T.C.
NECMETTİN ERBAKAN NİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BOYUT İNDİRGEME TEKNİKLERİNİN SINIFLANDIRMA PERORMANSLARININ KARŞILAŞTIRILMASI

Tenzile ERBAYRAM

YÜKSEK LİSANS TEZİ

İstatistik Anabilim Dalı

Temmuz -2020
KONYA

TEZ KABUL VE ONAYI

Tenzile ERBAYRAM tarafından hazırlanan "Boyut İndirgeme Tekniklerinin Sınıflandırma Performanslarının Karşılaştırılması" adlı tez çalışması 17/07/2020 tarihinde aşağıdaki jüri tarafından oy birliği ile Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı'nda YÜKSEK LİSANS olarak kabul edilmiştir.

Jüri Üyeleri	**İmza**
Başkan Dr. Öğr. Üyesi Selim GÜNDÜZ	……..
Danışman Prof. Dr. Murat ERİŞOĞLU	……..
Üye Dr. Öğr. Üyesi Aydın KARAKOCA	……..

Fen Bilimleri Enstitüsü Yönetim Kurulu'nun …./…/20.. gün ve …….. sayılı kararıyla onaylanmıştır.

Prof. Dr. S. Savaş DURDURAN
FBE Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

İmza

Tenzile ERBAYRAM

17.07.2020

ÖZET

YÜKSEK LİSANS TEZİ

BOYUT İNDİRGEME TEKNİKLERİNİN SINIFLANDIRMA PERFORMANSLARININ KARŞILAŞTIRILMASI

Tenzile ERBAYRAM

Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü
İSTATİSTİK Anabilim Dalı

Danışman: Prof. Dr. Murat ERİŞOĞLU

2020, 87 Sayfa

Prof. Dr. Murat ERİŞOĞLU
Dr. Öğr. Üyesi Aydın KARAKOCA
Dr. Öğr. Üyesi Selim GÜNDÜZ

Bu çalışmada boyut indirgeme tekniklerinin sınıflandırma performansları karşılaştırılmıştır. Boyut indirgeme teknikleri özellik seçimi ve özellik çıkarma olmak üzere iki kategoride incelenmiştir. Çalışmada sınıflara ait değişim katsayısına dayalı yeni bir özellik seçim yöntemi önerilmiştir. Boyut indirgeme tekniklerinin karşılaştırılmasında nicel verilerden oluşan gerçek veri setleri kullanılmıştır. Boyut indirme teknikleri, karesel diskriminant analizinde doğru sınıflandırma olasılığı, entropy ve kappa katsayısı bakımından karşılaştırılmıştır. Çalışma sonuçları önerilen özellik seçim yöntemlerinin boyut indirgeme amacıyla kullanılabileceğini göstermiştir.

Anahtar Kelimeler: Boyut İndirgeme, Entropi, Kappa katsayısı, Özellik Seçimi, Özellik Çıkarma, Sınıflandırma Doğruluğu

ABSTRACT

MS THESIS

COMPARISON OF THE CLASSIFICATION PERFORMANCES OF THE DIMENSIONALITY REDUCTION TECHNIQUES

Tenzile ERBAYRAM

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE OF NECMETTİN ERBAKAN UNIVERSITY
THE DEGREE OF MASTER OF SCIENCE IN STATISTICS

Advisor: Prof. Dr. Murat ERİŞOĞLU

2020, 87 Pages

Jury
Prof. Dr. Murat ERİŞOĞLU
Asist. Prof. Dr. Aydın KARAKOCA
Asist. Prof. Dr. Selim GÜNDÜZ

In this study, the classification performances of dimensionality reduction techniques were compared. Dimensionality reduction techniques are discussed in two categories as feature selection and feature extraction. Simulation study and real data sets were used to compare dimensionality reduction techniques. Dimension reduction techniques were compared in terms of classification accuracy, entropy and kappa coefficient in quadratic discriminant analysis. The results of the study showed that the proposed feature selection methods can be used for dimension reduction.

Keywords: Classification Accuracy, Dimensionality Reduction, Entropy, Feature Selection, Feature Extraction, Kappa Coefficient

ÖNSÖZ

Lisans ve yüksek lisans eğitimi boyunca benden her türlü yardımlarını esirgemeyen Necmettin Erbakan Üniversitesindeki hocalarıma verdikleri emeklerinden dolayı teşekkürlerimi iletmek isterim. Yanında çalışmaktan gurur duyduğum, bilgi ve tecrübelerini paylaşarak yetişmemde emeği olan, her aşamada hoşgörü, anlayış ve desteğiyle bana yol gösteren ve hiçbir bilgisini desteğini esirgemeden yardımcı olan tez danışmanım Prof. Dr. Murat ERİŞOĞLU hocama çok teşekkür ederim.

Beni bu yaşıma kadar getiren, sevgi ve varlıklarıyla her koşulda arkamda olan, mutluluk ve huzur kaynağım olan ailem ve daima yanımda olan sevdiklerime Sonsuz sevgi ve teşekkürlerimi sunarım.

Tenzile ERBAYRAM
KONYA-2020

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

Simgeler

w_i : i. değişken ait ağırlık katsayısı

σ : Kernel genişliği

ξ(w) : Tek gözlem dışarda sınıflandırma doğruluğu

∂ : Kısmi türev

λ : Düzenleme parametresi

η : Pozitif küçük sabit sayı

α : Başlangıç adım uzunluğu

I : Birim matris

$\Sigma = S$: Varyans Kovaryans Matrisi

λ_i : Öz değer

R : Korelasyon matrisi

δ : Konfigürasyon uzaklık

d : Orijinal uzaklık

Kısaltmalar

TDA : Tekil Değer Ayrıştırma

MSH : Minimum Sınıflandırma Hatası

GMSH : Genelleştirilmiş Minimum Sınıflandırma Hatası (GMSH).

LDA : Lineer Diskriminant Analizi

LLE :Yerel Doğrusal Eşleşme

TBA : Temel Bileşenler Analiz

DSO : Doğru Sınıflandırma Olasılığı

DVM : Destek Vektör Makinesi

FDA : Fisher Doğrusal Ayırıcılar

BBA : Bağımsız Bileşenler Analizi

DDA : Doğrusal Diskriminant Analizi

YDDA : Yerellik Duyarlı Diskriminant Analizi

NCFS : Komşuluk Bileşeni Özellik Seçimi

NCA : Komşuluk Bileşen Analizi

ÇBÖ : Çok Boyutlu Ölçekleme

ISOMAP :İzometrik Eşleme

1. GİRİŞ

Bilişim teknolojileri ve veri işleme alanında meydana gelen gelişmelerle birlikte, hem birim sayısında hem de özellik (değişken) sayısında artış olmuş ve büyük hacimli veri kümeleri oluşmuştur. Veri kümelerinin hacimlerinin büyümesi beraberinde verilerin yorumlanması ve modellenmesinde bir takım problemlere neden olmuştur. Araştırmacıların ne tür bir süreçle ilgilendiğine bağlı olarak çeşitli kaynaklardan yüksek boyutlu veriler elde edilebilir. Doğadaki birçok süreç farklı birçok değişkenin sonucu olarak meydana gelir. Birçok araştırmacı yararlı göstergeleri yakalamak ve daha doğru sonuçlar elde etmek için ilk başta mümkün olduğunca birçok özelliği çalışmalarına dahil etme eğilimindedir. Ancak belli bir noktadan sonra artan özellik sayısı modelin performansını düşürecektir. Özellik sayısının çok fazla olması durumunda veri kümeleri bazen istatistiksel analizler için gerekli varsayımları sağlayamaz bazen de aşırı uyum problemi neden ile oluşturulan modellerin genelleme performansı düşer. Çok sayıda özelliğin olduğu veri setlerinde çoğu özellik birbiri ile ilişkilidir ve bu modellin uyumunu anlamsız bir şekilde artırır. Böyle durumlarda modelin oluştuğu eğitim setlerinde model çok iyi bir performans gösterirken test verilerinde model performansı çok düşük gerçekleşir. Çünkü model veriye aşırı bağımlıdır. Bu nedenle aşırı uyumluluk durumunda model genelleme özelliğini kaybeder ve bu istenmeyen bir durumdur.

Özellik sayısının çok fazla olduğu veri kümelerinin görselleştirilebilmesi, analizlere uygun hale getirilebilmesi, veriden daha basit ve daha anlamlı modeller üretebilmesi için boyut indirgemek gerekir. Boyut indirgeme en basit anlatımla, orijinal verinin taşıdığı bilgiden mümkün olduğunca az bir kayıpla boyut sayısını azaltma sürecidir. Boyut indirgeme özellik seçimi veya özellik çıkarımına göre gerçekleştirilir. Özellik seçimi veri kümesini en iyi temsil edecek bir alt özellik kümesinin seçimi olarak tanımlanabilir. Özellik seçimi, verinin temsilinde daha az önemli olan özelliklerin veri kümesinden çıkartılması ile boyut indirgeme işlemini gerçekleştirir. Özellik çıkarımı tüm özellikleri dikkate alarak bu özelliklerin doğrusal veya doğrusal olmayan bileşenleri ile daha az boyutta veriyi temsil etme sürecidir. Özellik çıkarma yöntemlerinde elde edilen bileşenlerin birkaçı değişimin büyük bir kısmını açıklayabilmektedir. Özellik çıkarma yöntemlerinde değişimi açıklama yüzdesi çok düşük olan bileşenler dışarda tutularak boyut indirgeme gerçekleştirilir.

Hem özellik seçimi hem de özellik çıkarma yöntemleri içerisinde çok sayıda farklı yaklaşım bulunmaktadır. Verinin yapısı ve özelliklerin ölçümlerindeki farklılık gibi nedenlerle uygun boyut indirgeme yönteminin seçimi önemli bir problemdir. Bu çalışmada özellik sayısının birim sayısından fazla olduğu yüksek boyutlu veri setlerinde seçili özellik seçim ve özellik çıkarma yöntemlerinin karesel diskriminant analizindeki sınıflama performansı incelenecektir. Nicel verilerden oluşan gerçek veri setlerinde gerçekleştirilecek karşılaştırılmasında doğru sınıflandırma olasılığı, entropy ve kappa istatistiği kullanılacaktır. Sınıf bilgisine dayalı olarak iki yeni özellik seçim kriterinin önerileceği çalışmada, özellik seçimi ve özellik çıkarma yöntemlerinin bir arada kullanılmasının sınıflama performansı üzerindeki etkisi de incelenecektir.

Bu çalışma beş bölümden oluşmaktadır. Çalışmanın ikinci bölümünde boyut indirgeme yöntemleri ilgili literatür çalışması gerçekleştirilmiştir. Çalışmanın üçüncü bölümünde boyut indirgeme yöntemleri özellik seçim ve özellik çıkarma yöntemleri olarak iki sınıfta incelenmiştir. Özellik seçim yöntemleri bölümünde değişim katsayısı, F test istatistiği, küme merkezlerine olan uzaklık, Fisher skoru, t skoru, Welch'in t istatistiği, komşuluk bileşen analizi (NCA), Relief, ReliefF algoritmaları ve sınıf yapısına dayalı önerilen özellik seçim yöntemleri incelenmiştir. Özellik çıkarma yöntemlerinde temel bileşenler analizi (TBA), çok boyutlu ölçekleme analizi (ÇBÖ), yerel doğrusal eşleme (LLE) ve izometrik eşleme (ISOMAP) yöntemleri incelenmiştir. Çalışmanın dördüncü bölümünde öncelikle karesel diskriminant analizi ve karşılaştırma kriterleri hakkında bilgi verilmiştir. Daha sonra incelenen veri setleri hakkında bilgi verdikten sonra üç farklı uygulama ile boyut indirgeme yöntemlerinin sınıflama performansı incelenmiştir. Sonuç bölümünde çalışmadan elde edilen sonuçlar özetlenmiş ve bundan sonraki benzer çalışmalar için önerilerde bulunulmuştur.

2. KAYNAK ARAŞTIRMASI

Harsanyi ve ark.(1994), veri boyutlarını eş zamanlı olarak azaltan, istenmeyen veya karışan spektral imzaları baskılayan ve ilgilenilen bir spektral imzanın varlığını algılayan bir tekniği önermişlerdir. Her bir piksel vektörünü istenmeyen imzalara ortogonal olan bir alt uzay üzerine yansıtmışlardır. Ortogonal alt uzay izdüşümü operatörü ile ilgili k imzaların genişletilebilir böylece k boyutsallığı azaltıp hiperspektral görüntünün aynı anda sınıflandırılmasını sağlayabileceğini söylemişlerdir. Yaklaşım hem karışık olarak hem de karışık pikseller için geçerlidir.

Sarwar ve ark. (2000), öneri sistem veri tabanlarının boyutsallığını azaltmak için Tekil Değer Ayrıştırma (TDA) adlı bir teknolojiyi ile iki farklı deney sunmuşlardır. Her iki deneyde de ortak bir filtreleme kullanarak bir öneri sisteminin kalitesi ile TDA kullanarak bir öneri sisteminin kalitesini karşılaştırmışlardır. İlk deneyde, tüketici tercihlerinin tahmin edilmesinde iki öneri siteminin etkililiğini, ürünlerin açık derecelendirme veri tabanına dayanarak karşılaştırmışlardır. İkinci deneyde ise iki öneri sisteminin, bir E-Ticaret sitesinden gerçek hayattaki bir müşteri satın alma veri tabanına dayanarak Top –N listeleri üretme etkinliğini karşılaştırmışlardır. Deneye göre TDA ‘nın öneri sistemlerinde karşılaşılan zorlukların birçoğunu karşılama potansiyeline sahip olduğunu ortaya koymuşlardır.

Roweis ve ark. (2000),yüksek boyutlu verilerin düşük boyutlu, komşuluklarını koruyan bağlarını hesaplayan denetimsiz bir öğrenme algoritması olan lokal lineer yerleştirmeyi önermişlerdir.

Bingham ve ark. (2001), boyutsal bir indirgeme amacı olarak rastgele iz düşüm kullanımı üzerine deneyler önermişlerdir. Verilerin rastgele bir alt boyutlu alt uzay yerine yansıtılmasının, temel bileşen-analitik gibi geleneksel boyutsal indirgeme yöntemleri ile karşılaştırılabilir sonuçlar getirdiğini göstermişlerdir ve veri vektörlerinin benzerliğini rastgele yansıtma altında korunmuştur.

Keogh ve ark. (2001), yüksek boyutlu zaman sersi verilerinin boyut indirgemesinde yaygın kullanılan üç indirgeme tekniği ile parçalı kümeleme yaklaşımı ismini verdikleri yöntem ile karşılaştırmışlardır. Önerdikleri yaklaşımın üstünlüklerini teorik ve ampirik olarak ortaya koymuşlardır. Önerdikleri yöntemin diğer indirgeme yöntemlerine göre daha hızlı ve etkili olduklarını vurgulamışlardır.

Wang ve ark. (2003), özellik sınıflandırması için minimum sınıflandırma hatası (MSH) eğitim algoritması (başlangıçta sınıflandırıcıların optimizasyonu için önermişlerdir.)'nı incelmişlerdir. MSH eğitim algoritmasının eksikliklerini gidermek için genelleştirilmiş bir MSH(GMSH) eğitim algoritması önermişlerdir. Lineer Diskriminant Analizi(LDA), TBA, MSH ve GMSH algoritmaları, doğrusal dönüşüm yoluyla özeliklerini çıkarmışlardır. Destek vektör makinesi (DVM), parametrik uzaydaki doğrusal olmayan karar sınırlarını elde etmek için doğrusal olmayan öz fonksiyonlarını kullanan yeni geliştirilen bir model sınıflandırma algoritmasıdır. Bu çalışmada, DVM ayrıca doğrusal özellik çıkarma algoritmalarını araştırmış ve karşılaştırmışlardır.

Robnik Sikonja ve Kononenko (2003), değişken seçimi için Relief ve ReliefF'in Teorik ve Ampirik analizi üzerinde çalışmışlardır. Bu çalışmada teorik ve ampirik olarak nasıl çalıştıklarını ve neden çalıştıklarını, teorik ve pratik özelliklerini, parametrelerini, ne tür bağımlılıkları saptadıklarını, çok sayıda örnek ve özelliğe nasıl ölçeklediklerini, verilerin nasıl örnekleneceğini tartışmışlardır. Bunlar, değişken sayısı fazla olduğunda ne kadar sağlam oldukları, ilgisiz ve gereksiz değişkenlerin çıktılarını nasıl etkilediği ve farklı ölçümlerin onları nasıl etkilediğini incelemişlerdir.

Zhang ve ark. (2004) , manifold öğrenme ile doğrusal olmayan boyut indirgeme için yeni bir algoritma önermişlerdir. Algoritmada daha az bir hata analizi sunarak ve yeniden yapılandırma hatalarının ikinci dereceden doğrulukta olduğunu göstermişlerdir. Daha fazla araştırma ve iyileştirme için çeşitli teorik ve algoritmik konuları ele almışlardır.

Turhan (2004), örüntü tanıma için doğrusal olmayan boyut indirgeme yöntemleri üzerinde çalışmış ve tezinde gözetimsiz boyut indirgeme yöntemlerini çeşitli standart değerlendirme veri kümelerinde deneyerek incelemiş ve karşılaştırmıştır. Daha önce karşılaşılmamış veri noktaları sorununu çözmek için eşleme fonksiyonlarının öğrenimini önermiştir. Veri dağılımının doğasında bulunan ölçü birimlerinin kullanılmasının Öklid mesafesinden daha iyi modellediğini ve yüksek boyutlu veri modellerinin doğruluklarının artırdığını gözlemlemişlerdir.

Yakut (2008), gizliliği koruyarak boyut indirgeme tabanlı işbirlikçi filtreleme yöntemi üzerinde çalışmıştır. Çalışmasında kişisel gizliliğe zarar vermeden Eigentaste algoritmasına dayalı işbirlikçi hizmetleri sunmak için çözümler önermişlerdir. Önerilen

yöntemlerin doğruluk, gizlilik ve ek maliyet analizlerini yaparak sonuçlar çıkarılmış ve öneriler sunmuştur.

Börü (2009), TBA ile Türkiye'nin finansal gelişimi ile ilgili bir çalışma yapmıştır. Bu çalışmada, finansal gelişmenin değişik yönlerini literatürdeki diğer çalışmalarda kullanılan yaklaşık değişkenlerden daha iyi temsil ettiğini ve yaklaşık değişkenler sunarak ölçüm sorununu çözmeyi amaçlamıştır.

Van Der Maaten ve ark.(2009), TBA ve klasik ölçekleme tekniklerini inceleyerek sistematik bir şekilde karşılaştırılmasını sunmuşlardır. Doğrusal olmayan yöntemlerinin performanslarını yapay ve doğal görevler üzerinde incelemişlerdir. Bu deney sonuçları, doğrusal olmayan yöntemlerin seçilmiş yapay görevler üzerinde iyi performans gösterdiğini, fakat bu güçlü performansın gerçek dünyadaki görevlere uzanmayacağını ortaya koymuşlardır. Bu çalışmada mevcut doğrusal olmayan yöntemlerin zayıflıklarını tanımlayarak açıklamakta ve doğrusal olmayan boyutsal indirgeme yöntemlerinin performansının nasıl geliştirilebileceğini göstermişlerdir.

Carreira-Perpinan ve ark.(2010), verilerin düşük boyutlu koordinatlarının sezgisel, doğrusal olmayan objektif fonksiyonunu optimize eden yeni bir boyut indirgeme yöntemi olan elastik gömülmeyi önermişlerdir. Bu yöntemde spektral bir yöntem, Laplasiyen özdeşlikleri ve doğrusal olmayan bir yöntem, stokastik komşu gömülmesi gibi temel bir ilişki ortaya koymuşlardır. Elastik gömülmenin hem koordinatları hem de veri noktaları arasındaki ilişkilerini öğrenirken gözlemlenebileceğini göstermişlerdir. Elastik gömülmeyi eğitmek ve homotopi parametresinin kritik değerini karakterize etmek ve yöntemin davranışını incelemek için bir homotopi yöntemi sunmuşlardır. Sabit bir homotopi parametresi için, çok etkili ve kullanıcı parametrelerini gerektirmeyen, global olarak yakınsayan bir yineleme algoritması oluşturmuşlardır. Son olarak, örnek dışı noktalara bir uzantı vermişlerdir. Standart veri kümelerinde, elastik gömülme sonuçları SNE' den daha iyi ancak daha verimli ve sağlam bir şekilde elde ettiğini gözlemlemişlerdir.

Zhang ve ark. (2010) , orijinal özellik açıklaması ve ilişkili sınıf arasındaki bağımlılığı en üst düzeye çıkarırken, orijinal verileri daha düşük boyutlu bir özellik alanına yansıtmaya çalışan çok etiketli bir boyut indirgeme tekniğini önermişlerdir. Hilbert-Schmidt bağımsızlık kriterine dayanarak, boyut indirgeme sürecini daha iyi olmasını sağlayan kapalı bir çözümünü üretmişlerdir.

Rosman ve ark.(2010) , topolojik olarak kısıtlı izometrik yerleştirme ile doğrusal olmayan boyut indirgeme tekniğini önermişlerdir. Düz gömme işlemi genellikle komşu özellikler arasındaki mesafelere dayanan çoğunlukla birbirinden uzak olan özelikler arasındaki mesafeleri belirli manifoldun konveksizliğinden dolayı özellik manifoldundaki gerçek mesafenin bir tahminini sağlamaktadır. Sınırları olan yerel düz manifoldların içsel geometrisini öğrenmek için hem yerel hem de küresel mesafeleri kullanarak doğrusal olmayan boyut indirgeme için bir çerçeve oluşturmaktadır. Önerilen algoritma doğrusal olmayan yapılarla eşleştiğinde güçlü seslere dayanıklı olduğu gözlemlenmiştir.

Durmaz (2011), metin sınıflandırmada boyut azaltmanın etkisi ve özellik seçimi üzerinde çalışmıştır. Bu çalışmada metinlerin tümü terim frekansı – ters doküman frekansı (TF–IDF) vektörleri ile temsil etmiştir. Çalışmada uygulanan Ayrık Kosinüs Dönüşüm yöntemi ve Varyans Oranı ile özellik seçim yöntemi metin vektörlerinden oluşturulan TF–IDF vektör uzayının boyutunun indirgeyerek sınıflandırma için daha etkili sonuçların elde etmek amacıyla kullanmıştır. Boyutları indirgenmiş vektörlerle başarılı sonuçlar elde etmiştir.

Dehak ve ark.(2011), daha önce konuşmacı tanımlaması alanında geliştirilen toplam değişkenlik yaklaşımına dayalı yeni bir dil tanımlama sistemi önermişlerdir. Düşük boyutlu i-vektör uzayında en belirgin özellikleri çıkarmak için çeşitli teknikler kullanılan ve geliştirilen sistem herhangi bir işlem sonrası veya arka uç tekniğine gerek kalmadan 2009 LRE değerlendirme setinde mükemmel performans sağladığı gözlemlenmiştir. Bu sistem diğer akustik sistemlerle birleştirildiğinde ek performans kazandırdığı gözlemlenmiştir.

Akyürek (2012), hiperspektral görüntülerde boyut indirgeme yöntemlerinin karşılaştırılmasını yapmıştır. Bu tez çalışmasında hiperspektral görüntülerin boyutlarının doğrusal ve doğrusal olmayan yöntemler yardımıyla indirgenerek kullanılan boyut indirgeme yöntemlerinin karşılaştırılması yapılmıştır. Örnek bir hiperspektral görüntü verisi üzerinde seçilen üç adet görüntü parçasını doğrusal ve doğrusal olmayan yöntemleri incelemiştir.

Li ve ark.(2012), tek-Gauss varsayımından yola çıkarak, verilerin istatiksel yapısını kullanmak için tasarlanmış bir sınıflandırma paradigmasını önermişlerdir. Önerdikleri metodun, multimodal yapısını korurken verinin boyutsallığını azaltmak için

yerel Fısher ‘in Diskriminant analizini kullanmışlardır. Sonraki adımda Gaussian karışım modeli veya destek vektör makinesi ile indirgenmiş boyutlu çok modlu verilerin etkili bir şekilde sınıflandırılmasını sağlamışlardır. Birkaç farklı çoklu sınıf hiperspektral sınıflandırma görevlerindeki deneysel sonuçlar, önerilen yaklaşımın birçok geleneksel alternatifi önemli ölçüde geride bıraktığını gözlemlemişlerdir.

Yang ve ark.(2012),Çok boyutlu verilerde Komşuluk Bileşen Analizi ile değişken seçim yöntemi üzerinde çalışma yapmıştır. Bu çalışmada, normalizasyon terimiyle beklenen bir kez dışarıda bırakılan sınıflandırma doğruluğunu maksimize ederek bir özellik ağırlıklandırma vektörünü öğrenen komşu temelli bir özellik ağırlıklandırma algoritması önerilmektedir. Algoritma, verilerin dağıtımı hakkında herhangi bir parametrik varsayımda bulunmaz ve doğal olarak çok sınıflı problemlere ölçeklenir. Yapay ve gerçek veri setleri üzerinde yapılan deneyler, önerilen algoritmanın ilgisiz özelliklerin sayısındaki artışa büyük ölçüde duyarsız olduğunu ve çoğu durumda en gelişmiş yöntemlerden daha iyi performans gösterdiğini göstermektedir.

Kuş (2013), bireylerin ayrılmasında kulak biyometrisinin kullanımında temel bileşenler analizi (TBA) ile Fısher doğrusal ayırıcılar(FDA) yöntemini bir arada kullanmayı önermiştir. Çok sayıda özellik dikkate alınarak oluşan kulak biyometrisi verisinde öncelikle TBA kullanılarak hem boyut indirgenmiş hem de bağımsızlık kazandırılmıştır. TBA’den sonra indirgenmiş veride en iyi ayıraç olan özellikler FDA yöntemi belirlenmiştir. Çalışmada önerilen yaklaşım, histogram matrisi ile tanımlama yaklaşımı ile karşılaştırılmıştır. Karşılaştırma sonucunda histogram matrisi ile tanımlama yaklaşımının sonuçların güvenilir olmadığı doğru tanımlama olasılığın önerilen yönteme göre düşük olduğu belirtilmiştir.

Özgür (2013), kategorik verilerde boyut indirgeme yöntemiyle çoklu uyum analizi ve sağlık bilimlerinde beslenme üzerine bir uygulama yapmıştır. Bu çalışmada, Marmara Üniversitesi' nde ki öğrencilerin beslenme alışkanlıkları dört yapraklı yonca modeliyle incelenerek, öğrencilerin beslenme alışkanlıklarını etkileyen faktörler arasındaki ilişkileri ortaya koymayı amaçlamıştır. Çalışmaya göre beslenme alışkanlığına bakıldığı zaman, kız öğrencilerin erkek öğrencilere göre ve ailelerinin yanında yaşayan öğrencilerin ise diğer öğrencilere göre daha düzenli ve sağlıklı beslendiği sonucunu ortaya koymuştur.

Kurt (2013), Temel bileşen analiziyle öznitelik seçimi ve görsel nesne sınıflandırma ile ilgili bir çalışma yapmıştır. TBA'ya dayalı betimleyicinin ağırlıklandırılmış açılarının histogramlarını kullanan yeni bir betimleme tekniği önermişlerdir. Kullanılan diğer betimleme tekniklerine göre oldukça iyi sonuçlar verdiğini ve diğer detektörlerin bulduğu ilgi noktalarından elde edilen karesel çerçevenin üzerinde kullanılmasıyla birlikte sınıflandırma başarısının daha da artığını gözlemlemiştir.

Kozal (2014), Hiperspektral görüntülerin sınıflandırılması sürecinde yapılacak boyut indirgeme aşaması için farklı boyut indirgeme yöntemlerinin sınıflandırma açısından başarımları ve hesaplama performanslarını karşılaştırmışlardır. Sınıflandırıcı olarak DVM ve En Yakın Komşuluk Sınıflandırıcı yöntemlerini kullanmışlardır. Boyut indirgeme yöntemlerinin kullandığı veriyi indirgeyerek işlem süresi ve sınıflandırma başarımlarındaki performans değişiklikleri ortaya koymuştur. Yüksek boyutlu verilerin sınıflandırılmasında, eğitim verisinin yetersiz kalması ile sınıflandırma performansındaki düşüş etkisi incelenmiş, farklı boyut indirgeme yöntemlerinde bu olgunun etkisini azaltabildiğini gözlemlemiştir.

Çatalbaş (2014), çalışmada temel bileşenler ve kanonik korelasyon analizlerinin imge tanıma ve sınıflandırma problemlerindeki rolünü incelemişlerdir. Bu çalışmada ise kanonik bileşenler analizinin de imge tanıma ve sınıflandırma problemleri için etkin bir öznitelik belirleme ve boyut indirgeme yöntemi olarak kullanılabileceği gösterip TBA ile kıyaslama yapmışlardır. Çalışmada, çok sınıflı imge sınıflandırma problemlerine yönelik olarak, çoklu kanonik bileşenler analizinin bir öznitelik belirleme ve indirgeme paradigması olarak kullanılabileceğini göstermişlerdir. Örnek problemler üzerinden, kanonik korelasyon analizinin imge tanıma ve sınıflandırma problemlerinde TBA 'ya kıyasla daha etkin bir boyut indirgeme yöntemi olduğunu gözlemlemişlerdir. Yapılan çalışmalarda, LDA ve en yakın komşu gibi temel sınıflandırma algoritmalarının kanonik bileşenler analizi ile uyumlu olarak kullanılabileceği ve doğrusal olmayan sınıflandırma yöntemlerine başvurmadan daha yüksek tanıma başarımları elde edileceği gözlemlenmiştir.

Tilki (2014) , TBA tabanlı yüz tanımada bir uygulama yapmıştır. Bu çalışmada TBA tabanlı bir yüz tanımlama sistemi geliştirerek ve farklı alanlarda kullanılmak üzere önermiştir. Yüz tanımlama işleminin gerçekleştirmek için TBA analizini kullanarak bir

algoritma gerçekleştirmişlerdir. Sonucunda elde edilen bulgulara göre ileri araştırmalar için bazı önerilerde bulunmuştur.

Durgabaı (2014), Relief Algoritmasını kullanarak değişken seçimi üzerinde çalışmıştır. Bu yazıda, hata minimizasyonu ile değişken seçimini belirleyen yeni bir algoritma önermişlerdir. Önerilen algoritmik çerçeve, parametrik olmayan bir tahminci tarafından tahmin edilen Bayes hata oranını en aza indirerek bir özellikler alt kümesini seçer. Sonuç olarak iki özellik ağırlıklandırma algoritmasını karşılaştırmışlardır. Bu nedenle seçilen ilgili özellikler, daha iyi doğrulama için bazı kümeleme algoritmaları kullanılarak kümelerde gösterilmiştir. Büyük veri kümeleri için iyi bilinen kümelenme tekniklerinin sınırlamaları ve önerilen kümelenme yönteminin detayları, Liderler-Alt Klasörler sunmuşlardır. Sayısal veri setleri konusundaki deneysel sonuçları, Liderler - Alt Liderler algoritmasının iyi çalıştığını göstermektedir. Her bir kümedeki alt grupları / alt kümeleri düşük hesaplama maliyetiyle bulmak için önerilen yöntemle, gerekli seviyelerde hiyerarşik yapı üretilebilmiştir. Alt kümelerin temsilcileri, sınıflama doğruluğunun geliştirilmesinde yardımcı olmuştur. Davies-Bouldin endeksi de, farklı yarıçapta bile sonuçların eşdeğer olduğunu gösteren iyi bir performans göstermiştir.

Çukur (2015), sezgisel hiperspektral görüntülerde boyut indirgeme yöntemleri üzerinde çalışmıştır. Bu çalışmada, hiperspektral görüntülere değişik sezgisel yöntemler uygulanarak, bant seçimi yaklaşımıyla boyut indirgeme yapılmış ve veriyi en iyi temsil eden bantları bulmuştur. Değişik gruplama yöntemleri kullanılarak bantlar arası benzerliğe dayanılarak yeni gruplar oluşturmuş ve boyut indirgemeyi bu gruplar üzerinde gerçekleştirmiştir. Boyut indirgeme sonucu bulunan en iyi bantlar DVM algoritması ile sınamıştır. Bu sonuca göre de yöntemlerin hiperspektral görüntülerde kullanılacağını göstermişlerdir.

Pamukçu (2015), yüksek boyutlu kanser sınıflama probleminde bilgi karmaşıklığı kriteri ile aykırı gözlem tespiti ve boyut indirgeme yöntemini incelemiştir. Bu çalışmada, mikro dizilim verilerinin analizinde, Maksimum Entropi kovaryans matrisinin ve diğer bazı sağlam veya düzgünleştirilmiş kovaryans matrislerinin kullanımı ile S'in dejenere olmasının önüne geçilmiş ve dolayısıyla boyut indirgeme ve aykırı gözlemlerin tespitleri mümkün hale getirmiştir. Boyut sayısına karar verirken önemli bileşenler, klasik yöntemlerden farklı olarak bilgi karmaşıklığı kriteri ICOMP yardımıyla seçilmiş ve verinin boyutu indirgendikten sonra elde edilen alt uzay üzerinde

ICOMP yardımı ile verideki aykırı gözlemler tespit etmiştir. Bilgi Karmaşıklığı Kriteri ICOMP ile önerilen bu yaklaşımların hem benzetim verilerine hem de çeşitli mikro dizilim veri setlerine uygulanması sonucunda, boyut indirgemenin ve aykırı gözlemlerinin tespitinin başarılı bir şekilde yapılabildiği gözlemlemiştir.

Öztürk (2016), EEG sinyallerinde farklı boyut indirgeme ve sınıflandırma yöntemlerinin karşılaştırılması yapılmıştır. Bu çalışmada, epileptik ve epileptik olmayan EEG sinyallerinden elde edilen özniteliklerin boyutlarının TBA ve Bağımsız Bileşenler Analizi (BBA) yöntemleri ile indirgenmesinin sınıflandırma başarısı üzerine etkilerinin belirlenmesi ve Lineer Diskriminant Analizi (LDA) ile Destek Vektör Makinesi (DVM) yöntemlerinin sınıflandırma performanslarının karşılaştırılması yapılmıştır. Sonucunda ise, en yüksek sınıflandırma başarısı %92,2 duyarlılık %85,6 özgüllük ve %88,9 doğruluk oranlarıyla özniteliklerde boyut indirgenme yapılmadan ve radyal tabanlı çekirdek fonksiyonunun kullanıldığı DVM yöntemi ile elde edilirken BBA ve TBA ile boyutu indirgenen özniteliklerle yapılan sınıflandırmalarda da benzer sonuçlar elde edilmiştir.

Yıldız ve ark.(2016), Sınıflandırma yöntemleri üzerinde doğrusal boyut indirgeme yöntemlerinin karşılaştırılmasını incelemişlerdir. Doğrusal boyut indirgeme yöntemlerinin işlem süreleri açısından en başarılı metot olarak TBA ve Doğrusal Diskriminant Analizi (DDA) olarak gözlemlenirken en kötü performansı da Yerellik Duyarlı Diskriminant Analizi (YDDA) yöntemi olarak gözlemlenmiştir. Elde edilen sonuçlara göre, incelenen doğrusal boyut indirgeme yöntemleri sınıflandırma sürelerini kayda değer bir şekilde azaldığı gözlemlenirken YSA yöntemi ile sınıflandırma işleminin süresi çok ciddi bir azalma göstermiştir.

Çetin (2016), Kazalara neden olan sürücü alışkanlıklarını bulmak ve sağlıklı sınıflandırma tahminlemesi yapmak için boyut indirgeme yöntemlerini uygulamıştır. Bu çalışma boyunca veri sayısı ve özellik kümesi fazlalığından, sınıflandırma başarımı çeşitli nedenlerde düştüğü gözlemlenmiştir. Kazaya etki eden dinamiklerin bulunması ve sınıflandırma başarımını arttırmak için, veri madenciliği öncesi veriyi işlemeden önce, özellik boyut indirgemesi yöntemlerinden, öznitelik arama veya özellik alt küme seçimi yöntemlerini kullanmıştır.

Singh ve ark. (2016), Çok boyutlu veri için Değişken Seçme Yöntemleri Üzerine Literatür Taraması üzerine bir çalışma yapmıştır. Uygun özellik seçim yönteminin

belirlenmesi, çok boyutlu veri içeren belirli bir makine öğrenme görevi için çok önemlidir. Bu nedenle, araştırmanın, özellikle yüksek boyutlu verilerde makine öğrenim görevlerinin performansını artırmak için uygun özellik seçim yöntemini geliştirmeye adanmış araştırma topluluğu için çeşitli özellik seçim yöntemleri üzerinde yapılması gerekmektedir. Bu amacı gerçekleştirmek için, bu çalışmada yüksek boyutlu veri alanları için çeşitli değişken seçim yöntemleri hakkındaki literatür taramasının tamamını ele almışlardır.

Toktay (2017) ,faktör ve diskriminant analizinin Iğdır üniversitesi öğrencileri üzerinde bir uygulamasını yapmıştır. Öğrencilerin eğitim, öğretim, ders, okul memnuniyeti ve şehir hakkındaki görüşlerini içeren likert tipi sorular, Üniversite'ye girdikleri Puan türü, Barınma ve Ulaşım şekline göre uygulanmış olan diskriminant analizi için gerekli olan normallik varsayımı sebebiyle, ayrıca herhangi bir varsayım şartı bulunmayan Multinomiyal Regresyon Analizini de uygulamıştır.

Makul ve ark. (2017), çalışmada son zamanların popüler konularından olan akan verilerin kümelenmesi üzerine yeni yaklaşım önermişlerdir. Yaklaşım graf yapısı kullanılarak herhangi şekle sahip kümeleme işlemi gerçekleştiren CEDAS algoritmasına, doğrusal ayırtaç analizinden akan verilere uygulanabilmek için uyarlanarak elde edilen yerelleştirilmiş doğrusal ayırtaç analizi yaklaşımını entegre edilmesine dayandırmışlaradır. Önerilen yaklaşım sıkça kullanılan CoverType, DS1 ve Mackey-Glass akan veri setleri üzerinde denemişlerdir ve elde edilen sonuçlar hibrit yaklaşımının başarısını göstermişlerdir.

Yüksek ve ark. (2017), ANFIS modelinin eğitim performansının üzerindeki etkilerini karşılaştırmak için farklı boyut indirgeme yöntemlerini önermişlerdir. Farklı boyut indirgeme yöntemleri kullanılarak giriş değişkenlerinin sayılarının indirgenmesi ANFIS modeli ile probleme ait en uygun çözümün hangisi olduğunu araştırmışlardır. Bu çalışmada, farklı boyut indirgeme yöntemlerinin ürettiği sonuçlar karşılaştırılarak ANFIS'in eğitimi için hangi yöntemin kullanılmasının daha iyi olduğunu gözlemlemişlerdir.

Lai C ve ark (2017) , Temporal lob epilepsisinin diskriminant analizi için değişken seçim yöntemlerinin karşılaştırması üzerinde bir çalıştırma yapmışlardır. Bu çalışmada, sol Temporal lob epilepsili 41 hastanın yapısal MR görüntüleri, sağ Temporal lob epilepsili 34 hastanın ve 58 normal kontrolün elde edildiği ve kortikal

kalınlık, kortikal yüzey alanı, gri madde hacmi olmak üzere dört çeşit kortikal önlem alınmıştır ve ortalama eğrilik, diskriminant analiz için incelenmiştir. Sonuçlar, destek vektör makine - özyinelemeli özellik ortadan kaldırılması 'nın (% 84'den fazla doğrulukla sınıflandırmaların çoğu), seyrek kısıtlı boyutluluk azaltma modeli 'ni ve t-testinden sonra en yüksek performansı elde ettiğini göstermiştir. Özellikle, yüzey alanı ve gri madde hacmi belirgin bir ayırt edici yetenek sergiledi ve dört kortikal ölçü birleştirildiğinde destek vektör makinesi ‘nin performansı önemli ölçüde artırmıştır. Bu çalışma, kortikal özelliklerin anormal anatomik paternlerin tanınması için etkili bilgi sağladığı ve önerilen yöntemlerin Temporal lob epilepsisinin klinik tanısını iyileştirme potansiyeline sahip olduğu sonucuna varmıştır.

Castro ve ark.(2018), boyut indirgeme ve çoklu dizin parçalanması için yeni bir yöntem önermişlerdir. Eşit uzunluktaki m rasgele dizilimlerin n sonlu sayıda bağımsız bloklara bölünmesinde, her birinin pozisyonunun yanı sıra bağımsız noktalarının sayısın eşzamanlı olarak ortaya çıkarmak için cezalandırılmış bir maksimum olasılık ölçütü kullanmışlardır. Önerilen algoritmaların yakınsamalarını simülasyon ve ebola virüsünün gerçek protein sekansında göstermişlerdir.

Sellami ve ark.(2018), öncelikle boyut indirgeme yöntemlerini inceleyerek Destek vektör makinelerini sınıflandırıcısını kullanarak sınıflandırma görevi için kullanıldığında performanslarını kıyaslamışlar ve sınıflandırma için özellik çıkarma ve grup seçimi kombinasyonunu önermişlerdir. Bu yöntemlerin hepsinde performanslarını gerçek hiperspektral görüntülerini kullanarak hiperspektral görüntü sınıflandırması için etkinliğini göstermişlerdir. Bu çalışmada tensör yerel koruma projeksiyonunun hiperspektral görüntü sınıflandırması için daha iyi sonuçlar verdiğini gözlemlemişlerdir.

Guo ve ark. (2018), çalışmalarında etki sınıflandırması için boyutsal indirgeme ön çalışmasını yapmışlardır. En yaygın bilinen beş tane boyut indirgeme yöntemini anlatmışlar ve karşılaştırmışlardır. DEAP veri setiyle ilgili deneyler, hiçbir yaklaşımın evrensel olarak diğerlerinden daha iyi performans gösteremediğini ve doğrudan ham özellikleri kullanarak sınıflandırma yapmanın her zaman kötü bir seçim olamayacağını göstermişlerdir.

Budak (2018) ,Özellik seçimlerinde yeni bir yaklaşım önermiştir. Bu çalışmada, filtreleme yöntemleri içerisinden Fisher Skor yöntemine alternatif olabilecek yeni bir yöntem önermişler ve bu yöntemin başarılı olup olmadığını tespit edebilmek amacıyla

sınıflandırma doğruluk yüzdeleri kullanılarak karşılaştırma yapılmıştır. Yapılan karşılaştırma sonucunda, önerilen yöntem ile seçilen tüm veri kümelerinden hesaplanan sınıflandırma doğruluk yüzdeleri Fisher Skor yöntemi ile seçilen veri kümelerine ait yüzdelerden daha yüksek olduğu görülmüştür. Sonuç olarak, filtreleme özellik seçim yöntemleri arasında sıkça kullanılan Fisher Skor yöntemine alternatif olarak önerilen yöntemden elde edilen sınıflandırma sonuçlarının daha başarılı olduğunu tespit etmişlerdir. Dolayısıyla, önerilen yöntemin özellik seçim işleminde Fisher Skor yöntemine alternatif olarak kullanılabileceği ve daha iyi sonuçlar verebileceğini söylemişlerdir.

Çiğdem ve Demirel (2018), Parkinson hastalığının tespitinde farklı özellik seçim yöntemleri kullanılarak farklı sınıflandırma algoritmalarının performans analizini incelemişlerdir. Her biri adaptif Fisher durma kriterleri özellik seçim yöntemi ile takip edilen ve her biri farklı özellik sıralaması kullanan farklı sınıflandırma yaklaşımlarını gösteren performanslar değerlendirmişlerdir. Gri madde ile beyazı birleştiren bir kaynak füzyon tekniği olan saptama performansını geliştirmek için doku haritaları ve korelasyona dayalı özellik seçimi yöntemini kullanarak tüm sınıflandırıcıların çıktılarını çoğunluk oyuyla birleştiren bir karar füzyon tekniği kullanmışlardır. Beş değişken seçim yöntemi arasında korelasyona dayalı değişken seçimi, tüm beş sınıflandırma algoritmaları için en yüksek sonuçları sağladığını ve destek vektör makinesinin, beş farklı değişken seçim metodu için en iyi sınıflandırma performansını olduğu sonucuna varmışlardır.

Catalbas ve ark. (2015), cinsiyet tabanlı bir imge sınıflandırma uygulaması üzerinde çalışmışlardır. Boyut indirgeme sürecinde, kullanılan yöntemlerden farklı olarak, denetimli bir boyut indirgeme ve öznitelik ayrıştırma olan kanonik korelasyon analizini kullanmışlardır. Seçilen bileşen sayısı ile sınıflandırma başarısı arasındaki ilişkiyi farklı boyut indirgeme yöntemlerini kullanarak incelemişlerdir. Sınıflandırma için en uygun bileşen sayısının seçilimi ise kanonik yüzler ve kanonik vektörlerin karşılıklı bilgileri üzerinde çalışmışlardır. Sonuç olarak, kanonik korelasyon analizinin, TBA ‘ya oranla cinsiyet tabanlı imge sınıflandırmada daha başarılı sonuçlar verdiğini gözlemlemişlerdir. Bu yöntemin boyutları indirgenmiş uzayda daha başarılı temsil yeteneğine sahip olduğu sonucuna varmışlardır.

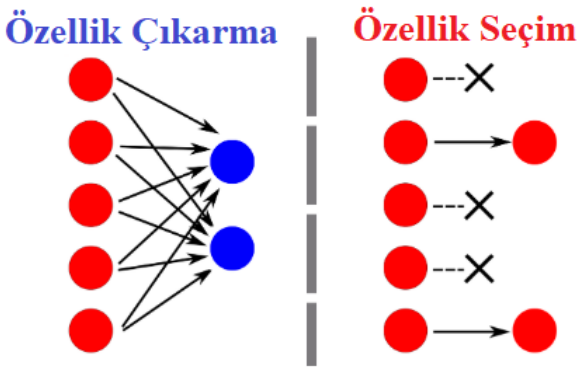
Konenko ve ark (1994), RELİEFF ile endüktif öğrenme algoritmalarının miyopinin üstesinden gelmek üzerine bir çalışma yapmışlardır. Miyopik safsızlık fonksiyonları ve görünümleri yerine, endüktif öğrenme algoritmalarının sezgisel rehberliği için Kira ve Rendell tarafından geliştirilen Relief'in bir uzantısı olan ReliefF'i kullanmayı önermişlerdir. Her seçim adımında özelliklerin tahmincisi olarak ReliefF'i kullanarak karar ağaçlarının yukarıdan aşağı indüksiyonu için bir sistem olan asistanı yeniden hayata geçirmişlerdir. Algoritma, birkaç yapay ve birkaç gerçek dünya problemi üzerinde test edilmiştir ve sonuçlar, diğer iyi bilinen makine öğrenme algoritmaları ile karşılaştırmışlardır. Yapay veri setlerinde mükemmel sonuçlar ve iki gerçek dünya problemi, indüktif öğrenmeye sunulan yaklaşımın avantajını göstermişlerdir.

Wang ve ark. (2003), yüksek çözünürlüklü uzaktan algılama görüntülerinin sınıflandırılması için değişkenlerin seçilmesine ilişkin algoritma uygulaması adlı bir uygulama yapmışlardır. Bu makalede, ReliefF algoritmasının bazı eksiklikleri, komşu örneklerin seçiminin zayıf stabilitesi problemi üzerine, algoritmanın anti-uçuculuğunu arttırmak için çoklu rasgele seçimin ortalama değerini kullanma yöntemini önererek geliştirilmiştir. Deneysel sonuçlar, geliştirilmiş ReliefF algoritmasının, sınıflandırma özellik kümelerini etkin bir şekilde oluşturabileceğini ve daha iyi sınıflandırma doğruluğunu sağladığını göstermektedir

3. BOYUT İNDİRGEME YÖNTEMLERİ

Bir veri kümesindeki değişkenlerin başka bir ifade ile özelliklerin sayısı boyutsallık olarak ifade edilmektedir. Boyut indirgeme en basit anlatımla bir veri kümesindeki özellik sayısının azaltılması şeklinde ifade edilebilir. Özellik sayısının birim sayısından fazla olduğu veri setlerinin analizi ve yorumlanması oldukça güç ve karmaşıktır. İstatistiksel analizlerin çoğu birim sayısının özellik sayısından fazla olduğu durumlar için gerçekleştirilebilmektedir. Yüksek boyutlu veri kümelerinde genellikle özellikler birbiriyle yüksek ilişkilidir ve dolayısıyla gereksizdir. Özellik sayısının fazla olması oluşturulacak tahmin modelinin parametre sayısını arttırmakta ve modelin yorumlanmasını güçleştirmektedir. Boyut indirgeme yöntemleri veri kümesinin istatistiksel analizler için uygun hale getirilmesi ve veri kümesinin yorumlanmasını daha basit hale getirebilmek için uygulanan yöntemlerden oluşmaktadır. Boyut indirgemede temel amaç orijinal veri kümesinin içerdiği bilgiden en az kayıp ile veri kümesini daha düşük boyutta temsil etmektir. Veri kümesinin özünü yakalayan daha düşük boyutlu bir alt uzaya yansıtarak boyutsallığı azaltmak genellikle yararlıdır (Murphy, 2012).

Boyut indirgeme yöntemleri özellik çıkarma ve özellik seçim olmak üzere iki alt sınıfta incelenebilir. Özellik seçim yöntemleri boyut indirgeme işlemini, veri kümesinde en önemli özellikleri belirleyerek daha az önemli olan özellikleri veri kümesinden ayıklayarak gerçekleştirir. Özellik çıkarım yöntemleri ise özelliklerin birleşim ile oluşturulan yeni değişkenler içerisinden en az bilgi kaybı ile daha az sayıda değişken kullanarak boyut indirgeme işlemini gerçekleştirir. Özellik seçim ve özellik çıkarma yöntemlerinin boyut indirgeme işlemini nasıl gerçekleştirdiğini gösteren grafik Şekil 3.1'de verilmiştir.



Şekil 3.1. Özellik seçim ve özellik çıkarma yöntemlerinin boyut indirgeme işlemini nasıl gerçekleştirdiklerini gösteren grafiksel gösterim (Cancele ve ark., 2020)

Özellik çıkarma, model doğruluğunun model yorumlanabilirliğinden daha önemli olduğu görüntü analizi, sinyal işleme ve bilgi alma gibi uygulamalarda tercih edilirken, özellik seçimi, metin madenciliği, genetik analiz ve sensör veri işleme gibi veri madenciliği uygulamalarında yaygın olarak kullanılır.

3.1. Özellik Seçim Yöntemleri

Özellik seçimi, tahmine dayalı bir model geliştirirken özellik sayısını azaltma işlemidir. Özellik seçimi, öncelikle bilgilendirici olmayan veya gereksiz tahmin unsurlarını modelden kaldırmaya odaklanır (Kuhn ve Johnson, 2013). Hem modelin hesaplanmasını basitleştirmek hem de bazı durumlarda modelin performansını iyileştirmek için özellik sayısının azaltılması arzu edilir. Özellik seçimi, veri kümesindeki en faydalı ve en önemli özellikleri seçerek veri kümesindeki özellik sayısını azaltmayı yani boyut indirgemeyi amaçlamaktadır.

Özellik seçimin yöntemleri sadece istatistiksel ölçütlere dayalı olan filtreleme yöntemleri, özellikler üzerinde arama işlemleri gerçekleştiren sarmal yöntemler ve en iyi bölen ölçütünü bulmaya dayalı olan gömülü yöntemler olmak üzere genel olarak üç grupta toplanmaktadır (Saeys ve ark., 2007).

İstatistik temelli filtreleme özellik seçim yöntemleri, istatistik ölçütler kullanılarak her bir özellikler ile hedef değişken arasındaki ilişkinin değerlendirilmesini ve hedef değişkenle en güçlü ilişkiye sahip olan özelliklerin seçilmesini içerir. Bu yöntemler hızlı ve etkili olabilir, ancak istatistiksel ölçülerin seçimi hem girdi hem de çıktı değişkenlerinin veri türüne bağlıdır. Bu nedenle, filtre tabanlı özellik seçimini gerçekleştirirken bir veri kümesi için uygun bir istatistiksel ölçümün seçilmesi zor olabilir. Bu çalışma kapsamında istatistiksel ölçütlere dayalı özellik seçim yöntemi olan filtreme yöntemlerinin sınıflandırma performansları incelenecektir

3.1.1. Değişim Katsayısı

X rassal değişkeninin ortalaması μ ve standart sapması σ olsun. Bu durumda X rassal değişkeni için değişim katsayısı

$$Dk = \frac{\sigma}{\mu} \tag{3.1}$$

eşitliği ile belirlenir. Değişim katsayısı ölçü birimi içermediği ve terim büyüklüklerinden etkilenmediği için iki kitlenin değişkenliğinin karşılaştırılmasında kullanılan istatistiksel bir ölçüdür. Değişkenliğin fazla olması veri seti içerisinde farklı özellikte birimlerin olabileceğinin yani kitle içerisinde alt kitlelerin olduğunun bir göstergesi olduğundan dolayı sınıflama için değişim katsayısı büyük olan değişkenlerin (özelliklerin) seçilmesi önerilir.

3.1.2. F Test İstatistiği

Bağımsız k grup ortalamasının eşitliğinin test edilmesinde kullanılan tek yönlü varyans analizindeki F test istatistiği

$$F = \frac{(n-k)\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_i\right)^2} \tag{3.2}$$

eşitliği ile hesaplanır. Eşitlikte yer alan x_{ij} gösterimi i. gruptaki j. birimin gözlem değerini ifade ederken $\bar{x}_i$ gösterimi ilgili özellik için i. grup ortalamasını ve $\bar{x}$ gösterimi ise genel ortalamayı ifade etmektedir. F test istatistiği gruplar arası kareler ortalamasının, gruplar içi kareler ortalamasına bölünmesi ile elde edilen istatistiksel bir ölçüttür. Birbirinden iyi ayırt edilebilen grup ya da diğer bir ifade ile kümeler için küme içi değişim az, kümeler arası değişim ise yüksektir. Buna göre F test istatistiğinin büyük değer alması ilgili özelliğin sınıflandırma performansının yüksek olduğunu gösterir.

3.1.3. Küme Merkezine Olan Uzaklık

Küme yapısı bilinen bir veri kümesinde, her bir özellik için küme merkezine olan uzaklıklara göre küme tahminleri gerçekleştirilerek özelliklerin doğru sınıflandırma performanslarına göre önem düzeyleri belirlenebilir. X özelliği için k kümeye ait küme merkezleri $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ olmak üzere i. birimin j. kümeye olan uzaklığı

$$d_{ij} = \left|x_i - \bar{x}_j\right| \tag{3.3}$$

eşitliği ile belirlenir. X özelliği için i. birimin ait olduğu küme, küme merkezlerine olan uzaklıklar arasında en küçük uzaklığa sahip olan küme olarak tahmin edilir. X özelliği için i. birimin ait olduğu küme g ise

$$g = arg\min_{j} d_{ij} \tag{3.4}$$

şeklinde gösterilir. Tüm birimler için küme tahminleri gerçekleştirildikten sonra küme üyeliği doğru olarak tahmin edilen birim sayısının genel birim sayısına oranı ile doğru sınıflandırma olasılığı hesaplanır. Küme merkezine olan uzaklığa göre belirlenen doğru sınıflandırma olasılığı ne kadar büyük ise ilgili özellik sınıflandırmada o kadar etkilidir. Dolayısıyla doğru önemli özellikler olarak sınıflandırma olasılığı yüksek olan özellikler seçilir.

3.1.4. Fisher Skoru

İki sınıf durumunda Fisher'in doğrusal diskriminant analizine dayalı Fisher skoru kullanılarak her bir özelliğin sınıflamadaki önemliliği belirlenebilir. İki sınıf için sınıf ortalamaları $\bar{x}_i^+$ ve $\bar{x}_i^-$, sınıflar için standart sapmalar s_i^+ ve s_i^- olmak üzere i. özellik için Fisher skoru

$$Fisher(x_i) = \frac{|\bar{x}_i^+ - \bar{x}_i^-|}{|s_i^+ + s_i^-|} \tag{3.5}$$

eşitliği ile belirlenir. Fisher skorunun büyük değer alması iki sınıfın biririnden iyi ayırt edilebildiğini göstermektedir. Bu nedenle veri kümesi içerisindeki önemli özellikler belirlenirken Fisher skoru yüksek olan özellikler tercih edilmektedir (Pai ve ark., 2014).

Bu çalışmada Fisher skoru k sınıf için, k adet kukla sınıf etiketi ile elde edilen k adet Fisher skorunun ortalamasını kullanılarak genelleştirilmiştir.

3.1.5. t Skoru

İki sınıflı veri kümelerinde özellik seçimi için kitle varyanslarının eşit olduğu varsayımı altında bağımsız iki grup ortalamasının eşitliğinin test edilmesinde kullanılan t test istatistiği kullanılabilir. Bu yaklaşımda t test istatistiğinin büyük değer alması iki sınıfın birbirinden iyi ayırt edilebildiğini ifade etmektedir. İki sınıf için sınıf ortalamaları $\bar{x}_i^+$ ve $\bar{x}_i^-$, sınıflar için standart sapmalar s_i^+ ve s_i^- olmak üzere i. özellik için t skoru

$$t(x_i) = \frac{|\bar{x}_i^+ - \bar{x}_i^-|}{\sqrt{\frac{(n^+ - 1)(s_i^+)^2 + (n^- - 1)(s_i^-)^2}{n - 2}}} \tag{3.6}$$

eşitliği ile belirlenir. Eşitlikte yer alan n^+ ve n^- sınıflardaki birim sayılarını göstermektedir. Bu çalışmada t skoru k sınıf için, k adet kukla sınıf etiketi ile elde edilen k adet t skorunun ortalamasını kullanılarak genelleştirilmiştir.

3.1.6. Welch 'in t İstatistiği

İki sınıflı veri kümelerinde özellik seçimi için kitle varyanslarının farklı olması durumunda bağımsız iki kitle ortalamasının eşitliğinin test edilmesinde kullanılan Welch'in t-istatistiği kullanılabilir. Welch'in t istatistiği

$$welch(x_i) = \frac{|\bar{x}_i^+ - \bar{x}_i^-|}{\sqrt{\frac{(s_i^+)^2}{n^+} + \frac{(s_i^-)^2}{n^-}}} \tag{3.7}$$

eşitliği ile hesaplanır. Bu yaklaşımda Welch'in t test istatistiğinin büyük değer alması iki sınıfın birbirinden iyi ayırt edilebildiğini ifade etmektedir. Dolayısıyla özellik seçimi, en yüksek skora sahip özelliklerin seçilmesi şeklinde yapılmaktadır. Bu çalışmada Welch'in t istatistiği k sınıf için, k adet kukla sınıf etiketi ile elde edilen k adet Welch'in t istatistiğinin ortalamasını kullanılarak genelleştirilmiştir.

3.1.7. Komşuluk Bileşen Analizi

Komşuluk bileşen analizi, sınıflama algoritmalarının doğru sınıflandırma olasılığını en büyüklemeyi amaçlayan parametrik olmayan özellik seçim algoritmasıdır. Komşuluk bileşen analizinde doğru sınıflandırma olasılığını en büyükleyecek şekilde özelliklerin ağırlıklandırılması gerçekleştirilir. Elde edilen ağırlıklarla özellik seçimi gerçekleştirilir.

$T = \{(\mathbf{x}_1, \mathrm{y}_1), \ldots\ldots(\mathbf{x}_\mathrm{i}, \mathrm{y}_\mathrm{i}), \ldots\ldots(\mathbf{x}_\mathrm{n}, \mathrm{y}_n)\}$ bir eğitim verisi olsun, burada $\mathbf{x}_\mathrm{i}$ p boyutlu bir değişken vektörü, $\mathrm{y}_\mathrm{i} \in \{1, 2, \ldots, \mathrm{C}\}$ ise sınıf etiketidir. Komşuluk bileşen analizinde amaç en yakın komşu sınıflandırma algoritmasının doğru sınıflandırma olasılığını optimize eden özellik alt kümesini seçecek bir ağırlık vektörü $\mathbf{w}$ bulmaktır. Ağırlık vektörü $\mathbf{w}$ olmak üzere $\mathbf{x}_\mathrm{i}$ ve $\mathbf{x}_\mathrm{j}$ birimleri arasındaki ağırlıklı uzaklık

$$d_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_\mathrm{j}) = \sum_{\mathrm{l}=1}^{\mathrm{d}} \mathrm{w}_\mathrm{l}^2 \left|\mathrm{x}_\mathrm{il} - \mathrm{x}_\mathrm{jl}\right| \tag{3.8}$$

eşitliği ile ifade edilir. Eşitlikte yer alan w_l gösterimi, l. özellikle ilişkili ağırlıktır. En yakın komşu sınıf algoritmasının başarılı olması için, sezgisel ve etkili bir strateji, eğitim verisi T'de doğru sınıflandırma olasılığını en üst düzeye çıkarmaktır. Komşuluk bileşen analizinde referans noktası bir olasılık ile belirlenir. Burada $\mathbf{x}_\mathrm{i}$ gözlem vektörünün referans noktası olarak $\mathbf{x}_\mathrm{j}$ gözlem vektörünü seçmesi olasılığı

$$p_{ij} = \begin{cases} \dfrac{k(d_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_j))}{\sum_{k \neq i} K(d_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_\mathrm{k}))}, & i \neq j \\ 0, & i = j \end{cases} \tag{3.9}$$

olarak tanımlanır. Bir $k(.)$ gösterimi $d_{\mathbf{w}}(\mathbf{x}_i, \mathbf{x}_j)$ ağırlı uzaklık fonksiyonunun büyük değerleri için küçük değerler veren kernel ve benzeri bir fonksiyondur. Yang ve ark. (2012) tarafından $k(z) = e^{-\frac{z}{\sigma}}$ şeklinde önerilmiştir. Eşitlik (3.9)'da tanımlanan olasılıklara dayalı olarak $\mathbf{x}_\mathrm{i}$ gözlemim vektörünün doğru sınıflandırılma olasılığı

$$p_i = \sum\nolimits_j \mathrm{y}_\mathrm{ij} \mathrm{p}_\mathrm{ij} \tag{3.10}$$

şeklinde tanımlanır. Eşitlikte yer alan y_{ij} katsayısı $\mathbf{x}_\mathrm{i}$ gözlem vektörü ile $\mathbf{x}_\mathrm{j}$ gözlem vektörü aynı sınıfta ise yani $\mathrm{y_i} = \mathrm{y_j}$ ise 1, aksi takdirde 0 değerini alır. Genel doğru sınıflandırma olasılığı

$$F(\mathbf{w}) = \frac{1}{n} \sum\nolimits_{i=1}^{n} p_i = \frac{1}{n} \sum\nolimits_{i=1}^{n} \sum\nolimits_j \mathrm{y}_\mathrm{ij} \mathrm{p}_\mathrm{ij} \tag{3.11}$$

eşitliği ile ifade edilir. Özellik seçimi yapmak ve aşırı uyumu azaltmak için Eşitlik (3.11)'e bir ceza terimi ekleyerek ilgili fonksiyon

$$\begin{aligned} F(\mathbf{w}) &= \frac{1}{n} \sum\nolimits_{i=1}^{n} p_i - \lambda \sum\nolimits_{l=1}^{p} w_l^2 = \frac{1}{n} \sum\nolimits_{i=1}^{n} \underbrace{\left[\sum\nolimits_j \mathrm{y}_\mathrm{ij} \mathrm{p}_\mathrm{ij} - \lambda \sum\nolimits_{l=1}^{p} w_l^2\right]}_{F_i(\mathbf{w})} \\ &= \frac{1}{n} \sum\nolimits_{i=1}^{n} F_i(\mathbf{w}) \end{aligned} \tag{3.12}$$

şeklinde düzenlenebilir. Burada λ düzeltme parametresidir. Komşuluk bileşen analizinde eşitlik (3.12) ile verilen amaç fonksiyonunu en büyükleyecek şekilde **w** ağırlık vektörü tahmin edilir.

$$\widehat{\mathbf{w}} = arg\max_{\mathbf{w}}\left(\frac{1}{n}\sum\nolimits_{i=1}^{n} F_i(\mathbf{w})\right) \tag{3.13}$$

3.1.8. Relief ve ReliefF Algoritmaları

Relief algoritması, iki sınıflı veri kümelerinde özellik seçimi için Kira ve Rendell (1992) tarafından önerilmiştir. Relief algoritması özelliklerin, birimlerin ait olduğu sınıftaki en yakın komşuları ile ait olmadığı sınıftaki en yakın komşuları arasındaki uzaklık farklılıklarına dayalı olarak ağırlıklandırılması esasına dayanmaktadır.

Kira ve Rendell (1992) tarafından önerilen orijinal Relief algoritmasında tüm birimler için ait oldukları sınıftaki en yakın komşu uzaklığı $\Delta(+) = (x_i - x^+)^2$ ve ait olmadıkları sınıftaki en yakın komşu uzaklığı $\Delta(-) = (x_i - x^-)^2$ olmak üzere ilgili özelliğin ağırlığı

$$W(x) = \frac{\sum_{i=1}^{n}\{\Delta(-) - \Delta(+)\}}{n} \tag{3.14}$$

eşitliği ile hesaplanır. Eşitliklerde yer alan x^+ gösterimi x_i gözlem değerinin aynı sınıfındaki en yakın komşu gözlemini, x^- gösterimi isex_i gözlem değerinin ait olmadığı sınıfdaki en yakın komşu gözlemini temsil etmektedir. Aynı sınıfta bulunan komşular arasındaki uzaklığın küçük, farklı sınıfta bulunan komşular arasındaki uzaklığın büyük olması beklenir. Buna göre Relief algoritmasındaki ağırlık değerinin büyük olması ilgili özelliğin sınıflama performansının yüksek olduğunu gösterir. Algoritmanın son adımında ilgili özellikler içerisinden belirlenen eşik değerini aşan özellikler seçilerek boyut azalta işlemi gerçekleştirilir.

İki sınıf problemi için önerilen Relief algoritması Kononenko (1994) tarafından k sınıf için genelleştirilmiştir. Kononenko (1994) tarafından önerilen algoritma ReliefF

olarak isimlendirilmiştir. ReliefF algoritmasında en yakın m komşu üzerinden ağırlıklar hesaplanmıştır. Aynı sınıfta bulunan en yakın komşu uzaklıkların ağırlığa katkısı

$$W(x)^r = W(x)^{r-1} - \frac{\Delta(x_i, x^+)}{m} d_{i,+} \tag{3.15}$$

eşitliği ile oluşturulur. Eşitlikte yer alan $\Delta(x_i, x^+)$ gösterimi x_i birimi ile aynı sınıfta yer alan m gözlem noktasından birini temsil eden x^+ gözlemi arasındaki uzaklığı ifade eder ve sayısal veriler için bu uzaklık

$$\Delta(x_i, x^+) = \frac{|x_i - x^+|}{\max(x) - \min(x)} \tag{3.16}$$

eşitliği ile hesaplanır. Eşitlik (8)'de yer alan $d_{i,+}$ gösterimi ise en yakın komşunun yakınlık derecesine göre ağırlığını ifade eder ve

$$d_{i,+} = \frac{\tilde{d}_{i,+}}{\sum_{l=1}^{m} \tilde{d}_{i,l}} \tag{3.17}$$

eşitliği ile hesaplanır. Eşitlikte yer alan $\tilde{d}_{i,l} = e^{-(rank(i,l)/sigma)^2}$ şeklinde hesaplanır. Eşitlikte yer alan $sigma$ değeri araştırmacı tarafından belirlenen pozitif bir sayıdan oluşan ölçeklendirme değeridir.

Farklı sınıfta bulunan en yakın komşu uzaklıkların ağırlığa katkısı

$$W(x)^r = W(x)^{r-1} + \frac{p_+}{1 - p_-} \frac{\Delta(x_i, x^-)}{m} d_{i,-} \tag{3.18}$$

eşitliği ile hesaplanır. Eşitlikte yer alan p_+ gösterimi x_i biriminin ait olduğu sınıfa ait önsel olasılığını gösterirken, p_- gösterimi x^- biriminin ait olduğu sınıfa ait önsel olasılığı göstermektedir. İlgili önsel olasılıklar $p_+ = \frac{n_+}{n}$ ve $p_- = \frac{n_-}{n}$ eşitlikleri ile elde edilir. ReliefF algoritmasında da ağırlık değerinin büyük olması ilgili özelliğin sınıflama performansının yüksek olduğunu göstermektedir.

3.1.9. Önerilen Özellik Seçim Yöntemleri

Sınıf üyeliklerinin bilindiği veri setlerinde sınıf bilgisi kullanılarak değişim katsayısının özellik seçiminde farklı bir kullanımı söz konusu olabilir. X rassal değişkeninin $\mu_1, \mu_{2,} \dots, \mu_k$ ortalamaları ve $\sigma_1, \sigma_{2,} \dots, \sigma_k$ standart sapmalarına sahip k tane alt kitleye sahip olduğu varsayılsın. Bu durumda i. alt kitleye ilişkin değişim katsayısı

$$Dk_i = \frac{\sigma_i}{\mu_i} \qquad i = 1, 2, \cdots, k \tag{3.19}$$

eşitliği ile hesaplanır. Alt kitleler içerisinde en büyük ve en küçük değişim katsayıları

$$Dk_{max} = max(Dk_1, Dk_2, \cdots, Dk_k) \tag{3.20}$$

$$\mathrm{Dk_{min}} = \min(\mathrm{Dk_1}, \mathrm{Dk_2}, \cdots, \mathrm{Dk_k}) \tag{3.21}$$

ve X rassal değişkeni için genel değişim katsayısı Dk_{Genel} olmak üzere aşağıda tanımlanan değişim katsayısı oranları

$$DkO_{Enk} = \frac{Dk_{Genel}}{Dk_{max}} \tag{3.22}$$

$$DkO_{Enb} = \frac{Dk_{Genel}}{Dk_{min}} \tag{3.23}$$

özellik seçiminde kullanılabilir. Yeni tanımlanan her iki kriter içinde kriterin büyük değer alması ilgili özelliğin sınıflamada daha etkili olabileceği anlamına gelir.

3.2. Özellik Çıkarma Yöntemleri

Çok değişkenli istatistiksel analizlerin temel amaçlarından biri de verileri önemli bilgileri kaybetmeden orijinal boyut sayısından daha az boyutta özetlemektir. Bir asırdan daha uzun bir süre önce Pearson (1901) ve Hotelling (1933) bu sorunu ele aldı ve değişkenlerle tek tek ilgilenmek yerine değişkenlerin lineer birleşimleri ile ilgilendiler. Başlangıçta tüm değişkenlerin ortalaması şeklinde bir özetleme düşünebilir ancak burada cevaplanması gereken iki temel soru vardır.

i. En uygun birleşim hangisidir?

ii. Kaç tane birleşim ile çalışılmalıdır?

Hotelling (1933) ilk sorunun cevabı olarak verilerin değişkenliğini en iyi şekilde açıklayacak lineer birleşimleri oluşturmayı önerdi. Orijinal değişkenlerin doğrusal birleşimlerinin oluşturulması ve yorumlanması nispeten daha kolaydır ve önemli matematiksel özelliklere sahiptir.

İkinci soru farklı niteliktedir ve veriye bağlı olarak farklı cevaplara sahiptir. İkinci sorunun cevabında temel yaklaşım verimliliktir. Kullanılan doğrusal birleşim sayısı arttıkça orijinal veriye o kadar yakın olunur ancak boyut indirgeme, basitleştirme amacından da o kadar uzaklaşılır ve hesaplama maliyetleri artar. Burada temel hedef basitleştirme, boyut indirgeme amacı gözönünde bulundurularak verideki önemli bilgileri kaybetmeden her iki amacı dengeleyecek bir çözüm oluşturmaktır. Bazen bu çözüm tek bir birleşim olabilirken bazen de daha fazla sayıda birleşim kullanılması şeklinde olabilir.

3.2.1. Temel Bileşenler Analizi

Temel bileşenler analizi muhtemelen çok değişkenli istatistiksel analizler içerisinde en eski ve en çok bilinen analizdir. İlk olarak Pearson (1901) tarafından önerilmiş ve Pearson'nın çalışmalarından bağımsız olarak Hotelling (1933) tarafından geliştirilmiştir. Temel bileşenler analizinde temel fikir, çok sayıda birbiri ile ilişkili değişkenlerden oluşan bir veri setini, veri setindeki değişimi mümkün olduğunca koruyarak, birbiri ile ilişkisiz daha az sayıda yeni değişkenle yani boyut indirgeyerek açıklamaktır. Boyut indirgeme; temel bileşenler olarak isimlendirilen birbiriyle ilişkisiz, orijinal veri setindeki değişimi açıklamadaki önemine göre büyükten küçüğe sıralanmış,

orijinal değişkenlerin lineer birleşiminden oluşan yeni değişkenlerden birkaçının kullanılmasıyla sağlanır. Temel bileşenlerin hesaplanması pozitif yarı tanımlı simetrik bir matrisin öz değer-öz vektör probleminin çözümü ile ilgilidir. Bu nedenle temel bileşenlerin tanımı ve hesaplanması basittir. Ancak görünüşte basit teknik, çok çeşitli farklı uygulamaların yanı sıra çok sayıda farklı türevlere sahiptir.

Temel bileşenler analizi birbiri ile ilişkili çok sayıda değişkenden oluşan bir veri setinin varyans-kovaryans yapısını önemli bir bilgi kaybetmeden, birbiriyle ilişkisiz, daha az sayıda orijinal değişkenlerin lineer birleşimi ile açıklayan çok değişkenli istatistiksel bir analizdir. Temel bileşenler analizinde, orijinal veri setinde çok sayıdaki değişken tarafından açıklanan değişim birkaç temel bileşenle açıklanmaya çalışılmaktadır. Uygulamada temel bileşenler orijinal veri setinin varyans-kovaryans matrisi veya korelasyon matrisi esas alınarak elde edilmektedir. Eğer veri setini oluşturan değişkenlerin terim büyüklükleri ve ölçekleri farklı ise analiz öncesi standartlaştırma işlemi uygulanmaktadır yani temel bileşenler korelasyon matrisi esas alınarak hesaplanmaktadır. Standartlaştırma işlemi, terim büyüklükleri yüksek olan değişkenlerin diğer değişkenler üzerindeki olumsuz etkisini yok etmek için gerçekleştirilir.

3.2.1.1. Temel Bileşenlerin Elde Edilmesi

Temel bileşenler analizinde $X_1, X_2, \ldots, X_p$ rassal değişkenlerden oluşan $\mathbf{x}$ rassal vektörünün varyans-kovaryans matrisi $\mathbf{\Sigma}$ veya korelasyon matrisi $\boldsymbol{\rho}$ kullanılır. Daha öncede ifade edildiği gibi değişkenlerin ölçü birimlerinin veya terim büyüklüklerinin farklı olması durumunda korelasyon matrisinin kullanılması önerilmektedir. Korelasyon matrisinin kullanılması aslında temel bileşenlerin hesaplamasında standartlaştırılmış rassal vektör $\mathbf{z}$ 'nin temel alındığı anlamına gelmektedir.

Temel bileşenlerin hesaplanmasında $\mathbf{x}$ rassal vektörünün varyans-kovaryans matrisi $\mathbf{\Sigma}$ 'nın kullanıldığını varsayalım. Bu durumda $p \times p$ boyutlu birim matris $\mathbf{I}$ olmak üzere

$$|\mathbf{\Sigma} - \lambda \mathbf{I}| = 0 \tag{3.24}$$

eşitliği ile elde edilen $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ özdeğerleri ve bu özdeğerlere karşılık gelen ve

$$(\boldsymbol{\Sigma} - \lambda_k \mathbf{I})\mathbf{a}_k = 0 \qquad k = 1, 2, \dots, p \tag{3.25}$$

eşitliği ile hesaplanan öz vektörler kullanılarak temel bileşenler elde edilir. $X_1, X_2, \dots, X_p$ rassal değişkenlerin ortalamaları $\mu_1, \mu_2 \dots, \mu_p$ olmak üzere temel bileşenler

$$\begin{aligned}
T_1 &= a_{11}(X_1 - \mu_1) + a_{12}(X_2 - \mu_2) + \cdots + a_{1p}(X_p - \mu_p) \\
T_2 &= a_{21}(X_1 - \mu_1) + a_{22}(X_2 - \mu_2) + \cdots + a_{2p}(X_p - \mu_p) \\
&\vdots \\
T_k &= a_{k1}(X_1 - \mu_1) + a_{k2}(X_2 - \mu_2) + \cdots + a_{kp}(X_p - \mu_p) \\
&\vdots \\
T_p &= a_{p1}(X_1 - \mu_1) + a_{p2}(X_2 - \mu_2) + \cdots + a_{pp}(X_p - \mu_p)
\end{aligned} \tag{3.26}$$

eşitlikleri ile elde edilir. Eşitlik (3.26)'dan anlaşılabileceği gibi temel bileşenler orijinal değişkenlerin lineer bir dönüşümü şeklinde elde edilmektedir. Varyans-kovaryans matrisinden elde edilen öz değerlere karşılık gelen öz vektörlerden oluşan katsayılar matrisi **A** olmak üzere temel bileşenler matris notasyonu ile

$$\mathbf{T} = (\mathbf{X} - \boldsymbol{\mu})\mathbf{A} \tag{3.27}$$

şeklinde gösterilir. Burada katsayılar matrisi $\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{p1} \\ \vdots & \ddots & \vdots \\ a_{1p} & \cdots & a_{pp} \end{bmatrix}$ olarak tanımlanmıştır.

Uygulamalarda genellikle örneklem ile çalışıldığından **Σ** varyans-kovaryans matrisinin tahmini örneklem varyans kovaryans matrisi **S** kullanılır. Bu durumda temel bileşenler benzer şekilde örneklem varyans kovaryans matrisi **S** 'nin öz değer ve öz vektörleri ile oluşturulur. Standartlaştırılmış rassal vektör **z** esas alınarak gerçekleştirilen temel bileşenler analizinde, temel bileşenler korelasyon matrisi **ρ** 'nun öz değerlerine karşılık gelen öz vektörleri ile oluşturulan katsayılar matrisi **A** olmak üzere

$$\mathbf{T} = \mathbf{ZA} \tag{3.28}$$

eşitliği ile hesaplanır.

3.2.1.2. Temel Bileşenlerin Özellikleri

x rassal vektörünün ortalama vektörü $\boldsymbol{\mu}$ ve varyans-kovaryans matrisi **Σ** olmak üzere eşitlik (4) ile hesaplanan temel bileşenler aşağıdaki özelliklere sahiptir (Mardia ve ark., 1979).

a. $E(T_k) = 0$

Temel bileşenlerin elde edilmesinde merkezileştirilmiş veriler kullanıldığında hesaplanan bileşenlerin ortalamaları sıfırdır. Bu özelliğin ispatında beklenen değer işleminin özellikleri kullanılır.

$$E(T_k) = E\{(\mathbf{x} - \boldsymbol{\mu})a_k\} = E(\mathbf{x} - \boldsymbol{\mu})a_k = 0$$

b. $Var(T_k) = \lambda_k$

Temel bileşenlerin varyansı, bileşenin hesaplanmasında kullanılan öz vektöre karşılık gelen öz değere eşittir. Bu özelliğin ispatında varyans işleminin özellikleri ve spektral ayrışım sonucu elde edilen öz değer ve öz vektör özellikleri

$$a_{k1}^2 + a_{k2}^2 + \cdots + a_{kp}^2 = 1, \qquad k = 1, 2, \dots, p$$

$$a_{k1}a_{l1} + a_{k2}a_{l2} + \cdots + a_{kp}a_{lp} = 0, \qquad k \neq l$$

$$\boldsymbol{\Sigma} = \mathbf{A}\boldsymbol{\Gamma}\boldsymbol{A}'$$

kullanılır. Burada **Γ** özdeğerlerden oluşan diagonal matrisi $\boldsymbol{\Gamma} = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{bmatrix}$ göstermektedir.

$$Var(T_k) = Var\{(\mathbf{x} - \boldsymbol{\mu})a_k\} = a_k' Var(\mathbf{x} - \boldsymbol{\mu})a_k$$

$$Var(T_k) = a_k' \boldsymbol{\Sigma} a_k = a_k' \mathbf{A}\boldsymbol{\Gamma}\boldsymbol{A}' a_k = \lambda_k$$

c. $Cov(T_k, T_l) = 0 \qquad k \neq l$

Temel bileşenler aralarında ilişkisizdir. Bu özelliğin ispatında kovaryans işleminin özellikleri ve spektral ayrışım sonucu elde edilen öz değer ve öz vektör özellikleri kullanılır.

$$Cov(T_k, T_l) = Cov\{(\mathbf{x} - \boldsymbol{\mu})a_k, (\mathbf{x} - \boldsymbol{\mu})a_l\} = a_k' Var(\mathbf{x} - \boldsymbol{\mu})a_l$$

$$Cov(T_k, T_l) = a_k' \boldsymbol{\Sigma} a_l = 0$$

d. $Var(T_1) > Var(T_2) > \cdots > Var(T_p)$

İlk temel bileşen en büyük öz değere karşılık gelen öz vektör kullanılarak elde edildiğinden en büyük varyansa sahiptir. Temel bileşenler $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ sıralamasına göre oluşturulduğundan dolayı bileşenlerin varyansı giderek küçülür. Temel bileşenler analizinde boyut indirgeme, çok küçük varyansa sahip bileşenlerin göz ardı edilmesi ile gerçekleştirilir.

e. $\sum_{i=1}^{p} \lambda_i = trace(\boldsymbol{\Sigma}) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2$

Çok değişkenli istatistiksel analizlerde varyans-kovaryans matrisinin esas köşegen üzerindeki elemanlarının yani varyansların toplamı, toplam varyans olarak isimlendirilir. Temel bileşenlerin varyanslarının toplamı başka bir ifade ile öz değerlerin toplamı orijinal değişkenlerin varyanslarının toplamına eşittir. Temel bileşenlerin hesaplanmasında korelasyon matrisinin kullanılması durumunda öz değerlerin toplamı yani bileşen varyanslarının toplamı değişken sayısı p değerine eşit olur. Tanımlanan bu özellik matris teorisinin temel özelliklerinden biridir. Bilindiği gibi $n \times n$ boyutlu $\mathbf{D}$ matrisinin özdeğerleri $\lambda_1, \lambda_2, \ldots, \lambda_n$olmak üzere $trace(\mathbf{D}) = \sum_{i=1}^{n} \lambda_i$ olur (Rencher, 2007).

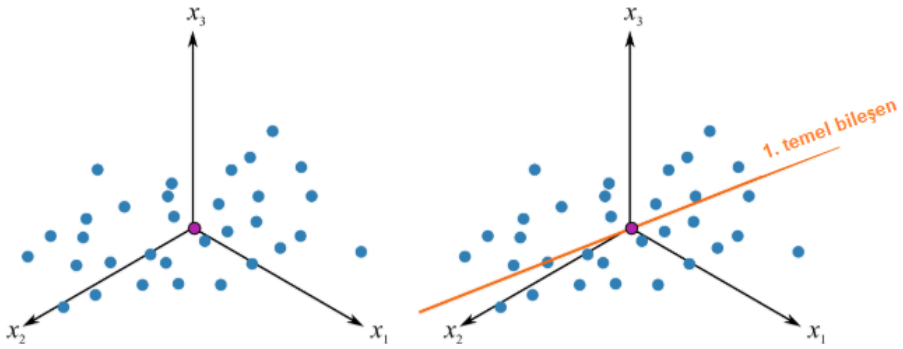
f. $\prod_{i=1}^{p} \lambda_i = |\boldsymbol{\Sigma}|$

Çok değişkenli istatistiksel analizlerde varyans-kovaryans matrisin,n determinantı genelleştirilmiş varyans olarak isimlendirilir. Buna göre orijinal değişkenlerin genelleştirilmiş varyansı ile temel bileşenlerin genelleştirilmiş varyansı birbirine eşittir. Tanımlanan bu özellik matris teorisinin temel özelliklerinden biridir. Bilindiği gibi $n \times n$ boyutlu $\mathbf{D}$ matrisinin özdeğerleri $\lambda_1, \lambda_2, \ldots, \lambda_n$olmak üzere $|\mathbf{D}| = \prod_{i=1}^{n} \lambda_i$ olur (Rencher, 2007). Temel bileşenler

için b ve c maddelerinde belirtilen özelliklere göre temel bileşenler rassal vektörünün varyans-kovaryans matrisi $\mathbf{\Gamma} = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{bmatrix}$ şeklinde tanımlıdır.

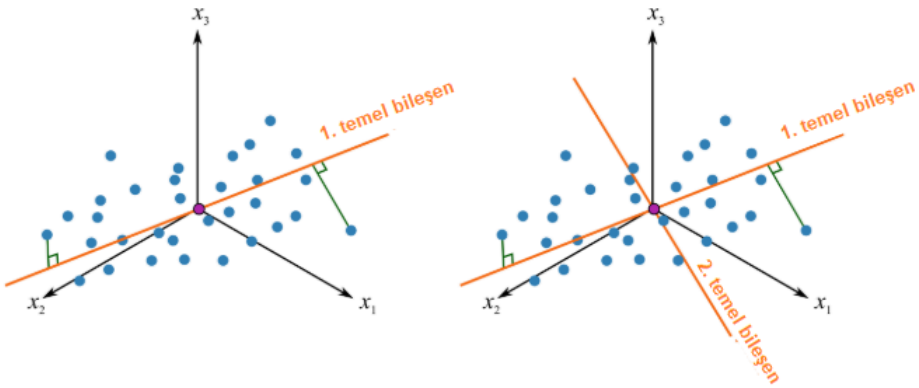
3.2.1.3. Temel Bileşenler Analizinin Geometrik Yorumu

Temel bileşenler analizinin geometrik yorumunda merkezileştirilmiş X_1, X_2, X_3 rassal değişkenlerini göz önünde bulunduralım. İlk temel bileşen en fazla değişimin olduğu yön boyunca veri noktalarının doğruya 90^o 'lik izdüşümlerinin kareler toplamını en küçükleyecek şekilde oluşturulur. Orijinal değişkenler arasındaki ilişki ne kadar yüksek ise oluşturulan doğrunun veri setini temsil etme yeteneği o kadar fazladır.



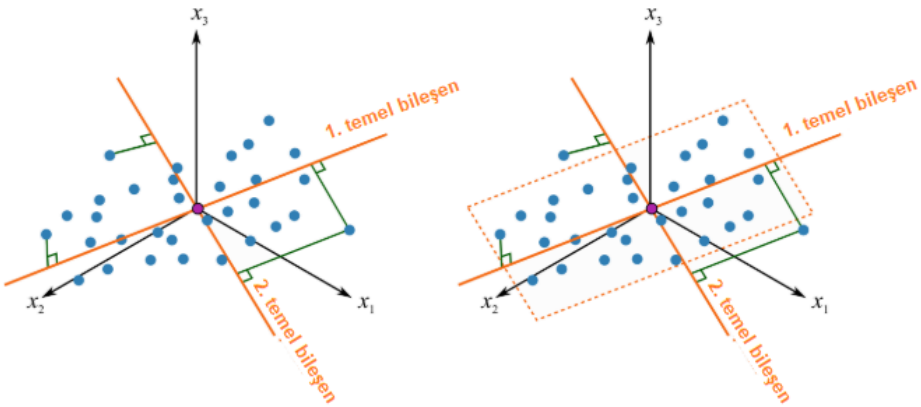
Şekil 3.2 İlk temel bileşenin oluşturulması (Dunn, 2020)

Veri noktalarının doğruya 90^o 'lik izdüşümlerinin gerçekleştiği nokta ile orjin arasındaki uzaklıklar, birinci temel bileşene ait skorları oluşturur. İkinci temel bileşen, ilk temel bileşenin yönüne dik olacak şekilde oluşturulur.



Şekil 3.3. İlk temel bileşen skorları ve ikinci temel bileşenin oluşturulması (Dunn, 2020)

İkinci temel bileşen skorları da veri noktalarının ikinci doğruya 90^o 'lik izdüşümlerinin gerçekleştiği nokta ile orjin arasındaki uzaklıklardan oluşur. Oluşturulan ikinci temel bileşendeki yayılım, en büyük değişkenliğe sahip ilk temel bileşenden daha azdır. Birbirine dik olan ilk iki temel bileşen ile veri noktaları yeni bir düzlemde tanımlanmıştır. Eğer üç veya daha fazla bileşen kullanılırsa veri noktaları oluşturulan hiper düzlemde tanımlanmış olur.



Şekil 3.4. İki temel bileşen ile tanımlanan düzlem (Dunn, 2020)

3.2.2. Çok Boyutlu Ölçekleme Analizi

Çok boyutlu ölçekleme analizi, n biriminin p değişken bakımından ölçülmesi sonucu oluşan çok boyutlu verinin birimler arasındaki ikili benzerlikleri koruyacak şekilde grafiksel olarak gösterilmesini hedefleyen istatistiksel bir analizdir. Çok boyutlu ölçekleme analizinde birimler arasındaki benzerlikler temel alınarak birimlerin grafiksel gösterimi sağlanabildiği gibi değişkenler arasındaki benzerliği temel alarak değişkenlerin grafiksel gösterimi de sağlanabilmektedir.

Çok boyutlu ölçekleme analizi, teknik olmayan bir bakış açısı ile benzer birimleri birbirlerine yakın, benzemeyen birimleri birbirlerine uzak olacak şekilde koordinat sistemine yerleştirerek çok boyutlu verinin grafiksel gösterimin sağlayan boyut indirgeme tekniği olarak da tanımlanabilir. Çok boyutlu verinin grafiksel gösterimini sağlamak için daha düşük boyutta benzerlikleri koruyacak şekilde boyut indirgeme işlem gerçekleştirilmektedir. Çok boyutlu ölçekleme analizi, birimler veya değişkenler arasındaki benzerlikleri esas aldığından dolayı çok boyutlu kategorik verilerin görselleştirilmesinde de yaygın olarak kullanılan istatistiksel bir araçtır.

Benzerlik kavramı literatürde benzememezlik olarak da tanımlanabilmektedir. Benzerliğin ölçülmesinde uzaklık türü veya ilişki türü benzerlik ölçüleri kullanılmaktadır. Benzerlik uzaklık türü ölçüler ile ifade edildiğinde iki birim arasındaki uzaklık ne kadar küçük ise iki birim o kadar benzerdir. Benzerlik ilişki türü ölçüler ile ifade edildiğinde ise iki birimin benzer olması için ilişkinin yüksek olması gerekir.

$\mathbf{X}$ $n \times p$ boyutlu bir veri matrisi olmak üzere $d: \mathbf{X} \times \mathbf{X} \rightarrow R$ şeklinde tanımlanan bir fonksiyon eğer tüm $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$ için

a) Pozitiflik: $d(\mathbf{x}_i, \mathbf{x}_j) \geq 0$,
b) Simetri: $d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i)$
c) Özdeşlik: $d(\mathbf{x}_i, \mathbf{x}_i) = 0$

koşullarını sağlıyorsa d uzaklık fonksiyonu olarak isimlendirilir. Belirtilen koşullar yanında eğer tüm $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k \in \mathbf{X}$ için

d) Belirlilik: $d(\mathbf{x}_i, \mathbf{x}_j) = 0$ sadece ve sadece $\mathbf{x}_i = \mathbf{x}_j$
e) Üçgen eşitsizliği: $d(\mathbf{x}_i, \mathbf{x}_j) \leq d(\mathbf{x}_i, \mathbf{x}_k) + d(\mathbf{x}_j, \mathbf{x}_k)$

şeklinde ifade edilen koşulları sağlıyorsa d uzaklık fonksiyonu metrik olarak isimlendirilir.

Nicel değişkenler ile çalışıldığında birimler arasındaki uzaklıkların belirlenmesinde en yaygın kullanılan uzaklık fonksiyonu Öklid uzaklık fonksiyonudur. İki birim arasındaki Öklid uzaklığı

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left\{ \sum_{l=1}^{p} (x_{il} - x_{jl})^2 \right\}^{\frac{1}{2}} \tag{3.28}$$

eşitliği ile hesaplanır.

Çok boyutlu ölçekleme analizinde çok boyutlu verinin grafiksel gösteriminin sağlanması için çok sayıda algoritma önerilmiştir. Algoritmalar benzerliğin belirlenmesinde kullanılan ölçüm türüne göre genel olarak metrik ve metrik olmayan algoritmalar olarak iki kategoride incelenmektedir. Bazı araştırmacılar ve kaynaklarda, hem metrik hem de metrik olamayan sınıflandırmada yer alan algoritmalar gözönünde bulundurularak bu iki kategoriye yarı metrik algoritmalar kategorisi eklemektedir. Çok boyutlu ölçekleme analizinde benzerlik metrik ölçüler ile belirleniyorsa verinin

grafiksel gösterimi için metrik algoritmalar kullanılmaktadır. Eğer çok boyutlu ölçekleme analizi nitel verilere uygulanıyorsa ya da benzerlik insan kararına dayanan bir sıralama ile belirleniyorsa metrik olmayan algoritmalara dayalı çok boyutlu ölçekleme analizi uygulanmaktadır.

Metrik çok boyutlu ölçekleme analizinde, $n \times p$ boyutlu orijinal veri matrisine dayalı olarak hesaplanan birimler arasındaki uzaklık değerleri ile indirgenmiş koordinat düzleminde ölçülen birimler arasındaki δ_{ij} uzaklıklar değerleri arasında

$$\delta_{ij} \approx f(d_{ij}) \tag{3.29}$$

şeklinde sürekli monoton bir fonksiyonel bir ilişki varsayılır (Cox ve Cox, 2001). Eşitlikte yer alan f fonksiyonunun farklı şekillerde tanımlanması ile farklı çok boyutlu ölçekleme modelleri tanımlanabilir. Bu modellerde d_{ij} orijinal uzaklıkları bağımsız değişken, indirgenmiş koordinat sistemindeki δ_{ij} uzaklıkları bağımlı değişken olarak ele alınarak iki uzaklık arasındaki fonksiyonel ilişki modellenmeye çalışılmaktadır. Bu modeller içerisinde en yaygın kullanılan ve oransal çok boyutlu ölçekleme modeli adı verilen model

$$f(d_{ij}) = bd_{ij} \tag{3.30}$$

eşitliği ile tanımlanır (Borg ve Groenen, 1997). Yaygın kullanılan diğer bir çok boyutlu ölçekleme modeli ise

$$f(d_{ij}) = a + bd_{ij} \tag{3.31}$$

eşitliği ile tanımlanan aralıklı çok boyutlu ölçekleme modelidir (Martinez ve ark., 2017). Bu modeller dışında yüksek dereceli polinom, üstel ve logaritmik fonksiyonlar ile tanımlanan çok boyutlu ölçekleme modelleri de kullanılmaktadır.

Metrik olmayan çok boyutlu ölçekleme algoritmalarında orijinal uzaklıklar arasındaki sıralamayı koruyacak şekilde

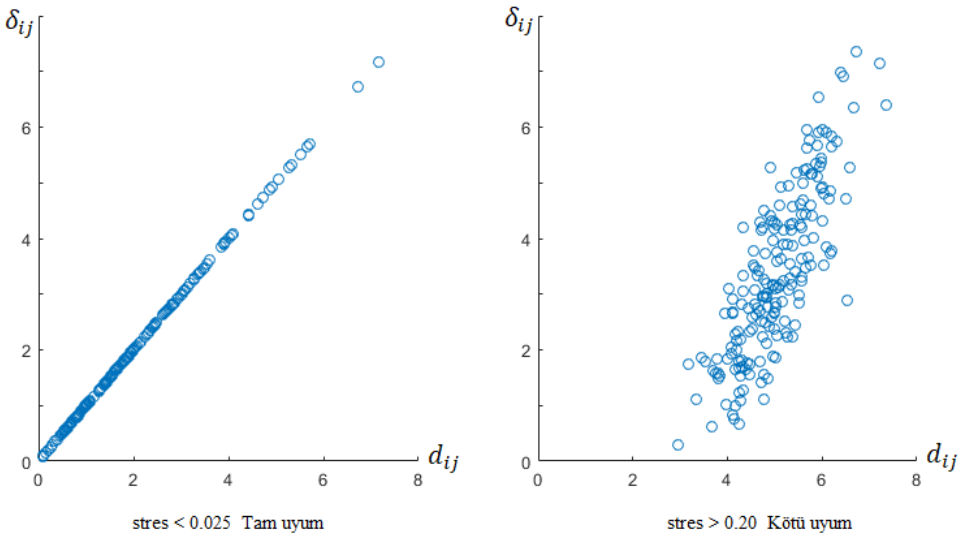
$$d_{rs} < d_{ab} \rightarrow f(d_{rs}) < f(d_{ab}) \tag{3.32}$$

boyut indirgeme sağlanmaktadır.

Çok boyutlu ölçekleme analizinde indirgenmiş koordinat sistemindeki grafiksel gösteriminin başarısı

$$stres = \left\{ \frac{\sum_i \sum_{j>i} \left(\delta_{ij} - f(d_{ij})\right)^2}{\sum_i \sum_{j>i} \delta_{ij}^2} \right\}^{\frac{1}{2}} \tag{3.33}$$

eşitliği ile tanımlanan stres fonksiyonu ile ölçülmektedir. Stres değeri ne kadar küçük ise orijinal uzaklıklar ile konfigürasyon uzaklıkları arasındaki uyum o kadar iyi olarak değerlendirilir. Uygulamada stres değerinin 0.025'den düşük olması tam uyum, stres değerinin 0.025 ile 0.05 arasında olması mükemmel uyum ve stres değerinin 0.10 ile 0.05 arasında olması iyi uyum olarak ifade edilmektedir. Stres değerinin 0.10 ile 0.20 arasında olması orta uyum olarak ifade edilirken, 0.20'den büyük stres değerleri kötü uyumu göstermektedir. Tam ve kötü uyum için örnek gösterimler Şekil 3.5'de verilmiştir.



Şekil 3.5. Orijinal uzaklıklar ile indirgenmiş koordinat sistemindeki uzaklıklar arasındaki uyum

3.2.2.1. Klasik Çok Boyutlu Ölçekleme Analizi

Metrik çok boyutlu ölçekleme yaklaşımı olan klasik çok boyutlu ölçekleme analizi Torgerson kk1952) tarından önerilmiştir. Temel bileşenler analizine olan benzerliğinden dolayı temel koordinatlar analizi olarak da isimlendirilen klasik çok boyutlu ölçekleme analizinde, **X** veri matrisindeki birimler arasındaki uzaklıklar Eşitlik (3.28)'de tanımlanan Öklid uzaklığı ile belirlenir.

Klasik çok boyutlu ölçekleme, Öklid uzaklığı ile oluşturulan **D** uzaklık matrisindeki her bir uzaklığının karesi alınarak oluşturulan karesel uzaklık matrisi $\mathbf{D}^{(2)}$ ve merkezileştirme matrisi **H** olmak üzere

$$\mathbf{B} = -\frac{1}{2}\mathbf{H}\mathbf{D}^{(2)}\mathbf{H} \tag{3.34}$$

eşitliği tanımlanan **B** matrisinin özdeğer ve özvektörlerini temel alan bir yaklaşımdır. Merkezileştirme matrisi **H**, birim matrisi **I** ve bir vektörü $\mathbf{1}$ olmak üzere $\mathbf{H} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}^{\boldsymbol{T}}$ eşitliği ile elde edilir. Klasik çok boyutlu ölçekleme analizinde indirgenmiş koordinat sistemi

$$Y = \mathbf{A}_d\mathbf{L}_d^{1/2} \tag{3.35}$$

eşitliği ile oluşturulur. Eşitlikte yer alan $\mathbf{A}_d$ gösterimi **B** matrisinin en büyük d özdeğerine karşılık gelen özvektörler matrisini, $\mathbf{L}_d^{1/2}$gösterimi ise **B** matrisinin en büyük d özdeğerinin karekök değerleri ile oluşturulan diagonal matrisi göstermektedir. Bazı veri setleri için **B** matrisi pozitif yarı tanımlı olmayabilir, bu durumda öz değerlerin bazıları negatif olmaktadır.

3.2.2.2. Karmaşık Fonksiyonların Optimizasyonuyla Çok Boyutlu Ölçekleme Analizi

Leeuw (1977) ve Groenen (1993) tarafından geliştirilen, hem metrik hem de metrik olmayan olarak uygulanabilen karmaşık fonksiyonların optimizasyonu ile çok boyutlu ölçekleme algoritması (SMACOF) uygulama basitliği ile yaygın bir kullanıma sahiptir. Borg ve Groenen (1997) metrik durum için önerdikleri optimizasyon yönteminin, seçili bir fonksiyonun en küçüklenmesi için gerekli koşulları sağladığını göstermişlerdir. Önerilen algoritmada indirgenmiş koordinat sisteminde sıra stres değeri olarak ifade edilen

$$\sigma(\mathbf{Y}) = \sum_{i<j} \omega_{ij}\left(\delta_{ij}(\mathbf{Y}) - d_{ij}\right)^2 \tag{3.36}$$

fonksiyon en küçüklenmeye çalışılır. Eşitlikte yer alan ω_{ij} kayıp gözlemler için 0, gözlemler için 1 değeri alan ağırlık katsayısıdır. $\delta_{ij}(\mathbf{Y})$ indirgenmiş koordinat

sisteminde birimler arasındaki uzaklıkların hesaplandığı uzaklık fonksiyonudur. İndirgenmiş koordinat sistemi **Y** başlangıçta rassal olarak seçilebileceği gibi belirli bir sistematiğe göre de oluşturulabilir. Algoritmanın $r.$ tekrarında kayıp gözlem olmadığı varsayımı altında indirgenmiş koordinat sistemi $\mathbf{Y}^r$

$$\mathbf{Y}^r = n^{-1}B(\mathbf{Y}^{r-1})\mathbf{Y}^{r-1} \tag{3.37}$$

Guttman dönüşümü ile güncellenir. Eşitlikte yer alan $B(\mathbf{Y}^{r-1})$ matrisinin elemanları

$$i \neq j \text{ için} \quad b_{ij} = \begin{cases} -\dfrac{d_{ij}}{\delta_{ij}(\mathbf{Y}^{r-1})}, & \delta_{ij}(\mathbf{Y}^{r-1}) \neq 0 \\ 0, & \delta_{ij}(\mathbf{Y}^{r-1}) = 0 \end{cases} \tag{3.38}$$

$$i = j \text{ için} \quad b_{ii} = \sum_{j=1, i\neq j}^{n} b_{ij}$$

şeklinde hesaplanır. Algoritma ardışık iki tekrarda elde edilen sıra stres değerleri arasındaki mutlak fark belirlenen kritik değere eşit veya küçük olunca sonlandırılır.

3.2.3. Yerel Doğrusal Eşleme (LLE)

Yerel doğrusal eşleme (Locally Linear Embedding, LLE) yöntemi manifold öğrenme temelli boyut indirgeme tekniğidir. Manifold öğrenme, doğrusal olmayan boyutsallığın azaltılması için yeni geliştirilen bir tekniktir. İlgilenilen verilerin daha yüksek boyutlu alan içindeki gömülü doğrusal olmayan bir manifoldda olduğu varsayılmaktadır. Manifold öğrenme algoritmaları, verilerin düşük boyutlu bir gösterimini bulmak için bu parametreleri ortaya çıkarmaya çalışır.

LLE algoritması ilk olarak Roweis ve Saul (2000) tarafından karmaşık yüksek boyutlu verilerin analiz için çok daha düşük boyutlu bir alana yansıtmanın bir yolu olarak tanımlanmıştır. Buradaki fikir, yüksek boyutlu uzaydan verilerin yerel, doğrusal modellerini oluşturmak ve daha sonra düşük boyutlu uzaya indirgeme sırasında yerel uzaklıkları korumaktır. LLE algoritması, aynı yerel olarak doğrusal ilişkilere sahip veri noktalarının yüksek ve düşük boyutlu gösterimleri arasındaki ilişkileri sağlamak için birincil bir yaklaşım sağlayan doğrusal olmayan manifold öğrenme algoritması olarak düşünülebilir. LLE algoritması çok yüksek boyutlu veriler için, özellikle yüz verileri gibi görüntüler için kullanılmıştır (Graff ve Wichmann, 2002).

LLE öz değer-öz vektör temelli bir yöntemdir. LLE yöntemi basit geometrik kavramlara dayanmaktadır. LLE algoritması, Öklid uzaklığına göre sadece en yakın komşularını kullanarak her bir veri noktasını yeniden yapılandıran doğrusal katsayıları bularak veri noktalarının düşük boyutta yerel geometrisini karakterize eder.

Yeniden ağırlıklandırmadaki hata

$$E(W) = \sum_i \left| \mathbf{x}_i - \sum_j w_{ij} \mathbf{x}_j \right|^2 \tag{3.39}$$

eşitliği ile ölçülür. Eşitlikte yer alan j indisi, $\mathbf{x}_i$ veri noktasının en yakın k komuluğundaki veri noktalarını göstermektedir. Hata fonksiyonundaki optimal ağırlıklar $\sum_j w_{ij} = 1$ kısıtı altında en küçük kareler yöntemi kullanılarak elde edilir.

Optimal ağırlıklar w_{ij} sabitlenerek orijinal boyuttaki uzaklıklar indirgenmiş boyutta temsil edilir. Orijinal uzaklıkların indirgenmiş boyutta temsili

$$\Phi(\mathbf{s}) = \sum_i \left| \mathbf{s}_i - \sum_j w_{ij} \mathbf{s}_j \right|^2 \tag{3.40}$$

eşitliği ile tanımlanan karesel amaç fonksiyonu

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \mathbf{s}_i^T \mathbf{s}_i &= \mathbf{I} \\ \sum_{i=1}^{n} \mathbf{s}_i &= 0 \end{aligned} \tag{3.41}$$

kısıtları altında en küçükleyerek gerçekleştirilir. En küçükleme probleminin çözümünde sparse öz değer-öz vektör yaklaşımı kullanılabilir. Öz değer ayrıştırmasının gerçekleştirileceği simetrik, yarı pozitif $n \times n$ boyutlu seyrek matris $\mathbf{M}$

$$\mathbf{M} = (\mathbf{I} - \mathbf{W})(\mathbf{I} - \mathbf{W})^T \tag{3.42}$$

eşitliği ile elde edilir. $\mathbf{M}$ matrisinin sıfır olmayan en küçük d özdeğere karşı gelen özvektörler orjinde merkezileşmiş bağımsız koordinatlar sağlar. Yerel doğrusal eşleme

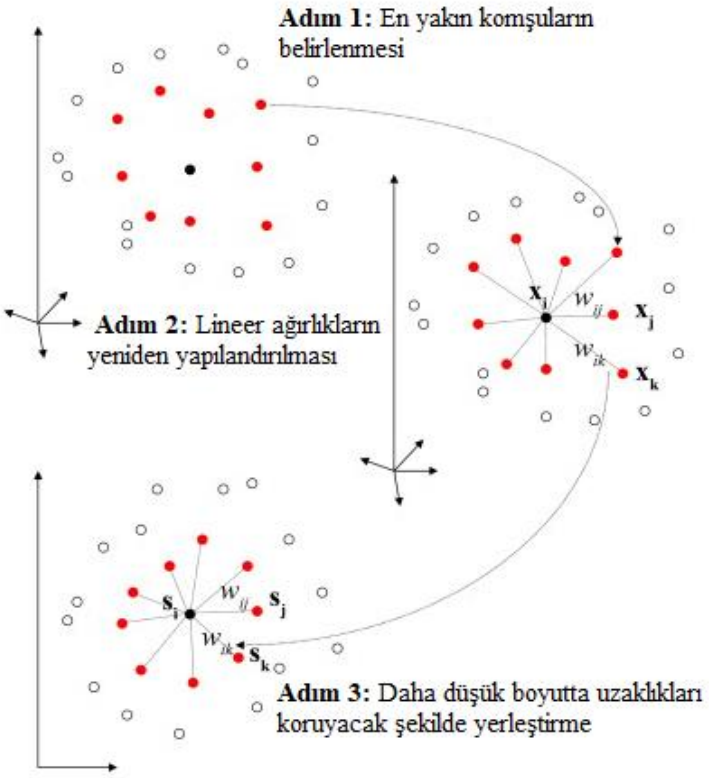
yöntemi için algoritma aşağıdaki gibidir. Algoritmanın grafiksel gösterimi Şekil 3.6'da verilmiştir.

Adım 1. Komşuluk için k ve indirgenmiş koordinat sistemindeki boyut sayısı d belirlenir.

Adım 2. Her bir $\mathbf{x}_i$ noktası için en yakın k komşusu belirlenir.

Adım 3. Her bir $\mathbf{x}_i$ noktasının en yakın komşularına olan ağırlıkları Eşitlik x ile yeniden yapılandırılarak doğrusal ağırlıklar w_{ij} hesaplanır.

Adım 4. İndirgenmiş d boyutlu uzayda Adım 3'de belirlenen ağırlıklar aynı kalacak şekilde Eşitlik xx ile $\mathbf{s}_i$ noktaları oluşturulur.



Şekil 3.6. Yerel doğrusal eşleme algoritmasının grafiksel gösterimi (Roweis ve Saul, 2000)

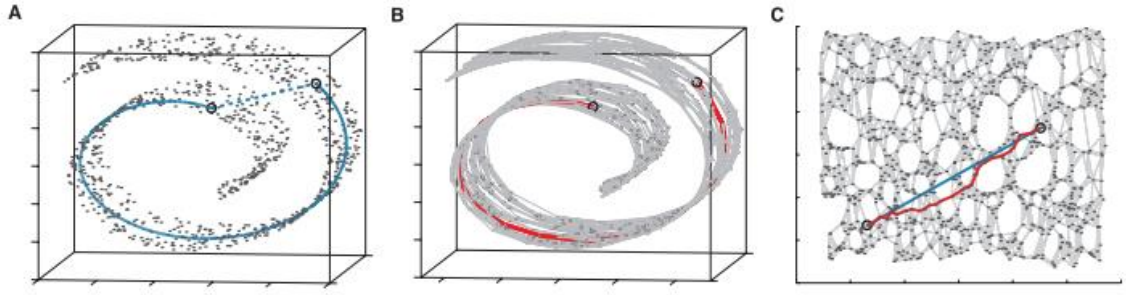
3.2.4. İzometrik Eşleme (ISOMAP)

İzometrik eşleme (Isometric Mapping, ISOMAP) yöntemi, yüksek boyutlu veriyi daha düşük boyutlu koordinat sisteminde temsil etmeyi amaçlayan grafik tabanlı, doğrusal olmayan boyut indirgeme yöntemidir. Temel bileşenler analizi ve çok boyutlu ölçekleme analizi ile veri noktalarının doğrusal olmayan ilişkilere sahip olduğu veri setlerinde etkili sonuçlar elde edilemeyebilir. Bu problemin çözümünde Tenenbaum ve ark. (2000), Öklid uzaklığı yerine jeodezik uzaklığın kullanıldığı klasik çok boyutlu

ölçekleme yönteminin bir uzantısı olan izometrik eşleme yöntemin önermişlerdir. Jeodezik uzaklık en temel tanımıyla bir yüzey üzerindeki manifold yol boyunca iki nokta arasındaki eğrisel en kısa uzunluk olarak tanımlanır.

Normalde iki birim arasındaki benzerliğini hesaplamak için genelde Öklid uzaklığı kullanılır. Bununla birlikte Öklid uzunluğunun kullanılması durumunda manifoldun iç geometrisi korunamaz hale gelmektedir. Öklid uzaklığı açısından benzer olan yani yakın olan iki nokta gerçekte uzak olabilir, çünkü gerçek uzaklıkları manifold boyunca bu noktalar arasındaki yolun uzunluğu olabilir.

Grafik tabanlı boyut indirgeme yöntemlerinden biri olan izometrik eşleme yönteminde boyut indirgeme işleminden sonra yakın mesafede olan noktalar yine birbirlerine yakın mesafede kalırlar. Uzak mesafeli noktalar da yine uzak olarak yerlerini korurlar. İki nokta arasındaki jeodezik uzaklık, izometrik eşleme ile bu uzaklığın belirlenmesi ve boyut indirgeme işlemlerinin grafiksel gösterimi Şekil 3.7'de verilmiştir.



Şekil 3.7. İzometrik eşleme yönteminin doğrusal olmayan boyutsallığı azaltmak için jeodezik yolları nasıl kullandığını gösteren "Swiss roll" veri seti (Tenenbaum ve ark. 2000)

(A) Doğrusal olmayan bir manifold üzerindeki iki rastgele nokta (daire içine alınmış) için, yüksek boyutlu giriş alanındaki Öklid uzaklığı (kesikli çizgi uzunluğu), düşük boyutlu manifold boyunca jeodezik uzaklık ile ölçülen içsel benzerliklerini doğru bir şekilde yansıtmayabilir.

(B) İki nokta arasındaki gerçek jeodezik uzunluk, k=7 ve n=1000 olmak üzere izometrik eşleme yöntemi ile doğru bir yaklaşım ile oluşturulan en yakın komşuluk grafiği G (kırmızı segmentler) ile verimli bir şekilde hesaplanmıştır.

(C) İzometrik eşleme yöntemi ile elde edilen iki boyutlu koordinat sisteminde, iki nokta arasındaki manifold boyunca jeodezik uzaklık (kırmızı çizgi), basit ve jeodezik uzaklığa benzer şekilde bir doğru (mavi çizgi) ile temsil edilir.

İzometrik eşleme yönteminde yerel doğrusallık ilkesi kullanılır ve komşu noktaların manifoldun doğrusal bir yamasının üzerinde bulunduğu varsayılır. Bu nedenle yakındaki noktalar için Öklid uzaklıklarının jeodezik uzaklıkları doğru olarak tahmin ettiği kabul edilir. Uzak noktalar için, jeodezik uzaklıklar manifold üzerinde komşu uzaklıklar eklenerek tahmin edilir.

İzometrik eşleme yönteminin algoritması aşağıdaki gibidir (Tan ve ark., 2006).

Adım 1. Tüm veri noktaları için en yakın komşularını belirlenir.

Adım 2. Her bir noktayı en yakın komşularına bağlayacak ve düğümler veri noktalarını, bağlantılar da noktalar arası uzaklıkları gösterecek şekilde ağırlıklandırılmış bir graf oluşturulur.

Adım 3. Oluşturulan komşuluk grafında, iki nokta arasındaki en kısa yolun uzunluğu olacak şekilde uzaklıkları yeniden tanımlanır.

Adım 4. Adım 3'de tanımlanan yeni uzaklık matrisine klasik çok boyutlu ölçekleme uygulanır.

İzometrik eşlemede ilk önce Öklid uzaklıkları kullanılarak her veri noktasının komşulukları belirlenir. Öklid uzaklığına göre veri noktası i'ye en yakın mesafede bulunan k veri noktasından biri j ise veya j noktası veri noktası i'nin ϵ yarıçapı içindeyse j noktası, i noktasının komşusu olarak belirlenir.

Daha sonra veri noktası i'nin en yakın k komşusu olarak belirlenen veri noktaları ile olan $d_{\mathbf{X}}(i,j)$ Öklid uzaklıklarının komşu noktaların kendi aralarındaki kenarları olduğu ve kenarların ağırlıkları $d_{\mathbf{X}}(i,j)$ olmak üzere tüm veri noktaları ağırlıklandırılmış grafik G olarak temsil edilir. Tüm veri nokta çiftleri arasında Floyd algoritması kullanılarak ağırlıklandırılmış G grafiğindeki en kısa yollar $d_G(i,j)$ hesaplanır. Hesaplanan $d_G(i,j)$ uzunlukları, M manifoldu üzerindeki jeodezik uzaklığının $d_M(i,j)$ tahmini olarak kullanılır. Son olarak, M manifoldu üzerindeki tüm veri nokta çiftleri arasındaki $n \times n$ boyutlu simetrik jeodezik uzaklık matrisi $D_G = \{d_G(i,j)\}$ temel alınarak klasik çok boyutlu ölçekleme ile boyut indirgeme gerçekleştirilir. İzometrik eşleme yönteminde veri nokta çiftleri arasındaki D_G jeodezik

uzaklık matrisini koruyacak boyut indirgeme amaçlanmaktadır. İndirgenmiş boyuttaki uzaklık matrisi D_Y olmak üzere izometrik eşleme yönteminde hata fonksiyonu

$$E = \|\tau(D_G) - \tau(D_Y)\| \tag{3.43}$$

en küçüklenmeye çalışılır.

İzometrik eşleme yeterli veri sağlandığı sürece doğrusal olmayan uzaklıkların gerçek boyutsallığını bulur. Yöntemin başarısı komşuluğu belirleyen k sayısına veya ε yarıçapına bağlıdır. İzometrik eşleme yöntemi girdi ve çıktı arasında matematiksel bir fonksiyon tanımlamadığından dolayı yeni bir veri noktası eklendiğinde tüm sürecin sıfırdan tekrarlanması yöntemin dezavantajlarından biridir (Turhan, 2004).

4. BOYUT İNDİRGEME YÖNTEMLERİNİN SINIFLANDIRMA PERFORMANSLARININ KARŞILAŞTIRILMASI

Bu bölümde incelenen özellik seçim ve özellik çıkarım yöntemlerinin karesel diskriminant analizdeki sınıflama performansları karşılaştırılacaktır. Boyut indirgeme yöntemlerinin sınıflama performanslarının karşılaştırılmasında karşılaştırma kriteri olarak doğru sınıflandırma olasılığı, entropy ve kappa katsayısı kullanılacaktır. Yüksek boyutlu gerçek veri setleri ile gerçekleştirilecek karşılaştırmalarda seçilen boyut sayısına göre yöntemler için karşılaştırma kriterlerinin değerleri hesaplanacaktır. Bu bölümde öncelikle karesel diskriminant analizi, karşılaştırma kriterleri ve uygulamada kullanılacak yüksek boyutlu gerçek veri setleri hakkında bilgi verilecektir.

Uygulamada öncelikle özellik seçimi için önerilen yeni yöntemlerin bilinen yöntemler karşısındaki performansı belirlenecektir. Uygulamanın ikinci aşamasında ele alınan dört özellik çıkarım yönteminin sınıflama performansı karşılaştırılacaktır. Uygulamanın üçüncü aşamasında ise özellik çıkarım yöntemleri öncesi uygulanacak özellik seçim yönteminin özellik çıkarım yöntemlerinin sınıflama performansı üzerindeki etkisi belirlenecektir.

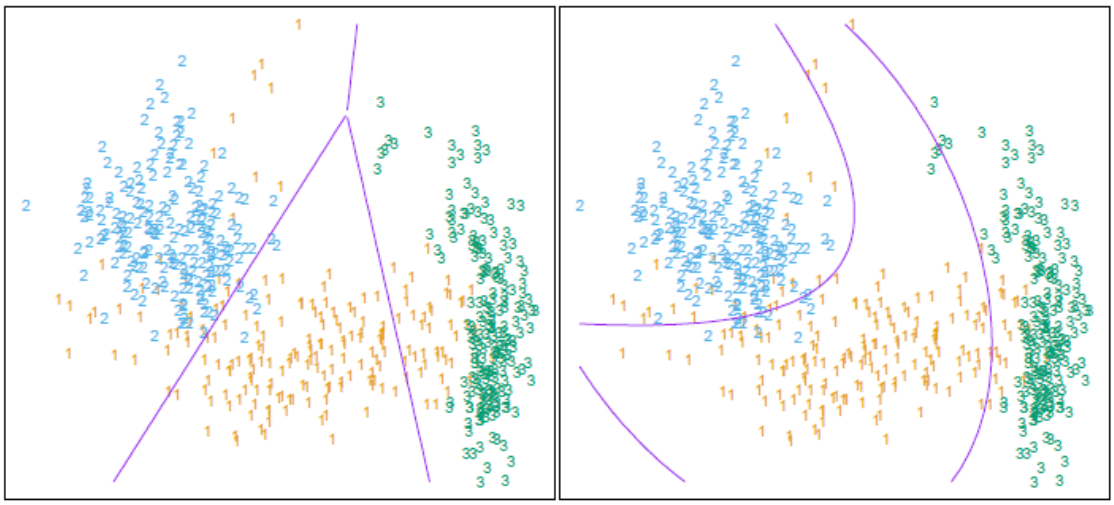
4.1. Karesel Diskriminant Analizi

Diskriminant analizi, en az hata ile birimleri ait oldukları sınıflara atamak için ayırma fonksiyonlarının oluşturulması ve oluşturulan ayırma fonksiyonları yardımı ile daha sonradan gözlemlenen, sınıf üyeliği bilinmeyen birimlerin sınıflandırılmasını gerçekleştirmeyi amaçlayan çok değişkenli istatistiksel bir analizdir (Rencher, 2002). Diskriminant analizinin temeli birimlerin gruplarının belirlenmesini sağlayacak ayırma fonksiyonlarının oluşturulmasıdır. Ayırma fonksiyonlarının oluşturulmasında sınıf ortalamaları arasındaki farklılığın maksimum olması amaçlanmaktadır.

Diskriminant analizinde sınıflara ait varyans-kovaryans matrisleri eşitse yani $\Sigma_1 = \cdots = \Sigma_K = \Sigma$ durumunda doğrusal diskriminant analizi ile ayırma fonksiyonları oluşturulmaktadır. Eğer sınıflara ait varyans-kovaryans matrisleri farklı ise yani $\Sigma_i \neq \Sigma_j$ durumunda karesel diskriminant analizi ile ayırma fonksiyonları oluşturulmaktadır. Sınıflara ait ortalama vektörü μ_k ve varyans-kovaryans matrisi Σ_k ve $k = 1, \dots, K$ olmak üzere karesel diskriminant fonksiyonu

$$Q_k(\mathbf{x}) = -\frac{1}{2} log|\Sigma_k| - \frac{1}{2}(\mathbf{x} - \mu_k)^T \Sigma_k^{-1}(\mathbf{x} - \mu_k) + log\pi_k \tag{4.1}$$

eşitliği ile oluşturulmaktadır. Eşitlikte yer alan π_k gösterimi k. gruba ait önsel olasılığı gösterir. Lineer ve karesel diskriminant analizleri yapı olarak birbirine oldukça benzerdir. İki ayırma fonksiyonu arasındaki temel fark varyans-kovaryans matrislerinin farklılığıdır. Karesel diskriminant analizinde varyans-kovaryans matrislerinin farklı kullanılması yöntemi daha esnek hale getirmektedir. Doğrusal ve karesel diskriminant analizinin grafiksel karşılaştırılması Şekil 4.1'de verilmiştir.



Şekil 4.1. Üç sınıflı bir veri kümesi için doğrusal diskriminant analizi (soldaki grafik) ve karesel diskriminant analizi (sağdaki grafik) ile elde edilen karar sınırları (Hastie ve ark., 2008)

Karesel diskriminant analizinde de sınıflama kuralı **x** gözlem vektörüne sahip birimi en büyük ayırma fonksiyonu değerinin elde edildiği gruba atama şeklindedir. Buna göre karesel diskriminant analizinde sınıflandırma kuralı

$$G(\mathbf{x}) = arg \max_k Q_k(\mathbf{x}) \tag{4.2}$$

şeklinde ifade edilir. Sınıflara ait ortalama vektörü μ_k ve varyans-kovaryans matrisi Σ_k olmak üzere Bayes kuralına göre **x** gözlem vektörünün k. sınıfa ait olma sonsal olasılığı

$$P(G = k|\mathbf{X} = \mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{\sum_{g=1}^{K} \pi_g f_g(\mathbf{x})} \tag{4.3}$$

eşitliği ile belirlenir. Eşitlikte yer alan $f_k(.)$ fonksiyonu ortalama vektörü μ_k ve varyans-kovaryans matrisi Σ_k olan çok değişkenli normal dağılıma ait olasılık yoğunluk fonksiyonudur. Eşitlik (4.3)'de verilen sonsal olasılığa göre **x** gözlem vektörü en yüksek olasılık değerinin elde edildiği sınıfa atanır. Buna göre karesel diskriminant analizinde sonsal olasılıklara göre sınıflandırma kuralı

$$G(\mathbf{x}) = arg\ \max_k P(G = k|\mathbf{X} = \mathbf{x}) \tag{4.4}$$

şeklinde ifade edilir.

4.2. Karşılaştırma Kriterleri

4.2.1. Doğru Sınıflandırma Olasılığı

Sınıflandırma analizi sonucu gerçek sınıf üyelikleri ve tahmini sınıf üyeliklerine göre oluşturulan sınıflandırma doğruluk tablosu $k = 1, \dots, K$ sınıf için

Tablo 4.1. Sınıflandırma doğruluk tablosu

Tahmini Sınıf Üyelikleri	**Gerçek Sınıf Üyelikleri**				**Toplam**
	1	2	⋯	K	
1	f_{11}	f_{12}	⋯	f_{1K}	$f_{1.}$
2	f_{21}	f_{22}	⋯	f_{2K}	$f_{2.}$
⋮	⋮	⋮	⋯	⋮	⋮
K	f_{K1}	f_{K2}	⋯	f_{KK}	$f_{K.}$
Toplam	n_1	n_2	⋯	n_k	n

şeklinde verilsin. Tabloda yer alan f_{KK} gösterimi gerçek sınıf üyeliği K olup sınıflama analizi sonucunda doğru olarak K sınıfına atanan birimlerin sayısını göstermektedir. Tabloda yer alan f_{K1} gösterimi gerçek sınıf üyeliği 1. sınıf iken sınıflama analizi sonucunda yanlış olarak K sınıfına atanan birimlerin sayısını göstermektedir. Doğru sınıflandırılan toplam birim sayısının, toplam gözlem sayısına bölünmesi ile

$$DSO = \frac{\sum_{k=1}^{K} f_{kk}}{n} \tag{4.5}$$

doğru sınıflandırma olasılığı ile elde edilir. Doğru sınıflandırma olasılığının 1'e yakın olması sınıflama performansının başarılı olduğunu gösterir.

4.2.2. Entropy

Entropy, sınıflama belirsizliğinin bir ölçüsüdür. Entropy kriteri $i = 1, 2, \cdots, n$ ve $k = 1, 2, \cdots, k$ olmak üzere

$$En(\tau) = -\sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik} \ln(\tau_{ik}) \tag{4.6}$$

eşitliği ile hesaplanır. Eşitlikte yer alan τ_{ik} gösterimi i. birimin k. kümeye ait olma olasılığıdır ve

$$\tau_{ik} = \frac{\pi_k f_k(\mathbf{x}_i)}{\sum_{l=1}^{K} \pi_l f_l(\mathbf{x}_i)} \tag{4.7}$$

eşitliği ile hesaplanır. Entropy değeri ne kadar küçük ise sınıflama o kadar başarılıdır.

4.2.3. Kappa Katsayısı

Kappa katsayısı κ, iki karar vericinin kararları arasındaki uyumu ölçen bir istatistik ölçüsüdür. κ katsayısı, Tablo 4.1'de verilen sınıflandırma doğruluk tablosunda gerçek sınıf üyelikleri ve tahmini sınıf üyelikleri arasındaki uyumun ölçülmesi için kullanılabilir. κ katsayısının değerinin büyük olması sınıflamanın başarılı olduğu gösterir. κ katsayısı

$$\kappa = \frac{p_0 - p_c}{1 - p_c} \tag{4.8}$$

eşitliği ile hesaplanır. Eşitlikte yer alan p_0 gösterimi gerçek ve tahmini sınıf üyelikleri arasında gerçekleşen uyum için bir olasılık değeridir ve Eşitlik (4.5)'de verilen doğru sınıflandırma olasılığına karşılık gelmektedir. Eşitlik (4.8)'de yer alan p_c gösterimi ise gerçek ve tahmini sınıf üyelikleri arasında beklenen uyum için bir olasılık değeridir ve

$$p_c = \frac{\sum_{k=1}^{K} n_k f_{k.}}{n^2} \tag{4.8}$$

eşitliği ile hesaplanır.

4.3. Veri Setleri

Boyut indirgeme yöntemlerinin karesel diskriminant analizinde sınıflandırma performanslarının karşılaştırılmasında özellik (değişken) sayısının birim sayısından fazla olduğu gerçek 10 adet yüksek boyutlu veri seti kullanılacaktır.

Arcene veri seti, spektrometrik veriler ile normal ve kanserli örnekleri birbirinden ayırt etmek amaçlı kullanılan bir veri setidir. Isabelle ve ark. (2004) tarafından oluşturulan Arcene veri seti 10000 özellik ve 100 birimden oluşmaktadır. Arcene veri setinde 44 birim kanser, 56 birim normal olmak üzere iki sınıftan oluşmaktadır.

Breast veri seti (Karakoca ve ark., 2013), üç farklı kanser türünü ayırt etmek için kullanılan 1213 özellik ve 98 birimden oluşan bir veri setidir. Her bir kanser türüne ait birim sayısı sırasıyla 11, 51 ve 36'dır.

Chowdary veri seti, lenf nodenegatif meme tümörleri ve Dukes'B kolon tümörlerinden elde edilen dokudan oluşur. Chowdary ve ark. (2006) tarafından oluşturulan veri seti 62 meme tümörü ve 42 kolon tümörü olmak üzere 104 birim ve 22283 özellikten oluşmaktadır.

Phoneme veri seti seti, beş sesbirimini ayırt etmek için 256 özellik ve 244 birimden oluşmaktadır. Hastie ve ark. (1995) tarafından kullanılan veri setindeki sınıfların birim sayıları sırasıyla 48, 56, 42, 38 ve 59'dur.

Prostate veri seti, Singh ve ark. (2002)'nın mikrodizi çalışmasından elde edilen gen ekspresyon verileri 50 normal ve 52 kanser birimi olmak üzere 102 birim ve 6033 özellikten oluşmaktadır.

Elma ve Kiraz veri setleri, Dedeoğlu (2011) tarafından elma ve kiraz ağaçlarında oluşan çinko eksikliğini görünür yakın kızılötesi yöntemle belirlendiği çalışma için oluşturulmuştur. Üç farklı bahçeden alınan elma ve kiraz yapraklarının spektral yansıma ölçümlerinden oluşan veri seteri 60 birim ve 701 özellik içermektedir.

Şeftali ve Şeftali Bahçe veri setleri Dedeoğlu (2020) tarafından şeftali ağaçlarındaki nitrojen düzeyini spektral yansıma ölçümleri ile belirlendiği çalışma için oluşturulmuştur. Her iki veri setinde de eksik, yeterli ve aşırı nitrojen düzeyi olmak üzere üç sınıftan oluşmaktadır.

Şekerpancarı veri seti Dedeoğlu ve ark. (2019) tarafından şekerpancarı bitkisinin yapraklarındaki azot içeriğini hiperspektral yansımalar ile belirlediği çalışma için oluşturulmuştur. Veri seti şekerpancarı bitkisinin yapraklarındaki azot düzeyi noksan, yeterli ve fazla olmak üzere üç sınıftan oluşmaktadır.

Çalışma kapsamında incelenen yüksek boyutlu veri setleri ile ilgili birim sayıları, özellik sayıları, küme ve kümelerdeki birim sayılarının yer aldığı bilgiler Tablo 4.2'de verilmiştir.

Tablo 4.2. İncelenen veri setlerinin temel özellikleri

Veri Seti	Birim Sayısı	Değişken Sayısı	Küme Sayısı	Küme Birim sayıları
Arcene	100	10000	2	56, 44
Breast	98	1213	3	11, 51, 36
Chowdary	104	22283	2	62, 42
Elma	60	701	3	20, 20, 20
Kiraz	60	701	3	20, 20, 20
Phoneme	244	256	5	48, 56, 42, 38, 59
Prostate	102	6033	2	50, 52
Şeftali	84	601	3	28, 27, 29
Şeftali Bahçe	96	601	3	45, 20 31
Şekerpancarı	72	601	3	24, 24,24

4.4. Özellik Seçim Yöntemlerinin Karşılaştırılması

Çalışma kapsamında incelenen özellik seçim yöntemleri ve sınıfların değişim katsayısına dayalı önerilen değişim katsayısı oranları ile özellik seçim yöntemlerinin karesel diskriminant analizindeki sınıflama performansları doğrusal sınıflandırma olasılığı, entropy ve kappa katsayıları bakımından karşılaştırılmıştır.

Breast veri seti için seçilen özellik sayısına göre hesaplanan doğru sınıflandırma olasılıkları, entropy ve kappa katsayı değerleri Tablo 4.3, 4-4 ve 4.5'de verilmiştir. Özellik seçim yöntemleri için seçilen özelliklere göre her sınıf için en az bir birimi doğru olarak sınıflandırıldığında doğru sınıflandırma olasılıkları hesaplanmıştır.

Tablo 4.3 incelendiğinde, genel olarak özellik seçim yöntemlerinin Breast veri setinde boyut indirgemede oldukça başarılı olduğu görülmektedir. Breast veri setinde boyut indirgemede doğru sınıflandırma olasılığı kriterine göre özellik seçim yöntemleri içerisinde komşuluk bileşen analizi (NCA) daha başarılı olmuştur. Hesaplanan doğru sınıflandırma olasılıkları incelendiğinde seçilen özellik sayısı arttığında yöntemler arasındaki farklılıkların azaldığı görülmüştür.

Tablo 4.3. Breast verisi için hesaplanan doğru sınıflandırma olasılıkları

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	-	-	-	-	-	-	0.510	**0.592**	0.561	-
2	0.745	0.867	-	-	0.745	**0.878**	0.735	0.622	0.602	0.847
3	0.816	0.878	0.847	0.776	0.755	**0.959**	0.745	0.827	0.735	0.827
4	0.867	0.929	0.867	0.755	0.816	**0.959**	0.776	0.888	0.755	0.847
5	0.867	0.908	0.888	0.796	0.867	**0.969**	0.806	0.878	0.776	0.918
6	0.929	0.959	0.888	0.867	0.898	**0.990**	0.847	0.918	0.755	0.929
7	0.929	**1.000**	0.898	0.898	0.969	**0.990**	0.878	0.929	0.867	0.929
8	0.918	**1.000**	0.918	0.929	0.959	**1.000**	0.908	0.949	0.898	0.949
9	0.949	**1.000**	0.969	0.959	0.969	**1.000**	0.959	0.949	0.898	0.990
10	0.949	0.990	0.959	0.949	0.959	**1.000**	0.969	0.969	0.918	0.959

Tablo 4.4. Breast verisi için hesaplanan Entropy değerleri

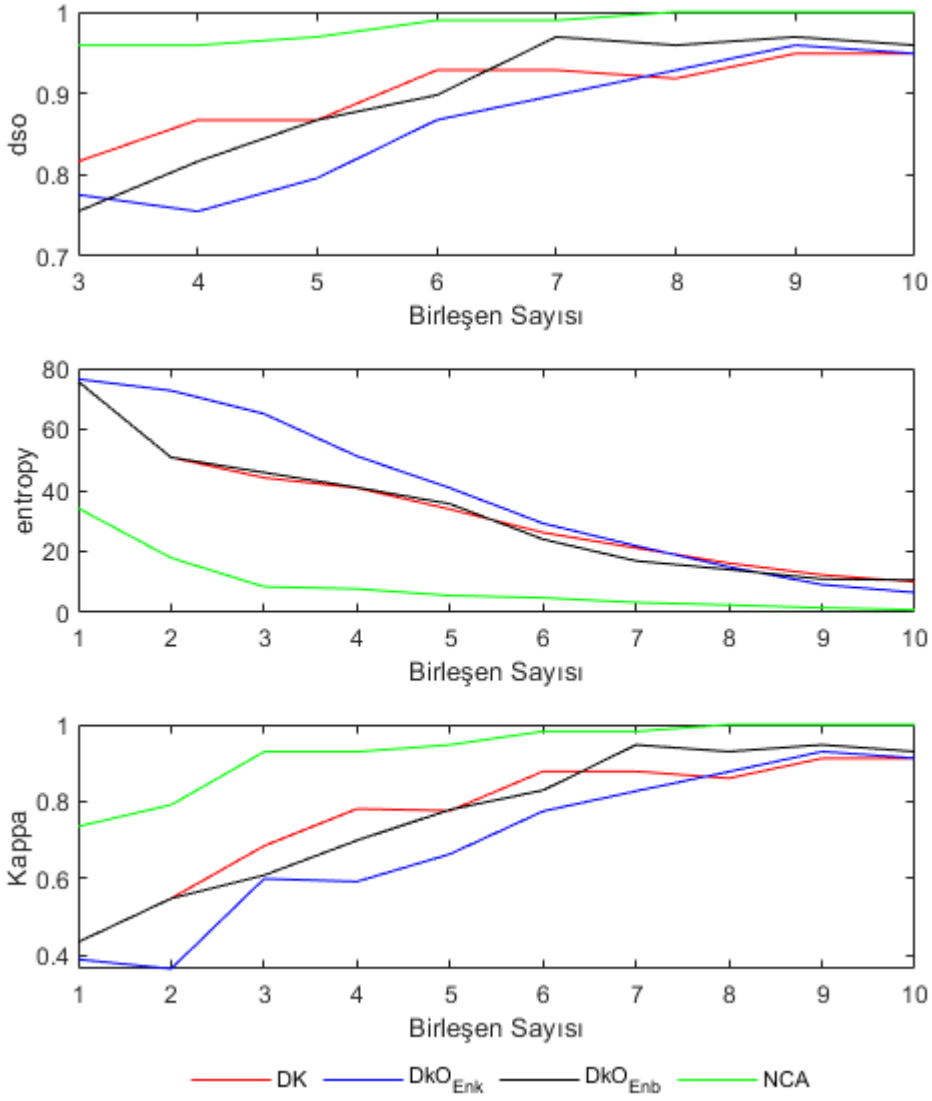
d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	75.69	62.80	42.70	76.47	75.69	**34.19**	82.56	79.52	77.37	55.56
2	50.82	24.31	32.67	72.69	50.82	**17.90**	61.22	64.96	69.08	44.01
3	44.16	18.55	28.36	65.09	45.85	**8.44**	53.03	38.06	53.65	38.06
4	40.80	15.95	26.16	51.21	40.96	**7.75**	47.07	27.54	51.65	30.85
5	33.84	14.09	20.97	40.78	35.57	**5.56**	35.32	23.77	46.80	23.17
6	26.22	11.08	17.32	29.21	23.99	**4.83**	30.16	19.91	44.00	19.73
7	21.16	5.90	16.78	21.84	16.91	**3.28**	22.51	17.29	34.54	15.97
8	16.15	3.21	11.32	14.94	14.09	**2.46**	19.73	13.25	30.93	11.21
9	12.33	2.33	7.70	9.13	10.99	**1.56**	9.05	10.11	28.07	11.27
10	10.06	0.99	6.59	6.55	10.62	**0.89**	7.33	7.63	20.23	10.75

Tablo 4.5. Breast verisi için hesaplanan Kappa istatistikleri

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	0.434	0.586	0.638	0.389	0.434	**0.735**	0.160	0.243	0.211	0.470
2	0.547	0.757	0.733	0.364	0.547	**0.791**	0.543	0.364	0.306	0.729
3	0.684	0.786	0.720	0.599	0.608	**0.929**	0.546	0.704	0.541	0.704
4	0.781	0.874	0.765	0.591	0.699	**0.929**	0.605	0.809	0.571	0.738
5	0.776	0.844	0.804	0.663	0.779	**0.947**	0.661	0.793	0.611	0.860
6	0.878	0.931	0.804	0.775	0.829	**0.982**	0.732	0.858	0.577	0.876
7	0.878	**1.000**	0.825	0.827	0.947	**0.982**	0.790	0.878	0.769	0.877
8	0.860	**1.000**	0.858	0.877	0.930	**1.000**	0.843	0.913	0.823	0.912
9	0.912	**1.000**	0.947	0.930	0.948	**1.000**	0.929	0.913	0.824	0.983
10	0.912	0.982	0.929	0.913	0.930	**1.000**	0.947	0.947	0.860	0.930

Entropy değerlerine göre Breast veri setinde boyut indirgemede NCA yöntemi diğer yöntemlere göre daha başarılı olmuştur. Entropy değerleri incelendiğinde yöntemlerin sınıflama performanslarını ayırt etmede diğer karşılaştırma kriterlerine göre daha başarılı olduğu görülmüştür.

Kappa katsayısına göre de Breast veri setinde boyut indirgemede NCA yöntemi diğer yöntemlere göre daha başarılı olmuştur. Genel olarak seçilen özellik sayısı arttığında sınıflama başarısının arttığı görülse de birkaç istisnai durum için seçilen özellik sayısının artması durumunda sınıflama performansının azaldığı görülmüştür. Breast veri setinde önerilen özellik seçim yöntemlerinin, NCA ve değişim katsayına göre sınıflama performansının grafiksel karılaştırılması Şekil 4.2'de verilmiştir.



Şekil 4.2. Önerilen özellik seçim yöntemlerinin seçili yöntemlere göre sınıflandırma performansı (Breast)

Chowdary veri seti için seçilen özellik sayısına göre hesaplanan doğru sınıflandırma olasılıkları, entropy ve kappa katsayı değerleri Tablo 4.6, 4-7 ve 4.8'de verilmiştir.

Tablo 4.6. Chowdary verisi için hesaplanan doğru sınıflandırma olasılıkları

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	0.731	0.942	0.865	0.856	**0.904**	0.683	0.683	0.808	0.865	0.865
2	0.769	**0.971**	0.933	0.846	0.942	0.692	0.663	0.952	0.933	0.933
3	0.760	0.933	0.923	0.865	0.952	0.663	0.663	**0.962**	0.923	0.942
4	0.779	0.933	0.942	0.856	**0.962**	0.673	0.683	**0.962**	0.942	0.942
5	0.769	0.952	0.962	0.894	**0.971**	0.721	0.875	0.952	0.962	0.962
6	0.817	0.952	**0.990**	0.894	0.971	0.721	0.865	0.952	**0.990**	0.952
7	0.846	0.952	**0.990**	0.904	0.971	0.712	0.904	0.971	**0.990**	0.962
8	0.856	0.971	**0.990**	0.904	0.971	0.760	0.913	0.981	**0.990**	0.981
9	0.865	0.971	**0.990**	0.904	0.981	0.846	0.962	**0.990**	**0.990**	**0.990**
10	0.865	0.971	**0.990**	0.923	0.981	0.865	0.942	**0.990**	**0.990**	**0.990**
15	0.913	**0.990**	**0.990**	0.933	0.981	0.942	**0.990**	**0.990**	**0.990**	**0.990**
20	0.962	0.990	0.990	0.962	0.981	0.952	**1.000**	0.990	0.990	0.990

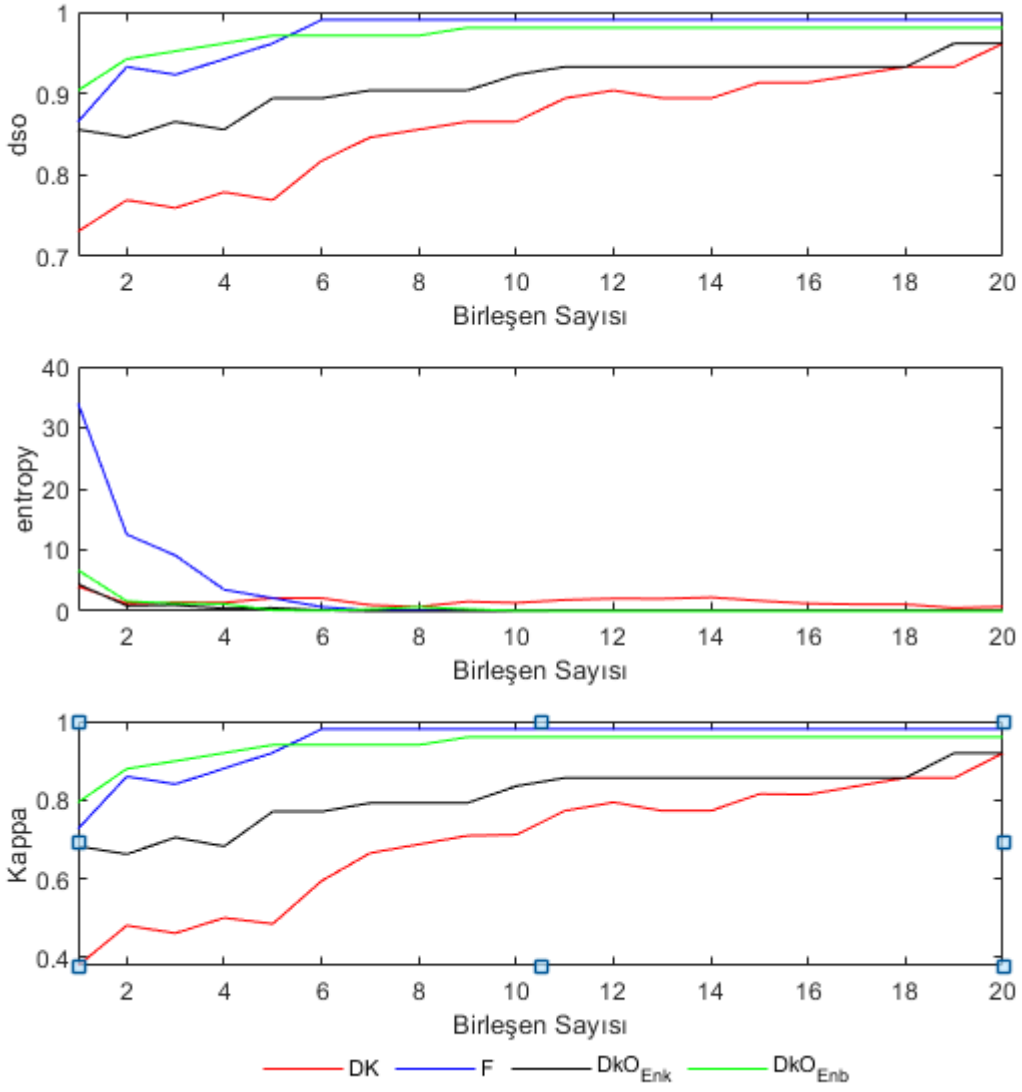
Tablo 4.7. Chowdary verisi için hesaplanan Entropy değerleri

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	**4.021**	22.338	34.010	4.412	6.680	39.680	49.091	21.359	34.010	34.010
2	**1.256**	6.590	12.533	0.878	1.619	34.038	34.210	8.170	12.533	12.533
3	1.436	3.728	9.104	**0.982**	1.260	30.046	33.983	1.134	9.104	5.635
4	1.386	3.572	3.490	**0.433**	1.179	28.173	17.120	0.970	3.490	3.490
5	2.111	0.486	2.122	0.518	**0.132**	26.375	7.530	0.524	2.122	1.211
6	2.120	0.062	0.680	0.140	**0.061**	24.468	7.796	1.456	0.680	1.210
7	0.982	0.431	0.124	**0.003**	0.296	22.481	7.507	3.306	0.124	1.794
8	0.706	0.112	0.173	**0.000**	0.691	18.002	6.604	2.378	0.173	1.347
9	1.556	0.091	0.003	**0.000**	0.311	11.988	5.759	1.301	0.003	0.993
10	1.345	0.020	**0.000**	**0.000**	0.046	12.615	5.161	**0.000**	**0.000**	0.559
15	1.686	0.296	**0.000**	**0.000**	**0.000**	10.090	0.017	**0.000**	**0.000**	**0.000**
20	0.727	**0.000**	**0.000**	0.003	**0.000**	7.695	0.005	**0.000**	**0.000**	**0.000**

Tablo 4.8. Chowdary verisi için hesaplanan Kappa istatistikleri

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	0.379	**0.879**	0.727	0.682	0.794	0.258	0.252	0.624	0.727	0.727
2	0.481	**0.940**	0.860	0.662	0.879	0.271	0.220	0.901	0.860	0.860
3	0.461	0.857	0.840	0.705	0.899	0.207	0.220	**0.920**	0.840	0.880
4	0.500	0.857	0.880	0.682	**0.920**	0.226	0.271	**0.920**	0.880	0.880
5	0.485	0.899	0.920	0.771	**0.940**	0.348	0.731	0.899	0.920	0.920
6	0.594	0.899	**0.980**	0.771	0.940	0.354	0.712	0.899	**0.980**	0.899
7	0.665	0.899	**0.980**	0.792	0.940	0.334	0.796	0.940	**0.980**	0.920
8	0.687	0.940	**0.980**	0.792	0.940	0.452	0.815	0.960	**0.980**	0.960
9	0.709	0.940	**0.980**	0.792	0.960	0.660	0.920	**0.980**	**0.980**	**0.980**
10	0.712	0.940	**0.980**	0.835	0.960	0.705	0.878	**0.980**	**0.980**	**0.980**
15	0.815	**0.980**	**0.980**	0.856	0.960	0.877	**0.980**	**0.980**	**0.980**	**0.980**
20	0.919	0.980	0.980	0.919	0.960	0.898	**1.000**	0.980	0.980	0.980

Tablo 4.6 incelendiğinde doğru sınıflandırma olasılığına göre Chowdary veri setinde boyut indirgemede en başarılı özellik seçim yöntemi F test istatistiğine göre filtreleme özellik seçim yöntemi olmuştur. Önerilen özellik seçim yöntemlerinden DkO_{enb} yöntemi özellikle seçilen ilk beş özellik için çok başarılı sonuçlar vermiştir. Entropy değerlerine göre önerilen özellik seçim yöntemlerinin genel başarısının diğer özellik seçim yöntemlerinden daha başarılı olduğu görülmüştür. Kappa katsayısı kriterine göre doğru sınıflandırma olasılığına benzer şekilde F test istatistiği genel olarak daha başarılı bulunmuştur. Chowdary veri seti için önerilen özellik seçim yöntemlerinin, F ve değişim katsayına göre sınıflama performansının grafiksel karılaştırılması Şekil 4.3'de verilmiştir.



Şekil 4.3. Önerilen özellik seçim yöntemlerinin seçili yöntemlere göre sınıflandırma performansı (Chowdary)

Elma veri seti için seçilen özellik sayısına göre hesaplanan doğru sınıflandırma olasılıkları, entropy ve kappa katsayı değerleri Tablo 4.9, 4-10 ve 4.11'de verilmiştir.

Tablo 4.9. Elma verisi için hesaplanan doğru sınıflandırma olasılıkları

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	0.300	0.550	0.550	**0.600**	0.433	0.533	0.550	0.550	0.550	0.550
2	0.517	**0.683**	0.633	**0.683**	0.550	0.717	0.667	0.633	0.667	0.633
3	0.617	**0.717**	0.633	0.700	0.583	0.683	0.700	0.683	**0.717**	0.683
4	0.767	0.750	0.717	**0.783**	0.683	0.650	0.733	**0.783**	0.767	**0.783**
5	0.767	0.750	0.667	**0.850**	0.717	0.683	0.767	0.817	0.800	0.817
6	**0.833**	0.800	0.683	0.817	0.750	0.783	0.783	0.800	0.800	0.817
7	0.883	0.817	0.800	0.783	0.783	0.850	0.867	0.833	0.833	**0.933**
8	0.883	0.867	**0.917**	0.817	0.817	0.867	0.850	**0.917**	**0.917**	0.850
9	0.900	0.900	0.900	0.867	0.850	0.867	**0.933**	0.883	**0.933**	0.883
10	0.933	0.900	0.933	**0.967**	0.867	0.900	**0.967**	0.917	0.917	0.917
15	0.983	**1.000**	**1.000**	**1.000**	0.983	**1.000**	0.983	0.983	**1.000**	0.983

Tablo 4.10. Elma verisi için hesaplanan Entropy değerleri

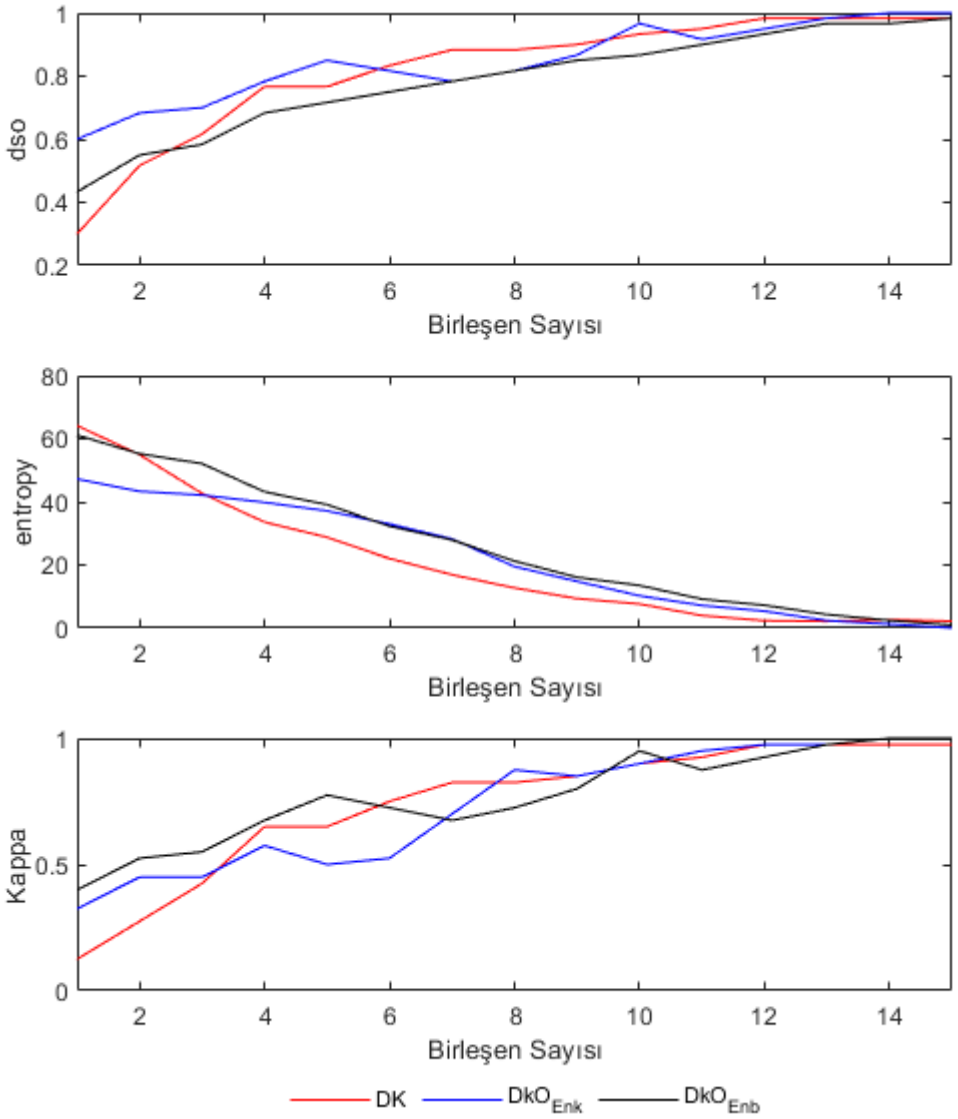
d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	64.13	52.70	52.37	**47.27**	61.15	55.65	52.37	52.37	52.37	52.37
2	54.94	47.40	46.75	**43.33**	55.25	44.46	47.38	46.75	47.38	46.75
3	42.65	44.03	43.73	42.15	52.11	42.70	40.55	43.52	**41.37**	43.52
4	**33.58**	38.16	38.27	39.83	43.18	37.42	34.43	37.32	37.72	37.32
5	28.76	30.12	33.08	37.14	39.12	32.77	**24.43**	34.21	31.09	34.21
6	22.03	25.42	29.91	32.99	32.30	27.20	**21.64**	28.93	28.93	30.31
7	**16.91**	21.27	25.71	28.25	27.95	23.41	20.39	24.61	24.61	25.42
8	**12.72**	13.29	16.02	19.50	21.24	21.67	15.56	19.91	19.91	21.06
9	**9.36**	10.73	13.82	14.80	16.12	12.76	10.80	15.61	13.06	15.61
10	7.64	8.16	10.58	10.25	13.52	9.94	**6.65**	9.64	8.86	9.64
15	2.20	0.79	0.32	**0.05**	0.99	0.30	0.74	0.95	0.85	1.65

Tablo 4.11. Elma verisi için hesaplanan Kappa istatistikleri

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	0.125	**0.325**	**0.325**	0.4	0.2	0.3	**0.325**	**0.325**	**0.325**	**0.325**
2	0.275	0.525	0.45	0.525	0.325	**0.575**	0.5	0.45	0.5	0.45
3	0.425	**0.575**	0.45	0.55	0.375	0.525	0.55	0.525	**0.575**	0.525
4	0.65	0.625	0.575	**0.675**	0.525	0.475	0.6	**0.675**	0.65	**0.675**
5	0.65	0.625	0.5	**0.775**	0.575	0.525	0.65	0.725	0.7	0.725
6	**0.75**	0.7	0.525	0.725	0.625	0.675	0.675	0.7	0.7	0.725
7	**0.825**	0.725	0.7	0.675	0.675	0.775	0.8	0.75	0.75	0.9
8	0.825	0.8	**0.875**	0.725	0.725	0.8	0.775	**0.875**	**0.875**	0.775
9	0.85	0.85	0.85	0.8	0.775	0.8	**0.9**	0.825	**0.9**	0.825
10	0.9	0.85	0.9	**0.95**	0.8	0.85	**0.95**	0.875	0.875	0.875
15	0.975	**1**	**1**	**1**	0.975	**1**	0.975	0.975	**1**	0.975

Tablo 4.9 incelendiğinde doğru sınıflandırma olasılığına göre Elma veri setinde boyut indirgemede genel olarak en başarılı özellik seçim yöntemi önerilen özellik seçim yöntemlerinden DkO_{enk} yöntemi olmuştur. Entropy kriterine göre daha başarılı olan özellik seçim yöntemleri değişim katsayısı, DkO_{enk} ve ReliefF yöntemleridir.

Elma veri seti için önerilen özellik seçim yöntemlerinin değişim katsayına göre sınıflama performansının grafiksel karılaştırılması Şekil 4.4'de verilmiştir.



Şekil 4.4. Önerilen özellik seçim yöntemlerinin seçili yöntemlere göre sınıflandırma performansı (Elma)

Kiraz veri seti için seçilen özellik sayısına göre hesaplanan doğru sınıflandırma olasılıkları, entropy ve kappa katsayı değerleri Tablo 4.12, 4-13 ve 4.14'de verilmiştir.

Tablo 4.12. Kiraz verisi için hesaplanan doğru sınıflandırma olasılıkları

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	0.467	0.600	**0.667**	0.517	0.533	0.483	0.600	0.650	0.650	0.650
2	0.550	0.600	0.700	0.500	**0.767**	0.533	0.633	0.633	0.633	0.633
3	0.633	0.567	0.800	0.617	**0.833**	0.550	0.633	0.633	0.633	0.633
4	0.683	0.650	**0.817**	0.650	0.800	0.667	0.717	0.717	0.717	0.717
5	0.750	0.633	**0.817**	0.700	**0.817**	0.700	0.767	0.783	0.750	0.783
6	0.800	0.767	0.850	0.767	0.867	0.833	**0.900**	0.783	0.783	0.883
7	0.850	0.800	0.900	0.800	0.883	0.850	**0.933**	0.900	0.833	0.900
8	0.867	0.867	0.950	0.867	0.900	0.900	**0.967**	0.950	0.900	0.950
9	0.917	0.883	0.950	0.867	0.950	0.900	0.967	**0.983**	0.900	0.950
10	0.883	0.917	0.967	0.933	0.950	0.917	**0.983**	**0.983**	0.933	**0.983**
15	0.983	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

Tablo 4.13. Kiraz verisi için hesaplanan Entropy değerleri

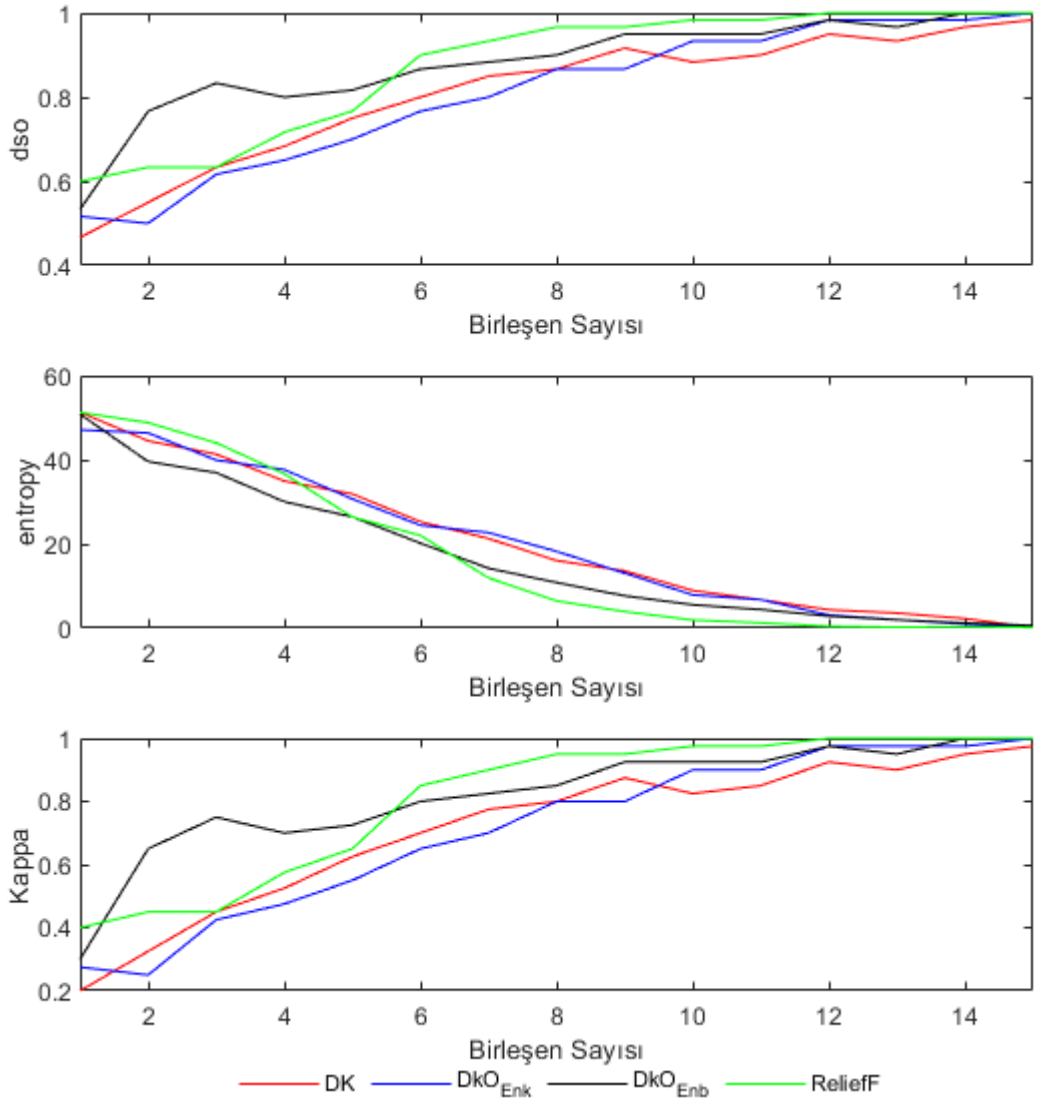
d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	51.30	45.16	**43.19**	47.11	50.83	50.53	51.26	51.36	51.36	51.36
2	44.51	44.39	**38.99**	46.38	39.57	48.34	48.82	48.82	48.82	48.82
3	41.32	41.71	**32.79**	39.93	36.94	46.83	44.01	44.01	44.01	44.01
4	34.94	34.87	**31.61**	37.67	30.05	38.47	36.68	36.68	36.68	36.68
5	31.88	30.62	**28.11**	30.70	26.44	31.86	26.40	30.24	33.47	30.24
6	25.29	26.09	22.19	24.44	**20.19**	25.86	22.02	26.92	29.48	13.38
7	21.28	20.42	16.29	22.69	14.24	21.33	11.94	**11.65**	23.04	**11.65**
8	16.06	17.04	11.80	18.25	10.82	18.43	**6.46**	6.99	15.63	6.99
9	13.61	15.53	10.46	13.07	7.68	15.07	**3.93**	4.73	12.00	4.52
10	8.94	10.74	7.86	7.89	5.56	10.88	**1.93**	3.51	8.99	3.51
15	0.12	0.49	0.49	0.18	0.63	0.54	0.19	0.27	0.21	**0.07**

Tablo 4.14. Kiraz verisi için hesaplanan Kappa istatistikleri

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	0.2	0.4	**0.5**	0.275	0.3	0.225	0.4	0.475	0.475	0.475
2	0.325	0.4	0.55	0.25	**0.65**	0.3	0.45	0.45	0.45	0.45
3	0.45	0.35	0.7	0.425	**0.75**	0.325	0.45	0.45	0.45	0.45
4	0.525	0.475	**0.725**	0.475	0.7	0.5	0.575	0.575	0.575	0.575
5	0.625	0.45	**0.725**	0.55	**0.725**	0.55	0.65	0.675	0.625	0.675
6	0.7	0.65	0.775	0.65	0.8	0.75	**0.85**	0.675	0.675	0.825
7	0.775	0.7	0.85	0.7	0.825	0.775	**0.9**	0.85	0.75	0.85
8	0.8	0.8	0.925	0.8	0.85	0.85	**0.95**	0.925	0.85	0.925
9	0.875	0.825	0.925	0.8	0.925	0.85	0.95	**0.975**	0.85	0.925
10	0.825	0.875	0.95	0.9	0.925	0.875	**0.975**	**0.975**	0.9	**0.975**
15	0.975	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**

Tablo 4.12 incelendiğinde doğru sınıflandırma olasılığına göre Kiraz veri setinde boyut indirgemede genel olarak en başarılı özellik seçim yöntemleri ReliefF, DkO_{enb} ve F yöntemleri olmuştur. Entropy kriterine göre ilk beş özellik seçiminde F yöntemi daha

başarılı olmuştur. Kiraz veri setinde Kappa istatistiği ve doğru sınıflandırma olasılığı kriterleri benzer sonuçlar vermiştir. Kiraz veri seti için önerilen özellik seçim yöntemlerinin, değişim katsayısı ve ReliefF yöntemine göre sınıflama performansının grafiksel karılaştırılması Şekil 4.5'de verilmiştir.



Şekil 4.5. Önerilen özellik seçim yöntemlerinin seçili yöntemlere göre sınıflandırma performansı (Kiraz)

Phoneme veri seti için seçilen özellik sayısına göre hesaplanan doğru sınıflandırma olasılıkları, entropy ve kappa katsayı değerleri Tablo 4.15, 4-16 ve 4.17'de verilmiştir.

Tablo 4.15. Phonome verisi için hesaplanan doğru sınıflandırma olasılıkları

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	-	**0.687**	-	-	0.605	-	**0.687**	-	-	-
2	0.588	0.749	0.510	0.757	0.712	**0.774**	0.741	0.650	0.646	0.560
3	0.601	0.807	0.683	0.794	0.782	**0.864**	0.823	0.650	0.695	0.551
4	0.613	0.835	0.728	0.798	0.807	**0.918**	0.844	0.691	0.679	0.560
5	0.626	0.860	0.720	0.807	0.835	**0.930**	0.868	0.687	0.708	0.613
6	0.646	0.872	0.720	0.815	0.872	**0.951**	0.881	0.695	0.737	0.658
7	0.704	0.909	0.757	0.827	0.889	**0.963**	0.893	0.737	0.737	0.700
8	0.712	0.922	0.786	0.840	0.897	**0.955**	0.901	0.782	0.749	0.778
9	0.728	**0.934**	0.790	0.885	0.893	0.951	0.926	0.798	0.761	0.782
10	0.770	0.938	0.827	0.905	0.934	**0.963**	0.934	0.815	0.778	0.790
15	0.877	0.963	0.922	0.947	0.951	**0.975**	0.963	0.864	0.856	0.914
20	0.951	0.988	0.967	0.971	0.984	0.984	**0.992**	0.930	0.914	0.947
25	0.979	0.996	0.984	0.992	0.996	0.992	**1.000**	0.959	0.959	0.984
30	0.992	1.000	1.000	0.996	1.000	**1.000**	**1.000**	0.975	0.984	0.992

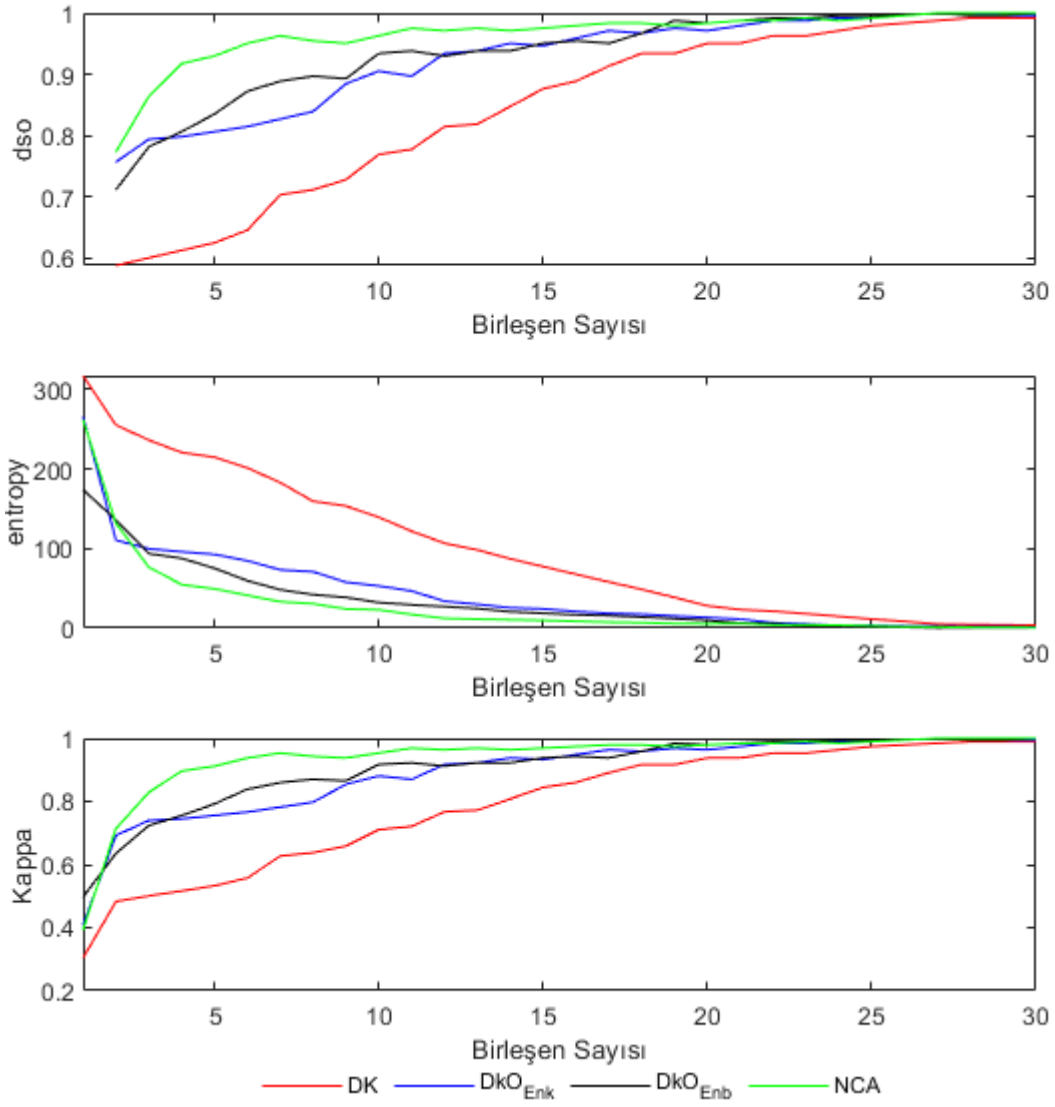
Tablo 4.16. Phonome verisi için hesaplanan Entropy değerleri

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	317.3	181.0	339.2	265.7	**174.1**	262.0	181.0	233.1	233.1	261.5
2	255.3	155.1	262.6	**110.5**	135.9	131.6	133.9	188.3	183.6	244.1
3	236.4	118.6	171.7	99.8	93.6	**76.8**	108.8	171.5	175.3	237.8
4	220.7	104.9	152.2	96.0	87.7	**54.7**	89.0	166.3	169.1	222.9
5	214.9	98.5	138.4	92.7	75.2	**49.3**	78.2	163.6	158.7	183.7
6	201.4	76.1	122.3	84.7	59.6	**41.3**	69.8	148.6	143.2	166.7
7	183.1	64.6	109.9	73.3	48.3	**33.3**	60.8	139.4	134.0	158.1
8	159.5	55.7	90.1	70.8	42.1	**30.7**	52.4	127.2	127.0	120.1
9	153.6	47.1	78.3	57.8	38.6	**24.1**	42.0	119.1	117.5	113.8
10	139.4	39.5	71.1	53.0	32.2	**23.1**	37.7	108.6	111.6	107.4
15	77.4	16.6	20.0	24.2	18.7	**9.7**	15.0	65.0	72.3	63.8
20	28.0	6.1	12.6	13.4	9.7	5.5	**4.6**	32.1	41.2	33.0
25	11.3	2.3	4.0	3.4	2.0	2.9	**1.7**	14.0	13.2	11.6
30	3.7	**0.0**	0.7	1.2	0.7	0.8	0.5	6.0	4.3	3.0

Tablo 4.17. Phonome verisi için hesaplanan Kappa istatistikleri

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	0.304	**0.603**	0.245	0.408	0.497	0.393	**0.603**	0.400	0.400	0.396
2	0.483	0.682	0.374	0.692	0.635	**0.714**	0.671	0.557	0.553	0.443
3	0.501	0.756	0.597	0.740	0.724	**0.829**	0.777	0.558	0.616	0.433
4	0.516	0.792	0.655	0.745	0.755	**0.896**	0.803	0.609	0.596	0.444
5	0.533	0.824	0.645	0.756	0.792	**0.912**	0.834	0.604	0.634	0.512
6	0.557	0.839	0.645	0.766	0.839	**0.938**	0.850	0.617	0.669	0.570
7	0.627	0.886	0.694	0.782	0.860	**0.953**	0.865	0.669	0.669	0.621
8	0.638	0.902	0.730	0.798	0.870	**0.943**	0.876	0.726	0.685	0.721
9	0.658	0.917	0.736	0.854	0.865	**0.938**	0.907	0.747	0.700	0.725
10	0.711	0.922	0.783	0.881	0.917	**0.953**	0.917	0.768	0.721	0.736
15	0.845	0.953	0.902	0.933	0.938	**0.969**	0.953	0.830	0.819	0.891
20	0.938	0.984	0.959	0.964	0.979	0.979	**0.990**	0.912	0.891	0.932
25	0.974	0.995	0.979	0.990	0.995	0.990	**1.000**	0.948	0.948	0.979
30	0.990	**1.000**	**1.000**	0.995	**1.000**	**1.000**	**1.000**	0.969	0.979	0.990

Phoneme veri seti için karşılaştırma kriterleri incelendiğinde boyut indirgemede genel olarak en başarılı özellik seçim yöntemi NCA yöntemi olmuştur. Phoneme veri seti için önerilen özellik seçim yöntemlerinin, değişim katsayısı ve NCA yöntemine göre sınıflama performansının grafiksel karılaştırılması Şekil 4.6'de verilmiştir.



Şekil 4.6. Önerilen özellik seçim yöntemlerinin seçili yöntemlere göre sınıflandırma performansı (Phoneme)

Şekil 4.6 incelendiğinde önerilen özellik seçim yöntemleri DkO_{enk} ve DkO_{enb} yöntemlerinin klasik değişim katsayısına göre daha başarılı olduğu görülmüştür.

Prostate veri seti için seçilen özellik sayısına göre hesaplanan doğru sınıflandırma olasılıkları, entropy ve kappa katsayı değerleri Tablo 4.18, 4-19 ve 4.20'de verilmiştir.

Tablo 4.18. Prostate verisi için hesaplanan doğru sınıflandırma olasılıkları

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	0.559	0.696	**0.735**	0.520	0.559	**0.735**	0.676	**0.735**	**0.735**	**0.735**
2	0.569	0.804	**0.824**	0.529	0.569	0.804	0.627	0.794	**0.824**	**0.824**
3	0.667	0.873	0.853	0.520	0.667	0.814	0.755	**0.882**	0.853	**0.882**
4	0.667	0.863	**0.873**	0.520	0.696	0.814	0.755	0.882	**0.873**	**0.873**
5	0.676	0.892	**0.931**	0.588	0.706	0.824	0.804	0.882	**0.931**	**0.931**
6	0.706	0.892	**0.922**	0.667	0.725	0.892	0.833	0.892	**0.922**	**0.922**
7	0.735	0.902	**0.922**	0.676	0.735	0.912	0.853	0.892	**0.922**	**0.922**
8	0.716	0.912	0.912	0.725	0.716	**0.931**	0.873	0.882	0.912	0.912
9	0.784	0.931	0.902	0.784	0.735	**0.941**	0.873	0.902	0.902	0.902
10	0.824	**0.941**	0.931	0.775	0.765	0.912	0.902	0.912	0.931	0.931
15	0.912	0.961	0.990	0.892	0.892	**1.000**	0.971	0.990	0.990	0.990
20	0.961	0.980	0.990	0.980	0.912	**1.000**	0.990	0.990	0.990	0.990
25	0.990	0.990	0.990	**1.000**	0.971	**1.000**	0.990	0.990	0.990	0.990
30	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.990	**1.000**	**1.000**

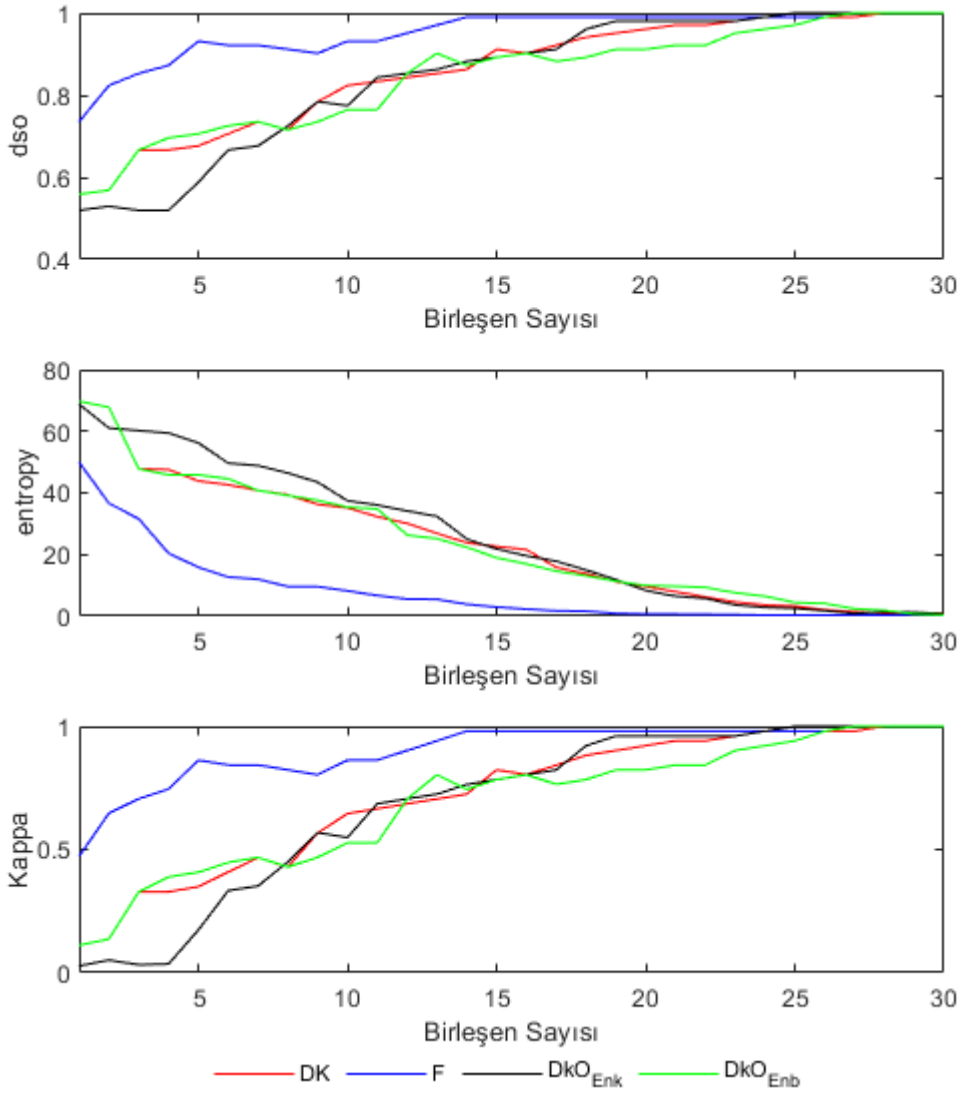
Tablo 4.19. Prostate verisi için hesaplanan Entropy değerleri

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	69.75	56.06	**49.94**	68.75	69.75	49.94	64.54	**49.94**	**49.94**	**49.94**
2	67.81	37.26	36.64	61.04	67.81	40.70	63.10	**29.78**	36.64	36.64
3	47.77	27.57	31.54	60.22	47.77	39.74	47.39	**22.73**	31.54	**22.73**
4	47.56	18.05	20.41	59.42	45.89	37.32	33.93	**16.74**	20.41	20.41
5	43.86	14.48	15.84	56.23	45.85	34.53	31.99	**12.94**	15.84	15.84
6	42.64	**11.74**	12.69	49.63	44.58	31.11	30.43	12.68	12.69	12.69
7	40.79	10.68	11.97	48.85	40.79	23.78	28.70	**10.51**	11.97	11.97
8	39.39	**9.29**	9.62	46.43	39.16	21.38	26.78	10.89	9.62	9.62
9	36.27	**8.53**	9.56	43.47	37.54	17.82	25.06	9.73	9.56	9.56
10	35.15	8.24	**8.19**	37.43	35.37	16.73	20.35	9.71	**8.19**	**8.19**
15	22.63	3.70	**2.93**	21.83	18.96	8.53	7.77	3.46	2.93	2.83
20	9.59	2.12	**0.65**	8.30	9.90	1.17	1.33	0.73	0.65	0.65
25	3.26	0.23	0.29	2.53	4.41	**0.03**	0.72	0.35	0.29	0.18
30	0.93	0.07	0.42	0.74	0.42	**0.00**	0.38	0.30	0.42	0.14

Tablo 4.20. Prostate verisi için hesaplanan Kappa istatistikleri

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	0.111	0.397	**0.472**	0.028	0.111	**0.472**	0.351	**0.472**	**0.472**	**0.472**
2	0.137	0.609	**0.647**	0.051	0.137	0.609	0.252	0.589	**0.647**	**0.647**
3	0.328	0.745	0.706	0.033	0.328	0.628	0.512	**0.765**	0.706	**0.765**
4	0.328	0.726	0.745	0.036	0.389	0.628	0.512	**0.765**	0.745	0.745
5	0.349	0.784	**0.863**	0.174	0.408	0.647	0.608	0.765	**0.863**	**0.863**
6	0.409	0.784	**0.843**	0.333	0.448	0.784	0.667	0.784	**0.843**	**0.843**
7	0.468	0.804	**0.843**	0.352	0.468	0.824	0.706	0.784	**0.843**	**0.843**
8	0.429	0.823	0.823	0.450	0.429	**0.863**	0.745	0.764	0.823	0.823
9	0.567	0.863	0.804	0.568	0.469	**0.882**	0.745	0.804	0.804	0.804
10	0.646	**0.882**	0.863	0.549	0.528	0.824	0.804	0.824	0.863	0.863
15	0.823	0.922	0.980	0.784	0.784	**1.000**	0.941	0.980	0.980	0.980
20	0.922	0.961	0.980	0.961	0.823	**1.000**	0.980	0.980	0.980	0.980
25	0.980	0.980	0.980	**1.000**	0.941	**1.000**	0.980	0.980	0.980	0.980
30	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.980	**1.000**	**1.000**

Prostate veri seti için karşılaştırma kriterleri incelendiğinde boyut indirgemede genel olarak en başarılı özellik seçim yöntemleri F, NCA, Fisher, t skor ve Welch'in t skor yöntemleri olmuştur. Prostate veri seti için önerilen özellik seçim yöntemlerinin, değişim katsayısı ve F yöntemine göre sınıflama performansının grafiksel karılaştırılması Şekil 4.7'de verilmiştir.



Şekil 4.7. Önerilen özellik seçim yöntemlerinin seçili yöntemlere göre sınıflandırma performansı (Prostate)

Şeftali veri seti için seçilen özellik sayısına göre hesaplanan doğru sınıflandırma olasılıkları, entropy ve kappa katsayı değerleri Tablo 4.21, 4-22 ve 4.23'de verilmiştir.

Tablo 4.21. Şeftali verisi için hesaplanan doğru sınıflandırma olasılıkları

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	-	0.500	0.536	**0.548**	0.452	0.417	-	0.452	0.536	0.452
2	0.381	0.560	0.548	0.548	0.500	0.429	0.452	**0.619**	0.548	0.464
3	0.440	0.619	0.536	0.560	0.560	0.524	0.536	**0.714**	0.536	0.524
4	0.476	0.619	**0.714**	0.548	0.560	0.560	0.583	**0.714**	0.548	0.560
5	0.488	0.619	**0.726**	0.500	0.595	0.667	0.595	**0.726**	0.607	0.595
6	0.512	0.690	**0.750**	0.607	0.655	0.679	0.655	**0.750**	0.631	0.690
7	0.548	0.702	0.726	0.655	**0.750**	**0.750**	0.726	0.726	0.690	0.762
8	0.583	0.762	0.750	0.690	0.750	0.774	0.774	0.750	0.655	**0.833**
9	0.619	**0.881**	0.762	0.810	0.845	0.869	0.810	0.762	0.679	0.845
10	0.655	**0.952**	0.833	0.798	0.881	0.869	0.833	0.833	0.714	0.869
15	0.893	**0.988**	0.976	0.964	0.964	0.940	0.976	0.976	0.917	0.976
20	0.988	**1.000**	**1.000**	0.988	**1.000**	0.988	**1.000**	**1.000**	**1.000**	**1.000**

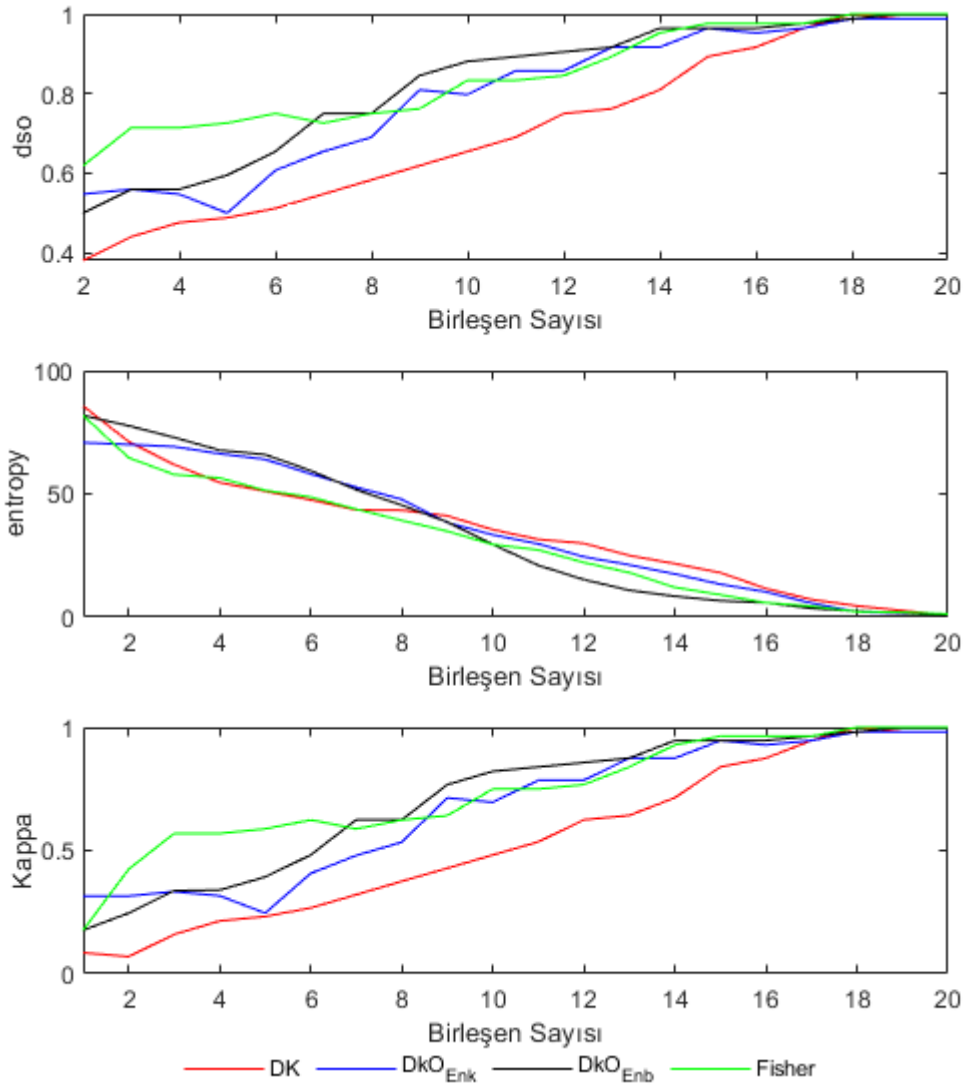
Tablo 4.22. Şeftali verisi için hesaplanan Entropy değerleri

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	85.64	81.74	70.83	**70.81**	81.79	91.60	85.83	81.79	70.83	81.79
2	71.23	75.75	69.21	70.03	77.64	87.52	81.03	**64.68**	69.21	76.66
3	61.89	72.34	68.87	69.17	72.91	73.88	76.76	**57.85**	68.87	73.86
4	**54.62**	67.75	56.46	66.30	67.71	72.17	73.60	56.46	66.75	67.71
5	**51.07**	63.64	51.39	64.00	65.96	69.53	67.47	51.39	62.56	65.96
6	**47.57**	51.45	48.71	58.08	59.41	58.37	57.64	48.71	57.68	53.39
7	43.50	46.34	43.76	52.68	51.73	**40.43**	53.01	43.76	53.24	42.43
8	43.34	**37.15**	39.23	47.79	45.41	38.07	43.83	39.23	47.38	35.43
9	41.10	33.37	34.85	38.34	38.59	**31.66**	36.27	34.85	40.19	29.80
10	35.48	25.44	29.46	33.39	29.47	25.73	30.77	29.46	38.03	**24.74**
15	18.00	8.12	9.08	13.33	6.61	10.32	8.91	9.08	12.05	**6.55**
20	0.86	0.82	1.01	1.00	0.65	0.91	**0.33**	1.26	1.26	0.64

Tablo 4.23. Şeftali verisi için hesaplanan Kappa istatistikleri

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	0.085	0.246	0.296	**0.315**	0.177	0.117	0.260	0.177	0.296	0.177
2	0.069	0.336	0.315	0.315	0.246	0.140	0.172	**0.424**	0.315	0.194
3	0.160	0.426	0.297	0.334	0.337	0.286	0.301	**0.569**	0.297	0.284
4	0.214	0.427	**0.569**	0.316	0.340	0.337	0.374	**0.569**	0.317	0.340
5	0.232	0.428	**0.588**	0.246	0.393	0.499	0.392	**0.588**	0.407	0.393
6	0.267	0.535	**0.624**	0.407	0.482	0.517	0.481	**0.624**	0.444	0.535
7	0.321	0.553	0.588	0.480	**0.625**	**0.625**	0.588	0.588	0.533	0.642
8	0.375	0.642	0.624	0.534	0.625	0.660	0.661	0.624	0.480	**0.750**
9	0.428	0.821	0.642	0.713	0.767	**0.804**	0.714	0.642	0.516	0.768
10	0.482	**0.929**	0.750	0.696	0.821	0.804	0.750	0.750	0.571	0.803
15	0.839	**0.982**	0.964	0.946	0.946	0.911	0.964	0.964	0.875	0.964
20	0.982	**1.000**	**1.000**	0.982	**1.000**	0.982	**1.000**	**1.000**	**1.000**	**1.000**

Şeftali veri seti için doğru sınıflandırma olasılığına göre boyut indirgemede özellikle ilk 6 özellik seçiminde Fisher yöntemi diğer özellik seçim yöntemlerine göre daha başarılı olmuştur. Şeftali veri seti için önerilen özellik seçim yöntemlerinin, değişim katsayısı ve Fisher yöntemine göre sınıflama performansının grafiksel karılaştırılması Şekil 4.8'de verilmiştir.



Şekil 4.8. Önerilen özellik seçim yöntemlerinin seçili yöntemlere göre sınıflandırma performansı (Şeftali)

Şekil 4.8 incelendiğinde önerilen özellik seçim yöntemleri DkO_{enk} ve DkO_{enb} yöntemlerinin klasik değişim katsayısına göre daha başarılı olduğu görülmüştür.

Şeftali Bahçe veri seti için seçilen özellik sayısına göre hesaplanan doğru sınıflandırma olasılıkları, entropy ve kappa katsayı değerleri Tablo 4.24, 4-25 ve 4.26'da verilmiştir.

Tablo 4.24. Şeftali Bahçe verisi için hesaplanan doğru sınıflandırma olasılıkları

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	-	0.246	0.296	**0.315**	0.177	0.117	0.260	0.177	0.296	0.177
2	0.069	0.336	0.315	0.315	0.246	0.140	0.172	**0.424**	0.315	0.194
3	0.160	0.426	0.297	0.334	0.337	0.286	0.301	**0.569**	0.297	0.284
4	0.214	0.427	**0.569**	0.316	0.340	0.337	0.374	**0.569**	0.317	0.340
5	0.232	0.428	**0.588**	0.246	0.393	0.499	0.392	**0.588**	0.407	0.393
6	0.267	0.535	**0.624**	0.407	0.482	0.517	0.481	**0.624**	0.444	0.535
7	0.321	0.553	0.588	0.480	0.625	0.625	0.588	0.588	0.533	**0.642**
8	0.375	0.642	0.624	0.534	0.625	0.660	0.661	0.624	0.480	**0.750**
9	0.428	**0.821**	0.642	0.713	0.767	0.804	0.714	0.642	0.516	0.768
10	0.482	**0.929**	0.750	0.696	0.821	0.804	0.750	0.750	0.571	0.803
15	0.839	**0.982**	0.964	0.946	0.946	0.911	0.964	0.964	0.875	0.964

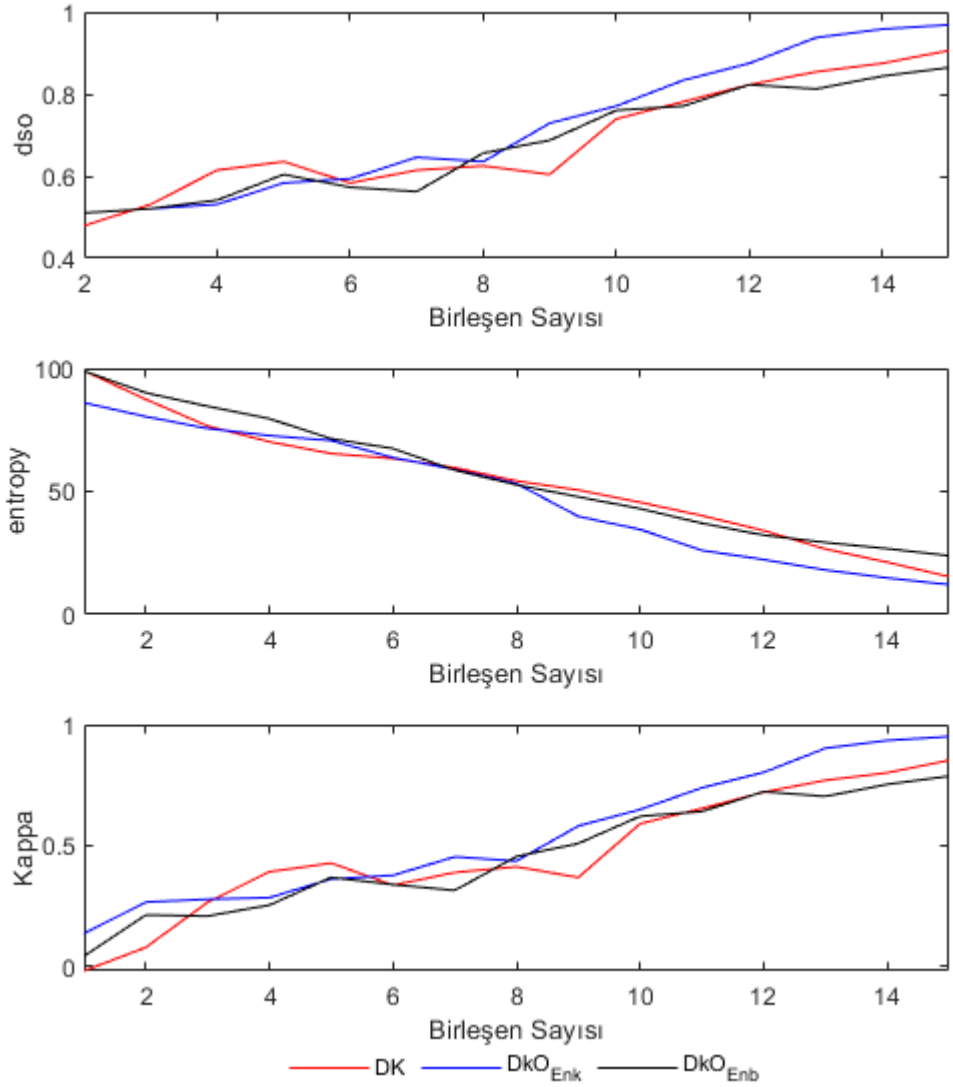
Tablo 4.25. Şeftali Bahçe verisi için hesaplanan Entropy Değerleri

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	**98.98**	93.64	94.32	86.01	98.83	98.11	94.67	96.78	96.78	96.78
2	87.40	81.19	83.41	80.40	90.06	**91.65**	86.86	91.64	91.64	91.64
3	76.57	78.89	80.77	75.58	84.58	84.63	82.22	**87.37**	**87.37**	**87.37**
4	70.17	74.42	73.85	72.68	79.54	81.05	74.97	**83.74**	**83.74**	**83.74**
5	65.39	69.27	71.55	70.65	71.40	78.54	66.43	**81.67**	**81.67**	**81.67**
6	63.39	58.39	67.28	63.75	67.42	**73.32**	61.42	71.24	71.24	71.24
7	59.66	55.78	61.85	58.90	58.64	65.00	54.64	**65.92**	**65.92**	**65.92**
8	54.25	51.55	57.30	53.21	52.69	**58.51**	43.40	52.79	52.79	52.79
9	**50.56**	46.46	48.10	39.85	47.80	47.00	36.89	47.68	47.68	47.68
10	**45.55**	36.94	44.61	34.58	43.00	37.36	32.04	41.68	41.68	41.40
15	15.39	10.83	13.08	12.13	**23.90**	12.97	10.76	16.10	16.10	13.01

Tablo 4.26. Şeftali Bahçe verisi için hesaplanan Kappa istatistikleri

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	-0.017	**0.232**	0.147	0.137	0.044	-	0.189	0.060	0.060	0.060
2	0.080	0.256	0.191	**0.267**	0.214	0.228	0.200	0.025	0.025	0.025
3	0.267	0.250	**0.359**	0.279	0.209	0.228	0.244	0.073	0.073	0.073
4	**0.392**	0.289	0.362	0.286	0.255	0.300	0.296	0.181	0.181	0.181
5	0.428	0.408	**0.452**	0.363	0.369	0.345	0.419	0.278	0.278	0.278
6	0.337	0.388	0.441	0.378	0.340	**0.471**	0.467	0.431	0.431	0.431
7	0.389	0.486	0.422	0.454	0.316	**0.555**	0.534	0.428	0.428	0.428
8	0.413	0.528	0.510	0.437	0.456	**0.619**	0.541	0.406	0.406	0.406
9	0.369	0.628	0.556	0.582	0.509	0.647	**0.703**	0.490	0.490	0.490
10	0.591	**0.691**	0.639	0.649	0.622	0.689	0.656	0.584	0.584	0.537
15	0.853	**0.951**	0.901	**0.951**	0.787	**0.951**	0.935	0.851	0.851	0.901

Tablo 4.24 incelendiğinde doğru sınıflandırma olasılığına göre Şeftali Bahçe veri setinde boyut indirgemede genel olarak en başarılı özellik seçim yöntemleri küme merkezine uzaklık, F ve Fisher skor yöntemleri olmuştur. Şeftali Bahçe veri setinde Entropy kriterine göre Fisher, t skor ve Welch'in t skor yöntemleri için aynı değerler elde edilmiştir. Şeftali Bahçe veri seti için önerilen özellik seçim yöntemlerinin değişim katsayısı yöntemine göre sınıflama performansının grafiksel karılaştırılması Şekil 4.9'da verilmiştir.



Şekil 4.9. Önerilen özellik seçim yöntemlerinin seçili yöntemlere göre sınıflandırma performansı (Şeftali Bahçe)

Şekerpancarı veri seti için seçilen özellik sayısına göre hesaplanan doğru sınıflandırma olasılıkları, entropy ve kappa katsayı değerleri Tablo 4.27, 4-28 ve 4.29'da verilmiştir.

Tablo 4.27.Şekerpancarı verisi için hesaplanan doğru sınıflandırma olasılıkları

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	0.361	0.875	0.875	**0.889**	0.847	0.875	0.861	0.861	0.861	0.861
2	0.361	0.847	0.847	0.847	0.847	0.847	0.847	**0.875**	**0.875**	**0.875**
3	0.375	**0.875**	**0.875**	0.861	0.889	0.847	0.833	0.861	0.861	0.861
4	0.417	0.875	0.875	0.931	0.903	0.861	0.875	**0.903**	**0.903**	**0.903**
5	0.444	0.903	0.903	**0.972**	0.903	0.903	0.931	0.903	0.944	0.903
6	0.597	0.931	0.903	**0.986**	0.944	0.931	0.944	0.931	0.931	0.917
7	0.611	0.986	0.931	**1.000**	0.944	0.944	0.931	0.944	0.972	0.944
8	0.625	0.972	0.958	**1.000**	0.972	0.986	0.972	0.958	0.958	0.958
9	0.708	**1.000**	0.931	**1.000**	0.958	0.986	0.986	0.944	0.972	0.944
10	0.708	**0.986**	0.972	**0.986**	0.972	0.972	0.972	0.972	0.972	0.972
15	0.931	**1.000**	0.986	**1.000**	0.986	**1.000**	0.986	**1.000**	**1.000**	**1.000**

Tablo 4.28. Şekerpancarı verisi için hesaplanan Entropy Değerleri

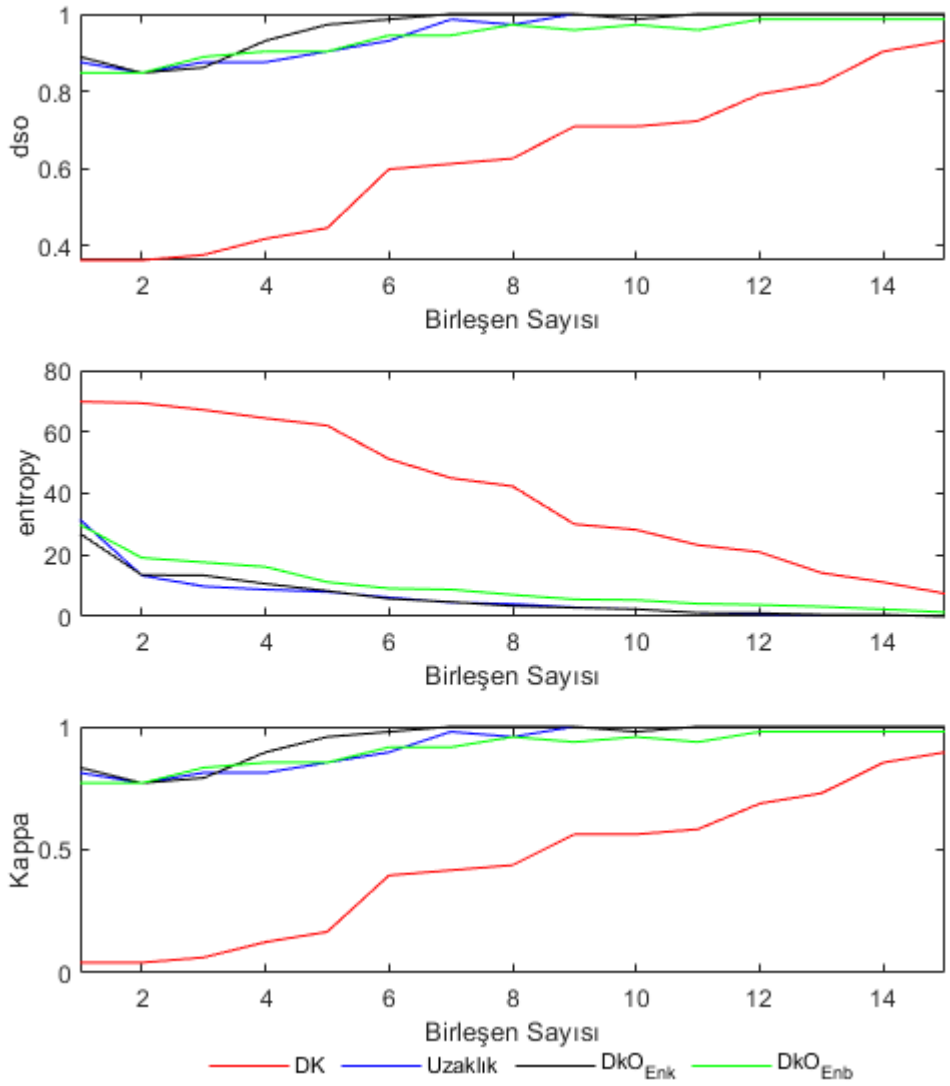
d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	69.81	31.49	26.07	26.89	29.83	31.49	28.30	**24.61**	**24.61**	**24.61**
2	69.35	**13.31**	14.45	13.50	19.09	**13.31**	19.09	19.37	19.37	19.37
3	67.14	**9.80**	14.98	13.38	17.62	12.61	18.37	16.81	16.81	16.35
4	64.44	**8.77**	14.21	10.66	16.14	10.83	16.53	15.01	15.01	15.01
5	62.06	**8.03**	12.70	8.32	11.20	**8.03**	14.66	14.13	11.68	14.13
6	51.19	6.21	12.13	**5.79**	9.11	6.21	12.90	11.50	11.50	12.22
7	45.01	**4.57**	10.65	4.75	8.80	5.50	11.22	8.20	9.39	8.20
8	42.33	4.08	8.80	**3.44**	7.06	4.16	8.10	6.48	6.48	6.48
9	30.02	2.97	6.67	2.89	5.60	**2.51**	7.64	5.40	4.80	5.40
10	28.21	**2.41**	4.76	**2.41**	5.35	2.93	4.95	3.86	3.86	3.86
15	7.45	**0.02**	1.34	**0.02**	1.37	0.52	1.34	0.49	0.49	0.49

Tablo 4.29. Şekerpancarı verisi için hesaplanan Kappa istatistikleri

d	DK	Uzaklık	F	DkO_{enk}	DkO_{enb}	NCA	ReliefF	Fisher	tskor	Welch
1	0.042	0.813	0.813	**0.833**	0.771	0.813	0.792	0.792	0.792	0.792
2	0.042	0.771	0.771	0.771	0.771	0.771	0.771	**0.813**	**0.813**	**0.813**
3	0.063	0.813	0.813	0.792	**0.833**	0.771	0.750	0.792	0.792	0.792
4	0.125	0.813	0.813	**0.896**	0.854	0.792	0.813	0.854	0.854	0.854
5	0.167	0.854	0.854	**0.958**	0.854	0.854	0.896	0.854	0.917	0.854
6	0.396	0.896	0.854	**0.979**	0.917	0.896	0.917	0.896	0.896	0.875
7	0.417	0.979	0.896	**1.000**	0.917	0.917	0.896	0.917	0.958	0.917
8	0.438	0.958	0.938	**1.000**	0.958	0.979	0.958	0.938	0.938	0.938
9	0.563	**1.000**	0.896	**1.000**	0.938	0.979	0.979	0.917	0.958	0.917
10	0.563	**0.979**	0.958	**0.979**	0.958	0.958	0.958	0.958	0.958	0.958
15	0.896	**1.000**	0.979	**1.000**	0.979	**1.000**	0.979	**1.000**	**1.000**	**1.000**

Şekerpancarı veri seti için doğru sınıflandırma olasılığı ve Kappa katsayısına göre boyut indirgemede genel olarak en başarılı özellik seçim yöntemi önerilen özellik seçim yöntemi DkO_{enk} yöntemi olmuştur. Şekerpancarı veri setinde entropy kriterine

göre boyut indirgemede daha başarılı olan yöntemler küme merkezine uzaklık ve önerilen özellik seçim yöntemi DkO_{enk} yöntemi olmuştur. Şekerpancarı veri seti için önerilen özellik seçim yöntemlerinin, değişim katsayısı ve küme merkezine uzaklık yöntemine göre sınıflama performansının grafiksel karılaştırılması Şekil 4.10'da verilmiştir.



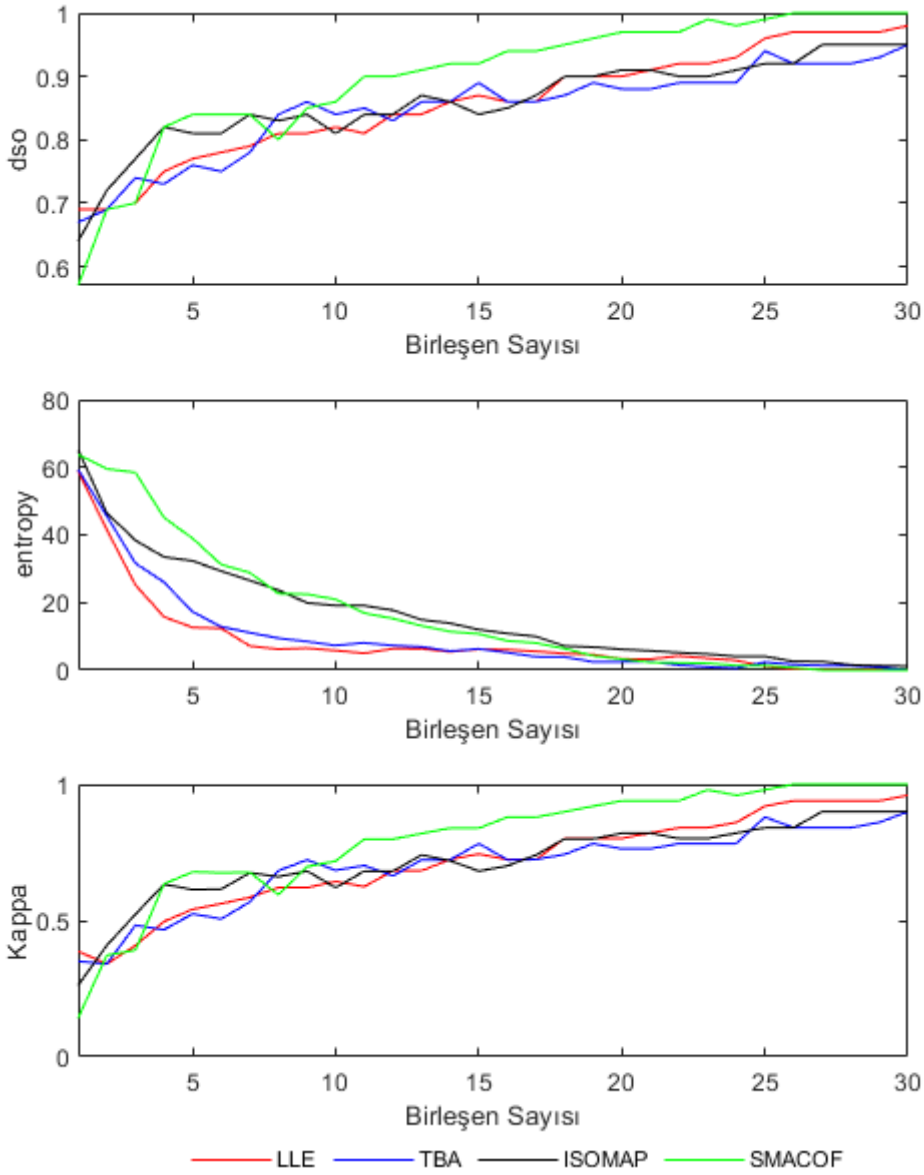
Şekil 4.10. Önerilen özellik seçim yöntemlerinin seçili yöntemlere göre sınıflandırma performansı (Şekerpancarı)

Şekil 4.10 incelendiğinde önerilen özellik seçim yöntemleri DkO_{enk} ve DkO_{enb} yöntemlerinin değişim katsayısına göre boyut indirgemede daha başarılı olduğu görülmüştür.

4.5. Özellik Çıkarma Yöntemlerin Karşılaştırılması

Çalışmanın bu bölümünde incelenen veri setleri için özellik çıkarma yöntemleri LLE, TBA, ISOMAP ve SMACOF yöntemlerinin karesel diskriminant analizindeki sınıflama performansları doğru sınıflandırma olasılığı, entropy ve kappa katsayısına göre karşılaştırılmıştır.

Arcene veri seti için bileşen sayısına göre özellik çıkarma yöntemleri LLE, TBA, ISOMAP ve SMACOF yöntemlerinin karesel diskriminant analizindeki sınıflama performanslarının grafiksel karşılaştırması Şekil 4.11'de verilmiştir.

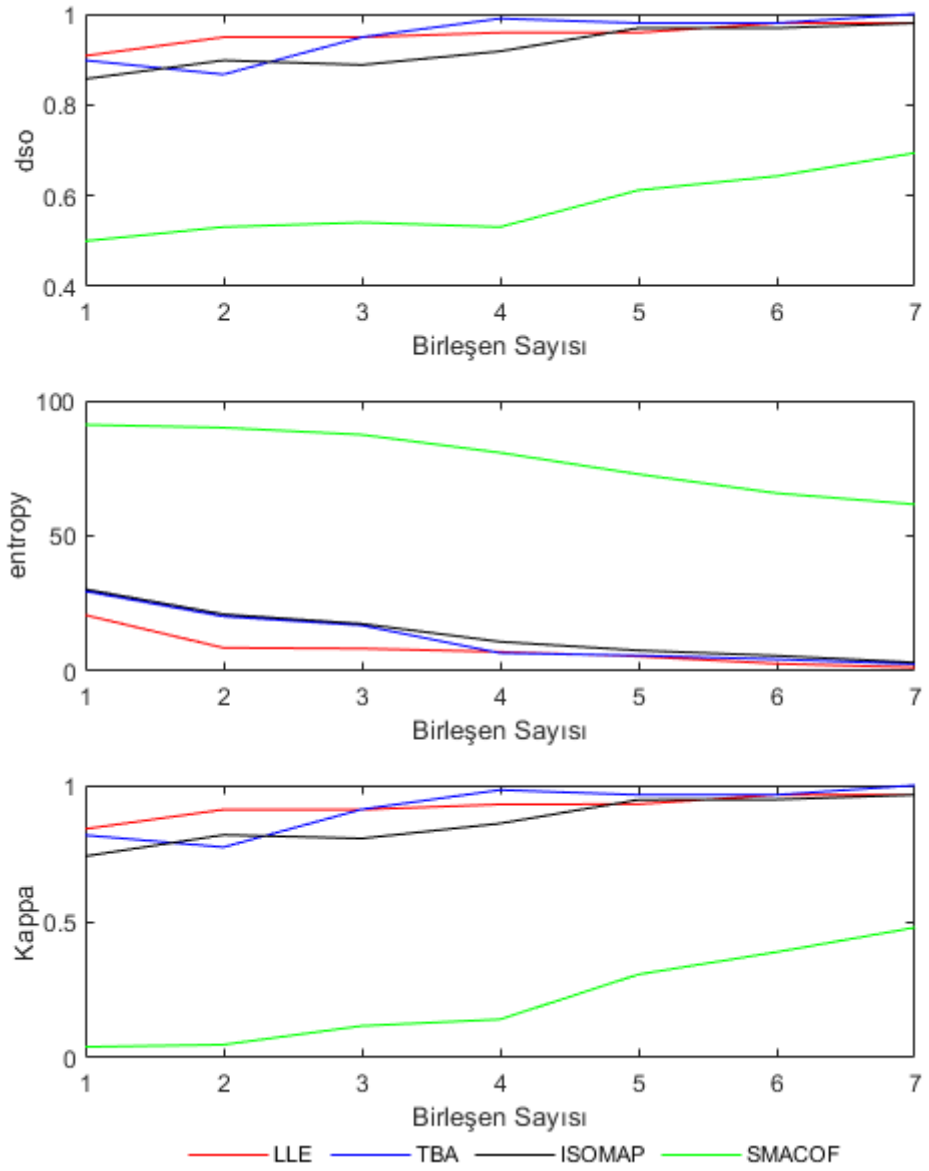


Şekil 4.11. Özellik çıkarma yöntemlerinin sınıflandırma performansları (Arcene)

Şekil 4.11 incelendiğinde Arcene veri setinde doğru sınıflandırma olasılık değerlerine göre SMACOF algoritması diğer özellik çıkarma yöntemlerinden daha

başarılı olduğu görülmektedir. Entropy kriterine göre özellikle ilk 15 birleşen sayısı için LLE yönteminin daha başarılı sonuçlar verdiği belirlenmiştir. Arcene veri seti için doğru sınıflandırma olasılığı ve kappa katsayısı kriterlerinin sonuçları benzerdir.

Breast veri seti için bileşen sayısına göre özellik çıkarma yöntemleri LLE, TBA, ISOMAP ve SMACOF yöntemlerinin karesel diskiriminant analizindeki sınıflama performanslarının grafiksel karşılaştırması Şekil 4.12'de verilmiştir.

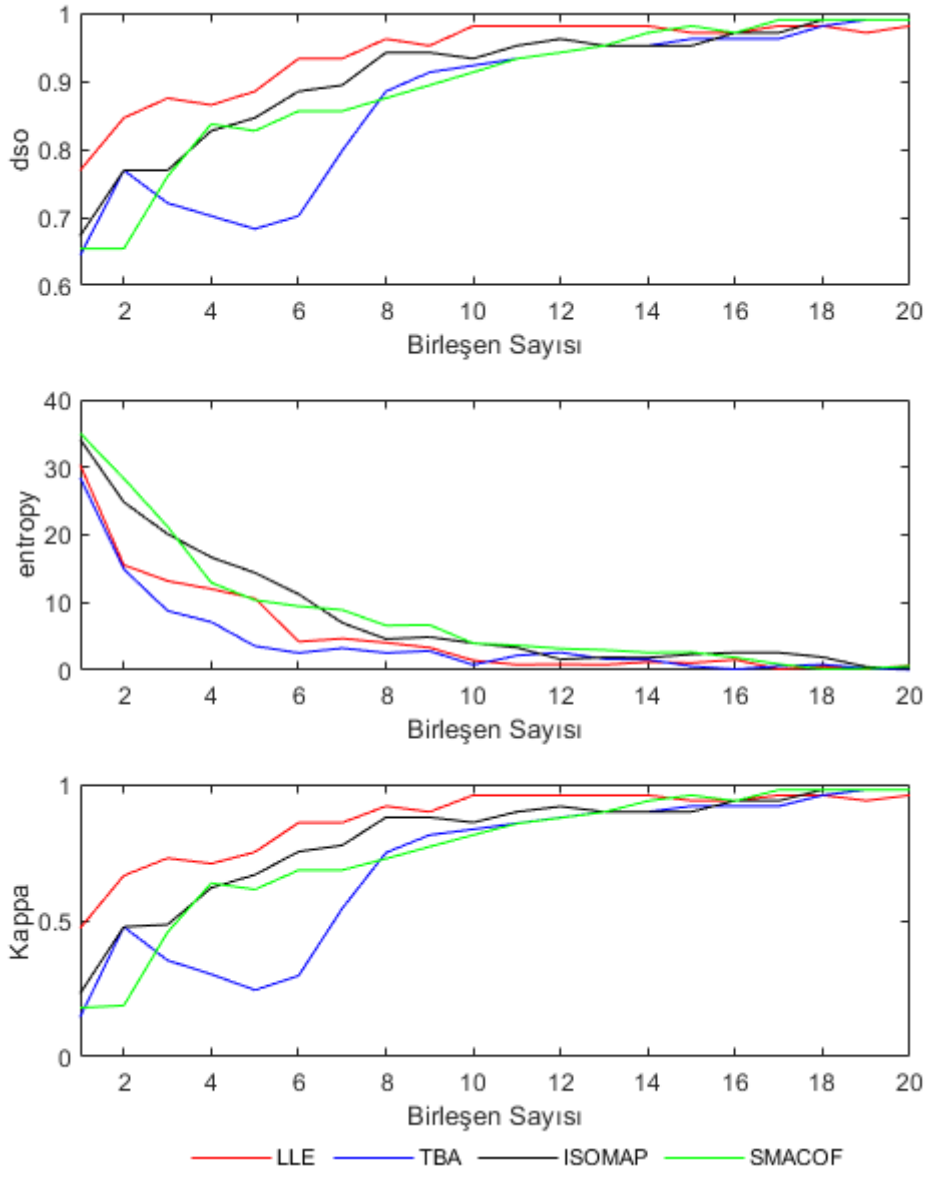


Şekil 4.12. Özellik çıkarma yöntemlerinin sınıflandırma performansları (Breast)

Şekil 4.12 incelendiğinde Breast veri setinde doğru sınıflandırma olasılık değerlerine göre LLE ve TBA yöntemlerinin diğer özellik çıkarma yöntemlerinden daha başarılı olduğu görülmektedir. Entropy kriterine göre LLE yönteminin daha başarılı

sonuçlar verdiği belirlenmiştir. Breast veri setinde SMACOF yöntemi tüm karşılaştırma kriterlerine göre diğer yöntemlere göre daha başarısız olmuştur.

Chowdary veri seti için bileşen sayısına göre özellik çıkarma yöntemlerinin karesel diskiriminant analizindeki sınıflama performanslarının grafiksel karşılaştırması Şekil 4.13'de verilmiştir.

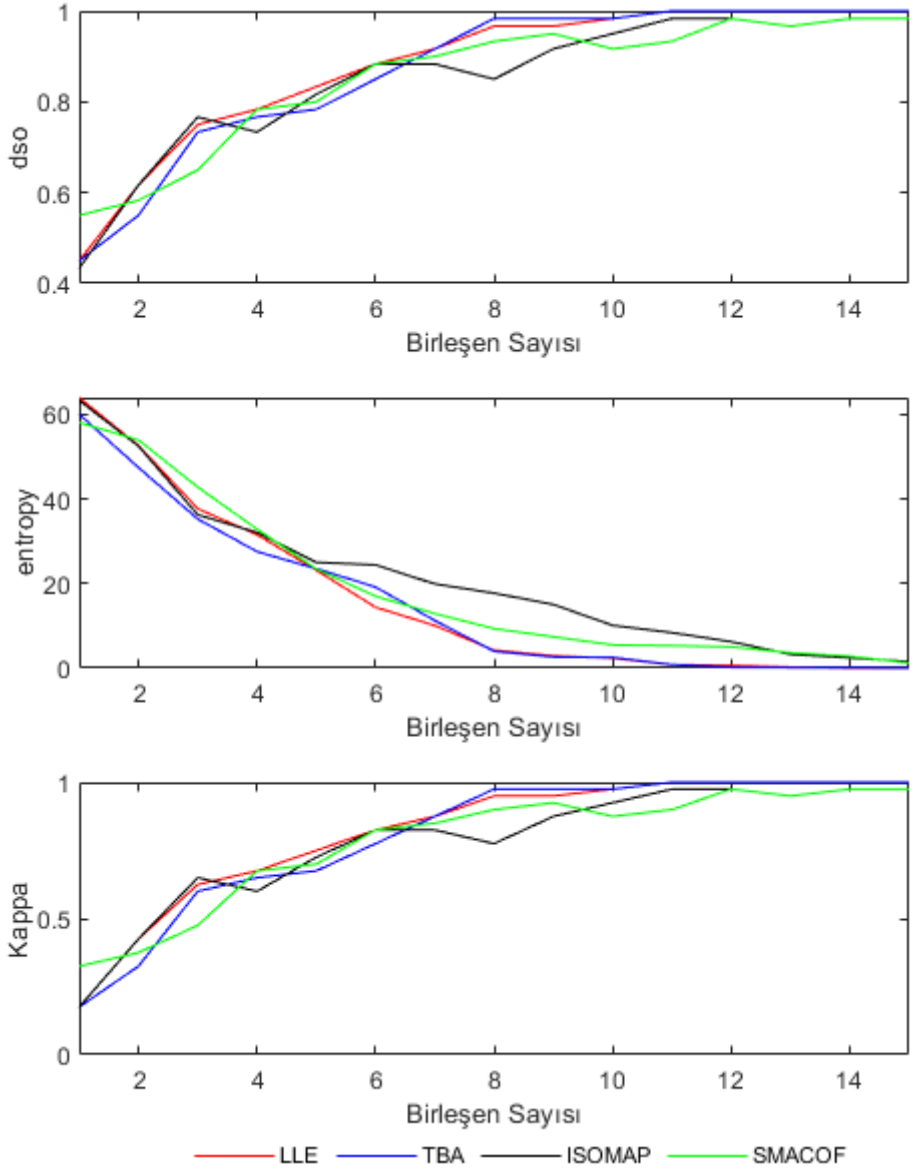


Şekil 4.13. Özellik çıkarma yöntemlerinin sınıflandırma performansları (Chowdary)

Şekil 4.13 incelendiğinde Chowdary veri setinde doğru sınıflandırma olasılığı ve kappa istatistiğine göre LLE yönteminin daha başarılı olduğu görülmektedir. Entropy kriterine göre TBA yönteminin daha başarılı sonuçlar verdiği belirlenmiştir. Breast veri

setinde SMACOF yöntemi tüm karşılaştırma kriterlerine göre diğer yöntemlere göre daha az başarılı olmuştur.

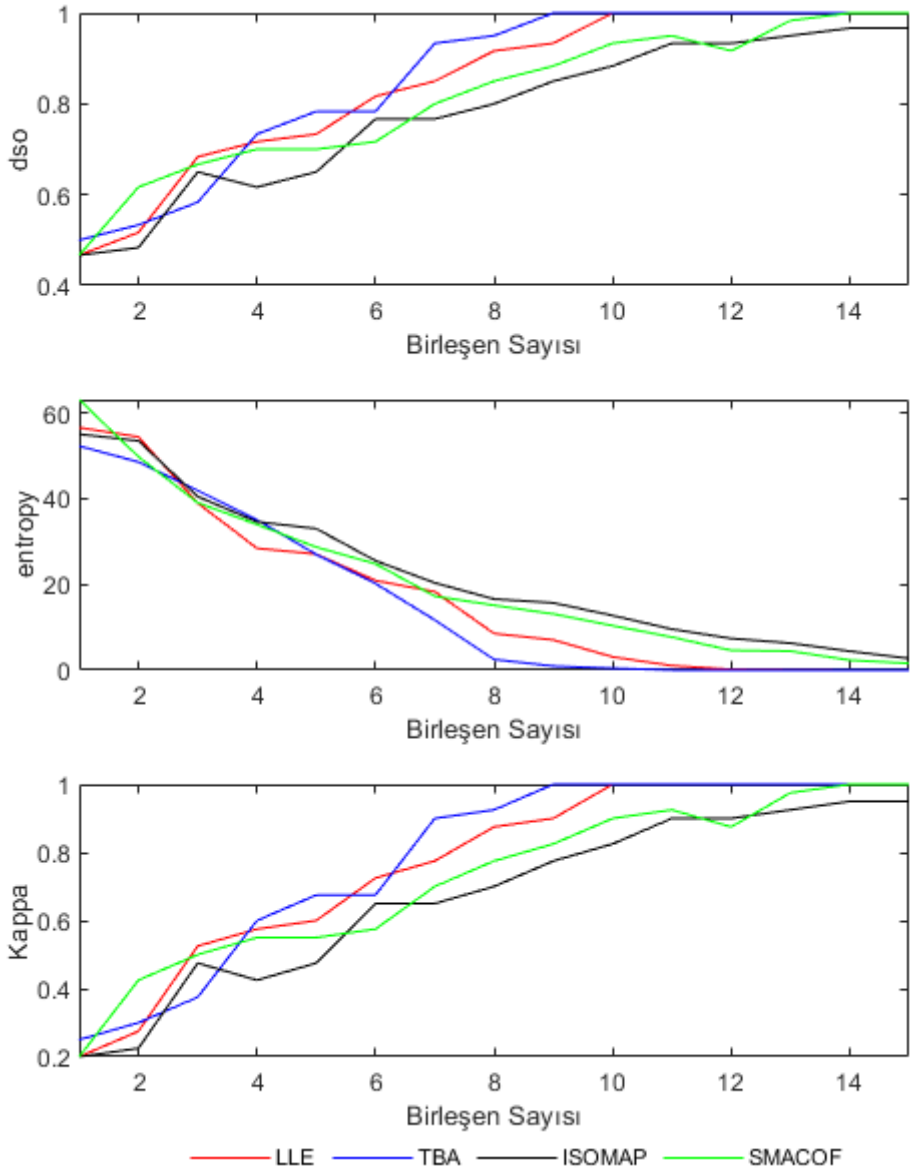
Elma veri seti için bileşen sayısına göre özellik çıkarma yöntemlerinin karesel diskiriminant analizindeki sınıflama performanslarının grafiksel karşılaştırması Şekil 4.14'de verilmiştir.



Şekil 4.14. Özellik çıkarma yöntemlerinin sınıflandırma performansları (Elma)

Şekil 4.14 incelendiğinde Elma veri setinde ele alınan karşılaştırma kriterlerine göre yöntemlerin performanslarının birbirine çok yakın olduğu görülmüştür. Entropy kriterine göre ilk 5 bileşen için boyut indirgemede TBA yöntemi daha başarılı olmuştur.

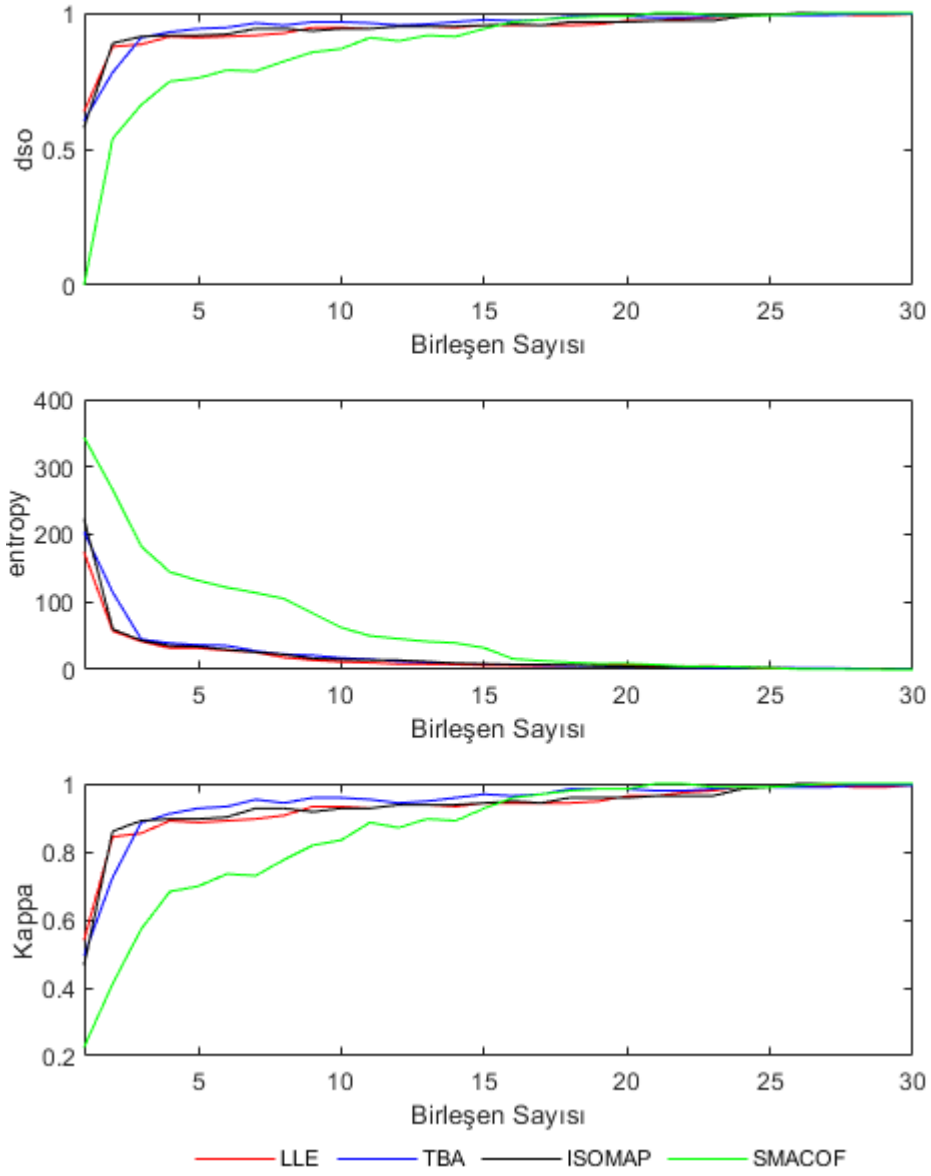
Kiraz veri seti için bileşen sayısına göre özellik çıkarma yöntemlerinin karesel diskiriminant analizindeki sınıflama performanslarının grafiksel karşılaştırması Şekil 4.15'de verilmiştir.



Şekil 4.15. Özellik çıkarma yöntemlerinin sınıflandırma performansları (Kiraz)

Şekil 4.15 incelendiğinde Kiraz veri setinde karşılaştırma kriterlerine göre bileşen sayısının 4 ve üzeri olduğu durumlarda TBA yönteminin diğer özellik çıkarma yöntemlerine göre daha başarılı olduğu görülmüştür.

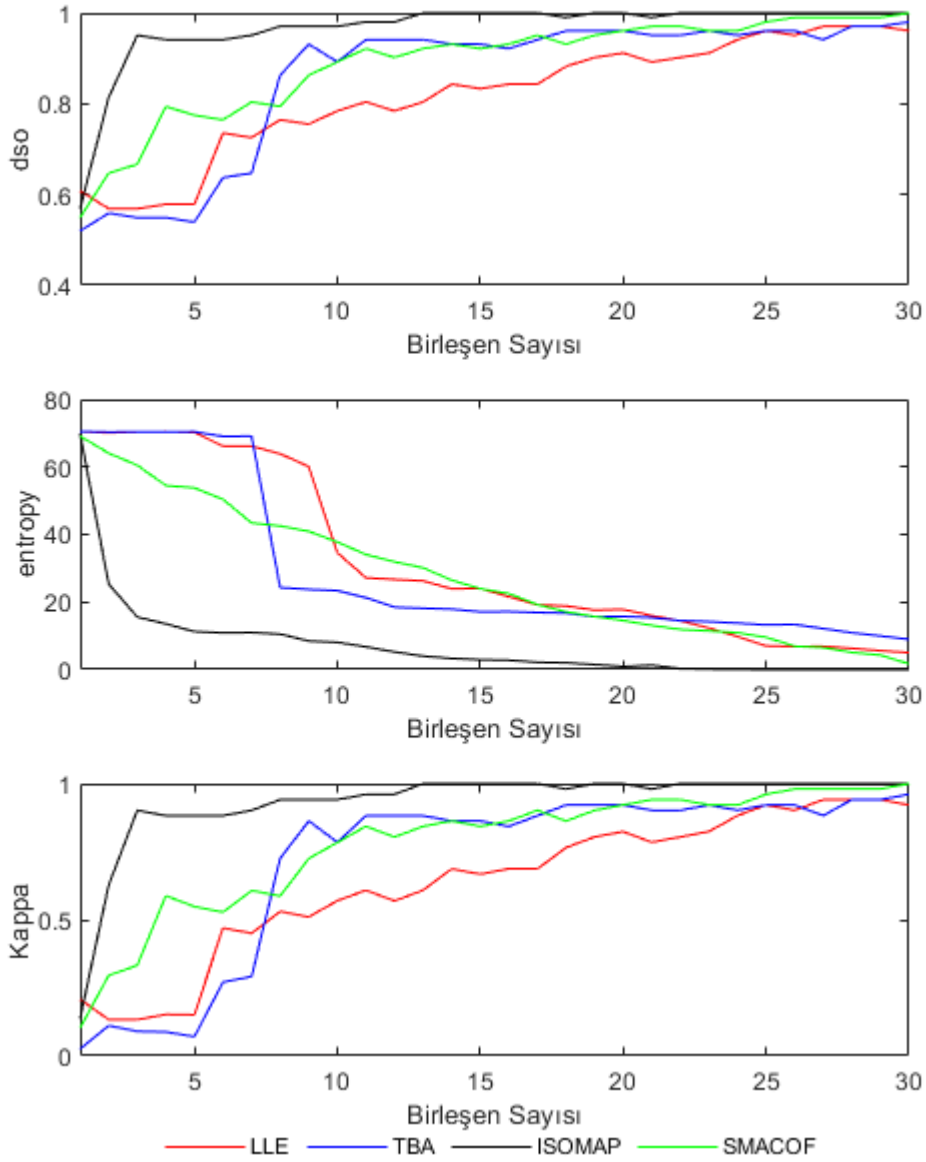
Phoneme veri seti için bileşen sayısına göre özellik çıkarma yöntemlerinin karesel diskiriminant analizindeki sınıflama performanslarının grafiksel karşılaştırması Şekil 4.16'de verilmiştir.



Şekil 4.16. Özellik çıkarma yöntemlerinin sınıflandırma performansları (Phoneme)

Şekil 4.16 incelendiğinde Phoneme veri setinde karşılaştırma kriterlerine göre LLE, TBA ve ISOMAP yöntemlerinin sınıflama performanslarının birbirine yakın olduğu görülmüştür. Phoneme veri setinde doğru sınıflandırma olasılığına göre TBA yöntemi, entropy kriterine göre de LLE yöntemi diğer yöntemlere göre daha başarılı olmuştur.

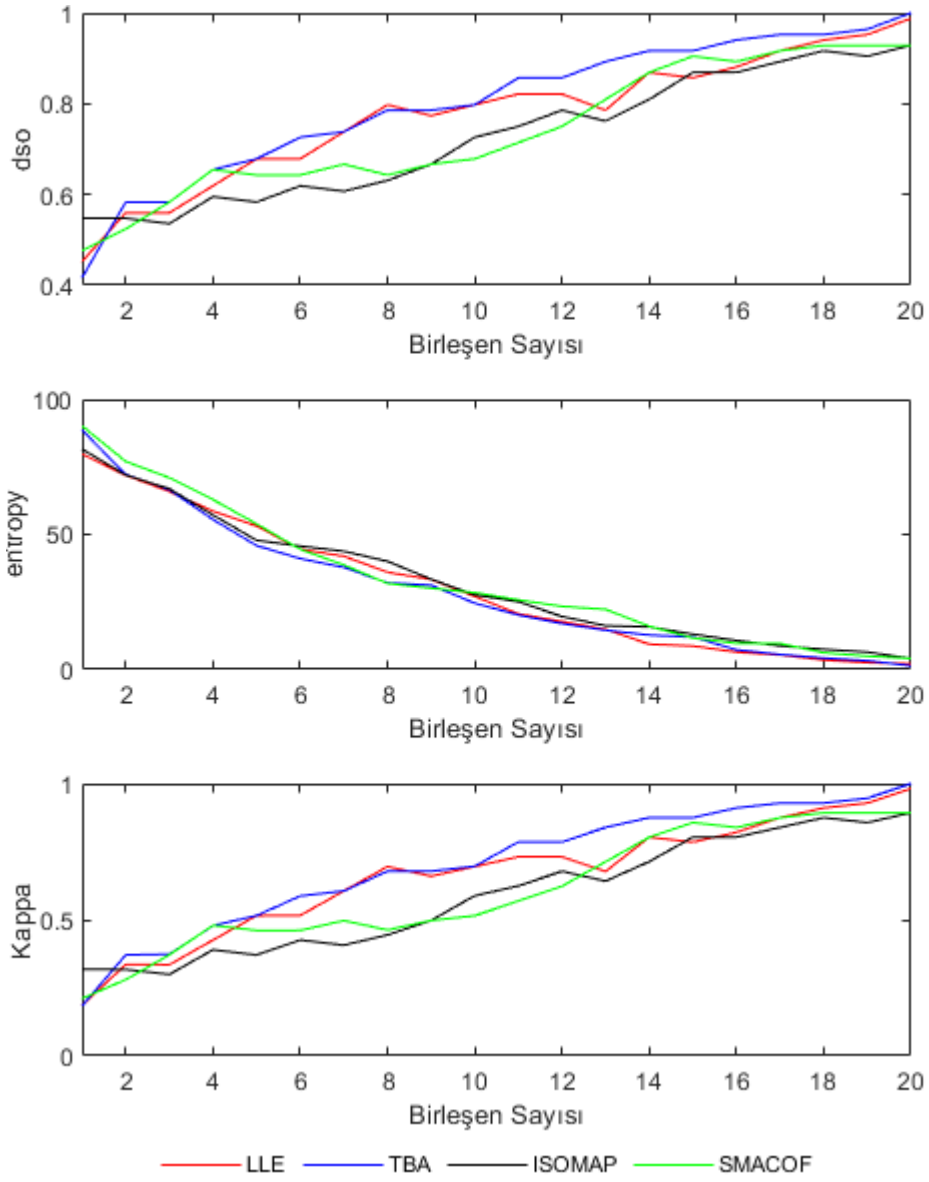
Prostate veri seti için bileşen sayısına göre özellik çıkarma yöntemlerinin karesel diskiriminant analizindeki sınıflama performanslarının grafiksel karşılaştırması Şekil 4.17'de verilmiştir.



Şekil 4.17. Özellik çıkarma yöntemlerinin sınıflandırma performansları (Prostate)

Şekil 4.17 incelendiğinde Prostate veri setinde tüm karşılaştırma kriterlerine göre ISOMAP yöntemi boyut indirgemede diğer yöntemlere göre daha başarılı olmuştur.

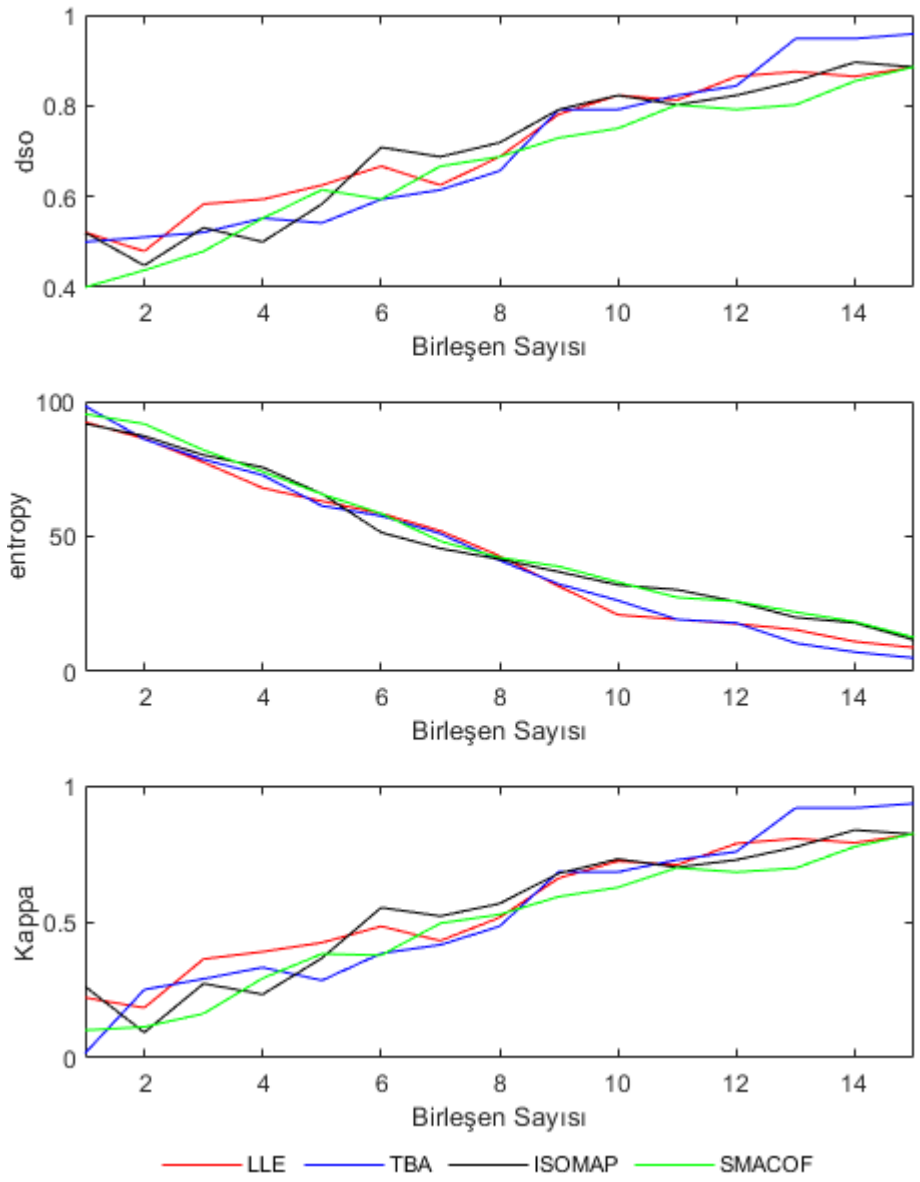
Şeftali veri seti için bileşen sayısına göre özellik çıkarma yöntemlerinin karesel diskiriminant analizindeki sınıflama performanslarının grafiksel karşılaştırması Şekil 4.18'de verilmiştir.



Şekil 4.18. Özellik çıkarma yöntemlerinin sınıflandırma performansları (Şeftali)

Şekil 4.18 incelendiğinde Şeftali veri setinde doğru sınıflandırma olasılığı ve kappa istatistiği kriterlerine göre TBA yönteminin boyut indirgemedeki sınıflama performansı diğer yöntemlere göre daha başarılı olmuştur. Entropy kriterine genel olarak LLE ve TBA yöntemleri diğer yöntemlere göre daha başarılı olmuştur.

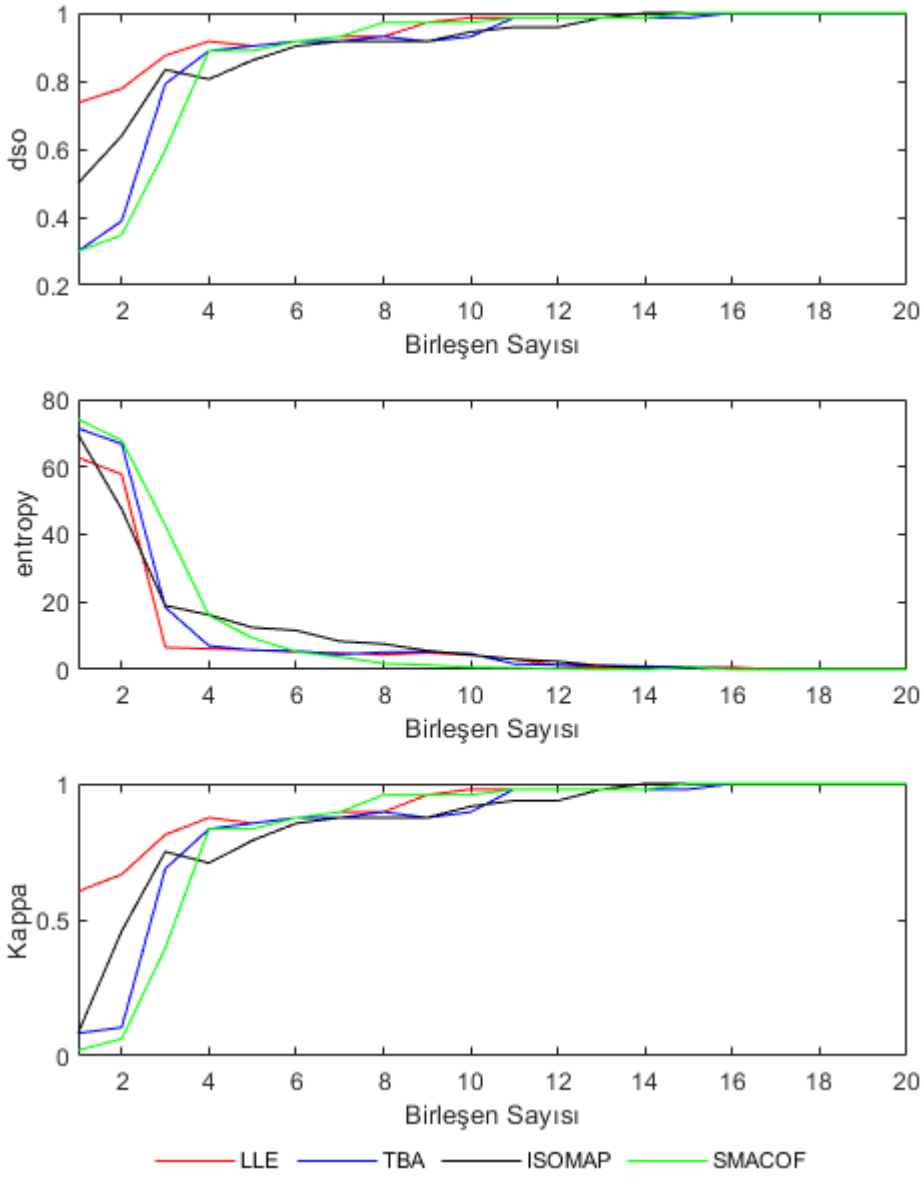
Şeftali Bahçe veri seti için bileşen sayısına göre özellik çıkarma yöntemlerinin karesel diskriminant analizindeki sınıflama performanslarının grafiksel karşılaştırması Şekil 4.19'da verilmiştir.



Şekil 4.19. Özellik çıkarma yöntemlerinin sınıflandırma performansları (Şeftali Bahçe)

Şekil 4.19 incelendiğinde Şeftali Bahçe veri setinde tüm karşılaştırma kriterlerine göre LLE, TBA ve ISOMAP yöntemleri SMACOF yöntemine göre daha başarılı olmuşlardır.

Şekerpancarı veri seti için bileşen sayısına göre özellik çıkarma yöntemlerinin karesel diskiriminant analizindeki sınıflama performanslarının grafiksel karşılaştırması Şekil 4.20'de verilmiştir.



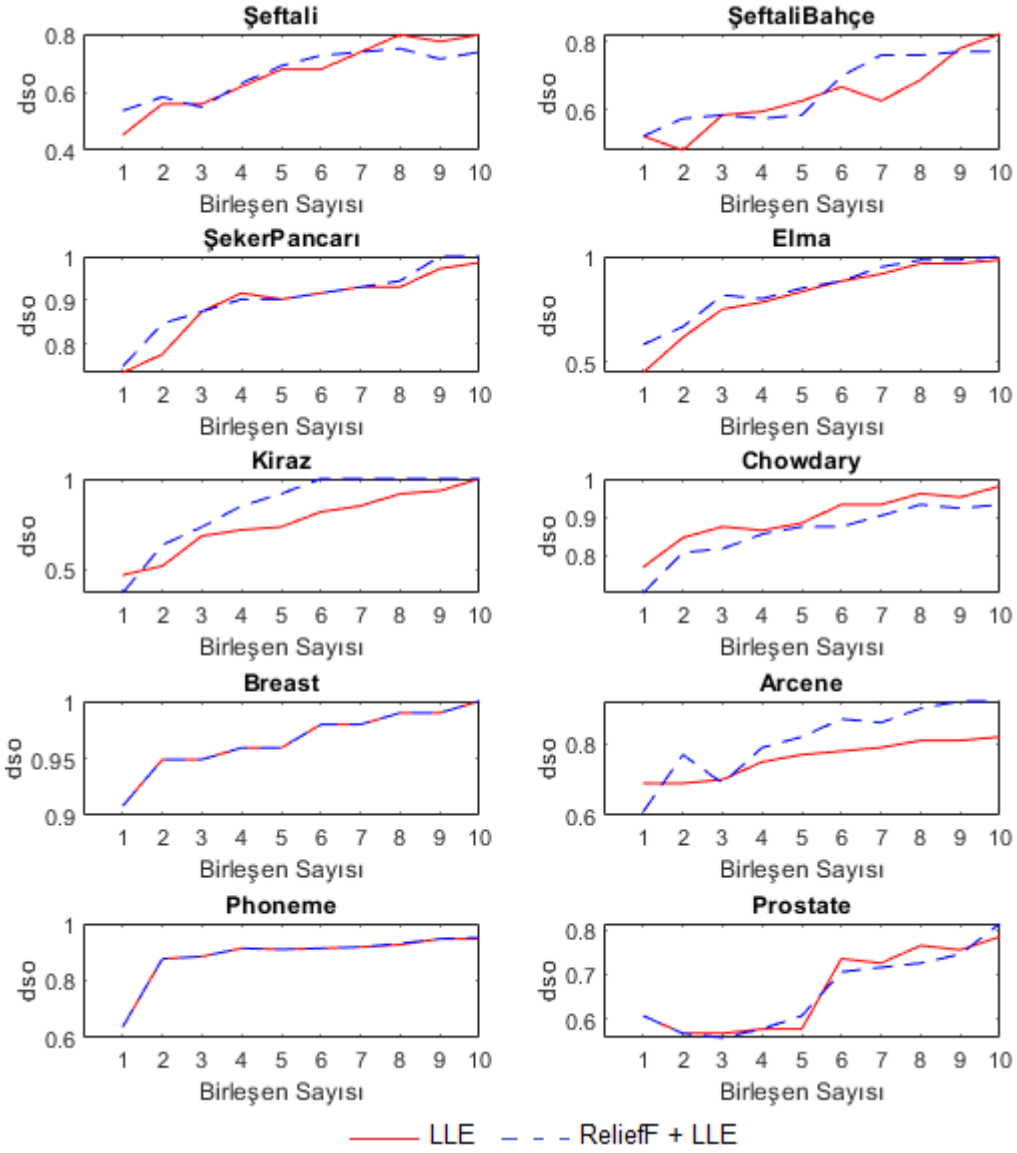
Şekil 4.20. Özellik çıkarma yöntemlerinin sınıflandırma performansları (Şekerpancarı)

Şekil 4.20 incelendiğinde Şekerpancarı veri setinde doğru sınıflandırma olasılığı ve kappa istatistiği kriterlerine göre genel olarak LLE yönteminin sınıflama performansının diğer özellik çıkarma yöntemlerine göre daha iyi olduğu görülmüştür. Entropy kriterine göre bileşen sayısının 7 ve üzeri olduğu durumlarda SMACOF yöntemi diğer özellik çıkarma yöntemlerine göre daha başarılı olmuştur.

Genel olarak değerlendirildiğinde boyut indirgeme işleminde özellik çıkarma yöntemlerinin sınıflama performansı bakımından oldukça başarılı olduğu gözlemlenmiştir. İncelenen dört özellik çıkarma yöntemi arasında TBA ve LLA yöntemleri diğer yöntemlere göre daha başarılı olmuştur.

4.6. Özellik Seçim ve Özellik Çıkarma Yöntemlerinin Birlikte Kullanımının Sınıflama Performansı Üzerindeki Etkileri

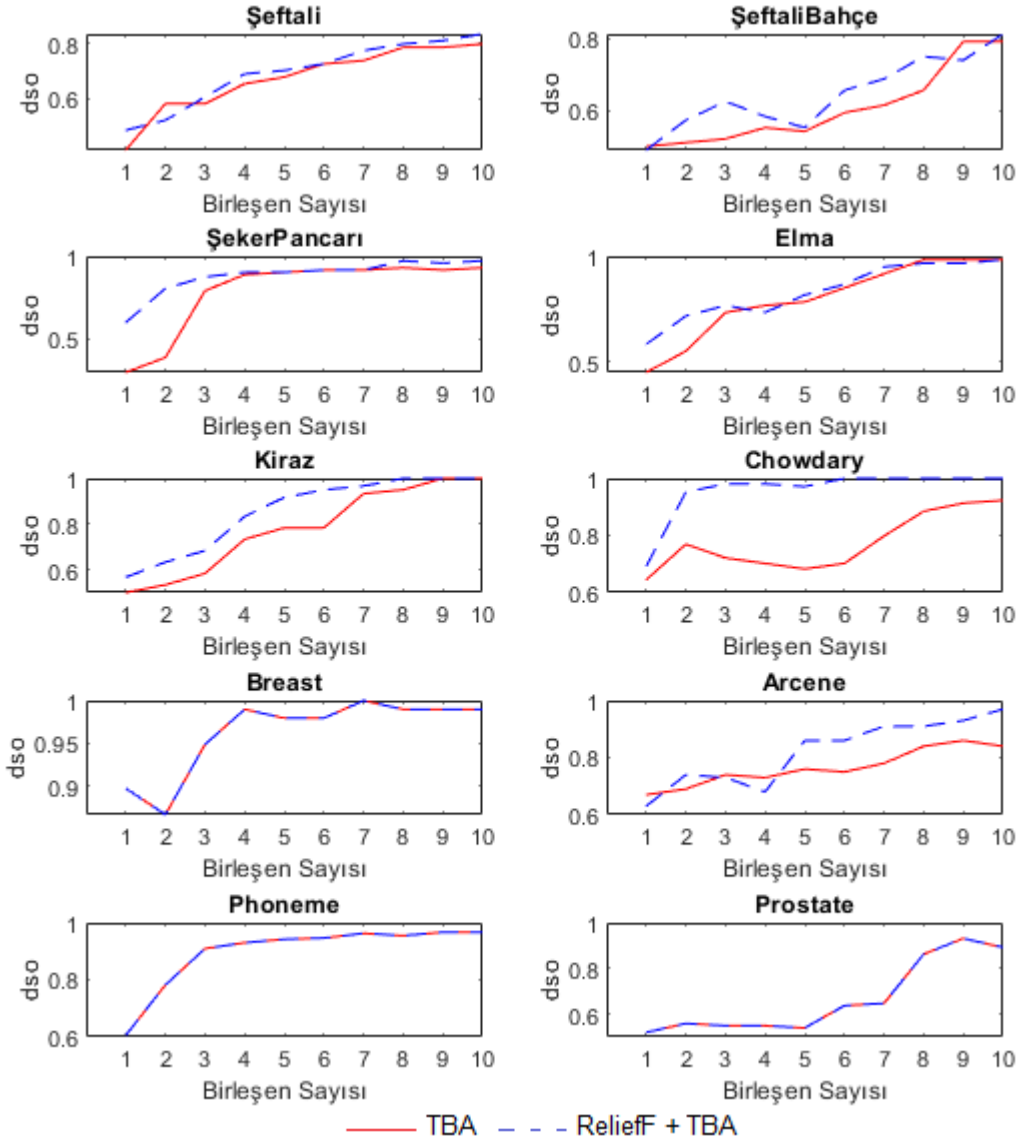
Çalışmanın bu bölümünde özellik çıkarma yöntemleri öncesinde uygulanacak özellik seçim yönteminin sınıflama performansı üzerindeki etkisi incelenecektir. Çalışmada özellik seçim yöntemlerinden ReliefF yöntemi ile 200 özellik belirlendikten sonra, belirlenen bu 200 özelliğe uygulanan özellik çıkarma yöntemlerinin karesel diskriminant analizindeki sınıflama performansları doğru sınıflandırma olasılığı bakımından ölçülmüştür. LLE yöntemi ve ReliefF ile LLE yöntemlerinin bir arada kullanılması ile elde edilen bileşenlere göre hesaplanan doğru sınıflandırma olasılıklarının çizgi grafikleri Şekil 4.21'de verilmiştir.



Şekil 4.21. LLE yöntemi ve ReliefF + LLE yöntemlerinin sınıflandırma performansları

Şekil 4.21 incelendiğinde doğru sınıflandırma olasılığı kriteri bakımından LLE yöntemi öncesi özellik seçim yöntemi ile boyut indirgemenin sınıflama performansı üzerinde olumlu etkileri olduğu gözlemlenmiştir. İncelenen veri setlerinde sadece Chowdary veri setinde, doğru sınıflandırma olasılığı kriteri bakımından LLE yöntemi öncesi uygulanan özellik seçim yöntemi ReliefF yönteminin sınıflama performansı üzerinde olumsuz etkisi olmuştur.

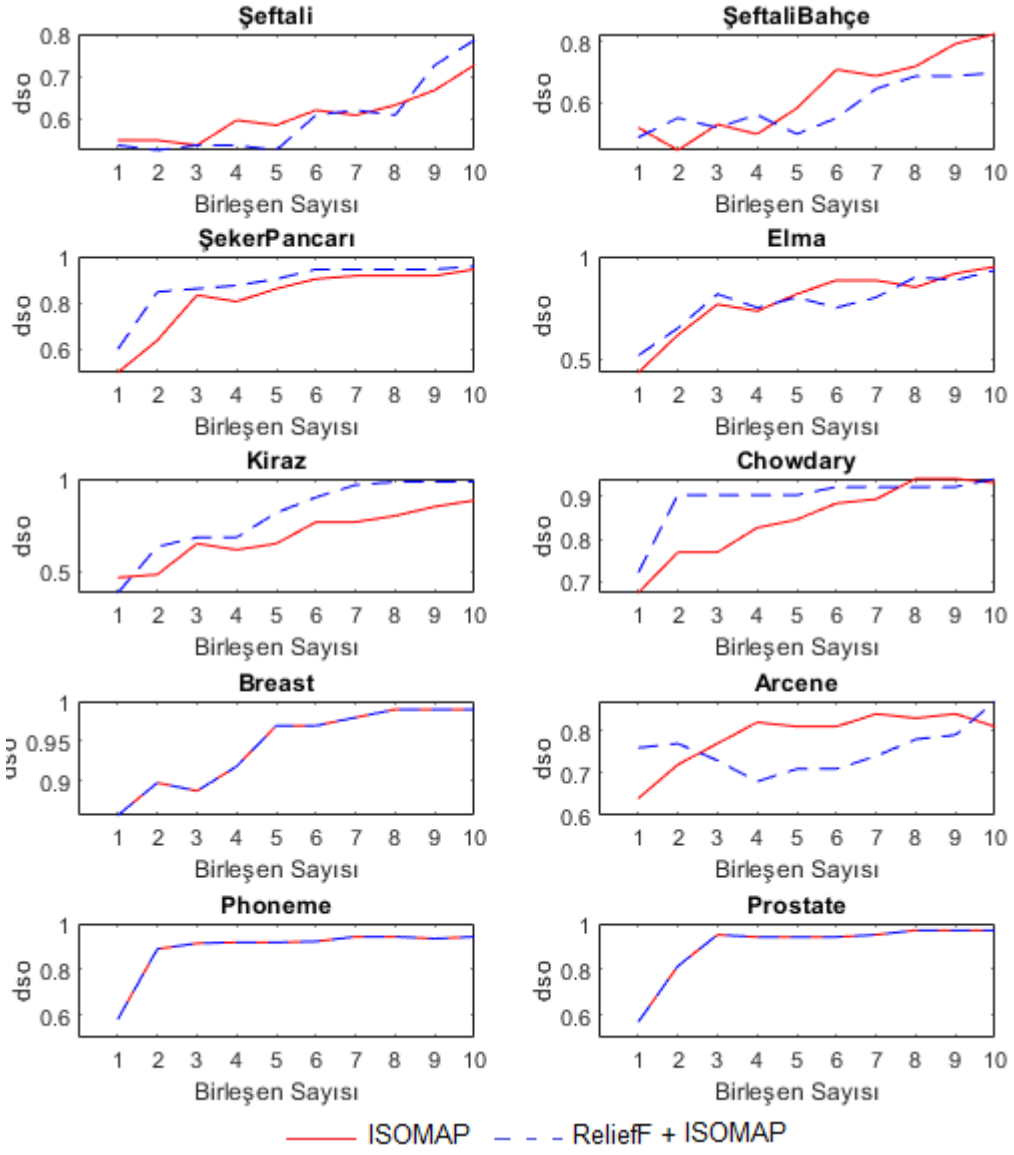
TBA yöntemi ve ReliefF ile TBA yöntemlerinin bir arada kullanılması ile elde edilen bileşenlere göre hesaplanan doğru sınıflandırma olasılıklarının çizgi grafikleri Şekil 4.22'de verilmiştir.



Şekil 4.22. TBA yöntemi ve ReliefF + TBA yöntemlerinin sınıflandırma performansları

Şekil 4.22 incelendiğinde doğru sınıflandırma olasılığı bakımından TBA yöntemi öncesi özellik seçim yöntemi ile boyut indirgemenin sınıflama performansı üzerinde olumlu etkileri olduğu gözlemlenmiştir. İncelenen veri setlerinden Breast, Phoneme ve Prostate veri setlerinde TBA yöntemi ile ReliefF + TBA yöntemlerinde aynı sonuçlar elde edilmiştir. İncelenen diğer veri setlerinde TBA öncesi uygulanan özellik seçim yöntemi ile TBA yönteminden daha yüksek doğru sınıflandırma olasılıkları elde edilmiştir.

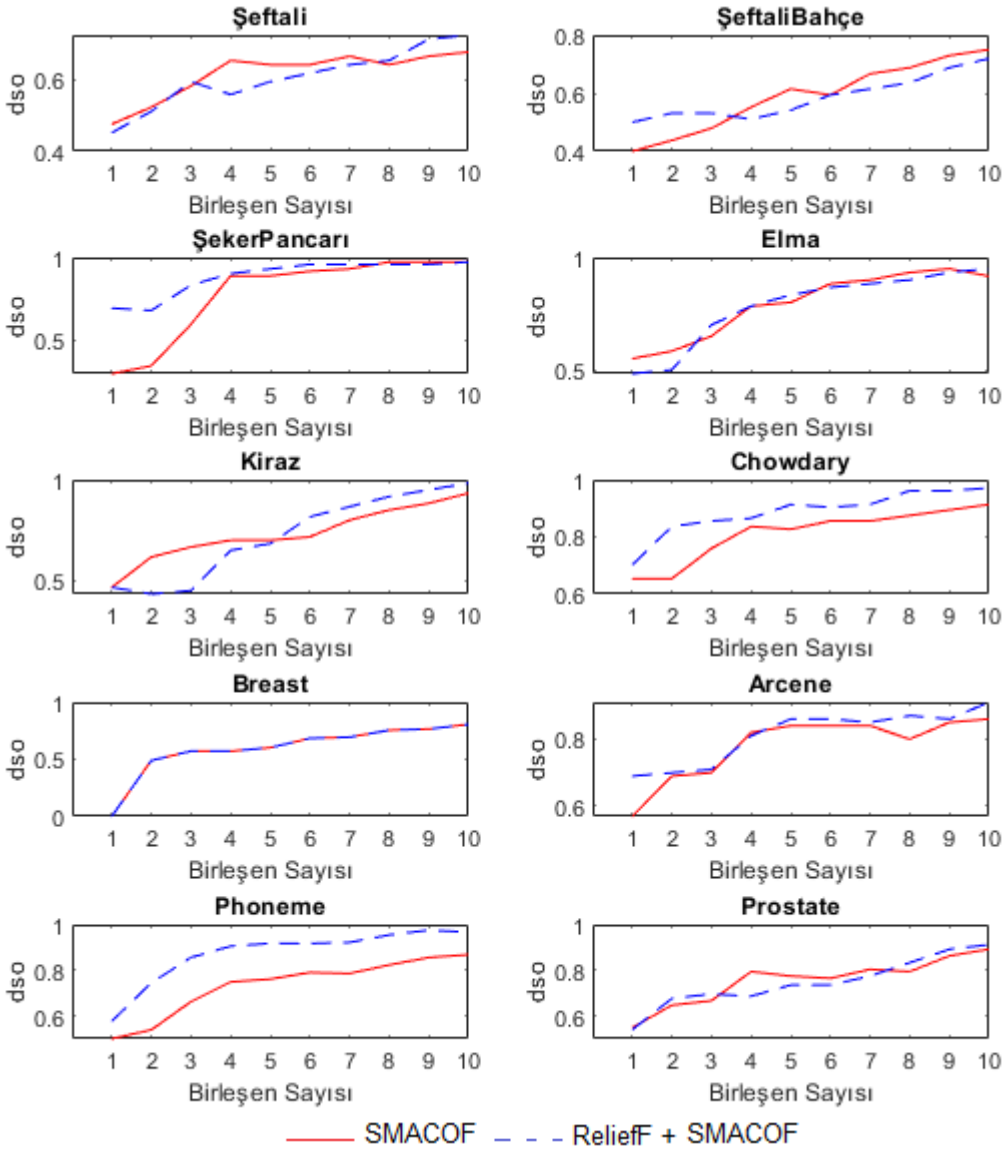
ISOMAP yöntemi ve ReliefF ile ISOMAP yöntemlerinin bir arada kullanılması ile elde edilen bileşenlere göre hesaplanan doğru sınıflandırma olasılıklarının çizgi grafikleri Şekil 4.23'de verilmiştir.



Şekil 4.23. ISOMAP yöntemi ve ReliefF + ISOMAP yöntemlerinin sınıflandırma performansları

Şekil 4.23 incelendiğinde doğru sınıflandırma olasılığı bakımından ISOMAP yöntemi öncesi özellik seçim yöntemi ile boyut indirgemenin sınıflama performansı üzerinde olumlu etkileri olduğu gözlemlenmiştir.

SMACOF yöntemi ve ReliefF ile SMACOF yöntemlerinin bir arada kullanılması ile elde edilen bileşenlere göre hesaplanan doğru sınıflandırma olasılıklarının çizgi grafikleri Şekil 4.24'de verilmiştir.



Şekil 4.24. SMACOF yöntemi ve ReliefF + SMACOF yöntemlerinin sınıflandırma performansları

Şekil 4.24 incelendiğinde doğru sınıflandırma olasılığı bakımından SMACOF yöntemi öncesi özellik seçim yöntemi ile boyut indirgemenin sınıflama performansı üzerinde olumlu etkileri olduğu gözlemlenmiştir.

5. SONUÇLAR VE ÖNERİLER

5.1 Sonuçlar

Bu çalışmada özellik seçim ve özellik çıkarma yöntemleri olmak üzere iki kategoride ele alınan boyut indirgenme tekniklerinin karesel diskriminant analizindeki sınıflama etkinliği incelenmiştir. Özelik seçim ve özellik çıkarma yöntemlerinin sınıflama performansı, özellik sayısının birim sayısından fazla olduğu nicel verilerden oluşan yüksek boyutlu 10 gerçek veri seti üzerinde incelenmiştir.

Karşılaştırma kriteri olarak doğru sınıflandırma olasılığı, entropy ve kappa katsayısının kullanıldığı çalışma sonucunda ele alınan özellik seçim ve özellik çıkarma yöntemlerinin boyut indirgemede oldukça etkili olduğu görülmüştür. Karşılaştırma sonuçlarına göre, en iyi performans gösteren boyut indirgeme yöntemlerinin veri setine göre farklılık gösterdiği tespit edilmiştir. Genel olarak beklenildiği gibi boyut indirgeme sonrası kullanılan boyut sayısı artığında yöntemler arasındaki farklılıkların azaldığı tespit edilmiştir.

Çalışmada ele alınan karşılaştırma kriterlerinden doğru sınıflandırma olasılığı ve kappa katsayısı sonuçlarının birbiri ile uyumlu iken entropy kriterinin sonuçlarının farklılık gösterebildiği tespit edilmiştir. Entropy kriteri, diğer karşılaştırma kriterlerine göre boyut indirgeme yöntemlerinin sınıflama performansını ayırt etmede daha etkili olmuştur.

Özellik seçim yöntemlerinde genel olarak ReliefF, F test istatistiği ve NCA yöntemlerinin sınıflama performansı diğer yöntemlere göre daha başarılı olmuştur. Özellik çıkarma yöntemleri arasında TBA ve LLE yöntemleri sınıflama performansı bakımından diğer yöntemlere göre öne çıkmıştır.

Çalışmada sınıf yapısına dayalı olarak önerilen değişim katsayısı oranlarının incelenen veri setlerinin genelinde klasik değişim katsayısına göre daha başarılı olduğu tespit edilmiştir. Önerilen özellik seçim yöntemlerinin karesel diskriminant analizindeki sınıflama performansı genel olarak başarılı olmuş ve önerilen yöntemlerin hesaplama kolaylığı ve etkinliği bakımından sınıflama analizlerinde boyut indirgeme amacıyla kullanılabileceği görülmüştür.

Çalışmada ayrıca ele alınan LLE, TBA, ISOMAP ve SMACOF yöntemleri öncesi uygulanacak olan özellik seçim yöntemlerinin sınıflama performansını genel olarak artırdığı gözlemlenmiştir.

5.2 Öneriler

Boyut indirgeme yöntemlerinin karesel diskriminant analizindeki sınıflandırma performanslarının farklı olduğunun ortaya konulduğu bu çalışma sonrası, ileri ki çalışmalarda farklı sınıflama analizi yöntemlerinde boyut indirgeme tekniklerinin sınıflama performansları karşılaştırılabilir.

Bu çalışmada boyut indirgeme yöntemlerinin, nicel veri setlerinde özellik sayısının birim sayısından daha fazla olduğu yüksek boyutlu gerçek veri setlerindeki sınıflama performansları incelenmiştir. Sonraki çalışmalarda hem nicel hem de nitel verilerden oluşan veri setlerinde boyut indirgeme tekniklerinin sınıflama performansı incelenebilir.

6. KAYNAKLAR

Akyürek Ö. 2012. Hipekstral Görüntülerde Boyut indirgeme yöntemlerinin Karşılaştırmalı Analizi. Kocaeli Üniversitesi Fen Bilimleri Enstitüsü. Kocaeli

Bingham E. Mannila H. 2001. Random projection in dimensionality reduction: applications to image and text data. Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining.245-250.ACM.

Borg, I., & Groenen, P. (1997). Multitrait-multimethod by multidimensional scaling. SoftStat, 97, 59-65.

Bolón-Canedo. V. Sánchez-Marono. N. Alonso-Betanzos. A. Benítez J. M. ve F. Herrera. "A review of microarray datasets and applied feature selection methods." Information Sciences. cilt 282. s. 111-135. 2014.

Budak H. 2015. Özellik Seçim Yöntemleri ve Yeni Bir Yaklaşım. Mimar Sinan Güzel Sanatlar Üniversitesi Fen Bilimleri Enstitüsü. Doktora Tezi

Budak H. 2018. Özellik Seçim Yöntemleri ve Yeni Bir Yaklaşım. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi.22. 21-31.İstanbul.

Cancela, B., Bolón-Canedo, V., Alonso-Betanzos, A., & Gama, J. (2020). A scalable saliency-based feature selection method with instance-level information. Knowledge-Based Systems, 192, 105326

Carreira-Perpinán M.A. 2010. The Elastic Embedding Algorithm for Dimensionality Reduction. ICML .10.167-174.

Castro B.M. Lemes R.B. Cesar J. Hünemeier T . Leonardi F .2018. A model selection approach for multiple sequence segmentation and dimensionality reduction. Journal of Multivariate Analysis.319-330. Elsevier

Catalbas M.C. Ozkazanc Y and Gulten A. 2015. Kanonik Korelasyon Analizi ile Cinsiyet Tabanlı İmge Sınıflandırma. Akademik Platform

Cigdem O. Demirel H.2018. Performance analysis of different classification algorithms using different feature selection methods on Parkinson's disease detection. Journal of neuroscience methods. 81-90. Elsevier.

Chowdary D, Lathrop J, Skelton J, Curtin K et al. Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. J Mol Diagn 2006 Feb;8(1):31-9. PMID: 16436632

Cortez P. ve A. M. G. Silva. "Using data mining to predict secondary school student performance." in: A. Brito & J. Teixeira (Eds.). Proceedings of 5th Annual Future Business Technology Conference. 2008. s. 5–12.

Cox, T.F. , Cox,M.A.A. (2001). Multidimensional scaling, Newcastle: CRC Press.

Critchley. F. (1985). “Influence in Principal Components Analysis.”Biometrika.72. 627–636

Çatalbaş M.C. 2014. Temel Bileşenler Analizi ve Kanonik Korelasyon Analizi ile İmge Tanıma ve Sınıflandırma. Hacettepe Üniversitesi Fen Bilimleri Enstitüsü. Ankara

Dedeoğlu, M. (2011), Elma ve kiraz ağaçlarında çinko noksanlığının görünür yakın kızılötesi (VNIR) spektroradyometrik yöntemle belirlenebilirliğinin araştırılması, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü Toprak Bilimi ve Bitki Besleme Anabilim Dalı Yüksek Lisans Tezi, Konya.

Dedeoğlu, M., Başayiğit, L., Erişoğlu, M. (2019). Şeker pancarı yapraklarında azot durumunun spektral diskriminant analizi ile belirlenmesi. Toprak Bilimi ve Bitki Besleme Dergisi, 7(2), 128-138.

Dedeoglu, M. (2020). Estimation of critical nitrogen contents in peach orchards using visible-near infrared spectral mixture analysis. Journal of Near Infrared Spectroscopy, 0967033520939319

Dehak N. Torres-Carrasquillo PA. Reynolds D. Dehak R.2011. Language recognition via i-vectors and dimensionality reduction. Twelfth annual conference of the international speech communication association.

Devlin. S. J.. Gnanadesikan. R.. and Kettenring. J. R. (1981). “Robust Estimation of Dispersion Matrices and Principal Components.”Journal of the American Statistical Association.76. 354–362.

Dietterich. T. G.: 1997. ‘Machine Learning Research: Four Current Directions’. AI Magazine.18(4). 97–136.

Dunn, Kevin G. 2020, Process Improvement Using Data, Release 04388a.

Durmaz O. Bilge HŞ . 2011. Metin sınıflandırmada boyut azaltmanın etkileri ve özellik seçimi. Signal Processing and Communications Applications (SIU 2011).21-24.

Durgabai R.P.L. Bhushan Y.R. 2014. Feature selection using ReliefF algorithm. International Journal of Advanced Research in Computer and Communication Engineering. 3(10). 8215-8218.

Erişoğlu M. 2011. Uzaklık Ölçülerinin Kümeleme Analizine Olan Etkilerinin İncelenmesi Ve Geliştirilmesi. Çukurova Üniversitesi Fen Bilimleri Enstitüsü İstatistik Ana Bilim Dalı. Basılmamış Doktora Tezi. Adana.

Gnanadesikan. R. (1997).Methods for Statistical Data Analysis of Multivariate Observations. 2nd ed.. New York: Wiley

Graf, AB ve Wichmann, FA (2002, Kasım). İnsan yüzlerinin cinsiyet sınıflandırması. In Biyolojik Uluslararası Çalıştay Bilgisayar Vizyon Motive (s. 491-500). Springer, Berlin, Heidelberg.

Gorostiaga A. ve J. L. Rojo-Álvarez. “On the use of conventional and statistical-learning techniques for the analysis of PISA results in Spain.” Neurocomputing. cilt 171. s. 625-637. 2016

Guo C. Wu D. 2018. Feature dimensionality reduction for video affect classification: A comparative study. 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia).1-6. IEEE.

Gümüşçü A. İ. B. Aydilek ve R. Taşaltın. “Mikro-dizilim Veri Sınıflandırmasında Öznitelik Seçme Algoritmalarının Karşılaştırılması.” Harran Üniversitesi Mühendislik Dergisi. cilt 1. sayı 1. s. 1-7. 2016.

Hastie, T., Buja, A., & Tibshirani, R. (1995). Penalized discriminant analysis. The Annals of Statistics, 73-102.

Harsanyi J.C. Chang C.I. 1994. Hyperspectral image classification and dimensionality reduction: An orthogonal subspace projection approach. IEEE Transactions on geoscience and remote sensing.32(4). 779-785. IEEE.

Hotellıng, Harold, 1933. Analysis of a Complex of Statistical Variables into Principal Components. Journal of Educational Psychology, 24(6 & 7), 417–441 & 498–520.

Isabelle Guyon, Steve R. Gunn, Asa Ben-Hur, Gideon Dror, 2004. Result analysis of the NIPS 2003 feature selection challenge. In: NIPS

Karakoca A., Ü. Erişoğlu, M. Erişoğlu, A. Pekgör, 26 th European Conference on Operational Research konferansı dahilinde "Abstract Book" bildiri kitapçığındaki "A New Dimension Reduction Approach in The Classification of High-Dimensional Data", 149 pp., Roma, İtalya, Temmuz 2013

Kaiser. H. F. (1970). “A Second Generation Little Jiffy.”Psychometrika.35. 401–415

Kaiser. H. F.. and Rice. J. (1974). “Little Jiffy. Mark IV.”Educational and Psychological Measurement.34. 111–117.

Keogh E. Chakrabarti K. Pazzani X. Mehrotra S.2001. Dimensionality reduction for fast similarity search in large time series databases. Knowledge and information Systems.3(3).263-286. Springer.

Kira. K. and L. A. Rendell: 1992b. ‘A practical approach to feature selection’. In: D.Sleeman and P.Edwards (eds.): Machine Learning: Proceedings of International Conference (ICML’92). pp. 249–256. Morgan Kaufmann.

Kononenko, I. (1994) Estimating Attributes: Analysis and Extensions of RELİEF. Machine Learning: ECML-94, Euro- pean Conference on Machine Learning, Secaucus, 6-8 April 1994, 171-182.

Kuhn, M. ve Johnson, K. (2013). Uygulamalı Tahmine Dayalı Modelleme. New York: Springer.

Kurt Z . 2013. Temel bileşen analiziyle öznitelik seçimi ve görsel nesne sınıflandırma. Master thesis. Eskişehir Osmangazi Üniversitesi. Fen Bilimleri Enstitüsü. Eskişehir.

Kuş M .2013. Temel Bileşen Analizi Ve Fisher Doğrusal Ayırıcılar Yöntemleri ile Kulak Biyometrisi. PhD thesis.Fen Bilimleri Enstitüsü.

Lai C. Guo S. Cheng L. Wang W. 2017. A comparative study of feature selection methods for the discriminative Analysis of Temporal Lobe Epilepsy. Frontiers in neurology. 8(633). Frontiers.

Li W. Prasad S. Fowler J.E. and Bruce L.M. .2012. Locality-preserving dimensionality reduction and classification for hyperspectral image analysis. IEEE Transactions on Geoscience and Remote Sensing. 50(4).1185-1198.

Makul Ö. Ekinci M. 2017. A graph form data stream clustering approach based on dimension reduction. Signal Processing and Communications Applications Conference (SIU). 2017 25th.1-4. IEEE.

Mardia, K., Kent, J. Bibby, 1979, Multivariate Analysis, Academic Press, London.

Murphy, K. P. (2012). Machine learning: a probabilistic perspective. MIT press.

Jolliffe. I. T. (1972). "Discarding Variables in a Principal Component Analysis. I: Artificial Data."Applied Statistics.21. 160–173

Öztürk H. 2016. EEG sinyallerinde farklı boyut indirgeme ve sınıflandırma yöntemlerinin karşılaştırılması. Adnan Menderes Üniversitesi Sağlık Bilimleri Enstitüsü. Aydın

Pai. P. F., Chen C. T. , Y. M. Hung. W. Z. Hung ve Y. C. Chang. "A group decision classifier with particle swarm optimization and decision tree for analyzing achievements in mathematics and science." Neural Computing and Applications. cilt 25. sayı 7-8. s. 2011-2023. 2014.

Pamukçu, Ö., Kayar, Y., Eroğlu, H., Kalkan Erol, K., İlhan, A. ve Kocaman, O. (2015). Diyabetik hastalarda Helicobacter pylori enfeksiyonları ile iltihaplar, metabolik sendrom ve komplikasyonlar arasındaki ilişki. Uluslararası kronik hastalıklar dergisi , 2015 .

Pearson, Karl, 1901. On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine, Series 6, 2(11), 559–572.

Rencher, A. C., & Schaalje, G. B. (2007). Linear models in statistics. New Jersey, Hoboken.

Robnik-Šikonja M. Kononenko I.2003. Theoretical and empirical analysis of ReliefF and RReliefF. Machine learning. 53(1-2). 23-69. Springer.

Rosman G. Bronstein M.M. Bronstein A.M. Kimmel R .2010. Nonlinear dimensionality reduction by topologically constrained isometric embedding. International Journal of Computer Vision.89(1).56-68. Springer.

Roweis S.T. Saul L.K .2000. Nonlinear dimensionality reduction by locally linear embedding. American Association for the Advancement of Science. 290(5500). 2323-2326.

Ruymgaart. F. H. (1981). "A Robust Principal Component Analysis."Journal of Multivariate Analysis.11. 485–497.

Saeys, Y., Inza, I., and Larranaga, P., 2007. A review of feature selection techniques in bioinformatics, Bioinformatics, 23(19), 2507-2517.

Sarwar B. Karypis G. Konstan J. and Riedl J .2000. Application of dimensionality reduction in recommender system-a case study. Minnesota Univ Minneapolis Dept of Computer Science

Sellami A ve Farah M. 2018. Comparative study of dimensionality reduction methods for remote sensing images interpretation. Advanced Technologies for Signal and Image Processing (ATSIP). 2018 4th International Conference on .1-6. IEEE.

Servi T .2009. Çok Değişkenli Karma Dağılım Modeline Dayalı Kümeleme Analizi. Yayımlanmamış Doktora Tezi. Çukurova Üniversitesi. Adana.

Singh G.D.A.A. Balamurugan S.A.A. Leavline E.J. 2016. Literature review on feature selection methods for high-dimensional data. International Journal of Computer Applications. 8887. Foundation of Computer Science.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., ... & Lander, E. S. (2002). Gene expression correlates of clinical prostate cancer behavior. Cancer cell, 1(2), 203-209.

Takane, Y., Young, FW ve De Leeuw, J. (1977). Metrik olmayan bireysel farklılıklar çok boyutlu ölçekleme: Optimal ölçekleme özelliklerine sahip alternatif bir en küçük kareler yöntemi. Psychometrika , 42 (1), 7-67.

Tenenbaum, JB , Bernstein, M., De Silva, V. Ve Langford, JC (2000). Gömülü manifoldlar üzerindeki jeodeziklere yönelik grafik yaklaşımları (sayfa 961-968). Teknik rapor, Psikoloji Bölümü, Stanford Üniversitesi.

Toktay Y. 2017. Çok Değişkenli İstatistik Analiz Yöntemler: Faktör Analizi Ve Diskriminant Analizinin Iğdır Üniversitesi Öğrencileri Üzerine Uygulaması. Iğdır Üniversitesi Fen Bilimleri Enstitüsü İstatistik Ana Bilim Dalı. Yüksek lisans Tezi. ığdır.

Van Der Maaten L. Postma E. Van den Herik J. 2009. Dimensionality reduction: a comparative. J Mach Learn Res.10.66-71.

Yakut İ. .2008. Prıvacy-Preservıng Dımensıonalıty Reductıon-Based Collaboratıve Fılterıng. Anadolu Üniversitesi. Fen Bilimleri Enstitüsü. Bilgisayar Mühendisliği Bölümü. Eskişehir .

Yang W. Wang K. Zuo W.2012. Neighborhood Component Feature Selection for High-Dimensional Data. JCP.7(1). 161-168.

Yıldız K. 2010. Veri madenciliğinde yüksek boyutlu veriler ile uygulama. Master thesis. Marmara Üniversitesi Fen Bilimleri Enstitüsü. İstanbul.

Yıldız E. Sevim Y. 2016. Comparison of linear dimensionality reduction methods on classification methods. Electrical. Electronics and Biomedical Engineering (ELECO). 2016 National Conference on.161-164. IEEE.

Yüksek AG. Arslan H. Kaynar O. Delibaş E. Şeker A .2017. Farklı Boyut İndirgeme Yöntemlerinin. Anfis Modelinin Eğitim Performansı üzerindeki Etkilerinin Karşılaştırılması. 38(4). 716-730.

Zhang Y. Zhou Z.H. .2010. Multilabel dimensionality reduction via dependence maximization. ACM Transactions on Knowledge Discovery from Data (TKDD).4(3).14.ACM.

Zhang Z. Zha H. 2004. Principal manifolds and nonlinear dimensionality reduction via tangent space alignment. SIAM journal on scientific computing.26(1).313-338. SIAM.

Dietterich T. G. “Ensemble methods in machine learning.” in International workshop on multiple classifier systems. 2000. s. 1-15.

Tilki Ö .2014. PCA based face recognition: An application. Çankaya Üniversitesi. Fen Bilimleri Enstitüsü. Bilgisayar Mühendisliği Bölümü. Ankara.

Thomson. G. H. (1951).The Factorial Analysis of Human Ability. London: London University Press.

Turhan B .2004. Nonlinear dimensionality reduction methods for pattern recognition. PhD thesis. Bogaziçi University. İstanbul.

Turgut S. Dağtekin M. Ensari T. 2018. Microarray breast cancer data classification using machine learning methods. 2018 Electric Electronics. Computer Science. Biomedical Engineerings' Meeting (EBBT). 1-3. IEEE.

Wang X. Paliwal K.K. 2003. Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. Pattern recognition.36(10). 2429-2439. Elsevier.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı	**:**	Tenzile ERBAYRAM
Uyruğu	**:**	T.C.
Doğum Yeri ve Tarihi	**:**	Kadınhanı /KONYA 22.10.1994
Telefon	**:**	05071695855
Faks	**:**	
e-mail	**:**	tnzle12@gmail.com.tr

EĞİTİM

Derece		**Adı, İlçe, İl**	**Bitirme Yılı**
Lise	:	Karatay Anadolu İmam Hatip Lisesi Karatay /Konya	2012
Üniversite	:	Necmettin Erbakan Üniversitesi Meram/Konya	2017
Yüksek Lisans	:	Necmettin Erbakan Üniversitesi Meram/Konya	2020

İŞ DENEYİMLERİ

Yıl	**Kurum**	**Görevi**
2020	Selçuk Üniversitesi Fen Fakültesi	Araştırma Görevlisi

UZMANLIK ALANI

İstatistik

YABANCI DİLLER

İngilizce