kategorik hale getirilmiştir. Bağımlı değişkenin oluşturulmasında prevalansın her tekrarda belirlenen düzeyde sabit kalmasını sağlamak için BinNor içinde yer alan veri üretme fonksiyonu kullanılmamıştır.

4. BULGULAR

Bu bölümde uygulama sonuçları tablo ve grafiklerle sunulmuştur. Senaryo sayısının fazla olması nedeniyle doğruluk ve F ölçütü dışındaki diğer performans ölçütlerine ilişkin tablolar Ek 2’de yer almaktadır.

Sınıf dağılımının dengeli olması durumu için çalışma kapsamında belirlenen senaryolara ilişkin sınıflama yöntemlerinin ortalama doğruluk değerleri Tablo 4.1.’de F ölçütü değerleri ise Tablo 4.2’de yer almaktadır. Tablolarda ilk sütun, senaryoları numaralandırmak için oluşturulmuştur. Gürültü oranına ilişkin sütun adı ‘G.O.’ olarak belirtilmiş olup yüksek gürültü oranına (% 90) sahip veri setlerini ifade etmek için ‘Y’, orta düzey gürültü oranına (%50) sahip veri setlerini ifade etmek için ‘O’ harfleri kullanılmıştır. Bağımlı değişken ile bağımsız değişkenler arasındaki korelasyonlar ‘BlıBsız’, bağımsız değişkenlerin kendi aralarındaki korelasyonlar ‘BsızBsız’ olarak isimlendirilmiştir. ‘D’ düşük düzey korelasyon, ‘O’ orta düzey korelasyon ve ‘Y’ yüksek düzey korelasyonu ifade etmektedir. Gözlem sayısına ilişkin sütuna ‘n’ ve boyuta ilişkin sütuna da ‘d’ adı verilmiştir.

Tablo 4.1. Dengeli Dağılım Durumunda Diğer Senaryolara Göre Doğruluk Değerleri

No	G.O.	Blı Bsız	Bsız Bsız	n	d	Doğruluk				
						DVM	UAA	RForest	RFerns	NB
1	Y	D	D	20	100	0,358	0,494	0,453	0,661	0,458
2	Y	D	D	20	500	0,350	0,462	0,434	0,666	0,415
3	Y	D	D	20	1000	0,338	0,415	0,412	0,671	0,388
4	Y	D	D	50	100	0,495	0,567	0,532	0,611	0,555
5	Y	D	D	50	500	0,452	0,540	0,502	0,595	0,520
6	Y	D	D	50	1000	0,432	0,513	0,485	0,594	0,504
7	Y	D	D	100	100	0,577	0,609	0,582	0,604	0,586
8	Y	D	D	100	500	0,521	0,566	0,536	0,589	0,558
9	Y	D	D	100	1000	0,489	0,529	0,515	0,574	0,542
10	Y	D	O	20	100	0,380	0,495	0,457	0,623	0,470
11	Y	D	O	20	500	0,369	0,465	0,447	0,643	0,447
12	Y	D	O	20	1000	0,355	0,429	0,432	0,633	0,424
13	Y	D	O	50	100	0,499	0,567	0,526	0,579	0,534
14	Y	D	O	50	500	0,452	0,513	0,494	0,569	0,514
15	Y	D	O	50	1000	0,437	0,494	0,475	0,570	0,501
16	Y	D	O	100	100	0,552	0,612	0,561	0,573	0,546
17	Y	D	O	100	500	0,498	0,553	0,525	0,553	0,526
18	Y	D	O	100	1000	0,485	0,528	0,508	0,550	0,526
19	Y	D	Y	20	100	0,402	0,502	0,464	0,602	0,487

Tablo 4.1. (Devam)

No	G.O.	Bl Bsiz	Bsiz Bsiz	n	d	Doğruluk				
						DVM	UAA	RForest	RFerns	NB
20	Y	D	Y	20	500	0,375	0,459	0,445	0,615	0,459
21	Y	D	Y	20	1000	0,360	0,437	0,429	0,616	0,441
22	Y	D	Y	50	100	0,476	0,574	0,508	0,550	0,512
23	Y	D	Y	50	500	0,450	0,518	0,496	0,552	0,507
24	Y	D	Y	50	1000	0,443	0,498	0,478	0,552	0,499
25	Y	D	Y	100	100	0,536	0,624	0,553	0,553	0,527
26	Y	D	Y	100	500	0,494	0,547	0,517	0,540	0,518
27	Y	D	Y	100	1000	0,484	0,518	0,507	0,534	0,517
28	Y	O	D	20	100	0,389	0,569	0,505	0,686	0,496
29	Y	O	D	20	500	0,349	0,476	0,443	0,666	0,438
30	Y	O	D	20	1000	0,341	0,441	0,422	0,667	0,402
31	Y	O	D	50	100	0,565	0,663	0,610	0,648	0,609
32	Y	O	D	50	500	0,472	0,584	0,532	0,621	0,550
33	Y	O	D	50	1000	0,450	0,546	0,510	0,605	0,530
34	Y	O	D	100	100	0,681	0,726	0,678	0,672	0,655
35	Y	O	D	100	500	0,562	0,618	0,577	0,608	0,585
36	Y	O	D	100	1000	0,520	0,573	0,539	0,579	0,562
37	Y	O	O	20	100	0,403	0,572	0,482	0,635	0,492
38	Y	O	O	20	500	0,368	0,478	0,450	0,639	0,449
39	Y	O	O	20	1000	0,358	0,449	0,442	0,643	0,443
40	Y	O	O	50	100	0,544	0,666	0,583	0,608	0,565
41	Y	O	O	50	500	0,462	0,568	0,518	0,585	0,525
42	Y	O	O	50	1000	0,443	0,522	0,483	0,581	0,512
43	Y	O	O	100	100	0,658	0,739	0,651	0,628	0,594
44	Y	O	O	100	500	0,535	0,614	0,558	0,575	0,552
45	Y	O	O	100	1000	0,496	0,555	0,529	0,561	0,533
46	Y	O	Y	20	100	0,420	0,568	0,490	0,602	0,507
47	Y	O	Y	20	500	0,383	0,479	0,449	0,615	0,459
48	Y	O	Y	20	1000	0,385	0,465	0,457	0,635	0,467
49	Y	O	Y	50	100	0,524	0,683	0,571	0,586	0,541
50	Y	O	Y	50	500	0,469	0,574	0,513	0,568	0,519
51	Y	O	Y	50	1000	0,449	0,530	0,489	0,566	0,510
52	Y	O	Y	100	100	0,631	0,750	0,634	0,595	0,549
53	Y	O	Y	100	500	0,516	0,614	0,544	0,547	0,523
54	Y	O	Y	100	1000	0,487	0,552	0,518	0,539	0,513
55	Y	Y	D	20	100	0,393	0,573	0,507	0,686	0,503
56	Y	Y	D	20	500	0,352	0,480	0,444	0,661	0,426
57	Y	Y	D	20	1000	0,342	0,458	0,429	0,662	0,413
58	Y	Y	D	50	100	0,607	0,699	0,644	0,664	0,638
59	Y	Y	D	50	500	0,490	0,614	0,558	0,623	0,574
60	Y	Y	D	50	1000	0,451	0,566	0,511	0,608	0,538
61	Y	Y	D	100	100	0,718	0,759	0,712	0,690	0,676
62	Y	Y	D	100	500	0,585	0,652	0,602	0,618	0,606

Tablo 4.1. (Devam)

No	G.O.	Bl Bsiz	Bsiz Bsiz	n	d	Doğruluk				
						DVM	UAA	RForest	RFerns	NB
63	Y	Y	D	100	1000	0,535	0,603	0,559	0,595	0,579
64	Y	Y	O	20	100	0,415	0,594	0,515	0,648	0,521
65	Y	Y	O	20	500	0,364	0,495	0,463	0,646	0,466
66	Y	Y	O	20	1000	0,363	0,463	0,443	0,638	0,448
67	Y	Y	O	50	100	0,585	0,713	0,624	0,639	0,598
68	Y	Y	O	50	500	0,476	0,602	0,524	0,589	0,541
69	Y	Y	O	50	1000	0,447	0,561	0,510	0,583	0,525
70	Y	Y	O	100	100	0,697	0,776	0,687	0,646	0,613
71	Y	Y	O	100	500	0,563	0,657	0,587	0,584	0,560
72	Y	Y	O	100	1000	0,517	0,592	0,548	0,569	0,546
73	Y	Y	Y	20	100	0,408	0,605	0,496	0,618	0,507
74	Y	Y	Y	20	500	0,373	0,498	0,448	0,625	0,474
75	Y	Y	Y	20	1000	0,370	0,469	0,440	0,626	0,444
76	Y	Y	Y	50	100	0,565	0,730	0,608	0,601	0,551
77	Y	Y	Y	50	500	0,472	0,600	0,524	0,572	0,524
78	Y	Y	Y	50	1000	0,455	0,553	0,504	0,565	0,515
79	Y	Y	Y	100	100	0,689	0,802	0,688	0,624	0,573
80	Y	Y	Y	100	500	0,549	0,666	0,584	0,565	0,542
81	Y	Y	Y	100	1000	0,503	0,586	0,538	0,549	0,528
82	O	D	D	20	100	0,410	0,559	0,515	0,692	0,511
83	O	D	D	20	500	0,353	0,495	0,452	0,673	0,446
84	O	D	D	20	1000	0,343	0,457	0,425	0,674	0,422
85	O	D	D	50	100	0,587	0,656	0,599	0,665	0,627
86	O	D	D	50	500	0,489	0,581	0,542	0,630	0,575
87	O	D	D	50	1000	0,456	0,544	0,505	0,614	0,542
88	O	D	D	100	100	0,678	0,710	0,644	0,669	0,652
89	O	D	D	100	500	0,581	0,615	0,576	0,614	0,601
90	O	D	D	100	1000	0,541	0,581	0,551	0,597	0,584
91	O	D	O	20	100	0,423	0,563	0,501	0,645	0,524
92	O	D	O	20	500	0,377	0,470	0,455	0,647	0,450
93	O	D	O	20	1000	0,356	0,446	0,433	0,657	0,430
94	O	D	O	50	100	0,546	0,647	0,564	0,617	0,571
95	O	D	O	50	500	0,477	0,555	0,514	0,588	0,541
96	O	D	O	50	1000	0,451	0,527	0,500	0,578	0,517
97	O	D	O	100	100	0,636	0,705	0,614	0,614	0,593
98	O	D	O	100	500	0,533	0,584	0,542	0,573	0,554
99	O	D	O	100	1000	0,513	0,562	0,525	0,568	0,549
100	O	D	Y	20	100	0,415	0,538	0,480	0,603	0,506
101	O	D	Y	20	500	0,385	0,470	0,459	0,621	0,468
102	O	D	Y	20	1000	0,373	0,454	0,441	0,617	0,449
103	O	D	Y	50	100	0,516	0,651	0,550	0,580	0,545
104	O	D	Y	50	500	0,468	0,541	0,503	0,567	0,521
105	O	D	Y	50	1000	0,456	0,510	0,481	0,566	0,510

Tablo 4.1. (Devam)

No	G.O.	Bl Bsız	Bsız Bsız	n	d	Doğruluk				
						DVM	UAA	RForest	RFerns	NB
106	O	D	Y	100	100	0,608	0,717	0,596	0,588	0,557
107	O	D	Y	100	500	0,515	0,581	0,530	0,553	0,535
108	O	D	Y	100	1000	0,499	0,545	0,518	0,543	0,528
109	O	O	D	20	100	0,484	0,652	0,593	0,725	0,586
110	O	O	D	20	500	0,390	0,548	0,502	0,690	0,495
111	O	O	D	20	1000	0,370	0,514	0,488	0,675	0,472
112	O	O	D	50	100	0,711	0,761	0,698	0,738	0,717
113	O	O	D	50	500	0,570	0,651	0,609	0,675	0,638
114	O	O	D	50	1000	0,515	0,611	0,569	0,650	0,605
115	O	O	D	100	100	0,777	0,802	0,733	0,738	0,727
116	O	O	D	100	500	0,662	0,695	0,647	0,669	0,657
117	O	O	D	100	1000	0,613	0,647	0,610	0,643	0,628
118	O	O	O	20	100	0,468	0,627	0,552	0,682	0,555
119	O	O	O	20	500	0,390	0,519	0,479	0,665	0,486
120	O	O	O	20	1000	0,372	0,495	0,477	0,670	0,478
121	O	O	O	50	100	0,640	0,738	0,642	0,674	0,637
122	O	O	O	50	500	0,518	0,629	0,558	0,615	0,580
123	O	O	O	50	1000	0,496	0,592	0,546	0,614	0,565
124	O	O	O	100	100	0,735	0,793	0,686	0,681	0,654
125	O	O	O	100	500	0,599	0,668	0,598	0,617	0,594
126	O	O	O	100	1000	0,557	0,616	0,571	0,597	0,581
127	O	O	Y	20	100	0,462	0,646	0,538	0,643	0,541
128	O	O	Y	20	500	0,414	0,538	0,492	0,629	0,506
129	O	O	Y	20	1000	0,383	0,475	0,455	0,630	0,470
130	O	O	Y	50	100	0,602	0,751	0,617	0,629	0,590
131	O	O	Y	50	500	0,514	0,623	0,549	0,592	0,555
132	O	O	Y	50	1000	0,480	0,576	0,522	0,585	0,537
133	O	O	Y	100	100	0,723	0,811	0,679	0,649	0,609
134	O	O	Y	100	500	0,571	0,667	0,580	0,585	0,565
135	O	O	Y	100	1000	0,538	0,608	0,550	0,578	0,559

Tablo 4.1’de dengeli dağılım durumu için oluşturulan 135 senaryoda, en yüksek sınıflama doğruluğunun (0,811); gürültü düzeyinin orta, bağımlı değişken ile bağımsız değişken arasındaki ilişki düzeyinin orta, bağımsız değişkenlerin kendi aralarındaki ilişki düzeyinin yüksek, gözlem sayısının 100 ve boyut sayısının 100 olarak belirlendiği senaryoda gerçekleştiği görülmektedir. Doğruluk değerleri genel olarak incelendiğinde UAA ve RFerns’in diğer yöntemlere göre daha iyi performans gösterdiği anlaşılmaktadır. Bununla birlikte DVM performanslarının gürültü düzeyinin düşmesi ile bir miktar arttığı görülmektedir.

Tablo 4.2. Dengeli Dağılım Durumunda Diğer Senaryolara Göre F Değerleri

No	G.O.	Bh Bsız	Bsız Bsız	n	d	F Ölçütü				
						DVM	UAA	RForest	RFerns	NB
1	Y	D	D	20	100	0,291	0,466	0,428	-	0,422
2	Y	D	D	20	500	0,275	0,434	0,405	-	0,381
3	Y	D	D	20	1000	0,259	0,381	0,386	-	0,345
4	Y	D	D	50	100	0,472	0,560	0,522	0,531	0,544
5	Y	D	D	50	500	0,396	0,525	0,486	0,487	0,513
6	Y	D	D	50	1000	0,362	0,497	0,467	0,461	0,490
7	Y	D	D	100	100	0,567	0,603	0,577	0,580	0,579
8	Y	D	D	100	500	0,497	0,560	0,531	0,550	0,554
9	Y	D	D	100	1000	0,456	0,520	0,509	0,522	0,533
10	Y	D	O	20	100	0,331	0,458	0,415	-	-
11	Y	D	O	20	500	0,308	0,439	0,425	-	0,411
12	Y	D	O	20	1000	0,297	0,391	0,399	-	0,384
13	Y	D	O	50	100	0,472	0,558	0,516	0,524	0,520
14	Y	D	O	50	500	0,408	0,505	0,483	0,491	0,501
15	Y	D	O	50	1000	0,375	0,482	0,458	0,482	0,485
16	Y	D	O	100	100	0,542	0,607	0,555	0,553	0,537
17	Y	D	O	100	500	0,474	0,544	0,517	0,523	0,516
18	Y	D	O	100	1000	0,449	0,514	0,498	0,508	0,512
19	Y	D	Y	20	100	0,351	0,472	0,432	-	0,445
20	Y	D	Y	20	500	0,312	0,426	0,417	-	0,417
21	Y	D	Y	20	1000	0,295	0,395	0,385	-	0,383
22	Y	D	Y	50	100	0,443	0,561	0,493	0,503	0,486
23	Y	D	Y	50	500	0,418	0,507	0,488	0,492	0,491
24	Y	D	Y	50	1000	0,393	0,486	0,465	0,486	0,482
25	Y	D	Y	100	100	0,523	0,621	0,546	0,536	0,513
26	Y	D	Y	100	500	0,467	0,538	0,510	0,515	0,503
27	Y	D	Y	100	1000	0,452	0,499	0,497	0,501	0,497
28	Y	O	D	20	100	0,321	0,536	0,474	-	0,467
29	Y	O	D	20	500	0,277	0,446	0,414	-	0,395
30	Y	O	D	20	1000	0,250	0,414	0,390	-	0,358
31	Y	O	D	50	100	0,545	0,653	0,597	0,569	0,594
32	Y	O	D	50	500	0,417	0,575	0,520	0,518	0,539
33	Y	O	D	50	1000	0,396	0,541	0,498	0,480	0,523
34	Y	O	D	100	100	0,677	0,721	0,673	0,651	0,647
35	Y	O	D	100	500	0,542	0,614	0,573	0,574	0,579
36	Y	O	D	100	1000	0,483	0,565	0,532	0,536	0,554
37	Y	O	O	20	100	0,356	0,538	0,449	-	0,453
38	Y	O	O	20	500	0,315	0,457	0,421	-	0,406
39	Y	O	O	20	1000	0,289	0,424	0,414	-	0,399
40	Y	O	O	50	100	0,523	0,657	0,572	0,556	0,549
41	Y	O	O	50	500	0,417	0,557	0,505	0,519	0,511
42	Y	O	O	50	1000	0,386	0,510	0,467	0,491	0,493

Tablo 4.2. (Devam)

No	G.O.	Bl Bsız	Bsız Bsız	n	d	F Ölçütü				
						DVM	UAA	RForest	RFerns	NB
43	Y	O	O	100	100	0,654	0,735	0,647	0,611	0,585
44	Y	O	O	100	500	0,518	0,609	0,554	0,546	0,542
45	Y	O	O	100	1000	0,460	0,546	0,522	0,527	0,525
46	Y	O	Y	20	100	0,374	0,537	0,454	-	0,461
47	Y	O	Y	20	500	0,316	0,435	0,404	-	0,410
48	Y	O	Y	20	1000	0,326	0,430	0,421	-	0,429
49	Y	O	Y	50	100	0,500	0,672	0,558	0,544	0,517
50	Y	O	Y	50	500	0,424	0,563	0,497	0,502	0,494
51	Y	O	Y	50	1000	0,405	0,515	0,472	0,492	0,488
52	Y	O	Y	100	100	0,625	0,747	0,630	0,577	0,537
53	Y	O	Y	100	500	0,495	0,609	0,537	0,527	0,510
54	Y	O	Y	100	1000	0,452	0,543	0,512	0,510	0,499
55	Y	Y	D	20	100	0,335	0,536	0,471	-	0,464
56	Y	Y	D	20	500	0,272	0,448	0,412	-	0,388
57	Y	Y	D	20	1000	0,263	0,429	0,392	-	0,374
58	Y	Y	D	50	100	0,588	0,690	0,634	0,601	0,623
59	Y	Y	D	50	500	0,450	0,608	0,550	0,518	0,565
60	Y	Y	D	50	1000	0,386	0,558	0,498	0,495	0,526
61	Y	Y	D	100	100	0,715	0,756	0,708	0,674	0,670
62	Y	Y	D	100	500	0,574	0,647	0,597	0,581	0,596
63	Y	Y	D	100	1000	0,506	0,597	0,553	0,552	0,571
64	Y	Y	O	20	100	0,368	-	0,482	-	
65	Y	Y	O	20	500	0,303	0,459	0,427	-	0,428
66	Y	Y	O	20	1000	0,302	0,433	0,410	-	
67	Y	Y	O	50	100	0,571	0,704	0,618	0,591	0,585
68	Y	Y	O	50	500	0,437	0,590	0,513	0,520	0,527
69	Y	Y	O	50	1000	0,397	0,550	0,497	0,493	0,508
70	Y	Y	O	100	100	0,695	0,773	0,684	0,631	0,606
71	Y	Y	O	100	500	0,548	0,653	0,580	0,559	0,552
72	Y	Y	O	100	1000	0,486	0,585	0,541	0,535	0,535
73	Y	Y	Y	20	100	0,361	0,569	0,471	-	0,473
74	Y	Y	Y	20	500	0,309	0,468	0,414	-	0,434
75	Y	Y	Y	20	1000	0,308	0,446	0,407	-	0,405
76	Y	Y	Y	50	100	0,553	0,722	0,597	0,564	0,530
77	Y	Y	Y	50	500	0,438	0,590	0,509	0,517	0,509
78	Y	Y	Y	50	1000	0,421	0,541	0,487	0,492	0,497
79	Y	Y	Y	100	100	0,683	0,800	0,685	0,610	0,558
80	Y	Y	Y	100	500	0,532	0,660	0,576	0,542	0,527
81	Y	Y	Y	100	1000	0,479	0,579	0,533	0,519	0,515
82	O	D	D	20	100	0,358	0,521	0,484	-	0,475
83	O	D	D	20	500	0,289	0,472	0,425	-	0,417
84	O	D	D	20	1000	0,273	0,426	0,396	-	0,386
85	O	D	D	50	100	0,571	0,646	0,588	0,601	0,615

Tablo 4.2. (Devam)

No	G.O.	Bl Bsız	Bsız Bsız	n	d	F Ölçütü				
						DVM	UAA	RForest	RFerns	NB
86	O	D	D	50	500	0,456	0,569	0,531	0,522	0,562
87	O	D	D	50	1000	0,408	0,536	0,494	0,483	0,530
88	O	D	D	100	100	0,672	0,705	0,639	0,647	0,644
89	O	D	D	100	500	0,567	0,608	0,571	0,588	0,595
90	O	D	D	100	1000	0,515	0,573	0,546	0,561	0,574
91	O	D	O	20	100	0,365	0,526	0,465	-	0,480
92	O	D	O	20	500	0,323	0,444	0,426	-	0,412
93	O	D	O	20	1000	0,288	0,410	0,401	--	0,389
94	O	D	O	50	100	0,527	0,637	0,550	0,557	0,550
95	O	D	O	50	500	0,445	0,543	0,504	0,506	0,525
96	O	D	O	50	1000	0,395	0,515	0,488	0,503	0,502
97	O	D	O	100	100	0,632	0,702	0,609	0,599	0,584
98	O	D	O	100	500	0,511	0,578	0,535	0,551	0,545
99	O	D	O	100	1000	0,488	0,557	0,522	0,538	0,540
100	O	D	Y	20	100	0,371	0,498	0,435	-	-
101	O	D	Y	20	500	0,340	0,439	0,424	-	0,418
102	O	D	Y	20	1000	0,318	0,416	0,402	-	0,412
103	O	D	Y	50	100	0,489	0,641	0,537	0,540	0,520
104	O	D	Y	50	500	0,432	0,534	0,489	0,512	0,503
105	O	D	Y	50	1000	0,414	0,495	0,464	0,493	0,492
106	O	D	Y	100	100	0,597	0,712	0,589	0,569	0,545
107	O	D	Y	100	500	0,497	0,575	0,524	0,527	0,520
108	O	D	Y	100	1000	0,466	0,531	0,509	0,520	0,512
109	O	O	D	20	100	0,448	0,619	0,565	-	0,552
110	O	O	D	20	500	0,334	0,519	0,471	-	0,462
111	O	O	D	20	1000	0,293	0,484	0,460	-	0,429
112	O	O	D	50	100	0,707	0,757	0,692	0,704	0,713
113	O	O	D	50	500	0,550	0,642	0,601	0,607	0,627
114	O	O	D	50	1000	0,471	0,600	0,561	0,573	0,593
115	O	O	D	100	100	0,773	0,798	0,728	0,728	0,721
116	O	O	D	100	500	0,659	0,692	0,646	0,649	0,651
117	O	O	D	100	1000	0,604	0,642	0,606	0,617	0,620
118	O	O	O	20	100	0,418	-	-	-	-
119	O	O	O	20	500	0,342	0,486	0,446	-	0,440
120	O	O	O	20	1000	0,316	0,469	0,445	-	0,448
121	O	O	O	50	100	0,628	0,727	0,628	0,631	0,616
122	O	O	O	50	500	0,484	0,620	0,550	0,555	0,562
123	O	O	O	50	1000	0,453	0,579	0,527	0,538	0,545
124	O	O	O	100	100	0,730	0,789	0,682	0,666	0,643
125	O	O	O	100	500	0,588	0,661	0,593	0,601	0,587
126	O	O	O	100	1000	0,541	0,612	0,566	0,575	0,572
127	O	O	Y	20	100	0,416	0,610	0,494	-	0,492
128	O	O	Y	20	500	0,372	0,506	0,459	-	0,468

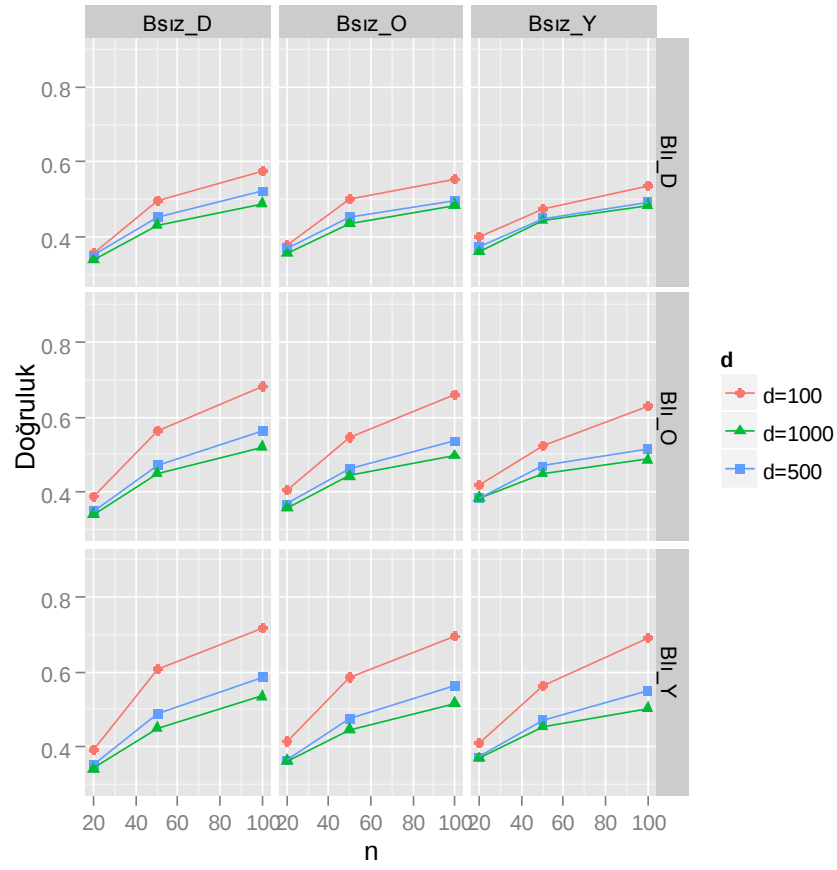
Tablo 4.2. (Devam)

No	G.O.	Blı Bsız	Bsız Bsız	n	d	F Ölçütü				
						DVM	UAA	RForest	RFerns	NB
129	O	O	Y	20	1000	0,328	0,433	0,420	-	0,420
130	O	O	Y	50	100	0,588	0,746	0,609	0,604	0,576
131	O	O	Y	50	500	0,485	0,614	0,534	0,544	0,533
132	O	O	Y	50	1000	0,443	0,567	0,512	0,523	0,522
133	O	O	Y	100	100	0,717	0,807	0,672	0,634	0,596
134	O	O	Y	100	500	0,557	0,663	0,574	0,564	0,552
135	O	O	Y	100	1000	0,515	0,601	0,545	0,554	0,547

Dengesiz dağılımlarda doğruluk değerlerinin yorumlama anlamında sakıncaları olduğu bilinmekte ve dengesiz dağılım durumlarında genellikle farklı ölçütler kullanılmaktadır. Çalışma kapsamında bu nedenle F ölçütü ve Matthews korelasyon katsayıları hesaplanmıştır. Ancak sınıf oranlarının (pozitif/negatif) 1:2 oranında olduğu durumlarda F ölçüsünün Matthews korelasyon katsayısından daha yüksek performans göstermesi nedeniyle (49) bu bölümde özellikle F ölçüsü üzerinde durulmuş Matthews korelasyon katsayıları Ek 2’de verilmiştir.

Tablo 4.2’de görüldüğü üzere dengeli dağılım durumunda F ölçütü değerleri genel olarak doğruluk değerlerine benzerlik göstermektedir. Burada da benzer şekilde gürültü düzeyinin orta, bağımlı değişken ile bağımsız değişkenler arasındaki ilişki düzeyinin orta, bağımsız değişkenlerin kendi aralarındaki ilişki düzeyinin yüksek, gözlem sayısının 100 ve boyut sayısının 100 olarak belirlendiği senaryoda en yüksek değer (0,807) elde edilmiştir.

Her iki ölçütün de genel ortalamaları incelendiğinde dengeli dağılım ve yüksek düzey gürültü durumunda iki ölçü arasındaki en belirgin farklılık RFerns’de görülmektedir ve ilgili senaryoların ortalama doğruluk değerlerinin ortalaması 0,61 iken ortalama F değerlerinin ortalaması 0,53’tür. DVM için bu değerler sırasıyla 0,47 ve 0,43’tür. Diğer üç yöntemde ise bu ortalamalar birbirine çok daha yakındır. Orta düzey gürültü için hem doğruluk hem de F değerlerinin genel olarak arttığı söylenebilir. Burada da yine doğruluk ve F değerleri arasındaki fark RFerns ve DVM’de diğer yöntemlere göre daha fazladır.

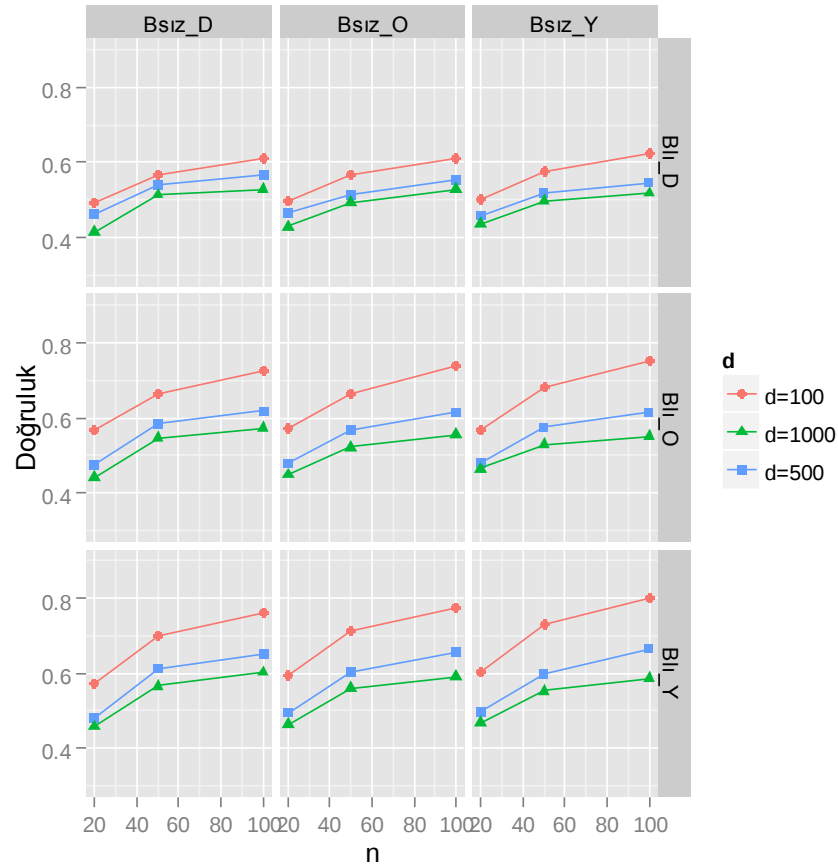


Şekil 4.1. Dengeli Dağılım ve Yüksek Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre DVM'nin Sınıflama Performansı

Bağımsız değişkenlerin %90'ının gürültü olarak belirlendiği ve sınıfların dengeli dağıldığı durumda gözlem sayısı, boyut sayısı ve korelasyon düzeylerindeki değişimlere göre DVM'lerden elde edilen ortalama doğruluk değerleri Şekil 4.1'de gösterilmektedir. Bağımsız değişkenlerin kendi aralarındaki korelasyon düzeylerinin DVM performansında ciddi bir değişikliğe neden olmadığı görülmektedir. Bağımlı değişken ile bağımsız değişkenler arasındaki korelasyonun ise doğruluk değerlerini bir miktar yükselttiği görülmektedir. Literatüre paralel olarak gözlem sayısı arttıkça performans artmaktayken boyut artışıyla birlikte performans azalmaktadır. Boyut sayısı 500 ve 1000 iken sonuçların birbirine daha yaklaşmış olması ve $d=100$ için DVM'nin bağımlı değişkenler ile bağımsız değişkenler arasındaki ilişkiden daha çok etkilenmesi dikkat çekmektedir.

UAA'nın dengeli dağılım ve yüksek gürültü düzeyi durumunda performansı Şekil 4.2'de görüldüğü üzere bağımlı değişken ile bağımsız değişkenler arasındaki

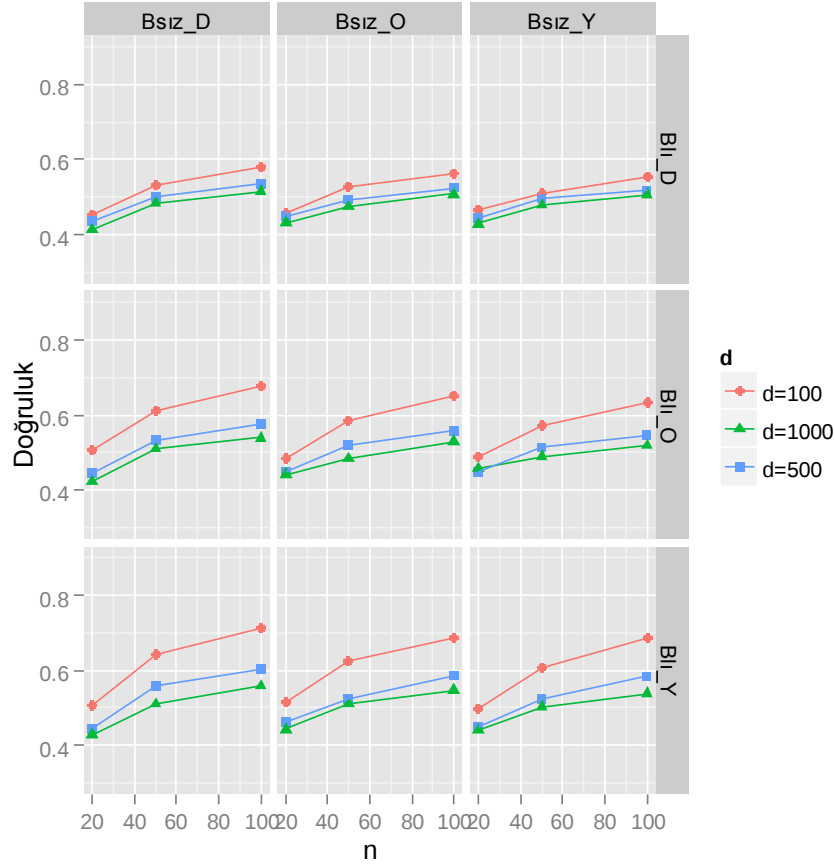
korelasyondan etkilenmektedir. Her ne kadar bağımsız değişkenlerin sadece %10'u bağımlı değişken ile ilişkili olsa da bu ilişki miktarı 0,70 ile 0,90 arasında olduğunda UAA'nın doğru sınıflama oranı 0,80'lere ulaşmaktadır. Bu başarı görece boyutun düşük olduğu $d=100$ durumu için geçerli olmakla birlikte, $d=1000$ için bağımlı değişken ile bağımsız değişkenler arasında düşük ilişki olduğunda 0,50'lerde olan doğruluk oranı yüksek ilişki olduğunda 0,60'lara çıkmaktadır.



Şekil 4.2. Dengeli Dağılım ve Yüksek Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre UAA'nın Sınıflama Performansı

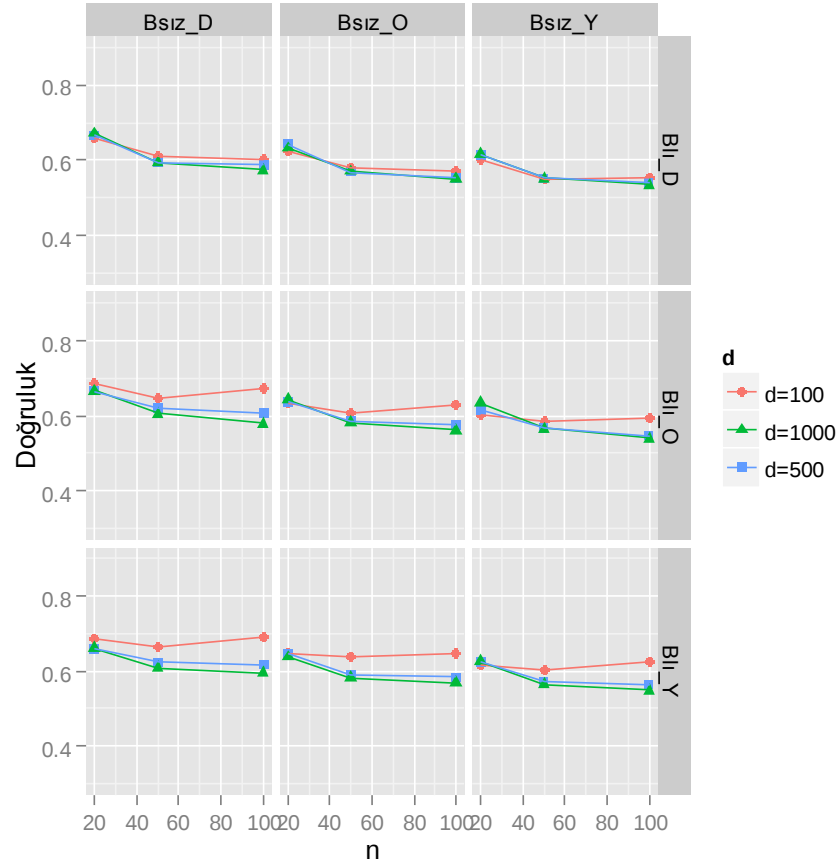
Şekil 4.3'te ise RForest'in dengeli dağılım ve yüksek gürültü düzeyinde farklı senaryolara göre oluşturulmuş veri yapılarındaki performansı görülmektedir. Burada doğruluk oranları 0,412 ile 0,733 arasında değişmektedir. En başarılı performans gözlem ve boyut sayısının 100, bağımlı değişken ile bağımsız değişken arasındaki ilişki düzeyinin orta ve bağımsız değişkenler arasında zayıf ilişkinin olduğu senaryo için gerçekleşmiştir. $d=100$ için bağımlı değişkenin bağımsız değişkenler ile olan

ilişkisinin miktar olarak küçük olsa da diğer boyutlara göre daha belirgin etki yarattığı görülmektedir.



Şekil 4.3. Dengeli Dağılım ve Yüksek Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre RForest'in Sınıflama Performansı

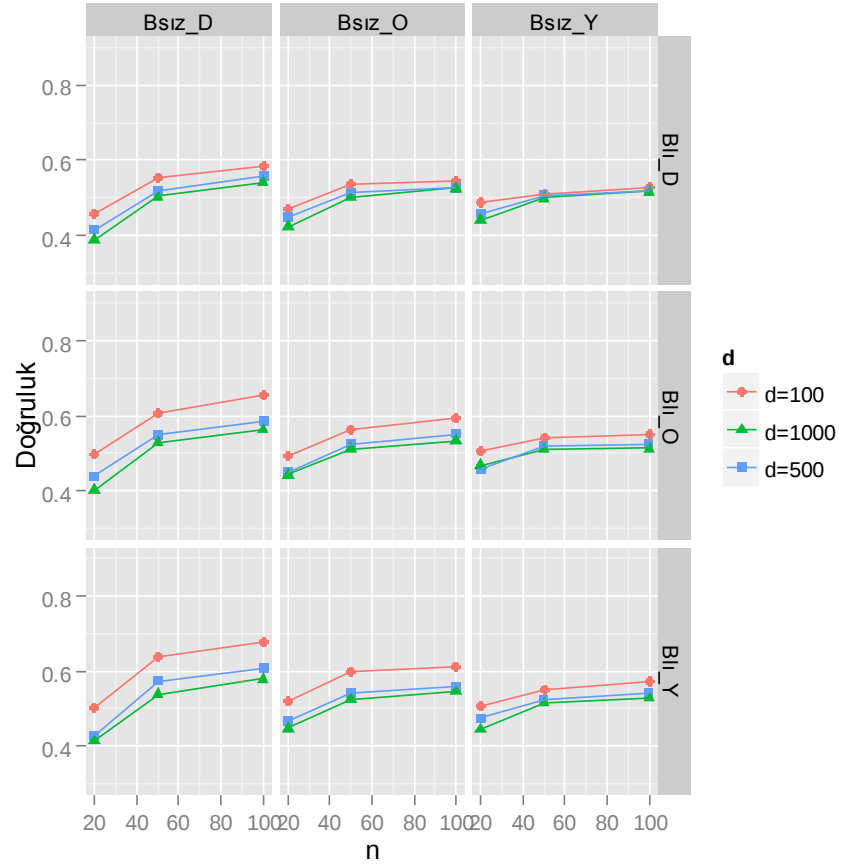
Şekil 4.4'te RFerns yönteminin ve Şekil 4.5'te NB yönteminin ortalama doğru sınıflama oranlarının senaryolara göre değişimi gösterilmektedir. Şekilden de görüldüğü üzere RFerns'in performans değişimi önceki üç yönteme göre farklılık göstermektedir. RFerns'in düşük gözlem sayısında ($n=20$) daha iyi performans göstermiş olması en belirgin özellik olarak görülmektedir. Bağımlı değişken ile bağımsız değişkenler arasındaki ilişki düşükken farklı boyutlarda sınıflama performansları benzerlik gösterirken söz konusu ilişki arttığında düşük boyut sayısında RFerns performansı da iyileşme göstermektedir. Dikkat çeken noktalardan bir diğeri ise bağımsız değişkenler arasındaki ilişki miktarı artarken sınıflama performansında küçük miktarda da olsa gözlenen azalmadır.



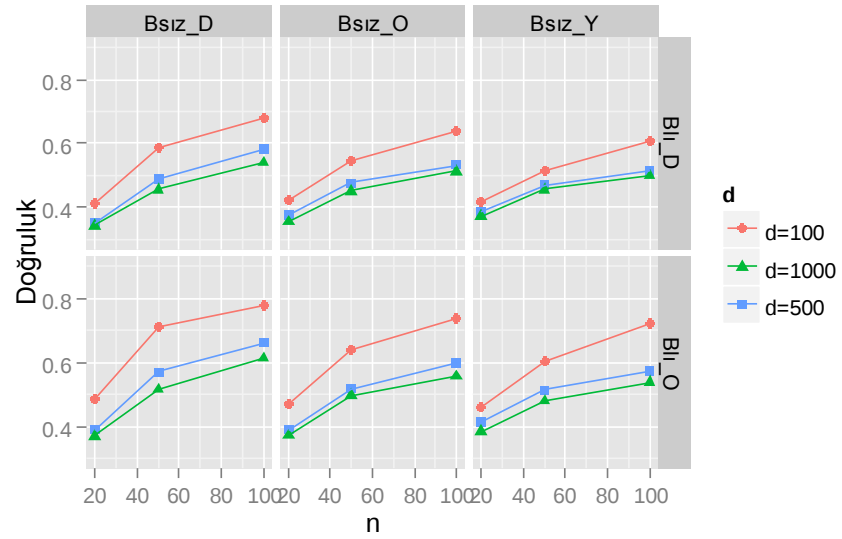
Şekil 4.4. Dengeli Dağılım ve Yüksek Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre RFerns'in Sınıflama Performansı

NB sınıflayıcı incelendiğinde (Şekil 4.5) dengeli dağılım ve yüksek gürültü düzeyine sahip veri setlerinde bağımlı değişken ile bağımsız değişkenler arasındaki ilişki arttışının sınıflama performansında çok küçük miktarda artışa neden olduğu görülmektedir. Yine RFerns dışındaki diğer yöntemlerde olduğu gibi genel olarak gözlem sayısındaki artış doğru sınıflama başarısını artırırken boyut sayısındaki artış doğru sınıflama başarısını düşürmektedir. Burada dikkat çeken bir nokta ise bağımsız değişkenler arasındaki ilişki arttıkça özellikle yüksek gözlem sayıları için ($n=50$ ve $n=100$) farklı boyutlarda sınıflama performanslarının birbirine yaklaştığı olmasıdır.

Gürültü düzeyinin orta düzeyde (%50) olduğu ve yine dengeli dağılım durumu için diğer bileşenlere göre oluşturulan veri yapılarında yöntemlerden elde edilen ortalama doğru sınıflama sonuçları Şekil 4.6 – Şekil 4.10 arasında gösterilemektedir.

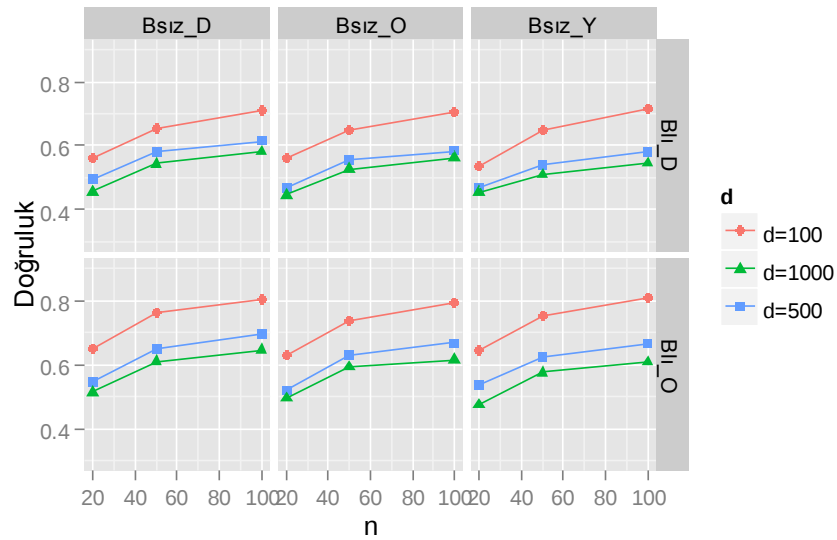


Şekil 4.5. Dengeli Dağılım ve Yüksek Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre NB'nin Sınıflama Performansı



Şekil 4.6. Dengeli Dağılım ve Orta Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre DVM'nin Sınıflama Performansı

Dengeli dağılım durumunda gürültü düzeyi azaltıldığında DVM'nin performanslarında genel olarak iyileşme görülmektedir. Örneğin Şekil 4.6'da görüldüğü üzere $n=100$, $d=100$, bağımlı değişken ile bağımsız değişken arasında yüksek korelasyon, bağımsız değişkenler arasında düşük korelasyon olduğu senaryo DVM'nin en başarılı olduğu senaryodur. Burada doğruluk değerleri ortalaması 0,777 iken gürültü düzeyi arttırıldığında aynı koşullar için DVM'nin ortalama doğruluk değeri 0,681'dir (Bkz. Tablo 4.1) Orta düzey gürültü varlığında diğer etkenler değerlendirildiğinde ise DVM'nin sınıflama performansının yüksek gürültü düzeyi ile benzer eğilim gösterdiği görülmektedir.

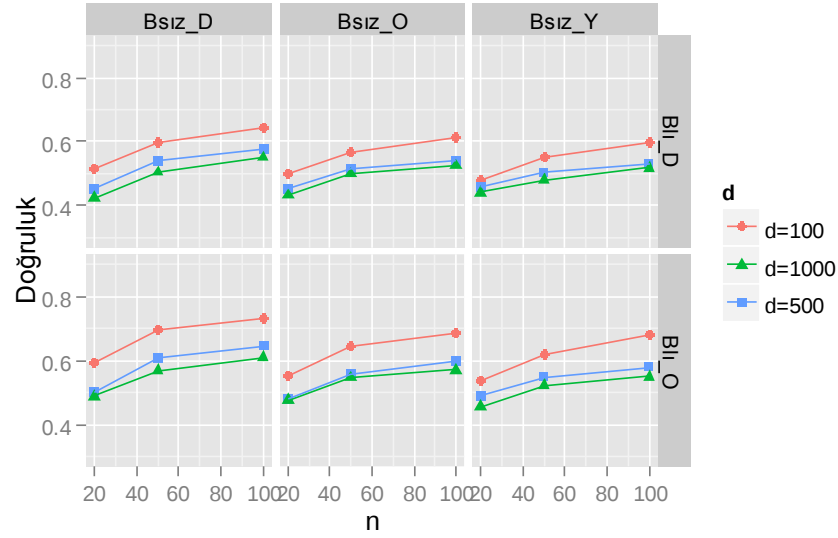


Şekil 4.7. Dengeli Dağılım ve Orta Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre UAA'nın Sınıflama Performansı

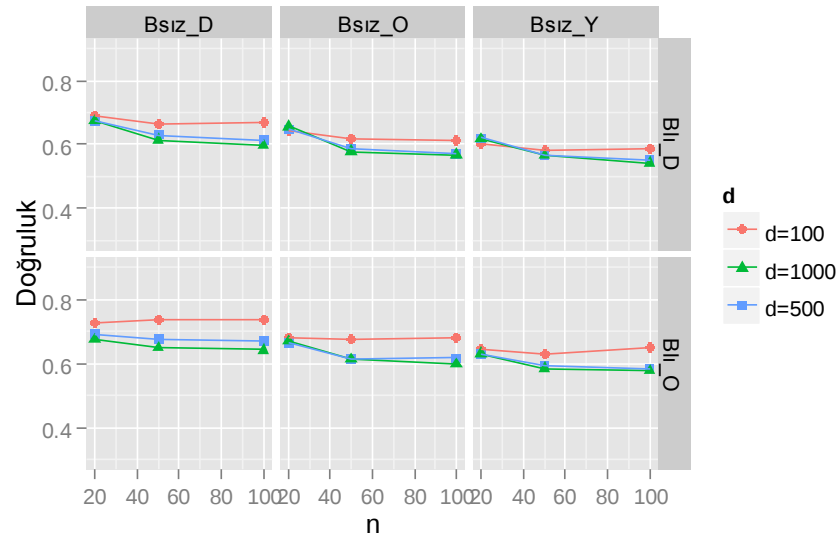
UAA'nın da gürültü oranındaki azalmadan pozitif yönde etkilendiği Şekil 4.2 ile Şekil 4.7 kıyaslandığında açıkça görülmektedir. Bağımsız değişkenler arasındaki ilişkinin belirgin bir etkisi görülmemekle birlikte bağımlı değişken ile bağımsız değişkenler arasındaki ilişki düzeyi sınıflama performansı açısından olumlu bir etki göstermektedir.

Şekil 4.8'de dengeli dağılım ve orta düzey gürültü varlığında RForest'in sınıflama başarısının UAA yönteminin altında olduğu görülmektedir. Ayrıca gözlem sayısı dikkate alındığında aynı koşullarda RForest gözlem sayısı değişikliğinden DVM ve UAA'ya göre daha az etkilenmektedir. Örneğin bağımsız değişkenlerin

hem kendi aralarında hem de bağımlı değişken ile aralarındaki ilişki düzeyi düşük ve $d=100$ iken gözlem sayısı 20’den 100’e çıktığında DVM’nin ortalama doğruluk değerinde yaklaşık 0,30’luk artış olurken, RForest’ta bu artış miktarı yaklaşık 0,15 civarındadır.

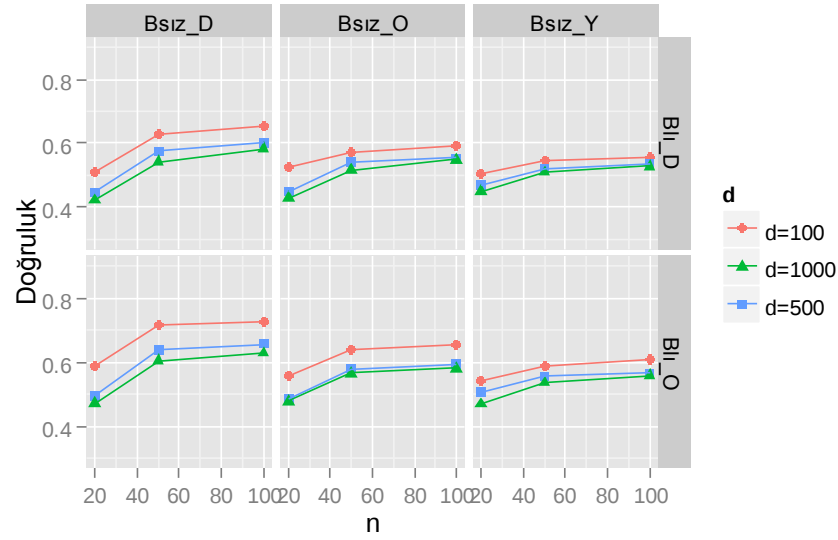


Şekil 4.8. Dengeli Dağılım ve Orta Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre RForest’in Sınıflama Performansı



Şekil 4.9. Dengeli Dağılım ve Orta Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre RFerns’in Sınıflama Performansı

RFerns yüksek gürültü düzeyi varlığında olduğu gibi orta düzey gürültü varlığında da dengeli dağılım için belirlenen koşullarda diğer yöntemlere göre farklı eğilim göstermektedir (Şekil 4.9). Yine yüksek boyutlarda performans değerleri birbirine yaklaşmakta ve yine düşük gözlem sayısında daha iyi performans göstermektedir. Diğer taraftan bağımsız değişkenlerin kendi aralarındaki ilişki miktarındaki artışın yüksek gürültü düzeyine karşın burada bir miktar daha belirgin etki yarattığı görülmektedir.



Şekil 4.10. Dengeli Dağılım ve Orta Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre NB'nin Sınıflama Performansı

Şekil 4.10'da görüldüğü üzere bağımsız değişkenler arasındaki ilişki düzeyi arttıkça NB yönteminin sınıflama performansı bir miktar azalmakta ve boyut sayısının sınıflama performansı üzerinde etkisi azalmaktadır. Bununla birlikte yine bağımsız değişkenler arasındaki ilişki arttıkça farklı gözlem sayıları için sonuçlar birbirine yaklaşmaktadır. Bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin ise sınıflama performansını pozitif yönde etkilediği görülmektedir.

Dengesiz sınıf dağılımı durumunda diğer bileşenlere göre oluşturulan senaryolara ilişkin sınıflama yöntemlerinin ortalama doğruluk değerleri Tablo 4.3'te ortalama F değerleri ise Tablo 4.4'te gösterilmektedir. UAA yöntemi literatürde de dengesiz dağılım durumu için kabul görmüş başarısızlığı nedeniyle karşılaştırmaya dahil edilmemiştir.

Tablo 4.3. Dengesiz Dağılım Durumunda Diğer Senaryolara Göre Doğruluk Değerleri

No	G.O.	Blı Bsız	Bsız Bsız	n	d	Doğruluk			
						DVM	RForest	RFerns	NB
1	Y	D	D	20	100	0,630	0,577	0,374	0,572
2	Y	D	D	20	500	0,648	0,601	0,353	0,609
3	Y	D	D	20	1000	0,648	0,599	0,358	0,612
4	Y	D	D	50	100	0,631	0,623	0,388	0,602
5	Y	D	D	50	500	0,640	0,618	0,365	0,592
6	Y	D	D	50	1000	0,638	0,619	0,365	0,594
7	Y	D	D	100	100	0,652	0,648	0,397	0,608
8	Y	D	D	100	500	0,649	0,640	0,366	0,601
9	Y	D	D	100	1000	0,645	0,637	0,363	0,599
10	Y	D	O	20	100	0,624	0,576	0,394	0,549
11	Y	D	O	20	500	0,643	0,583	0,369	0,573
12	Y	D	O	20	1000	0,625	0,582	0,383	0,572
13	Y	D	O	50	100	0,629	0,607	0,411	0,555
14	Y	D	O	50	500	0,630	0,599	0,382	0,552
15	Y	D	O	50	1000	0,634	0,598	0,382	0,551
16	Y	D	O	100	100	0,647	0,629	0,428	0,556
17	Y	D	O	100	500	0,646	0,632	0,395	0,554
18	Y	D	O	100	1000	0,645	0,628	0,384	0,561
19	Y	D	Y	20	100	0,607	0,561	0,437	0,519
20	Y	D	Y	20	500	0,621	0,569	0,397	0,542
21	Y	D	Y	20	1000	0,630	0,574	0,388	0,554
22	Y	D	Y	50	100	0,627	0,594	0,439	0,528
23	Y	D	Y	50	500	0,623	0,592	0,417	0,532
24	Y	D	Y	50	1000	0,626	0,599	0,405	0,540
25	Y	D	Y	100	100	0,647	0,634	0,457	0,536
26	Y	D	Y	100	500	0,643	0,622	0,417	0,531
27	Y	D	Y	100	1000	0,647	0,621	0,404	0,535
28	Y	O	D	20	100	0,645	0,618	0,368	0,601
29	Y	O	D	20	500	0,647	0,610	0,356	0,618
30	Y	O	D	20	1000	0,641	0,598	0,363	0,610
31	Y	O	D	50	100	0,644	0,651	0,382	0,631
32	Y	O	D	50	500	0,635	0,621	0,372	0,603
33	Y	O	D	50	1000	0,639	0,622	0,364	0,606
34	Y	O	D	100	100	0,673	0,692	0,424	0,673
35	Y	O	D	100	500	0,649	0,654	0,373	0,618
36	Y	O	D	100	1000	0,646	0,640	0,364	0,612
37	Y	O	O	20	100	0,617	0,588	0,405	0,564
38	Y	O	O	20	500	0,628	0,575	0,378	0,556
39	Y	O	O	20	1000	0,623	0,571	0,383	0,570
40	Y	O	O	50	100	0,630	0,641	0,436	0,591
41	Y	O	O	50	500	0,628	0,605	0,395	0,558
42	Y	O	O	50	1000	0,638	0,609	0,378	0,568

Tablo 4.3. (Devam)

No	G.O.	Bl Bsız	Bsız Bsız	n	d	Doğruluk			
						DVM	RForest	RFerns	NB
43	Y	O	O	100	100	0,674	0,680	0,461	0,604
44	Y	O	O	100	500	0,649	0,640	0,399	0,573
45	Y	O	O	100	1000	0,648	0,633	0,386	0,566
46	Y	O	Y	20	100	0,614	0,575	0,433	0,537
47	Y	O	Y	20	500	0,610	0,563	0,407	0,544
48	Y	O	Y	20	1000	0,628	0,569	0,384	0,545
49	Y	O	Y	50	100	0,630	0,632	0,460	0,549
50	Y	O	Y	50	500	0,633	0,605	0,412	0,537
51	Y	O	Y	50	1000	0,627	0,595	0,402	0,535
52	Y	O	Y	100	100	0,667	0,679	0,486	0,565
53	Y	O	Y	100	500	0,642	0,631	0,431	0,545
54	Y	O	Y	100	1000	0,643	0,624	0,413	0,540
55	Y	Y	D	20	100	0,634	0,614	0,376	0,596
56	Y	Y	D	20	500	0,646	0,636	0,369	0,625
57	Y	Y	D	20	1000	0,643	0,609	0,360	0,615
58	Y	Y	D	50	100	0,653	0,679	0,397	0,667
59	Y	Y	D	50	500	0,649	0,678	0,409	0,683
60	Y	Y	D	50	1000	0,643	0,632	0,362	0,617
61	Y	Y	D	100	100	0,693	0,722	0,445	0,702
62	Y	Y	D	100	500	0,681	0,701	0,478	0,693
63	Y	Y	D	100	1000	0,654	0,655	0,359	0,632
64	Y	Y	O	20	100	0,638	0,617	0,398	0,589
65	Y	Y	O	20	500	0,631	0,585	0,379	0,568
66	Y	Y	O	20	1000	0,648	0,600	0,362	0,596
67	Y	Y	O	50	100	0,644	0,656	0,441	0,617
68	Y	Y	O	50	500	0,628	0,613	0,399	0,572
69	Y	Y	O	50	1000	0,639	0,615	0,378	0,578
70	Y	Y	O	100	100	0,686	0,701	0,480	0,634
71	Y	Y	O	100	500	0,651	0,650	0,410	0,584
72	Y	Y	O	100	1000	0,649	0,637	0,387	0,576
73	Y	Y	Y	20	100	0,607	0,593	0,444	0,553
74	Y	Y	Y	20	500	0,617	0,569	0,397	0,543
75	Y	Y	Y	20	1000	0,627	0,583	0,387	0,562
76	Y	Y	Y	50	100	0,641	0,651	0,470	0,561
77	Y	Y	Y	50	500	0,631	0,608	0,416	0,544
78	Y	Y	Y	50	1000	0,628	0,603	0,404	0,540
79	Y	Y	Y	100	100	0,681	0,705	0,503	0,573
80	Y	Y	Y	100	500	0,648	0,641	0,433	0,556
81	Y	Y	Y	100	1000	0,644	0,629	0,413	0,546
82	O	D	D	20	100	0,631	0,606	0,381	0,592
83	O	D	D	20	500	0,644	0,611	0,362	0,610
84	O	D	D	20	1000	0,643	0,609	0,359	0,620
85	O	D	D	50	100	0,645	0,647	0,396	0,640

Tablo 4.3. (Devam)

No	G.O.	Bl Bsız	Bsız Bsız	n	d	Doğruluk			
						DVM	RForest	RFerns	NB
86	O	D	D	50	500	0,639	0,634	0,377	0,623
87	O	D	D	50	1000	0,637	0,623	0,369	0,615
88	O	D	D	100	100	0,671	0,676	0,451	0,659
89	O	D	D	100	500	0,651	0,653	0,397	0,638
90	O	D	D	100	1000	0,646	0,646	0,380	0,623
91	O	D	O	20	100	0,616	0,591	0,410	0,566
92	O	D	O	20	500	0,634	0,585	0,382	0,572
93	O	D	O	20	1000	0,639	0,593	0,366	0,597
94	O	D	O	50	100	0,630	0,622	0,444	0,586
95	O	D	O	50	500	0,632	0,614	0,403	0,563
96	O	D	O	50	1000	0,628	0,606	0,394	0,567
97	O	D	O	100	100	0,657	0,651	0,482	0,598
98	O	D	O	100	500	0,647	0,636	0,430	0,580
99	O	D	O	100	1000	0,651	0,637	0,407	0,570
100	O	D	Y	20	100	0,615	0,569	0,437	0,530
101	O	D	Y	20	500	0,619	0,576	0,408	0,558
102	O	D	Y	20	1000	0,622	0,573	0,394	0,563
103	O	D	Y	50	100	0,627	0,611	0,474	0,553
104	O	D	Y	50	500	0,625	0,597	0,420	0,536
105	O	D	Y	50	1000	0,629	0,601	0,407	0,548
106	O	D	Y	100	100	0,650	0,649	0,491	0,556
107	O	D	Y	100	500	0,644	0,628	0,440	0,545
108	O	D	Y	100	1000	0,646	0,629	0,421	0,546
109	O	O	D	20	100	0,650	0,662	0,381	0,651
110	O	O	D	20	500	0,643	0,629	0,366	0,620
111	O	O	D	20	1000	0,635	0,607	0,367	0,612
112	O	O	D	50	100	0,684	0,712	0,464	0,721
113	O	O	D	50	500	0,643	0,662	0,397	0,666
114	O	O	D	50	1000	0,638	0,645	0,379	0,644
115	O	O	D	100	100	0,755	0,741	0,572	0,736
116	O	O	D	100	500	0,672	0,686	0,456	0,680
117	O	O	D	100	1000	0,654	0,666	0,410	0,653
118	O	O	O	20	100	0,643	0,635	0,428	0,608
119	O	O	O	20	500	0,633	0,606	0,393	0,593
120	O	O	O	20	1000	0,629	0,595	0,389	0,584
121	O	O	O	50	100	0,659	0,674	0,510	0,647
122	O	O	O	50	500	0,632	0,632	0,431	0,607
123	O	O	O	50	1000	0,633	0,622	0,404	0,594
124	O	O	O	100	100	0,717	0,705	0,572	0,652
125	O	O	O	100	500	0,657	0,659	0,479	0,609
126	O	O	O	100	1000	0,651	0,646	0,438	0,605
127	O	O	Y	20	100	0,618	0,622	0,471	0,578
128	O	O	Y	20	500	0,630	0,593	0,420	0,571

Tablo 4.3. (Devam)

No	G.O.	Bl Bsız	Bsız Bsız	n	d	Doğruluk			
						DVM	RForest	RFerns	NB
129	O	O	Y	20	1000	0,623	0,587	0,398	0,561
130	O	O	Y	50	100	0,648	0,655	0,526	0,606
131	O	O	Y	50	500	0,627	0,618	0,456	0,567
132	O	O	Y	50	1000	0,628	0,615	0,425	0,563
133	O	O	Y	100	100	0,706	0,700	0,567	0,611
134	O	O	Y	100	500	0,648	0,647	0,483	0,580
135	O	O	Y	100	1000	0,648	0,641	0,453	0,572

Sınıf dağılımının dengesiz olduğu durumlar için belirlenen senaryolara bakıldığında genel olarak DVM yönteminin diğer üç yönteme göre daha yüksek sınıflama doğruluğuna sahip olduğu görülmektedir (Tablo 4.2). Diğer taraftan bağımsız değişkenler arasındaki korelasyon düzeyi düşükken NB yönteminin de performansında artış olduğu görülmektedir. Ancak daha öncede belirtildiği gibi doğruluk değerleri ile dengesiz dağılım durumunda yapılacak bu gibi yorumlar yanlış yönlendirmelere neden olabilmektedir. Ek 2 incelendiğinde görüleceği üzere özellikle dengesiz dağılım durumu için yüksek doğruluk elde edilmiş senaryoların çoğunda duyarlılık ve seçicilik değerleri arasında büyük farklılıklar bulunmaktadır. Örneğin gürültü düzeyinin orta, bağımsız değişkenlerin kendi aralarındaki korelasyon düzeyinin orta, bağımlı değişkenler ile arasında korelasyon düzeyinin ise yüksek olduğu 100 gözlem ve 100 değişken içeren senaryo için doğruluk değeri 0,71 civarında iken duyarlılık değeri (0,27) oldukça düşük, seçicilik değeri (0,95) ise oldukça yüksektir. Bu nedenle dengesiz dağılım durumu ile dengeli dağılım durumunu karşılaştırmak için Tablo 4.4 üzerinden senaryoları değerlendirmenin faydalı olduğu düşünülmektedir. Bununla birlikte bütünlük sağlamak açısından grafiklerde doğruluk değerleri kullanılmış ancak her grafik için değerlendirmeler kendi içinde yapılmıştır. Her bir tekrar (iterasyon) için F değerlerinin elde edilmediği senaryolar tutarlı karşılaştırma yapılabilmesi için Tablo 4.4’de gösterilmemiştir.

Tablo 4.4’te F değerleri incelendiğinde doğruluk değerlerinde gözlenenin aksine, dengesiz dağılım durumunda DVM’nin sınıflama performansının oldukça düştüğü görülmektedir. Buna karşın RFerns’in sınıflama performansının dengeli dağılım durumuna göre çok değişmediği ve F değerlerinin diğer yöntemlere göre daha yüksek olduğu görülmektedir. Ancak genel olarak 0,50 civarındaki bu

değerlerin yüksek başarı göstermediği açıktır. Diğer taraftan DVM, RForest ve NB'nin boyut artışından olumsuz etkilendiği, RFerns'in ise bu artıştan çok fazla etkilenmediği dikkat çekmektedir.

Tablo 4.4. Dengesiz Dağılım Durumunda Diğer Senaryolara Göre F Değerleri

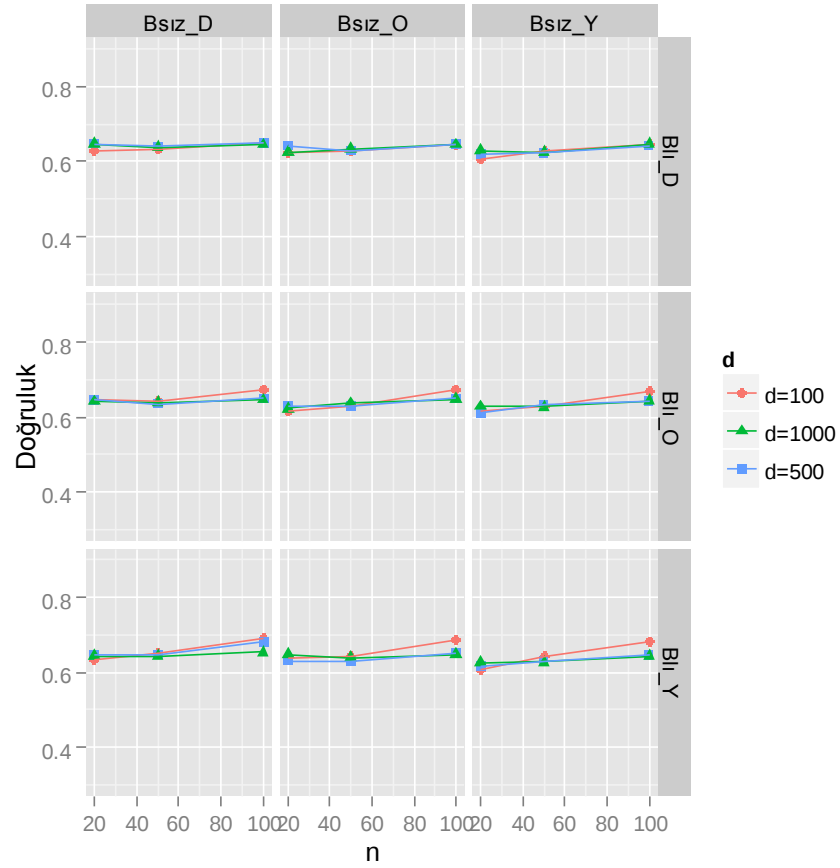
No	G.O.	Blı Bsız	Bsız Bsız	n	d	F Ölçütü			
						DVM	RForest	RFerns	NB
4	Y	D	D	50	100	0,038	-	0,525	0,366
5	Y	D	D	50	500	0,012	0,136	0,517	0,259
6	Y	D	D	50	1000	-	0,117	0,519	0,197
7	Y	D	D	100	100	0,068	0,219	0,519	0,421
8	Y	D	D	100	500	0,013	0,138	0,514	0,369
9	Y	D	D	100	1000	0,007	0,106	0,517	0,330
13	Y	D	O	50	100	0,052	0,218	0,508	0,367
14	Y	D	O	50	500	0,020	0,161	0,509	0,324
15	Y	D	O	50	1000	0,019	0,139	0,515	0,283
16	Y	D	O	100	100	0,054	0,194	0,502	0,400
17	Y	D	O	100	500	0,021	0,160	0,505	0,397
18	Y	D	O	100	1000	0,014	0,140	0,511	0,373
22	Y	D	Y	50	100	0,055	0,220	0,486	0,375
23	Y	D	Y	50	500	0,044	0,198	0,506	0,362
24	Y	D	Y	50	1000	0,024	0,178	0,517	0,349
25	Y	D	Y	100	100	0,056	0,228	0,483	0,399
26	Y	D	Y	100	500	0,024	0,167	0,495	0,397
27	Y	D	Y	100	1000	0,019	0,140	0,495	0,387
31	Y	O	D	50	100	0,070	0,267	0,520	0,399
32	Y	O	D	50	500	-	0,165	0,521	0,292
33	Y	O	D	50	1000	0,006	0,120	0,518	0,220
34	Y	O	D	100	100	0,186	0,347	0,534	0,517
35	Y	O	D	100	500	0,021	0,174	0,518	0,402
36	Y	O	D	100	1000	0,010	0,130	0,517	0,361
40	Y	O	O	50	100	0,098	0,313	0,526	0,432
41	Y	O	O	50	500	-	0,174	0,518	0,335
42	Y	O	O	50	1000	-	0,142	0,515	0,298
43	Y	O	O	100	100	0,186	0,327	0,518	0,460
44	Y	O	O	100	500	0,025	0,184	0,508	0,414
45	Y	O	O	100	1000	0,020	0,144	0,509	0,379
49	Y	O	Y	50	100	0,078	0,287	0,507	0,402
50	Y	O	Y	50	500	0,046	0,204	0,499	0,370
51	Y	O	Y	50	1000	0,027	0,178	0,514	0,350
52	Y	O	Y	100	100	0,149	0,342	0,509	0,434
53	Y	O	Y	100	500	0,032	0,188	0,507	0,420
54	Y	O	Y	100	1000	0,019	0,153	0,507	0,401
58	Y	Y	D	50	100	0,102	0,341	0,523	0,461

Tablo 4.4. (Devam)

No	G.O.	Bh Bsız	Bsız Bsız	n	d	F Ölçütü			
						DVM	RForest	RFerns	NB
59	Y	Y	D	50	500	0,123	0,345	0,530	0,494
60	Y	Y	D	50	1000	-	0,144	0,515	0,239
61	Y	Y	D	100	100	0,272	0,431	0,543	0,561
62	Y	Y	D	100	500	0,248	0,404	0,553	0,570
63	Y	Y	D	100	1000	0,013	0,144	0,509	0,381
67	Y	Y	O	50	100	0,128	0,322	0,523	0,445
68	Y	Y	O	50	500	0,028	0,196	0,522	0,354
69	Y	Y	O	50	1000	0,023	0,166	0,513	0,320
70	Y	Y	O	100	100	0,256	0,391	0,539	0,500
71	Y	Y	O	100	500	0,039	0,207	0,514	0,424
72	Y	Y	O	100	1000	0,017	0,154	0,508	0,391
76	Y	Y	Y	50	100	0,131	0,338	0,510	0,424
77	Y	Y	Y	50	500	0,046	0,203	0,506	0,377
78	Y	Y	Y	50	1000	-	0,184	0,510	0,354
79	Y	Y	Y	100	100	0,238	0,417	0,518	0,447
80	Y	Y	Y	100	500	0,034	0,204	0,506	0,422
81	Y	Y	Y	100	1000	0,025	0,172	0,505	0,400
85	O	D	D	50	100	0,097	0,281	0,524	0,441
86	O	D	D	50	500	0,027	0,201	0,520	0,339
87	O	D	D	50	1000	0,015	0,149	0,521	0,281
88	O	D	D	100	100	0,181	0,316	0,536	0,510
89	O	D	D	100	500	0,059	0,225	0,524	0,475
90	O	D	D	100	1000	0,026	0,173	0,521	0,423
94	O	D	O	50	100	0,090	0,266	0,526	0,427
95	O	D	O	50	500	0,045	0,212	0,518	0,362
96	O	D	O	50	1000	0,027	0,181	0,524	0,331
97	O	D	O	100	100	0,143	0,283	0,530	0,468
98	O	D	O	100	500	0,053	0,210	0,516	0,436
99	O	D	O	100	1000	0,038	0,186	0,509	0,406
103	O	D	Y	50	100	0,110	0,276	0,511	0,421
104	O	D	Y	50	500	0,048	0,211	0,504	0,375
105	O	D	Y	50	1000	0,031	0,180	0,515	0,361
106	O	D	Y	100	100	0,127	0,292	0,513	0,441
107	O	D	Y	100	500	0,044	0,199	0,500	0,418
108	O	D	Y	100	1000	0,035	0,183	0,503	0,401
112	O	O	D	50	100	0,294	0,459	0,555	0,583
113	O	O	D	50	500	0,093	0,304	0,528	0,454
114	O	O	D	50	1000	0,049	0,228	0,522	0,380
115	O	O	D	100	100	0,527	0,518	0,601	0,633
116	O	O	D	100	500	0,193	0,352	0,547	0,555
117	O	O	D	100	1000	0,084	0,260	0,531	0,495
121	O	O	O	50	100	0,234	0,399	0,551	0,504
122	O	O	O	50	500	0,082	0,283	0,529	0,441

Tablo 4.4. (Devam)

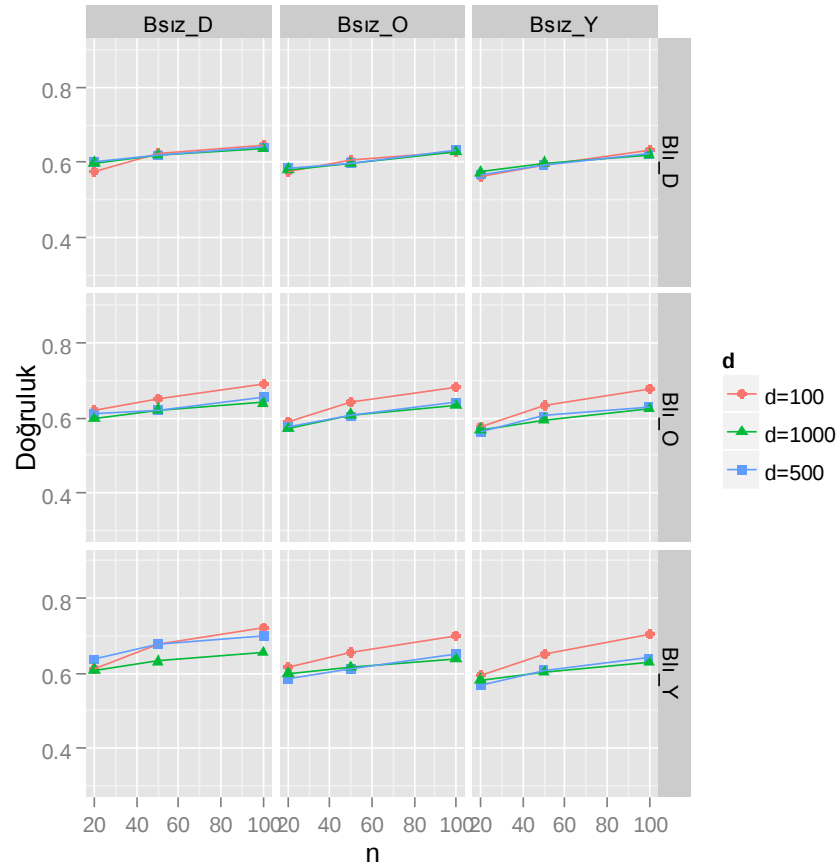
No	G.O.	Bl Bsız	Bsız Bsız	n	d	F Ölçütü			
						DVM	RForest	RFerns	NB
123	O	O	O	50	1000	0,047	0,217	0,520	0,377
124	O	O	O	100	100	0,415	0,457	0,573	0,547
125	O	O	O	100	500	0,118	0,305	0,529	0,481
126	O	O	O	100	1000	0,058	0,234	0,520	0,462
130	O	O	Y	50	100	-	0,382	0,545	0,485
131	O	O	Y	50	500	0,091	0,266	0,519	0,424
132	O	O	Y	50	1000	0,049	0,230	0,517	0,388
133	O	O	Y	100	100	0,353	0,428	0,555	0,506
134	O	O	Y	100	500	0,089	0,267	0,519	0,462
135	O	O	Y	100	1000	0,051	0,226	0,515	0,441



Şekil 4.11. Dengesiz Dağılım ve Yüksek Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre DVM'nin Sınıflama Performansı

Yüksek düzey gürültü ve dengesiz dağılım durumu için DVM (Şekil 4.11), RForest (Şekil 4.12), RFerns (Şekil 4.13) ve NB (Şekil 4.14) yöntemlerinin ortalama

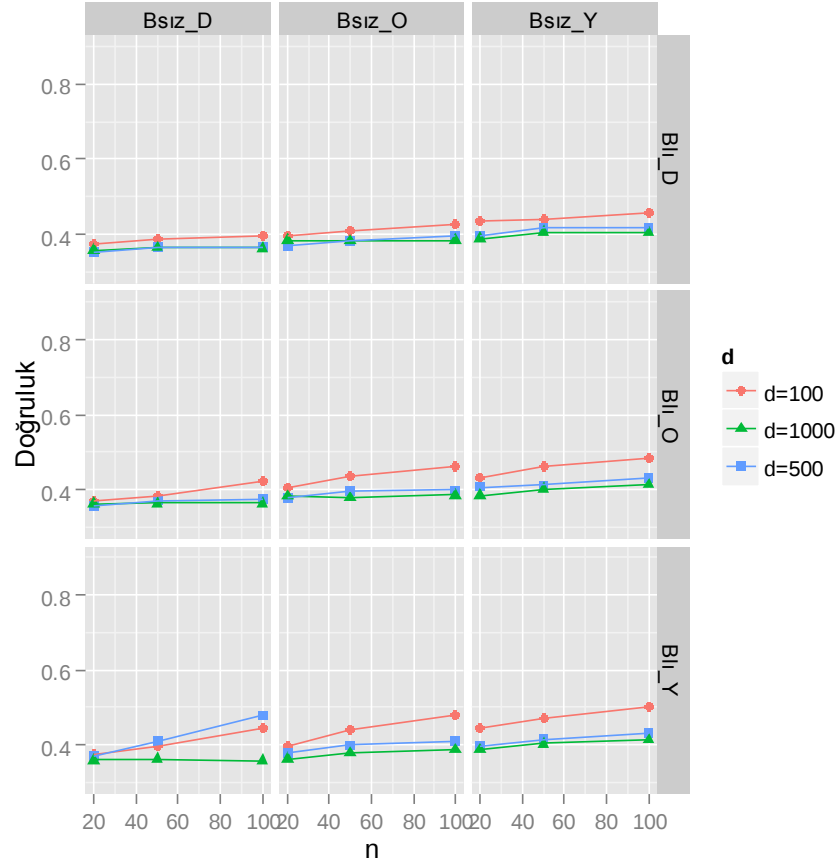
doğru sınıflama değerlerinin veri yapısını belirleyen diğer etkenlerden nasıl etkilendiği grafiklerde görülmektedir. Bağımlı değişken ile bağımsız değişkenler arasındaki korelasyonlar, bağımsız değişkenlerin kendi aralarındaki korelasyonlar, farklı gözlem sayıları ve farklı boyut sayıları için DVM'nin sınıflama eğiliminde ciddi farklılıklar olmadığı görülmektedir. Tablo 4.3'te de rahatlıkla görüleceği üzere söz konusu senaryolar için DVM'nin doğru sınıflama oranı 0,60 civarındadır.



Şekil 4.12. Dengesiz Dağılım ve Yüksek Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre RForest'in Sınıflama Performansı

DVM'nin aksine RForest söz konusu koşullar altında gözlem sayısındaki değişimden bir miktar etkilenmektedir. Gözlem sayısı artışı sınıflama doğruluğunda da artışa neden olmaktadır. Diğer taraftan bağımlı değişken ile bağımsız değişken arasındaki ilişki arttıkça düşük boyutlarda elde edilen doğru sınıflama oranının da küçük de olsa artış gösterdiği görülmektedir.

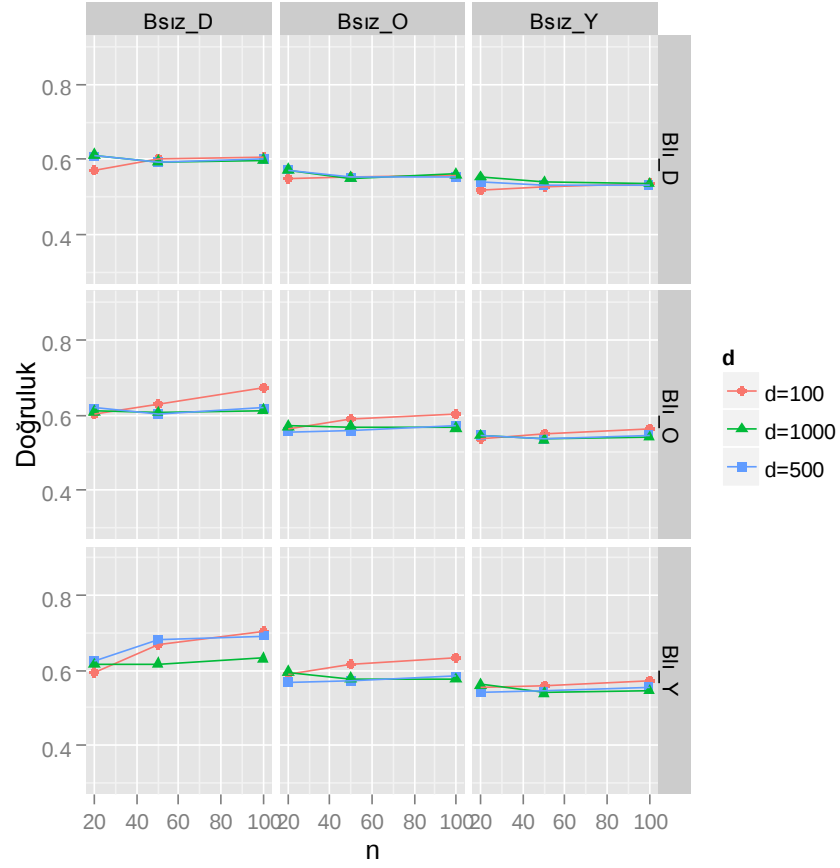
Gürültü düzeyinin yüksek, sınıf dağılımlarının ise dengesiz olduğu senaryolar için RFerns'in sınıflama doğrulukları 0,35-0,50 arasında değişmektedir. Bağımsız değişkenler arasındaki ilişki artışının sınıflama doğruluğunu artırıcı yönde etkisi olduğu ve gözlem sayısı arttıkça RFerns'in sınıflama performansının da bir miktar arttığı Tablo 4.13'te görülmektedir.



Şekil 4.13. Dengesiz Dağılım ve Yüksek Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre RFerns'in Sınıflama Performansı

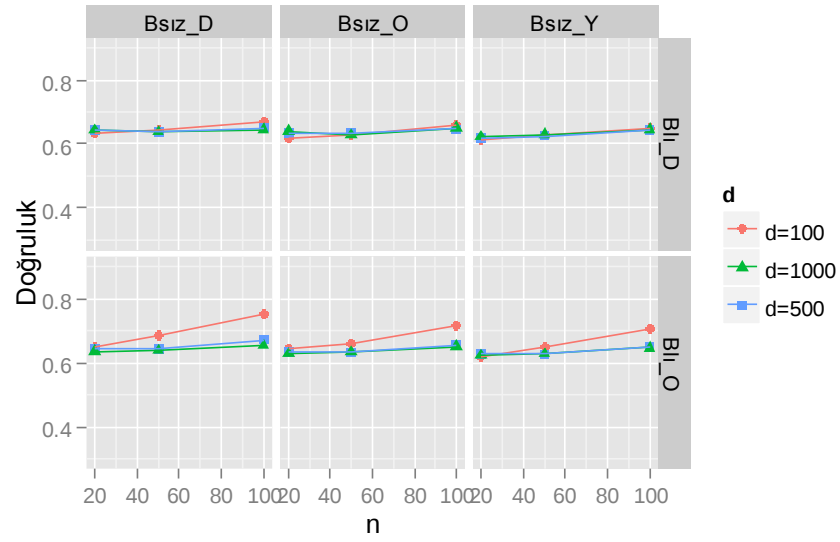
Şekil 4.14'te görüldüğü üzere dengesiz dağılım ve yüksek gürültü düzeyine sahip veri setlerinde NB yöntemi ile sınıflama yapıldığında bağımsız değişkenler arasındaki ilişki arttıkça sınıflama performansında küçük azalmalar yaşanmaktadır. Bununla birlikte bağımlı değişken ile bağımsız değişkenler arasındaki ilişki miktarı arttıkça ise sınıflama performansında küçüğe olsa artışlar yaşandığı görülmektedir. Gözlem ve boyut sayısındaki değişimin ise belirgin bir etkisi görülmemektedir. Yüksek düzey gürültü için belirlenen senaryolar göz önüne alındığında dengesiz

dağılım durumunda NB için doğru sınıflama oranı 0,52 ile 0,70 aralığında değişmektedir.

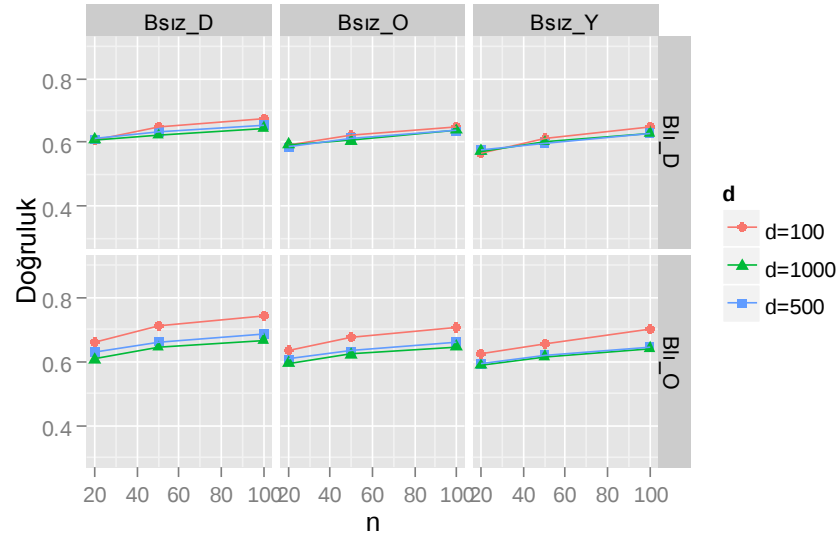


Şekil 4.14. Dengesiz Dağılım ve Yüksek Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre NB'nin Sınıflama Performansı

Gürültü düzeyi %50 olarak ayarlandığında dengesiz dağılıma sahip farklı karakterlerdeki verilerde DVM, RForest, RFerns ve NB yöntemlerinin ortalama sınıflama doğrulukları sırasıyla Şekil 4.15, Şekil 4.16, Şekil 4.17 ve Şekil 4.18'de verilmiştir. Şekil 4.15 incelendiğinde bağımlı değişkenin bağımsız değişkenler ile orta düzey korelasyona sahip olduğu ve boyut sayısının düşük ($d=100$) tutulduğu durumlarda DVM'nin sınıflama performansının gözlem sayısı arttıkça arttığı görülmektedir. Bağımsız değişkenler arasındaki ilişkinin ise performans üzerinde etkisi olmadığı görülmektedir.



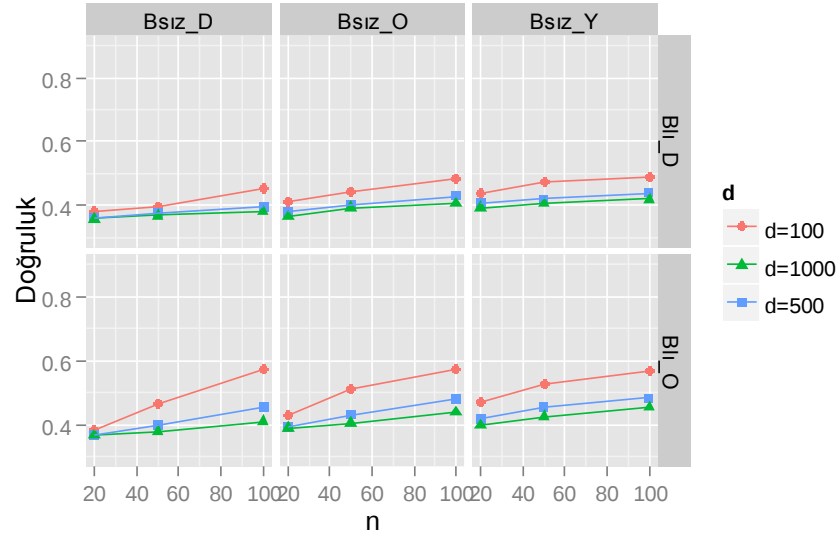
Şekil 4.15. Dengesiz Dağılım ve Orta Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre DVM'nin Sınıflama Performansı



Şekil 4.16. Dengesiz Dağılım ve Orta Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre RForest'in Sınıflama Performansı

Şekil 4.16'da görüldüğü gibi belirlenen tüm koşullar için gözlem sayısı arttıkça RForest'in sınıflama performansı artmaktadır. Bağımlı değişken ile bağımsız değişkenler arasındaki düşük düzeyde ilişki varken bağımsız değişkenlerin kendi aralarındaki ilişkinin sınıflama performansı üzerinde belirgin bir etkisi görülmemekle birlikte bağımlı değişken ile bağımsız değişkenler arasındaki ilişki orta düzeye

yükseldiğinde bağımsız değişkenlerin kendi aralarındaki ilişkinin performans üzerinde ters yönlü bir etkisi olduğu görülmektedir. Diğer taraftan boyut etkisi de bağımlı değişken ile bağımsız değişkenler arasında orta düzey bir ilişki varken daha çok ortaya çıkmaktadır.

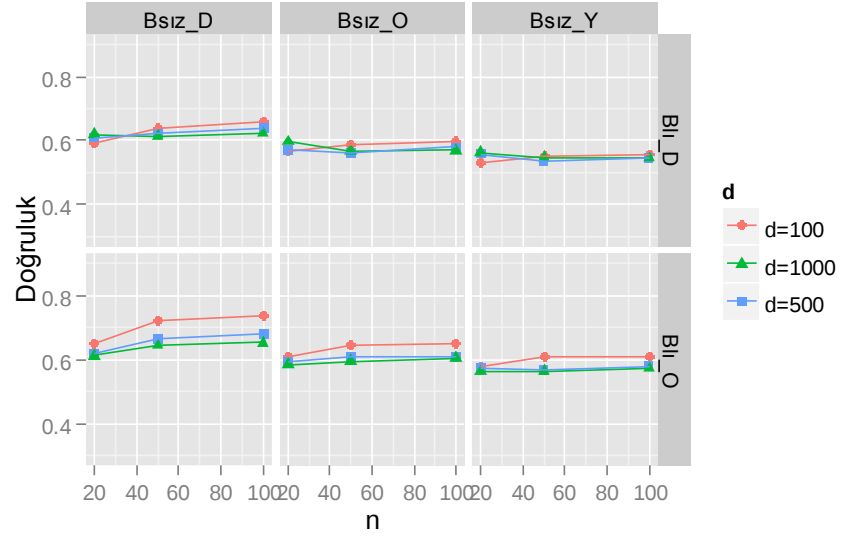


Şekil 4.17. Dengesiz Dağılım ve Orta Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre RFerns'in Sınıflama Performansı

Dengesiz Dağılım ve Orta Gürültü Düzeyi Durumunda RFerns'in davranışları incelendiğinde gözlem sayısından genel olarak etkilendiği görülmektedir. DVM'ye benzer şekilde özellikle bağımlı değişken ile bağımsız değişkenler arasında orta düzey ilişki olduğu ve boyutun düşük olduğu durumda gözlem sayısı artışının performansta belirgin artış yarattığı görülmektedir. Diğer taraftan Şekil 4.17'de de görüldüğü üzere RFerns'in belirlenen tüm koşullar için sağladığı ortalama doğru sınıflama oranı genel olarak 0,60'ın altında kalmaktadır. Bağımsız değişkenlerin kendi aralarındaki ilişkideki ve bağımlı değişkenle aralarındaki ilişkideki artışın sınıflama performansında küçük miktarda artış yarattığını söylemek mümkündür.

NB yönteminin sınıflama performansı sınıf dağılımı dengesiz ve orta düzey gürültü varlığında bağımlı değişken ile bağımsız değişken arasındaki korelasyonun artışına paralel olarak bir miktar artış göstermektedir (Tablo 4.18). Bağımsız değişkenler arasındaki korelasyon düzeyinin artması durumunda ise sınıflama performansında azalma yaşanmaktadır. Bağımlı değişken ile bağımsız değişken

arasında ilişki düşük seviyedeysen boyut değişimlerinin performansı etkilemediği, orta seviyedeysen ise düşük boyutta sınıflama performansının diğer boyutlara göre daha yüksek olduğu görülmektedir.



Şekil 4.18. Dengesiz Dağılım ve Orta Gürültü Düzeyi Durumunda Gözlem ve Boyut Sayısı ile Korelasyon Düzeylerine Göre NB'nin Sınıflama Performansı

5. TARTIŞMA

YBDÖG yapıları sahip veri setlerinde sınıflama problemleri ile ilgili olarak literatürde genellikle bu veri yapılarının klasik yapılardan farklı olmasının göz ardı edildiği dikkat çekmektedir. Bu tez çalışması ile ihmal edilen bu durumun uygulamada yaratabileceği etkiler gözlenmeye çalışılmıştır. Bu bağlamda YBDÖG verilerin sınıflandırılmasında hem sıklıkla kullanılan makine öğrenmesi yöntemlerinin hem de henüz yaygın uygulamaya sahip olmayan yöntemlerin performansları incelenerek bu alandaki boşluğu doldurmak ve öneriler geliştirmek amaçlanmıştır. Daha ileri çalışmalara bir basamak oluşturması açısından da kapsamlı senaryolar kurgulanarak çok yönlü karşılaştırmalar yapılmaya çalışılmıştır.

Literatürde YBDÖG yapıları üzerinde benzetim uygulamaları ile performans karşılaştırılması çok fazla bulunmamaktadır. Var olan çalışmalarda (13,23) da gözlem ve boyut sayısı ile bağımlı değişken ile bağımsız değişkenler arasındaki ilişkilerin ön planda tutulduğu, özellikle mikro dizilim verilerde karşılaşılan en önemli sorunlardan biri olan gürültü değişkenlerinin varlığının benzetim uygulamalarında çoğunlukla ihmal edildiği gözlenmektedir. Dolayısıyla veri yapılarını belirleyen birçok etkeni göz önüne alması açısından bu çalışma diğer çalışmalardan ayrılmaktadır.

Araştırmacıların genellikle klasik uygulamalarda başarılı olan sınıflama yöntemlerini kullanarak farklı tekniklerle boyut sayısı indirgenmiş veriler üzerinde çalıştıkları Bölüm 2’de örnekleri ile ortaya konmuştur. Özellikle DVM ve RForest yöntemlerinin bu çalışmalarda önemli bir yerinin olduğu açıktır (15-19). Ancak tez çalışması sonucunda elde edilen bulgular incelendiğinde araştırmacıların bu konuda daha dikkatli olması gerektiği ortaya çıkmaktadır. DVM açısından sonuçlara bakıldığında boyut artarken yöntemin sınıflama başarısının azaldığı görülmektedir. Gerçek veri uygulamalarında da araştırmacıları yönlendiren etkilerden birinin bu olduğu düşünülmektedir. Çünkü daha önce değinildiği gibi bu çalışmaların çoğunda öncelikle boyut indirgemesi yapılmaktadır, bu da düşük boyutta DVM’nin daha başarılı olması sonucunu doğurmaktadır. Ancak Pamukçu’nun (50) çalışmasında değindiği gibi boyut indirgeme ile genellikle tek değişkenli seçim yapılmasından dolayı değişkenler arası etkileşimler göz ardı edilmektedir. Diğer taraftan gürültü içeren yapılarda PCA gibi yöntemlerin performanslarında kayıplar yaşanmaktadır.

Dolayısıyla boyut indirgeme işleminden kaynaklanabilecek bu olumsuzlukların göz ardı edilmesi literatürdeki uygulamaların da sonuçlarının dikkatle incelenmesi gerektiğini göstermektedir. Bu tez çalışmasında hem gürültü oranlarının dikkate alınmış olmasının hem de tüm değişkenler ile çalışılmış olmasının bu açıdan faydalı bilgiler ortaya koyduğu düşünülmektedir.

Çalışmada dikkate değer noktalardan birisi de görece yeni olan iki sınıflama yönteminin gösterdiği sınıflama performanslarıdır. UAA zaten YBDÖG düzenlere özel olarak geliştirilmiş olmakla birlikte kapsamlı senaryolarda da başarılı sonuçlar vermektedir. Daha çok boyut ve gözlem sayısı üzerine odaklanan çalışmalarda (23) yüksek performansı gösterilmiş olan UAA yönteminin, bağımlı değişken ile bağımsız değişkenler arasındaki ilişkiden ve gürültü düzeyinden etkilendiği bu çalışma ile ortaya konmuş olmaktadır.

Diğer bir yeni yöntem olan RFerns'e bakıldığında ise dengeli dağılım ve düşük gözlem sayısında diğer yöntemlere göre daha başarılı olduğu görülmektedir. RForest'a göre daha basit ve etkili bir yöntem geliştirmek amacıyla Semi-Naive Bayes temeline dayanarak geliştirilen bu sınıflayıcının birçok senaryo için her iki yöntemden de daha yüksek performans göstermesi, temelinde yatan rasgelelik fikrinin özellikle YBDÖG yapılarda olumlu etkisi olduğunu düşündürmektedir. Bu yöntemin dengesiz dağılım durumunda performansında önemli değişikliklerin olmaması da dikkat çeken diğer bir özelliğidir.

6. SONUÇ VE ÖNERİLER

Çalışma sonuçları literatürde sıklıkla kullanılan ve özellikle mikro dizilim veriler ile sınıflama problemlerinde karşımıza çıkan makine öğrenmesi algoritmalarının sınıflama performanslarının farklı YBDÖG veri yapılarında farklı eğilimler gösterdiğini ortaya koymaktadır. Bununla birlikte doğru sınıflama başarılarına ilişkin ölçütler dikkatle incelendiğinde söz konusu algoritmaların yüksek performansa sahip olmadığı da görülmektedir. Aşağıda çalışma sonuçları genel olarak özetlenmekte ve bunlara ilişkin öneriler sunulmaktadır.

Klasik yapılarda oldukça başarılı ve üstün özellikleri olan DVM yöntemi YBDÖG veri yapılarında tahmin edildiği ölçüde başarı gösterememektedir. Özellikle mikro dizilim çalışmalarında bu yöntemin sıklıkla tercih ediliyor olması bu anlamda üzerinde dikkatle durulmasını gerektirmektedir.

Sınıflayıcılar açısından genel yapı gözlem sayıları ile artan performans eğilimi göstermekteyken RFerns bu yapıdan farklı davranış sergilemektedir. Özellikle dengeli dağılım ve düşük örneklem genişliğine sahip yapılarda bu yöntemin diğerlerinden daha yüksek performans göstermesi dikkat çekmektedir. Henüz sağlık alanında çok sayıda uygulaması bulunmayan bu yöntemin gözlem sayısının oldukça düşük olduğu mikro dizilim ve tıbbi görüntüleme çalışmalarına katkı sağlayacağı düşünülmektedir.

F ölçütü dikkate alınarak dengeli dağılım ve dengesiz dağılım durumunda sınıflama performanslarındaki değişim incelendiğinde prevalans değişikliğinden en fazla DVM'nin etkilendiği görülmektedir. RFerns ise bu durumdan en az etkilenen yöntem olarak öne çıkmaktadır. Bununla birlikte dengesiz dağılım durumunda sınıflama performanslarının büyük oranda 0,50'ler altında kalması aslında farklı yaklaşımlar ile bu problemlerin çözülmesi gerektiğini de düşündürmektedir.

UAA yönteminin sınıf dağılımlarının dengeli olduğu senaryolardaki yüksek başarısı, yöntemin gerçek uygulamalarda araştırmacılara daha doğru bilgiye ulaşma konusunda yardımcı olacağı kanaatini oluşturmaktadır. Bununla birlikte bu yöntemin dengesiz dağılımlar için geliştirilmiş hali olan AUAA'nın da bilgisayar ortamında geliştirilerek kolay uygulanabilir hale getirilmesi ve dengesiz dağılımlarda sağladığı sınıflama performansının incelenmesi önem arz etmektedir.

Burada çok sayıda senaryo üzerinde çalışılmış olsa da daha genel kanılara ulaşabilmek için ileri çalışmalarla bu kapsamın daha da genişletilmesi gerektiği düşünülmektedir. Özellikle yüksek kapasiteli bilgisayarlarla çalışarak boyut sayısının on binlerin üzerinde olduğu senaryoların geliştirilmesiyle mikro dizilim verilerine daha benzer yapıların ortaya konması faydalı olacaktır. Boyutun yanı sıra YBDÖG veri yapısı kapsamında daha farklı gözlem sayıları ve farklı gürültü düzeylerinde çalışmaların tekrarlanması da önemlidir. Bu çalışmada yalnızca sürekli veri tipinde bağımsız değişkenler ve iki sınıflı sınıflama problemleri ile çalışıldığı da göz önünde bulundurulmalıdır. Kategorik bağımsız değişkenler içeren ve çok sınıflı veri yapıları için performans karşılaştırmasının literatüre katkı sağlayacağı düşünülmektedir.

Yalnızca beş sınıflama algoritmasının dikkate alındığı bu çalışmanın diğer makine öğrenmesi yöntemlerini kapsayacak şekilde genişletilmesi de YBDÖG veri setleri ile sınıflama problemleri konusunda literatüre katkı sağlayacaktır. Diğer bir önemli nokta ise bu çalışma kapsamında R yazılımında standart (default) olan model parametrelerinin kullanılmış olmasıdır. Model parametrelerini de dikkate alarak yapılacak olan ileri çalışmalar ile bu konuda daha farklı sonuçların elde edilebileceği düşünülmektedir.

KAYNAKLAR

1. Donoho, D.L. (2000). *High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality*. Stanford University, Eriřim: 10 Haziran 2015, <http://statweb.stanford.edu/~donoho/Lectures/AMS2000/Curses.pdf>
2. Tan, P.N., Steinbach, M. ve Kumar, V. (2006). *Introduction to Data Mining*. NewYork: Pearson Education.
3. Han, J. ve Kamber, M. (2006). *Data Mining: Concept and Techniques*. Second Edition, San Francisco: Elsevier Inc.
4. Ye, J. ve Wang, T. (2006). Regularized Discriminant Analysis for High Dimensional, Low Sample Size Data. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (s. 454-463). Philadelphia. Eriřim: 04 Ağustos 2014, <http://dl.acm.org/citation.cfm?id=1150453&dl=ACM&coll=DL&CFID=542775688&CFTOKEN=78954282>
5. Muller, K.,E., Chi, Y., Ahn, J. ve Marron, J.S. (2008). Limitations of High Dimension, Low Sample Size Principal Components for Gaussian Data. Eriřim: 11 Mayıs 2014, http://midag.cs.unc.edu/pubs/papers/Muller_HDLSS_PCA_Rev208.pdf
6. Sen, P.K., Tsai, M-T. ve Jou, Y-S. (2007). High Dimension, Low-Sample Size Perspectives in Constrained Statistical Inference: The SARSCoV RNA Genome Illustration. *Journal of the American Statistical Association*, 102(478), 686-694.
7. GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012, Eriřim: 7 Haziran 2015, <http://globocan.iarc.fr/Default.aspx>
8. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P. ve diğ erleri. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439), 531-537.
9. Ramey, J.A. ve Young, P.D. (2011). A Comparison of Regularization methods Applied to the Linear Discriminant Function With High-Dimensional Microarray Data. *Journal of Statistical Computation and Simulation*, 83(3), 581-596.
10. Xing E.P., Jordan, M.I. ve Karp, R.M. (2001). Feature Selection for High-Dimensional Genomic Microarray Data. Proceedings of the Eighteenth International Conference on Machine Learning. (s. 601-608) Eriřim: 20 Eylül 2014, <http://www.public.asu.edu/~huanliu/papers/icml03.pdf>
11. Qiao. Z., Zhou, L. ve Huang, J.Z. (2008). Effective Linear Discriminant Analysis for High Dimensional, Low Sample Size Data. Proceedings of the World Congress of Engineering. London. (s.1070-1075). Eriřim: 10 Mayıs 2014, <https://www.stat.tamu.edu/~jianhua/paper/iccsde-sparseLDA.pdf>

12. Hingdon, R., Foster, N.L., Koeppe, R.A., DeCarli, C.S., Jagust, W.J., Clark, C.M. ve diğerleri. (2004) A Comparison of Classification Methods for Differentiating Fronto-Temporal Dementia from Alzheimer's Disease Using FDG-PET Imaging. *Statistics in Medicine*, 23, 315-326.
13. Hall, P., Marron, J.S. ve Neeman, A. (2005). Geometric Representation of High Dimension, Low Sample Size Data. *Journal of the Royal Statistical Society, Series B*. 67, 476-498.
14. Qiao, X., Zhang, H.H., Liu, Y., Todd, M.J. ve Marron, J.S. (2010). Asymptotic Properties of Distance-Weighted Discrimination. *Journal of the American Statistical Association* 105, 401-414.
15. Dudoit, S., Fridlyand, J., Speed, T.P. (2002). Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*, 97(457), 77-87.
16. Cho, S.B. ve Won, H.H. (2003). Machine Learning in DNA Microarray Analysis for Cancer Classification. Proceedings of the First Asia-Pacific Bioinformatics Conference. Australia. (s.189-198). Erişim: 17 Mart 2014, <http://dl.acm.org/citation.cfm?id=820213&dl=ACM&coll=DL&CFID=543586538&CFTOKEN=46532454>
17. Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M. ve Haussler, D. (2000). Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data. *Bioinformatics*, 16(10), 906-914.
18. Pirooznia, M., Yang, J.Y., Yang, M.Q. ve Deng, Y. (2008). A Comparative Study of Different Machine Learning Methods on Microarray Gene Expression Data. *BMC Genomics*, 9, 13.
19. Wang, Y., Tetko, I.V., Hall, M.A., Frank, E., Facius, A., Mayer, K.F.X. ve diğerleri. (2005). Gene Selection From Microarray Data for Cancer Classification- A Machine Learning Approach. *Computational Biology and Chemistry*, 29, 37-46.
20. Ahn, J. (2006). *High Dimension Low Sample Size Data Analysis*. Doktora Tezi, North Carolina Üniversitesi, Chapel Hill.
21. Bickel, P.J., ve Levina, E. (2004). Some Theory for Fisher's Linear Discriminant Function, 'Naive Bayes', and Some Alternatives When There Are Many More Variables Than Observations. *Bernoulli*, 10(6), 989-1010.
22. Tamatani, M. (2015). Asymptotic Theory for Discriminant Analysis in High Dimension Low Sample Size. *Memoirs of the Graduate School of Science and Engineering. Shiman University. Series B: Mathematics*, 48, 15-26.
23. Marron, J.S., Todd, M.J. ve Ahn, J. (2007). Distance-Weighted Discrimination. *Journal of the American Statistical Association*, 102(480), 1267-1271.

24. Liu, Y., Zhang, H.H. ve Wu, Y. (2011). Hard of Soft classification? Large-margin Unified Machines. *Journal of American Statistical Association*, 106(493), 166-177.
25. Abe, S. (2005). *Support Vector Machine for Pattern Classification*, Springer.
26. Wang, L. (2005). *Support Vector Machines: Theory and Application*, London: Studies in Fuzziness and Soft Computing, Springer.
27. Schölkopf, B., Burges, C.J.C. ve Smola, A.J. (1998). *Advances in Kernel Methods: Support Vector Learning*. Cambridge: The MIT Press.
28. Aydoğan, Ü. (2011). *Destek Vektör Makinalarında Kullanılan Çekirdek Fonksiyoların Sınıflama Performanslarının Karşılaştırılması*. Yüksek lisans tezi, Hacettepe Üniversitesi, Ankara.
29. Huang, H., Liu, Y., Du, Y., Perou, C.M., Hayes, D.N., Todd, M.J. ve diğerleri. (2013). Multiclass Distance-Weighted Discrimination. *Journal of Computational and Graphical Statistics*, 22(4), 953-969.
30. Huang, H., Lu, X., Liu, Y., Marron, J.S. ve Haaland, P. (2011). Distance Weighted Discrimination and Second Order Cone Programming.
31. Huang, H., Lu, X., Liu, Y., Haaland, P. ve Marron, J.S. (2012). R/DWD: Distance Weighted Discrimination for Classification, Visualization and Batch Adjustment. *Bioinformatics*, 28, 1182-1183.
32. Benito, M., Parker, J., Du, Q., Wu, J., Xiang, D., Perou, C.M. ve diğerleri. (2004). Adjustment of Systematic Microarray Data Biases. *Bioinformatics*, 20, 105-114.
33. Qiao, X., Zhang, H.H., Liu, Y., Todd, M.J. ve Marron, J.S. (2010). Weighted Distance Weighted Discrimination and Its Asymptotic Properties. *Journal of the American Statistical Association*, 105(489), 401-414.
34. Qiao, X. ve Zhang, H.H. (2014). Distance-weighted Support Vector Machine. *Statistics and Its Interface*. 8(3), 331-345.
35. Qiao, X. (2010). *Weighted Distance Weighted Discrimination and Pairwise Variable Selection for Classification*. Doktora tezi. North Carolina Üniversitesi. Chapel Hill.
36. Rish, I. (2001). An Empirical Study of the Naive Bayes. IBM Technical Report (Rapor No: 22230), Watson Research Center. Erişim: 22 Mayıs 2015, <http://www.research.ibm.com/people/r/rish/papers/RC22230.pdf>
37. Korkem, E. (2013). *Mikroarray Gen Ekspresyon Veri Setlerinde Random Forest ve Naive Bayes Sınıflama Yöntemleri Yaklaşımı*. Yüksek lisans tezi. Hacettepe Üniversitesi. Ankara.
38. Breimen, L. (2001). Random Forest. *Machine Learning*. 45, 5-32.
39. Boulesteix, A.L., Janitza, S., Kruppa, J., König, I.R. (2012). Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2(6), 493-507.

40. Hapfelmeier, A. (2012). Analysis of Missing Data with Random Forests. Doktora Tezi. University of Munich. Munich.
41. Yılmaz, H. (2014). *Random Forest Yönteminde Kayıp Veri Probleminin İncelenmesi ve Sağlık Alanında Bir Uygulama*. Yüksek lisans tezi, Eskişehir Osmangazi Üniversitesi, Eskişehir.
42. Özuysal, M., Calonder, M., Lepetit, V. Ve Fua, P. (2010). Fast Keypoint Recognition Using Random Ferns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 32(3), 448- 461.
43. Bosch, A., Zisserman, A. ve Munoz, X. (Ekim 2007) Image Classification Using Random Forests and Ferns. [Bildiri], IEEE 11th International Conference on Computer Vision, Brezilya.
44. Dong, Y., Zhang, Y., Yue, J. ve Hu, Z. (2015). *Comparison of Random Forest, Random Ferns and Support Vector Machine for Eye State Classification*. Multimedia Tools and Applications, Springer. (s.1-21). Erişim: 14 Mayıs 2015,
<http://link.springer.com/article/10.1007/s11042-015-2635-0>
45. Kursu, M.B. (2014). rFerns- Random Ferns Method Implementation for the General Purpose Machine Learning. *Journal of Statistical Software*, 61(10), 1-13.
46. Demirtaş, H. ve Doğanay, B. (2012). Simultaneous Generation of Binary and Normal Data with Specified Marginal and Association Structures. *Journal of Biopharmaceutical Statistics*. 22(2), 223-236.
47. Demirtaş, H., Amatya, A. ve Doğanay, B. (2014). BinNor: An R Package for Concurrent Generation of Binary and Normal Data. *Communication in Statistics- Simulation and Computation*, 43(3), 569-579.
48. Demirtaş, H. ve Amatya, A. (2012). Package ‘BinNor’, Simultaneous Generation of Multivariate Binary and Normal Variables. Erişim: 5 Ocak 2014,
<http://cran.r-project.org/web/packages/BinNor/index.html>
49. Straube, S. Ve Krell M.M. (2014). How to Evaluate an Agent’s Behavior to Infrequent Events? Reliable Performance Estimation Insensitive to Class Distribution. *Frontiers in Computational Neuroscience*, 8(43). Erişim: 30 Eylül 2015,
<http://journal.frontiersin.org/article/10.3389/fncom.2014.00043/abstract>
50. Pamukçu, E. (2015). *Yüksek Boyutlu Kanser Sınıflama Probleminde Bilgi Karmaşıklığı Kriteri ile Aykırı Gözlem Tespiti ve Boyut İndirgeme*. Doktora tezi, Fırat Üniversitesi, Elazığ.

EKLER

EK 1: Benzetim Çalışması Kod Örneği

#Senaryo Girdileri Tanımlama

```
n=20
nbin=1
nnor=99
d=nbin+nnor
prop.bin=rep(0.5,nbin)
mean=rep(0,d)
d_c1=ceiling((d-1)*0.10)
c1=runif(d_c1,0.10,0.30)
c2=runif(d-1-d_c1,0,0.10)
c3=runif(((d*(d-1)/2)-(d-1)),0.10,0.30)
corr.vec1=append(c1,c2)
corr.vec=append(corr.vec1,c3)
cmat1=lower.tri.to.corr.mat(corr.vec,d)
cmat2=make.positive.definite(cmat1)
cmat3=(cmat2+t(cmat2))/2
cmat4=make.positive.definite(cmat3)
cmat=cov2cor(cmat4)
sigma.star=compute.sigma.star(no.bin=nbin,no.nor=nnor,prop.vec.bin=prop.bin,corr.
mat=cmat)
fonksiyon=function(n,d,sigma.star,num.sim)
```

#Veri Türetme

```
mydata=rmvnorm(n,mean,sigma.star$sigma_star)
```

#Eğitim ve Test Setleri Oluşturma

```
myBreaks= quantile(mydata[,1],c(0,.5,1))
y=cut(mydata[,1],breaks=myBreaks,include.lowest=TRUE,labels=c("0","1"))
mydata=data.frame(y,mydata[,-1])
smp_size <- floor(0.75 * nrow(mydata))
train_ind <- sample(seq_len(nrow(mydata)), size = smp_size)
mydata_train <- mydata[train_ind, ]
mydata_test <- mydata[-train_ind, ]
y_test=factor(mydata_test[,1],levels=c(0,1),labels=c("0","1"))
y_train=factor(mydata_train[,1],levels=c(0,1),labels=c("0","1"))
mydata_train=data.frame(y_train,mydata_train[,-1])
mydata_test=data.frame(y_test,mydata_test[,-1])
```

#Modelleme (Öğrenme Süreci)

```
model.svm= svm(y_train~.,data=mydata_train,na.action=na.omit,type='C')
model.dwd=kdwd(y_train~.,data=mydata_train,na.action=na.omit)
model.RForest=randomForest(y_train~.,data=mydata_train)
model.RFerns=rFerns(y_train~.,data=mydata_train)
model.NBayes=naiveBayes(y_train~.,data=mydata_train)
pred1.1=factor(predict(model.svm,mydata_test),levels=c(0,1),labels=c("0","1"))
```

```

tablo.svm=table(pred1.1,y_test)
pred2.1=factor(predict(model.dwd,mydata_test),levels=c(0,1),labels=c("0","1"))
tablo.dwd=table(pred2.1,y_test)
pred3.1=factor(predict(model.RForest,mydata_test),levels=c(0,1),labels=c("0","1"))
tablo.RForest=table(pred3.1,y_test)
pred4.1=factor(predict(model.RFerns,mydata_test),levels=c(0,1),labels=c("0","1"))
tablo.RFerns=table(pred4.1,y_test)
pred5.1=factor(predict(model.NBayes,mydata_test),levels=c(0,1),labels=c("0","1"))
tablo.NBayes=table(pred5.1,y_test)

```

#Grafik Oluşturma

```

graf=read.csv2(file.choose(), header=TRUE)
a=(qplot(n,DVM,data=graf,color=d,shape=d,facets=B1|Bs1z~Bs1zBs1z,xlab="DVM",
ylab="Doğruluk",ylim=c(0.30,0.90)))+geom_line()
b=(qplot(n,UAA,data=graf,color=d,shape=d,facets=B1|Bs1z~Bs1zBs1z,xlab="UAA",
ylab="Doğruluk",ylim=c(0.30,0.90)))+geom_line()
c=(qplot(n,Rforest,data=graf,color=d,shape=d,facets=B1|Bs1z~Bs1zBs1z,xlab="Rfore
st",ylab="Doğruluk",ylim=c(0.30,0.90)))+geom_line()
d=(qplot(n,Rferns,data=graf,color=d,shape=d,facets=B1|Bs1z~Bs1zBs1z,xlab="Rferns
",ylab="Doğruluk",ylim=c(0.30,0.90)))+geom_line()
e=(qplot(n,Nbayes,data=graf,color=d,shape=d,facets=B1|Bs1z~Bs1zBs1z,xlab="Nbay
es",ylab="Doğruluk",ylim=c(0.30,0.90)))+geom_line()

```

EK 2: Senaryolara İlişkin Duyarlılık, Seçicilik ve Matthews Korelasyon Katsayısı Değerleri

No	Dağılım	G.O.	Blı Bsız	Bsız Bsız	N	d	DVM			UAA			RForest			RFerns			NB		
							Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk
4	Dengeli	Y	D	D	50	100	0,562	0,554	-	0,598	0,585	-	0,564	0,569	-	0,558	0,542	-	0,574	0,567	-
5	Dengeli	Y	D	D	50	500	0,523	0,526	-	0,562	0,574	-	0,534	0,544	-	0,525	0,531	-	0,544	0,558	-
6	Dengeli	Y	D	D	50	1000	0,518	0,518	-	0,542	0,557	-	0,525	0,535	-	0,509	0,522	-	0,536	0,532	-
7	Dengeli	Y	D	D	100	100	0,595	0,605	-	0,618	0,625	0,247	0,597	0,598	0,198	0,589	0,581	0,178	0,591	0,592	0,184
8	Dengeli	Y	D	D	100	500	0,557	0,563	-	0,586	0,591	0,182	0,558	0,557	0,119	0,564	0,553	-	0,568	0,560	0,129
9	Dengeli	Y	D	D	100	1000	0,547	0,527	-	0,562	0,554	-	0,542	0,534	0,079	0,536	0,544	-	0,548	0,550	0,100
13	Dengeli	Y	D	O	50	100	0,547	0,548	-	0,595	0,582	-	0,559	0,545	-	0,543	0,539	-	0,549	0,534	0,085
14	Dengeli	Y	D	O	50	500	0,519	0,517	-	0,545	0,544	-	0,529	0,527	-	0,513	0,526	-	0,529	0,520	-
15	Dengeli	Y	D	O	50	1000	0,497	0,524	-	0,532	0,533	-	0,513	0,516	-	0,515	0,507	-	0,517	0,517	-
16	Dengeli	Y	D	O	100	100	0,572	0,585	-	0,623	0,631	0,258	0,574	0,580	0,157	0,561	0,552	-	0,546	0,552	0,099
17	Dengeli	Y	D	O	100	500	0,534	0,539	-	0,574	0,584	0,167	0,542	0,549	0,095	0,535	0,528	-	0,532	0,532	0,064
18	Dengeli	Y	D	O	100	1000	0,524	0,526	-	0,551	0,562	-	0,523	0,533	0,059	0,517	0,534	-	0,521	0,536	0,058
22	Dengeli	Y	D	Y	50	100	0,524	0,529	-	0,600	0,600	-	0,534	0,534	-	0,518	0,533	-	0,515	0,519	-
23	Dengeli	Y	D	Y	50	500	0,519	0,498	-	0,556	0,546	-	0,532	0,519	-	0,515	0,505	-	0,515	0,501	-
24	Dengeli	Y	D	Y	50	1000	0,499	0,516	-	0,535	0,538	-	0,509	0,515	-	0,510	0,507	-	0,507	0,509	0,017
25	Dengeli	Y	D	Y	100	100	0,560	0,568	-	0,641	0,643	0,289	0,562	0,575	0,140	0,543	0,542	0,088	0,529	0,531	0,063
26	Dengeli	Y	D	Y	100	500	0,524	0,534	-	0,574	0,576	-	0,534	0,539	0,075	0,525	0,524	-	0,517	0,521	0,039
27	Dengeli	Y	D	Y	100	1000	0,519	0,519	-	0,543	0,555	-	0,520	0,530	0,051	0,507	0,528	-	0,508	0,527	0,036
31	Dengeli	Y	O	D	50	100	0,612	0,621	-	0,677	0,685	-	0,631	0,639	-	0,586	0,600	-	0,618	0,627	-
32	Dengeli	Y	O	D	50	500	0,537	0,558	-	0,608	0,615	-	0,567	0,574	-	0,556	0,542	-	0,568	0,574	-
33	Dengeli	Y	O	D	50	1000	0,553	0,505	-	0,592	0,562	-	0,556	0,541	-	0,528	0,539	-	0,569	0,542	-
34	Dengeli	Y	O	D	100	100	0,702	0,697	0,405	0,738	0,734	0,474	0,694	0,686	0,384	0,653	0,656	0,324	0,659	0,660	0,321
35	Dengeli	Y	O	D	100	500	0,588	0,608	-	0,638	0,638	0,283	0,596	0,599	0,200	0,585	0,573	-	0,589	0,594	0,184
36	Dengeli	Y	O	D	100	1000	0,550	0,570	-	0,593	0,600	0,201	0,555	0,564	0,124	0,547	0,551	-	0,564	0,573	0,139

No	Dağılım	G.O.	Bh Bsız	Bsız Bsız	n	d	DVM			UAA			RForest			RFerns			NB		
							Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk
40	Dengeli	Y	O	O	50	100	0,588	0,590	-	0,687	0,681	-	0,605	0,607	-	0,569	0,575	-	0,571	0,571	-
41	Dengeli	Y	O	O	50	500	0,525	0,533	-	0,595	0,602	-	0,551	0,558	-	0,547	0,521	-	0,538	0,536	-
42	Dengeli	Y	O	O	50	1000	0,505	0,533	-	0,559	0,563	-	0,519	0,533	-	0,521	0,518	-	0,524	0,539	-
43	Dengeli	Y	O	O	100	100	0,676	0,674	0,359	0,749	0,751	0,502	0,665	0,661	0,330	0,619	0,611	-	0,599	0,593	0,194
44	Dengeli	Y	O	O	100	500	0,579	0,561	-	0,637	0,636	0,281	0,584	0,570	-	0,557	0,553	-	0,557	0,553	0,112
45	Dengeli	Y	O	O	100	1000	0,533	0,542	-	0,585	0,584	-	0,550	0,551	0,106	0,541	0,533	-	0,540	0,536	0,077
49	Dengeli	Y	O	Y	50	100	0,572	0,569	-	0,705	0,702	-	0,598	0,594	-	0,559	0,562	-	0,546	0,544	-
50	Dengeli	Y	O	Y	50	500	0,521	0,532	-	0,604	0,611	-	0,539	0,550	-	0,515	0,543	-	0,515	0,529	-
51	Dengeli	Y	O	Y	50	1000	0,515	0,513	-	0,563	0,570	-	0,520	0,528	-	0,509	0,530	-	0,510	0,528	0,038
52	Dengeli	Y	O	Y	100	100	0,656	0,650	0,315	0,766	0,761	0,529	0,654	0,641	0,298	0,587	0,584	0,175	0,558	0,545	0,105
53	Dengeli	Y	O	Y	100	500	0,543	0,551	-	0,636	0,638	-	0,558	0,564	0,126	0,535	0,533	-	0,523	0,525	0,048
54	Dengeli	Y	O	Y	100	1000	0,514	0,532	-	0,586	0,578	-	0,537	0,536	0,076	0,519	0,524	-	0,513	0,518	0,033
58	Dengeli	Y	Y	D	50	100	0,636	0,667	-	0,709	0,723	-	0,660	0,679	-	0,629	0,592	-	0,642	0,657	0,303
59	Dengeli	Y	Y	D	50	500	0,567	0,554	-	0,646	0,633	-	0,596	0,590	-	0,556	0,553	-	0,598	0,585	-
60	Dengeli	Y	Y	D	50	1000	0,528	0,530	-	0,602	0,590	-	0,549	0,551	-	0,541	0,532	-	0,563	0,559	-
61	Dengeli	Y	Y	D	100	100	0,731	0,736	-	0,769	0,769	0,537	0,722	0,726	0,450	0,680	0,668	-	0,679	0,681	0,361
62	Dengeli	Y	Y	D	100	500	0,623	0,617	-	0,671	0,670	0,346	0,625	0,618	0,250	0,590	0,587	-	0,609	0,614	0,223
63	Dengeli	Y	Y	D	100	1000	0,575	0,580	-	0,627	0,628	-	0,581	0,580	0,166	0,568	0,556	-	0,582	0,589	0,173
67	Dengeli	Y	Y	O	50	100	0,627	0,623	-	0,734	0,726	0,461	0,660	0,637	-	0,609	0,594	-	0,612	0,596	0,213
68	Dengeli	Y	Y	O	50	500	0,537	0,541	-	0,620	0,638	-	0,551	0,564	-	0,544	0,534	-	0,550	0,554	-
69	Dengeli	Y	Y	O	50	1000	0,526	0,511	-	0,594	0,594	-	0,548	0,547	-	0,520	0,529	-	0,538	0,537	-
70	Dengeli	Y	Y	O	100	100	0,716	0,710	0,433	0,788	0,785	0,572	0,700	0,698	0,401	0,640	0,630	-	0,619	0,615	0,236
71	Dengeli	Y	Y	O	100	500	0,589	0,596	-	0,675	0,679	0,363	0,598	0,609	0,213	0,567	0,562	-	0,564	0,563	0,128
72	Dengeli	Y	Y	O	100	1000	0,552	0,558	-	0,620	0,616	-	0,568	0,567	0,140	0,542	0,553	-	0,547	0,553	0,102
76	Dengeli	Y	Y	Y	50	100	0,614	0,600	-	0,750	0,748	-	0,634	0,633	-	0,579	0,579	-	0,559	0,557	-

No	Dağılım	G.O.	Blı Bsız	Bsız Bsız	n	d	DVM			UAA			RForest			RFerns			NB		
							Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk
77	Dengeli	Y	Y	Y	50	500	0,534	0,524	-	0,630	0,631	-	0,551	0,557	-	0,538	0,524	-	0,540	0,515	0,058
78	Dengeli	Y	Y	Y	50	1000	0,538	0,503	-	0,596	0,588	-	0,541	0,539	-	0,517	0,520	-	0,530	0,517	-
79	Dengeli	Y	Y	Y	100	100	0,697	0,713	0,417	0,813	0,812	0,625	0,696	0,703	0,401	0,616	0,613	0,236	0,566	0,582	0,152
80	Dengeli	Y	Y	Y	100	500	0,579	0,577	-	0,686	0,689	0,384	0,598	0,600	0,204	0,551	0,554	-	0,542	0,547	0,091
81	Dengeli	Y	Y	Y	100	1000	0,549	0,529	-	0,621	0,609	-	0,562	0,551	0,117	0,531	0,532	-	0,532	0,529	0,063
85	Dengeli	O	D	D	50	100	0,626	0,635	-	0,671	0,670	0,344	0,618	0,631	-	0,617	0,615	-	0,637	0,635	0,276
86	Dengeli	O	D	D	50	500	0,573	0,545	-	0,609	0,610	0,223	0,583	0,576	-	0,551	0,575	-	0,598	0,589	-
87	Dengeli	O	D	D	50	1000	0,554	0,519	-	0,588	0,565	-	0,556	0,537	-	0,525	0,550	-	0,577	0,552	-
88	Dengeli	O	D	D	100	100	0,691	0,694	0,389	0,716	0,724	0,440	0,657	0,657	0,317	0,649	0,655	-	0,654	0,654	0,308
89	Dengeli	O	D	D	100	500	0,606	0,607	-	0,626	0,635	0,267	0,593	0,588	0,185	0,595	0,591	-	0,606	0,600	0,208
90	Dengeli	O	D	D	100	1000	0,573	0,576	-	0,602	0,599	-	0,574	0,565	0,143	0,570	0,575	-	0,586	0,591	0,178
94	Dengeli	O	D	O	50	100	0,599	0,589	-	0,676	0,661	-	0,589	0,593	-	0,570	0,597	-	0,578	0,580	-
95	Dengeli	O	D	O	50	500	0,553	0,523	-	0,588	0,576	-	0,557	0,534	-	0,524	0,553	-	0,558	0,540	0,099
96	Dengeli	O	D	O	50	1000	0,498	0,535	-	0,556	0,564	-	0,533	0,537	-	0,531	0,518	-	0,525	0,532	-
97	Dengeli	O	D	O	100	100	0,655	0,650	0,312	0,715	0,716	0,434	0,623	0,629	0,255	0,607	0,601	0,215	0,595	0,593	0,190
98	Dengeli	O	D	O	100	500	0,558	0,570	-	0,605	0,604	0,215	0,559	0,560	0,122	0,561	0,554	-	0,560	0,552	0,114
99	Dengeli	O	D	O	100	1000	0,556	0,542	-	0,598	0,578	-	0,553	0,535	0,091	0,550	0,546	-	0,554	0,551	0,107
103	Dengeli	O	D	Y	50	100	0,557	0,564	-	0,675	0,675	-	0,572	0,574	-	0,555	0,555	-	0,541	0,554	-
104	Dengeli	O	D	Y	50	500	0,531	0,520	-	0,589	0,564	-	0,537	0,534	-	0,534	0,529	-	0,534	0,522	-
105	Dengeli	O	D	Y	50	1000	0,525	0,511	-	0,542	0,553	-	0,511	0,517	-	0,512	0,527	-	0,519	0,515	-
106	Dengeli	O	D	Y	100	100	0,627	0,630	-	0,730	0,731	-	0,607	0,611	0,220	0,579	0,581	0,164	0,562	0,555	0,121
107	Dengeli	O	D	Y	100	500	0,557	0,538	-	0,615	0,597	-	0,553	0,542	0,097	0,539	0,541	0,082	0,540	0,534	0,076
108	Dengeli	O	D	Y	100	1000	0,518	0,542	-	0,566	0,578	-	0,529	0,540	0,509	0,529	0,530	0,062	0,521	0,540	0,061
112	Dengeli	O	O	D	50	100	0,735	0,739	-	0,776	0,768	0,546	0,717	0,717	-	0,714	0,685	-	0,733	0,711	-

No	Dağılım	G.O.	Blı Bsız	Bsız Bsız	n	d	DVM			UAA			RForest			RFerns			NB		
							Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk
113	Dengeli	O	O	D	50	500	0,626	0,625	-	0,670	0,677	0,350	0,638	0,641	-	0,628	0,609	-	0,653	0,648	-
114	Dengeli	O	O	D	50	1000	0,558	0,601	-	0,624	0,652	-	0,598	0,610	-	0,609	0,562	-	0,613	0,630	-
115	Dengeli	O	O	D	100	100	0,783	0,785	0,570	0,806	0,810	0,616	0,741	0,739	0,483	0,733	0,726	0,469	0,730	0,726	0,459
116	Dengeli	O	O	D	100	500	0,686	0,670	-	0,711	0,703	0,416	0,670	0,651	0,324	0,657	0,651	-	0,665	0,652	0,320
117	Dengeli	O	O	D	100	1000	0,635	0,638	-	0,661	0,664	-	0,628	0,625	0,257	0,622	0,622	-	0,630	0,632	0,264
121	Dengeli	O	O	O	50	100	0,681	0,664	-	0,756	0,748	0,504	0,668	0,654	-	0,643	0,648	-	0,641	0,641	0,284
122	Dengeli	O	O	O	50	500	0,568	0,572	-	0,657	0,647	-	0,590	0,581	-	0,574	0,576	-	0,587	0,587	0,177
123	Dengeli	O	O	O	50	1000	0,549	0,564	-	0,615	0,625	-	0,565	0,592	-	0,560	0,567	-	0,566	0,588	-
124	Dengeli	O	O	O	100	100	0,742	0,752	0,497	0,801	0,802	0,602	0,697	0,696	0,395	0,672	0,677	0,356	0,650	0,662	0,315
125	Dengeli	O	O	O	100	500	0,622	0,622	-	0,679	0,690	0,373	0,615	0,608	0,226	0,612	0,596	0,213	0,604	0,587	0,193
126	Dengeli	O	O	O	100	1000	0,591	0,580	-	0,639	0,634	0,281	0,589	0,587	0,179	0,584	0,575	-	0,582	0,585	0,168
130	Dengeli	O	O	Y	50	100	0,637		-	0,769	-	-	0,640	-	-	0,623	-	-	0,599	-	-
131	Dengeli	O	O	Y	50	500	0,558	0,558	-	0,650	0,645	-	0,569	0,573	-	0,556	0,567	-	0,554	0,559	-
132	Dengeli	O	O	Y	50	1000	0,539	0,534	-	0,610	0,607	-	0,556	0,550	-	0,548	0,537	-	0,549	0,540	-
133	Dengeli	O	O	Y	100	100	0,732	0,734	0,471	0,818	0,819	0,637	0,687	0,688	0,378	0,642	0,643	0,292	0,607	0,610	0,221
134	Dengeli	O	O	Y	100	500	0,601	0,588	-	0,692	0,678	0,376	0,599	0,589	0,192	0,572	0,577	0,154	0,570	0,564	0,135
135	Dengeli	O	O	Y	100	1000	0,558	0,575	-	0,629	0,634	0,274	0,566	0,567	0,136	0,563	0,563	0,130	0,560	0,560	0,121
139	Dengesiz	Y	D	D	50	100	-	0,982	0,038				-	0,896	-	-	0,038	-	-	0,755	-
140	Dengesiz	Y	D	D	50	500	-	0,994	0,012				-	0,917	-	-	0,012	-	-	0,803	-
141	Dengesiz	Y	D	D	50	1000	-	0,996	-				-	0,935	-	-	0,007	-	-	0,849	-
142	Dengesiz	Y	D	D	100	100	0,049	0,980	0,068				0,167	0,915	-	0,959	0,089	-	0,441	0,706	-
143	Dengesiz	Y	D	D	100	500	0,010	0,993	0,013				0,099	0,938	-	0,981	0,032	-	0,371	0,733	-
144	Dengesiz	Y	D	D	100	1000	0,005	0,996	0,007				0,075	0,952	-	0,989	0,018	-	0,323	0,759	-
148	Dengesiz	Y	D	O	50	100	-	0,964	0,052				-	0,856	-	-	0,117	-	-	0,647	-

No	Dağılım	G.O.	Bh Bsız	Bsız Bsız	n	d	DVM			UAA			RForest			RFerns			NB		
							Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk
149	Dengesiz	Y	D	O	50	500	-	0,979	0,020				-	0,872	-	-	0,055	-	-	0,676	-
150	Dengesiz	Y	D	O	50	1000	-	0,984	0,019				-	0,881	-	-	0,047	-	-	0,705	-
151	Dengesiz	Y	D	O	100	100	0,039	0,977	0,054				0,150	0,895	-	0,855	0,190	-	0,455	0,615	-
152	Dengesiz	Y	D	O	100	500	0,016	0,990	0,021				0,121	0,916	-	0,906	0,111	-	0,449	0,616	-
153	Dengesiz	Y	D	O	100	1000	0,011	0,992	0,014				0,104	0,921	-	0,940	0,074	-	0,407	0,652	-
157	Dengesiz	Y	D	Y	50	100	-	0,954	0,055				-	0,828	-	-	0,225	-	-	0,578	-
158	Dengesiz	Y	D	Y	50	500	0,041	0,965	0,044				0,180	0,848	-	0,870	0,140	-	0,413	0,608	-
159	Dengesiz	Y	D	Y	50	1000	0,025	0,977	0,024				0,158	0,869	-	0,923	0,094	-	0,400	0,634	-
160	Dengesiz	Y	D	Y	100	100	0,042	0,975	0,056				0,182	0,888	-	0,763	0,284	-	0,478	0,570	0,048
161	Dengesiz	Y	D	Y	100	500	0,019	0,986	0,024				0,131	0,898	-	0,842	0,176	-	0,472	0,567	0,038
162	Dengesiz	Y	D	Y	100	1000	0,015	0,987	0,019				0,108	0,903	-	0,870	0,145	-	0,453	0,584	-
166	Dengesiz	Y	O	D	50	100	-	0,981	0,070				-	0,908	-	-	0,040	-	-	0,777	-
167	Dengesiz	Y	O	D	50	500	-	0,991	-				-	0,916	-	-	0,015	-	-	0,808	-
168	Dengesiz	Y	O	D	50	1000	0,005	0,997	0,006				-	0,932	-	0,994	0,006	-	0,210	0,852	-
169	Dengesiz	Y	O	D	100	100	0,132	0,977	0,186				0,264	0,935	-	0,965	0,122	-	0,528	0,759	-
170	Dengesiz	Y	O	D	100	500	0,014	0,994	0,021				0,124	0,949	-	0,982	0,039	-	0,400	0,745	-
171	Dengesiz	Y	O	D	100	1000	0,007	0,996	0,010				0,091	0,948	-	0,988	0,021	-	0,355	0,765	-
175	Dengesiz	Y	O	O	50	100	-	0,958	0,098				-	0,872	-	-	0,143	-	-	0,666	-
176	Dengesiz	Y	O	O	50	500	-	0,979	-				-	0,878	-	-	0,065	-	-	0,685	-
177	Dengesiz	Y	O	O	50	1000	-	0,988	-				-	0,894	-	-	0,040	-	-	0,723	-
178	Dengesiz	Y	O	O	100	100	0,133	0,972	0,186				0,259	0,916	-	0,864	0,234	-	0,516	0,654	0,170
179	Dengesiz	Y	O	O	100	500	0,020	0,988	0,025				0,139	0,916	-	0,914	0,117	-	0,465	0,637	0,101
180	Dengesiz	Y	O	O	100	1000	0,015	0,992	0,020				0,108	0,923	-	0,937	0,083	-	0,411	0,657	-
184	Dengesiz	Y	O	Y	50	100	0,070	0,956	0,078				0,261	0,862	-	0,828	0,234	-	0,477	0,594	-
185	Dengesiz	Y	O	Y	50	500	0,045	0,963	0,046				0,188	0,850	-	0,881	0,139	-	0,437	0,600	-

No	Dağılım	G.O.	Bh Bsız	Bsız Bsız	n	d	DVM			UAA			RForest			RFerns			NB		
							Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk
186	Dengesiz	Y	O	Y	50	1000	0,027	0,977	0,027				0,158	0,864	-	0,918	0,093	-	0,398	0,626	-
187	Dengesiz	Y	O	Y	100	100	0,108	0,971	0,149				0,274	0,904	-	0,800	0,311	-	0,515	0,592	0,105
188	Dengesiz	Y	O	Y	100	500	0,024	0,984	0,032				0,146	0,905	-	0,853	0,191	-	0,493	0,576	0,069
189	Dengesiz	Y	O	Y	100	1000	0,014	0,989	0,019				0,116	0,911	-	0,882	0,147	-	0,468	0,585	-
193	Dengesiz	Y	Y	D	50	100	0,087	0,978	0,102				0,298	0,914	-	0,990	0,061	-	0,470	0,794	-
194	Dengesiz	Y	Y	D	50	500	-	0,971	0,123				-	0,915	-	-	0,072	-	-	0,800	-
195	Dengesiz	Y	Y	D	50	1000		0,997	-				-	0,936	-	-	0,007	-	-	0,855	-
196	Dengesiz	Y	Y	D	100	100	0,193	0,977	0,272				0,335	0,942	-	0,962	0,152	-	0,570	0,779	0,354
197	Dengesiz	Y	Y	D	100	500	0,195	0,957	0,248				0,322	0,917	-	0,952	0,210	-	0,615	0,738	0,346
198	Dengesiz	Y	Y	D	100	1000	0,009	0,996	0,013				0,101	0,956	-	0,987	0,024	-	0,371	0,779	-
202	Dengesiz	Y	Y	O	50	100	-	0,965	0,128				-	0,887	-	-	0,152	-	-	0,704	-
203	Dengesiz	Y	Y	O	50	500	0,027	0,980	0,028				0,171	0,886	-	0,950	0,068	-	0,382	0,696	-
204	Dengesiz	Y	Y	O	50	1000	0,021	0,990	0,023				0,143	0,895	-	0,958	0,044	-	0,344	0,727	-
205	Dengesiz	Y	Y	O	100	100	0,181	0,972	0,256				0,305	0,926	-	0,886	0,248	-	0,544	0,685	0,229
206	Dengesiz	Y	Y	O	100	500	0,028	0,991	0,039				0,155	0,925	-	0,917	0,129	-	0,469	0,650	-
207	Dengesiz	Y	Y	O	100	1000	0,012	0,991	0,017				0,114	0,924	-	0,933	0,088	-	0,423	0,665	-
211	Dengesiz	Y	Y	Y	50	100	-	0,949	0,131				-	0,860	-	-	0,247	-	-	0,598	-
212	Dengesiz	Y	Y	Y	50	500	0,043	0,970	0,046				0,180	0,865	-	0,882	0,137	-	0,434	0,613	-
213	Dengesiz	Y	Y	Y	50	1000	-	0,972	-				-	0,868	-	-	0,100	-	-	0,631	-
214	Dengesiz	Y	Y	Y	100	100	0,172	0,964	0,238				0,338	0,912	-	0,798	0,336	-	0,527	0,600	0,126
215	Dengesiz	Y	Y	Y	100	500	0,025	0,986	0,034				0,157	0,910	-	0,858	0,196	-	0,492	0,591	0,082
216	Dengesiz	Y	Y	Y	100	1000	0,019	0,987	0,025				0,132	0,906	-	0,881	0,148	-	0,461	0,594	0,056
220	Dengesiz	O	D	D	50	100	0,090	0,971	0,097				0,248	0,889	-	0,978	0,060	-	0,458	0,755	-
221	Dengesiz	O	D	D	50	500	-	0,989	0,027				-	0,913	-	-	0,027	-	-	0,802	-
222	Dengesiz	O	D	D	50	1000	-	0,992	0,015				-	0,926	-	-	0,012	-	-	0,833	-

No	Dağılım	G.O.	Bh Bsız	Bsız Bsız	n	d	DVM			UAA			RForest			RFerns			NB		
							Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk
223	Dengesiz	O	D	D	100	100	0,130	0,966	0,181				0,245	0,913	-	0,940	0,180	-	0,540	0,726	0,266
224	Dengesiz	O	D	D	100	500	0,043	0,985	0,059				0,168	0,925	-	0,971	0,080	-	0,501	0,721	-
225	Dengesiz	O	D	D	100	1000	0,020	0,992	0,026				0,125	0,939	-	0,980	0,047	-	0,431	0,736	-
229	Dengesiz	O	D	O	50	100	-	0,959	0,090				-	0,863	-	-	0,159	-	-	0,660	-
230	Dengesiz	O	D	O	50	500	0,042	0,972	0,045				0,188	0,868	-	0,940	0,086	-	0,406	0,664	-
231	Dengesiz	O	D	O	50	1000	0,025	0,981	0,027				0,152	0,882	-	0,959	0,057	-	0,358	0,704	-
232	Dengesiz	O	D	O	100	100	0,105	0,966	0,143				0,227	0,892	-	0,856	0,270	-	0,532	0,638	0,166
233	Dengesiz	O	D	O	100	500	0,042	0,979	0,053				0,167	0,899	-	0,892	0,171	-	0,497	0,632	-
234	Dengesiz	O	D	O	100	1000	0,030	0,985	0,038				0,147	0,908	-	0,914	0,127	-	0,460	0,634	-
238	Dengesiz	O	D	Y	50	100	-	0,944	0,110				-	0,834	-	-	0,260	-	-	0,590	-
239	Dengesiz	O	D	Y	50	500	-	0,964	0,048				-	0,843	-	-	0,150	-	-	0,603	-
240	Dengesiz	O	D	Y	50	1000	0,030	0,975	0,031				0,160	0,866	-	0,922	0,101	-	0,405	0,634	-
241	Dengesiz	O	D	Y	100	100	0,094	0,965	0,127				0,232	0,889	-	0,779	0,324	-	0,524	0,576	0,099
242	Dengesiz	O	D	Y	100	500	0,038	0,977	0,044				0,161	0,890	-	0,829	0,219	-	0,501	0,571	0,070
243	Dengesiz	O	D	Y	100	1000	0,027	0,983	0,035				0,144	0,899	-	0,863	0,174	-	0,466	0,593	0,059
247	Dengesiz	O	O	D	50	100	0,260	0,952	0,294				0,407	0,906	-	0,981	0,154	-	0,608	0,799	-
248	Dengesiz	O	O	D	50	500	-	0,975	0,093				-	0,911	-	-	0,052	-	-	0,802	-
249	Dengesiz	O	O	D	50	1000	-	0,984	0,049				-	0,923	-	-	0,027	-	-	0,818	-
250	Dengesiz	O	O	D	100	100	0,437	0,941	0,527				0,435	0,915	-	0,938	0,362	-	0,681	0,769	0,442
251	Dengesiz	O	O	D	100	500	0,147	0,968	0,193				0,277	0,919	-	0,961	0,171	-	0,601	0,727	0,323
252	Dengesiz	O	O	D	100	1000	0,062	0,983	0,084				0,192	0,934	-	0,971	0,095	-	0,514	0,734	-
256	Dengesiz	O	O	O	50	100	-	0,934	0,234				-	0,864	-	-	0,266	-	-	0,699	-
257	Dengesiz	O	O	O	50	500	-	0,962	0,082				-	0,868	-	-	0,127	-	-	0,684	-
258	Dengesiz	O	O	O	50	1000	0,050	0,970	0,047				0,194	0,879	-	0,954	0,085	-	0,414	0,710	-
259	Dengesiz	O	O	O	100	100	0,332	0,940	0,415				0,389	0,887	-	0,844	0,415	-	0,626	0,667	0,285

No	Dağılım	G.O.	Bh Bsız	Bsız Bsız	n	d	DVM			UAA			RForest			RFerns			NB		
							Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk	Duy.	Seç.	Mkk
260	Dengesiz	O	O	O	100	500	0,093	0,965	0,118				0,249	0,886	-	0,869	0,262	-	0,552	0,643	0,191
261	Dengesiz	O	O	O	100	1000	0,044	0,979	0,058				0,185	0,901	-	0,900	0,182	-	0,518	0,655	-
265	Dengesiz	O	O	Y	50	100	-	0,931	-				-	0,846	-	-	0,327	-	-	0,629	-
266	Dengesiz	O	O	Y	50	500	-	0,946	0,091				-	0,847	-	-	0,205	-	-	0,615	-
267	Dengesiz	O	O	Y	50	1000	0,050	0,968	0,049				0,211	0,868	-	0,904	0,131	-	0,448	0,648	-
268	Dengesiz	O	O	Y	100	100	0,274	0,948	0,353				0,358	0,891	-	0,798	0,437	-	0,602	0,616	0,212
269	Dengesiz	O	O	Y	100	500	0,069	0,968	0,089				0,219	0,889	-	0,822	0,289	-	0,547	0,600	0,142
270	Dengesiz	O	O	Y	100	1000	0,041	0,979	0,051				0,181	0,897	-	0,856	0,227	-	0,519	0,603	-