İstatistikçiler Dergisi: İstatistik & Aktüerya

Journal of Statisticians: Statistics and Actuarial Sciences

IDIA 16, 2023, 2, 100-115

Geliş / Received: 05.12.2023, **Kabul** / Accepted: 29.12.2023

Araştırma Makalesi / Research Article

Veri madenciliği yöntemleri ile bir melez sınıflandırma yaklaşımı ve uygulaması

Gözde Ulu Metin[1]

Ankara Üniversitesi, Fen Bilimleri Enstitüsü,
İstatistik Anabilim Dalı,
Ankara, Türkiye
ulumetin@ankara.edu.tr
ORCID: 0000-0003-0384-9504

Özlem Türkşen

Ankara Üniversitesi, Fen Fakültesi,
İstatistik Bölümü,
Ankara, Türkiye
turksen@ankara.edu.tr
ORCID: 0000-0002-5592-1830

Öz

Son yıllarda hızla artan büyüklükteki veri setlerinden bilgi keşfetmek oldukça değerlidir. Veri madenciliği yöntemleri, sınıflandırma problemlerinde, büyük ve karmaşık veri setlerindeki gizli örüntünün ortaya çıkarılarak verilerin belli bir sınıfa atanması amacıyla kullanılır. Bu çalışmada, kurumların başarım değerlendirilmesi sürecine istatistiksel bakış açısı kazandırmak amacıyla veri madenciliği yöntemleri ile Analitik Hiyerarşi Süreci (AHP) ve CODAS yöntemleri kullanılarak bir melez sınıflandırma yaklaşımı önerilmiştir. Uygulama amacıyla bir kurum verisi ele alınmıştır. Veri seti ön işleme aşamasından geçirilerek, veri setindeki değişkenler, uzman bilgisi dikkate alınarak AHP yöntemi ile ağırlıklandırılmıştır. Ağırlıklandırılmış gerçek veri setine, veri madenciliği sınıflandırma yöntemlerinden Lojistik Regresyon (LR), K-En Yakın Komşu (KNN) algoritması, Destek Vektör Makineleri (SVM) ve Rastgele Orman (RF) algoritması uygulanmıştır. Sınıflandırma yöntemleri, 5-kat çapraz doğrulama sonucu elde edilen doğruluk, kesinlik, duyarlılık ve F_1-skor performans ölçütlerine göre hesaplanmıştır. Elde edilen performans ölçütleri, çok ölçütlü karar verme yöntemi olan CODAS'a göre değerlendirilmiştir. Yapılan melez sınıflandırma yaklaşımına göre, Ar-Ge ve Tasarım merkezlerinin faaliyetlerinin değerlendirilmesi konusunda RF yönteminin daha iyi sınıflandırma performansına sahip olduğu görülmüştür.

Anahtar sözcükler: AHP, CODAS, Çok ölçütlü karar verme, Melez sınıflandırma, Performans ölçütleri, Veri madenciliği yöntemleri, Veri ön işleme

[1] Bu çalışma, birinci yazarın, ikinci yazarın danışmanlığında hazırladığı doktora tezinden üretilmiştir.

Abstract

A hybrid classification approach with data mining methods and an application

In recent years, it is very valuable to discover information from data sets of rapidly increasing size. Data mining methods are used in classification problems to assign data to a certain class by revealing the hidden pattern in large and complex data sets. In this study, a hybrid classification approach is proposed by using data mining methods with Analytic Hierarchy Process (AHP) and CODAS methods in order to gain a statistical perspective on the performance evaluation process of the instutitions. An instutition data is taken as a basis for the application. The data set is preprocessed and the variables in the data set are weighted by AHP method by taking into account expert knowledge. Logistic Regression (LR), K-Nearest Neighbour (KNN) algorithm, Support Vector Machines (SVM) and Random Forest (RF) algorithm, data mining classification methods, were applied to the weighted real data set. The classification methods were calculated according to the accuracy, precision, sensitivity and F_1-score performance measures obtained from 5-fold cross-validation. The obtained performance criteria were evaluated according to the CODAS, a multi-criteria decision making method. As a result of the hybrid classification approach, it was seen that the RF method has better classification performance about the evaluation of the activities of R&D and Design centers.

Keywords: *AHP, CODAS, Multi criteria decision making, Hybrid classification, Performance metrics, Data mining methods, Data preprocessing*

1. Giriş

Günümüzde depolanmış veri setlerinde mevcut olan ve saklı kalan bilgileri ortaya çıkarmak oldukça kritik bir rol oynar. Fakat, oluşan büyük veri yığınlarında geleneksel veri analizi yöntemleri yetersiz kalmaktadır. Veri madenciliği, ham veride bulunan örüntüleri ortaya çıkarmak ve keşfetmek adına, özellikle sınıflandırma problemleri üzerinde etkili bir biçimde kullanılır. Veri madenciliği yöntemleri ile sınıflandırma yapılmadan önce keşfedici veri analizi ile veri ön işleme aşamalarının uygulanması gerekir. Çetin ve Yıldız [1] çalışmalarında, literatürde bulunan çok sayıda veri ön işleme yöntemleri ve algoritmaları üzerine kapsamlı bir inceleme yapmışlardır. Emeç ve Özcanhan [2] çalışmalarında, veri ön işleme yöntemlerini ayrıntılı olarak ele alıp yapılan uygulama sonucunda, veri ön işlemenin karar vermede daha doğru sonuçlar elde edilmesine yardımcı olduğunu belirtmişlerdir.

Veri ön işleme aşamasında, veri setinin yapısını anlamak, boyut azaltmak ya da gruplamak amacıyla farklı yöntemler kullanılabilir. Veri setindeki değişkenlerin önem ağırlıklarının hesaplanması ve subjektif değerlendirmelerden yararlanılması amacıyla Çok Ölçütlü Karar Verme (Multi Criteria Decision Making-MCDM) yöntemlerinin kullanılması da veri ön işleme sürecine dahil edilebilir. Bu çalışmada, bir MCDM yöntemi olan Analitik Hiyerarşi Süreci (Analytic Hierarchy Process-AHP) yöntemi ile uzman görüşü dikkate alınarak değişkenlerin önemine göre değişken ağırlıkları belirlenmiştir. Böylece, veri ön işleme sürecinde, MCDM yöntemleri kullanılarak veri setindeki değişkenlerin önemi hakkında önsel bilgi elde edinilmesi sağlanmıştır.

Veri madenciliği, veri yığınlarında veriye dayalı derinlemesine keşifler yapmayı, istatistiksel yöntemler ile örtülü bilgileri, veriden çıkarmayı amaçlar [3]. Han vd. [4] çalışmalarında, veri madenciliği kavramı ve yöntemleri detaylı olarak anlatılarak örneklerle açıklanmıştır. Çınar ve Silahtaroğlu [5] çalışmalarında, bir anket veri seti üzerinde veri madenciliği yöntemlerini uygulayarak, gizli kalmış örüntü ve nedenleri keşfetmişlerdir.

Veri madenciliği yöntemleri, denetimli öğrenme (supervised learning) ve denetimsiz öğrenmeden (unsupervised learning) oluşur. Denetimli öğrenme, sınıflandırma yöntemlerini, denetimsiz öğrenme de kümeleme ve birliktelik kurallarını içermektedir. Sınıflandırma çalışması yapılması istenen ön işlemesi yapılmış veri setinde, veri madenciliğinin denetimli öğrenme başlığı altında sınıflandırma algoritmaları kullanılır. Sınıflandırma performansının ölçülmesinde, Doğruluk (Accuracy), Duyarlılık (Sensivity),

Kesinlik (Precision) ve F_1-Skor (F_1-Score) ölçütleri hesaplanır. Nieto vd. [6] çalışmalarında, stratejik karar vermede denetimli sınıflandırma yöntemlerini kullanmışlardır. Gerçek veri seti ile yapılan çalışmada, performans ölçütlerine göre RF algoritmasının diğer yöntemlere göre daha iyi performansa sahip olduğu görülmüştür. Öztürk Zan [7] çalışmasında, gerçek veri seti üzerinde denetimli öğrenme algoritmasını uygulamıştır. 5-kat çapraz doğrulama (Cross Validation-CV) sonucu elde edilen performans ölçütlerine göre RF algoritmasının en iyi performansı gösterdiği belirtilmiştir. Yavuz vd. [8] çalışmalarında, karar destek sisteminin geliştirilmesi amacıyla, sınıflandırma yöntemlerinden Naive Bayes, KNN, Karar Ağacı ve RF kullanılarak performansları değerlendirilmiştir.

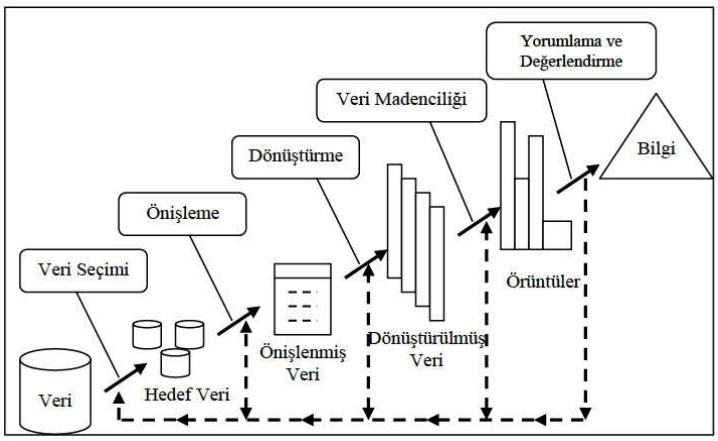
Yapılan çalışmalarda, çok ölçüte sahip karar seçenekleri arasından birine karar vermek oldukça zordur. MCDM, belirlenen bir amacı gerçekleştirmek için mevcut seçenekler arasından, belirlenen ölçütler dikkate alınarak en uygun olanın seçimine karar verme sürecidir. Karar verme sürecinde, MCDM yöntemlerinden CODAS (Combinative Distance-based Assesment) uygulanarak, sınıflandırma yöntemlerinin öncelikli sıralaması yapılır. Ulutaş [9] çalışmasında, bir tekstil şirketi için lojistik sağlayıcı seçiminde AHP ve CODAS yöntemlerini birlikte kullanmıştır. AHP ile elde edilen kriter ağırlıkları CODAS yöntemi ile birleştirilerek alternatifler sıralanmıştır. Can vd. [10] çalışmalarında, sağlık sektöründe altı sigma projelerinin önceliklendirilmesi ve seçimi için AHP ve CODAS yöntemlerini entegre eden bir hibrit karar verme modeli önermektedir. AHP yöntemi ile elde edilen kriter ağırlıkları kullanılarak CODAS uygulanmıştır.

Bu çalışmada, sınıflandırma yöntemlerinin MCDM yöntemleri ile ağırlıklandırılarak performans ölçütlerinin hesaplandığı ve performans ölçütlerinin değerlendirilmesinde MCDM yöntemi kullanılarak karar verilmesi konusunda, veri madenciliği yöntemleri ile MCDM yöntemleri kullanılarak bir melez sınıflandırma yaklaşımı önerilmiştir. Çalışmanın ikinci bölümünde, veri ön işleme ile veri setini etkileyen değişken ağırlıklarının hesaplanmasında kullanılan AHP yöntemine yer verilmiştir. Çalışmanın üçüncü bölümünde, veri madenciliği ile sınıflandırma başlığında Lojistik Regresyon (LR), K-En Yakın Komşu (KNN) algoritması, Destek Vektör Makineleri (SVM) ve Rastgele Orman (RF) algoritması açıklanmıştır. Sınıflandırma performans ölçütleri tanımlanarak performans ölçütlerinin karar verilmesinde kullanılan CODAS yöntemi anlatılmıştır. Çalışmanın dördüncü bölümünde, bir gerçek veri seti üzerinde uygulama yapılmıştır. Çalışmanın beşinci bölümünde ise sonuçlara yer verilmiştir.

2. Veri ön işleme

Bilgi teknolojisinin gelişimi ile verinin depolanma kapasitesi artarak büyük veri (big data) yapıları oluşmuştur. Elde edilen her yeni veri saklanmakta fakat, oluşan veri yığınından anlamlı bilgi çıkarmak zorlaşmaktadır. Veriden bilgi elde edilmesi süreci, bilgi keşfi olarak adlandırılır. Fayyad [11] çalışmasında, bilgi keşfi sayesinde oluşan veri yığınlarının etkili bir biçimde kullanılmasıyla değer elde edilmesinin önemi vurgulanmıştır. Kavurkacı vd. [12] çalışmalarında büyük veri işlemede kullanılan yöntemlere genel bir bakış açısı sunmuştur.

Veri madenciliği kavramı, yüksek kapasiteli verinin içerisindeki keşfedilmemiş bilgiyi ortaya çıkarmayı hedefler [13]. Veriden bilgi keşfi süreci olarak adlandırılan bu süreç aşamaları Şekil 1'de özetlenmiştir. Birinci aşamada araştırma konusuna yönelik elde edilen ham veriden, hedeflenen veri seçimi yapıldığı Şekil 1'den açıkça görülmektedir. Verideki bilgi keşfi sürecinin ön işleme aşamasında (veri temizleme, veri bütünleştirme, boyut azaltma, veri seçme) istatistiksel yaklaşımlara dayalı veri analizi yapılarak, hedef veride, eksik değer tamamlanarak hatalı, anlamsız değerler çıkartılır. Eğer, veri dönüşüm gerektiriyorsa, dönüştürme işlemi yapılarak dönüştürülmüş verilere ulaşılır.



Şekil 1. Veri-Bilgi Keşfi Süreci [18]

Çizelge 1'de verilen veri setindeki değişkenlerin ölçüldüğü birimler arasındaki farklılıkların giderilmesi amacıyla Z-skor, Min-Max gibi standartlaştırma yöntemleri kullanılarak veri seti, ölçü biriminden bağımsız hale getirilir. Veri hazırlama aşamasından sonra büyük verilerin analizini kolaylaştıran, gizli örüntü keşfini sağlayan temeli istatistiksel yöntemlere dayalı MCDM yöntemleri kullanılır. Keleş ve Tunca [14] çalışmalarında, bir Ar-Ge firmasının kuruluş aşamasında, işletmelerin görüşüne göre önemli değişkenlerin AHP yöntemi ile ağırlıklarını belirlemişlerdir. Arslan ve Belgin [15] çalışmalarında, AHP yöntemini, imalat sanayisindeki öncelikli teknoloji alanlarını etkileyen değişkenlerin ağırlıklandırılmasında kullanmıştır. Çalışma sonucunda elde edilen sıralamaya göre ilgili sektörlere sağlanan Ar-Ge, yenilik ve girişimcilik desteklerinde öncelik tanınması önerilmiştir. Güryeli [16] çalışmasında, Bilim, Sanayi ve Teknoloji Bakanlığı (günümüzde Sanayi ve Teknoloji Bakanlığı) tarafından yürütülen Teknolojik Ürün Yatırım Destek Programı desteği için sunulan Ar-Ge projelerinin seçim sürecini incelemiştir. Bu süreçte, alanında uzman akademisyenlerin görüşlerine göre AHP uygulanarak Ar-Ge projelerinin değerlendirilmesinde dikkate alınan değişkenlerin göreceli önem seviyeleri elde edilmiştir. Subjektif değerlendirmelerden yararlanılarak, veri setindeki değişkenlerin önem ağırlıklarının hesaplanması amacıyla AHP yöntemi uygulanır.

Çizelge 1. Çok değişkenli veri seti

	Bağımsız değişkenler				***Bağımlı değişken***
No.	X_1	X_2	...	X_m	Y
1	x_{11}	x_{12}	...	x_{1m}	Y_1
2	x_{21}	x_{22}	...	x_{2m}	Y_2
⋮	⋮	⋮	⋱	⋮	⋮
n	x_{n1}	x_{n2}	...	x_{nm}	Y_n

AHP yöntemi, değişkenlerin ikili kıyaslamasını yaparken sözel ifadeleri sayısal değerler kullanarak ifade edip karşılaştırma yapan MCDM yöntemlerinden biridir [17]. AHP'de, Saaty tarafından önerilmiş olan karşılaştırma ölçeği ile subjektif değerlendirmeler, matrisler yardımıyla matematiksel olarak ifade edilir [18]. AHP uygulamak için Çizelge 1'de görülen m sayıda değişken, Saaty [18] çalışmasında önerilen karşılaştırma ölçeğiyle birbirlerine göre önem değerleri dikkate alınarak

$$A=\begin{bmatrix} 1 & a_{12} & \cdots & a_{1m} \\ 1/a_{12} & 1 & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1/a_{1m} & 1/a_{2m} & \cdots & 1 \end{bmatrix} \tag{1}$$

biçiminde bir karşılaştırma matrisi oluşturulur. Eşitlik (1) ile tanımlı karşılaştırma matrisinin her bir elemanı

$$a'_{ij} = \frac{a_{ij}}{\sum_{j=1}^{m} a_{ij}}, \ i = 1, 2, \ldots, m \tag{2}$$

olacak biçimde normalleştirilir**.** Normalleştirilmiş karşılaştırma matrisinin satır ortalamaları değişkenlerin önem ağırlıkları olup

$$w_j = \frac{1}{m} \sum_{j=1}^{m} a'_{ij}, \ i = 1, 2, \ldots, m \tag{3}$$

olur. Elde edilen değişken ağırlıkları ile veri seti Çizelge 2'deki biçimde ağırlıklandırılır.

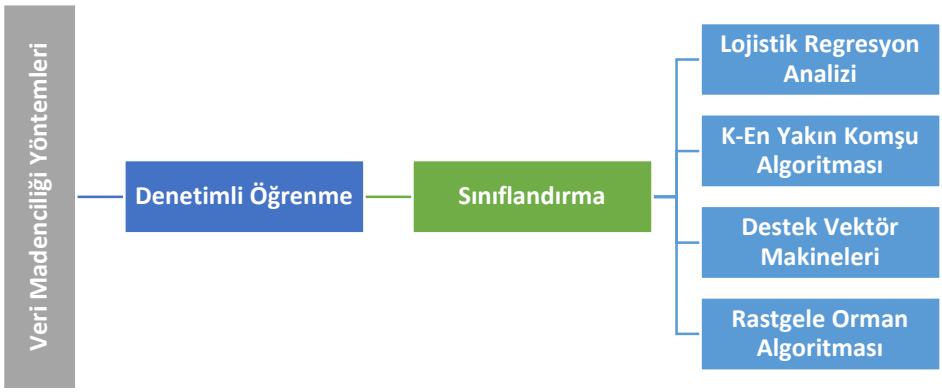
Çizelge 2. AHP ile ağırlıklandırılmış çok değişkenli bir veri seti

	Bağımsız değişkenler				***Bağımlı değişken***
No.	w_1X_1	w_2X_2	...	w_mX_m	Y
1	w_1x_{11}	w_2x_{12}	...	w_mx_{1m}	Y_1
2	w_1x_{21}	w_2x_{22}	...	w_mx_{2m}	Y_2
⋮	⋮	⋮	⋱	⋮	⋮
n	w_1x_{n1}	w_2x_{n2}	...	w_mx_{nm}	Y_n

Ağırlıklandırılmış veri setinin elde edilmesi ile veri ön işleme süreci tamamlanır. İstatistiksel bakış açısı ile daha esnek hesaplama kolaylığı sağlayan veri madenciliği sınıflandırma yöntemleri uygulanarak veri yapısına uygun biçimde analiz yapılır.

3. Veri madenciliği sınıflandırma yöntemleri

Veri madenciliği sınıflandırma yöntemleri, verideki gizli örüntülerin ortaya çıkarılarak gözlemlerin hangi sınıfa ait olduğunun tahmin edilmesini sağlayan denetimli öğrenme yöntemleridir. Bağımlı değişkenin kategorik ve bağımsız değişkenlerin sürekli ve kategorik değişkenlerden oluştuğu veri setleri için Şekil 3'te verilen veri madenciliği sınıflandırma yöntemleri kullanılmaktadır.



Şekil 3. Veri madenciliği sınıflandırma yöntemleri

Veri madenciliği sınıflandırma yöntemlerinin uygulanabilmesi için, yöntemlere ilişkin ayarlanabilir parametrelerin (tuning parameters) ilgilenilen veri setine yönelik uygun biçimde belirlenmiş olması gerekir. Ayarlanabilir parametrelerin seçimi yöntemlerin sınıflandırma performansında etkindir. Uzman görüşü alınarak da belirlenen bu parametrelerin seçimi önemlidir.

3.1. Lojistik Regresyon

Lojistik regresyon (LR), veri setindeki gözlemlerin bir sınıfa ait olup olmama olasılığını hesaplayarak yeni gelecek gözlemleri sınıflandıran bir veri madenciliği yöntemidir. $Y_i \in \{0,1\}$ olmak üzere $\{(X_i, Y_i)\}_{i=1}^{n}$, etiketli bir veri seti olsun. Bağımlı değişken kategorik olduğundan dolayı bağımlı değişkenin olasılığı hesaplanıp lojit dönüşüm yapılarak lojistik regresyon modeli

$$f(x_i) = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{X}_i \boldsymbol{\beta})}},\ i = 1, 2, \ldots, n \tag{4}$$

biçiminde elde edilir. Burada

$$\beta_0 + \boldsymbol{X}_i \boldsymbol{\beta} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{nm},\ i = 1, 2, \ldots, n \tag{5}$$

dir. Model parametrelerinin tahmin edilmesinde maksimum olabilirlik yöntemi kullanılarak tahmin olasılıklarının maksimize edilmesi istenir [19]. Eşitlik (4) bir olasılık değeri olup, eşik değeri 0.5 alındığında sınıflandırma

$$\begin{aligned} f(x_i) \geq 0.5,\ &\Rightarrow \hat{Y}_i = 1 \\ f(x_i) < 0.5,\ &\Rightarrow \hat{Y}_i = 0 \end{aligned} \tag{6}$$

biçiminde yapılır. Lojistik regresyonun ayarlanabilir parametreleri, kısıtlama yapılacak yöntemi (Lasso, Ridge, None) belirleyen ceza değeri (penalty) ve kısıtlama oranı (c)'dır.

3.2. K-En Yakın Komşu Algoritması

K-En Yakın Komşu Algoritması (KNN), uygulaması kolay ve anlaşılır sınıflandırma yöntemlerinden biridir. KNN yönteminde, sınıfı belli olan gözlemlerden yararlanılarak yeni katılacak gözlemin hangi sınıfa ait olup olmadığı belirlenir. Bu yöntemde, gözlemlerin her biri ile yeni gelecek gözlem arasındaki uzaklıklar hesaplanarak en küçük uzaklığa sahip *k* sayıda gözlemin seçimi yapılır. Hesaplanan uzaklık değerleri arasında en çok tekrar eden sınıf, yeni gözlem değerinin sınıfıdır. Gözlemler arasındaki uzaklıklar hesaplanırken yaygın olarak Öklid uzaklık formülü kullanılır. *m* değişken sayılı veri setindeki *i*. ve *j*. gözlemler arasındaki uzaklık

$$d_{ij} = \sqrt{\sum_{K=1}^{m} (x_{ik} - x_{jk})^2},\ \ i, j = 1, 2, \ldots, m,\ i \neq j \tag{7}$$

biçiminde hesaplanır. Burada, *k* ayarlanabilir parametresinin belirlenmesi önemlidir. En küçük uzaklığın elde edilmesi amaçlanır.

3.3. Destek Vektör Makineleri

Destek vektör makineleri (SVM) veri madenciliği yöntemlerinden biridir [20]. Bu yöntem, veriyi doğrusal olarak iki sınıfa ayırabilmek için en uygun fonksiyonun (hiperdüzlemin) tahmin edilmesi esasına dayanır. $Y_i \in \{-1,1\}$ olmak üzere $\{(X_i, Y_i)\}_{i=1}^{n}$, etiketli bir veri seti olsun. Nokta çarpımları Eşitlik (8) biçiminde ifade edilir.

$$< \boldsymbol{\beta},\ \boldsymbol{X} > = \boldsymbol{\beta} * \boldsymbol{X} = \boldsymbol{\beta}^T \boldsymbol{X} = \sum_{j=1}^{m} \beta_j X_j \tag{8}$$

İki sınıf arasındaki karar sınırı olan hiperdüzlem denklemi

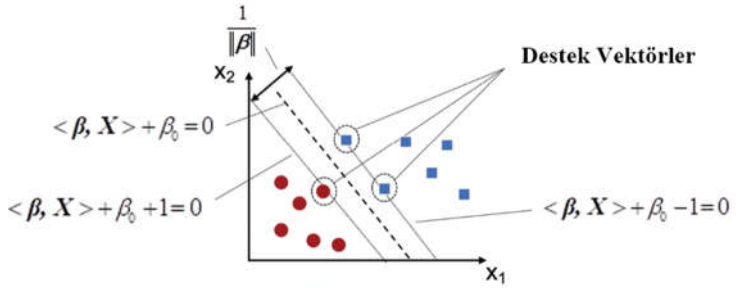
$$H\ düzlemi: <\boldsymbol{\beta},\ \boldsymbol{X}> + \beta_0 = 0 \tag{9}$$

biçiminde yazılır.

Şekil 4'te verilen $\boldsymbol{\beta}$ ağırlık vektörü, hiperdüzleme dik yönde olup eğimi belirlemektedir. Burada, β_0 sabit terimdir. H düzlemi optimal hiperdüzlem olup H_1 düzlemi ve H_2 düzlemi

$$\begin{aligned} H_1\ düzlemi &: <\boldsymbol{\beta},\ \boldsymbol{X}> + \beta_0 - 1 = 0 \\ H_2\ düzlemi &: <\boldsymbol{\beta},\ \boldsymbol{X}> + \beta_0 + 1 = 0 \end{aligned} \tag{10}$$

biçiminde olur. H_1 düzlemi ve H_2 düzlemi üzerindeki gözlemler, sınırları belirleyen destek vektörleri (support vectors) olarak tanımlanmaktadır. Destek vektörlerinin seçimi yapılırken H_1 düzlemi ve H_2 düzlemleri arasındaki mesafenin en büyük olması istenmektedir. Bu mesafeye kenar payı (marjin) adı verilmekte olup kenar payı $2d = \frac{2}{\|\boldsymbol{\beta}\|}$ değerinin en büyük yapılması $\min \frac{1}{2}\|\boldsymbol{\beta}\|^2$ amaçlanmaktadır.



Şekil 4. İki sınıflı veri setinin SVM yöntemi ile sınıflandırılması

SVM problemi

$$\begin{aligned} &\min_{\beta}\ f(\boldsymbol{\beta}) = \frac{1}{2}\langle \boldsymbol{\beta}, \boldsymbol{\beta} \rangle \\ &g(\boldsymbol{\beta},\ \beta_0) = Y_i\left(<\boldsymbol{\beta}, \boldsymbol{X}_i> + \beta_0\right) - 1 \geq 0,\ \ i = 1,2,..,n \end{aligned} \tag{11}$$

biçimde eşitsizlik kısıtlı optimizasyon problemi yazılır. Yeni gelen gözlemin sınıflandırılmasında

$$Y' = \text{sgn}(\boldsymbol{\beta}^T \boldsymbol{X} + \beta_0) \tag{12}$$

ifadesi kullanılır [21,22]. Doğrusal sınıflamanın mümkün olmadığı ya da değişken sayısının fazla olduğu durumlarda, SVM yöntemi çekirdek fonksiyonları kullanarak sınıflandırma yapar [23]. Çekirdek fonksiyonu

$$K\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) = <\phi_{X_i}, \phi_{X_j}> = \left(\left\langle \boldsymbol{X}_i, \boldsymbol{X}_j \right\rangle\right)^2 \tag{13}$$

biçiminde tanımlanır. Herhangi bir doğrusal sınıflandırma probleminin amaç fonksiyonunda $\left\langle \boldsymbol{X}_i, \boldsymbol{X}_j \right\rangle$ biçiminde vektörlerin iç çarpımı yer alıyorsa, bu ifade yerine uygun bir $K\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right)$ çekirdek fonksiyonu yazılarak problem güncellenir. Çekirdek fonksiyonların seçimi, SVM performansını etkilemektedir. Yaygın olarak kullanılan çekirdek fonksiyonları Çizelge 3'te verilmiştir.

Çizelge 3. Yaygın olarak kullanılan çekirdek fonksiyonları

Doğrusal: $K\left(\boldsymbol{X}_i,\boldsymbol{X}_j\right)=\left(\left\langle \boldsymbol{X}_i,\boldsymbol{X}_j\right\rangle\right)$
Polinom: $K\left(\boldsymbol{X}_i,\boldsymbol{X}_j\right)=\left(\left\langle \boldsymbol{X}_i,\boldsymbol{X}_j\right\rangle+1\right)^m$
Dairesel (*Radyal*) *Tabanlı*: $K(\boldsymbol{X}_i,\boldsymbol{X}_j)=\exp(-\gamma\left\|\boldsymbol{X}_i-\boldsymbol{X}_j\right\|^2)$
Sigmoid: $K(\boldsymbol{X}_i,\boldsymbol{X}_j)=\tanh(\gamma\boldsymbol{X}_i^T\boldsymbol{X}_j+r)$

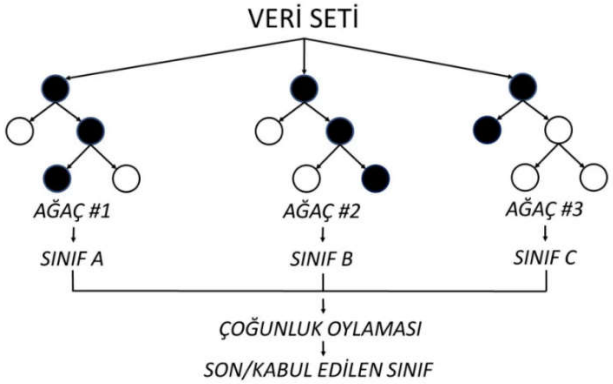
SVM için ayarlanabilir parametreler Çizelge 4'te verilmiştir.

Çizelge 4. SVM'de kullanılan ayarlanabilir parametreler

c: Kısıtlama oranı
Kernel: Çekirdek fonksiyon (*Polynomial*, *Rbf*, *Sigmoid*, *Linear*)
Gamma: *c* oranını düzenleyici değer

3.4. Rastgele Orman Algoritması

Rastgele Orman Algoritması (RF), verilen veri kümesinin çeşitli alt kümelerinde birden çok sayıda karar ağacı içeren sınıflandırma yöntemidir [23]. Bu yöntem, veri kümesinin tahmin doğruluğunu iyileştirmek, varyansı ve yanlılığı azaltmak amacıyla kullanılır. RF, tek bir karar ağacına güvenmek yerine Şekil 5'teki gibi her ağaçtan tahminleri toplar ve tahminlerin çoğunluğuna dayanarak nihai sonucu tahmin eder.



Şekil 5. RF Algoritması [24]

Bootstrap yöntemi ile eğitim verisinin 2/3'ü (In-bag(IB)) ile örneklemlerden ağaçlar oluşturulur. Eğitim veri setinin geriye kalan 1/3'ü (Out-of-bag(OOB)), ağaçların performans değerlendirmesi için hataların kestirim hesabında kullanılır. Her bir ağaç ikili bölünme yapısı ile alt düğümlere ayrılır. Oluşturulacak ağaç sayısı araştırmacı tarafından belirlenmektedir. Her bir örneklem için her düğümde *m* değişken arasından $s=\sqrt{m}$ adet değişken belirlenir. Her eğitim setinden elde edilen tahminlerin çoğunluk oylamasına göre sınıflandırma yapılır [21]. B oluşturulan ağaç sayısı olmak üzere $(b=1,2,\ldots,B)$, b. ağacın sınıf tahmini

$$\hat{C}_{RF}^{B}(x)=\text{çoğunluk oylaması}\left\{\hat{C}_b(x)\right\}_1^B \tag{14}$$

biçiminde hesaplanır. Performansın değerlendirilmesinde OOB hata oranı

$$E_{OOB}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{I}(Y_i\neq\hat{C}_{RF}^{B}(x)) \tag{15}$$

dir. OOB hata oranının minimum yapılması istenir. RF için ayarlanabilir parametreler Çizelge 5'te verilmiştir.

Çizelge 5. RF'de kullanılan ayarlanabilir parametreler

Maksimum derinlik (*Max dept*): Ardışık soru sayısı
Maksimum değişken (*Max features*): Her düğümde değerlendirilen değişken sayısı
Minumum yaprak örneği (*Min samples leaf*): Bir yapraktaki minimum gözlem sayısı
Minimum örneklem (*Min samples split*): Bir düğüm bölünmeden önce gerekli gözlem sayısı
n tahmin edici (*n estimators*): Oluşan ağaç sayısı (*B*)
Kriter (*Criterion*): İndeks hesaplama yöntemi (Gini, Entropy)

Uygulanan sınıflandırma yöntemlerinde, algoritmanın performansının karşılaştırılabilmesi için performans ölçütleri bulunmaktadır. Sınıflandırma performans ölçütleri temel alınarak MCDM ile sınıflandırma yöntemine karar verilir. Sınıflandırma yöntemlerinde, performans ölçütleri hesaplanmadan önce, veri seti, eğitim seti ve test seti (genellikle %80 eğitim seti ile %20 test seti) olmak üzere ikiye ayrılır. Veri seti *k* sayıda parçaya bölünür. Bölünen her bir parçadan birisi test diğer *k*-1 parça eğitim verisi olarak kullanılır. Eğitilen her bir parça test edilerek performans ölçütü hesaplanır. Hesaplanan *k* tane performans ölçütünün ortalamasıyla, *k*-kat Çapraz Doğrulama (*k*-fold Cross Validation-CV) ile performans ölçütleri elde edilir. Sınıflandırma yöntemlerinin performans ölçütlerinin hesaplanabilmesi amacıyla Çizelge 6'da verilen karışıklık matrisi kullanılmaktadır.

Çizelge 6. Karışıklık matrisi

		Tahmin Sınıfı	
	Sınıf	**Pozitif**	**Negatif**
Gerçek Sınıf	**Pozitif**	Doğru Pozitif (DP)	Yanlış Negatif (YN)
	Negatif	Yanlış Pozitif (YP)	Doğru Negatif (DN)

Doğruluk (Accuracy): Doğru olarak sınıflandırılmış örneklerin toplam örnek sayısına oranı

$$Doğruluk = \frac{DP + DN}{DP + DN + YP + YN} \tag{16}$$

ile hesaplanır.

Duyarlılık (Sensivity): Doğru olarak sınıflandırılmış pozitif örnek sayısının toplam pozitif örnek sayısına oranı olarak

$$Duyarlılık = \frac{DP}{DP + YN} \tag{17}$$

biçiminde hesaplanır.

Kesinlik (Precision): Pozitif olarak tahmin edilen değerlerin ne oranda doğru olarak tahmin edildiği ölçütü

$$Kesinlik = \frac{DP}{DP + YP} \tag{18}$$

ile hesaplanmaktadır.

F_1-skor (F_1-score): Kesinlik ve duyarlılık ölçütünün harmonik ortalaması olan ölçüt

$$F_1\text{-}skor = 2\frac{(Duyarlılık)\,x\,(Kesinlik)}{(Duyarlılık)+(Kesinlik)} \tag{19}$$

biçiminde hesaplanır. Sınıflandırma yöntemlerine karar verilebilmesi için sınıflandırma performans ölçütlerine göre MCDM yöntemlerinden CODAS uygulanacaktır.

CODAS yöntemi, 2016 yılında geliştirilen MCDM yöntemlerinden biridir [25]. Bu yöntemde alternatiflerin değerlendirilmesi negatif ideal çözüme olan uzaklıklara dayanır. Alternatiflerin negatif ideal çözüme olan uzaklıklarının hesaplanmasında Öklid ve Taxicab uzaklıkları kullanılır. CODAS işleyiş adımları aşağıdaki gibi tanımlanır.

Adım 1: CODAS yönteminde n tane sınıflandırma yöntemi (seçenekler), m tane performans ölçütleri (ölçütler) olacak şekilde $n \times m$ boyutlu bir karar matrisi Eşitlik (20)'de verilen biçimde belirlenir.

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nm} \end{bmatrix} \tag{20}$$

Adım 2: Normalleştirilmiş matris Eşitlik (21) kullanılarak hesaplanır.

$$d'_{ij} = \frac{d_{ij}}{maxd_{ij}},\ i = 1,2,\ldots,n,\quad j = 1,2,\ldots,m \tag{21}$$

Eşitlik (22)'de verilen V normalleştirilmiş karar matrisi elde edilir.

$$V = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1m} \\ v_{21} & v_{22} & \cdots & v_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nm} \end{bmatrix} \tag{22}$$

Performans ölçütlerini en büyük yapacak negatif ideal çözüm $V^- = \left\{v_1^-, v_2^-, \ldots, v_m^-\right\}$, V matrisinin sütunlarının en küçük değerleri ile elde edilir.

Adım 3: Öklid uzaklığı kullanılarak her bir yönteme ilişkin Eşitlik (23)'teki gibi negatif ideal çözüm değerleri uzaklıkları hesaplanır.

$$E_i = \sqrt{\sum_{j=1}^{m}(v_{ij} - v_j^-)^2},\ i = 1,2,\ldots,n \tag{23}$$

Adım 4: Taxicab uzaklığı kullanılarak her bir yönteme ilişkin Eşitlik (24)'teki gibi negatif ideal çözüm değerleri uzaklıkları hesaplanır.

$$T_i = \sum_{j=1}^{m}\left|v_{ij} - v_j^-\right|,\ i = 1,2,\ldots,n \tag{24}$$

Adım 5: Göreli değerlendirme matrisi Eşitlik (25)'teki gibi elde edilir.

$$h_{ik} = (E_i - E_k) + \varphi(E_i - E_k)x(T_i - T_k) \tag{25}$$

φ ile iki alternatif arasındaki Öklid uzaklık değeri için bir eşik fonksiyonu Eşitlik (26) ile elde edilir.

$$\varphi(x)=\begin{cases}0, & |x|<0.2\\ 1, & |x|\geq 0.2\end{cases} \tag{26}$$

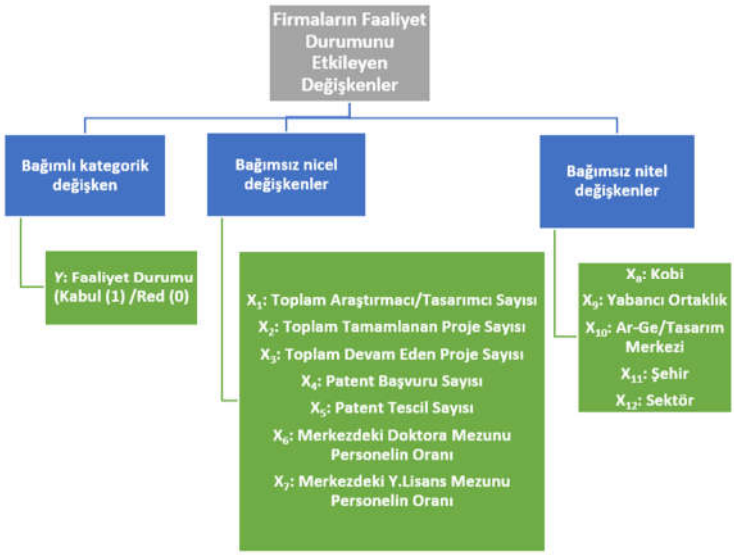
Adım 6: Değerlendirme puanlarının hesaplanması için Eşitlik (27) kullanılır.

$$H_i=\sum_{k=1}^{m} h_{ik} \tag{27}$$

Değerlendirme puanı büyük olan sınıflandırma yöntemi öncelikli olarak tercih edilir.

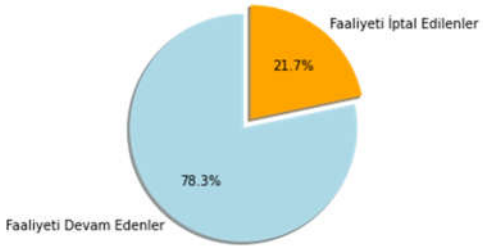
4. Sayısal uygulama

Bu çalışmada, Ar-Ge ve Tasarım merkezi belgesi alan firmaların çalışmalarına göre faaliyetlerinin devam ettirilmesi ya da iptal edilmesi (belge iptali) durumlarına (faaliyet sonucuna) göre sınıflandırılması istenmiştir. Buna göre, yetkililerin bilgisi dahilinde Ar-Ge ve Tasarım Merkezleri Daire Başkanlığı'ndan veri seti talep edilmiştir. Veri seti, 2008-2021 yılları arasında faaliyeti devam eden 2334 adet Ar-Ge ve Tasarım Merkezleri'ni kapsamakta olup uygun biçimde veri tabanından temin edilmiştir. Veri setinde her bir bilginin değerli olması ve bilgi kaybının istenmemesi nedeniyle veri setinde örnekleme yapılmadan verinin tamamı değerlendirilmiştir. Çalışma kapsamında, ilgilenilen veri setine, öncelikli olarak veri ön işleme uygulanmıştır. Aynı bilgiyi içeren değişkenler birleştirilip bazı değişkenlerin daha anlamlı olması adına oransal değişkenlere dönüştürülerek değişken seçimi ve boyut indirgeme yapılmıştır. Çalışmada yapılan analizler için Python 3.11.3 programı ve kütüphaneleri kullanılmıştır. İlgilenilen veri setine yönelik, bağımlı ve bağımsız değişkenler belirlenerek, değişkenlerin aldığı değerlere ilişkin kategorileri Şekil 6'da detaylı biçimde belirtilmiştir.



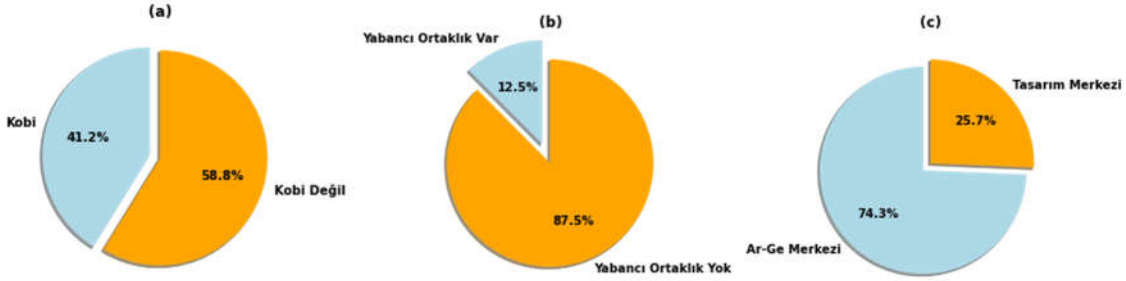
Şekil 6. Firmaların faaliyet durumunu etkileyen değişkenler

Şekil 6'dan, veri setinin bir kategorik bağımlı değişken, yedi nicel ve beş nitel değişken olmak üzere on iki (12) bağımsız değişkenden oluştuğu görülmektedir. Betimsel istatistikler kullanılarak veri seti hakkında özet bilgiler elde edilmiştir.



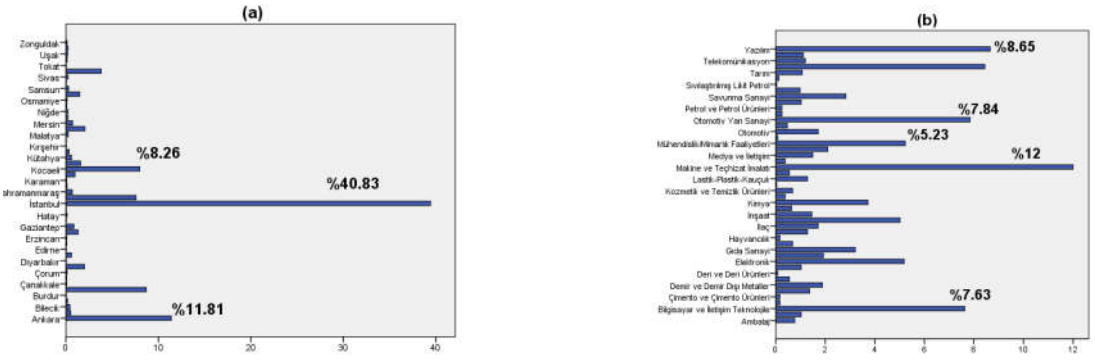
Şekil 7. Veri setinde kategorik bağımlı değişkene ait pasta grafiği

Şekil 7'ye göre Ar-Ge ve Tasarım merkezlerinin yaklaşık %78'inin faaliyetinin devam ettiği, yaklaşık %22'sinin ise faaliyeti iptal edilmiştir. Buna göre yanıt değişkeninin Y=1 olma olasılığı p=0.78 olarak da ifade edilebilir.



Şekil 8. (a) Kobi, (b) Yabancı Ortaklık ve (c) Ar-Ge/Tasarım Merkezi bilgisine ait pasta grafiği

Şekil 8'e göre, Ar-Ge ve Tasarım merkezlerinin %58.83'ü büyük ölçekli işletme sınıfında yer alırken, %41.17'si Kobi'dir. Ar-Ge ve Tasarım merkezlerinin %87.53'ünün yabancı ortaklığı bulunmazken, %12.47'sinin yabancı ortaklığı bulunmaktadır. Firmaların %73.34'ü Ar-Ge merkezi iken, %25.66'sı Tasarım merkezidir.



Şekil 9. Ar-Ge ve Tasarım merkezlerinin (a) Şehir ve (b) Sektör değişkenlerine göre yüzdelik değerleri

Şekil 9'a göre, Ar-Ge ve Tasarım merkezi kuran firmaların bulunduğu şehirlerin %40.83'ü İstanbul'da, %11.81'i Ankara'da, %8.26'sı Kocaeli'nde bulunmaktadır. Ar-Ge ve Tasarım merkezlerinin %12'si makine ve teçhizat imalatı sektöründe bulunurken, %8.65'i yazılım, %7.84'i ise otomotiv yan sanayi, %7.63'ü bilgisayar ve iletişim teknolojileri, %5.23 mühendislik/mimarlık sektöründe faaliyet göstermektedir.

Çizelge 8. Veri setindeki nicel değişkenlere ilişkin betimsel istatistikler

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
n	2334	2334	2334	2334	2334	2334	2334
min	1	0	0	0	0	0	0
Q_1	11	5	4	0	0	0	0.030
Q_2	15	12	6	0	0	0	0.080
Q_3	24	26	11	2	1	0	0.140
max	2710	1056	485	2127	693	0.43	0.710

Nicel değişkenlere ait minimum, maksimum ve çeyreklik değerleri Çizelge 8'de verilmiştir. Ar-Ge ya da Tasarım merkezi olan firmaların yarısında 15'ten fazla Araştırmacı/Tasarımcı personelin bulunduğu, en fazla Araştırmacı/Tasarımcı personeli bulunan firmada ise bu sayının 2710 olduğu, toplam tamamlanan ve toplam devam eden proje sayısı sıfır olan firmaların bulunduğu fakat bu firmaların faaliyette olmadığı, en fazla 2127 patent başvurusu yapıldığı, en fazla 693 patente tescil alındığı, en fazla doktoralı oranının 0.43, en fazla yüksek lisanslı oranının 0.71 olduğu görülmektedir.

Ar-Ge ve Tasarım merkezlerinin faaliyet durumunu etkileyen değişkenlerin, Ar-Ge ve Tasarım Merkezi Dairesi'nde çalışan uzmanların görüşleri dikkate alınarak melez bir sınıflandırma yaklaşımı uygulanması istenmiştir. Bu amaçla, nitel ve nicel değişkenleri değerlendirebilen hem objektif hem subjektif bilgileri karar sürecine dahil edebilen AHP yöntemi seçilmiştir. Ar-Ge ve Tasarım Merkezlerinin faaliyetlerini etkileyen değişkenlerin önemlerine göre ağırlıklarının hesaplanması istenmiştir. AHP yönteminde her bir değişkenin birbirine göre doğrudan öneminin belirlenebilmesi amacıyla on beş uzmanın görüşü alınmıştır. Buna göre, Eşitlik (1)'de hesaplaması gösterilen karşılaştırma matrisi

$$A = \begin{bmatrix}
1.0000 & 2.0000 & 1.0000 & 2.0000 & 1.0000 & 0.5000 & 1.0000 & 9.0000 & 3.0000 & 4.0000 & 5.0000 & 4.0000 \\
0.5000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 6.0000 & 3.0000 & 2.0000 & 4.0000 & 3.0000 \\
1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 5.0000 & 3.0000 & 2.0000 & 4.0000 & 2.0000 \\
0.5000 & 1.0000 & 1.0000 & 1.0000 & 0.5000 & 0.3333 & 0.3333 & 6.0000 & 3.0000 & 2.0000 & 4.0000 & 2.0000 \\
1.0000 & 1.0000 & 1.0000 & 2.0000 & 1.0000 & 1.0000 & 1.0000 & 7.0000 & 3.0000 & 3.0000 & 6.0000 & 3.0000 \\
2.0000 & 1.0000 & 1.0000 & 3.0000 & 1.0000 & 1.0000 & 5.0000 & 9.0000 & 5.0000 & 4.0000 & 6.0000 & 5.0000 \\
1.0000 & 1.0000 & 1.0000 & 3.0000 & 1.0000 & 0.2000 & 1.0000 & 8.0000 & 4.0000 & 3.0000 & 5.0000 & 3.0000 \\
0.1111 & 0.1667 & 0.2000 & 0.1667 & 0.1429 & 0.1111 & 0.1250 & 1.0000 & 0.3333 & 0.5000 & 1.0000 & 1.0000 \\
0.3333 & 0.3333 & 0.3333 & 0.3333 & 0.3333 & 0.2000 & 0.5000 & 3.0000 & 1.0000 & 1.0000 & 2.0000 & 1.0000 \\
0.2500 & 0.5000 & 0.5000 & 0.5000 & 0.3333 & 0.2500 & 0.6667 & 2.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 \\
0.2000 & 0.2500 & 0.2500 & 0.2500 & 0.1667 & 0.1667 & 0.2000 & 1.0000 & 0.5000 & 1.0000 & 1.0000 & 1.0000 \\
0.2500 & 0.3333 & 0.5000 & 0.5000 & 0.3333 & 0.4000 & 0.3333 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000
\end{bmatrix}$$

biçiminde oluşturulmuştur.

Çizelge 9. AHP ile belirlenen değişken ağırlıkları

	Değişkenler	w
Nicel	Araştırmacı/Tasarımcı Sayısı (X_1)	0.1310
	Toplam Tamamlanan Proje Sayısı (X_2)	0.1020
	Toplam Devam eden Proje Sayısı (X_3)	0.1026
	Patent Başvuru Sayısı (X_4)	0.0800
	Patent Tescil Sayısı (X_5)	0.1218
	Merkezdeki Doktora Mezunu Personelin Oranı (X_6)	0.1835
	Merkezdeki Y. Lisans Mezunu Personelin Oranı (X_7)	0.1189
Nitel	Sektör (X_8)	0.0373
	Şehir (X_9)	0.0249
	Kobi (X_{10})	0.0186
	Ar-Ge/Tasarım Merkezi (X_{11})	0.0404
	Yabancı Ortaklık (X_{12})	0.0390

Eşitlik (2) ve (3) kullanılarak hesaplanan değişken ağırlıkları Çizelge 9'da verilmiştir. Buna göre Ar-Ge ve Tasarım merkezlerinin faaliyet durumunu etkilemede en önemli değişkenlerin Merkezdeki Doktora Mezunu Personelin Oranı ile Araştırmacı/Tasarımcı Sayısı olduğu uzmanlar tarafından düşünülmektedir. Merkeze ait Sektör, Şehir, Kobi, Ar-Ge/Tasarım Merkezi ve Yabancı Ortaklık bilgilerinin Ar-Ge ve Tasarım merkezlerinin faaliyet durumunu etkilemede en az öneme sahip olduğu görülmektedir. AHP sonucunda en az öneme sahip olduğu düşünülen merkezlere ait bilgileri içeren değişkenler (Kobi, Yabancı Ortaklık Ar-Ge/Tasarım Merkezi, Şehir, Sektör) veri madenciliği sınıflandırma yöntemlerinde dikkate alınmamıştır. AHP sonucu elde edilen değişken ağırlıkları kullanılarak veri seti Çizelge 2'deki biçimde ağırlıklandırılır. Veri ön işleme sonucunda veri seti, yedi bağımsız nicel değişken ve bir kategorik bağımlı değişken ile analize hazır hale getirilmiştir. Çalışmanın bir sınıflandırma problemi olması sebebiyle veri madenciliği sınıflandırma yöntemlerinden LR, KNN, SVM ve RF kullanılmıştır. Sınıflandırma yöntemlerinin performansının ölçülebilmesi amacıyla veri seti, %80 eğitim seti ile %20 test seti olmak üzere ayrılmıştır. Python 3.11.3 sürümünde Scikit-learn kütüphanesinde sınıflandırma yöntemlerinin performansı 5-kat CV kullanılarak test edilmiş, ızgara arama yöntemi ile optimal parametreler elde edilmiştir.

Çizelge 10. Izgara arama ile belirlenen ayarlanabilir parametre değerleri

Yöntemler	***Ayarlanabilir Parametreleri***
LR	*c* = 1e-05, *Penalty* = None
KNN	*Metric* = Euclidean, *Neighbors* (k) = 5
SVM	*c* = 100, *Gamma* = 10, *Kernel* = Rbf'
RF	*Bootstrap* = True, *Criterion* = Gini, *Max Depth* = 3, *Max Features* = 3, *Min Samples Leaf* = 5, *Min Samples Split* = 3, *n Estimator* = 250

Her bir sınıflandırma yöntemi için kullanılan ayarlanabilir parametre değerleri Çizelge 10'da verilmiştir. Sınıflandırma yöntemlerini değerlendirmek için performans ölçütlerinden doğruluk, kesinlik, duyarlılık ve F_1-skor test veri seti için hesaplanmıştır. Test veri setine ait performans ölçütleri Çizelge 11'de verilmiştir.

Çizelge 11. Test veri seti sonuçları

	Performans Ölçütleri			
Yöntemler	***Doğruluk***	***Kesinlik***	***Duyarlılık***	***F_1-skor***
LR	0.85	0.89	0.92	0.90
KNN	0.83	0.87	0.91	0.89
SVM	0.85	0.90	0.90	0.90
RF	0.86	0.89	0.93	0.91

Çizelge 11'de görülen değerlere göre performans ölçütleri bakımından sınıflandırma yöntemlerini karşılaştırmak zordur. Objektif karşılaştırma yapılabilmesi için CODAS çok ölçütlü karar verme yöntemi uygulanmıştır. Çizelge 11'de verilen sonuçlar kullanılarak Eşitlik (20)'ye göre karar matrisi

$$D = \begin{bmatrix} 0.85 & 0.89 & 0.92 & 0.90 \\ 0.83 & 0.87 & 0.91 & 0.89 \\ 0.85 & 0.90 & 0.90 & 0.90 \\ 0.86 & 0.89 & 0.93 & 0.91 \end{bmatrix} \tag{28}$$

biçiminde oluşturulur. Eşitlik (21-27) kullanılarak CODAS yöntemine göre, değerlendirme puanı büyük olan sınıflandırma yöntemi öncelikli olarak tercih edilir.

Çizelge 12. Test veri seti için CODAS'a göre sıralama

Yöntemler	*Değerlendirme Puanları*	*Sıralama*
LR	0.0129	3
KNN	-0.1069	4
SVM	0.0202	2
RF	0.0821	1

Çizelge 12'ye göre, RF sınıflandırma yöntemi öncelikli tercih edilir. Buna göre, sınıflandırma yöntemlerinin öncelikli tercih sıralamasının RF >> SVM >> LR >> KNN biçiminde olduğu söylenir.

5. Sonuç

Bu çalışmada, sınıflandırma problemlerinin çözümünde, karar verme sürecine katkı sağlayacağı düşünülen bir melez sınıflandırma yaklaşımı önerilmiştir. Oluşturulan melez sınıflandırma yaklaşımında, veriden bilgi elde etmeye yönelik subjektif değerlendirmenin AHP yöntemi ile dikkate alınmasının yanı sıra veri madenciliği yöntemleri ile sınıflandırma yaparken MCDM yöntemi ile objektif olarak sınıflandırma yöntemlerinin performansına karar verilmiştir. Titizlikle veri ön işleme aşamasından geçirilen veri setindeki değişkenler için Ar-Ge ve Tasarım Merkezi Dairesi'nde çalışan uzmanların görüşleri alınarak AHP yöntemi değişkenlerin ağırlıkları belirlenmiştir. Ağırlıklandırılmış veri setine, LR, KNN, SVM ve RF sınıflandırma yöntemleri uygulanarak yöntemlerin performans ölçütleri hesaplanmıştır. Elde edilen sınıflandırma performans değerleri bakımından karar verilmesi çok ölçütlü bir karar verme problemi olduğundan, objektif karar verebilmek için performans ölçütü hesaplama sonuçları bir karar matrisi olarak ele alınıp CODAS uygulanmıştır. CODAS sonucuna göre sınıflandırma yöntemlerinden RF'nin öncelikli tercih edilebileceği kararı elde edilerek, Ar-Ge ve Tasarım merkezlerinin faaliyetlerinin değerlendirilmesinde RF yönteminin SVM, KNN ve LR yöntemlerine göre daha iyi sınıflandırma performansı gösterdiği sonucuna ulaşılmıştır.

Kaynaklar

[1] V. Çetin ve O. A. Yıldız, 2022, A Comprehensive review on data preprocessing techniques in data analysis, *Pamukkale University Journal of Engineering Sciences*, 28(2), 299-312.

[2] M. Emeç ve M. H. Özcanhan, 2023, Veri Ön İşleme ve Öznitelik Mühendisliğinin Yapay Zekâ Yöntemlerine Uygulanması, *Mühendislikte Öncü ve Çağdaş Çalışmalar*, 33-54.

[3] A. Burkov, "*The Hundred-Page Machine Learning Book*" kitabından çeviri, Çeviren: A. Okatan, T. Karatekin ve K. Okatan, 2021 ,100 Sayfada Makine Öğrenmesi Kitabı, (1), *Papatya Yayıncılık Eğitim*, İstanbul.

[4] J. Han, M. Kamber and J. Pei, 2012, Data mining concepts and techniques, University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University.

[5] A. Çınar ve G. Silahtaroğlu, 2012, Veri madenciliği teknikleri ile müşteri memnuniyetine etki eden gizli nedenlerin keşfi, *Marmara Üniversitesi İktisadi ve İdari Bilimler Dergisi*, 33(2), 309-330.

[6] Y. Nieto, V. Gacía-Díaz, C. Montenegro, C. C. González and R. G. Crespo, 2019, Usage of machine learning for strategic decision making at higher educational institutions, *IEEE Access*, 7, 75007-75017.

[7] Ç. Öztürk Zan, 2021, Prediction of Soil Radon Gas Using Meteorological Parameters with Machine Learning Algorithms, *M.Sc Thesis*, Dokuz Eylül University Graduate School of Natural and Applied Sciences.

[8] Ö. Ç. Yavuz, E. Karaman ve C. Yeşilyaprak, 2022, Makine öğrenmesi algoritmalarıyla astronomik gözlem kalitesi tahminine yönelik karar destek sistemi geliştirilmesi ve uygulanması, *Trends in Business and Economics*, 36 (3), 289-303.

[9] A. Ulutaş, 2019, Third-Party Logistics Provider Selection By Using AHP and CODAS Methods, *SETSCI Conference Proceedings*, 4 (8), 36-38.

[10] G. F. Can, P. Toktaş ve F. Pakdil, 2021, Six Sigma Project Prioritization and Selection Using AHP–CODAS Integration: A Case Study in Healthcare Industry, *IEEE Transactions on Engineering Management*, 70 (10), 3587-3600.

[11] U. Fayyad, 1997, Knowledge discovery in databases: An overview, *In International Conference on Inductive Logic Programming*, 1-16, Berlin, Heidelberg: Springer Berlin Heidelberg.

[12] Ş. Kavurkacı, Z. K. Aydın ve R. Şamlı, 2011, Büyük ölçekli veri tabanlarında bilgi keşfi, *Akademik Bilişim Konferansları*, 2-4.

[13] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, 1996, From data mining to knowledge discovery in databases, *AI magazine*, 17(3), 37-37.

[14] K. Keleş ve P. Z. Tunca, 2015, Hiyerarşik Electre Yönteminin Teknokent Seçiminde Kullanımı Üzerine Bir Çalışma, *Süleyman Demirel Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 20 (1), 199-223.

[15] S. Arslan ve Ö. Belgin, 2020, Yüksek ve Orta-Yüksek Teknoloji Alanındaki Sektörlerin Çok Kriterli Karar Verme Teknikleri ile Önceliklendirilmesi, *Verimlilik Dergisi*, (4), 7-23. DOI: 10.51551/verimlilik.556526.

[16] M. Güryeli, 2016, Ar-Ge Projeleri Seçim Probleminin AHP Yöntemi ile İncelenmesi: Kamu Destekli Teknolojik Ürün Yatırım Destek Programı Üzerine Bir Uygulama", *Yüksek Lisans Tezi*, Adnan Menderes Üniversitesi, Sosyal Bilimler Enstitüsü.

[17] T. L. Saaty, 2008, Decision making with The Analytic Hierarchy Process, *International Journal Services Sciences*, 1(1), 83-98.

[18] T. L. Saaty, 1990, The Analytıc Hıerarchy Process In Conflıct Management, *International Journal of Conflict Management*, 1(1), 47-68. https://doi.org/10.1108/eb022672

[19] M. Ö. Dolgun, T. G. Özdemir ve D. Oğuz, 2009, Veri madenciliğinde yapısal olmayan verinin analizi: Metin ve web madenciliği. *İstatistikçiler Dergisi: İstatistik ve Aktüerya*, *2*(2), 48-58.

[20] C. Cortes and V. Vapnik, 1995, Support-vector networks, *Machine learning*, 20, 273–297. https://doi.org/10.1007/BF00994018

[21] N. Bayram Arlı, M. Engin ve S. Gürsakal, 2022, Random Forest. Supervised Machine Learning Algorithms R and Python Applications, *Nobel Yayınevi*, Ankara.

[22] A. Géron, 2019, Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow, *O'Reilly Media,* Sebastopol, CA.

[23] L. Breiman, 2001, Random Forest, *Machine learning*, 45, 5-32.https://doi.org/10.1023/A:1010933404324

[24] M. Öztürk, Python ile Sınıflandırma Analizleri – Rastgele Orman (Random Forest) Algoritması- Miraç ÖZTÜRK (miracozturk.com), Erişim tarihi:04.10.2023.

[25] M. K. Ghorabaee, E. K. Zavadskas, Z. Turskis and J. Antucheviciene, 2016, A new combinative distance-based assessment (CODAS) method for multi-criteria decision-making. *Economic Computation and Economic Cybernetics Studies and Research*, 50, 25–44.