*SDÜİİBFD*
*2024*
*29(1), ss. 1-12*

**ISSN: 1301-0603**
**e-ISSN: 2980-3896**
https://dergipark.org.tr/tr/pub/sduiibfd

*SDUJFEAS*
*2024*
*29(1), pp. 1-12*

**Araştırma Makalesi** / *Research Article*

# Entropi Tabanlı Gri İlişkisel Analiz ile Sınıflamada Değişken Seçimi İçin Bir Yaklaşım

*An Approach for Variable Selection in Classification with Entropy-Based Grey Relational Analysis*

Serkan Akoğul[1]

**MAKALE BİLGİSİ**

**Başvuru:** *06.09.2023*
**Revizyon:** *19.10.2023*
**Kabul:** *07.12.2023*
**Yayım:** *30.04.2024*

**Anahtar Kelimeler**
*Çok kriterli karar verme (ÇKKV)*
*Entropi*
*Filtreleme tabanlı değişken seçimi*
*Gri İlişkisel Analiz (GİA)*

**Jel Kodları**
*C10, C38, C44*

**ÖZ**

Günümüzde bilgi ve ölçüm teknolojilerindeki gelişmelere bağlı olarak veri boyutlarında da hızlı bir artış olmuştur. Yüksek boyutluluk nedeniyle özellikle sınıflandırma ve kümeleme gibi birçok analiz için, analiz öncesi veri indirgeme yöntemlerinin kullanılması gerekmektedir. Veri setinin sınıflandırma ve kümeleme sonuçlarının yorumlanması ve daha az değişken ile başarılı sonuçlar elde edilmesi için değişken seçimi önemli rol oynamaktadır. Bu çalışmada, sınıflandırmada kullanılacak değişken seçimi problemi çok kriterli karar verme (ÇKKV) problemi olarak düşünülmüştür. Alternatif olarak değişkenler, kriterler ise filtreleme tabanlı değişken seçim algoritmaları alınarak karar matrisi oluşturulmuş. Bu karar matrisi, ÇKKV yöntemlerinden birisi olan Gri İlişkisel Analiz (GİA) yöntemi kullanarak analiz edilmiştir. Böylelikle değişkenlerin sıralanması ve seçimi için yeni bir yaklaşım ortaya konulmuştur. Ayrıca kriterlerin ağırlıklandırılmasında objektif bir yöntem olan Entropi den yararlanılmıştır. Önerilen yöntem sınıflandırma literatüründe sıkça kullanılan Wine (Şarap) ve Ionosphere (İyonosfer) veri setlerinde uygulanmış olup, filtreleme tabanlı değişken seçim algoritmalarının sınıflandırma başarılarıyla karşılaştırılmıştır. Sonuç olarak önerilen yaklaşım kullanılarak az değişken ile başarılı sonuçlar elde edilmiştir.

**MANUSCRIPT INFO**

**Submitted:** *06.09.2023*
**Revised:** *19.10.2023*
**Accepted:** *07.12.2023*
**Published:** *30.04.2024*

**Keywords**
*Entropy*
*Filter-based variable selection*
*Grey Relational Analysis (GRA)*
*Multi-criteria decision making (MCDM)*

**Jel Codes**
*C10, C38, C44*

**ABSTRACT**

Nowadays, there has been a rapid increase in data size due to developments in information and measurement technologies. Due to the high dimensionality, many analyses, such as classification and clustering, require the use of data reduction methods before analysis. Variable selection plays an important role in interpreting the classification and clustering results of the data set and obtaining successful results with fewer variables. In this study, the variable selection problem to be used in classification is considered a multi-criteria decision-making (MCDM) problem. The decision matrix was created by taking variables as alternatives and filtering-based variable selection algorithms as criteria. This decision matrix was analyzed using the Grey Relational Analysis (GRA) method, which is one of the MCDM methods. Thus, a new approach for ranking and selecting variables has been proposed. In addition, entropy, which is an objective method, is used for weighting the criteria. The proposed method is applied to the Wine and Ionosphere datasets, which are frequently used in the classification literature, and compared with the classification success of filtering-based variable selection algorithms. As a result, successful results were obtained with fewer variables using the proposed approach.

**Önerilen Atıf**
*Suggested Citation*

Akoğul, S. (2024). Entropi tabanlı gri ilişkisel analiz ile sınıflamada değişken seçimi için bir yaklaşım. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, *29*(1), 1-12.

[1] Doç. Dr., Pamukkale Üniversitesi, Fen Fakültesi, İstatistik Bölümü, sakogul@pau.edu.tr, https://orcid.org/0000-0002-0346-4308

## EXTENDED SUMMARY

### Introduction and Research Purpose

Among variable selection methods, filter-based methods are widely used in the literature due to their ease of calculation and speed. These methods select features with the help of functions based on statistical criteria such as distance and information measurements. Variable selection algorithms available in the literature can select different variables in the same data set. This study aims to determine the most appropriate variables using MCDM methods by considering the variable selection in the data set as a decision problem. For the first time in the literature, entropy-based Gray Relational Analysis (E-GRA) was used for variable selection. Additionally, two different methods were applied to determine the appropriate number of variables.

The study was structured as follows: The variables of the data set were taken as alternatives, and filter-based variable selection methods were taken as criteria. Criteria are weighted using the entropy method. Thus, an important ranking of the criteria, that is, variable selection methods, was obtained. Then, the variables were ranked using the E-GRA method. Thus, different variable selection methods were brought together, and variable selection was determined with multi-criteria.

### Methodology

Filter-based variable selection methods: In filter-based methods, variable selection is made based on some statistical criteria without using any classifiers. These methods calculate a score with the help of a function calculated according to the statistical criterion determined for each variable in the data set. With this calculated score, variables are ranked in order of importance, and subsets are created by selecting the desired number of variables. In the study, filter-based Chi-square, Gain Ratio, One-R, and Random Forest methods, which are frequently used in the literature and are easy and fast to calculate, were used.

Classification Algorithm: The C4.5 (J48) decision tree classification algorithm developed by Quinlan in 1993 was used to calculate the classification success of the determined variables.

Criterion weighting method: The entropy method, an objective weighting method that uses a decision matrix without resorting to the personal judgments and opinions of experts, was used to calculate the weights of the criteria.

Gray Relational Analysis: Entropy-based GRA was used as the MCDM method in the study. This method provides the opportunity to rank and classify alternatives by calculating the gray relational degree between each alternative in the decision matrix and the reference values determined.

Proposed Approach: In our study, we consider variable selection methods as a decision criterion and propose a multi-criteria approach for the selection of variables. In this way, the information obtained from variable selection algorithms is considered as a whole, and a common decision mechanism is obtained. The steps of the proposed approach are given below.

Step 1. Variable selection algorithms (KK, GR, OR, RF) were applied to the data set, and scores were calculated for each variable.

Step 2. A decision matrix was created, including variable selection algorithms, criteria, and variables of the data set as alternatives.

Step 3. Considering the created decision matrix, our problem is considered an MCDM problem. Criteria are weighted by the entropy method, and alternatives (variables) are ranked according to the gray relational degree calculated by the GRA method.

### Findings

The approach proposed in the study was applied to the Wine and Ionosphere datasets in the UCI library, which are frequently used in classification and variable selection methods, and the results are given. According to the findings obtained for the Wine dataset, the entropy weights of the variable selection methods are 05.06% for KK, 16.55% for GR, 21.32% for OR, and 57.07% for RF, and RF is determined to be the most important criterion. The decision matrix was analyzed by E-GRA, and the gray relational degrees of the variables were calculated. Using the significantly better variable subset identification method, the selected variable subsets were determined for all methods. Classification was performed with the J48 decision tree algorithm using the identified variables. The DSO and SDO values of the proposed E-GRA approach are 98.88% and 23%, respectively, indicating that it has high classification success and a low variable selection rate. Similarly, the entropy weights for the variables in the Ionosphere dataset are 19.56% for KK, 30.15% for GR, 06.48% for OR, and 43.81% for RF. With the proposed E-GIA approach, the variables were ranked according to other methods, and 8 variables were selected with the top 25% (quartile) selection approach. The classification result shows that the proposed approach has the highest DSO, with a classification rate of 98.29%..

### Conclusion and Discussion

Variable selection methods can give different scores to variables in the same data set and determine different subsets of variables suitable for classification. To overcome this problem, the variable selection was considered as an MCDM problem, thus the variables were re-scored according to multiple criteria using the scores assigned to the variables by different variable selection algorithms. In the study, filter-based Chi-square, Gain Ratio, One-R, and Random Forest variable selection methods, which are frequently used in the literature and are easy to calculate and fast, were considered as criteria, and the alternatives were determined as variables in the data set. This multi-criteria structure was analyzed with Entropy-based Gray Relational Analysis (E-GRA), thus variable selection methods were weighted and variables were ranked according to their importance level.

In this study, the variable selection problem is considered as a decision problem and, unlike the literature, an approach to ranking variables with E-GRA is presented. In addition, the use of two different methods for feature subset selection of filter-based feature algorithms, namely variable subset determination and first percentile selection of the number of significantly better variables, is explained. In this respect, the study proposed an innovative approach to the literature. The proposed approach was tested on two real data sets used in classification and the results were examined in detail. It is intended to be a reference study for future studies and researchers that can be used in classification, clustering, and other multivariate statistical methods using different criteria and different MCDM methods.

## Giriş

Teknolojik gelişmelerle birlikte birçok alanda büyük veri tabanları oluşmuş ve depolanan veri miktarı daha da artmıştır. Bu durum yüksek boyutlu veri ya da büyük veri teriminin ortaya çıkmasına neden olmuştur. Geleneksel yöntemlerin yüksek boyutlu verileri analiz etmedeki yetersizliği ve uzun zaman alması nedeniyle birçok veri madenciliği teknikleri geliştirilmiştir.

Veri madenciliği, büyük boyutlu veriler arasından faydalı bilgiye ulaşma, bilgiyi modelleme ve sonuç çıkarma işlemidir. Bilgisayar tabanlı teknolojiler ile verilerde saklı olan ilişkileri, örüntüleri ve bilgileri ortaya çıkarmayı amaçlayan çok aşamalı bir süreç olarak da tanımlanabilir. Bu sürecin önemli adımlarından biri de değişken seçme sürecidir. Değişken seçimi, orijinal veri kümesinin tamamıyla uyumlu sonuçlar üreten en kullanışlı bir alt kümenin belirlenmesidir (Dash ve Liu, 1997; Ladha ve Deepa, 2011; Wu vd., 2013).

Sınıflandırma ve kümeleme performansını etkileyen faktörlerin başında veri setindeki değişken sayısı gelmektedir. Değişken sayısı yetersiz olduğu durumlarda sınıflandırmadaki ayırım zorlaşırken, fazla olması durumunda da modelleme süresinin artmasına ve gürültülü değişkenlerin yüzünden sınıflandırma başarısının düşmesine sebep olmaktadır. Bu sebeplerden dolayı optimum değişkenle çalışmak daha doğru sonuçlar üretilmesine imkân sağlamaktadır. (Kaynar vd., 2018, Akogul vd., 2020).

Değişken seçim yöntemleri arasında filtre tabanlı yöntemler hesaplama kolaylığı ve hızı nedeniyle literatürde yaygın olarak kullanılmaktadır. Filtre tabanlı değişken seçim yöntemleri uzaklık ve bilgi ölçümleri gibi istatistiksel kriterlere dayalı fonksiyonlar yardımıyla özellik seçimi yapmaktadır. Literatürde var olan değişken seçim algoritmaları farklı yaklaşımlar ile amaca uygun değişkenler seçtikleri için aynı veri setinde farklı değişkenler seçebilmektedirler.

Çok kriterli karar verme (ÇKKV), alternatiflerin birçok kritere göre değerlendirilmesi ve en uygun alternatifin seçilmesi, sıralanması ve sınıflandırılmasını içerir. Bu çalışmanın odak noktası, veri setindeki değişken seçimini bir karar problemi olarak düşünmek ve ÇKKV yöntemleriyle en uygun değişkenleri belirlemektir. Bu amaç doğrultusunda literatür araştırması yapılmış ve literatür üç ana başlıkta toplanmıştır.

Son yıllarda yapılan çalışmalar arasından ilk olarak filtre tabanlı değişken seçimi kullanılan çalışmalar verilmiştir.

Onan ve Korukoğlu (2016) çalışmalarında, metin madenciliğinde filtre tabanlı ve sarmalama-tabanlı öznitelik seçim yöntemlerinin farklı algoritmalar ile kullanımının sınıflandırma başarını karşılaştırmışlardır.

Emhan ve Mehmet (2019) çalışmalarında, filtreleme tabanlı öznitelik seçme yöntemlerinin internet ağlarında anomali tabanlı saldırı tespit sistemlerine etkisini araştırmışlardır 41 özniteliğe sahip veri seti öznitelik seçim yöntemleri ile 8 boyuta indirgenmiş ve boyut indirgemenin sınıflandırma başarısını arttırdığını belirtmişlerdir.

Eryılmaz vd. (2020), Türkçe e-posta verileri üzerinde gereksiz e-postaları tespit etmek için farklı makine öğrenmesi yöntemlerini kullanmışlardır. Analizlerden önce öznitelik çıkarımı ve öznitelik seçimi yaparak yöntemlerin başarılarını karşılaştırmışlardır.

Bulut vd. (2023) çalışmalarında EEG ile elde edilen bant gücü özelliklerinin, öznitelik seçim yöntemleri ve bu yöntemlerin sınıflandırma performanslarına etkisi incelenmiştir.

Sağbaş (2023) çalışmasında metinlerde duygu sınıflandırmasında filtre tabanlı öznitelik seçim yöntemlerinin sınıflandırma performanslarını farklı veri setlerinde karşılaştırmıştır.

İkinci olarak literatürde son yıllarda Entropi tabanlı GİA nın kullanıldığı çalışmalara yer verilmiştir.

Ersoy (2018) çalışmasında, Türkiye'de beyaz eşya sektöründe faaliyet gösteren bir firmanın kurumsal sürdürülebilirlik performansını değerlendirmek için Entropy tabanlı TOPSIS ve GİA yöntemleri kullanmıştır.

Altıntaş (2020), G7 ülkelerinin 2020 yılı Küresel İnovasyon Endeksi verilerine göre Entropi tabanlı Gri İlişkisel Analiz yöntemi ile ülkelerin performanslarını sıralamıştır.

Akbulut (2020) çalışmasında, Gri Entropi, PSI ve ARAS yöntemleri ile 2018 yılında Türkiye'de faaliyet gösteren bazı mevduat bankalarının performansını karşılaştırmıştır.

Ecemiş vd. (2021) çalışmalarında futbol takımlarının performanslarını ÇKKV yöntemleriyle karşılaştırmışlardır. Bu amaçla TFF Süper Lig 2018-2019 Sezonunda mücadele eden futbol takımlarının performanslarını Entropi-Gri İlişkisel Analiz yöntemiyle analiz etmişlerdir.

Song vd. (2022) çalışmalarında 6-UPS paralel mekanizmanın performansını artırmak için Taguchi yöntemi ve E-GİA yöntemini temel alan çok amaçlı bir optimizasyon tasarımı önermişlerdir.

Şekkeli ve Güçlü (2023), katılım bankalarının finansal performanslarının değerlendirilmesinde E-GİA yöntemini kullanmışlardır.

Haseki ve Avşar (2023) çalışmalarında Avrupa Birliği ve seçili diğer ülkelerinin teknoloji üretim verilerine göre E-GİA yöntemi ile analizlerini gerçekleştirmişlerdir.

Son olarak ÇKKV de değişken seçim yöntemlerinin uygulandığı son yıllardaki literatür yer alan bazı çalışmalar aşağıda verilmiştir.

Hashemi vd., (2020) çalışmalarında, çok etiketli değişken seçim yöntemlerini ÇKKV süreci olarak tanımlamışlardır. Değişkenlere uygulanan ridge regresyon yardımı ile karar matrisi oluşturulmuş ve Entropi tabanlı TOPSIS yöntemi ile değişken seçimine karar vermişlerdir. Değişken alt kümesi belirlemede ise istenilen sayıda amaca uygun olarak değişkenlerin seçilimini sunmuşlardır.

Hashemi vd. (2021) çalışmalarında, çok hedefli regresyon sorununu çözmek için ÇKKV yöntemlerinden VIKOR'u kullanarak öznitelik sıralama yaklaşımı sunmuşlardır.

Hashemi vd. (2022) çalışmalarında, topluluk öznitelik seçimini ÇKKV süreci olarak modellenmiştir. Bu amaçla, çeşitli özellik seçim yöntemlerini karar kriteri alarak VIKOR yöntemini ile öznitelik sıralamaları elde etmişlerdir. Özniteliklerin seçimi için kullanıcıların istenilen sayıda seçebileceğini belirtmişlerdir.

Literatürde yer alan çalışmalardan farklı olarak bu çalışmada değişken seçimi bir ÇKKV süreci olarak düşünülmüş, ilk defa değişken seçimi için Entropi tabanlı Gri İlişkisel Analiz (E-GİA) kullanılmıştır. Ayrıca uygun değişken sayısı belirlemek için de iki farklı yöntem uygulanmıştır. Çalışma şu şekilde kurgulanmıştır: Veri setinin değişkenleri birer alternatif, filtre tabanlı değişken seçim yöntemleri birer kriter olarak alınmış, kriter ağırlıklandırılması için Entropi kullanılmış böylece kriterlerin yani değişken seçim yöntemlerinin bir önem sıralaması elde edilmiştir. Daha sonra E-GİA yöntemiyle değişkenler sıralanmıştır. Böylelikle farklı değişken seçme yöntemleri bir araya getirilerek çok kriterli olarak değişken seçimi belirlenmiştir. Ayrıca önerilen yaklaşımın diğer değişken seçim yöntemleri ile sınıflandırma sonuçlarına göre karşılaştırılmıştır. Sonuç olarak önerilen yaklaşım kullanılarak daha az değişken ile başarılı sonuçlar elde edilebileceği gösterilmiştir.

## 1. Materyal ve Metot

### *1.1. Değişken Seçim Yöntemleri: Filtre Tabanlı Algoritmalar*

Değişkenlerin sıralanmasını ya da bir alt kümesinin seçilmesini içerir. Değişken seçiminin avantajları: değişken sayısı azaltır ve algoritma hızını arttırır, veriyi gürültülü değişkenlerden temizler ve veri kalitesini arttırır, veri depolama alanını küçültür, veri kümesini daha basit ve yorumlanması kolaylaşır, görselleştirilebilir ve anlaşılabilir hale getirir aşağıdaki şekilde sıralayabiliriz.

Değişken seçimi; filtreleme tabanlı, sarmal ve gömülü yöntemler olmak üzere üç ana başlıkta incelenir. Filtreleme tabanlı yöntemlerde, herhangi bir sınıflandırıcı kullanılmadan bazı istatistiksel ölçütlere dayalı olarak değişken seçimi yapılmaktadır. Bu yöntemler, veri kümesinde bulunan her bir değişken için belirlenen istatistiksel ölçüte göre hesaplanan bir fonksiyon yardımıyla bir skor hesaplar. Hesaplanan bu skor ile değişkenler önem sırasına göre sıralanır ve istenilen sayıda değişken seçilerek alt kümeler oluşturulmuş olur (Dash ve Liu, 1997; Guyon ve Elisseeff, 2003; Budak, 2018).

Çalışmada literatürde sıklıkla kullanılan, hesaplaması kolay ve hızlı olan filtreleme tabanlı Ki-kare, Kazanç Oranı, One-R ve Rasgele Orman yöntemleri kullanılmıştır.

- **Ki-Kare testi:** Veriden elde edilen gözlenen değerler ile dağılımdan elde edilen beklenen değerler arasındaki farkın anlamlılığını test eden Ki-kare testi (KK), amaç değişkeni ile diğer değişkenler arasındaki ilişkiyi hesaplayarak, ilişki derecesine göre değişkenleri sıralayan fitreleme tabanlı değişken seçme yöntemidir (Liu ve Setiono, 1995).
- **Kazanç Oranı:** Kazanç Oranı (Gain Ratio-GR) yöntemi, entropiye dayalı hesaplanan değişken seçim yöntemlerinden Bilgi kazancının, farklı değerlere sahip ilişkili değişkenleri seçebilme durumunu engellemek için önerilmiştir. Kazanç oranı, bilgi kazanç değerleri ile bölünmüş bilgilerin oranı alınarak elde edilen normalleştirmedir (Karegowda vd., 2010).
- **One-R:** One-R (OR) algoritması (Holte, 1993), veri setini test ederek her bir değişken için bir kural belirler ve bu kural ile değişkenlerin sınıflandırma başarılarını hesaplar. Böylelikle hata oranı en düşük ve en yüksek değişkenleri belirleyerek değişkenleri sıralar (Akmaz, 2022).
- **Rasgele orman algoritması:** Rasgele orman (Random Forest-RF) algoritması, sadece bir karar ağacı oluşturmak yerine veri setinden elde edilen farklı eğitim verisinde eğitilmiş olan fazla sayıda karar ağacını birleştirerek karar ağaçları oluşturur. Bilgi kazancı hesabına göre, her bir değişkenin göreceli önemini hesaplayarak bir skor ataması yapar. Karar ağacının bölünme noktalarını seçmek için kullanılan GINI'ya göre bu skorlar belirlenir (Breiman, 2001).

### *1.2. Sınıflandırma Yöntemi: C4.5 Algoritması*

C4.5 (J48) karar ağacı, 1993 yılında Quinlan tarafından geliştirilen ve bilgi entropisini kullanarak ikili bir karar ağacı oluşturan sınıflandırma algoritmasıdır (Quinlan, 1993). Her bir değişkenin entropisi ile sınıfların entropisi karşılaştırılarak bilgi kazancı hesaplanır ve kök düğümler belirlenir. Verileri en iyi şekilde kategorize edene kadar karar ağaç üzerinde budama işlemi yapılarak nihai karar ağacı belirlenir (Breiman, 2017).

### *1.3. Kriter ağırlıklandırma Yöntemi: Entropi*

Nesnel (objektif) bir ağırlıklandırma yöntemi olan Entropi, enformasyon teorisine göre belirsizlik düzeyinin ölçülmesidir. Bu yöntem ÇKKV problemlerinde uzmanların kişisel yargı ve düşüncelerine başvurmadan karar matrisinden kriterlerin ağırlıklarının hesaplanmasına imkân sağlayan Entropi yönteminin adımları sırasıyla aşağıda ifade edilmiştir. (Shannon, 1948; Karaatlı, 2016; Bircan, 2020)

**Adım 1.** $X = \left[x_{ij}\right]_{nxm}$ karar marisi $r_{ij} = x_{ij} / \sum_{i=1}^{n} x_{ij}$ formülasyonu kullanılarak $R = \left[r_{ij}\right]_{nxm}$ normalleştirilmiş karar matrisi hesaplanmaktadır.

**Adım 2**. $e_j = -\frac{1}{\ln n} \sum_{i=1}^{n} r_{ij} \ln r_{ij}$ formülü ile her bir kriterin entropisi bulunmaktadır.

**Adım 3**. $w_j = (1 - e_j) / \sum_{j=1}^{m} (1 - e_j)$ ile $j$'inci kritere ait ağırlık değeri hesaplanmaktadır.

### *1.4. ÇKKV Yöntemi: Gri İlişkisel Analiz*

1982'de Ju Long Deng tarafından ortaya atılan gri sistem teorisi; zayıf, yetersiz ve belirsiz olan bilgileri sayısallaştırılmasında kullanılan bir yöntemdir. Bu teoriyi kullanan GİA, karar matrisinde yer alan her bir alternatif ile belirlenen referans değerler arasındaki gri ilişkisel dereceyi hesaplayarak, alternatifler için sıralama ve sınıflandırma imkânı sunun ÇKKV yöntemidir. Yöntemin adımları aşağıda kısaca özetlenmiştir (Karaatlı, 2016; Özdemir ve Kılıçarslan, 2021; Lu vd., 2009; Yıldırım ve Önder, 2015).

**Adım 1.** Karar matrisinin oluşturulması: $X = \left[x_{ij}\right]_{nxm}$ karar matrisi, karar verici tarafından oluşturulan *n* alternatifin $(i = 1, \dots, n)$ ve *m* kriterin $(j = 1, \dots, m)$ bulunduğu veri matrisidir.

**Adım 2**. Referans serisinin belirlenmesi: Karar matrisindeki alternatifleri kıyaslamak üzere amaca uygun belirlenen $x_0 = (x_0(j))$ referans serisidir.

**Adım 3.** Normalizasyon işlemi: Kriterlerin alternatifler üzerindeki etkisinin fayda $x_i^* = \frac{x_i(j) - min\, x_i(j)}{max\, x_i(j) - min\, x_i(j)}$ ya da maliyet $x_i^* = \frac{max\, x_i(j) - x_i(j)}{max\, x_i(j) - min\, x_i(j)}$ olması durumuna göre karar matrisi ve referans serisinin normalize edilmesidir.

**Adım 4.** Mutlak farklar matrisinin oluşturulması: Normalize matrisdeki her bir alternatifin $x_i^*(j)$, normalleştirilmiş referans seriye $x_0^*(j)$ mutlak uzaklıkları $\Delta_i(j) = |x_0^*(j) - x_i^*(j)|$ hesaplanarak mutlak farklar matrisi $\Delta_{ij}$ oluşturulur.

**Adım 5**. Gri ilişkisel katsayı matrisinin oluşturulması: Mutlak farklar matrisi $\Delta_{ij}$ nin tüm elemanlarının minimumu $\Delta_{min}$ ve maksimumu $\Delta_{max}$ olmak üzere Gri ilişkisel katsayı matrisinin elemanları $\gamma_i(j) = \frac{\Delta_{min} + \zeta\, \Delta_{max}}{\Delta_i(j) + \zeta\, \Delta_{max}}$ hesaplanarak gri ilişkisel katsayı matrisi $\gamma_{ij}$ oluşturulur. Burada $\zeta$ parametresi [0,1] aralığında değerler alan ayırıcı katsayısıdır. Genellikle $\zeta = 0.5$ değeri ile orta düzeyde bir ayırt edici olup, iyi bir kararlılık sunar.

**Adım 6.** Gri ilişkisel derecelerin hesaplanması. Gri ilişkisel derece $\Gamma_i$, kriterlerin eşit $\Gamma_i = \frac{1}{n} \sum_{j=1}^{m} \gamma_i(j)$ ya da farklı $\Gamma_i = \sum_{j=1}^{m} [w_i(j) \gamma_i(j)]$ ağırlıklara sahip olması durumuna göre eşitlikler yardımıyla hesaplanır ($\sum_{j=1}^{m} w_j = 1$). Hesaplanan gri ilişkisel dereceye göre alternatifler sıralanır ve derecesi en yüksek olan alternatif amaca uygun en iyi alternatif olarak seçilir.

### *1.5. Önerilen Yaklaşım*

Değişken seçim algoritmaları daha az değişken ile başarılar sonuçlar elde etmeyi amaçlayan yöntemlerdir. Anlaşılması kolay ve hızlı olan filtre tabanlı değişken seçim yöntemleri literatürde sıklıkla kullanılmaktadır. Bu yöntemlerde kullanılan farklı yaklaşımlardan dolay aynı veri seti için farklı değişken sıralanması ve seçimi gerçekleşebilmektedir. Her bir değişken seçim algoritmasının kendine göre üstün özellikleri mevcuttur. Çalışmamızda değişken seçim yöntemlerini bir karar kriteri olarak ele alıp değişkenlerin seçimi için çok kriterli bir yaklaşım öneriyoruz. Böylelikle değişken seçim algoritmalarından elde edilen bilgi bir bütün olarak ele alınıp ortak bir karar mekanizması elde etmiş oluyoruz. Şekil 1'de önerilen yaklaşım gösterilmiştir.

Çalışmamızda filtre tabanlı değişken seçim algoritmalarından farklı özelliklere sahip Ki-Kare (KK), Kazanç Oranı (GR), One-R (OR) ve Rasgele Orman (RF) yöntemleri kullanılmıştır. Bu değişken seçim yöntemleri değişkenlerdeki bilgi, Entropi, ilişki gibi istatistiksel bilgilere dayalı hesaplamalar yapmaktadır. Bu nedenle algoritmaların ağırlıklandırılmasında Entropi yönteminden yararlanılmıştır. Yine gri sistem teorisine sahip GİA yöntemiyle de değişkenler ÇKKV ile sıralanmıştır. Değişken seçim algoritmaları ve önerilen yöntem ile sıralanan değişkenlerden önemli ölçüde daha iyi olan değişken alt kümesi belirlenmiştir. Belirlenen değişkenler ile sınıflandırma yapılıp yöntemlerin doğru sınıflandırma oranları (DSO-Doğruluk) ve seçilen değişken oranları (SDO) karşılaştırılmıştır. DSO değeri maksimum SDO değeri minimum olan yöntem başarılı olarak kabul edilmektedir.

**Şekil 1.** Değişken Seçimi İçin Önerilen Yaklaşım


Önerilen yaklaşımın adımları aşağıdaki verilmiştir.

**Adım 1.** Veri setine değişken seçim algoritmaları (KK, GR, OR, RF) uygulanmış her bir değişken için skorlar hesaplanmıştır.

**Adım 2.** Değişken seçim algoritmaları kriterler ve veri setine ait değişkenler birer alternatif olmak üzere karar matrisi oluşturulmuştur.

**Adım 3.** Oluşturulan karar matrisi göze alınarak problemimiz bir ÇKKV problemi olarak düşünülmüştür. Kriterler Entropi yöntemi ile ağırlıklandırılmış, alternatifler (değişkenler) de GİA yöntemine ile hesaplanan gri ilişkisel dereceye göre sıralanmıştır.

## 2. Uygulama

### *2.1. Veri Seti*

Önerilen yöntem kümeleme ve sınıflandırma literatürde yaygın kullanıma sahip Wine (Şarap) ve Ionosphere (iyonosfer) veri setileri üzerinde test edilmiştir (Lichman, 2013). Wine veri seti İtalya'nın aynı bölgesinde 3 farklı çeşit yetiştirilen ürünlerden elde edilen şarapların kimyasal analizinin sonuçlarıdır. 13 farklı değişkene bakılarak şarapların kalite skalalarına göre sınıflandırılmasına dair 178 gözlemden oluşmaktadır. Değişkenler sırasıyla, Alcohol (alkol oranı), Malic (malik asit derecesi), Ash (sodyum karbonat oranı), Alcalinity (sodyum karbonat alkanilitesi), Magnesium (magnezyum değeri), Phenols (fenoller), Flavanoids (flavanoid değeri), Nonflavanoids (flavonoid içermeyen fenoller), Proanthocyanins (proantosiyanidinler), Color (renk yoğunluğu), Hue (renk tonu), Dilution (seyreltme) ve Proline (prolin değeri) olarak adlandırılmaktadır. Wine veri seti sırasıyla 59, 71 ve 48 gözlem bulunan 3 kümeden oluşmaktadır.

Ionosphere veri seti ise Kanada'daki Goose Bay Labrador sistemi tarafından toplanan radar yansıma verilerinden oluşmaktadır. Sistemdeki antene gelen radar dönüşleri "iyi" ve "kötü" olarak sınıflandırılmıştır. Ionosphere verisi 351 adet örnekten oluşmakta olup toplam 34 [X1-X34] değişken bakımından veriler toplanmıştır. Verilerin 223 tanesi "iyi", 128 tanesi de "kötü" olarak etiketlendirilmiştir.

### *2.2. Bulgular*

Veri setleri için hesaplanan değişken seçim algoritmalarının sonuçları RStudio (Allaire, 2012) da bulunan "FSelector" paketi (Romanski vd., 2013) ile hesaplanmıştır. Yöntemler tarafından sıralanan değişkenlerin sınıflandırma başarılarını karşılaştırmak için yöntemlerin doğru sınıflandırma oranları (DSO) kullanılmıştır. Sınıflandırmada kullanılacak değişkenler için, R da "FSelector" paketi içinde yer alan "cutoff.biggest.diff()" ve "cutoff.k.percent" komutları kullanılmıştır. Komutlar sırasıyla önemli ölçüde daha iyi olan değişken alt küme belirleme ve değişken sayısının ilk yüzde k'ınci değişkeni seçme anlamına gelmektedir. Eğer tüm değişkenler arasında önemli farklar oluşmuyorsa yüzde seçim komutu ile en iyi ilk yaklaşık %25 (birinci çeyreklik) seçim yöntemi kullanılmıştır.

Wine verisine uygulanan KK, GR, OR ve RF algoritmalarının hesapladığı skor değerleri ile karar matrisi Tablo 1'de oluşturulmuştur. Karar matrisi yardımıyla hesaplanan Entropi ve GİA sonuçları için RStudio'da komutlar yazılmış olup analizler yapılmıştır.

**Tablo 1.** Wine Veri Seti İçin Karar Matrisi

| | **KK** | **GR** | **OR** | **RF** |
|---|---|---|---|---|
| **Alcohol** | 0,5759 | 0,4373 | 0,4101 | 29,5630 |
| **Malic** | 0,5395 | 0,2922 | 0,3820 | 15,3090 |
| **Ash** | 0,4368 | 0,3252 | 0,1742 | 8,4650 |
| **Alcalinity** | 0,5997 | 0,3155 | 0,2865 | 14,5630 |
| **Magnesium** | 0,5469 | 0,3139 | 0,2809 | 15,3810 |
| **Phenols** | 0,5901 | 0,3909 | 0,4101 | 17,8650 |
| **Flavanoids** | 0,8044 | 0,5445 | 0,5562 | 33,9380 |
| **Nonflavanoids** | 0,5266 | 0,2282 | 0,2360 | 5,4100 |
| **Proanthocyanins** | 0,5982 | 0,3097 | 0,2416 | 11,8710 |
| **Color** | 0,6859 | 0,5251 | 0,4888 | 35,1320 |
| **Hue** | 0,6265 | 0,3633 | 0,3876 | 26,6500 |
| **Dilution** | 0,6536 | 0,5730 | 0,3876 | 25,6510 |
| **Proline** | 0,7061 | 0,4453 | 0,4551 | 36,4980 |

Karar matrisi kullanılarak kriterlere ait Entropi ağırlıkları Tablo 2'de verilmiştir. Tabloya göre Ki Kare (KK) %05,06; Kazanç Oranı (GR) %16,55; One-R (OR) %21,32 ve Rasgele Orman (RF) %57,07 önem ağırlıklarına sahip olduğu belirlenmiştir.

**Tablo 2.** Wine Veri Setine ait Entropi Bulguları

| **Entropi Parametreleri** | **KK** | **GR** | **OR** | **RF** |
|---|---|---|---|---|
| $e_j$ | 0,9959 | 0,9864 | 0,9825 | 0,9532 |
| $d_j$ | 0,0041 | 0,0136 | 0,0175 | 0,0468 |
| $w_j$ | 0,0506 | 0,1655 | 0,2132 | 0,5707 |

Entropi tabanlı GİA (E-GİA) için karar matrisinden sonraki adım referans serisinin belirlenmesidir. Burada tüm kriterlerin büyük değerleri alternatifler için istenilen durum olmasından dolay referans serisi tüm kriterlerin maksimum değeri $x_0 = \{0{,}8044;\ 0{,}5730;\ 0{,}5562;\ 36{,}4980\}$ olarak alınmıştır. Karar matrisinin ve referans serisinin normalizasyonunda fayda durumu göz önünde bulundurularak hesaplanan normalize matris Tablo 3'te verilmiştir.

**Tablo 3.** Normalize Matris

| | **KK** | **GR** | **OR** | **RF** |
|---|---|---|---|---|
| ***Referans Seri*** | *1,0000* | *1,0000* | *1,0000* | *1,0000* |
| **Alcohol** | 0,3784 | 0,6064 | 0,6175 | 0,7769 |
| **Malic** | 0,2794 | 0,1856 | 0,5440 | 0,3184 |
| **Ash** | 0,0000 | 0,2813 | 0,0000 | 0,0983 |
| **Alcalinity** | 0,4431 | 0,2532 | 0,2940 | 0,2944 |
| **Magnesium** | 0,2995 | 0,2485 | 0,2793 | 0,3207 |
| **Phenols** | 0,4170 | 0,4719 | 0,6175 | 0,4006 |
| **Flavanoids** | 1,0000 | 0,9173 | 1,0000 | 0,9177 |
| **Nonflavanoids** | 0,2443 | 0,0000 | 0,1618 | 0,0000 |
| **Proanthocyanins** | 0,4391 | 0,2364 | 0,1764 | 0,2078 |
| **Color** | 0,6776 | 0,8611 | 0,8236 | 0,9561 |
| **Hue** | 0,5161 | 0,3918 | 0,5586 | 0,6832 |
| **Dilution** | 0,5898 | 1,0000 | 0,5586 | 0,6511 |
| **Proline** | 0,7326 | 0,6296 | 0,7353 | 1,0000 |

Normalize matrisin elemanlarının referans seriye olan mutlak uzaklıkları hesaplanarak Tablo 4'te verilmiştir.

**Tablo 4.** Mutlak Farklar Matrisi

| | **KK** | **GR** | **OR** | **RF** |
|---|---|---|---|---|
| **Alcohol** | 0,6216 | 0,3936 | 0,3825 | 0,2231 |
| **Malic** | 0,7206 | 0,8144 | 0,4560 | 0,6816 |
| **Ash** | 1,0000 | 0,7187 | 1,0000 | 0,9017 |
| **Alcalinity** | 0,5569 | 0,7468 | 0,7060 | 0,7056 |
| **Magnesium** | 0,7005 | 0,7515 | 0,7207 | 0,6793 |
| **Phenols** | 0,5830 | 0,5281 | 0,3825 | 0,5994 |
| **Flavanoids** | 0,0000 | 0,0827 | 0,0000 | 0,0823 |
| **Nonflavanoids** | 0,7557 | 1,0000 | 0,8382 | 1,0000 |
| **Proanthocyanins** | 0,5609 | 0,7636 | 0,8236 | 0,7922 |
| **Color** | 0,3224 | 0,1389 | 0,1764 | 0,0439 |
| **Hue** | 0,4839 | 0,6082 | 0,4414 | 0,3168 |
| **Dilution** | 0,4102 | 0,0000 | 0,4414 | 0,3489 |
| **Proline** | 0,2674 | 0,3704 | 0,2647 | 0,0000 |

Gri ilişkisel katsayı matrisinin hesaplanmasında $\Delta_{min} = 0$, $\Delta_{max} = 1$ ve $\zeta = 0{,}5$ değeri kullanılmış olup matris Tablo 5'te verilmiştir.

**Tablo 5.** Gri İlişkisel Katsayı Matrisi

| | **KK** | **GR** | **OR** | **RF** |
|---|---|---|---|---|
| **Alcohol** | 0,4458 | 0,5596 | 0,5666 | 0,6915 |
| **Malic** | 0,4096 | 0,3804 | 0,5230 | 0,4232 |
| **Ash** | 0,3333 | 0,4103 | 0,3333 | 0,3567 |
| **Alcalinity** | 0,4731 | 0,4010 | 0,4146 | 0,4147 |
| **Magnesium** | 0,4165 | 0,3995 | 0,4096 | 0,4240 |
| **Phenols** | 0,4617 | 0,4863 | 0,5666 | 0,4548 |
| **Flavanoids** | 1,0000 | 0,8581 | 1,0000 | 0,8586 |
| **Nonflavanoids** | 0,3982 | 0,3333 | 0,3736 | 0,3333 |
| **Proanthocyanins** | 0,4713 | 0,3957 | 0,3778 | 0,3869 |
| **Color** | 0,6080 | 0,7826 | 0,7392 | 0,9192 |
| **Hue** | 0,5082 | 0,4512 | 0,5311 | 0,6122 |
| **Dilution** | 0,5493 | 1,0000 | 0,5311 | 0,5890 |
| **Proline** | 0,6515 | 0,5745 | 0,6539 | 1,0000 |

Alternatiflerin sıralanmasında son olarak gri ilişki dereceleri hesaplanmıştır. Gri ilişki derecesini belirlemek için kriterlere ait ağırlıklar $w_j = \{0{,}0506;\ 0{,}1655;\ 0{,}2132;\ 0{,}5707\}$ göz önünde bulundurulmuştur. Her bir alternatife ait gri ilişkisel dereceler ve alternatiflerin sıralanması Tablo 6 da gösterilmiştir.

**Tablo 6.** Wine Veri Seti İçin Gri İlişkisel Dereceler ve Değişkenlerin Sıralanması

| **Değişkenler** | $\Gamma_i$ | **Sıra** |
|---|---|---|
| **Flavanoids** | 0,8958 | 1 |
| **Color** | 0,8425 | 2 |
| **Proline** | 0,8382 | 3 |
| **Dilution** | 0,6427 | 4 |
| **Alcohol** | 0,6306 | 5 |
| **Hue** | 0,5630 | 6 |
| **Phenols** | 0,4842 | 7 |
| **Malic** | 0,4367 | 8 |
| **Magnesium** | 0,4165 | 9 |
| **Alcalinity** | 0,4154 | 10 |
| **Proanthocyanins** | 0,3907 | 11 |
| **Ash** | 0,3594 | 12 |
| **Nonflavanoids** | 0,3452 | 13 |

E-GİA sonuçlarına göre en başarılı değişken "Flavanoids" olarak belirlenmiştir. Diğer değişkenlerde önem derecelerine göre sıralanmıştır. Değişken seçim yöntemlerine ve E-GİA sonuçlarına göre seçilen en önemli değişken alt kümeleri ("cutoff.biggest.diff" komutu) Tablo 7'de verilmiştir.

**Tablo 7.** Yöntemlerin ve Önerilen Yaklaşımın Sınıflandırma İçin Seçilen Değişkenleri

| **Yöntemler** | **Seçilen Değişkenler** |
|---|---|
| **KK** | Flavanoids |
| **GR** | Dilution, Flavanoids, Color |
| **OR** | Flavanoids, Color, Proline, Alcohol, Phenols, Hue, Dilution, Malic |
| **RF** | Proline, Color, Flavanoids, Alcohol, Hue, Dilution |
| **E-GİA** | Flavanoids, Color, Proline |

R'da bulunan "RWeka" paketi (Hornik vd., 2023) içindeki J48 karar ağacı sınıflandırma algoritması kullanılarak elde edilen sonuçlar Tablo 8'de verilmiştir. Tabloya göre önerilen yaklaşım yüksek DSO içerisinde ve düşük SDO sahiptir.

**Tablo 8.** Yöntemlerin Sınıflandırma Başarıları

| | **KK** | **GR** | **OR** | **RF** | **E-GİA** |
|---|---|---|---|---|---|
| **DSO** | %82,58 | %95,51 | %98,88 | %98,88 | %98,88 |
| **SDO** | %8 | %23 | %62 | %46 | %23 |

Önerilen yöntemin Ionosphere veri setine uygulanması sonucu elde edilen karar matrisi Tablo 9'da verilmiştir Karar matrisi kullanılarak kriterlere ait Entropi ağırlıkları KK için %19,56; GR için %30,15; OR için %06,48 ve RF için de %43,81 olarak bulunmuştur.

**Tablo 9.** Ionosphere Veri Seti İçin Karar Matrisi

| | **KK** | **GR** | **OR** | **RF** | | **KK** | **GR** | **OR** | **RF** |
|---|---|---|---|---|---|---|---|---|---|
| **X1** | 0,4656 | 0,3590 | 0,2308 | 13,2820 | **X18** | 0,4923 | 0,2752 | 0,2450 | 15,5520 |
| **X2** | 0,0000 | 0,0000 | 0,1225 | 0,0000 | **X19** | 0,5808 | 0,1099 | 0,2735 | 9,8760 |
| **X3** | 0,6673 | 0,1970 | 0,2934 | 29,8420 | **X20** | 0,4972 | 0,2411 | 0,2507 | 12,2310 |
| **X4** | 0,6484 | 0,1642 | 0,3134 | 21,2580 | **X21** | 0,6870 | 0,1906 | 0,3162 | 12,0280 |
| **X5** | 0,7290 | 0,2659 | 0,3134 | 32,7330 | **X22** | 0,6723 | 0,1753 | 0,3305 | 13,0090 |
| **X6** | 0,7449 | 0,2119 | 0,3590 | 23,8150 | **X23** | 0,6236 | 0,1507 | 0,2877 | 11,6700 |
| **X7** | 0,6722 | 0,2574 | 0,2991 | 23,7340 | **X24** | 0,5017 | 0,2353 | 0,2536 | 17,7600 |
| **X8** | 0,6843 | 0,1886 | 0,3305 | 23,1590 | **X25** | 0,6193 | 0,1528 | 0,2991 | 12,4770 |
| **X9** | 0,5521 | 0,2132 | 0,2621 | 14,1810 | **X26** | 0,4712 | 0,2100 | 0,2422 | 12,4880 |
| **X10** | 0,4892 | 0,1501 | 0,2308 | 16,2580 | **X27** | 0,6432 | 0,2104 | 0,2991 | 24,0720 |
| **X11** | 0,5792 | 0,1357 | 0,2764 | 10,5960 | **X28** | 0,5995 | 0,2928 | 0,2963 | 15,3350 |
| **X12** | 0,6235 | 0,1523 | 0,3048 | 14,5580 | **X29** | 0,6849 | 0,1957 | 0,3219 | 14,1180 |
| **X13** | 0,6718 | 0,1595 | 0,2991 | 9,9070 | **X30** | 0,3964 | 0,1415 | 0,2165 | 9,6260 |
| **X14** | 0,5295 | 0,1946 | 0,2536 | 20,7210 | **X31** | 0,6636 | 0,1785 | 0,3077 | 14,8690 |
| **X15** | 0,6338 | 0,1513 | 0,2849 | 9,2540 | **X32** | 0,4767 | 0,2068 | 0,2450 | 13,1130 |
| **X16** | 0,6441 | 0,1675 | 0,3162 | 19,0890 | **X33** | 0,6953 | 0,2150 | 0,3191 | 11,4790 |
| **X17** | 0,5975 | 0,1299 | 0,2792 | 9,6390 | **X34** | 0,6964 | 0,1911 | 0,3419 | 17,0440 |

Karar matrisinden hesaplanan E-GİA sonuçlarına göre değişkenler için hesaplanan gri ilişkisel dereceler ve sıralamaları Tablo 10' da verilmiştir.

**Tablo 10.** Ionosphere Veri Seti İçin Gri ilişkisel dereceler ve değişkenlerin sıralanması

| **Değişkenler** | $\Gamma_i$ | **Sıra** | **Değişkenler** | $\Gamma_i$ | **Sıra** |
|---|---|---|---|---|---|
| **X5** | 0,8710 | 1 | **X22** | 0,5636 | 18 |
| **X3** | 0,7344 | 2 | **X24** | 0,5598 | 19 |
| **X6** | 0,7097 | 3 | **X12** | 0,5397 | 20 |
| **X7** | 0,6819 | 4 | **X9** | 0,5362 | 21 |
| **X8** | 0,6515 | 5 | **X13** | 0,5323 | 22 |
| **X27** | 0,6481 | 6 | **X20** | 0,5277 | 23 |
| **X1** | 0,6446 | 7 | **X25** | 0,5254 | 24 |
| **X28** | 0,6156 | 8 | **X23** | 0,5191 | 25 |
| **X34** | 0,6091 | 9 | **X15** | 0,5102 | 26 |
| **X4** | 0,6042 | 10 | **X32** | 0,5091 | 27 |
| **X16** | 0,5863 | 11 | **X26** | 0,5059 | 28 |
| **X29** | 0,5806 | 12 | **X10** | 0,5047 | 29 |
| **X33** | 0,5789 | 13 | **X11** | 0,4941 | 30 |
| **X18** | 0,5688 | 14 | **X17** | 0,4930 | 31 |
| **X14** | 0,5682 | 15 | **X19** | 0,4824 | 32 |

| | | | | | |
|---|---|---|---|---|---|
| **X21** | 0,5658 | 16 | **X30** | 0,4484 | 33 |
| **X31** | 0,5655 | 17 | **X2** | 0,3333 | 34 |

E-GİA sonuçlarına göre en başarılı değişken "X5" olarak belirlenmiş ve diğer değişkenlerde önem derecelerine göre sıralanmıştır. Değişken seçim yöntemlerinin sonuçları incelendiğinde tüm değişkenler için yakın skorlar hesaplandığından en iyi değişken alt kümesi belirlenememiş, tüm değişkenlerin seçilmesi eğilimi göstermiştir. Bu durumda alternatif yöntem olan ilk yüzde k'inci değişken seçim yöntemi ("cutoff.k.percent" komutu %25) kullanılmıştır. Böylelikle yöntemlerin aynı oranda değişken seçimlerinin DSO karşılaştırılmıştır. Yöntemlerin ve E-GİA sonuçlarına göre seçilen ilk 8 değişkenleri Tablo 11'de verilmiştir.

**Tablo 11.** Yöntemlerin ve Önerilen Yaklaşımın Sınıflandırma İçin Seçilen Değişkenleri

| **Yöntemler** | **Seçilen değişkenler** |
|---|---|
| **KK** | X6, X5, X34, X33, X21, X29, X8, X22 |
| **GR** | X1, X28, X18, X5, X7, X20, X24, X33 |
| **OR** | X6, X34, X8, X22, X29, X33, X16, X21 |
| **RF** | X5, X3, X27, X6, X7, X8, X4, X14 |
| **E-GİA** | X5, X3, X6, X7, X8, X27, X1, X28 |

Tablo 11'e göre tüm değişken seçme yöntemlerinin farklı değişken alt kümleri belirledikleri görülmektedir. Veri seti seçilen değişkenler ile J48 karar ağacı sınıflandırma algoritması kullanılarak sınıflandırılmış ve elde edilen sonuçlar Tablo 12'de verilmiştir. Tablo12'ye göre aynı sayıda seçilen değişkenler bakımından önerilen yaklaşımın yüksek DSO sahip olduğu belirlenmiştir.

**Tablo 12.** Yöntemlerin Sınıflandırma Başarıları

| | **KK** | **GR** | **OR** | **RF** | **E-GİA** |
|---|---|---|---|---|---|
| **DSO** | %95,73 | %96,01 | %97,44 | %96,87 | %98,29 |

## Sonuç

Veri madenciliğinin önemli aşamasından biri de veri setindeki değişkenlerin seçimidir. Özellikle büyük verilerde bulanıklık, fazla ve gürültülü değişkenlerin olması analizleri olumsuz etkilemektedir. Makine öğrenmesi yöntemlerinde değişken sayısının azaltılması, modelin daha hızlı ve daha kısa sürede eğitilmesine imkân sağlar. Sınıflandırma problemlerinde gözlemlerin sınıflara doğru şekilde atanabilmesi için az değişkenle yüksek başarı elde etmek değişken seçim yöntemlerinin temel amacıdır. Kullanım kolaylığı ve hızlı olmasından dolayı filtre tabanlı değişken seçim yöntemleri literatürde sıklıkla kullanılmaktadır. Değişken seçim yöntemleri farklı algoritmalar kullanarak sınıflandırma başarısı en yüksek olacak şekilde değişkenleri belirler. Bu değişken seçim yöntemleri aynı veri setinde değişkenlere farklı skorlar verebilmekte ve sınıflandırmaya uygun farklı değişken alt kümeleri belirleyebilmektedirler. Bu sonunun üstesinden gelmek için değişken seçimini ÇKKV problemi olarak düşünülmüş, böylece farklı değişken seçim algoritmalarının değişkenlere atadıkları skorlar kullanılarak değişkenler bir den çok kritere göre yeniden skorlanmıştır. Çalışmada literatürde sıklıkla kullanılan, hesaplaması kolay ve hızlı olan filtreleme tabanlı Ki-kare, Kazanç Oranı, One-R ve Rasgele Orman değişken seçim yöntemleri birer kriter olarak düşünülmüş, alternatiflerde veri setindeki değişkenler olarak belirlenmiştir. Oluşturulan bu çok kriterli yapı Entropi tabanlı Gri İlişkisel Analiz (E-GİA) ile çözümlenmiş böylece değişken seçim yöntemleri ağırlıklandırılmış ve değişkenler de önem düzeyine göre sıralanmıştır.

Çalışmada önerilen yaklaşımın tüm adımları UCI kütüphanesinde yer alan sınıflandırma ve değişken seçim yöntemlerinde sıklıkla kullanılan Wine veri setinde detaylıca anlatılmış olup ayrıca Ionosphere veri seti içinde sonuçlar verilmiştir. Wine veri setine göre elde edilen bulgular, değişken seçim yöntemlerinin Entropi ağırlıkları KK için %05,06; GR için %16,55; OR için %21,32 ve RF için %57,07 şeklinde olup en önemli kriter olarak RF olarak belirlenmiştir. Oluşturulan karar matrisi E-GİA ile çözümlenmiş değişkenlere ait gri ilişkisel dereceler hesaplanmıştır. Önemli ölçüde daha iyi olan değişken alt küme belirleme yöntemi kullanılarak tüm yöntemler için seçilen değişken alt kümeleri belirlenmiştir. Önerilen yaklaşım için önemli seçilen ilk üç değişken "Flavanoids (0,8958), Color (0,8425), Proline (0,8382)" olarak belirlenmiştir. Seçilen değişkenlerin sınıflandırmadaki başarılarını karşılaştırmak için J48 karar ağacı algoritması kullanılmış ve değişken seçim yöntemleri ile önerilen yaklaşımın doğru sınıflandırma oranları karşılaştırılmıştır. Önerilen E-GİA yaklaşımı %98,88 DSO ve %23 SDO ile yüksek sınıflandırma başarısı ve düşük değişken seçim oranına sahip olduğu görülmektedir. Benzer şekilde Ionosphere veri setindeki değişkenler için Entropi ağırlıkları KK için %19,56; GR için %30,15; OR için %06,48 ve RF için de %43,81 olarak bulunmuştur. Burada da yine en önemli kriter RF olduğu gözükmektedir. Bu durum RF algoritmasının değişkenleri açıklamada yüksek bilgi içerdiğini göstermektedir. Önerilen E-GİA yaklaşım ile diğer yöntemlere göre değişkenler sıralanmış ve değişken sayısının ilk yüzde %25 (çeyreklik) seçim yaklaşımı ile seçilen 8 değişkenler belirlenmiştir. Önerilen yaklaşımın sınıflandırma sonucunda en yüksek DSO (%98,29) sahip olduğu görülmüştür.

Bu çalışmada değişken seçim problemi bir karar problemi olarak düşünülmüş olup literatürden farklı olarak E-GİA ile değişkenlerin sıralanması yaklaşımı sunmuştur. Ayrıca filtre tabanlı öznitelik algoritmalarının öznitelik alt kümesi seçimi için de önemli ölçüde daha iyi olan değişken alt küme belirleme ve değişken sayısının ilk yüzde seçim olmak üzere iki farklı yöntemin de kullanımı anlatılmıştır. Bu yönüyle çalışma literatüre yenilikçi bir yaklaşım önermiştir. Önerilen yaklaşım sınıflandırmada kullanılan iki

gerçek veri seti üzerinde test edilmiş sonuçlar detaylı olarak incelenmiştir. Daha sonraki çalışmalar ve araştırmacılar için, farklı kriterler ve farklı ÇKKV yöntemleri kullanılarak sınıflandırma, kümeleme ve diğer çok değişkenli istatistiksel yöntemlerde de kullanılabilecek referans bir çalışma olması düşünülmektedir.

## Ek Bilgiler/Yazar Beyanları


| | |
|---|---|
| **Araştırma ve Yayın Etiği Beyanı** | Çalışma, etik kurul onayı gerektirmemektedir. |
| **Çıkar Çatışması** | Yazar(lar) açısından ya da üçüncü taraflar açısından çalışmadan kaynaklı çıkar çatışması bulunmamaktadır. |
| **Teşekkür veya Destek Beyanı** | Bu araştırmayı desteklemek için dış fon kullanılmamıştır. |
| **Yazar Katkıları** | Yazar 1'in makaleye katkısı %100'dür. |


## Kaynakça


Akbulut, O. Y. (2020). Gri entropi temelli PSI ve ARAS ÇKKV yöntemleriyle Türk mevduat bankalarının performans analizi. *Finans Ekonomi ve Sosyal Araştırmalar Dergisi*, *5*(2), 171-187.

Akmaz, D. (2022). Stockwell Dönüşümü, ONE-R özellik seçme yöntemi ve rastgele orman algoritması ile güç kalitesi bozulumu sinyallerinin sınıflandırılması. *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, *34*(1), 267-276.

Akoğul, S., Erişoğlu, M., & Erişoğlu, Ü. (2020). Çok değişkenli normal dağılımların karmasına dayalı kümelemede TOPSIS yöntemiyle ile küme sayısının belirlenmesi. *Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi*, *36*(3), 472-480.

Allaire, J. (2012). *RStudio: Integrated development environment for R*. Boston, MA, *770*(394), 165-171.

Altıntaş, F. F. (2020). İnovasyon performanslarının Entropi tabanlı gri ilişkisel analiz yöntemi ile değerlendirilmesi: G7 Grubu Ülkeleri Örneği. Adnan *Menderes Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, *7*(2), 151-172.

Bircan, H. (2020). *Çok kriterli karar verme problemlerinde kriter ağırlıklandırma yöntemleri*. Nobel Yayınevi.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32.

Budak, H. (2018). Özellik seçim yöntemleri ve yeni bir yaklaşım. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *22*, 21-31.

Bulut, C., Altuğlu, T. B., & Yetkin, E. F. (2023). Filtre modelli öznitelik seçim algoritmalarının EEG tabanlı beyin bilgisayar arayüzü sistemindeki karşılaştırmalı sınıflandırma performansları. *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, *38*(4), 2397-2408.

Dash, M., & Liu, H. (1997). Feature selection for classification. Intelligent data analysis, *1*(1-4), 131-156. https://doi.org/10.1016/S1088-467X(97)00008-5

Ecemiş, O., Akçan, F., & Abakay, U. (2021). Analysis of the performance of football teams via the entropy-gray relational analysis method: Turkish super league model. *SPORMETRE Beden Eğitimi ve Spor Bilimleri Dergisi*, *19*(3), 51-59. https://doi.org/10.33689/spormetre.854446

Emhan, Ö., & Mehmet, A. (2019). Filtreleme tabanlı öznitelik seçme yöntemlerinin anomali tabanlı ağ saldırısı tespit sistemlerine etkisi. *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, *10*(2), 549-559. https://doi.org/10.24012/dumf.565842

Ersoy, N. (2018). Entropy tabanlı bütünleşik ÇKKV yaklaşımı ile kurumsal sürdürülebilirlik performans ölçümü. *Ege Akademik Bakış Dergisi*, *18*(3), 367-385. https://doi.org/10.21121/eab.2018339487

Eryılmaz, E. E., Şahin, D. Ö., & Kılıç, E. (2020). Türkçe istenmeyen e-postaların farklı öznitelik seçim yöntemleri kullanılarak makine öğrenmesi algoritmaları ile tespit edilmesi. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*, *13*(2), 57-77.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157-1182.

Hashemi, A., Dowlatshahi, M. B., & Nezamabadi-Pour, H. (2020). MFS-MCDM: Multi-label feature selection using multi-criteria decision making. *Knowledge-Based Systems*, *206*, 106365.

Hashemi, A., Dowlatshahi, M. B., & Nezamabadi-Pour, H. (2021). VMFS: A VIKOR-based multi-target feature selection. *Expert Systems With Applications*, *182*, 115224.

Hashemi, A., Dowlatshahi, M. B., & Nezamabadi-Pour, H. (2022). Ensemble of feature selection algorithms: A multi-criteria decision-making approach. *International Journal of Machine Learning and Cybernetics*, *13*(1), 49-69.

Haseki, M. İ., & Avşar, İ. İ. (2023). Avrupa Birliği ve seçili ülkelerinin teknoloji üretim odaklı verilerinin entropi ve gri ilişkiler analiz modelleriyle incelenmesi. *Uluslararası İktisadi ve İdari İncelemeler Dergisi*, *39*, 154-169. https://doi.org/10.18092/ulikidince.1214069

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, *11*, 63-91. https://doi.org/10.1023/A:1022631118932

Hornik, K., Buchta, C., Hothorn, T., Karatzoglou, A., Meyer, D., Zeıleıs, A., & Hornik, M. K. (2023). *Package 'RWeka'*.

Karaatlı, M. (2016). Entropi-Gri ilişkisel analiz yöntemleri ile bütünleşik bir yaklaşım: Turizm sektöründe uygulama. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, *21*(1), 63-77.

Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, *2*(2), 271-277.

Kaynar, O., Arslan, H., Görmez, Y., & Işık, Y. E. (2018). Makine öğrenmesi ve öznitelik seçim yöntemleriyle saldırı tespiti. *Bilişim Teknolojileri Dergisi*, *11*(2), 175-185. https://doi.org/10.17671/gazibtd.368583

Ladha, L., & Deepa, T. (2011). Feature selection methods and algorithms. I*nternational Journal on Computer Science and Engineering*, *3*(5), 1787-1797.

Lichman, M. (2013). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml adresinden 6 Haziran 2023 tarihinde alınmıştır.

Liu, H., & Setiono, R. (1995). Chi2: Feature selection and discretization of numeric attributes. In *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence* (p. 388-391). IEEE.

Lu, H. S., Chang, C. K., Hwang, N. C., & Chung, C. T. (2009). Grey relational analysis coupled with principal component analysis for optimization design of the cutting parameters in high-speed end milling. *Journal of materials processing technology*, 209(8), 3808-3817. https://doi.org/10.1016/j.jmatprotec.2008.08.030

Onan, A., & Korukoğlu, S. (2016). *Metin sınıflandırmada öznitelik seçim yöntemlerinin değerlendirilmesi*. Akademik Bilişim.

Özdemir, O., & Kılıçarslan, Ş. (2021). Entropi temelli gri ilişkisel analiz tekniği ile hayat ve emeklilik şirketlerinin finansal performansları üzerine bir araştırma. *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, *26*(4), 413-434.

Quinlan, J. R. (1993). *Programs for machine learning*. Morgan Kaufmann Publishers.

Romanski, P., Kotthoff, L., & Kotthoff, M. L. (2013). *Package "FSelector"*. http://cran/r-project. org/web/packages/FSelector/index.html

Sağbaş, E. A. (2023). Filtre tabanlı öznitelik seçim yöntemleri kullanılarak metinlerde duygu sınıflandırması üzerine karşılaştırmalı bir çalışma. *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, *35*(1), 239-250. https://doi.org/10.35234/fumbd.1195908

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Song, H., Chen, X., Zhang, S., & Xu, L. (2022). Multi-objective optimization design of 6-ups parallel mechanism based on Taguchi method and entropy-weighted gray relational analysis. *Applied Sciences*, *12*(12), 5836. https://doi.org/10.3390/app12125836

Şekkeli, F. E., & Güçlü, F. (2023). Katılım bankalarının finansal performanslarının entropi tabanlı gri ilişkisel analiz (GİA) yöntemiyle değerlendirilmesi. *TESAM Akademi Dergisi*, *10*(2), 489-511. https://doi.org/10.30626/tesamakademi.1253985

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2013). Data mining with big data. *IEEE transactions on knowledge and data engineering*, *26*(1), 97-107.

Yıldırım, B. F., & Önder, E. (2015). *Çok kriterli karar verme yöntemleri*. Dora Basım.