Int. J. Adv. Eng. Pure Sci. 2025, 37(4): <422-429>
DOI: 10.7240/jeps.1785840

RESEARCH ARTICLE / ARAŞTIRMA MAKALESİ

Özellik Seçimi Temelli Yaklaşımların Çok Değişkenli Regresyon Modellerine Etkisi

The Effect of Feature Selection Based Approaches on Multivariate Regression Models

Ramazan AKMAN [1], **Nihat TAK** [1]

[1]*Marmara Üniversitesi, Fen Fakültesi, İstatistik Bölümü, 34700, İstanbul, Türkiye*

Öz

Bu çalışmada yüksek boyutlu veri setlerinde boyut indirgemeyi ve indirgenen modellerin tahmin performansını artırmayı hedefleyen K-Ortalamalar kümeleme temelli bir özellik seçimi yöntemi önerilmektedir. Önerilen yöntemde her bir bağımsız değişken özellik olarak tanımlanmaktadır. Tanımlanan bu özellikler K-Ortalamalar kümeleme algoritmasıyla kümelenir, her kümeden kümeyi temsil düzeyi en yüksek olan özellik seçilerek hafızaya alınır. Sonraki adımda hafızaya alınan yani kümeleri temsil eden bu özellikler ile çok değişkenli doğrusal regresyon, Ridge regresyon ve LASSO regresyon yöntemleri kullanılarak regresyon modelleri oluşturulur. Gerçekleştirilen boyut indirgeme işlemi çoklu bağlantı sorununu azaltmaktadır. Ayrıca önerilen indirgenmiş çok değişkenli doğrusal regresyon modeli, indirgenmiş Ridge regresyon modeli ve indirgenmiş LASSO regresyon modeli, çok değişkenli regresyon yöntemiyle karşılaştırılmıştır. c Elde edilen bulgular, önerilen boyut indirgeme modellerinin yüksek boyutlu veri ortamlarında hem etkinlik hem de verimlilik açısından kayda değer performans sergilediğini kanıtlamaktadır.

Anahtar Kelimeler: K-Ortalamalar kümeleme, Özellik seçimi, Makine Öğrenmesi, Çok değişkenli regresyon

Abstract

This study proposes a K-Means Cluster based feature selection method that aims to reduce the dimensionality of high-dimensional data sets and improve the prediction performance of the reduced models. In the proposed method, each independent variable is defined as a feature. These defined features are clustered using the K-Means algorithm, and the feature with the highest cluster representation level is selected from each cluster and stored in memory. In the next step, regression models are created using multivariate linear regression, Ridge regression, and LASSO regression methods with these features stored in memory, which represent the clusters. The dimension reduction process reduces the multicollinearity problem. Additionally, the proposed reduced multivariate linear regression model, reduced Ridge regression model, and reduced LASSO regression model were compared with the multivariate regression method. In comparison based on actual data, the reduced models showed an improvement of 10% to 38% over the unreduced model according to the OMYH criterion and an improvement of 8% to 50% according to the HKOK criterion. The findings demonstrate that the proposed dimension reduction models exhibit remarkable performance in terms of both effectiveness and efficiency in high-dimensional data environments.

Keywords: K-Means clustering, Feature selection, Dimension reduction, Machine learning, Multivarite regression

I. GİRİŞ

Günümüzde sağlık hizmetlerinden eğitim sistemlerine, finansal analizlerden perakende ve üretim süreçlerine kadar geniş bir yelpazede yüksek boyutlu veri setleri karşımıza çıkmaktadır. Ortaya çıkan bu veri setleri yüzlerce, hatta binlerce bağımsız değişkenin bir arada değerlendirilmesi sonucunda hem çoklu bağlantı (multicollinearity) sorununa hem de modelin genel performansını olumsuz etkileyen birçok problemin oluşmasına zemin hazırlar. Çoklu bağlantı belirtilerinden biri olan değişkenler arasındaki yüksek korelasyon, model katsayılarının belirsizleşmesine, modelin tahmin değerlerinin yanlış olmasına ve sonuçların güvenilirliğinin azalmasına yol açar [1]. Bu koşullar altında, gereksiz veya bilgi katkısı düşük değişkenlerin elenmesi hem modelin yorumlanabilirliğini artırmak hem de aşırı öğrenme (overfitting) riskini azaltmak için kaçınılmaz bir gereklilik haline gelir [2]. Özellik seçimi, tam da bu noktada, modele dahil edilecek değişken alt kümesinin dikkatli bir biçimde belirlenmesini sağlayarak veri boyutunun indirgenmesine, model eğitim sürecindeki hesaplama maliyetinin düşürülmesine ve öngörü doğruluğunun istatistiksel anlamlılık düzeylerinde iyileştirilmesine olanak sağlar [3, 4].

Sorumlu Yazar: RAMAZAN AKMAN, **Tel:** 0531 503 24 99, **E-posta:** ramazan.akman@marmara.edu.tr

Gönderilme: 17.09.2025, **Düzenleme:** 12.11.2025, **Kabul:** 29.11.2025

Bu sayede, büyük ölçekli veri ortamlarında hem model verimliliği hem de genelleme yeteneği kayda değer ölçüde artırılmış olur. Özellik seçimi, çeşitli alanlardaki araştırmacılar arasında yaygın olarak tercih edilen bir yöntem haline gelmiştir. Saeys, ve ark. [5], Chandrashekar ve Sahin [6], Li ve ark. [7] J., Khaire ve Dhanalakshmi [8], Yang ve ark. [9] ve Guyon ve Elisseeff [2] tarafından özellik seçim yöntemleri üç ana başlık altında toplanmıştır. Birinci ana başlık, filtre yöntemleri olup, değişkenlerin korelasyon katsayısı, ki-kare testi, bilgi kazancı gibi bağımsız istatistiki ölçütler temelinde sıralanarak en yüksek puanlıların seçilmesini öngörür [6]. İkinci ana başlık olan sarmalayıcı (wrapper) yöntemleri; belirli bir öğrenme algoritmasını dış döngü şeklinde kullanarak farklı alt kümelerin performansını karşılaştırır ve en uygun alt küme yapısını ileri-geri seçim, boşluk arama vb. iteratif olarak belirler [10]. Üçüncü ana başlık ise gömülü (embedded) yöntemlerdir; bu yöntemler, özellik seçme işlemini modelin öğrenme mekanizmasının bir parçası haline getirerek hem tahmin doğruluğunu hem de modelin genelleme yeteneğini yükseltir. Özellikle Tibshirani [11] tarafından önerilen En Küçük Mutlak Küçülme ve Seçim Operatörü (Least Absolute Shrinkage and Selection Operator, LASSO) yöntemi, regresyon katsayılarına L1 ceza terimi uygulayarak gereksiz değişkenlerin ağırlık matrisinden uzaklaştırılmasını sağlar. Zou ve Hastie [12] tarafından tanımlanan Elastic Net ise L1 ve L2 normlarının birlikte kullanıldığı ceza terimiyle hem değişken seçimini hem de model kararlılığını eş zamanlı olarak gerçekleştirir. Bu yaklaşımlar, ceza terimi kullanan doğrusal modellerde gereksiz değişkenleri ağırlık matrisinden uzaklaştırmaktadır [13]. Ayrıca, Breiman [14] tarafından geliştirilen Rasgele Orman (Random Forest) yöntemi, topluluk öğrenmesi kapsamında birden çok karar ağacı kullanarak değişken önemine dayalı seçim mekanizmalarıyla öne çıkan özellikleri belirlemeyi mümkün kılar. Friedman [15] ise Gradyan Güçlendirme Makinesi (Gradient Boosting Machine) yaklaşımıyla, zayıf öğrenicileri ardışık olarak birleştirerek her adımda değişken ağırlıklarını güncellemek suretiyle güçlü tahmin modelleri oluşturur. Böylece gömülü yöntemler, özellik seçimini ön işleme adımından çıkartıp modelin parametre optimizasyonu süreciyle bütünleştirerek hem hesaplama verimliliği sağlar hem de aşırı öğrenme riskini azaltır.

Çoğu geleneksel filtre yöntemi her bir özelliği bağımsız ele almakta ve bu nedenle karmaşık etkileşimleri göz ardı ederek çoklu bağlantı sorununu tam olarak gideremektedir [16]. Minimum Artıklık Maksimum İlgililik (Minimum Redundancy Maximum Relevance, mRMR), özellikle yüksek boyutlu veri setlerinde kullanılan ve her bir özelliğin hedef değişkenle olan ilişkisini Karşılıklı Bilgi (Mutual Information, MI) ölçütüyle değerlendirerek maksimum alaka ve minimum fazlalık ilkelerine dayanan bir filtre tabanlı özellik seçimi yöntemidir. Sınıflayıcıdan bağımsız olarak çalışması ve hesaplama açısından görece verimli olması nedeniyle yaygın olarak tercih edilmektedir. Ancak mRMR, yalnızca birinci dereceden ilişkileri dikkate alması sebebiyle değişkenler arası karmaşık bağıntıları ve çoklu bağlantı (multicollinearity) gibi yapısal sorunları göz ardı edebilir; ayrıca MI hesaplamalarının sürekli değişkenler üzerinde yüksek hesaplama maliyeti doğurabileceği bilinmektedir [17]. Bilgi Kazancı (Information Gain, IG) yöntemi, her bir özelliği hedef değişkenle olan bağımsız ilişkisine göre değerlendirerek sıralayan bir diğer filtre yaklaşımıdır [18]. Bu yöntemin temel sınırlılıklarından biri, yalnızca bireysel bilgi katkısını dikkate alması nedeniyle özellikler arası korelasyonu göz ardı etmesidir. Bu durum, yüksek derecede ilişkili (çoklu bağlantılı) özelliklerin birlikte seçilmesine ve böylece modele fazladan ve yinelenen bilgi taşınmasına neden olabilir. Dolayısıyla IG, çoklu bağlantı problemini doğrudan ele almadığı için yüksek boyutlu ve korelasyonlu veri yapılarında etkin bir özellik seçimi sağlamaktan uzaktır [5]. Korelasyon tabanlı Özellik Seçimi (Correlation-based Feature Selection, CFS) gruplar arası ilişkileri dikkate alsa da çoklu bağlantı sorununu tek başına gideremektedir [19].

Bu çalışma özellik seçimini veri setindeki değişkenler arasındaki benzerlik ilişkilerine dayanarak gerçekleştiren bir filtreleme yaklaşımıdır. Tüm bağımsız değişkenler ‘özellik’ olarak tanımlanmış ve K-Ortalamalar kümeleme [20] algoritmasıyla birbirine benzer davranış sergileyen özellikler kümelenmiştir. Her kümeden küme merkezine en yakın konumdaki değişken seçilerek elde edilen indirgenmiş özellik kümeleri, Çok Değişkenli Doğrusal Regresyon (ÇDDR) [21], Ridge regresyon [22] ve LASSO regresyon [11] modellerinin eğitiminde kullanılmıştır. Böylece boyut indirgeme adımının yalnızca veri yapısına dayalı olarak uygulandığı ve modele ek bir cezalandırma mekanizması eklemediği açıkça ortaya konmuştur. Buna ek olarak, çalışmada kullanılan verilerdeki çoklu doğrusal bağlantı düzeyindeki azalmayı ortaya koymak amacıyla, özellik seçimi adımı uygulanmadan önce ve uygulandıktan sonra koşul indeksi analizi gerçekleştirilmiştir [28, 29].

Çalışmada kullanılan veri setlerinden ilki, R programının ‘base’ paketinde bulunan, otomobil özelliklerini içeren ‘mtcars’ veri setidir. Bu veri seti 32 gözlem ve 11 bağımsız değişkenden oluşmaktadır ve analizlerde bağımlı değişken olarak ‘mpg’ değişkeni kullanılmaktadır. İkinci veri seti, R programının ‘rattle’ paketinde yer alan ‘Wine’ veri setidir [30]. Bu veri

setinde 178 gözlem ve 14 değişken yer almakta ve 'proline düzeyi' bağımlı değişken olarak ele alınmaktadır. Üçüncü olarak kullanılan veri seti 'fat' Faraway [23] veri setidir. Bu veri seti 252 bireye ait 18 antropometrik ölçüm içermekte olup özellikle bilek çevresi ölçümü dikkate alınmaktadır. Dördüncü ve beşinci veri setleri ise kontrollü koşullarda oluşturulan simülasyon verilerinden oluşmaktadır. Bu veri setleri, önerilen yöntemin hem gerçek hem de sentetik veriler üzerindeki performansını kapsamlı bir şekilde değerlendirebilmek amacıyla seçilmiştir. Performans ölçütleri olarak Hata Kareler Ortalamasının Karekökü (HKOK) [24, 25] ve Ortalama Mutlak Yüzde Hata (OMYH) [26] kriterleri belirlenmiş olup bu sayede modellerin hem mutlak hem de göreceli hata düzeyleri objektif biçimde karşılaştırılmıştır. Bu çalışmanın ikinci bölümünde, önerilen yöntemin algoritması detaylı şekilde sunulmuştur. Üçüncü bölümde, yöntemin farklı gerçek veri setleri ve simülasyon verileri üzerindeki performansı değerlendirilmiş ve çoklu bağlantı problemine ilişkin koşul indeksi analizi sonuçları yorumlanmıştır. Son olarak, dördüncü bölümde elde edilen bulgular tartışılmış ve genel sonuçlara yer verilmiştir.

II. YÖNTEM

Adım 1. Parametrelerin Belirlenmesi

c: Küme sayısı

Test_Size: Test seti oranı

Validation_Size: Doğrulama seti oranı

Train_Size: Test seti oranı ve doğrulama oranı girildiğinde otomatik olarak girilmiş olacaktır. (1- Test_Size + Validation_Size)

Adım 2. Verinin Bölünmesi

Veri seti, girilen parametre oranlarına göre rastgele olarak eğitim (train), doğrulama (validation) ve test alt kümelerine ayrılır.

Veri Matrisi:

$X = [x_{ij}], i = 1, 2, \dots, p; j = 1, 2, \dots, n$

Burada;

X: Tüm veri matrisi

p: Özellik (değişken) sayısı

n: Toplam gözlem sayısı

şeklinde tanımlanır.

Eğitim seti (Training Set):

$X_{eğitim} = [x_{ij}], i = 1, 2, \dots, p;\ j = 1, 2, \dots, n_{eğitim}$

Doğrulama seti (Validation Set):

$X_{doğrulama} = [x_{ij}], i = 1, 2, \dots, p;$

$j = 1, 2, \dots, n_{doğrulama}$

Test seti (Test Set):

$X_{test} = [x_{ij}], i = 1, 2, \dots, p; j = 1, 2, \dots, n_{test}$

Adım 3. Özellik seçimi için K-Ortalamalar kümeleme yöntemi

Adım 3.1. Tüm bağımsız değişkenler transpoze edilir.

Adım 3.2. Başlangıçta küme merkezleri rasgele olarak belirlenir.

$v_k^{(0)}, k = 1, 2, \dots, c$

Burada c küme sayısını temsil etmektedir.

Adım 3.3. Nesnelerin kümelere atanması:

- Bağımsız değişkenler ile küme merkezleri arasındaki Öklid uzaklıkları hesaplanır (alternatif olarak Manhattan uzaklığı veya Minkowski uzaklığı da kullanılabilir). Hesaplama için kullanılan Öklid uzaklığı formülü;

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad \textbf{(1)}$$

şeklindedir.

- Her nesne, uzaklığı en küçük olan kümeye atanır.

Adım 3.4. Küme merkezleri güncellenir:

Her küme için, o kümeye atanan tüm veri noktalarının ortalaması hesaplanır. Yeni ortalamalar, yeni küme merkezleri olur.

$$c_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_i \quad \textbf{(2)}$$

Burada;

c_k: k. kümenin merkezi

n_k: k. kümedeki veri noktalarının sayısı

x_i: k. kümeye atanan i. veri noktası

olarak tanımlanır.

Adım 4. Her küme için en iyi temsilci özellik belirlenir.

Önceki aşamada oluşturulan özellik kümeleri içinde, her bir değişkenin ait olduğu küme merkezi ile Öklid uzaklığı hesaplanır. Her küme için bu mesafe ölçümlerinden en düşük değeri veren değişken, ilgili kümenin en iyi temsilcisi olarak seçilir.

Her küme k için o kümedeki özellikler arasından

$$j_k = \arg\min_{1\le j\le p} d_{kj} \quad \textbf{(3)}$$

ifadesiyle en küçük d_{kj} değerini veren j_k indeksi bulunur. Buradaki d_{kj} k. küme merkezi ile j. özelliğin (değişkenin) vektörü arasındaki Öklid uzaklığını, j_k ise k. küme için "küme merkezine en küçük uzaklığa sahip" değişkenin indeksini ifade eder.

Adım 5. Modellerin kurulumu:

Seçilen değişkenlerle Çok Değişkenli Doğrusal Regresyon (ÇDDR), İndirgenmiş Çok Değişkenli Doğrusal Regresyon (İÇDDR), İndirgenmiş Ridge Regresyon (İRR) ve İndirgenmiş LASSO Regresyon (İLR) modelleri kurulur.

Doğrulama/Test seti kullanılarak her bir model için tahmin yapılır.

Doğrusal regresyon amaç fonksiyonu:

$$\min_{\beta} \left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2\right\} \quad \textbf{(4)}$$

şeklindedir.

Ridge regresyon amaç fonksiyonu:

$$\min_{\beta} \left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\} \quad \textbf{(5)}$$

şeklindedir.

LASSO Regresyon amaç fonksiyonu:

$$\min_{\beta} \left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \quad \textbf{(6)}$$

şeklindedir. Buradaki;

β_j: Katsayılar

λ: Ceza (Penalty) parametresi

olarak tanımlanır.

Adım 6. Performans Değerlendirmesi

Her model için denklem (7) ve (8) kullanılarak performans değerleri hesaplanır.

$$HKOK = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \textbf{(7)}$$

$$OMYH = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad \textbf{(8)}$$

Burada;

y_i: Gerçek gözlem değeri

$\hat{y}_i$: Modelin tahmin değeri

n: Gözlem sayısı

olarak tanımlanır [27].

III. DEĞERLENDİRME

Tablo 1, K-Ortalamalar kümeleme tabanlı özellik seçimi yönteminin kullanıldığı beş ayrı veri setine ilişkin özellik sayısı (p), bağımlı değişken, gözlem sayısı (n), test seti uzunluğu ve ele alınacak küme sayılarını ayrıntılı biçimde sunmaktadır. İlk veri seti olarak değerlendirilen 'mtcars' (n = 32, p = 11) üzerinde, motorlu taşıtların mil başına galon (mpg) değeri tahmin edilmek üzere küme sayısı (c) 2 ile 5 aralığında ızgara arama metoduyla optimize edilmiş; veri setinin %20'sini oluşturan yedi gözlem, model performansının değerlendirilmesi için test kümesi olarak ayrılmıştır. İkinci veri seti olarak ele alınan 'Wine' veri setinde (n = 178, p = 14), şaraplardaki 'proline düzeyi' bağımlı değişken olarak belirlenmiş; c, 2 ile 7 aralığında ızgara arama yöntemi ile optimize edilmiş ve modelin performansını ölçmek üzere 18 gözlem test kümesi olarak ayrılmıştır. Üçüncü veri seti, R'nin 'faraway' kütüphanesinden 'fat' (n = 252, p = 18) veri seti olup bilek ölçümü (wrist) tahmini için c, 2 ile 9 aralığında ızgara arama yöntemi ile optimize edilmiştir; bu aşamada 25 gözlem (%10) test kümesi olarak kullanılmıştır. Dördüncü ve beşinci veri setleri ise, kendi içerisinde ilişkili değişkenlerin senaryolarını modellemek amacıyla yapay simülasyonlarla üretilmiştir. İlk veri seti, n = 300 gözlem ve p = 300 özellik, ikinci veri seti ise n = 500 gözlem ve p = 400 özelliğe sahip veri setleridir. İlk simülasyon senaryosunda küme sayısı (2, 150), ikinci senaryoda ise (2, 200) aralığında ızgara arama yöntemiyle optimize edilmiş; her yinelemede ilgili veri setinin %10'u (ilk senaryoda 30, ikinci senaryoda 50 gözlem) test verisi olarak ayrılarak yöntemlerin genelleme yeteneği karşılaştırılabilir hale getirilmiştir. Böylece, orta boyutlu gerçek veri örneklerinden yüksek boyutlu simülasyon senaryolarına kadar, özellik sayısının ve kümeleme parametresinin model performansına etkisi hem OMYH hem de HKOK gibi objektif hata ölçütleriyle sistematik olarak değerlendirilmiştir.

Tablo 1. Beş Veri Setinde Uygulanan K-Ortalamalar Tabanlı Özellik Seçimi Sonuçları

Veri Seti	Özellik Sayısı	Bağımlı Değişken	c Aralığı	Gözlem Sayısı	Test Seti Uzunluğu
mtcars	11	mpg	(2, 5)	32	7
Wine	14	Propline	(2, 7)	178	18
fat	18	wrist	(2, 9)	252	25
Simülasyon Verisi 1	300	y	(2, 150)	300	30
Simülasyon Verisi 2	400	y	(2, 200)	500	30

Tablo 2. Yöntemlerin Ortalama Mutlak Yüzde Hata (OMYH) Sonuçları

Veri Seti	Optimal c	ÇDDR	İÇDDR	İRR	İLR
mtcars	5	0.2191	0.1499	**0.0819**	0.0936
Wine	7	0.2550	**0.2415**	0.2688	0.2661
fat	6	0.0278	0.0206	**0.0194**	0.0203
Simülasyon Verisi 1	148	1.0658	0.4787	0.4898	**0.4030**
Simülasyon Verisi 2	175	0.9929	1.1144	**0.9571**	11.5301

Tablo 3. Yöntemlerin Hata Kareler Ortalamasının Karekökü (HKOK) Sonuçları

Veri Seti	Optimal c	ÇDDR	İÇDDR	İRR	İLR
mtcars	5	6.5521	2.8834	**1.7131**	2.0095
Wine	7	216.2393	**198.7030**	218.1752	214.7946
fat	6	0.6715	0.4646	**0.4414**	0.4517
Simülasyon Verisi 1	148	371.5158	1.8178	1.8289	**1.4089**
Simülasyon Verisi 2	177	87.6107	1.4231	1.4472	**1.2993**

Tablo 4. Koşul İndeksi Analizi Sonuçları

	mtcars	Wine	fat	Simülasyon Verisi 1	Simülasyon Verisi 2
Önce	15.5583	9.9640	208.6360	2123.5120	28.2518
Sonra	6.6618	2.3671	8.8049	6.1636	4.5655

Tablo 2'de beş farklı veri seti için K-Ortalamalar kümeleme tabanlı özellik seçimi ve ızgara arama (Grid Search) yöntemiyle elde edilen optimum küme sayısı (c) ve bu c değeri altında Çok Değişkenli Doğrusal Regresyon (ÇDDR), İndirgenmiş Çok Değişkenli Doğrusal Regresyon (İÇDDR), İndirgenmiş Ridge Regresyon (İRR) ve İndirgenmiş LASSO Regresyon (İLR) yöntemlerine ilişkin OMYH sonuçları özetlenmektedir. Tablo 2'de, her bir veri setindeki en düşük OMYH sonucunu veren yöntem kalın ve siyah bir biçimde vurgulanmıştır. 'mtcars' veri setinin c = 5 için, ÇDDR modelin 0.2191'lik OMYH'sini İÇDDR modelde 0.1499'a (%31.6 iyileşme) düşürmüştür. Bu süreçte K-Ortalamalar kümeleme tabanlı İÇDDR, değişken boyutunu azaltarak hem aşırı uyum riskini düşürmüş hem de hata oranını anlamlı biçimde iyileştirmiştir. Ardından İRR (0.0819) ve İLR (0.0936) yöntemleri, seçilen beş değişken üzerindeki katsayıları ceza terimleriyle yeniden düzenleyerek ek hata düşüşü sağlamış; özellikle İRR %62.6'ya varan iyileştirmesiyle, küçük örneklem ve sınırlı değişken yapısının etkilerine karşı dayanıklılık verdiğini göstermiştir. 'Wine' veri seti üzerinde c = 7 ile indirgeme uygulandığında, ÇDDR modelinin 0.2550'lik OMYH değeri İÇDDR'de 0.2415'e düşmüştür; bu da boyut indirgemenin orta ölçekli veri setlerinde dahi tutarlı fayda sunduğunu ortaya koymuştur. Ancak İRR (0.2688) ve İLR (0.2661) düzenlemelerinde, test örneklem büyüklüğü ve model karmaşıklığı arasındaki denge gereği, ceza terimlerinin

bazen İÇDDR'den daha yüksek hata üretmesi dikkat çekmektedir.

'fat' veri seti üzerinde yapılan analizde, c = 6 için, İÇDDR model, ÇDDR'nin 0.0278'lik OMYH'sini 0.0206'ya (%25 iyileşme) düşürmüştür. Burada İRR (0.0194) ve İLR (0.0203) modelleri, aralarındaki çok küçük performans farkıyla, orta boyutlu çoklu regresyonda K-Ortalamalar kümeleme ve uygun değişken seçiminin güçlü bir temel sunduğunu teyit etmiştir.

Simülasyon Verisi 1'de c = 148 değeri için İLR, ÇDDR'nin 1.0658'lik OMYH'sini 0.4030'a (%62.2'lik iyileştirme) düşürmüştür. Buna karşın özellik seçimi sonrasında kurulan cezasız İÇDDR (0.4787) ve ceza terimi içeren İRR (0.4898) yüksek boyutlu uzayın karmaşık gürültü yapısına karşı değişken setini İLR kadar kararlı biçimde tutamamış ve hata oranını artırmıştır.

Simülasyon Verisi 2'de ise c = 148 için İRR model, ÇDDR modelde 0.9929 olan OMYH'yi 0.9571'e (%3.61 iyileşme) düşürmüştür. İÇDDR (1.1144) ve İLR (11.5301) modeller ise yüksek boyutlu gürültüyü tamamen sıfırlayamamış, İRR yaklaşımının sağladığı temel başarıyı aşamamıştır.

Tablo 3'te yer alan sonuçlar incelendiğinde, beş farklı veri seti için küme sayısı göz önünde bulundurularak yöntemlerin HKOK sonuçları verilmiştir. Tablo 3'te, her bir veri seti için En düşük HKOK sonucunu veren yöntem kalın ve siyah bir biçimde vurgulanmıştır.

'mtcars' veri seti üzerinde gerçekleştirilen analizde, ÇDDR'nin HKOK'si 6.5521 olarak bulunurken, yalnızca beş özelliğin seçildiği İÇDDR modelinde bu değer 2.8834'e düşürerek hatada %56 oranında anlamlı bir iyileşme sağlamıştır. Ardından İRR, HKOK'yi 1.7131'e, İLR ise 2.0095'e indirgemiştir ki bu durum, küçük örneklem boyutuna sahip ve orta büyüklükte değişken seti içeren yapılarda önce boyut indirgeme yapılmasının, sonrasında ceza terimli yöntemlerle genelleme performansını daha da artırmanın etkili bir strateji olduğunu göstermektedir.

'Wine' veri seti bağlamında, ÇDDR'nin HKOK'si 216.2393 iken İÇDDR modeli 198.7030'a gerilemiş ve hata yaklaşık %8.2 oranında azalmıştır. İRR uygulandığında HKOK 218.1752'ye yükselmiş; İLR'de ise 214.7946 değerine ulaşmıştır. Bu sonuçlar, orta boyutlu değişken kümeleri söz konusu olduğunda K-Ortalamalar kümeleme tabanlı özellik seçiminin tutarlı bir avantaj sağladığını, ancak ceza terimli yöntemlerin ek faydasının örneklem ve değişken dengesiyle sınırlı kalabileceğini ortaya koymaktadır.

'fat' veri seti üzerinde ÇDDR HKOK'si 0.6715 iken, İÇDDR'de 0.4646 düzeyine; İRR'de 0.4414, İLR'de ise 0.4517 seviyesine inmiştir. Özellikle ÇDDR ile İRR modeli arasında elde edilen yaklaşık %34.3'lük düşüş, orta boyutlu çoklu regresyon problemlerinde ÇDDR tabanlı boyut indirgeme ve hemen ardından kontrollü düzenleme stratejisinin birlikte güçlü bir temel oluşturduğunu doğrulamaktadır.

Simülasyon Verisi 1'de, ÇDDR HKOK'si 371.5158 iken İÇDDR'de 1.8178'e (%99.5 azalma) gerilemiş; İRR ve İLR modelleri ise sırasıyla 1.8289, 1.4089 değerleriyle benzer ya da daha düşük hatalar üretmiştir. Özellikle İLR, İÇDDR modeline kıyasla %22.5'lik bir azalma sağlayarak, p = n koşullarında öncelikle özellik sayısının azaltılmasının zorunlu olduğunu, ardından ise düzenleme yöntemlerinin modelin genelleme yeteneğini güçlendirdiğini göstermektedir.

Simülasyon Verisi 2 için ÇDDR'nin HKOK'si 87.6107; İÇDDR'nin 1.4231; İRR'nin 1.4472; İLR'nin ise 1.2993 olarak bulunmuştur. ÇDDR ile İLR modelleri arasındaki elde edilen %98'lik hata düşüşü ve ceza terimli yöntemlerle sağlanan ek iyileştirmeler, aşırı gürültülü ve yüksek boyutlu veri setlerinde boyut indirgeme ve ardından düzenleme yaklaşımlarının model performansını dengeli ve anlamlı biçimde geliştirdiğini ortaya koymaktadır.

Tablo 4, bu çalışmada ele alınan veri setleri için, özellik seçimi uygulanmadan önceki tam model ile K-Ortalamalar kümeleme tabanlı indirgeme uygulanmış modelin koşul indeksi değerlerini karşılaştırmaktadır. Koşul indeksi, tasarım matrisinin sayısal olarak ne ölçüde iyi koşullandığını özetleyen bir ölçüt olup, genellikle 10–30 arası değerler zayıf, 30–100 arası değerler orta–güçlü düzeyde, 100'ün üzerindeki değerler ise aşırı düzeyde çoklu doğrusal bağlantı problemine işaret etmektedir.

Tablo 4'e göre, K-Ortalamalar özellik seçimi yapılmadan önce, özellikle 'fat' veri setinde (208.64) ve Simülasyon Verisi 1'de (2123.51) koşul indeksleri son derece yüksektir ve ilgili verilerde aşırı düzeyde çoklu doğrusal bağlantı problemi bulunduğunu göstermektedir. Simülasyon Verisi 2'de 28.25'lik koşul indeksi, eşik değerlere oldukça yakın olup belirgin bir çoklu doğrusal bağlantı problemi varlığına işaret ederken, 'mtcars' (15.56) ve 'Wine' (9.96) veri setlerinde ise daha düşük düzeyde bir çoklu doğrusal bağlantı problemi gözlenmektedir.

Özellik seçimi sonrasında tüm veri setlerinde koşul indeksinin belirgin biçimde azaldığı görülmektedir: 'mtcars' ve 'Wine' veri setleri için koşul indeksi sırasıyla 6.66 ve 2.37 seviyelerine gerileyerek yaklaşık

%57 ve %76 oranında düşmüş; 'fat' veri setinde 208.64'ten 8.80'e (%95.8), Simülasyon Verisi 1'de ise 2123.51'den 6.16'ya (%99.7) inerek başlangıçtaki aşırı derecede olan çoklu doğrusal bağlantılı yapıyı neredeyse tamamen ortadan kaldırmıştır. Simülasyon Verisi 2'deki koşul indeksinin 28.25'ten 4.57'ye düşmesi de yaklaşık %83.8'lik bir azalmaya karşılık gelmekte ve tasarım matrisinin belirgin biçimde daha iyi koşullandığını göstermektedir.

Bu bulgular, hem orta boyutlu gerçek veri örneklerinde hem de büyük boyut yapısına sahip simülasyon senaryolarında, kümeleme tabanlı özellik seçiminin çoklu doğrusal bağlantıyı önemli ölçüde azalttığını ve regresyon modelleri için daha istikrarlı bir tasarım matrisi sağladığını ortaya koymaktadır.

IV. TARTIŞMA ve SONUÇ

Bu çalışmada, küçük örneklem veri setlerinden büyük karmaşık veri setlerine kadar model karmaşıklığını azaltmak ve tahmin performansını artırmak amacıyla K-Ortalamalar kümeleme tabanlı bir özellik seçimi yöntemi önerilmiş, seçilen değişkenler Çok Değişkenli Doğrusal regresyon, Ridge regresyon ve LASSO regresyon üzerinde test edilmiştir. Elde edilen sonuçlar, özellik seçimi uygulanmış modellerin özellikle HKOK ve OMYH performans ölçütü açısından tüm değişkenlerin kullanıldığı modellere kıyasla daha başarılı performans sergilediğini göstermektedir.

K-Ortalamalar kümeleme temelli seçim, her kümeden yalnızca temsil gücü en yüksek özelliğin modele dahil edilmesi sayesinde gereksiz ve tekrarlı bilgilerin etkisini azaltmıştır. Bu durum, Ridge ve LASSO Regresyon modellerinin cezalandırma tekniği ile birleştiğinde, model katsayılarının kararlılığını artırmış ve aşırı uyum (overfitting) riskini önemli ölçüde düşürmüştür. Özellikle LASSO regresyonunun, seçilen özellikler üzerinden ek bir değişken eliminasyonu sağlaması, modelin sadeliğini ve yorumlanabilirliğini daha da artırmıştır. Sonuç olarak, önerilen yaklaşım; yüksek boyutlu verilerde hem boyut indirgeme hem de modelleme performansını iyileştirme açısından etkili ve uygulanabilir bir çözüm sunmaktadır. Bu çalışmada geliştirilen K-Ortalamalar kümeleme özellik seçimi yöntemi, farklı metodolojik veya uygulamalı araştırma alanlarına genişletilebilecek potansiyele sahiptir. Gelecek çalışmalarda özellik seçim sürecinin, Bulanık C-Ortalamalar kümeleme veya hiyerarşik kümeleme gibi farklı kümeleme algoritmaları ile çeşitlendirilmesi ve metin, görüntü, zaman serisi vb. farklı veri tipleri üzerinde testler yapılması, yöntemin genellenebilirliğini ve etkinliğini artırabilir. Ayrıca, seçilen özelliklerin anlamlılığına ilişkin derinlemesine analizlerin yapılması, yöntemin yorumlanabilirliğini güçlendirecektir.

TEŞEKKÜR

Bu çalışma, Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK)
tarafından 123F266 numaralı proje ile desteklenmiştir. Projeye verdiği destekten ötürü TÜBİTAK'a teşekkürlerimizi sunarız.

KAYNAKÇA

[1] Farrar, D.E. ve Glauber, R.R. (1967). Multicollinearity in regression analysis: the problem revisited. The review of economic and statistics, 92-107.

[2] Guyon, I. ve Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

[3] Dash, M. ve Liu, H. (1997). Feature selection for classification. Intelligent data analysis, 1(1-4), 131-156.

[4] Liu, H. ve Motoda, H. (2012). Feature selection for knowledge discovery and data mining (Vol. 454). Springer science ve business media.

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

[6] Chandrashekar, G. ve Sahin, F. (2014). A survey on feature selection methods. Computers and electrical engineering, 40(1), 16-28.

[7] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J. ve Liu, H. (2017). Feature selection: A data perspective. ACM computing surveys (CSUR), 50(6), 1-45.

[8] Khaire, U.M. ve Dhanalakshmi, R. (2022). Stability investigation of improved whale optimization algorithm in the process of feature selection. IETE Technical Review, 39(2), 286-300.

[9] Yang, P., Huang, H. ve Liu, C. (2021). Feature selection revisited in the single-cell era. Genome Biology, 22(1), 321.

[10] Kohavi, R. ve John, G.H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1-2), 273-324.

[11] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1), 267-288.

[12] Zou, H. ve Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B: Statistical Methodology, 67(2), 301-320.

[13] Ng, A. Y. (2004, July). Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In Proceedings of the twenty-first international conference on Machine learning (p. 78).

[14] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[15] Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

[16] Bolón-Canedo, V., Sánchez-Maroño, N. ve Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. Knowledge and information systems, 34(3), 483-519.

[17] Peng, H., Long, F. ve Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency. max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence, 27(8), 1226-1238.

[18] Quinlan, J.R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.

[19] Hall, M.A. (2000). Correlation-based feature selection of discrete and numeric class machine learning.

[20] Jain, A.K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.

[21] Marill, K.A. (2004). Advanced statistics: linear regression, part II: multiple linear regression. Academic emergency medicine, 11(1), 94-102.

[22] Hoerl, A.E. ve Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

[23] Faraway, J. J. (2002). Practical regression and ANOVA using R (Vol. 168). Bath: University of Bath.

[24] Chai, T. ve Draxler, R.R. (2014). Root mean square error (RMSE) or mean absolute error (MAE). Geoscientific model development discussions, 7(1), 1525-1534.

[25] Hyndman, R.J. ve Koehler, A.B. (2006). Another look at measures of forecast accuracy, International journal of forecasting. 22(4), 679-688.

[26] Makridakis, S. (1993). Accuracy measures: theoretical and practical concerns. International journal of forecasting, 9(4), 527-529.

[27] Tak, N. ve İnan, D. (2022). Type-1 fuzzy forecasting functions with elastic net regularization. Expert Systems with Applications, 199, 116916.

[28] Belsley, D. A. (1991). A guide to using the collinearity diagnostics. Computer Science in Economics and Management, 4(1), 33–50.

[29] Adkins, L. C. (2022). Weak identification in nonlinear econometric models. Southern University College of Business E-Journal, 17(3), 2.

[30] Williams, G. J. (2009). Rattle: a data mining GUI for R.