

T.C.
İSTANBUL KÜLTÜR ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

**MÜŞTERİ İLİŞKİLERİ YÖNETİMİ İÇİN MANİFOLD ÖĞRENME
İLE DENETİMLİ DOĞRUSAL OLMAYAN BOYUT İNDİRGEME**

YÜKSEK LİSANS TEZİ
BEYTULLAH YILMAZ
2000001760

Anabilim Dalı: Bilgisayar Mühendisliği
Programı: Bilgisayar Mühendisliği

Tez Danışmanı: Doç. Dr. Fatma PATLAR AKBULUT

EYLÜL 2023

T.C.
İSTANBUL KÜLTÜR ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

**MÜŞTERİ İLİŞKİLERİ YÖNETİMİ İÇİN MANİFOLD ÖĞRENME
İLE DENETİMLİ DOĞRUSAL OLMAYAN BOYUT İNDİRGEME**

YÜKSEK LİSANS TEZİ
BEYTULLAH YILMAZ
2000001760

Anabilim Dalı: Bilgisayar Mühendisliği
Programı: Bilgisayar Mühendisliği

Tez Danışmanı: Doç. Dr. Fatma PATLAR AKBULUT
Jüriler: Prof. Dr. Güray Yılmaz
Doç. Dr. Akhan Akbulut

EYLÜL 2023

ÖNSÖZ ve TEŞEKKÜR

Tez konusu seçiminde bana yol gösteren, çalışma sürecinde beni destekleyen, yönlendiren ve tez yazımında yardımını esirgemeyen tez danışmanım Doç. Dr. Fatma PATLAR AKBULUT hocama, sevgili eşim Hatice Hazal Hacibeyoğlu Yılmaz'a ve İstanbul Kültür Üniversitesi'nin değerli bölüm hocalarına teşekkür ederim.



İÇİNDEKİLER

Sayfa

ÖNSÖZ ve TEŞEKKÜR	i
ŞEKİLLER LİSTESİ	iv
TABLolar LİSTESİ	v
RESİMLER LİSTESİ	vii
SİMGELER VE KISALTMALAR LİSTESİ	viii
ÖZET	x
ABSTRACT	xi
1. GİRİŞ	1
2. VERİ SETİ ve VERİ ÖN İŞLEME TEKNİKLERİ	3
2.1. Sayısal (Nicel) Değişkenler	3
2.1.1. Sayısal Sürekli Değişkenler	4
2.1.2. Sayısal Ayırık Değişkenler	4
2.2. Kategorik Değişkenler	4
2.2.1. Sınıflayıcı (Nominal) Değişkenler	4
2.2.2. Sıralayıcı (Ordinal) Değişkenler	4
2.3. Veri Ön İşleme Teknikleri	5
2.3.1. Veriyi Yeniden Ölçeklendirme (Rescaling Data)	5
2.3.2. Ortalama Kaldırma (Mean Removal)	5
2.3.3. Veriyi Standartlaştırma (Standartizing Data)	5
2.3.4. Tek Dönüşüm (One Hot Encoding)	6
2.3.5. Veriyi Normalleştirme (Normalizing Data)	6
2.3.6. Eksik Değerleri Tamamlama (Replacing Missing Values)	6
2.4. Kullanılan Veri Setleri	6
3. BOYUT İNDİRGEME	8
3.1. Özellik Seçimi (Feature Selection)	8
3.1.1. Filtre Yöntemi (Filter Method)	9
3.1.2. Sarma Yöntemi (Wrapper Method)	9
3.1.3. Gömülü Yöntem (Embedded Method)	9
3.2. Özellik Çıkarma (Feature Extraction)	9
3.3. Doğrusal Boyut İndirgeme Yöntemleri	10
3.3.1. Temel Bileşenler Analizi (Principal Component Analysis, PCA)	10

3.3.2.	Doğrusal Ayırıştırma Analizi (Linear Discriminant Analysis, LDA)	11
3.3.3.	Tekil Değer Ayırışımı (Singular Value Decomposition, SVD)	13
4.	MANİFOLD ÖĞRENME	14
4.1.	Manifold Kavramı ve Manifold Öğrenme	14
4.2.	Manifold Öğrenme Algoritmaları.....	15
4.2.1.	Çok Boyutlu Ölçekleme (Multidimensional Scaling, MDS).....	15
4.2.2.	Çekirdek Temel Birleşen Analizi (Kernel Principal Component Analysis)....	15
4.2.3.	İzomap (Isomap).....	16
4.2.4.	Lokal Doğrusal Gömme (Locally Linear Embedding, LLE)	17
4.2.5.	T-Dağıtılmış Stokastik Komşu Gömme (T-Distributed Stochastic Neighbor Embedding, t-SNE)	18
4.2.6.	Standart Manifold Yaklaşımı ve Projeksiyonu (Uniform manifold approximation and projection, UMAP).....	19
5.	DEĞERLENDİRME ÖLÇÜM TEKNİKLERİ	20
5.1.	Sheppard Diyagramı (Sheppard Diagram, SD)	20
5.2.	Kruskal Stres Ölçümü (Kruskal Stress Measure, KrM)	21
5.3.	Sammon Stres (Sammon Stress, Ss).....	22
5.4.	Spearman'ın Rho'su (Spearman's Rho, Sr)	22
5.5.	Topolojik Çarpımı (Topological Product, Tpr)	22
5.6.	Artık Varyans (Residual Variance, Rv).....	23
5.7.	König Ölçüsü (König's Measure, KoM)	23
5.8.	Güvenirlilik ve Süreklilik (Trustworthiness & Continuity, Tw & Cnt)	23
5.9.	Yerel Süreklilik Meta Kriteri (Local Continuity Meta-Criterion, LCMC).....	24
5.10.	Yeniden Düzenlenmiş Ölçeklendirme (Refactored Scaling, RES)	24
6.	ANALİZ	25
6.1.	Çekirdek Temel Birleşen Analizi (Kernel PCA) Algoritmasının Sonuçları	25
6.2.	İzomap Algoritmasının Sonuçları.....	30
6.3.	Lokal Doğrusal Gömme (LLE) Algoritmasının Sonuçları.....	35
6.4.	T-Dağıtılmış Stokastik Komşu Gömme (t-SNE) Algoritmasının Sonuçları	41
6.5.	Standart Manifold Yaklaşımı ve Projeksiyonu (UMAP) Algoritmasının Sonuçları	47
7.	TARTIŞMA.....	52
8.	SONUÇ.....	56
9.	KAYNAKLAR.....	57

ŞEKİLLER LİSTESİ

Şekil

Sayfa

Şekil 1 Yıllara göre üretilen, elde edilen kopyalanan ve tüketilen veri miktarı (2004-2020)1

Şekil 2 UCI Makine Öğrenme Kaynağındaki Veri Setlerinin Değişken Tipine Göre Dağılımı
.....2



TABLÖLAR LİSTESİ

Tablo	Sayfa
Tablo 1 Analizde Kullanılan veri setleri	7
Tablo 2 Kernel PCA için kullanılan parametre ve değerleri.....	25
Tablo 3 “Lenses” Veri Setinin Kernel PCA için sonuçları	25
Tablo 4 “Soybean (Small)” Veri Setinin Kernel PCA için sonuçları	26
Tablo 5 “Molecular Biology (Promoter Gene Sequences)” Veri Setinin Kernel PCA için sonuçları	26
Tablo 6 “Hayes-Roth” Veri Setinin Kernel PCA için sonuçları	26
Tablo 7 “Lymphography” Veri Setinin Kernel PCA için sonuçları	27
Tablo 8 “Audiology (Original)” Veri Setinin Kernel PCA için sonuçları	27
Tablo 9 “Breast Cancer” Veri Setinin Kernel PCA için sonuçları	27
Tablo 10 “Soybean (Large)” Veri Setinin Kernel PCA için sonuçları	28
Tablo 11 “Primary Tumor” Veri Setinin Kernel PCA için sonuçları	28
Tablo 12 “Congressional Voting Records” Veri Setinin Kernel PCA için sonuçları	28
Tablo 13 “Balance Scale” Veri Setinin Kernel PCA için sonuçları	29
Tablo 14 “Tic-Tac-Toe Endgame” Veri Setinin Kernel PCA için sonuçları	29
Tablo 15 “Car Evaluation” Veri Setinin Kernel PCA için sonuçları	29
Tablo 16 Isomap için kullanılan parametre ve değerleri	30
Tablo 17 “Lenses” Veri Setinin Isomap için sonuçları	30
Tablo 18 “Soybean (Small)” Veri Setinin Isomap için sonuçları	30
Tablo 19 “Molecular Biology (Promoter Gene Sequences)” Veri Setinin Isomap için sonuçları	31
Tablo 20 “Hayes-Roth” Veri Setinin Isomap için sonuçları	31
Tablo 21 “Audiology (Original)” Veri Setinin Isomap için sonuçları	32
Tablo 22 “Breast Cancer” Veri Setinin Isomap için sonuçları	32
Tablo 23 “Soybean (Large)” Veri Setinin Isomap için sonuçları	32
Tablo 24 “Primary Tumor” Veri Setinin Isomap için sonuçları	33
Tablo 25 “Congressional Voting Records” Veri Setinin Isomap için sonuçları	33
Tablo 26 “Balance Scale” Veri Setinin Isomap için sonuçları	34
Tablo 27 “Tic-Tac-Toe Endgame” Veri Setinin Isomap için sonuçları	34
Tablo 28 “Car Evaluation” Veri Setinin Isomap için sonuçları	35
Tablo 29 LLE için kullanılan parametre ve değerleri	35
Tablo 30 “Lenses” Veri Setinin LLE için sonuçları	35
Tablo 31 “Soybean (Small)” Veri Setinin LLE için sonuçları	36
Tablo 32 “Molecular Biology (Promoter Gene Sequences)” Veri Setinin LLE için sonuçları	36
Tablo 33 “Hayes-Roth” Veri Setinin LLE için sonuçları	37
Tablo 34 “Lymphography” Veri Setinin LLE için sonuçları	37
Tablo 35 “Audiology (Original)” Veri Setinin LLE için sonuçları	38
Tablo 36 “Breast Cancer” Veri Setinin LLE için sonuçları	38
Tablo 37 “Soybean (Large)” Veri Setinin LLE için sonuçları	39
Tablo 38 “Primary Tumor” Veri Setinin LLE için sonuçları	39
Tablo 39 “Congressional Voting Records” Veri Setinin LLE için sonuçları	40
Tablo 40 “Balance Scale” Veri Setinin LLE için sonuçları	40
Tablo 41 “Tic-Tac-Toe Endgame” Veri Setinin LLE için sonuçları	41
Tablo 42 t-SNE için kullanılan parametre ve değerleri	42

Tablo 43 “Lenses” Veri Setinin t-SNE için sonuçları	42
Tablo 44 “Soybean (Small)” Veri Setinin t-SNE için sonuçları.....	42
Tablo 45 “Molecular Biology (Promoter Gene Sequences)” Veri Setinin t-SNE için sonuçları	43
Tablo 46 “Hayes-Roth” Veri Setinin t-SNE için sonuçları	43
Tablo 47 “Lymphography” Veri Setinin t-SNE için sonuçları.....	43
Tablo 48 “Audiology (Original)” Veri Setinin t-SNE için sonuçları	44
Tablo 49 “Breast Cancer” Veri Setinin t-SNE için sonuçları	44
Tablo 50 “Soybean (Large)” Veri Setinin t-SNE için sonuçları.....	45
Tablo 51 “Primary Tumor” Veri Setinin t-SNE için sonuçları.....	45
Tablo 52 “Congressional Voting Records” Veri Setinin t-SNE için sonuçları	46
Tablo 53 “Balance Scale” Veri Setinin t-SNE için sonuçları	46
Tablo 54 “Tic-Tac-Toe Endgame” Veri Setinin t-SNE için sonuçları	46
Tablo 55 “Car Evalution” Veri Setinin t-SNE için sonuçları	47
Tablo 56 UMAP için kullanılan parametre ve değerleri.....	48
Tablo 57 “Lenses” Veri Setinin UMAP için sonuçları.....	48
Tablo 58 “Soybean (Small)” Veri Setinin UMAP için sonuçları	48
Tablo 59 “Molecular Biology (Promoter Gene Sequences)” Veri Setinin UMAP için sonuçları.....	49
Tablo 60 “Hayes-Roth” Veri Setinin UMAP için sonuçları.....	49
Tablo 61 “Lymphography” Veri Setinin UMAP için sonuçları	50
Tablo 62 “Audiology (Original)” Veri Setinin UMAP için sonuçları.....	50
Tablo 63 “Breast Cancer” Veri Setinin UMAP için sonuçları	51
Tablo 64 “Soybean (Large)” Veri Setinin UMAP için sonuçları	51
Tablo 65 “Primary Tumor” Veri Setinin UMAP için sonuçları	52
Tablo 66 Bütün Veri Setlerinin Kernel PCA için sonuçları	53
Tablo 67 Bütün Veri Setlerinin Isomap için sonuçları	54
Tablo 68 Bütün Veri Setlerinin LLE için sonuçları.....	54
Tablo 69 Bütün Veri Setlerinin t-SNE için sonuçları	55
Tablo 70 Bütün Veri Setlerinin UMAP için sonuçları.....	55

RESİMLER LİSTESİ

Resim	Sayfa
Resim 1 İsviçre Rulosu	14
Resim 2 Sheppard Diyagramı	21



SİMGELER VE KISALTMALAR LİSTESİ

Çalışmanın içeriğinde geçen simgeler ve kısaltmalar açıklamaları ile aşağıdaki bölüme eklenmiştir.

Kısaltmalar	Açıklama
Cnt	Süreklilik (Continuity)
Dist	Uzaklık Ölçüm Metodu (Distance)
KoM	König Ölçütü (König's Measure)
KrM	Kruskal Stres Ölçümü (Kruskal Stress Measure)
LCMC	Yerel Süreklilik Meta Kriteri (Local Continuity Meta-Criterion)
NoC	Boyut Sayısı (Number of Components)
NoN	Komşu Sayısı (Number of Neighbors)
NoPn	Parametrelerin RES algoritmasına toplam etkisi
NoPd	Kullanılan parametre sayısı
RES	Yeniden Düzenlenmiş Ölçeklendirme (Refactored Scaling)
Rv	Artık Varyans (Residual Variance)
SD	Sheppard Diyagramı (Sheppard Diagram)
Sr	Spearman'ın Rho'su (Spearman's Rho)
Ss	Sammon Stres (Sammon Stress)
SVD	Tekil Değer Ayrışımı (Singular Value Decomposition)
Tpr	Topolojik Çarpımı (Topological Product)
Tw	Güvenirlilik (Trustworthiness)
TB	Terabayt
ZB	Zetabayt

Simgeler	Açıklama
D_M^o	$T \times 1$ boyutlu d^o setin ortalama değeri
D_M^r	$T \times 1$ boyutlu d^r setin ortalama değeri
d_i^o	$T \times 1$ boyutlu orijinal setin uzaklık matrisinin i 'nci değeri
d_i^r	$T \times 1$ boyutlu boyut indirgenmiş setin uzaklık matrisi i 'nci değeri
d_{ij}^o	$N \times N$ boyutlu orijinal setin uzaklık matrisi i 'nci ve j 'nci kayıtlar arası değeri
d_{ij}^r	$N \times N$ boyutlu boyut indirgenmiş setin uzaklık matrisi i 'nci ve j 'nci kayıtlar arası değeri
d_{ij}^*	$N \times N$ boyutlu orijinal setteki uzaklık matrisi i 'nci ve j 'nci kayıtlar arası değerinin monotonik regresyona göre çıkan değeri
k	Çekirdek fonksiyonu (Kernel function)
o_{ij}^o	$N \times N$ boyutlu orijinal setin uzaklık matrisin i 'nci ve j 'nci kayıtlar arası uzaklığın diğer uzaklıklar göre sırası
o_{ij}^r	$N \times N$ boyutlu boyut indirgenmiş setteki uzaklık matrisin i 'nci ve j 'nci kayıtlar arası uzaklığın diğer uzaklıklar göre sırası
o_i^o	$T \times 1$ boyutlu orijinal setin uzaklık matrisinin i 'nci değerinin diğer kayıtların uzaklığına göre sırası
o_i^r	$T \times 1$ boyutlu orijinal setin uzaklık matrisinin i 'nci değerinin diğer kayıtların uzaklığına göre sırası
X	$M \times N$ boyutlu orijinal veri seti
x	$M \times 1$ boyutlu değişken
x_i	$M \times 1$ boyutlu i 'nci değişken
x_{max}	$M \times 1$ boyutundaki değişkenin maksimum değerinin $1_{M \times 1}$ matrisi ile çarpılması
x_{min}	$M \times 1$ boyutundaki değişkenin minimum değerinin $1_{M \times 1}$ matrisi ile çarpılması
x_{mr}	$M \times 1$ boyutunda ortalaması çıkartılmış değişken
x_{norm}	$M \times 1$ boyutunda verisi yeniden ölçeklendirilmiş değişken
x_{st}	$M \times 1$ boyutunda verisi normalleştirilmiş değişken
μ	$M \times 1$ boyutundaki değişkenin ortalaması
σ	$M \times 1$ boyutundaki değişkenin standart sapması

Üniversite	: T.C. İstanbul Kültür Üniversitesi
Enstitü	: Lisansüstü Eğitim Enstitüsü
Anabilim Dalı	: Bilgisayar Mühendisliği
Program	: Bilgisayar Mühendisliği
Tez Danışmanı	: Doç. Dr. Fatma Patlar AKBULUT
Tez Türü ve Tarihi	: Yüksek Lisans – Eylül 2023

ÖZET

Veri hacimleri katlanarak artmasından dolayı, yüksek boyutlu veri setleri, analiz ve modellemenin yanı sıra görselleştirme ve yorumlama adımlarındaki performanslarında zorluklar ortaya çıkmaya başladı. Bu zorluğun bir sonucu olarak, onu tanımlamak için “boyutsallığın laneti” türetilmiştir. Çeşitli öğrenme yöntemleri, içsel bağlantılarını korurken veri setlerinin boyutlarını azaltarak boyutsallığın lanetini hafifletmeyi amaçladıkları için bu çalışmanın konusudur. Çoklu öğrenme algoritmalarının boyutsallığı düşürmedeki başarısına ilişkin çalışmalar cesaret verici olsa da bu çalışmanın büyük çoğunluğu ya hayali ya da sayısal veri setleriyle yürütülmüştür. Sonuç olarak, bu algoritmaların kategorik verilerle kullanım için uyarlanmasına yönelik araştırmalarda çok büyük bir boşluk vardır. Bu boşluğu doldurmak için, bu tezde önce kategorik veri kümeleri üzerinde öne çıkan manifold öğrenme algoritmalarını seçilmiş ve değerlendirilmiş, ardından sonuçlar derinlemesine analiz edilerek sunulmuştur. Manifold öğrenme teknikleri Çekirdek Temel Birleşen Analizi, İzomap, Lokal Doğrusal Gömme (LLE) (diğer 3 varyantı ile), t-Dağıtılmış Stokastik Komşu Gömme (t-SNE) ve Standart Manifold Yaklaşımı ve Projeksiyonu (UMAP) bu tez için seçilmiştir. Araştırmacılar, bir dizi farklı alandaki başarıları nedeniyle bu algoritmalara çok dikkat etmişlerdir. Bu çalışma, kategorik veriler bağlamında faydalarını keşfederek, boyut indirgeme yöntemlerine ilişkin artan bilgi birikimine katkıda bulunmaktadır. Bu çeşitli öğrenme algoritmalarının, kategorik veri setlerindeki temel yapıyı ve ilişkileri ne kadar iyi yakaladığını ve koruduğunu incelemek, bu yöntemleri değerlendirmenin önemli bir parçasıdır. Performansı değerlendirmek için çeşitli değerlendirme ölçüm teknikleri kullanılmıştır. Bulgular, kategorik verilere uygulandığında manifold öğrenme algoritmalarının güçlü yanlarını ve sınırlamalarını ortaya çıkararak, çeşitli veri türleriyle uğraşan araştırmacılar için yararlı bilgiler sağlamayı hedefler. Bu çalışmadan elde edilen bulgular, boyutluluk indirgeme yöntemlerinin kullanımını geliştirerek, yüksek boyutlu kategorik veri kümelerinin daha doğru modellenmesine ve yorumlanmasına yol açacaktır.

Anahtar Kelimeler: Boyut İndirgeme, Manifold Öğrenme, Çekirdek Temel Birleşen Analizi, İzomap, Lokal Doğrusal Gömme, t-Dağıtılmış Stokastik komşu Gömme, Standart Manifold Yaklaşımı ve Projeksiyonu

University	: İstanbul Kültür University
Institute	: Institute of Graduate Studies
Department	: Computer Engineering
Program	: Computer Engineering
Supervisor	: Assoc. Prof. Fatma Patlar AKBULUT
Degree Awarded and Date	: Master's degree – September 2023

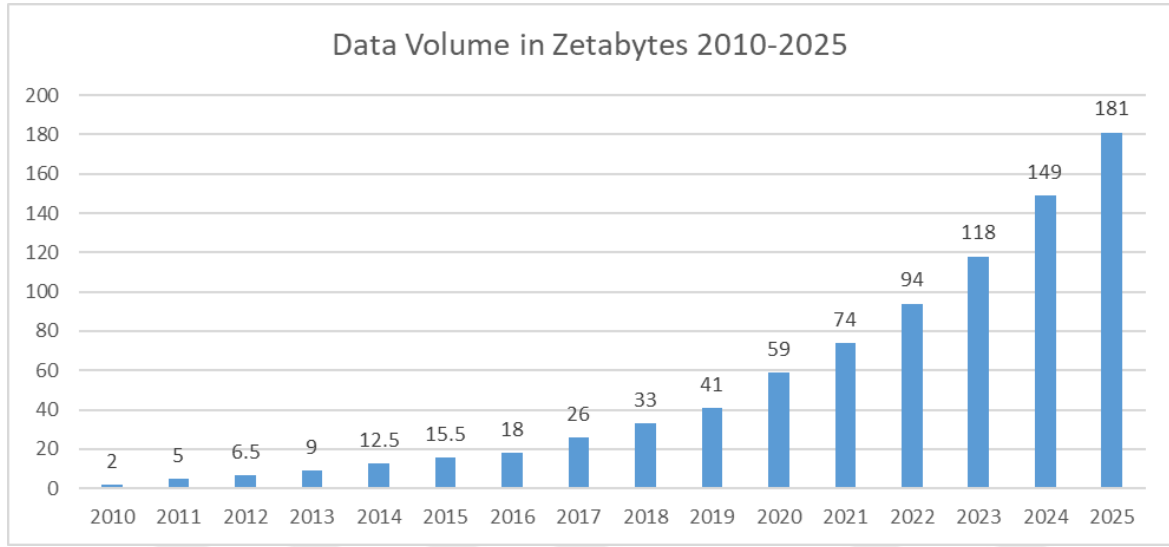
ABSTRACT

High-dimensional datasets have presented difficulties in terms of performance during analysis and modeling, as well as visualization and interpretation, as data volumes have increased at an exponential rate. As a result of this difficulty, the curse of dimensionality has been coined to describe it. Manifold learning methods are the topic of this study since they seek to alleviate the curse of dimensionality by decreasing the data sets' dimensions while maintaining their inherent linkages. While studies on the effectiveness of manifold learning algorithms in lowering dimensionality have been encouraging, most of this work has been conducted with either image or numerical datasets. As a result, there is an enormous gap in the research on adapting these algorithms for use with categorical data. To fill this void, this thesis first selects and evaluates prominent manifold learning algorithms on categorical datasets, then analyzes and discusses these results in depth. Manifold learning techniques Kernel PCA, Isomap, Locally Linear Embedding (LLE) (and its 3 variants) -, t-Distribution Stochastic Neighbor Embedding (t-SNE), and Uniform manifold approximation and projection (UMAP) were selected for this investigation. Researchers have paid close attention to these algorithms because of their success in several different fields. This study contributes to the growing body of knowledge regarding dimensionality reduction methods by exploring their utility in the context of categorical data. Examining how well these manifold learning algorithms capture and retain the underlying structure and relationships in categorical datasets is an essential part of evaluating these methods. Different evaluation measurement techniques are utilized to evaluate performance. The findings provide useful insights for researchers dealing with varied data types by revealing the strengths and limitations of manifold learning algorithms when applied to categorical data. The findings from this study will enhance the use of dimensionality reduction methods, leading to more accurate modeling and interpretation of high-dimensional categorical datasets.

Keywords: Dimensionality Reduction, Manifold Learning, Kernel PCA, Isomap, Locally Linear Embedding (LLE), t-Distribution Stochastic Neighbor Embedding (t-SNE), Uniform manifold approximation and projection (UMAP)

1. GİRİŞ

2010 yılında dünya çapında üretilen, elde edilen, kopyalanan ve tüketilen yaklaşık veri 2 ZB (10e9 TB) iken 2020 yılında yaklaşık 59 ZB olmuş ve yapılan tahminlere göre de 2025 yılında 181 ZB olması beklenmektedir [1]. Verilerin büyümesi ile veri analizi ve makine öğrenmesi üzerinde çalışmalar yoğunlaşmış ve yeni yöntemler keşfedilmeye başlanmıştır.



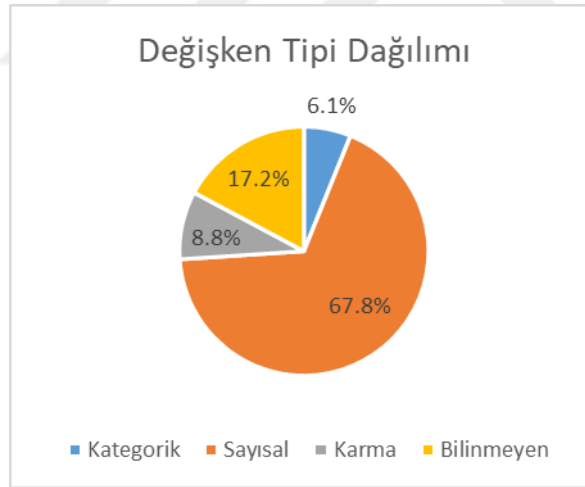
Şekil 1 Yıllara göre üretilen, elde edilen kopyalanan ve tüketilen veri miktarı (2004-2020)

Veri seti paylaşım sitelerinden Kaggle'a bakıldığında yaklaşık 224.000 veri seti kullanıcılarına sunulmuştur [2]. Benzer bir amaç ile veri seti paylaşan UCI makine öğrenme kaynağı sitesi ise 622 veri seti olmasına rağmen veri setleri hakkında kullanıcılarına bilgi vermektedir [3]. UCI sitesinde veri setinin örnek sayısı (number of instances) 1 milyondan fazla olan 31 veri seti bulunmakta ve değişken sayısı (attributes) yüzden fazla olan 110 veri seti bulunmaktadır. Veri analizi ve modellemelerde verinin sahip olduğu örnek sayısı arttıkça daha fazla tutarlı ve yansız sonuçlar elde edilebilmektedir. Aynı şekilde bağımsız değişken sayısı arttıkça da eğer bağımlı değişken ise denetimli makine öğrenmesi modellerinde sonuçlar daha tutarlı ve yansız çıkmakta ve gerçek dünyayı daha iyi mimik edebilmekte ve eğer bağımlı değişken yoksa da denetimsiz makine öğrenmesi modellerinde daha gruplama ve çıkarımlar yapılabilir.

Bununla beraber yüksek sayılı değişkenli veya bundan bahsedilecek şekliyle yüksek boyutlu veri setlerinde veriyi modellemek ve veriyi görselleştirmek zor bir süreçtir. Buna boyutsallık laneti (The Curse of Dimensionality) denmektedir [4]. Boyutsallık lanetinin üstesinden gelmek

için boyut indirgeme yöntemleri ortaya çıkmış ve ilk başta doğrusal yöntemler ile boyut indirgeme yöntemleri modellerde kullanılmaya başlanmıştır. Daha sonra ise yüksek boyutlu veri setlerindeki ilişkiyi bozmadan düşük boyutlarda da korumak için doğrusal olmayan boyut indirgeme algoritmaları ortaya çıkmış ve bu yöntemlere manifold öğrenme ismi de verilmiştir.

Manifold öğrenme yöntemleri daha çok resim temelli sayısal veri setleri üzerinde kullanılmıştır ve müşteri ilişkileri yönetimi veri setleri de hem kategorik hem de sayısal değişkenler içermektedir. UCI makine öğrenme kaynağı veri seti sitesine bakıldığında ise 622 veri setinin 38'i sadece kategorik değişkenlerden; 55'i ise hem kategorik hem de sayısal değişkenlerden oluşan veri setlerine sahiptir. Bu sebeple bu yüksek lisans tezi ile sadece kategorik değişkenlerden oluşan veri setleri için manifold öğrenme yöntemleri denenmiş ve sonuçları paylaşılmıştır. Böylece manifold öğrenme yöntemlerinde kategorik veri setleri üzerinden bir başarı yakalanırsa müşteri ilişkileri yönetimi veri setleri için de kullanılabileceği gösterilecektir.



Şekil 2 UCI Makine Öğrenme Kaynağındaki Veri Setlerinin Değişken Tipine Göre Dağılımı

Bu tezin ikinci bölümünde veri setlerindeki değişken tipleri ve veri ön işleme yöntemleri anlatılmış ve belirlenen modellerde kullanmadan önce veri setlerindeki verinin geçtiği adımlar anlatılmıştır. Ayrıca tez içerisinde kullanılan veri setleri de bu bölümde paylaşılmıştır.

Üçüncü bölümde boyut indirgeme yöntemleri ve doğrusal boyut indirgeme yöntemleri anlatılmıştır. Dördüncü bölümde manifold kavramı ve doğrusal olmayan boyut indirgeme yöntemleri (manifold öğrenme) açıklanmıştır.

Beşinci bölümde ise doğrusal olmayan boyut indirgeme yöntemlerini ölçmek için kullanılacak ölçüm teknikleri anlatılmış ve bu teknikler üzerinden üretilen yeni ölçüm tekniği de bu kısma eklenmiştir.

Altıncı bölümde belirlenen veri setlerinin yine belirlenen algoritmalar üzerinden çıkan sonuçları beşinci bölümde anlatılan ölçüm teknikleri ile ne kadar başarılı oldukları ve başarılı oldukları parametre değerleri paylaşılmıştır.

Yedinci ve Sekizinci bölümlerde tez sonuçları ve çıkan sonuçlar tartışılmıştır. Çıkan sonuçlar ile kategorik veri seti üzerinden yapılan çalışmaların müşteri ilişkileri yönetimi verileri için de kullanılabileceği gösterilmiştir.

2. VERİ SETİ ve VERİ ÖN İŞLEME TEKNİKLERİ

Veri seti, tablo formunda düşünülecek olursa her bir sütunun farklı bir değişkeni her bir satırın ise veri setine işlenen farklı bir kayıt olarak düşünülebilir [5]. Veri setleri içerisindeki değişkenler sayısal ve kategorik olarak iki ana gruba ayrılabilir [6]. Veri setleri sadece kategorik veya sayısal değişkenlerden oluşuyorsa sırasıyla kategorik ve sayısal veri setleri olarak adlandırılırken hem kategorik hem de sayısal veri seti içeren veri setleri de karma veri setleri olarak adlandırılmaktadır.

2.1. Sayısal (Nicel) Değişkenler

Kategoride sınıflandırma için kullanılan sayı değerleri hariç ölçülebilir karakteristiği ve değerleri sayı olan bütün değişkenler sayısal değişken olarak adlandırılır. Sayısal değişkenler de kendi içerisinde ikiye ayrılır:

2.1.1. Sayısal Sürekli Değişkenler

Sayısal değişkenin aralığı sonsuz ve reel sayılardan oluşuyorsa bu tür sayısal değişkenlere sürekli değişkenler denir. Örneğin sınıftaki öğrencilerin boyu bir sayısal sürekli değişkenlerdir.

2.1.2. Sayısal Ayırık Değişkenler

Sayısal sürekli değişkenlerin aksine sayısal ayırık değişkenlerde belirtilen aralık sınırlı reel sayılardan oluşmaktadır. Örneğin sınıftaki öğrencilerin final sınavından alacağı not 0 ve 100 arasındadır fakat 101 olamaz ve bu sayısal ayırık değişken olarak adlandırılır.

2.2. Kategorik Değişkenler

Sayısal değişkenlerin aksine sayısal olarak ölçülebilir karakteristiği olmayan nitel değişkenlere kategorik değişkenler adı verilir. Kategorik değişkenler de kendi içerisinde ikiye ayrılır.

2.2.1. Sınıflayıcı (Nominal) Değişkenler

Hiçbir şekilde sıralayıcı bir değer vermeden değişkeni isim, etiket veya kategori olarak tanımlayan değişkenlere sınıflayıcı değişkenler denir. Örneğin işe gitmek için kullanılan taşıt türlerini veren değişkenin değerleri “araba, otobüs, minibüs” olarak tanımlanıyorsa bu kategorik sınıflayıcı değişken olarak konumlandırılabilir.

2.2.2. Sıralayıcı (Ordinal) Değişkenler

Değişkenin alabileceği değerler belli bir sıra üzerinden anlam ifade ediyorsa buna sıralayıcı değişkenler denir. Örneğin davranış değerlerini alan bir değişken “mükemmel, iyi, ve kötü” değerlerini alıyorsa her ne kadar ölçülenebilir bir şekilde kıyaslayamasak bile kendi içinde birinin diğerinden daha iyi olduğu diğer bir deyişle bir sıralamanın mevcut olduğu bilgisine varabiliriz.

2.3. Veri Ön İşleme Teknikleri

Veri ön işleme ilgili veri setini kullanarak yapılacak analiz veya modeli yanlış sonuçlara yönlendirmemesi, modelin performansından emin ve model performansının yüksek olması için ham veriyi anlaşılabilir bir formata çevirdiğimiz işlemlere denir [7].

2.3.1. Veriyi Yeniden Ölçeklendirme (Rescaling Data)

Bu işlemde değişken aldığı minimum ve maksimum değer üzerinden oranlanıp 0 ve 1 arası değer üzerinde tekrar konumlandırılır. Değişken değerlerinin dağılımını ve pozisyonlarını değiştirir. Aykırı değerlere karşı hassas bir işlemdir.

$$x_{norm} = (x - x_{min}) / (x_{max} - x_{min})$$

2.3.2. Ortalama Kaldırma (Mean Removal)

Ortalama kaldırma işleminde değişkenin ortalama değer hesaplanıp bütün kayıtlardan ortalama çıkartılır ve değişken değerlerinin merkezi 0 olup 0 etrafında değişkenin aldığı değerler dağılır.

$$x_{mr} = x - \mu$$

2.3.3. Veriyi Standartlaştırma (Standartizing Data)

Veriyi standartlaştırma işleminde değişkenin normal (Gaussian) dağılıma sahip olduğu kabul edilip ortalaması 0 ve standart sapması 1 olarak alınır. Bu kabul üzerinden veri yeniden ölçeklendirilir.

$$x_{stand} = (x - \mu) / \sigma$$

2.3.4. Tek Dönüşüm (One Hot Encoding)

Kategorik verileri bir düzlem veya grafik üzerinde göstermek için ilgili verileri ikili sistemde vektörlere çevrilmesine nedir. Kategori değişkenindeki bütün değerler değerin atandığı vektör sırasında 1, vektörün diğer değerlerinde 0 olacak şekilde dönüşümü yapılır.

2.3.5. Veriyi Normalleştirme (Normalizing Data)

Bu işlemde veri seti içindeki her değişken değerinin bulunduğu kaydın büyüklüğüne bölünmesi ile veri seti orijine doğru konumlandırılır.

$$x'_i = x_i / \sqrt{\sum_{j=1}^n x_j^2}$$

2.3.6. Eksik Değerleri Tamamlama (Replacing Missing Values)

Veri setindeki bir kayıta yer alan değişken değerinin eksik olması yapılacak analizi veya modeli yanlış yönlendirebileceği ihtimali olduğu için değişkenin aldığı bütün değerler dikkate alınarak bir strateji belirlenmelidir. Buradaki bakış açısı veri setini çok yanlış hale getirmeden gerekirse değişkenlerin ortalaması veya medyan gibi değerler ile doldurulmasıdır.

2.4. Kullanılan Veri Setleri

Müşteri İlişkileri Yönetimi veri setleri sadece sayısal değil aynı zaman da kategorik değişkenler içermektedir. Manifold öğrenme için kullanılan algoritmaları incelendiğinde kullanılan veri setleri genellikle imaj (resim) veri setleri ve sayısal değişkenleri içeren veri setleri olduğu gözlemlenmiştir. Bunun üzerine çalışma kategorik değişkenler içeren veri setleri üzerinden ilerlenmiştir.

UCI Makine Öğrenmesi Deposunda yer alan sınıflandırma problemi için eklenmiş ve bütün değişkenleri kategorik olan 13 adet veri seti seçilmiştir. “Audiology (Original) [8]”,

“Balance Scale [9]”, “Breast Cancer [10]”, “Car Evaluation [11]”, “Congressional Voting Records [12]”, “Hayes-Roth [13]”, “Lenses [14]”, “Lymphography [15]”, “Molecular Biology (Promoter Gene Sequences) [16]”, “Primary Tumor [17]”, “Soybean (Large) [18]”, “Soybean (Small) [19]”, “Tic-Tac-Toe Endgame [20]” isimindeki veri setleri belirlenen manifold öğrenme modelleri için kullanılmıştır.

Kayıt sayısı 10 ile 100 arası olan 2, 100 ile 250 arası 4, 250 ile 500 arası 4, 500 ile 1000 arası 2 ve 1000 ile üzeri 1 veri seti bulunmaktadır. Değişken sayısı 10’dan az olan 6 ve 10 ile 100 arası 7 veri seti bulunmaktadır. Değişkenler kategorik olduğu için algoritmalarda kullanılması için tek dönüşüm veri ön işleme tekniği kullanılarak değişkenler düzenlenmiş ve her bir veri setindeki değişken sayısında artış olmuştur. Tek dönüşümden sonra değişken sayısı 10’dan az olan 1 veri seti, 10 ile 100 arası 9 veri seti ve 100 ile üzeri 3 veri seti elde edilmiştir.

Çekirdek Temel Birleşen Analizi ve t-Dağıtılmış Stokastik Komşu Gömme (t-SNE) algoritmaları için 13 veri seti; İzomap ve Lokal Doğrusal Gömme (LLE) (diğer 3 varyantı ile) için 12 veri seti, Standart Manifold Yaklaşımı ve Projeksiyonu (UMAP) için ise 9 veri setinin sonuçları paylaşılmıştır. Burada bütün veri setlerinin bütün algoritmalar için sonuçlarının paylaşılmamasının sebebi bazılarının ilgili algoritma için çok uzun süren öğrenme sürecinin olması veya değerlendirme ölçüm tekniklerinin tutarlı ve yayımlanabilir sonuçlar vermemesinden kaynaklanmaktadır.

Tablo 1 Analizde Kullanılan veri setleri

#	Veri Seti Adı	Kayıt Sayısı	Kayıt Grubu	Değişken Grubu	Değişken Sayısı	Değişken Sayısı (Tek Dönüşüm)
1	Audiology (Original)	200	100 ve 250 arası	10 ve 100 arası	69	354
2	Balance Scale	625	500 ve 1000 arası	10'dan az	4	20
3	Breast Cancer	286	250 ve 500 arası	10'dan az	9	43
4	Car Evaluation	1,728	1000 ve üzeri	10'dan az	6	21
5	Congressional Voting Records	435	250 ve 500 arası	10 ve 100 arası	16	48
6	Hayes-Roth	160	100 ve 250 arası	10'dan az	5	15
7	Lenses	24	10 ve 100 arası	10'dan az	4	9
8	Lymphography	148	100 ve 250 arası	10 ve 100 arası	18	59
9	Mol. Biol. (Promoter Gene Seqs.)	106	100 ve 250 arası	10 ve 100 arası	58	234

10	Primary Tumor	339	250 ve 500 arası	10 ve 100 arası	17	43
11	Soybean (Large)	307	250 ve 500 arası	10 ve 100 arası	35	132
12	Soybean (Small)	47	10 ve 100 arası	10 ve 100 arası	35	72
13	Tic-Tac-Toe Endgame	958	500 ve 1000 arası	10'dan az	9	27

3. BOYUT İNDİRGEME

Boyut, ilgili veri setinin değişkenlerinin sayısına denir. Boyut arttıkça modelleme zorlaşır, bu durum genel bir durumdur ve buna boyutluluk laneti denir [4]. Veriyi görselleştirmek için de yüksek boyutlu veriyi azaltmak gerekmektedir. Boyut indirgeme, yüksek boyutlu veri kümesindeki değişken sayılarını azaltmayı amaçlamaktadır. Bu yöntem istatistikte, makine öğreniminde ve veri analizi projelerinde oldukça kullanılmaktadır. Kullanım alanına bir örnek vermek gerekirse; makine öğrenimi sınıflandırma problemlerinden bahsedebiliriz. Genelde son sınıflandırma sırasında çok fazla faktör oluşturulmuş olur ancak bu faktörlerin artımı modelin eğitilmesini ve üzerinde çalışılmasını zorlaştırmaktadır. Bu nedenle ilişki bakımından gereksiz görülen bazı özelliklere karşı boyut indirgeme yöntemleri kullanılmaktadır. Boyut indirgeme yönteminde özellik seçimi ve özellik çıkarma olarak iki yaklaşım söz konusudur.

Özellik seçimi, değişkenlerinin alt kümesine ulaşarak modeli daha kısa sürede eğitilmesini sağlamakta, modeli basitleştirerek anlaşılabilirliğini arttırmaktadır. Özellik çıkarma ise var olan değişkenler kullanılarak yeni bir değişken elde edilmesidir.

3.1. Özellik Seçimi (Feature Selection)

Temel amacı değişkenlerin alt kümesine ulaşmaktır. Bu yöntemin kullanım nedeni ise modeli daha kısa sürede eğitebilmek, modelin basitleştirilmesi ve bu sayede anlaşılabilirliğinin artmasını sağlamak, çok sayıda boyut çıkmazından kurtulmak, varyansın azaltılması olarak sıralanabilir [21]. Filtre yöntemi, sarma yöntemi ve gömülü yöntem olarak temelde kullanılan 3 yöntem mevcuttur.

3.1.1. Filtre Yöntemi (Filter Method)

Filtre yönteminde yapılan işlem değişkenleri kendi içerisinde değerlendirip belli bir eşik değeri altındaki değişkenleri eleyerek boyut sayısını düşürmeyi hedeflemektedir. Burada kullanılan kriterler değişmek ile örnek olarak ilgili değişkenlerin varyansı üzerinden bir çıkarım yaparak; her bir kayıt için çok fazla değişmeyen yani varyansı düşük olan değişkenleri çıkartmak verilebilir.

3.1.2. Sarma Yöntemi (Wrapper Method)

Sarma yönteminde değişkenler içerisinde ilgili modelleme veya analiz için eğer bütün değişkenler kullanılsaydı aynı veya yakın sonucu verecek alt küme bulunması amaçlanır. Burada ideal sonucun yakalanması için bazen sırasıyla bir değişken çıkartılarak iteratif adımlar ile ileni veya modelin veya analizin gücünü arttırana kadar değişken eklemesi yapılır ve bu iteratif süreç model veya analiz ideal sonuç verdiğinde değişken ekleme işlemi durdurulur.

3.1.3. Gömülü Yöntem (Embedded Method)

Bu yöntem ise maliyet fonksiyonu üzerinden kullanılan değişken sayısına bağlı bir parametre eklenir. Ve değişken sayısı arttıkça da doğru orantılı olarak bu parametrenin değeri artar. Bu maliyet fonksiyonun amacı maliyeti minimize etmek olduğu için model her zaman değişken sayısını azaltmaya çalışacaktır. Bu yöntem ile özellik (değişken) seçimi model geliştirmesi ile kullanılır.

3.2. Özellik Çıkarma (Feature Extraction)

Özellik çıkarma yönteminin özellik seçiminden farkı, var olan değişkenler üzerinden seçim yapmak yerine var olan değişkenleri kullanarak yeni bir değişken üretmektedir [22]. Bu yöntem ile hem toplam değişken sayısı azaltılır hem de değişkenlerin özellikleri yeni oluşturulan değişkene aktarıldığı için veri setinin model ve analize katkısı azaltılmadan değişken sayısı azaltılmış olur. Özellik çıkarma adımları doğrusal ve doğrusal olmayanlar

olarak ikiye ayrılır. Ve bu tezin de konusu olan manifold öğrenme aslında doğrusal olmayan boyut indirgeme adımlarına verilen genel isimdir.

3.3. Doğrusal Boyut İndirgeme Yöntemleri

3.3.1. Temel Bileşenler Analizi (Principal Component Analysis, PCA)

Temel bileşenler analizi (PCA), görüntü boyut indirgeme ya da sıkıştırma ve sınıflandırma gibi alanlarda kullanılan bir istatistik yöntemidir. Yani fazla boyutlu bir verinin az boyutlu veriye dönüşümünü en yüksek varyans olacak şekilde yapmayı amaçlamaktadır. Birbiri ile ilişkili çok sayıda değişkenin doğrusal fonksiyonlarına bakarak daha az sayıda ve ilişkisiz ama aynı zamanda da önceki sistemdeki değişimleri büyük oranda açıklayan yeni değişkenler oluşturulur [23]. Bu yeni değişkenlere de temel bileşenler ismi verilir. Örnek vermek gerekirse 100 adet başlangıç değişkeni aynı zamanda 100 adet temel bileşene sahiptir çünkü her temel bileşen bütün değişkenlerden bilgiye sahiptir. Böylece PCA 100 boyutlu veri yerine 70 adet önemli temel bileşeni kullanarak boyut indirgemesi sağlamaktadır. Bahsedilen 70 temel bileşen toplam varyansın büyük bir kısmını karşılayabilir ise 30 adet bileşenin ihmal edilebileceği görülür. Doğru ve klasik değişken seçimi yöntemleri kullanılırsa PCA ile veri kaybının oldukça aza indirgenebileceği görülür.

Temel bileşenler analizi; standardizasyon, kovaryans matrisin hesaplanması, özdeğer (eigenvalue) ve özvektör (eigenvektör) değerlerinin hesaplanması, temel bileşenleri belirlemek için özellik vektörlerini (feature vectors) oluşturma ve temel bileşenlerin eksenleri üzerinden veriyi tekrar oluşturma olarak 5 adımdan oluşmaktadır.

Birinci adım yani standardizasyon adımıda veri seti içerisinde yer alan her değişken; ortalama ve standart sapma üzerinden standardize edilir. Bu adımı yapmaktaki ana sebep aldığı değerlerin aralığı geniş olan bir değişkenin aralığı dar olan bir değişken üzerinde etkisini azaltmasıdır. Bunun önüne geçmek için her değişkenin aldığı değer aralıklarını ortalamayı merkezi noktasına çekerek aynı yapmaktır.

İkinci adımda kovaryans matrisi hesaplanır. Bu adımda bütün değişkenlerin birbiri ile olan ilişkileri kovaryans üzerinden hesaplanır ve ilgili matris oluşturulur. Burada kovaryansın pozitif ve yüksek olması 2 değişken arasında doğru orantılı yüksek bir ilişki olduğunu, negatif ve yüksek olması 2 değişken arasında ters orantılı yüksek bir ilişki olduğunu göstermektedir. Bu adım ile yüksek ilişkisi olan değişkenlerden birinin fazla ve gereksiz olduğu bilgisi elde edilir.

Üçüncü adımda kovaryans matrisi üzerinden özdeğerler ve özvektörler hesaplanır. Kovaryans matrisinin özvektörleri en büyük varyansın olduğu eksenin yönünün bilgisini verirken ilgili özvektör ile ilişkili özdeğer varyansın miktarını vermektedir. Bu yüzden orijinal veri setindeki bilgileri en çok tutan bilgileri tutan eksenleri (öz vektörleri) bulmak için öz değerler büyüktür küçüğe önem sırasına göre sıralanır.

Dördüncü adımda özellik vektörlerinin seçilmesidir. Bu adım direkt olarak boyut indirgeme adımının gerçekleştiği adım olarak özetlenebilir. Yüksek boyutlu bir veri setinden özdeğerlerin büyüklük sırasına göre boyut sayısından daha düşük sayı kadar ilişki öz vektörleri seçilerek boyut indirgeme gerçekleştirilir. Örneğin 10 boyutlu (10 değişkene sahip) veri setinden yukarıdaki adımlar takip edilip özdeğerleri en yüksek 2 özvektör seçilirse boyut iki boyuta inmiş olur.

Beşinci ve son adımda da belirlenen özellik vektörleri ile belirlenen eksenleri standardize edilmiş orijinal veri seti ile çarparak (ikisinin de transpoze edilmiş hali ile) temel bileşenler analizinin sonucu elde edilmiş olur.

2013-2023 yılları arasında IEEE Xplore’da anahtar kelimesinde temel bileşenler analizi (Principal Component Analysis, PCA) içeren çalışmalar araştırıldığında 6,798 sonuç gelerek hala çalışmalarda kullanılan bir yöntem olduğunu göstermektedir [24].

3.3.2. Doğrusal Ayırıştırma Analizi (Linear Discriminant Analysis, LDA)

Doğrusal ayırıştırma analizi (LDA) sınıflandırma metodu olarak kullanılır ve oldukça basit olmasına rağmen karışık sorunlarda çok iyi sonuçlar sağlamaktadır [25]. Yöntemdeki temel amaç, hesaplama maliyetlerini düşürürken aşırı uyumu (overfitting) da engellemektir. Temel

bileşenler analizinden (PCA) farkı ise Doğrusal ayırıştırma analizi (LDA) sınıflar arasındaki uzaklığı maksimize etmesi ve LDA denetimli öğrenme modeli iken PCA denetimsiz öğrenme modelidir. LDA, çeşitli veri öğelerini kullanarak birçok nesne ve öge sınıfını ayrı ayrı analiz ederek bu kümeye işlevler uygular. Doğrusal ayırıştırma ve mantıksal regresyonlarının yetersiz kaldığı ikiden fazla sınıf analizleri olduğunda ihtiyaç duyulur. Özellikle veri ön işleme aşamasında boyut azaltma yöntemi olarak makina öğrenimi projelerinde tercih edilir. Ayrıca görüntü tanıma ya da tahmine dayalı analitik gibi alanlarda kullanılabilir.

Doğrusal ayırıştırma analizi; sınıfların ortalama vektörlerinin hesaplanması, sınıflar içinde ve arasındaki dağılım matrislerinin hesaplanması, dağılım matrislerinin özdeğerler (eigenvalues) ve özvektörlerinin (eigenvectors) hesaplanması, Özdeğerlere göre sıralanan özvektörlerden düşürülmesi istenen boyuta kadar olanların seçilmesi ve seçilen özvektörler ve orijinal veri seti üzerinden boyut indirgenmiş veriyi oluşturma olarak 5 adımdan oluşmaktadır.

Sınıfların ortalama vektörlerinin hesaplanması olan birinci adımda, veri setindeki bağımlı değişkenler üzerinden belirlenen sınıflar için ortalama vektörler hesaplanır. Ortalama vektörler ilgili sınıf içerisinde kalan kayıtlar üzerinden her bir değişkenin o sınıf için ortalama değerinin hesaplanmasıdır.

İkinci adımda sınıflar içerisinde ve sınıflar arasında dağılımı veren 2 matris oluşturulur. Her sınıf için kendi içerisinde dağılım matrisleri hesaplanır ve böylece sınıflar içerisinde kovaryans değerleri hesaplanır. Sınıflar arasında ise sınıf içerisinde uygulanan adımın aynısı burada da tekrarlanıp kovaryans ilişkisi yakalanmaya çalışılır.

Üçüncü adımda sınıflar içerisinde ve sınıflar arası hesaplanan matrislerin çarpımı üzerinden özdeğerler ve özvektörler hesaplanarak elde edilen özvektörler ile eksenler belirlenir. Özvektörlerin varyansı en çok yansıtan değerler için özdeğerler büyükten küçüğe sıralanarak ilgili özvektörler bulunur.

Dördüncü adımda ise sıralanan özdeğerler üzerinden indirgenecek boyut sayısı kadar özvektörler seçilir. Burada seçilen özvektörler yeni tanımlanacak eksenlerin yönünü

vermektedir. Özdeğerler büyükten küçüğe göre sıralandığı için de seçilen ilgili özvektörler veri seti hakkında en fazla bilgiyi içeren vektörlerdir.

Son ve beşinci adımda da dördüncü adımda seçilen özvektörler üzerinden oluşturulan matris ile orijinal veri seti ile çarparak doğrusal ayrıştırma analizinin sonucu elde edilmiş olur.

2013-2023 yılları arasında IEEE Xplore’da anahtar kelimesinde doğrusal ayrıştırma analizi (Linear Discriminant Analysis, LDA) içeren çalışmalar araştırıldığında 226 sonuç gelerek hala çalışmalarda kullanılan bir yöntem olduğunu fakat PCA kadar popüler olmadığını göstermektedir [24].

3.3.3. Tekil Değer Ayrışımı (Singular Value Decomposition, SVD)

Tekil değer ayrışımı (SVD), gerçek olan ama kompleks olmayan bir matrisin ayrıştırılma sürecinde kullanılabilmektedir. Eğer elimizde bir m satır ve n kolona sahip M matrisi var ise $M=U\Sigma V^T$ olarak ifade edilebilmektedir. U , Σ ve V matrisleri tekil olup U ve V matrisleri ortonormal matrislerdir. Ortonormal matris ise bütün kolonların Öklid uzunluğu 1 olan ve bütün kolonlarının birbirine ortogonal olması veya diğer bir deyişle nokta çarpımlarının 0 olmasıdır. Σ ise diyagonal matris olup bütün değerleri pozitif ve azalan sıra ile değerler sıralanmıştır. U matrisi m satır ve r kolondan oluşur ve bu matrise sol tekil vektör ismi verilir. Σ matrisi tekil değerlerden ve r satır ve kolondan oluşmaktadır. V matrisi n satır ve r kolondan oluşur ve bu matrise sağ tekil vektör ismi verilir.

Boyut indirgeme işlemleri buradaki Σ diyagonal matrisin azalan değer ile sıralanan değerler içerisinde en güçlü tekil değerleri seçilmesiyle başlanmaktadır. Seçilen tekil değerler ile ilişki U matrisindeki vektörler bize orijinal M matrisindeki en yüksek ve önemli bilgiyi verecektir. U_b seçilen tekil değerler ile ilişkili sol tekil vektörler iken Σ_b seçili tekil değerleri içeren matrise karşılık gelmektedir. İkisinin matris çarpımı yani $B = U_b \Sigma_b$, orijinal M matrisinin boyut indirgenmiş ve içindeki bilgiyi en çok koruyan matrisi vermiş olmaktadır.

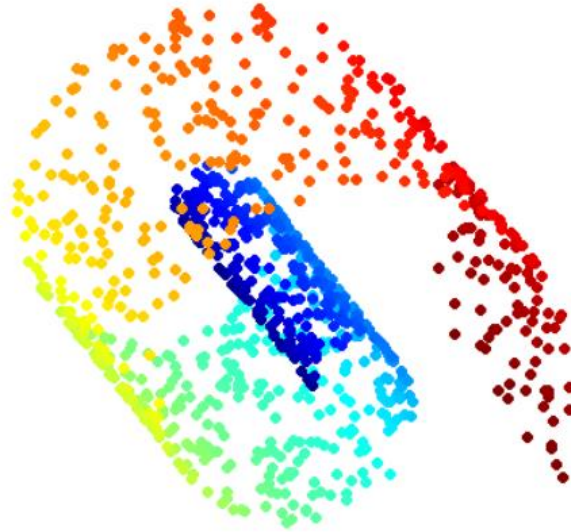
2013-2023 yılları arasında IEEE Xplore’da anahtar kelimesinde tekil değer ayrışımı (singular value decomposition, SVD) içeren çalışmalar araştırıldığında 2.530 sonuç gelmektedir. Fakat boyut indirgeme (dimensionality reduction) eklendiğinde ise 17 adet

sonuç gelmektedir. Bu bize çalışmalarda kullanılan bir yöntem olduğunu fakat PCA kadar popüler olmadığını göstermektedir [24].

4. MANİFOLD ÖĞRENME

4.1. Manifold Kavramı ve Manifold Öğrenme

Matematikte manifold yerel olarak her noktanın yakınında Öklid uzayına benzeyen topolojik bir uzaydır [26]. Kabaca herhangi bir şeklin yüzeyi olarak düşünülebilir. Bu tanımdan yola çıkarak yüksek boyutlu uzay içerisinde yüksek boyutlu veri setlerinin gerçek hayatta daha düşük boyutlu manifold üzerinde yer alabilmektedir [27]. Böylece yüksek boyutlu veri setleri daha düşük boyutlu veri setleri üzerinden veri setinin verdiği özellikler ve bilgi kaybolmadan gösterilebilmektedir. Resim 1’de 3 boyutlu uzayda yer alan şeklin aslında bir manifold olduğu ve 2 boyutlu düzlemde de verilerin ilişkisi bozulmadan da gösterilebileceği görülmektedir.



Resim 1 İsviçre Rulosu

Matris formunda satırların kayıtları sütunların değişkenlere karşılık geldiği bir veri seti düşünüldüğünde bütün değişkenleri analiz veya modelde kullanmak yerine bütün

değişkenlerin ilişkisini tutan daha az değişken ile analiz ve modelde kullanılarak algoritmaları daha etkili ve daha az kompleks yapmaktadır. Ayrıca yüksek boyutlu verileri de daha düşük boyutlarda daha rahat görselleştirilmesini sağlamaktadır.

4.2. Manifold Öğrenme Algoritmaları

4.2.1. Çok Boyutlu Ölçekleme (Multidimensional Scaling, MDS)

Çok Boyutlu Ölçekleme (MDS) modelinde amaç ilgili kayıt için diğer noktalar ile arasındaki “uzaklık” ilişkisini koruyarak daha düşük boyutlu uzayda veri setini temsil etmektir [28]. Kayıtlar arasındaki ilişkinin daha düşük boyutlarda korunması hedeflenmektedir. MDS kayıtlar arası uzaklıkların belirlenmesi ve optimizasyon problemin çözülmesi olarak iki adımdan oluşmaktadır.

İlk adımda satırların kayıt sütunların değişken olduğu tablo formunda bir veri seti düşünüldüğünde her bir kaydın ya da noktanın diğer kayıtlar ve noktalar arası Öklid uzayda uzaklıkları hesaplanır. MDS için bütün bu uzaklıkların bir bütün olarak daha düşük boyutlarda korunması hedeflenmektedir.

İkinci adımda ise bu yukarıdaki ilişkiyi korumayı hedeflenen bir maliyet fonksiyonu oluşturulur ve bu maliyet fonksiyonunu minimize eden optimizasyon problemi çözülür. Burada ilk adımda hesaplanan uzaklıklar düşük boyuttaki uzaklıklar ile arasındaki toplam farkı minimize edilmesi hedeflenmektedir.

4.2.2. Çekirdek Temel Birleşen Analizi (Kernel Principal Component Analysis)

Temel birleşen analizi algoritması üzerinden türeyen çekirdek temel birleşen analizi doğrusal olmayan boyut indirgeme algoritmalarındandır [29].

Temel birleşen analizi ile arasındaki tek fark ilk adımda yapılan çekirdek fonksiyonunun (Kernel Function) kullanılmasıdır. Doğrusal olan temel birleşen analizinde değişkenlerin birbirileri ile olan ilişkileri kovaryans üzerinden giderken çekirdek temel birleşen analizinde

kovaryans yerine kayıtlar arası ilişkiyi verecek bir çekirdek fonksiyonu işleminden geçmektedir.

$$K = k(x_i, x_j)$$

Kernel veri setinde yakalanacak ilişkiye seçilebilmektedir. ($k(x_i, x_j) = (x_i^T x_j + 1)^2$ ya da $e^{-(x_i - x_j)^2 / 2\sigma^2}$ olabilmektedir.

4.2.3. İzomap (Isomap)

İzomap algoritması MDS ile aynı bakış açısı ile çalışmaktadır [30]. İzomap algoritması, komşulukların belirlenmesi, noktalar arasındaki en kısa yolun bulunması ve düşük boyutlu gömme oluşturma olarak 3 adımdan oluşmaktadır.

Birinci adımda yani komşulukların belirlenmesinde ya belirli sayıda komşu seçilerek ya da belirli bir uzaklık belirlenerek ilgili kayıt veya noktanın komşulukları bulunur. Burada belirlenen komşular üzerinden graf (graph) tanımlanır. Bu graf için her bir nokta köşe (vertice) ve her iki nokta eğer komşu ise aralarındaki uzaklık kenar (edge) olarak belirlenir. Burada komşuluk sayısının veya komşulukların belirlenmesi için seçilen uzaklığın iyi seçilmesi gerekmektedir. Komşuluk sayısı arttıkça veya uzaklık büyüdükçe düşük boyutluda yakalanmak istenen manifold yapısını bulmak zorlaşacak ve tam tersi komşuluk sayısı düştükçe veya uzaklık küçüldükçe veri setinin içerdiği bilgi ve düşük boyutluda korunması istenen bilgi daha az korunacaktır.

İkinci adımda tanımlanan grafın köşelerinin birbirileri olan en kısa yol bulunur. Köşelerin veya veri setindeki noktaların diğer bütün köşelere veya veri setindeki diğer noktalara olan uzaklıkları bulunur. Bunun için ilk adımda oluşturulan yapıda noktaları graf teorisindeki köşe, birbirileri arasındaki uzaklıkları kenar (vertice) olarak düşünülebilir. Eğer iki kayıt arasında komşuluk yok ise bu kısımda aralarındaki uzaklık sonsuz yani kenar yok gibi hesaplanır. Dijkstra algoritması [31] gibi algoritma ile bütün köşelerin birbirine olan en kısa mesafesi bulunur ve final graf üzerinden uzaklıkları içeren matris oluşturulmuş olur.

Üçüncü ve son adımda ise düşük boyutlu gömme oluşturulur. Burada kullanılan matris veri setinin kendisi değil ilk iki adımda hesaplanan noktaların birbirine olan uzaklıklarından veya oluşturulan graf kenarlarından ortaya çıkmıştır. Burada doğrusal boyut indirgeme adımlarındaki SVD algoritmasındaki gibi matris çözümlenir. Varyansı yani içerisindeki bilgiyi gösteren en yüksek özdeğerler ve ilgili özvektörlerin çarpımı ile boyut indirgeme gerçekleştirilir.

İzomap algoritmasının kullanımı incelendiğinde 2013-20223 yılları arasında Google Scholar verilerine göre alıntılanma sayısı 9,700 olarak gelmektedir [32]. Bu doğrusal olmayan boyut indirgeme algoritma içerisinde popüler bir algoritma olduğunu gösterdiği için bu tez içerisinde analiz adımlarında kullanılmıştır.

4.2.4. Lokal Doğrusal Gömme (Locally Linear Embedding, LLE)

Lokal doğrusal gömme (LLE) algoritması doğrusal olmayan boyut indirgeme algoritmalarındandır [33]. Lokal doğrusal gömme algoritması her bir noktaya ait komşuların bulunması, her nokta ve ilgili komşuların arasındaki doğrusal ağırlıkların hesaplanması ve daha düşük boyutlu uzayda bu noktalara ve ilgili komşularına bağlı doğrusal ağırlıkları koruyan noktaların bulunması olarak 3 adımdan bulunmaktadır.

İlk adımda yani her noktaya ait komşulukların belirlenmesinde ya belirli sayıda komşu seçilerek ya da belirli bir uzaklık belirlenerek ilgili kayıt veya noktanın komşulukları bulunur. İzomap algoritmasındaki gibi algoritmadan düşük boyuttaki manifoldu yakalamak için komşuluk seçimi önemli bir adımdır.

İkinci adımda ise her bir nokta ve ilgili noktanın komşuları arasındaki ilişki üzerinden doğrusal ağırlıklar hesaplanmasıdır. Bu doğrusal ağırlıklar öyle bir seçilmelidir ki ilgili noktaların ağırlıkları ile doğrusal kombinasyonları o noktayı vermelidir. Böylece yüksek boyuttaki noktaların ilişkileri algoritmada kullanılmak üzere belirlenmiş olmaktadır.

Üçüncü ve son adımda ise daha düşük boyutta bir önceki adımda hesaplanan doğrusal ağırlıkları koruyan yeni noktaların bulunmasıdır. Bu işlemin yapılması için maliyet

fonksiyonu oluşturulur ve maliyet fonksiyonunu minimize edilmesi hedeflenir. Maliyet fonksiyonu:

$$\phi(Y) = \sum_i |Y_i - \sum_j W_{ij} Y_j|^2$$

Bu fonksiyonda, Y yeni oluşturan noktaları içeren matrisi, W yüksek boyutlu uzaydaki veri setindeki noktaları ve ilgili noktaların komşulukları üzerinden belirlenen doğrusal ağırlıkları ifade etmektedir.

Lokal doğrusal gömme algoritmasının kullanımı incelendiğinde 2013-20223 yılları arasında Google Scholar verilerine göre alıntılanma sayısı 11,300 olarak gelmektedir [34]. Bu doğrusal olmayan boyut indirgeme algoritma içerisinde popüler bir algoritma olduğunu gösterdiği için bu tez içerisinde analiz adımlarında kullanılmıştır.

4.2.5. T-Dağıtılmış Stokastik Komşu Gömme (T-Distributed Stochastic Neighbor Embedding, t-SNE)

T-dağıtılmış stokastik komşu gömme (T-SNE) doğrusal olmayan boyut indirgeme yöntemlerindendir [35]. T-SNE de diğer manifold öğrenme modelleri gibi noktalar arası uzaklığı algoritmasında dikkate almakta ve her bir noktanın diğer noktaları komşu olarak seçip seçmeyeceğini olasılık dağılımı üzerinden bulmaktadır. Bu dağılımı hesaplarken Gaussian dağılım üzerinden ortalamayı komşulukları hesaplanacak nokta olarak belirlenir ve aday komşuluklar bu noktaya olan uzaklığa göre Student'in t dağılımı üzerinde konumlandırılır.

Yüksek boyutlu uzayda yer alan noktaların birbirlerine olan uzaklıklara göre hesaplanan olasılık dağılımları düşük boyutlu uzayda da devam ettirilmesi hedeflenmekte ve t-SNE algoritmasında korunması gereken bu bilgi üzerinden boyut indirgenmesi yapılmaktadır. Burada p_{ij} yüksek boyutlu uzayda x_i ve x_j noktalarının komşu olma olasılıklarını verirken, q_{ij} düşük boyutlu uzayda y_i ve y_j noktalarının komşu olma olasılıklarını vermektedir.

$$p_{ij} = \frac{e^{(-|x_i - x_j|^2 / 2\sigma_i^2)}}{\sum_{k \neq l} e^{(-|x_k - x_l|^2 / 2\sigma_i^2)}}$$

$$q_{ij} = \frac{(1 + |y_i - y_j|^2)^{-1}}{\sum_{k \neq l} (1 + |y_k - y_l|^2)^{-1}}$$

T-SNE algoritması bu ilişkiyi kurduktan sonra Kullback-Leibler sapması [36] kullanarak bu dağılımı optimize etmeye çalışmaktadır.

$$\phi = \sum_{i \neq j} p_{ij} \log p_{ij} / q_{ij}$$

T-dağıtılmış stokastik komşu gömme (T-SNE) algoritmasının kullanımı incelendiğinde 2013-20223 yılları arasında Google Scholar verilerine göre alıntılanma sayısı 32,600 olarak gelmektedir [37]. Bu doğrusal olmayan boyut indirgeme algoritma içerisinde popüler bir algoritma olduğunu gösterdiği için bu tez içerisinde analiz adımlarında kullanılmıştır.

4.2.6. Standart Manifold Yaklaşımı ve Projeksiyonu (Uniform manifold approximation and projection, UMAP)

Standart manifold yaklaşımı ve projeksiyonu algoritması (UMAP); görsel kalite için t-SNE algoritması ile yarışabilecek, yüksek boyutlu uzaydaki yapıyı koruyup boyut indirgemeyi üstün çalışma süresi performansı ile gerçekleştirebilecek bir algoritma olarak sunulmuştur [38].

UMAP algoritmasının t-SNE algoritmasından farklılaştığı yer komşuluk seçiminde yapılan adımlardır. T-SNE algoritmasında Gaussian dağılımı üzerinden belirli bir varyansa bağlı olarak ilgili noktanın diğer noktalar ile komşu olabilme olasılığı üzerinden dururken, UMAP algoritmasında varyans her bir nokta için farklılaşabilir.

$$p_{j|i} = e^{-\frac{(d(x_i - x_j) - \rho_i)}{\sigma_i}}$$

$$p_{ij} = (v_{j|i} - v_{i|j}) - v_{j|i}v_{i|j}$$

$$q_{ij} = (1 + a|y_i - y_j|^{2b})^{-1}$$

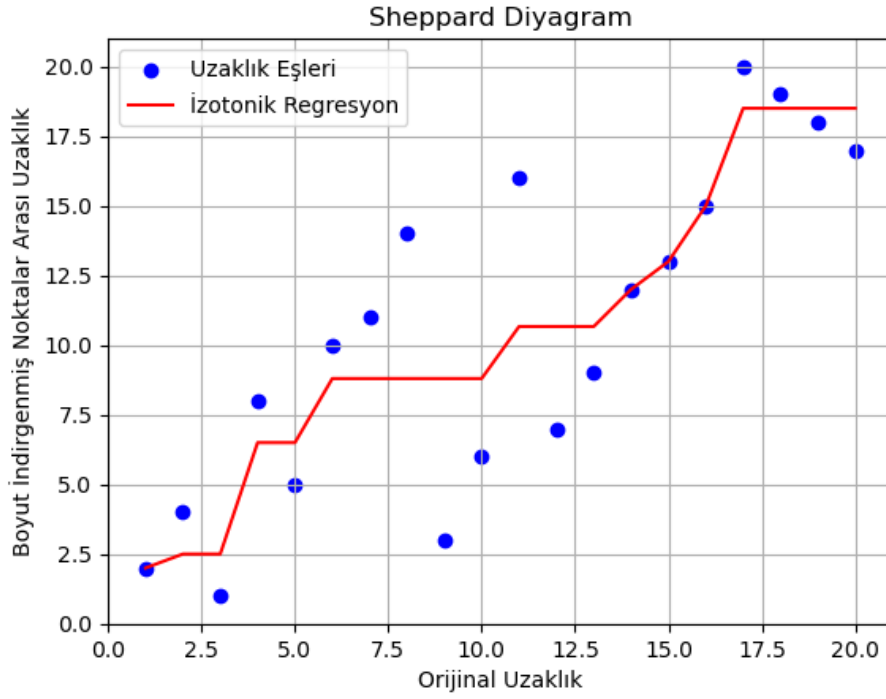
$$\phi = \sum_{i \neq j} p_{ij} \log q_{ij}$$

5. DEĞERLENDİRME ÖLÇÜM TEKNİKLERİ

Farklı boyutlarda kayıt ve değişken içeren 15 veri seti için belirlenen manifold algoritmaları kullanıldıktan sonra manifold algoritmalarının farklı parametreler için çıkardığı sonuçları karşılaştırmak için belli ölçüm teknikleri kullanılmıştır. Bu ölçüm teknikleri farklı bakış açıları ile orijinal veri seti ve manifold algoritmaları sonucunda değişken sayısı azaltılmış (boyut indirgemesi gerçekleştirilmiş) veri setlerini karşılaştırıp orijinal veri seti içerisindeki kayıtların ilişkileri ve buna benzer bilgilerinin ne kadar başarılı bir şekilde tutulduğunu ölçmektedir. Toplamda 9 değerlendirme ölçüm tekniği kullanılmıştır.

5.1. Sheppard Diyagramı (Sheppard Diagram, SD)

Sheppard diyagramı orijinal veri setindeki bütün kayıtların birbirleri ile olan uzaklıkları hesaplanarak bir eksene; boyutu indirgenmiş veri setindeki bütün kayıtların aynı şekilde birbirleri ile olan uzaklıkları hesaplanarak diğer bir eksene yerleştirilerek 2 boyutlu bir grafik üzerinden gösterimi ile elde edilir [39]. Bu grafik eğer bütün noktalar diyagonal olarak grafik üzerinde yerleşirse bütün noktaların uzaklıkları aynı şekilde boyutu indirgenmiş veri setinde de korunduğu anlamı çıkmaktadır.



Resim 2 Sheppard Diyagramı

5.2. Kruskal Stres Ölçümü (Kruskal Stress Measure, KrM)

Kruskal stres ölçümü, Sheppard diyagramını kullanarak boyutu indirgenmiş veri setindeki noktaların ne kadar orijinal veri setinden saptığını vermektedir. Bu sapmayı hesaplarken orijinal veri seti üzerinden bir regresyon kullanarak monotonik artışa uygun bir fonksiyon elde edilir [40] [41]. Bu fonksiyon kullanılarak orijinal veri setinden elde edilen uzaklıkların regresyon üzerinde nereye düştüğü bulunur. Bu bulunan değerler ile boyutu indirgenmiş veri setindeki uzaklıklar karşılaştırılarak ne kadar sapma olduğu üzerinden ölçüm yapılır. Bu metrik ne kadar 0 değerine yakın olursa sapmanın o kadar az olduğu ve yapının o kadar iyi korunduğu anlamı çıkmaktadır.

$$KrM = \sqrt{\frac{\sum_{ij} (d_{ij}^r - d_{ij}^*)^2}{\sum_{ij} d_{ij}^r}}$$

d_{ij}^r i ve j noktaları arasında boyut indirgenmiş veri setindeki uzaklığı gösterirken, d_{ij}^* monotonik regresyon üzerindeki değeri göstermektedir.

5.3. Sammon Stres (Sammon Stress, Ss)

Sammon stres ölçümü de aynı şekilde orijinal ve boyutu indirgenmiş veri seti üzerinden hesaplanan kayıtların birbirleri ile olan uzaklıklarının ne kadar saptığını belirlemede kullanılmaktadır [42]. Metrik pozitif tam sayı ve 0 değerini almaktadır. 0 değerine ne kadar yakınsa yapının o kadar iyi korunduğunu göstermektedir.

$$Ss = \frac{1}{\sum_{i < j} d_{ij}^o} \sum_{ij} \frac{(d_{ij}^o - d_{ij}^r)^2}{d_{ij}^o}$$

5.4. Spearman'ın Rho'su (Spearman's Rho, Sr)

Spearman'ın Rho'su orijinal veri seti içindeki kayıtların birbirleri ile olan uzaklıklarının; boyut indirgeme sonrası oluşturulmuş veri setindeki noktaların birbirleri ile olan uzaklıkları arasındaki sıralamaya bakarak bir değer çıkartmaktadır [43]. Kayıtlar arası uzaklığı hesapladığı için veri setindeki toplam kayıt sayısına “n” dersek hesaplanan toplam uzaklıkların sayısı, $T = n(n - 1)/2$ olacaktır.

$$Sr = 1 - \frac{6 \sum_{i=1}^T (o_i^o - o_i^r)^2}{(T^3 - T)}$$

Sr metriğinin alacağı maksimum değer 1, minimum değer ise -1 olacaktır. 1 değerine ne kadar yakınsa orijinal veri setinin o kadar iyi korunduğunu gösterecektir.

5.5. Topolojik Çarpımı (Topological Product, Tpr)

Topolojik çarpımı, orijinal veri seti ile boyutu indirgenmiş veri seti arasındaki ilişkiyi her bir kaydın orijinal veri seti içerisindeki komşulukları boyutu indirgenmiş veri seti ile karşılaştırması ile bulmaktadır [44]. Sperman's Rho'daki gibi burada da komşuluklar sıralanır ve diğer veri setine göre ne kadar saptığı bulunur. Burada sadece tek yönlü değil boyutu indirgenmiş veri setinin kayıtları üzerinden elde komşuluklar da orijinal veri seti üzerinden karşılaştırılmaktadır. Bu metriğin alacağı değer 0 çıkması durumunda veri setinin çok iyi korunmuş olduğu anlamı çıkmaktadır.

5.6. Artık Varyans (Residual Variance, Rv)

Artık varyans metriği, orijinal veri seti ile boyutu indirgenmiş veri setindeki kayıtların birbirlerine olan uzaklıklarının korelasyonu üzerinden elde edilen bir metriktir [30]. 0 ve 1 arasında değer almakta ve 0 değerine ne kadar yakınsa o kadar yapının iyi korunduğu anlamı çıkmaktadır.

$$Rv = 1 - \left(\frac{\sum_{i=0}^T (d_i^o - D_M^o)(d_i^r - D_M^r)}{\sqrt{\sum_{i=0}^T (d_i^o - D_M^o)^2} \sqrt{\sum_{i=0}^T (d_i^r - D_M^r)^2}} \right)^2$$

5.7. König Ölçüsü (König's Measure, KoM)

König ölçüsü, Tpr metriğindeki mantık ile çalışmaktadır. Buradaki tek fark bütün kayıtların diğer kayıtlar ile olan komşulukları değil belirlenen komşuluk sayıları üzerinden kayıtları orijinal veri seti ve boyutu indirgenmiş veri seti ile karşılaştırmaktadır [45]. KoM 0 ve 1 arasında değerler almaktadır. 1 değerine ne kadar yakınsa orijinal veri setinin o kadar iyi korunduğunu gösterecektir.

$$Q_{ij} = \begin{cases} 3 & \text{eğer } o_{ij}^o = o_{ij}^r \\ 2 & \text{eğer } o_{ij}^o \in \{o_{il}^r \mid l < n\} \\ 1 & \text{eğer } o_{ij}^o \in \{o_{it}^r \mid n < t < k\} \\ 0 & \text{hiçbiri değilse} \end{cases}$$

$$KoM = \frac{1}{(3nN)} \sum_{j=1}^n \sum_{i=1}^N Q_{ij}$$

5.8. Güvenirlilik ve Süreklilik (Trustworthiness & Continuity, Tw & Cnt)

Güvenirlilik ve süreklilik metrikleri aynı şekilde komşuluklar üzerinden gitmektedir [46]. Güvenirlilik metriği orijinal veri setinde bir kaydın (noktanın) uzağında olan diğer noktaların

boyutu indirgenmiş veri setinde ilgili kaydın komşuluğuna girip girmediğinin ölçütünü vermektedir. Süreklilik ise benzer mantık ile orijinal veri setinde bir kaydın (noktanın) komşuluğu içinde olan diğer noktaların boyutu indirgenmiş veri setinde ne kadar ilgili kayıttan uzaklaştığının ölçütünü vermektedir. Bu iki metrik de 0 ve 1 arasında değerler almaktadır. 1 değerine ne kadar yakınsa orijinal veri setinin o kadar iyi korunduğunu gösterecektir.

5.9. Yerel Süreklilik Meta Kriteri (Local Continuity Meta-Criterion, LCMC)

Yerel süreklilik meta kriteri ise orijinal veri setindeki kayıtların komşulukları ile boyutu indirgenmiş veri setindeki komşulukları her bir kayıt için karşılaştırmakta ve komşulukların ne kadar örtüştüğünü göstermektedir [47].

5.10. Yeniden Düzenlenmiş Ölçeklendirme (Refactored Scaling, RES)

Yukarıda anlatılan 9 değerlendirme ölçüm teknikleri farklı algoritma ve farklı parametreler ile değerlendirileceği için hangi algoritmanın hangi parametreler için ve hangi algoritmanın hangi parametre için iyi sonuç verdiğini görmek “yeniden düzenlenmiş ölçeklendirme (Refactored Scaling, RES)” metriği üretilmiştir. Bütün değerlendirme ölçüm teknikleri 0 ve 1 arasında eğer gerekirse normalize edilmiş ve 1 en iyi sonucu verecek şekilde düzenlenmiştir.

$$RES = \frac{NoPn + ((1 - KrM) + \left(1 - \left(\frac{\max(Ss, 100)}{100}\right)\right) + \frac{(Sr + 1)}{2} + (1 - Rv) + KoM + Tw + Cnt)}{9 + NoPd}$$

RES metriğinin pay kısmında kalan *NoPn* kullanılan algoritma için değişecek olup parametre olup analiz kısmında açıklanacaktır. *NoPd* ise kullanılan parametre sayısına karşılık gelmektedir. Buradaki *NoPn* RES metriği için cezalandırıcı bir değişken olarak eklenmiştir. Buradaki amaç ise de boyut indirgemedi ne kadar çok boyut indirgemesi gerçekleştirilirse veya algoritma çalışma süresini kısaltacak bir parametre seçimi olursa cezalandırıcı etkisi o kadar az olacak; tersi olur ise cezalandırıcı etkisi fazla olacak ve sifıra daha yakın olacaktır.

6. ANALİZ

Boyut indirgeme (Manifold öğrenme) algoritmalarının analizini yapmak için 14 veri seti kullanılmıştır. Bu 14 veri seti sırasıyla çekirdek temel birleşen analizi (Kernel PCA), izomap (Isomap), lokal doğrusal gömme (LLE), t-dağıtılmış stokastik komşu gömme (t-SNE) ve standart manifold yaklaşımı ve projeksiyonu (UMAP) algoritmaları üzerinden farklı parametreler için çalıştırılmış ve çıkan sonuçlar 9 farklı değerlendirme ölçüm teknikleri ile karşılaştırılmıştır.

Topolojik Çarpımı (Tpr) için bütün sonuçlar 0 ortalama ve çok düşük varyansa sahip olduğu için yeniden ölçeklendirme (RES) metriğinin hesaplamasına katılmamış ve aşağıdaki sonuçlar için en iyi Tpr değeri veren parametre değerleri paylaşılmamıştır.

6.1. Çekirdek Temel Birleşen Analizi (Kernel PCA) Algoritmasının Sonuçları

Çekirdek Temel Birleşen Analizi indirgenmesi istenen boyut sayısını parametre olarak almaktadır. Boyut sayısı (n_components, NoC) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 40, 60, 80, 100, 250, 500, 750 ve 1000 olarak veri setinin sahip olduğu değişken sayısına göre parametre değeri olarak algoritmaya verilmiştir.

Tablo 2 Kernel PCA için kullanılan parametre ve değerleri

Kısaltma	Parametre Adı	Aldığı Değerler
NoC	İndirgenecek Boyut Sayısı	2,3,4,5,6,7,8,9,10,20,40,60,80,100,250,500,750,1000

Tablo 3 “Lenses” Veri Setinin Kernel PCA için sonuçları

Veri Seti	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
lenses_adjusted	4	0.13	0.24	0.09	-0.06	0.26	0.73	1.00	1.00	0.82	0.73
lenses_adjusted	5	0.00	0.00	-0.04	0.00	0.00	0.75	1.00	1.00	0.82	0.75
lenses_adjusted	7	0.00	0.00	0.07	0.00	0.00	0.76	1.00	1.00	0.82	0.73
lenses_adjusted	8	0.00	0.00	0.14	0.00	0.00	0.74	1.00	1.00	0.82	0.72

Tek dönüşüm adımı uygulandıktan sonra “Lenses” veri seti 24 kayıt ve 10 değişkenli bir yapıya dönmüştür. 3 metrik; boyut sayısı 4 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre boyut sayısı 5 olan en iyi sonucu vermektedir.

Tablo 4 “Soybean (Small)” Veri Setinin Kernel PCA için sonuçları

Veri Seti	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
soybean-small	2	0.19	3.52	0.42	-0.03	0.19	0.72	0.96	0.97	0.51	0.74
soybean-small	20	0.01	0.00	0.32	0.00	0.00	0.82	1.00	1.00	0.83	0.80
soybean-small	40	0.00	0.00	0.30	0.00	0.00	0.86	1.00	1.00	0.84	0.78
soybean-small	60	0.00	0.00	0.34	0.00	0.00	0.87	1.00	1.00	0.81	0.75

Tek dönüşüm adımı uygulandıktan sonra “Soybean (Small)” veri seti 307 kayıt ve 72 boyuttan oluşmaktadır. 4 metrik; boyut sayısı 40 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre boyut sayısı 20 olan en iyi sonucu vermektedir.

Tablo 5 “Molecular Biology (Promoter Gene Sequences)” Veri Setinin Kernel PCA için sonuçları

Veri Seti	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
promoters_adjusted	80	0.01	0.00	0.02	0.00	0.01	0.88	1.00	1.00	0.83	0.79
promoters_adjusted	100	0.00	0.00	0.03	0.00	0.00	0.87	1.00	1.00	0.85	0.78

Tek dönüşüm adımı uygulandıktan sonra “Molecular Biology (Promoter Gene Sequences)” veri seti 106 kayıt ve 234 boyuttan oluşmaktadır. 1 metrik hariç bütün metrikler boyut sayısı 100 için en iyi değeri verirken; bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre boyut sayısı 80 olan en iyi sonucu vermektedir.

Tablo 6 “Hayes-Roth” Veri Setinin Kernel PCA için sonuçları

Veri Seti	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
hayes-roth	7	0.10	N/A	0.01	N/A	0.17	0.72	0.96	0.97	0.42	0.58
hayes-roth	9	0.06	N/A	0.00	N/A	0.06	0.74	0.97	0.97	0.45	0.59
hayes-roth	10	0.04	N/A	0.01	N/A	0.03	0.73	0.97	0.97	0.45	0.59

Tek dönüşüm adımı uygulandıktan sonra “Hayes-Roth” veri seti 160 kayıt ve 15 boyuttan oluşmaktadır. 5 metrik; boyut sayısı 10 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre boyut sayısı 9 olan en iyi sonucu vermektedir.

Tablo 7 “Lymphography” Veri Setinin Kernel PCA için sonuçları

Veri Seti	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
lymphography	3	0.24	1.23	0.02	-0.05	0.45	0.69	0.85	0.92	0.25	0.65
lymphography	20	0.03	0.01	0.00	0.00	0.04	0.79	0.99	0.99	0.72	0.76
lymphography	40	0.00	0.00	0.00	0.00	0.00	0.79	1.00	1.00	0.76	0.74

Tek dönüşüm adımı uygulandıktan sonra “Lymphography” veri seti 148 kayıt ve 59 boyuttan oluşmaktadır. 7 metrik; boyut sayısı 40 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre boyut sayısı 20 olan en iyi sonucu vermektedir.

Tablo 8 “Audiology (Original)” Veri Setinin Kernel PCA için sonuçları

Veri Seti	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
audiology.standardized	40	0.03	N/A	0.03	N/A	0.02	0.79	1.00	1.00	0.77	0.69
audiology.standardized	60	0.01	N/A	0.01	N/A	0.01	0.80	1.00	1.00	0.77	0.69
audiology.standardized	100	0.01	0.00	0.01	0.00	0.00	0.80	1.00	1.00	0.76	0.78
audiology.standardized	250	0.00	0.00	0.01	0.00	0.00	0.80	1.00	1.00	0.77	0.74

Tek dönüşüm adımı uygulandıktan sonra “Audiology (Original)” veri seti 200 kayıt ve 354 boyuttan oluşmaktadır. 5 metrik; boyut sayısı 250 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre boyut sayısı 100 olan en iyi sonucu vermektedir.

Tablo 9 “Breast Cancer” Veri Setinin Kernel PCA için sonuçları

Veri Seti	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
breast-cancer	9	0.10	0.07	0.02	N/A	0.20	0.72	0.98	0.98	0.47	0.72
breast-cancer	20	0.03	N/A	0.04	N/A	0.04	0.76	0.99	0.99	0.63	0.64
breast-cancer	40	0.00	N/A	0.03	N/A	0.00	0.76	0.99	0.99	0.64	0.60

Tek dönüşüm adımı uygulandıktan sonra “Breast Cancer” veri seti 286 kayıt ve 43 boyuttan oluşmaktadır. 5 metrik; boyut sayısı 40 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre boyut sayısı 9 olan en iyi sonucu vermektedir.

Tablo 10 “Soybean (Large)” Veri Setinin Kernel PCA için sonuçları

Veri Seti	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
soybean-large	40	0.01	0.00	0.18	N/A	0.00	0.85	1.00	1.00	0.84	0.80
soybean-large	60	0.00	0.00	0.20	N/A	0.00	0.85	1.00	1.00	0.85	0.78
soybean-large	80	0.00	0.00	0.18	N/A	0.00	0.84	1.00	1.00	0.85	0.77
soybean-large	100	0.00	0.00	0.18	N/A	0.00	0.84	1.00	1.00	0.85	0.75

Tek dönüşüm adımı uygulandıktan sonra “Soybean (Large)” veri seti 307 kayıt ve 132 boyuttan oluşmaktadır. 4 metrik; boyut sayısı 60 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre boyut sayısı 40 olan en iyi sonucu vermektedir.

Tablo 11 “Primary Tumor” Veri Setinin Kernel PCA için sonuçları

Veri Seti	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
primary-tumor	9	0.10	0.06	0.06	N/A	0.21	0.74	0.97	0.99	0.53	0.72
primary-tumor	40	0.00	N/A	0.09	N/A	0.00	0.77	1.00	1.00	0.67	0.61

Tek dönüşüm adımı uygulandıktan sonra “Primary Tumor” veri seti 339 kayıt ve 43 boyuttan oluşmaktadır. 7 metrik; boyut sayısı 40 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre boyut sayısı 9 olan en iyi sonucu vermektedir.

Tablo 12 “Congressional Voting Records” Veri Setinin Kernel PCA için sonuçları

Veri Seti	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
house-votes-84	8	0.10	0.07	0.01	N/A	0.12	0.72	0.97	0.99	0.45	0.72
house-votes-84	20	0.03	0.00	0.00	N/A	0.02	0.75	0.99	1.00	0.62	0.74
house-votes-84	40	0.00	0.00	0.01	N/A	0.00	0.75	1.00	1.00	0.64	0.71

Tek dönüşüm adımı uygulandıktan sonra “Congressional Voting Records” veri seti 435 kayıt ve 48 boyuttan oluşmaktadır. 7 metrik; boyut sayısı 40 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre boyut sayısı 20 olan en iyi sonucu vermektedir.

Tablo 13 “Balance Scale” Veri Setinin Kernel PCA için sonuçları

Veri Seti	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
balance-scale	8	0.17	0.27	-0.01	-0.02	0.63	0.68	0.93	0.96	0.21	0.61
balance-scale	9	0.16	0.18	0.02	-0.02	0.59	0.68	0.93	0.98	0.19	0.61
balance-scale	10	0.14	0.11	0.01	-0.01	0.51	0.68	0.96	0.98	0.23	0.62

Tek dönüşüm adımı uygulandıktan sonra “Balance Scale” veri seti 625 kayıt ve 20 boyuttan oluşmaktadır. 6 metrik; boyut sayısı 10 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman da RES metriğine göre gene boyut sayısı 10 için en iyi sonucu vermektedir.

Tablo 14 “Tic-Tac-Toe Endgame” Veri Setinin Kernel PCA için sonuçları

Veri Seti	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
tic-tac-toe	5	0.18	0.44	0.01	-0.03	0.50	0.68	0.95	0.95	0.16	0.64
tic-tac-toe	20	0.00	0.00	0.01	0.00	0.00	0.68	0.99	0.99	0.24	0.67

Tek dönüşüm adımı uygulandıktan sonra “Tic-Tac-Toe Endgame” veri seti 958 kayıt ve 27 boyuttan oluşmaktadır. 7 metrik; boyut sayısı 20 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman da RES metriğine göre gene boyut sayısı 20 için en iyi sonucu vermektedir.

Tablo 15 “Car Evaluation” Veri Setinin Kernel PCA için sonuçları

Veri Seti	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
car	10	0.11	0.06	0.04	-0.01	0.34	0.68	0.97	0.99	0.16	0.64
car	20	0.00	0.00	0.00	0.00	0.00	0.69	1.00	1.00	0.32	0.66

Tek dönüşüm adımı uygulandıktan sonra “Car Evaluation” veri seti 1728 kayıt ve 21 boyuttan oluşmaktadır. 7 metrik; boyut sayısı 20 için en iyi değerleri verirken bütün

metrikleri ve parametreleri hesaba kattığımız zaman da RES metriğine göre gene boyut sayısı 20 için en iyi sonucu vermektedir.

6.2. İzomap Algoritmasının Sonuçları

İzomap algoritması komşuluk sayısını ve boyut indirgenmesi istenen boyut sayısını parametre olarak almaktadır. Komşuluk sayısı (n_neighbors, NoN) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 40, 60, 80, 100, 250, 500, 750 ve 1000 ve boyut sayısı (n_components, NoC) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 40, 60, 80, 100, 250, 500, 750 ve 1000 olarak veri setinin sahip olduğu değişken sayısına göre parametre değeri olarak algoritmaya verilmiştir.

Tablo 16 Isomap için kullanılan parametre ve değerleri

Kısaltma	Parametre Adı	Aldığı Değerler
NoC	İndirgenecek Boyut Sayısı	2,3,4,5,6,7,8,9,10,20,40,60,80,100,250,500,750,1000
NoN	Komşuluk Sayısı	1,2,3,4,5,6,7,8,9,10,20,40,60,80,100,250,500,750,1000

Tablo 17 “Lenses” Veri Setinin Isomap için sonuçları

Veri Seti	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
lenses_adjusted	5	4	0.14	0.38	0.05	-0.07	0.26	0.73	1.00	1.00	0.82	0.80
lenses_adjusted	5	5	0.01	0.26	0.04	0.00	0.00	0.72	1.00	1.00	0.82	0.83
lenses_adjusted	5	6	0.01	0.26	0.17	0.00	0.00	0.72	1.00	1.00	0.82	0.82
lenses_adjusted	7	6	0.07	0.15	-0.10	-0.01	0.13	0.76	0.92	0.92	0.50	0.74
lenses_adjusted	20	4	0.12	0.03	0.05	-0.03	0.21	0.74	0.89	0.88	0.09	0.65

“Lenses” veri seti için 3 metrik; boyut sayısı 4 ve komşuluk sayısı 5 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre boyut sayısı 4 ve komşuluk sayısı 5 için en iyi sonucu vermektedir.

Tablo 18 “Soybean (Small)” Veri Setinin Isomap için sonuçları

Veri Seti	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
soybean-small	2	20	0.28	0.22	0.02	-0.02	0.99	0.78	0.99	0.94	0.75	0.74
soybean-small	4	20	0.16	0.06	-0.03	0.01	0.99	0.79	0.87	0.91	0.63	0.72

soybean-small	10	3	0.19	0.26	0.33	-0.04	0.23	0.73	0.96	0.97	0.59	0.82
soybean-small	10	5	0.14	0.22	0.27	-0.02	0.19	0.75	0.98	0.98	0.67	0.84
soybean-small	10	20	0.11	0.26	0.29	0.00	0.20	0.74	0.99	0.99	0.70	0.83
soybean-small	20	8	0.09	0.10	0.26	0.00	0.16	0.75	0.96	0.97	0.46	0.80
soybean-small	20	20	0.07	0.15	0.21	0.00	0.16	0.78	0.97	0.98	0.48	0.79
soybean-small	40	10	0.14	0.04	0.16	0.00	0.36	0.74	0.87	0.87	0.10	0.67

“Soybean (Small)” veri seti için 2 metrik; boyut sayısı 20 ve komşuluk sayısı 2 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre boyut sayısı 5 ve komşuluk sayısı 10 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

Tablo 19 “Molecular Biology (Promoter Gene Sequences)” Veri Setinin Isomap için sonuçları

Veri Seti	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
promoters_adjusted	3	40	0.12	0.60	0.05	-0.02	0.77	0.74	0.98	0.97	0.75	0.79
promoters_adjusted	6	20	0.11	0.45	0.02	-0.02	0.66	0.74	0.98	0.97	0.71	0.80
promoters_adjusted	20	40	0.05	0.33	0.01	0.00	0.42	0.71	0.98	0.98	0.63	0.80
promoters_adjusted	40	40	0.05	0.34	0.03	0.00	0.35	0.70	0.96	0.97	0.49	0.77
promoters_adjusted	60	20	0.09	0.24	0.07	-0.01	0.49	0.70	0.88	0.90	0.28	0.71
promoters_adjusted	80	7	0.20	0.07	0.04	-0.05	0.74	0.68	0.65	0.66	0.10	0.59

“Molecular Biology (Promoter Gene Sequences)” veri seti için 3 metrik; boyut sayısı 40 ve komşuluk sayısı 3 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre boyut sayısı 20 ve komşuluk sayısı 6 için en iyi sonucu vermektedir.

Tablo 20 “Hayes-Roth” Veri Setinin Isomap için sonuçları

Veri Seti	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
hayes-roth	1	10	0.70	N/A	0.03	N/A	0.97	0.70	0.76	0.88	0.23	0.47
hayes-roth	10	10	0.15	0.29	0.00	N/A	0.37	0.75	0.95	0.91	0.45	0.73
hayes-roth	20	10	0.08	0.21	0.00	N/A	0.12	0.73	0.98	0.98	0.68	0.79
hayes-roth	40	10	0.07	0.12	0.00	N/A	0.09	0.73	0.96	0.96	0.48	0.75
hayes-roth	100	6	0.12	0.04	-0.01	N/A	0.23	0.71	0.88	0.89	0.17	0.66

“Hayes-Roth” veri seti için 3 metrik; boyut sayısı 10 ve komşuluk sayısı 20 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman da RES metriğine göre gene boyut sayısı 10 ve komşuluk sayısı 20 için en iyi sonucu vermektedir.

Tablo 21 “Audiology (Original)” Veri Setinin Isomap için sonuçları

Veri Seti	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
audiology.standardized	4	4	0.30	N/A	0.03	N/A	0.52	0.72	0.90	0.95	0.40	0.66
audiology.standardized	9	100	0.11	0.39	0.01	0.00	0.29	0.74	0.99	0.98	0.68	0.82
audiology.standardized	10	100	0.11	N/A	0.01	N/A	0.28	0.74	0.99	0.99	0.70	0.72
audiology.standardized	40	40	0.07	N/A	0.01	0.00	0.17	0.75	0.98	0.98	0.65	0.73
audiology.standardized	60	40	0.08	N/A	0.02	N/A	0.21	0.76	0.97	0.98	0.55	0.71
audiology.standardized	60	100	0.06	N/A	0.01	N/A	0.21	0.75	0.97	0.98	0.56	0.69

“Audiology (Original)” veri seti için 2 metrik; boyut sayısı 100 ve komşuluk sayısı 10 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre boyut sayısı 100 ve komşuluk sayısı 9 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

Tablo 22 “Breast Cancer” Veri Setinin Isomap için sonuçları

Veri Seti	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
breast-cancer	1	8	0.78	N/A	0.08	N/A	1.00	0.68	0.73	0.77	0.11	0.49
breast-cancer	3	20	0.15	N/A	0.03	N/A	0.40	0.74	0.99	0.97	0.58	0.68
breast-cancer	20	9	0.10	0.17	0.02	N/A	0.18	0.72	0.97	0.98	0.58	0.82
breast-cancer	20	20	0.06	N/A	0.03	N/A	0.10	0.76	0.99	0.99	0.70	0.72
breast-cancer	60	20	0.04	N/A	0.04	N/A	0.06	0.76	0.98	0.99	0.63	0.71
breast-cancer	250	10	0.11	0.02	0.04	N/A	0.19	0.72	0.85	0.86	0.09	0.66

“Breast Cancer” veri seti için 3 metrik; boyut sayısı 20 ve komşuluk sayısı 60 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre boyut sayısı 9 ve komşuluk sayısı 20 için en iyi sonucu vermektedir.

Tablo 23 “Soybean (Large)” Veri Setinin Isomap için sonuçları

Veri Seti	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
soybean-large	5	20	0.28	N/A	0.11	N/A	0.85	0.74	0.99	0.99	0.64	0.66

soybean-large	6	9	0.25	0.56	0.17	N/A	0.30	0.74	0.97	0.98	0.58	0.82
soybean-large	6	20	0.23	N/A	0.16	N/A	0.31	0.74	0.99	0.99	0.66	0.73
soybean-large	20	20	0.17	0.35	0.14	N/A	0.21	0.75	0.99	0.99	0.70	0.84
soybean-large	40	20	0.09	0.22	0.16	N/A	0.10	0.77	0.98	0.99	0.69	0.85
soybean-large	80	20	0.08	0.14	0.14	N/A	0.10	0.78	0.98	0.98	0.62	0.83
soybean-large	100	20	0.08	0.13	0.13	N/A	0.12	0.79	0.98	0.98	0.56	0.82
soybean-large	200	9	0.11	0.04	0.12	N/A	0.13	0.74	0.94	0.94	0.27	0.75

“Soybean (Large)” veri seti için 2 metrik; boyut sayısı 20 ve komşuluk sayısı 80 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre boyut sayısı 20 ve komşuluk sayısı 40 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

Tablo 24 “Primary Tumor” Veri Setinin Isomap için sonuçları

Veri Seti	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
primary-tumor	3	10	0.19	N/A	0.04	N/A	0.57	0.74	0.98	0.98	0.52	0.67
primary-tumor	9	10	0.11	N/A	0.06	N/A	0.23	0.73	0.98	0.98	0.58	0.72
primary-tumor	20	9	0.11	0.17	0.06	N/A	0.21	0.73	0.97	0.98	0.61	0.82
primary-tumor	20	10	0.10	N/A	0.05	N/A	0.18	0.74	0.98	0.98	0.64	0.73
primary-tumor	80	10	0.09	N/A	0.06	N/A	0.16	0.75	0.96	0.97	0.55	0.70
primary-tumor	100	10	0.09	0.09	0.06	N/A	0.16	0.74	0.96	0.97	0.52	0.79
primary-tumor	200	6	0.14	0.04	0.08	N/A	0.23	0.72	0.91	0.91	0.29	0.73
primary-tumor	200	10	0.09	0.07	0.07	N/A	0.15	0.73	0.94	0.94	0.31	0.74

“Primary Tumor” veri seti için 2 metrik; boyut sayısı 6 ve komşuluk sayısı 200 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre boyut sayısı 9 ve komşuluk sayısı 20 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

Tablo 25 “Congressional Voting Records” Veri Setinin Isomap için sonuçları

Veri Seti	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
house-votes-84	40	2	0.22	N/A	0.02	N/A	0.26	0.69	0.88	0.92	0.32	0.67
house-votes-84	40	9	0.09	0.16	0.02	N/A	0.11	0.72	0.97	0.98	0.58	0.83
house-votes-84	40	40	0.07	0.25	0.02	N/A	0.08	0.74	0.99	0.99	0.69	0.78

house-votes-84	80	20	0.06	0.18	0.01	N/A	0.07	0.73	0.99	0.99	0.66	0.81
house-votes-84	100	20	0.06	0.17	0.01	N/A	0.06	0.74	0.99	0.99	0.64	0.81
house-votes-84	100	40	0.06	0.21	0.02	N/A	0.09	0.73	0.99	0.99	0.64	0.76
house-votes-84	250	6	0.10	0.05	0.01	N/A	0.11	0.71	0.96	0.97	0.37	0.76
house-votes-84	250	20	0.08	0.10	0.01	N/A	0.08	0.74	0.98	0.97	0.37	0.74

“Congressional Voting Records” veri seti için hiçbir metrik için en iyi değeri vermezken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre boyut sayısı 9 ve komşuluk sayısı 40 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

Tablo 26 “Balance Scale” Veri Setinin Isomap için sonuçları

Veri Seti	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
balance-scale	2	10	0.24	N/A	0.00	N/A	0.77	0.68	0.97	0.86	0.10	0.56
balance-scale	4	4	0.29	0.15	0.05	-0.07	0.86	0.67	0.82	0.83	0.07	0.66
balance-scale	8	10	0.14	0.22	0.03	-0.01	0.52	0.69	0.94	0.98	0.37	0.73
balance-scale	10	10	0.14	0.22	0.03	-0.01	0.51	0.68	0.95	0.99	0.42	0.74
balance-scale	20	10	0.13	0.15	0.02	-0.01	0.50	0.68	0.93	0.97	0.51	0.74
balance-scale	60	8	0.17	0.04	0.02	-0.02	0.62	0.68	0.88	0.89	0.37	0.70
balance-scale	250	10	0.13	0.07	0.02	-0.02	0.48	0.69	0.87	0.87	0.30	0.67

“Balance Scale” veri seti için 2 metrik; boyut sayısı 10 ve komşuluk sayısı 20 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman da RES metriğine göre gene boyut sayısı 10 ve komşuluk sayısı 20 için en iyi sonucu vermektedir.

Tablo 27 “Tic-Tac-Toe Endgame” Veri Setinin Isomap için sonuçları

Veri Seti	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
tic-tac-toe	4	10	0.14	0.36	0.00	-0.01	0.51	0.68	0.97	0.96	0.16	0.72
tic-tac-toe	8	10	0.11	0.29	0.00	-0.01	0.35	0.68	0.98	0.99	0.33	0.76
tic-tac-toe	10	3	0.31	0.16	0.01	-0.07	0.74	0.67	0.78	0.91	0.12	0.68
tic-tac-toe	20	10	0.11	0.25	0.00	-0.01	0.32	0.68	0.97	0.99	0.59	0.79
tic-tac-toe	40	10	0.09	0.18	0.00	-0.01	0.25	0.68	0.97	0.98	0.57	0.79
tic-tac-toe	60	10	0.09	0.16	0.01	-0.01	0.27	0.68	0.96	0.98	0.59	0.79
tic-tac-toe	750	7	0.14	0.03	0.00	-0.01	0.40	0.68	0.74	0.73	0.12	0.62

“Tic-Tac-Toe Endgame” veri seti için 2 metrik; boyut sayısı 10 ve komşuluk sayısı 40 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman da RES metriğine göre gene boyut sayısı 10 ve komşuluk sayısı 40 için en iyi sonucu vermektedir.

Tablo 28 “Car Evaluation” Veri Setinin Isomap için sonuçları

Veri Seti	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
car	20	20	0.02	0.30	0.01	0.00	0.02	0.69	0.99	0.99	0.76	0.79
car	100	20	0.00	0.27	0.02	0.00	0.00	0.69	1.00	1.00	0.87	0.81
car	1000	6	0.18	0.42	0.05	-0.07	0.50	0.68	0.81	0.79	0.21	0.65
car	1000	8	0.14	0.06	0.03	-0.02	0.44	0.68	0.85	0.85	0.24	0.66

“Car Evaluation” veri seti için 5 metrik; boyut sayısı 20 ve komşuluk sayısı 100 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman da RES metriğine göre gene boyut sayısı 20 ve komşuluk sayısı 100 için en iyi sonucu vermektedir.

6.3.Lokal Doğrusal Gömme (LLE) Algoritmasının Sonuçları

Lokal doğrusal gömme (LLE) algoritması komşuluk sayısını ve boyut indirgenmesi istenen boyut sayısını parametre olarak almaktadır. Komşuluk sayısı (n_neighbors, NoN) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 40, 60, 80, 100, 250, 500, 750 ve 1000 ve boyut sayısı (n_components, NoC) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 40, 60, 80, 100, 250, 500, 750 ve 1000 olarak veri setinin sahip olduğu değişken sayısına göre parametre değeri olarak algoritmaya verilmiştir. Ayrıca LLE algoritmasının kendisi (standard) ve farklı varyansları hessian, ltsa ve modified için de algoritma verilen parametreler için çalıştırılmıştır.

Tablo 29 LLE için kullanılan parametre ve değerleri

Kısaltma	Parametre Adı	Aldığı Değerler
NoC	İndirgenecek Boyut Sayısı	2,3,4,5,6,7,8,9,10,20,40,60,80,100,250,500,750,1000
NoN	Komşuluk Sayısı	1,2,3,4,5,6,7,8,9,10,20,40,60,80,100,250,500,750,1000
Mthd	Farklı LLE varyansları	hessian, ltsa, modified, standard

Tablo 30 “Lenses” Veri Setinin LLE için sonuçları

Veri Seti	Mthd	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
-----------	------	-----	-----	--------	----	----	-----	----	-----	----	-----	------	-----

lenses_adjusted	ltsa	1	3	0.87	N/A	0.19	N/A	1.00	0.70	0.63	0.56	0.17	0.44
lenses_adjusted	standard	5	5	0.05	4.97	0.01	-0.01	0.05	0.73	1.00	1.00	0.82	0.81
lenses_adjusted	standard	5	8	0.08	2.46	0.01	-0.01	0.45	0.73	0.87	0.99	0.64	0.71
lenses_adjusted	standard	6	5	0.05	4.98	0.01	-0.01	0.06	0.76	0.95	0.95	0.61	0.78

“Lenses” veri seti için 5 metrik; metodu “standard”, boyut sayısı 5 ve komşuluk sayısı 5 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman da RES metriğine göre gene metodu “standard”, boyut sayısı 250 ve komşuluk sayısı 5 için en iyi sonucu vermektedir.

Tablo 31 “Soybean (Small)” Veri Setinin LLE için sonuçları

Veri Seti	Mthd	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
soybean-small	hessian	40	2	0.20	N/A	0.22	-0.03	0.22	0.72	0.50	0.50	0.04	0.51
soybean-small	ltsa	20	8	0.13	48.91	0.07	0.00	0.46	0.78	0.79	0.82	0.26	0.66
soybean-small	modified	9	2	0.32	N/A	0.43	N/A	0.35	0.68	0.97	0.98	0.66	0.71
soybean-small	modified	9	3	0.46	N/A	0.45	N/A	0.61	0.69	0.81	0.80	0.36	0.61
soybean-small	standard	1	10	0.30	N/A	-0.25	N/A	0.77	0.71	0.99	0.97	0.59	0.64
soybean-small	standard	20	40	0.05	6.46	0.01	0.06	0.95	0.70	0.58	0.52	0.04	0.53
soybean-small	standard	40	40	0.04	6.47	-0.03	0.07	0.98	0.69	0.20	0.22	-0.01	0.41

“Soybean (Small)” veri seti için 2 metrik; metodu “modified”, boyut sayısı 2 ve komşuluk sayısı 9 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman da RES metriğine göre gene metodu “modified”, boyut sayısı 2 ve komşuluk sayısı 9 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

Tablo 32 “Molecular Biology (Promoter Gene Sequences)” Veri Setinin LLE için sonuçları

Veri Seti	Mthd	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
promoters_adjusted	ltsa	100	40	0.05	91.41	0.03	0.00	0.43	0.73	0.50	0.51	0.02	0.48
promoters_adjusted	modified	60	60	0.44	N/A	0.08	N/A	0.91	0.66	0.51	0.49	0.03	0.41
promoters_adjusted	standard	1	4	0.51	N/A	0.00	N/A	0.97	0.69	0.94	0.98	0.52	0.61
promoters_adjusted	standard	1	60	0.17	N/A	0.03	N/A	0.96	0.74	1.00	0.94	0.79	0.66
promoters_adjusted	standard	3	60	0.07	58.47	0.02	0.00	0.89	0.77	0.96	0.95	0.77	0.71
promoters_adjusted	standard	80	100	0.02	31.97	0.05	0.01	0.92	0.67	0.29	0.31	0.01	0.44

promoters_adjusted	standard	100	100	0.01	31.97	0.02	0.01	0.97	0.67	0.21	0.21	0.00	0.39
--------------------	----------	-----	-----	-------------	-------	------	------	------	------	------	------	------	------

“Molecular Biology (Promoter Gene Sequences” veri seti için 2 metrik; metodu “standard”, boyut sayısı 60 ve komşuluk sayısı 1 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre metodu “standard”, boyut sayısı 60 ve komşuluk sayısı 3 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

Tablo 33 “Hayes-Roth” Veri Setinin LLE için sonuçları

Veri Seti	Mthd	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
hayes-roth	hessian	100	8	0.10	32.97	0.03	N/A	0.22	0.72	0.74	0.72	0.13	0.59
hayes-roth	ltsa	20	5	0.18	61.88	0.00	N/A	0.36	0.72	0.94	0.95	0.54	0.70
hayes-roth	ltsa	80	9	0.11	28.86	-0.01	N/A	0.26	0.74	0.86	0.78	0.22	0.62
hayes-roth	modified	10	7	0.84	N/A	0.08	N/A	1.00	0.66	0.56	0.57	0.05	0.40
hayes-roth	modified	20	7	0.15	40.07	0.02	N/A	0.39	0.71	0.93	0.93	0.54	0.71
hayes-roth	standard	1	5	0.37	N/A	0.01	N/A	0.98	0.71	0.98	0.93	0.32	0.58
hayes-roth	standard	2	5	0.44	N/A	-0.01	N/A	0.99	0.73	0.93	0.96	0.42	0.58
hayes-roth	standard	40	10	0.11	25.37	0.00	N/A	0.28	0.73	0.92	0.91	0.41	0.68

“Hayes-Roth” veri seti için hiçbir metrik için en iyi değeri vermezken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre metodu “modified”, boyut sayısı 7 ve komşuluk sayısı 20 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

Tablo 34 “Lymphography” Veri Setinin LLE için sonuçları

Veri Seti	Mthd	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
lymphography	ltsa	1	4	0.97	N/A	0.05	N/A	1.00	0.67	0.65	0.51	0.01	0.43
lymphography	ltsa	40	20	0.17	51.39	-0.01	0.00	0.53	0.75	0.90	0.90	0.43	0.67
lymphography	modified	40	20	0.11	48.79	0.00	0.00	0.44	0.74	0.89	0.89	0.41	0.68
lymphography	modified	100	9	0.13	N/A	0.01	-0.01	0.36	0.73	0.78	0.79	0.17	0.57
lymphography	standard	1	9	0.40	N/A	0.00	N/A	0.98	0.69	0.99	0.90	0.57	0.61
lymphography	standard	1	20	0.33	N/A	0.01	N/A	0.98	0.69	0.99	0.91	0.58	0.60
lymphography	standard	3	10	0.25	N/A	0.01	-0.03	0.91	0.71	0.91	0.92	0.44	0.61
lymphography	standard	60	40	0.08	21.25	-0.01	0.01	0.62	0.74	0.76	0.74	0.21	0.60

“Lymphography” veri seti için 2 metrik; metodu “standard”, boyut sayısı 40 ve komşuluk sayısı 60 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre metodu “modified”, boyut sayısı 20 ve komşuluk sayısı 40 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

Tablo 35 “Audiology (Original)” Veri Setinin LLE için sonuçları

Veri Seti	Mthd	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
audiology.standardized	ltsa	80	20	0.23	94.95	0.01	-0.01	0.57	0.77	0.88	0.89	0.35	0.62
audiology.standardized	modified	40	40	0.80	N/A	0.09	N/A	1.00	0.65	0.38	0.55	-0.06	0.39
audiology.standardized	modified	100	9	0.16	N/A	0.00	-0.01	0.41	0.74	0.86	0.86	0.29	0.61
audiology.standardized	standard	1	3	0.57	N/A	0.01	N/A	0.97	0.71	0.96	0.95	0.60	0.62
audiology.standardized	standard	1	100	0.18	N/A	0.00	N/A	0.95	0.73	0.99	0.90	0.72	0.64
audiology.standardized	standard	10	100	0.07	N/A	0.00	N/A	0.83	0.74	0.80	0.80	0.31	0.59
audiology.standardized	standard	20	80	0.09	14.84	0.01	0.01	0.74	0.75	0.78	0.77	0.30	0.68
audiology.standardized	standard	20	100	0.07	11.07	0.00	0.02	0.78	0.74	0.78	0.77	0.29	0.67

“Audiology (Original)” veri seti için 2 metrik; metodu “standard”, boyut sayısı 100 ve komşuluk sayısı 1 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre metodu “standard”, boyut sayısı 80 ve komşuluk sayısı 20 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

Tablo 36 “Breast Cancer” Veri Setinin LLE için sonuçları

Veri Seti	Mthd	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
breast-cancer	ltsa	20	10	0.43	N/A	0.04	N/A	0.76	0.72	0.95	0.94	0.51	0.61
breast-cancer	modified	10	10	0.96	N/A	0.08	N/A	0.99	0.66	0.45	0.47	-0.01	0.39
breast-cancer	modified	40	10	0.25	N/A	0.04	N/A	0.62	0.73	0.94	0.94	0.51	0.64
breast-cancer	modified	60	20	0.11	58.04	0.03	N/A	0.42	0.73	0.91	0.90	0.44	0.67
breast-cancer	modified	100	8	0.14	N/A	0.03	N/A	0.33	0.72	0.92	0.91	0.40	0.65
breast-cancer	modified	100	20	0.10	57.72	0.04	N/A	0.44	0.72	0.88	0.87	0.35	0.64
breast-cancer	standard	1	40	0.32	N/A	0.03	N/A	0.94	0.69	0.99	0.87	0.40	0.53
breast-cancer	standard	60	40	0.12	26.08	0.02	N/A	0.66	0.74	0.83	0.81	0.35	0.61
breast-cancer	standard	100	40	0.13	26.32	0.04	N/A	0.66	0.74	0.78	0.77	0.25	0.57

“Breast Cancer” veri seti için hiçbir metrik için en iyi değeri vermezken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre metodu “modified”, boyut sayısı 20 ve komşuluk sayısı 60 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

Tablo 37 “Soybean (Large)” Veri Setinin LLE için sonuçları

Veri Seti	Mthd	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
soybean-large	ltsa	80	40	0.12	N/A	0.08	N/A	0.71	0.80	0.80	0.80	0.29	0.58
soybean-large	ltsa	250	5	0.20	N/A	0.14	N/A	0.23	0.73	0.81	0.80	0.11	0.58
soybean-large	modified	20	8	0.30	N/A	0.14	N/A	0.33	0.74	0.91	0.90	0.50	0.69
soybean-large	standard	1	60	0.26	N/A	-0.10	N/A	1.00	0.71	1.00	0.94	0.54	0.59
soybean-large	standard	2	100	0.14	N/A	0.00	N/A	0.96	0.74	0.97	0.90	0.61	0.59
soybean-large	standard	3	20	0.27	N/A	0.14	N/A	0.68	0.72	0.96	0.97	0.50	0.66
soybean-large	standard	7	10	0.39	N/A	0.25	N/A	0.76	0.71	0.92	0.91	0.44	0.64
soybean-large	standard	40	80	0.09	48.67	0.07	N/A	0.90	0.77	0.80	0.74	0.32	0.59
soybean-large	standard	40	100	0.09	37.85	0.07	N/A	0.94	0.77	0.73	0.72	0.27	0.57

“Soybean (Large)” veri seti için hiçbir metrik için en iyi değeri vermezken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre metodu “modified”, boyut sayısı 8 ve komşuluk sayısı 20 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

Tablo 38 “Primary Tumor” Veri Setinin LLE için sonuçları

Veri Seti	Mthd	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
primary-tumor	hessian	250	9	0.12	N/A	0.04	N/A	0.31	0.73	0.79	0.79	0.15	0.56
primary-tumor	ltsa	1	9	0.97	N/A	0.13	N/A	0.99	0.67	0.68	0.53	0.01	0.43
primary-tumor	modified	20	10	0.15	N/A	0.04	N/A	0.42	0.73	0.95	0.96	0.56	0.69
primary-tumor	modified	40	10	0.13	N/A	0.05	N/A	0.33	0.73	0.95	0.96	0.56	0.69
primary-tumor	modified	60	40	0.19	40.70	0.08	N/A	0.56	0.76	0.88	0.84	0.41	0.62
primary-tumor	standard	1	9	0.42	N/A	0.01	N/A	0.99	0.69	0.99	0.87	0.39	0.58
primary-tumor	standard	60	40	0.11	37.89	0.09	N/A	0.55	0.75	0.85	0.84	0.37	0.62

primary-tumor	standard	80	20	0.11	83.01	0.07	N/A	0.46	0.75	0.91	0.90	0.44	0.64
---------------	----------	----	----	-------------	-------	------	-----	------	------	------	------	------	------

“Primary Tumor” veri seti için hiçbir metrik için en iyi değeri vermezken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre metodu “modified”, boyut sayısı 10 ve komşuluk sayısı 40 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

Tablo 39 “Congressional Voting Records” Veri Setinin LLE için sonuçları

Veri Seti	Mthd	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
house-votes-84	modified	40	10	0.24	N/A	-0.01	N/A	0.75	0.72	0.93	0.95	0.52	0.63
house-votes-84	modified	40	40	0.38	N/A	0.03	N/A	0.99	0.68	0.59	0.62	0.13	0.42
house-votes-84	modified	100	2	0.48	N/A	0.01	N/A	0.56	0.68	0.91	0.93	0.43	0.61
house-votes-84	standard	1	20	0.36	N/A	0.01	N/A	1.00	0.69	0.99	0.90	0.32	0.56
house-votes-84	standard	60	40	0.17	79.26	0.00	N/A	0.89	0.72	0.76	0.85	0.43	0.54
house-votes-84	standard	100	20	0.14	N/A	0.00	N/A	0.82	0.72	0.82	0.82	0.38	0.56
house-votes-84	standard	250	20	0.15	N/A	0.00	N/A	0.80	0.74	0.45	0.49	0.03	0.43

“Congressional Voting Records” veri seti için 2 metrik; metodu “modified”, boyut sayısı 10 ve komşuluk sayısı 40 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman da RES metriğine göre metodu “modified”, boyut sayısı 10 ve komşuluk sayısı 40 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

Tablo 40 “Balance Scale” Veri Setinin LLE için sonuçları

Veri Seti	Mthd	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
balance-scale	ltsa	10	10	0.33	N/A	0.11	N/A	0.91	0.68	0.91	0.85	0.29	0.55
balance-scale	modified	40	10	0.14	N/A	-0.01	-0.02	0.51	0.69	0.92	0.91	0.37	0.62
balance-scale	modified	100	10	0.15	N/A	0.01	-0.02	0.56	0.68	0.90	0.92	0.46	0.61
balance-scale	standard	1	9	0.29	N/A	-0.04	N/A	0.87	0.68	0.99	0.77	0.08	0.54
balance-scale	standard	10	10	0.19	N/A	0.07	-0.02	0.70	0.68	0.94	0.96	0.39	0.61
balance-scale	standard	20	10	0.14	N/A	0.05	N/A	0.53	0.68	0.94	0.95	0.38	0.63
balance-scale	standard	80	10	0.13	N/A	0.02	-0.02	0.49	0.69	0.91	0.91	0.44	0.62

“Balance Scale” veri seti için 2 metrik; metodu “standard”, boyut sayısı 10 ve komşuluk sayısı 80 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman da RES metriğine göre metodu “standard”, boyut sayısı 10 ve komşuluk sayısı 20 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

Tablo 41 “Tic-Tac-Toe Endgame” Veri Setinin LLE için sonuçları

Veri Seti	Mthd	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
tic-tac-toe	hessian	750	20	0.26	N/A	-0.01	0.00	0.84	0.68	0.80	0.76	0.13	0.42
tic-tac-toe	ltsa	1	20	0.98	N/A	0.13	N/A	1.00	0.67	0.52	0.51	0.00	0.35
tic-tac-toe	ltsa	60	10	0.11	N/A	0.00	-0.01	0.34	0.68	0.96	0.97	0.56	0.68
tic-tac-toe	ltsa	60	20	0.07	N/A	0.01	0.00	0.31	0.68	0.98	0.98	0.66	0.66
tic-tac-toe	modified	40	20	0.10	N/A	0.00	0.00	0.43	0.68	0.99	0.97	0.67	0.65
tic-tac-toe	standard	750	20	0.07	N/A	0.00	0.00	0.28	0.68	0.72	0.70	0.11	0.49

“Tic-Tac-Toe Endgame” veri seti için 2’şer metrik; metodu “modified”, boyut sayısı 20, komşuluk sayısı 40 ve metodu “standard”, boyut sayısı 20, komşuluk sayısı 750 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman da RES metriğine göre metodu “standard”, boyut sayısı 20 ve komşuluk sayısı 40 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

6.4. T-Dağıtılmış Stokastik Komşu Gömmе (t-SNE) Algoritmasının Sonuçları

T-Dağıtılmış Stokastik Komşu Gömmе (t-SNE) algoritması komşuluk sınırlarını belirlemek için “perplexity” ve boyut indirgenmesi istenen boyut sayısını parametre olarak almaktadır. Perplexity (Perp) 5, 10, 15, 20, 25, 30, 35, 40, 45 ve 50 ile boyut sayısı (n_components, NoC) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 40, 60, 80, 100, 250, 500, 750 ve 1000 olarak veri setinin sahip olduğu değişken sayısına göre parametre değeri olarak algoritmaya verilmiştir. Ayrıca t-SNE algoritmasının başlangıç adımı için (init) “pca” ve “random” olmak üzere parametre ile de diğer kombinasyonlar denenmiştir.

Tablo 42 t-SNE için kullanılan parametre ve değerleri

Kısaltma	Parametre Adı	Aldığı Değerler
Init	Başlangıç Adımı	pca, random
NoC	İndirgenecek Boyut Sayısı	2,3,4,5,6,7,8,9,10,20,40,60,80,100,250,500,750,1000
Perp	Perplexity	5,10,15,20,25,30,35,40,45,50

Tablo 43 “Lenses” Veri Setinin t-SNE için sonuçları

Veri Seti	Init	NoC	Perp	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
lenses_adjusted	pca	2	15	0.32	25.99	0.11	-0.23	0.59	0.69	0.88	0.82	0.40	0.60
lenses_adjusted	pca	3	5	0.27	0.95	0.10	-0.04	0.70	0.73	0.92	0.82	0.55	0.62
lenses_adjusted	random	2	10	0.33	0.17	0.00	-0.08	0.65	0.73	0.90	0.85	0.46	0.62
lenses_adjusted	random	3	10	0.35	0.98	0.12	-0.06	1.00	0.68	0.66	0.58	0.16	0.50

“Lenses” veri seti için 3 metrik; başlangıç metodu “pca”, boyut sayısı 3 ve perplexity 5 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman da RES metriğine göre gene metodu “pca”, boyut sayısı 3 ve perplexity 5 için en iyi sonucu vermektedir.

Tablo 44 “Soybean (Small)” Veri Setinin t-SNE için sonuçları

Veri Seti	Init	NoC	Perp	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
soybean-small	pca	2	5	0.21	0.77	0.37	-0.02	0.21	0.81	0.98	0.98	0.70	0.77
soybean-small	pca	2	10	0.28	0.56	0.40	-0.02	0.32	0.80	0.99	0.98	0.71	0.75
soybean-small	pca	2	15	0.21	0.28	0.38	-0.02	0.21	0.80	0.99	0.99	0.70	0.77
soybean-small	pca	2	30	0.20	3.12	0.26	-0.02	0.21	0.78	0.99	0.99	0.74	0.77
soybean-small	random	2	25	0.20	1.03	0.31	-0.02	0.22	0.78	0.99	0.99	0.74	0.77
soybean-small	random	2	40	0.18	41.70	0.31	-0.03	0.17	0.71	0.97	0.97	0.55	0.71
soybean-small	random	2	45	0.18	N/A	0.35	-0.04	0.17	0.71	0.96	0.96	0.47	0.64

“Soybean (Small)” veri seti için 2 metrik; başlangıç metodu “pca”, boyut sayısı 2 ve perplexity 30 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman da RES metriğine göre gene metodu “pca”, boyut sayısı 2 ve perplexity 5 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

Tablo 45 “Molecular Biology (Promoter Gene Sequences)” Veri Setinin t-SNE için sonuçları

Veri Seti	Init	NoC	Perp	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
promoters_adjusted	pca	2	15	0.40	0.29	0.02	-0.14	0.91	0.77	0.83	0.79	0.43	0.60
promoters_adjusted	pca	2	35	0.39	2.58	0.00	-0.13	0.87	0.76	0.87	0.81	0.43	0.61
promoters_adjusted	pca	2	50	0.38	22.42	-0.01	-0.13	0.83	0.72	0.84	0.83	0.37	0.58
promoters_adjusted	pca	3	5	0.40	0.86	0.03	-0.13	0.98	0.72	0.72	0.66	0.28	0.55
promoters_adjusted	pca	3	30	0.33	0.87	-0.01	-0.14	0.97	0.70	0.70	0.63	0.20	0.54
promoters_adjusted	random	2	15	0.40	0.28	-0.02	-0.14	0.92	0.77	0.83	0.80	0.44	0.60
promoters_adjusted	random	2	35	0.39	2.18	-0.01	-0.13	0.89	0.75	0.87	0.82	0.46	0.61

“Molecular Biology (Promoter Gene Sequences)” veri seti için 2 metrik; başlangıç metodu “pca”, boyut sayısı 2 ve perplexity 50 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre metodu “random”, boyut sayısı 2 ve perplexity 35 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

Tablo 46 “Hayes-Roth” Veri Setinin t-SNE için sonuçları

Veri Seti	Init	NoC	Perp	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
hayes-roth	pca	2	10	0.38	N/A	-0.01	N/A	0.78	0.74	0.95	0.89	0.44	0.52
hayes-roth	pca	3	45	0.26	N/A	0.01	N/A	0.58	0.74	0.94	0.90	0.41	0.55
hayes-roth	pca	3	50	0.27	N/A	0.00	N/A	0.63	0.73	0.94	0.91	0.42	0.54
hayes-roth	random	2	10	0.38	0.73	0.02	N/A	0.74	0.74	0.95	0.90	0.44	0.63
hayes-roth	random	2	45	0.34	N/A	-0.01	N/A	0.66	0.72	0.96	0.90	0.42	0.54
hayes-roth	random	3	25	0.34	0.96	0.02	N/A	0.88	0.72	0.91	0.82	0.36	0.59

“Hayes-Roth” veri seti için 3 metrik; başlangıç metodu “pca”, boyut sayısı 3 ve perplexity 45 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre metodu “random”, boyut sayısı 2 ve perplexity 10 için en iyi sonucu vermektedir.

Tablo 47 “Lymphography” Veri Setinin t-SNE için sonuçları

Veri Seti	Init	NoC	Perp	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
lymphography	pca	2	10	0.35	0.77	0.01	-0.08	0.65	0.77	0.94	0.89	0.48	0.65

lymphography	pca	2	20	0.33	0.53	0.00	-0.07	0.56	0.76	0.94	0.90	0.50	0.67
lymphography	pca	2	35	0.31	0.31	0.02	-0.07	0.50	0.74	0.93	0.91	0.45	0.67
lymphography	pca	2	40	0.30	0.28	0.01	-0.06	0.50	0.74	0.94	0.91	0.47	0.67
lymphography	pca	2	45	0.31	0.23	0.00	-0.07	0.51	0.73	0.93	0.91	0.43	0.66
lymphography	random	2	15	0.32	0.65	0.00	-0.07	0.56	0.76	0.95	0.90	0.50	0.67
lymphography	random	2	50	0.31	0.23	-0.01	-0.07	0.51	0.72	0.93	0.91	0.42	0.66
lymphography	random	3	40	0.27	0.92	0.01	-0.05	0.62	0.76	0.94	0.89	0.49	0.66

“Lymphography” veri seti için hiçbir metrik için en iyi değeri vermezken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre metodu “pca”, boyut sayısı 2 ve perplexity 40 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

Tablo 48 “Audiology (Original)” Veri Setinin t-SNE için sonuçları

Veri Seti	Init	NoC	Perp	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
audiology.standardized	pca	2	45	0.29	N/A	0.03	N/A	0.45	0.76	0.93	0.93	0.49	0.59
audiology.standardized	pca	3	10	0.28	N/A	0.02	N/A	0.56	0.78	0.95	0.92	0.57	0.59
audiology.standardized	pca	3	20	0.26	N/A	0.03	N/A	0.50	0.78	0.96	0.95	0.57	0.60
audiology.standardized	random	2	15	0.32	0.66	0.01	-0.07	0.50	0.77	0.95	0.94	0.54	0.69
audiology.standardized	random	2	20	0.32	N/A	0.01	N/A	0.52	0.78	0.94	0.93	0.52	0.58
audiology.standardized	random	2	30	0.32	0.63	0.02	-0.07	0.52	0.77	0.94	0.93	0.55	0.68

“Audiology (Original)” veri seti için 4 metrik; başlangıç metodu “pca”, boyut sayısı 3 ve perplexity 20 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre metodu “random”, boyut sayısı 2 ve perplexity 15 için en iyi sonucu vermektedir.

Tablo 49 “Breast Cancer” Veri Setinin t-SNE için sonuçları

Veri Seti	Init	NoC	Perp	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
breast-cancer	pca	2	50	0.33	0.54	0.01	N/A	0.57	0.74	0.97	0.92	0.49	0.67
breast-cancer	pca	3	20	0.30	N/A	0.00	N/A	0.68	0.75	0.98	0.92	0.57	0.57
breast-cancer	random	2	50	0.35	0.55	0.03	N/A	0.62	0.74	0.97	0.92	0.50	0.66
breast-cancer	random	3	20	0.30	0.84	0.01	N/A	0.66	0.76	0.98	0.92	0.57	0.67
breast-cancer	random	3	45	0.27	0.87	0.00	N/A	0.62	0.75	0.97	0.94	0.54	0.67

“Breast Cancer” veri seti için 3 farklı parametre konfigürasyonu 2’şer metrik için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre metodu “random”, boyut sayısı 3 ve perplexity 45 için en iyi sonucu vermektedir.

Tablo 50 “Soybean (Large)” Veri Setinin t-SNE için sonuçları

Veri Seti	Init	NoC	Perp	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
soybean-large	pca	2	20	0.30	N/A	0.12	N/A	0.37	0.80	0.99	0.98	0.69	0.63
soybean-large	pca	2	30	0.31	0.58	0.10	N/A	0.39	0.79	0.99	0.98	0.70	0.73
soybean-large	pca	3	15	0.29	N/A	0.08	N/A	0.51	0.81	0.99	0.98	0.73	0.62
soybean-large	pca	3	20	0.29	N/A	0.08	N/A	0.52	0.81	0.99	0.98	0.73	0.62
soybean-large	pca	3	25	0.28	N/A	0.08	N/A	0.56	0.81	0.99	0.98	0.71	0.62
soybean-large	random	2	20	0.30	N/A	0.11	N/A	0.36	0.79	0.99	0.98	0.69	0.63
soybean-large	random	3	5	0.27	0.55	0.09	N/A	0.59	0.80	0.99	0.98	0.70	0.71
soybean-large	random	3	15	0.28	0.48	0.10	N/A	0.51	0.80	0.99	0.98	0.72	0.72

“Soybean (Large)” veri seti için 2 metrik; başlangıç metodu “pca”, boyut sayısı 3 ve perplexity 20 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre metodu “pca”, boyut sayısı 2 ve perplexity 30 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

Tablo 51 “Primary Tumor” Veri Setinin t-SNE için sonuçları

Veri Seti	Init	NoC	Perp	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
primary-tumor	pca	3	15	0.39	N/A	0.00	N/A	0.85	0.75	0.98	0.92	0.55	0.54
primary-tumor	pca	3	20	0.32	N/A	0.02	N/A	0.73	0.75	0.98	0.92	0.53	0.56
primary-tumor	random	2	20	0.38	0.73	0.05	N/A	0.73	0.74	0.96	0.91	0.49	0.65
primary-tumor	random	2	50	0.34	0.50	0.03	N/A	0.65	0.73	0.94	0.92	0.46	0.65
primary-tumor	random	3	50	0.27	0.85	0.05	N/A	0.60	0.74	0.96	0.93	0.50	0.67

“Primary Tumor” veri seti için 3 metrik; başlangıç metodu “random”, boyut sayısı 3 ve perplexity 50 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman da RES metriğine göre gene metodu “random”, boyut sayısı 3 ve perplexity 50 için en iyi sonucu vermektedir.

Tablo 52 “Congressional Voting Records” Veri Setinin t-SNE için sonuçları

Veri Seti	Init	NoC	Perp	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
house-votes-84	pca	2	50	0.25	0.57	0.01	N/A	0.30	0.73	0.96	0.95	0.44	0.70
house-votes-84	pca	3	15	0.27	0.57	0.01	N/A	0.40	0.74	0.98	0.95	0.53	0.70
house-votes-84	random	2	5	0.29	N/A	0.01	N/A	0.44	0.74	0.97	0.94	0.50	0.59
house-votes-84	random	3	30	0.23	0.47	0.01	N/A	0.32	0.74	0.98	0.96	0.51	0.71

“Congressional Voting Records” veri seti için 4 metrik; başlangıç metodu “random”, boyut sayısı 3 ve perplexity 30 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman da RES metriğine göre gene metodu “random”, boyut sayısı 3 ve perplexity 30 için en iyi sonucu vermektedir.

Tablo 53 “Balance Scale” Veri Setinin t-SNE için sonuçları

Veri Seti	Init	NoC	Perp	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
balance-scale	pca	2	25	0.41	0.60	0.00	-0.14	0.91	0.68	0.97	0.80	0.21	0.57
balance-scale	pca	3	10	0.33	0.88	0.05	-0.08	0.88	0.68	0.98	0.81	0.27	0.59
balance-scale	pca	3	15	0.33	0.86	-0.01	-0.08	0.87	0.69	0.99	0.82	0.28	0.59
balance-scale	pca	3	40	0.32	0.89	0.02	-0.08	0.86	0.69	0.98	0.83	0.28	0.59
balance-scale	random	2	15	0.40	1.16	-0.01	-0.14	0.87	0.69	0.98	0.83	0.27	0.59
balance-scale	random	3	15	0.33	0.87	0.02	-0.08	0.87	0.69	0.98	0.82	0.31	0.60
balance-scale	random	3	20	0.33	0.87	-0.01	-0.09	0.89	0.69	0.98	0.80	0.28	0.59

“Balance Scale” veri seti için 2 metrik; başlangıç metodu “pca”, boyut sayısı 3 ve perplexity 40 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre metodu “random”, boyut sayısı 3 ve perplexity 15 için en iyi sonucu vermektedir. Burada görüldüğü gibi bütün metrikler için tek bir parametre seti bulunmamaktadır.

Tablo 54 “Tic-Tac-Toe Endgame” Veri Setinin t-SNE için sonuçları

Veri Seti	Init	NoC	Perp	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
tic-tac-toe	pca	3	50	0.32	0.82	0.00	-0.07	0.81	0.69	0.98	0.87	0.23	0.60
tic-tac-toe	random	2	25	0.40	0.61	0.01	-0.13	0.89	0.68	0.98	0.85	0.20	0.59
tic-tac-toe	random	3	25	0.32	0.42	0.00	-0.08	0.83	0.68	0.99	0.87	0.21	0.60
tic-tac-toe	random	3	30	0.32	0.50	-0.01	-0.07	0.82	0.68	0.99	0.87	0.24	0.60

tic-tac-toe	random	3	35	0.32	0.45	0.00	-0.07	0.82	0.68	0.99	0.88	0.23	0.60
-------------	--------	---	----	------	------	------	-------	------	------	------	-------------	------	-------------

“Tic-Tac-Toe Endgame” veri seti için 3 metrik; başlangıç metodu “pca”, boyut sayısı 3 ve perplexity 50 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman da RES metriğine göre gene metodu “random”, boyut sayısı 3 ve perplexity 35 için en iyi sonucu vermektedir.

Tablo 55 “Car Evaluation” Veri Setinin t-SNE için sonuçları

Veri Seti	Init	NoC	Perp	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
car	pca	2	50	0.39	0.75	0.00	-0.13	0.85	0.68	0.99	0.86	0.22	0.59
car	pca	3	10	0.31	N/A	0.00	N/A	0.75	0.67	0.92	0.88	0.05	0.48
car	pca	3	15	0.31	0.76	-0.01	-0.08	0.80	0.69	1.00	0.87	0.30	0.61
car	pca	3	40	0.32	0.79	-0.03	-0.08	0.83	0.69	0.99	0.87	0.30	0.60
car	random	2	40	0.39	0.72	-0.01	-0.13	0.84	0.68	0.99	0.86	0.22	0.59
car	random	3	15	0.31	0.76	0.00	-0.07	0.79	0.69	1.00	0.88	0.30	0.61
car	random	3	20	0.31	0.77	-0.01	-0.08	0.81	0.69	1.00	0.87	0.30	0.61

“Car Evaluation” veri seti için 3 metrik; başlangıç metodu “pca”, boyut sayısı 3 ve perplexity 10 için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman da RES metriğine göre gene metodu “random”, boyut sayısı 3 ve perplexity 15 için en iyi sonucu vermektedir.

6.5. Standart Manifold Yaklaşımı ve Projeksiyonu (UMAP) Algoritmasının Sonuçları

Standart manifold yaklaşımı ve projeksiyonu (UMAP) algoritması komşuluk sayısı, boyut indirgemedeki kayıtların birbirine olan mesafesi, boyut indirgenmesi istenen boyut sayısını ve kayıtları arası mesafeyi hesaplamak için kullanacağı fonksiyonu parametre olarak almaktadır. Komşuluk sayısı (n_neighbors, NoN) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 40, 60, 80, 100, 250, 500, 750 ve 1000 ile boyut sayısı (n_components, NoC) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 40, 60, 80, 100, 250, 500, 750 ve 1000 olarak veri setinin sahip olduğu değişken sayısına göre parametre değeri olarak algoritmaya verilmiştir. Ayrıca UMAP algoritması kayıtlar arası mesafeyi hesaplaması için “euclidean”, “canberra”, “mahalanobis” ve “cosine” metrikleri diğer parametreler ile verilmiştir.

Tablo 56 UMAP için kullanılan parametre ve değerleri

Kısaltma	Parametre Adı	Aldığı Değerler
Dist	Noktalar arası uzaklık	0.0, 0.1, 0.25, 0.5, 0.8, 0.99
Met	Uzaklık Ölçüm Metodu	euclidean, canberra, mahalanobis, cosine
NoC	İndirgenecek Boyut Sayısı	2,3,4,5,6,7,8,9,10,20,40,60,80,100,250,500,750,1000
NoN	Komşuluk Sayısı	1,2,3,4,5,6,7,8,9,10,20,40,60,80,100,250,500,750,1000

Tablo 57 “Lenses” Veri Setinin UMAP için sonuçları

Veri Seti	NoN	Dist	NoC	Met	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
lenses_adjusted	6	0.25	4	cosine	0.13	0.06	0.27	-0.02	0.28	0.72	0.96	0.96	0.69	0.79
lenses_adjusted	6	0.25	5	correlation	0.08	0.05	0.18	-0.02	0.12	0.74	1.00	1.00	0.82	0.81
lenses_adjusted	6	0.50	3	canberra	0.21	0.05	0.19	-0.03	0.49	0.78	0.94	0.87	0.61	0.75
lenses_adjusted	6	0.80	5	euclidean	0.03	0.02	0.14	-0.01	0.02	0.73	1.00	1.00	0.82	0.83
lenses_adjusted	8	0.50	5	canberra	0.04	0.00	0.06	-0.01	0.04	0.73	1.00	1.00	0.82	0.81
lenses_adjusted	20	0.99	8	correlation	0.02	0.04	0.08	0.00	0.01	0.75	1.00	1.00	0.82	0.73

“Lenses” veri seti için 3 metrik; komşuluk sayısı 6, noktalar arası uzaklık 0.25, boyut sayısı 5 ve mesafe hesaplama algoritması “correlation” için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre komşuluk sayısı 6, noktalar arası uzaklık 0.80, boyut sayısı 5 ve mesafe hesaplama algoritması “euclidean” için en iyi sonucu vermektedir.

Tablo 58 “Soybean (Small)” Veri Setinin UMAP için sonuçları

Veri Seti	NoN	Dist	NoC	Met	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
soybean-small	6	0.5	5	cosine	0.19	0.51	0.42	-0.01	0.20	0.78	0.99	0.99	0.76	0.86
soybean-small	7	0.99	10	euclidean	0.24	0.62	0.01	-0.02	0.33	0.81	0.99	1.00	0.81	0.83
soybean-small	9	0.8	20	canberra	0.21	0.71	0.35	-0.01	0.25	0.83	0.99	0.99	0.77	0.83
soybean-small	10	0.8	20	euclidean	0.28	0.66	0.25	-0.02	0.36	0.79	1.00	0.99	0.80	0.81
soybean-small	20	0.1	5	canberra	0.34	0.71	0.47	-0.02	0.40	0.76	0.99	0.99	0.71	0.79
soybean-small	20	0.99	10	correlation	0.17	0.15	0.36	-0.01	0.22	0.77	0.99	0.99	0.78	0.83
soybean-small	20	0.99	20	correlation	0.22	0.18	0.38	-0.01	0.31	0.79	0.99	1.00	0.79	0.80
soybean-small	40	0.99	20	euclidean	0.09	0.45	0.34	0.00	0.15	0.77	0.99	0.99	0.77	0.78
soybean-small	40	0.99	40	mahalanobis	0.06	1.29	-0.02	0.08	1.00	0.68	0.55	0.53	0.02	0.48

“Soybean (Small)” veri seti için metriklere göre en iyi değeri veren ortak sonuç yokken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre komşuluk sayısı 6, noktalar arası uzaklık 0.5, boyut sayısı 5 ve mesafe hesaplama algoritması “cosine” için en iyi sonucu vermektedir.

Tablo 59 “Molecular Biology (Promoter Gene Sequences)” Veri Setinin UMAP için sonuçları

Veri Seti	NoN	Dist	NoC	Met	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
promoters_adjusted	2	0.1	80	canberra	0.26	N/A	0.07	0.13	0.91	0.73	0.74	0.68	0.28	0.54
promoters_adjusted	2	0.99	3	canberra	0.47	0.41	0.02	0.18	0.98	0.72	0.74	0.67	0.29	0.64
promoters_adjusted	20	0.99	20	euclidean	0.15	6.66	0.00	0.02	0.69	0.76	0.97	0.96	0.63	0.76
promoters_adjusted	80	0.99	20	correlation	0.11	7.89	0.01	0.01	0.56	0.76	0.98	0.97	0.67	0.73
promoters_adjusted	100	0.99	60	canberra	0.10	8.03	0.03	0.01	0.56	0.78	0.98	0.97	0.66	0.69
promoters_adjusted	100	0.99	60	euclidean	0.10	8.10	0.01	0.01	0.55	0.77	0.98	0.97	0.64	0.69
promoters_adjusted	100	0.99	80	cosine	0.10	8.03	0.01	0.01	0.56	0.77	0.98	0.97	0.66	0.69
promoters_adjusted	100	0.99	100	canberra	0.10	8.03	0.01	0.01	0.56	0.78	0.98	0.97	0.66	0.68
promoters_adjusted	100	0.99	100	mahalanobis	0.07	9.79	0.01	0.00	1.00	0.67	0.51	0.52	0.00	0.47

“Molecule Biology (Promoters Gene Sequences)” veri seti için 2 metrik; komşuluk sayısı 100, noktalar arası uzaklık 0.99, boyut sayısı 60 ve mesafe hesaplama algoritması “canberra” için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre komşuluk sayısı 20, noktalar arası uzaklık 0.99, boyut sayısı 20 ve mesafe hesaplama algoritması “euclidean” için en iyi sonucu vermektedir.

Tablo 60 “Hayes-Roth” Veri Setinin UMAP için sonuçları

Veri Seti	NoN	Dist	NoC	Met	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
hayes-roth	6	0.99	10	cosine	0.19	0.16	0.03	N/A	0.49	0.72	0.97	0.96	0.48	0.72
hayes-roth	6	0.99	10	euclidean	0.19	0.16	0.01	N/A	0.48	0.72	0.97	0.96	0.48	0.72
hayes-roth	8	0.8	4	canberra	0.21	0.13	0.01	N/A	0.44	0.72	0.95	0.95	0.41	0.76
hayes-roth	20	0.99	9	canberra	0.16	0.09	0.01	N/A	0.33	0.73	0.95	0.96	0.39	0.73
hayes-roth	80	0.99	9	canberra	0.16	0.09	0.00	N/A	0.35	0.73	0.94	0.96	0.39	0.68
hayes-roth	100	0.25	6	mahalanobis	0.42	N/A	0.17	N/A	1.00	0.67	0.53	0.56	0.03	0.38
hayes-roth	100	0.5	6	canberra	0.21	0.07	0.00	N/A	0.43	0.74	0.95	0.95	0.41	0.68

hayes-roth	100	0.5	7	cosine	0.20	0.06	-0.01	N/A	0.41	0.74	0.95	0.95	0.41	0.67
------------	-----	-----	---	--------	------	-------------	-------	-----	------	------	------	------	------	------

“Hayes-Roth” veri seti için 2 metrik; komşuluk sayısı 20, noktalar arası uzaklık 0.99, boyut sayısı 9 ve mesafe hesaplama algoritması “canberra” için en iyi değerleri verirken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre komşuluk sayısı 8, noktalar arası uzaklık 0.80, boyut sayısı 4 ve mesafe hesaplama algoritması “euclidean” için en iyi sonucu vermektedir.

Tablo 61 “Lymphography” Veri Setinin UMAP için sonuçları

Veri Seti	NoN	Dist	NoC	Met	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
lymphography	100	0.99	20	cosine	0.15	N/A	0.01	-0.01	0.46	0.75	0.98	0.99	0.63	0.62
lymphography	10	0.1	4	mahalanobis	0.64	1.15	0.13	N/A	1.00	0.67	0.53	0.48	-0.01	0.54
lymphography	3	0.99	3	euclidean	0.30	0.07	0.00	-0.06	0.72	0.73	0.91	0.89	0.46	0.74
lymphography	4	0.8	9	correlation	0.23	0.12	0.00	-0.03	0.66	0.76	0.96	0.93	0.56	0.76
lymphography	7	0.1	40	mahalanobis	0.10	0.92	0.01	0.02	1.00	0.67	0.52	0.50	0.01	0.54
lymphography	7	0.99	8	correlation	0.20	0.12	0.00	-0.02	0.57	0.74	0.98	0.97	0.64	0.79
lymphography	80	0.99	10	euclidean	0.15	0.28	0.00	-0.01	0.44	0.75	0.98	0.98	0.62	0.75
lymphography	80	0.99	40	cosine	0.16	N/A	0.01	-0.01	0.44	0.75	0.98	0.98	0.64	0.60

“Lymphography” veri seti için metriklerle göre en iyi değeri veren ortak sonuç yokken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre komşuluk sayısı 7, noktalar arası uzaklık 0.99, boyut sayısı 8 ve mesafe hesaplama algoritması “correlation” için en iyi sonucu vermektedir.

Tablo 62 “Audiology (Original)” Veri Setinin UMAP için sonuçları

Veri Seti	NoN	Dist	NoC	Met	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
audiology.standardized	20	0.99	100	mahalanobis	0.06	0.93	-0.03	0.04	1.00	0.67	0.50	0.51	0.01	0.57
audiology.standardized	2	0.5	40	cosine	0.34	0.67	0.04	-0.06	1.00	0.73	0.79	0.69	0.35	0.66
audiology.standardized	4	0.99	100	canberra	0.27	0.10	0.00	-0.03	0.61	0.75	0.96	0.95	0.56	0.75
audiology.standardized	60	0.99	40	euclidean	0.20	0.28	0.02	-0.01	0.49	0.75	0.97	0.98	0.57	0.77
audiology.standardized	60	0.99	80	canberra	0.21	0.26	0.01	-0.02	0.49	0.75	0.97	0.98	0.57	0.76
audiology.standardized	6	0.99	10	euclidean	0.26	0.11	0.01	-0.03	0.59	0.76	0.97	0.97	0.60	0.79
audiology.standardized	7	0.8	60	canberra	0.29	0.16	0.00	-0.03	0.58	0.76	0.97	0.97	0.61	0.77
audiology.standardized	80	0.99	80	canberra	0.20	0.30	0.01	-0.01	0.49	0.75	0.97	0.98	0.57	0.75

audiology.standardized	8	0.99	7	euclidean	0.25	0.12	0.01	-0.03	0.55	0.75	0.97	0.97	0.59	0.79
------------------------	---	------	---	-----------	------	------	------	-------	------	------	------	------	------	-------------

“Audiology (Original)” veri seti için metriklerle göre en iyi değeri veren ortak sonuç yokken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre komşuluk sayısı 8, noktalar arası uzaklık 0.99, boyut sayısı 7 ve mesafe hesaplama algoritması “euclidean” için en iyi sonucu vermektedir.

Tablo 63 “Breast Cancer” Veri Setinin UMAP için sonuçları

Veri Seti	NoN	Dist	NoC	Met	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
breast-cancer	10	0.1	4	mahalanobis	0.52	N/A	0.16	N/A	1.00	0.66	0.47	0.52	0.00	0.46
breast-cancer	250	0.99	20	euclidean	0.15	0.05	-0.02	N/A	0.51	0.74	0.99	0.98	0.57	0.68
breast-cancer	250	0.99	40	cosine	0.16	0.05	-0.01	N/A	0.51	0.73	0.99	0.98	0.56	0.63
breast-cancer	250	0.99	40	euclidean	0.15	0.05	-0.02	N/A	0.51	0.74	0.99	0.98	0.58	0.63
breast-cancer	40	0.1	40	mahalanobis	0.10	N/A	0.00	N/A	1.00	0.67	0.49	0.51	-0.01	0.40
breast-cancer	5	0.8	20	cosine	0.23	0.07	0.00	N/A	0.65	0.75	0.98	0.96	0.57	0.74
breast-cancer	5	0.99	10	canberra	0.22	0.08	-0.01	N/A	0.65	0.74	0.98	0.96	0.59	0.77
breast-cancer	7	0.99	7	cosine	0.21	0.06	0.00	N/A	0.61	0.73	0.98	0.97	0.57	0.78
breast-cancer	80	0.99	40	canberra	0.18	0.04	-0.02	N/A	0.54	0.73	0.98	0.98	0.56	0.68

“Breast Cancer” veri seti için metriklerle göre en iyi değeri veren ortak sonuç yokken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre komşuluk sayısı 7, noktalar arası uzaklık 0.99, boyut sayısı 7 ve mesafe hesaplama algoritması “cosine” için en iyi sonucu vermektedir.

Tablo 64 “Soybean (Large)” Veri Setinin UMAP için sonuçları

Veri Seti	NoN	Dist	NoC	Met	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
soybean-large	4	0.99	60	cosine	0.37	0.39	0.24	N/A	0.54	0.76	0.99	0.97	0.70	0.77
soybean-large	5	0.99	40	canberra	0.39	0.35	0.21	N/A	0.51	0.77	0.99	0.98	0.74	0.79
soybean-large	6	0.99	7	canberra	0.29	0.29	0.17	N/A	0.34	0.75	0.99	0.99	0.76	0.84
soybean-large	8	0.99	7	cosine	0.23	0.29	0.17	N/A	0.26	0.76	0.99	0.99	0.73	0.85
soybean-large	9	0.99	20	euclidean	0.20	0.15	0.17	N/A	0.22	0.76	0.99	0.99	0.72	0.85
soybean-large	10	0.99	9	cosine	0.23	0.14	0.12	N/A	0.29	0.76	0.99	0.99	0.72	0.84
soybean-large	60	0.99	20	correlation	0.23	0.27	0.11	N/A	0.34	0.75	0.99	0.99	0.70	0.81
soybean-large	250	0.1	100	mahalanobis	0.07	N/A	0.00	N/A	1.00	0.67	0.50	0.51	0.00	0.35

soybean-large	250	0.99	100	euclidean	0.18	0.95	0.07	N/A	0.42	0.76	0.99	1.00	0.71	0.68
---------------	-----	------	-----	-----------	------	------	------	-----	------	------	------	-------------	------	------

“Soybean (Large)” veri seti için metriklerle göre en iyi değeri veren ortak sonuç yokken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre komşuluk sayısı 8, noktalar arası uzaklık 0.99, boyut sayısı 7 ve mesafe hesaplama algoritması “cosine” için en iyi sonucu vermektedir.

Tablo 65 “Primary Tumor” Veri Setinin UMAP için sonuçları

Veri Seti	NoN	Dist	NoC	Met	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
primary-tumor	4	0.99	40	canberra	0.24	0.10	0.01	N/A	0.77	0.74	0.97	0.95	0.56	0.68
primary-tumor	5	0.99	8	euclidean	0.23	0.08	0.01	N/A	0.74	0.74	0.98	0.96	0.57	0.76
primary-tumor	5	0.99	9	cosine	0.24	0.08	0.02	N/A	0.75	0.74	0.98	0.96	0.57	0.75
primary-tumor	40	0.99	40	correlation	0.22	0.06	0.01	N/A	0.69	0.73	0.98	0.98	0.53	0.68
primary-tumor	100	0.1	40	mahalanobis	0.10	N/A	-0.01	N/A	1.00	0.67	0.51	0.51	0.00	0.39
primary-tumor	250	0.1	2	mahalanobis	1.00	1.00	0.16	N/A	1.00	0.67	0.63	0.53	0.01	0.46
primary-tumor	250	0.99	20	correlation	0.20	0.08	0.01	N/A	0.66	0.73	0.98	0.98	0.55	0.67
primary-tumor	250	0.99	40	canberra	0.20	0.08	0.01	N/A	0.65	0.74	0.98	0.98	0.56	0.62
primary-tumor	250	0.99	40	euclidean	0.20	0.09	0.01	N/A	0.68	0.74	0.98	0.98	0.57	0.62

“Primary Tumor” veri seti için metriklerle göre en iyi değeri veren ortak sonuç yokken bütün metrikleri ve parametreleri hesaba kattığımız zaman RES metriğine göre komşuluk sayısı 5, noktalar arası uzaklık 0.99, boyut sayısı 8 ve mesafe hesaplama algoritması “euclidean” için en iyi sonucu vermektedir.

7. TARTIŞMA

Boyut indirgeme için belirlenen algoritmalarının 9 metrik içerisinde 8 metrik üzerinden çıkan sonuçları karşılaştırılmış ve bütün metriklerin mutabık kaldığı bir parametre konfigürasyonu bu algoritmalar için kullanılan 13 veri seti için belirlenememiştir. Her metrik kendi içerisinde farklı veri setlerin üzerinde değerlendirilen farklı algoritmalarda farklı parametre değerleri için en iyi sonuçları vermiştir.

Bütün metrikler için bir mutabakat sağlanamadığı için bütün metriklerin hesapladığı değerleri normalize ederek eşit olarak alan ve ayrıca algoritmanın aldığı parametre

değerlerini de hesaba katan “Yeniden Düzenlenmiş Ölçeklendirme (Refactored Scaling, RES)” metriği oluşturulmuştur. Her bir algoritma için parametre değerlerinin etkisi bir maliyet fonksiyonu olarak RES metriğine eklemiştir. Örneğin UMAP algoritması için komşuluk sayısı ne kadar fazla ise RES algoritması o kadar kötü sonuç çıkartacaktır. Buradaki amaç boyut indirgeme olduğu için mümkün olduğu kadar az değişken sayısına sahip (az boyutlu) veri seti elde etmek veya daha az komşuluk içeren parametre konfigürasyonları hesaplama içerisinde daha az maliyet yaratacağı için RES metriğinde bu durum fazla komşuluk kötü sonuç çıkartacak şekilde düzenlenmiştir.

RES metriği için değerlendirme ölçüm metriği içerisinde anlatılan dokuz metrikten içerisinde anlamlı ve farklı veri setleri ve farklı parametre değerleri için varyansı yüksek metrikler seçilmiştir. Bu yüzden topolojik çarpımın (topological product, Tpr) açıklama gücü veya diğer bir deyişle varyansı düşük olduğu için RES içerisinde yer almamıştır. Diğer parametreler değerleri de bulunduğu veri seti ve seçilen algoritma için normalize edilerek minimum 0, maksimum 1 olacak şekilde metrik üretilmiştir.

Her metrik ve parametre konfigürasyonları eşit bir şekilde katkıda bulunarak her bir veri seti için en iyi sonuç veren parametre değerleri bulunmuştur. Çıkan sonuçlar da dikkate alındığında her bir veri seti için farklı algoritmalara göre tek bir parametre konfigürasyonu bulunamamış veya bütün veri setleri için genel bir algoritma iyi olarak seçilememiştir. Buradaki bakış açısı her kategorik veri seti kendi içerisinde farklı bir yapısı olduğu için analiz kısmında denenen algoritmalar için genel bir çözüm bulunmamaktadır.

Tablo 66 Bütün Veri Setlerinin Kernel PCA için sonuçları

Veri Seti	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
lenses_adjusted	5	0.00	0.00	-0.04	0.00	0.00	0.75	1.00	1.00	0.82	0.75
soybean-small	20	0.01	0.00	0.32	0.00	0.00	0.82	1.00	1.00	0.83	0.80
promoters_adjusted	80	0.01	0.00	0.02	0.00	0.01	0.88	1.00	1.00	0.83	0.79
hayes-roth	9	0.06	N/A	0.00	N/A	0.06	0.74	0.97	0.97	0.45	0.59
lymphography	20	0.03	0.01	0.00	0.00	0.04	0.79	0.99	0.99	0.72	0.76
audiology.standardized	100	0.01	0.00	0.01	0.00	0.00	0.80	1.00	1.00	0.76	0.78
breast-cancer	9	0.10	0.07	0.02	N/A	0.20	0.72	0.98	0.98	0.47	0.72
soybean-large	40	0.01	0.00	0.18	N/A	0.00	0.85	1.00	1.00	0.84	0.80
primary-tumor	9	0.10	0.06	0.06	N/A	0.21	0.74	0.97	0.99	0.53	0.72

house-votes-84	20	0.03	0.00	0.00	N/A	0.02	0.75	0.99	1.00	0.62	0.74
balance-scale	10	0.14	0.11	0.01	-0.01	0.51	0.68	0.96	0.98	0.23	0.62
tic-tac-toe	20	0.00	0.00	0.01	0.00	0.00	0.68	0.99	0.99	0.24	0.67
car	20	0.00	0.00	0.00	0.00	0.00	0.69	1.00	1.00	0.32	0.66

Çekirdek temel birleşen analizi (Kernel PCA) algoritmasına bütün veri setleri için bakıldığında RES metriği için minimum 0.59 maksimum 0.80 değerini almıştır.

Tablo 67 Bütün Veri Setlerinin Isomap için sonuçları

Veri Seti	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
lenses_adjusted	5	5	0.01	0.26	0.04	0.00	0.00	0.72	1.00	1.00	0.82	0.83
soybean-small	10	5	0.14	0.22	0.27	-0.02	0.19	0.75	0.98	0.98	0.67	0.84
promoters_adjusted	6	20	0.11	0.45	0.02	-0.02	0.66	0.74	0.98	0.97	0.71	0.80
hayes-roth	20	10	0.08	0.21	0.00	N/A	0.12	0.73	0.98	0.98	0.68	0.79
audiology.standardized	9	100	0.11	0.39	0.01	0.00	0.29	0.74	0.99	0.98	0.68	0.82
breast-cancer	20	9	0.10	0.17	0.02	N/A	0.18	0.72	0.97	0.98	0.58	0.82
soybean-large	40	20	0.09	0.22	0.16	N/A	0.10	0.77	0.98	0.99	0.69	0.85
primary-tumor	20	9	0.11	0.17	0.06	N/A	0.21	0.73	0.97	0.98	0.61	0.82
house-votes-84	40	9	0.09	0.16	0.02	N/A	0.11	0.72	0.97	0.98	0.58	0.83
balance-scale	20	10	0.13	0.15	0.02	-0.01	0.50	0.68	0.93	0.97	0.51	0.74
tic-tac-toe	40	10	0.09	0.18	0.00	-0.01	0.25	0.68	0.97	0.98	0.57	0.79
car	100	20	0.00	0.27	0.02	0.00	0.00	0.69	1.00	1.00	0.87	0.81

İzomap (Isomap) algoritmasına bütün veri setleri için bakıldığında RES metriği için minimum 0.74 maksimum 0.85 değerini almıştır.

Tablo 68 Bütün Veri Setlerinin LLE için sonuçları

Veri Seti	Mthd	NoN	NoC	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
lenses_adjusted	standard	5	5	0.05	4.97	0.01	-0.01	0.05	0.73	1.00	1.00	0.82	0.81
soybean-small	modified	9	2	0.32	N/A	0.43	N/A	0.35	0.68	0.97	0.98	0.66	0.71
promoters_adjusted	standard	3	60	0.07	58.47	0.02	0.00	0.89	0.77	0.96	0.95	0.77	0.71
hayes-roth	modified	20	7	0.15	40.07	0.02	N/A	0.39	0.71	0.93	0.93	0.54	0.71
lymphography	modified	40	20	0.11	48.79	0.00	0.00	0.44	0.74	0.89	0.89	0.41	0.68
audiology.standardized	standard	20	80	0.09	14.84	0.01	0.01	0.74	0.75	0.78	0.77	0.30	0.68
breast-cancer	modified	60	20	0.11	58.04	0.03	N/A	0.42	0.73	0.91	0.90	0.44	0.67

soybean-large	modified	20	8	0.30	N/A	0.14	N/A	0.33	0.74	0.91	0.90	0.50	0.69
primary-tumor	modified	40	10	0.13	N/A	0.05	N/A	0.33	0.73	0.95	0.96	0.56	0.69
house-votes-84	modified	40	10	0.24	N/A	-0.01	N/A	0.75	0.72	0.93	0.95	0.52	0.63
balance-scale	standard	20	10	0.14	N/A	0.05	N/A	0.53	0.68	0.94	0.95	0.38	0.63
tic-tac-toe	ltsa	60	10	0.11	N/A	0.00	-0.01	0.34	0.68	0.96	0.97	0.56	0.68

Lokal Doğrusal Gömme (LLE) algoritmasına bütün veri setleri için bakıldığında RES metriği için minimum 0.63 maksimum 0.81 değerini almıştır.

Tablo 69 Bütün Veri Setlerinin t-SNE için sonuçları

Veri Seti	Init	NoC	Perp	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
lenses_adjusted	pca	3	5	0.27	0.95	0.10	-0.04	0.70	0.73	0.92	0.82	0.55	0.62
soybean-small	pca	2	5	0.21	0.77	0.37	-0.02	0.21	0.81	0.98	0.98	0.70	0.77
promoters_adjusted	random	2	35	0.39	2.18	-0.01	-0.13	0.89	0.75	0.87	0.82	0.46	0.61
hayes-roth	random	2	10	0.38	0.73	0.02	N/A	0.74	0.74	0.95	0.90	0.44	0.63
lymphography	pca	2	40	0.30	0.28	0.01	-0.06	0.50	0.74	0.94	0.91	0.47	0.67
audiology.standardized	random	2	15	0.32	0.66	0.01	-0.07	0.50	0.77	0.95	0.94	0.54	0.69
breast-cancer	random	3	45	0.27	0.87	0.00	N/A	0.62	0.75	0.97	0.94	0.54	0.67
soybean-large	pca	2	30	0.31	0.58	0.10	N/A	0.39	0.79	0.99	0.98	0.70	0.73
primary-tumor	random	3	50	0.27	0.85	0.05	N/A	0.60	0.74	0.96	0.93	0.50	0.67
house-votes-84	random	3	30	0.23	0.47	0.01	N/A	0.32	0.74	0.98	0.96	0.51	0.71
balance-scale	random	3	15	0.33	0.87	0.02	-0.08	0.87	0.69	0.98	0.82	0.31	0.60
tic-tac-toe	random	3	35	0.32	0.45	0.00	-0.07	0.82	0.68	0.99	0.88	0.23	0.60
car	random	3	15	0.31	0.76	0.00	-0.07	0.79	0.69	1.00	0.88	0.30	0.61

t-SNE algoritmasına bütün veri setleri için bakıldığında RES metriği için minimum 0.60 maksimum 0.77 değerini almıştır.

Tablo 70 Bütün Veri Setlerinin UMAP için sonuçları

Veri Seti	NoN	Dist	NoC	Met	SD&KrM	Ss	Sr	Tpr	Rv	KoM	Tw	Cnt	LCMC	RES
lenses_adjusted	6	0.80	5	euclidean	0.03	0.02	0.14	-0.01	0.02	0.73	1.00	1.00	0.82	0.83
soybean-small	6	0.5	5	cosine	0.19	0.51	0.42	-0.01	0.20	0.78	0.99	0.99	0.76	0.86
promoters_adjusted	20	0.99	20	euclidean	0.15	6.66	0.00	-0.02	0.69	0.76	0.97	0.96	0.63	0.76
hayes-roth	8	0.8	4	canberra	0.21	0.13	0.01	N/A	0.44	0.72	0.95	0.95	0.41	0.76
lymphography	7	0.99	8	correlation	0.20	0.12	0.00	-0.02	0.57	0.74	0.98	0.97	0.64	0.79

audiology.standardized	8	0.99	7	euclidean	0.25	0.12	0.01	-0.03	0.55	0.75	0.97	0.97	0.59	0.79
breast-cancer	7	0.99	7	cosine	0.21	0.06	0.00	N/A	0.61	0.73	0.98	0.97	0.57	0.78
soybean-large	8	0.99	7	cosine	0.23	0.29	0.17	N/A	0.26	0.76	0.99	0.99	0.73	0.85
primary-tumor	5	0.99	8	euclidean	0.23	0.08	0.01	N/A	0.74	0.74	0.98	0.96	0.57	0.76

UMAP algoritmasına bütün veri setleri için bakıldığında RES metriği için minimum 0.76 maksimum 0.86 değerini almıştır.

RES metriği bütün diğer kullanılan metrik ve parametre etkilerini eşit olarak almış olup bütün metriklerin RES üzerine etkisi aynı olacak şekilde üretilmiştir. Metriklerin her biri orijinal veri seti içerisinde bir ilişkinin boyut indirgenmiş veri setinde korunup korunmadığını ölçmektedir. Bu yüzden RES gibi direkt olarak eşit ağırlıklar değil ama veri seti ve istenen yapı özelinde ağırlıklar değiştirilebilir.

8. SONUÇ

Farklı kayıt ve değişken sayısına sahip 13 kategorik veri seti; kPCA, izomap, lle (ve diğer varyansları), t-SNE ve UMAP algoritmaları üzerinde uygulanmıştır. 9 metrik üzerinden farklı parametreler kullanılarak elde edilen değerlendirme ölçüm metriklerinin sonuçları analiz kısmında paylaşılmış ve bütün değerlendirme ölçüm metriklerin mutabık kaldığı bir parametre konfigürasyonu bulunamamıştır. Bunun üzerine yeniden düzenlenmiş ölçeklendirme (refactored scaling, RES) isminde bütün metrikleri içine alan ve kullanılan diğer algoritma parametrelerini maliyet fonksiyonu olarak ekleyen bir metrik yaratılmıştır. Böylece farklı parametrelere göre farklı metrikler için iyi çıkan sonuçlar yerine bütün parametreleri ve bütün metriklerin ölçtüğü bakış açısını koruyan genel bir metrik üretilmiştir. Bu sayede boyut indirgemedeki tek bir ilişki yerine hem lokal hem global ölçekli orijinal veri setinde korunacak doğrusal olmayan ilişkiler dikkate alınarak en iyi sonucu verecek parametreler belirlenen doğrusal olmayan boyut indirgeme algoritmaları için denmiştir.

Bütün veri setleri kategorik veri setleri olduğu için bu tez içerisinde yapılan çalışma -seçilen boyut indirgeme algoritmaları, kullanılan değerlendirme ölçüm teknikleri ve bu tez içinde türetilen yeniden ölçeklendirme metriği (Refactored scaling, RES) – müşteri ilişkileri yönetimi verileri içinde uygulanabileceği gösterilmiştir.

9. KAYNAKLAR

- [1] «Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025,» statista, 8 9 2022. [Çevrimiçi]. Available: <https://www.statista.com/statistics/871513/worldwide-data-created/>. [Erişildi: 31 5 2023].
- [2] «Datasets,» 31 5 2023. [Çevrimiçi]. Available: <https://www.kaggle.com/datasets>. [Erişildi: 31 5 2023].
- [3] «Machine Learning Repository,» 31 5 2023. [Çevrimiçi]. Available: <https://archive.ics.uci.edu/ml/datasets.php>. [Erişildi: 31 5 2023].
- [4] R. Bellman, Dynamic programming, Princeton University Press, 1957.
- [5] «Data set,» Wikipedia, [Çevrimiçi]. Available: https://en.wikipedia.org/wiki/Data_set. [Erişildi: 31 5 2023].
- [6] «Types of variables,» Statistics Canada, [Çevrimiçi]. Available: <https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch8/5214817-eng.htm>. [Erişildi: 31 5 2023].
- [7] «Data pre-processing,» Wikipedia, [Çevrimiçi]. Available: https://en.wikipedia.org/wiki/Data_pre-processing. [Erişildi: 31 5 23023].
- [8] U. M. L. Repository, «Audiology (Original),» 1987. [Çevrimiçi]. Available: <https://doi.org/10.24432/C5M59M..> [Erişildi: 6 9 2023].
- [9] R. Siegler, «Balance Scale,» UCI Machine Learning Repository, 1994. [Çevrimiçi]. Available: <https://doi.org/10.24432/C5488X..> [Erişildi: 6 9 2023].
- [10] M. a. S. Zwitter, «Breast Cancer,» UCI Machine Learning Repository, 1988. [Çevrimiçi]. Available: <https://doi.org/10.24432/C51P4M>. [Erişildi: 6 9 2023].
- [11] M. Bohanec, «Car Evaluation,» UCI Machine Learning Repository, 1997. [Çevrimiçi]. Available: <https://doi.org/10.24432/C5JP48>. [Erişildi: 6 9 2023].
- [12] «Congressional Voting Records,» UCI Machine Learning Repository, 1987. [Çevrimiçi]. Available: <https://doi.org/10.24432/C5C01P>. [Erişildi: 6 9 2023].
- [13] B. Hayes-Roth ve F. Hayes-Roth, «Hayes-Roth,» UCI Machine Learning Repository, 1989. [Çevrimiçi]. Available: <https://doi.org/10.24432/C5501T>. [Erişildi: 6 9 2023].
- [14] J. Cendrowska, «Lenses,» UCI Machine Learning Repository, 1990. [Çevrimiçi]. Available: <https://doi.org/10.24432/C5K88Z>. [Erişildi: 6 9 2023].
- [15] M. Zwitter ve M. Soklic, «Lymphography,» UCI Machine Learning Repository, 1988. [Çevrimiçi]. Available: <https://doi.org/10.24432/C54598>. [Erişildi: 6 9 2023].
- [16] C. Harley, R. Reynolds ve M. Noordewier, «Molecular Biology (Promoter Gene Sequences),» UCI Machine Learning Repository, 1990. [Çevrimiçi]. Available: <https://doi.org/10.24432/C5S01D>. [Erişildi: 6 9 2023].
- [17] M. Zwitter ve M. Soklic, «Primary Tumor,» UCI Machine Learning Repository, 1988. [Çevrimiçi]. Available: <https://doi.org/10.24432/C5WK5Q>. [Erişildi: 6 9 2023].
- [18] R. Michalski ve R. Chilausky, «Soybean (Large),» UCI Machine Learning Repository, 1988. [Çevrimiçi]. Available: <https://doi.org/10.24432/C5JG6Z>. [Erişildi: 6 9 2023].
- [19] R. Michalski, «Soybean (Small),» UCI Machine Learning Repository, 1987. [Çevrimiçi]. Available: <https://doi.org/10.24432/C5DS3P>. [Erişildi: 6 9 2023].

- [20] D. Aha, «Tic-Tac-Toe Endgame,» UCI Machine Learning Repository, 1991. [Çevrimiçi]. Available: <https://doi.org/10.24432/C5688J>. [Erişildi: 6 9 2023].
- [21] «Feature Selection,» Wikipedia, [Çevrimiçi]. Available: https://en.wikipedia.org/wiki/Feature_selection. [Erişildi: 31 5 2023].
- [22] «Feature Extraction,» Wikipedia, [Çevrimiçi]. Available: https://en.wikipedia.org/wiki/Feature_extraction. [Erişildi: 31 5 2023].
- [23] H. Hotelling, « Analysis of a Complex of Statistical Variables into Principal Components,» *The Journal of Educational Psychology*, pp. 417-441, 1993.
- [24] «ieeexplore,» IEEE, [Çevrimiçi]. Available: <https://ieeexplore.ieee.org>. [Erişildi: 15 6 2023].
- [25] T. G. A. I. A. E. H. Alaa Tharwart, «Linear discriminant analysis: A detailed tutorial,» *AI Communicaitons*, pp. 169-190, 1 1 2017.
- [26] «Manifold,» Wikipedia, [Çevrimiçi]. Available: <https://en.wikipedia.org/wiki/Manifold>. [Erişildi: 16 6 2023].
- [27] L. Cayton, «Algorithms for Manifold Learning,» University of California at San Diego, San Diego, 2005.
- [28] P. A. J. Douglas Carroll, «Multidimensional Scaling,» *Measurement, Judgment and Decision Making*, pp. 179-250, 1998.
- [29] B. A. S.-R. M. Schölkopf, «Kernel Principal Component Analysis,» *International conference on artificial neural networks*, pp. 583-588, 1997.
- [30] V. d. S. J. C. L. Joshua B. Tenenbaum, «A global geometric framework for nonlinear dimensionality reduction,» *sciencemag*, pp. 2319-2323, 22 12 2000.
- [31] «Dijkstra's algorithm,» Wikipedia, [Çevrimiçi]. Available: https://en.wikipedia.org/wiki/Dijkstra%27s_algorithm. [Erişildi: 17 6 2023].
- [32] «Google Scholar,» Google, [Çevrimiçi]. Available: <https://t.ly/hdR5Y>. [Erişildi: 17 6 2023].
- [33] S. T. Roweis ve L. K. Saul, «Nonlinear dimensionality reduction by locally linear embedding,» *Science Magazine*, pp. 2323-2325, 22 12 2000.
- [34] «Google Scholar,» Google, [Çevrimiçi]. Available: <https://t.ly/dAaF>. [Erişildi: 17 6 2023].
- [35] L. v. d. Maaten ve G. Hinton, «Visualizing Data using t-SNE,» *Journal of Machine Learning Research*, pp. 2579-2605, 2008.
- [36] «Kullback-Leibler Divergence,» Wikipedia, [Çevrimiçi]. Available: https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence. [Erişildi: 16 6 2023].
- [37] «Google Scholar,» Google, [Çevrimiçi]. Available: <https://t.ly/Ka9S>. [Erişildi: 16 6 2023].
- [38] J. H. J. M. Leland McInnes, «UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,» 21 9 2020.
- [39] R. N. Shepard, «The analysis of proximities: multidimensional scaling with an unknown distance function,» *Psychometrika*, 1962.
- [40] J. B. Kruskal, «Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis,» *Psychometrika*, 1964.
- [41] J. B. Kruskal, «Nonmetric multidimensional scaling: a numerical method,» *Psychometrika*, 1964.

- [42] J. W. Sammon, «A nonlinear mapping for data structure analysis,» *IEEE Transactions on computers*, 1969.
- [43] S. Sidney, «Nonparametric statistics for the behavioral sciences,» *The Journal of Nervous and Mental Disease*, 1957.
- [44] H. U. & P. K. R. Bauer, « Quantifying the neighborhood preservation of self-organizing feature maps,» *IEEE Transactions on neural networks*, 1992.
- [45] A. Konig, «Interactive visualization and analysis of hierarchical neural projections for data mining,» *IEEE transactions on neural networks*, 2000.
- [46] J. K. S. Venna, «Local multidimensional scaling,» *Neural Networks*, 2006.
- [47] L. B. A. Chen, «Local multidimensional scaling for nonlinear dimension reduction, graph layout and proximity analysis,» University of Pennsylvania, 2006.
- [48] G. H. Lincoff, «Mushroom records drawn from The Audubon Society Field Guide to North American Mushrooms,» New York, 1981.
- [49] J. Schlimmer, «Congressional Quarterly Almanac, 98th Congress, 2nd session 1984,» Volume XL: Congressional Quarterly Inc. , Washington, D.C., 1985.
- [50] A. Shapiro, «Chess (King-Rook vs. King-Pawn),» UCI Machine Learning Repository, 1989. [Çevrimiçi]. Available: <https://doi.org/10.24432/C5DK5C>. [Erişildi: 6 9 2023].
- [51] «Molecular Biology (Splice-junction Gene Sequences),» UCI Machine Learning Repository, 1992. [Çevrimiçi]. Available: <https://doi.org/10.24432/C5M888>. [Erişildi: 6 9 2023].