DİCLE ÜNİVERSİTESİ
İKTİSADİ VE İDARİ BİLİMLER FAKÜLTESİ
DERGİSİ
Dicle University
Journal of Economics and Administrative Sciences
E-ISSN: 2587-0106

Cilt / Volume: 12, Sayı / Issue: 24, Sayfalar / Pages: 480-498

Araştırma Makalesi / Research Article

Received / Alınma: 29.06.2022 Accepted / Kabul: 31.10.2022

SINIFLANDIRMA AMAÇLI DEĞİŞKEN ALT KÜMESİ SEÇİMİ: BİR BANKACILIK UYGULAMASI*

Emrah SEZER[1]

Özgür ÇAKIR[2]

Öz

Teknolojik gelişmelerin etkisi ile kaydedilen operasyonel veriler giderek artmaktadır. Veri miktarı ve çeşitliliğindeki artış nedeni ile analiz aşamasında ve analiz sonuçlarının değerlendirilmesi aşamasında birçok zorluk yaşanmaktadır. İlgili ve ilgisiz birçok verinin analiz aşamasına aktarılmasının sonucunda analizlerin yapılabilmesi için gerekli zaman ve kaynak gereksinimleri artmaktadır. Kaynakların ve zamanın daima sınırlı olacağı aşikardır. Bu çalışmanın amacı, bankacılık müşteri verileri üzerinde sınıflandırma amaçlı değişken seçimi uygulamaları yaparak ilgisiz değişkenleri elemek ve sınıflandırma çalışmasına katkıda bulunmaktır. Farklı değişken seçimi yöntemleri kullanılarak seçilen değişken alt kümeleri üzerinde sınıflandırma uygulaması yapılmıştır. Sınıflandırma sonuçları karşılaştırılarak değişken seçim yöntemlerinin başarısı ölçülmüştür.

Anahtar Kelimeler: Veri Madenciliği, Sınıflandırma, Değişken Seçimi, Korelasyon Bazlı Değişken Seçimi, Sezgisel Arama.

Jel Kodları: C00, C02, C60, C61.

* Bu çalışma,"Sınıflandırma sorunu için en uygun değişken alt kümesi seçimi üzerine bir uygulama" başlıklı tez'den üretilmiştir.

[1]İstanbul Üniversitesi, E-posta: emrah.sezer@ogr.iu.edu.tr, ORCID: 0000-0002-5078-9463.
[2]Doç.Dr. , Marmara Üniversitesi, E-posta: ocakir@marmara.edu.tr, ORCID: 0000-0003-1410-8162.

Atıf/Citation
Sezer, E. & Çakır, Ö. (2022). Sınıflandırma amaçlı değişken alt kümesi seçimi: bir bankacılık uygulaması. *Dicle Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 12*(24), 480-498.

A FEATURE SELECTION APPLICATION FOR CLASSIFICATION: A BANKING APPLICATION

Abstract

Operational data is gradually increasing with the effect of technological developments. Due to the increase in the amount and diversity of data, many difficulties are faced in data analysis and the evaluation of its results. Since the data transferred to the analysis stage consists of both relevant and irrelevant variables, time and resources required for analysis increases. It is obvious that resources and time are always limited. The aim of this study is to increase the success of classification methods under the limitation of time and resources by applying various feature selection methods. Feature selection methods were used on a set of banking customer data in order to determine the appropriate subsets for classification. Then, a classification method was applied on these selected subsets of features. By comparing the classification results, the contribution of the feature selection methods to the classification success was measured.

Keywords: Data Mining, Classification, Feature Selection, Correlation Based Feature Selection, Heuristic Search.

Jel Codes: C00, C02, C60, C61.

1. GİRİŞ

Son birkaç on yılda, hem üretilen verilerin boyutlarının ve çeşitliliğinin artması hem de bu verilerin saklanmasına ilişkin maliyetlerin azalması nedeniyle işletmeler mümkün olan tüm kayıtları saklama eğilimindedirler. İşletmelerde kaydedilen büyük miktarlardaki verilerden anlamlı ve ilgi çekici bilgi çıkarımı veri madenciliğinin konusudur. Veri madenciliği ile tanımlayıcı çalışmalar yapılabildiği gibi öngörü modelleri de üretilebilmektedir.

Sınıflandırma, gerçekleşmiş verilerden hareketle kategorik bir hedef değişkene yönelik öngörü modelleri üretmek amacı ile kullanılan bir veri madenciliği yöntemidir. Sınıflandırma uygulamasında kategorik ve sürekli verilerle çalışma imkanı bulunmaktadır. Sınıflandırma, karar ağaçları, yapay sinir ağları, lojistik regresyon, destek vektör makinesi gibi farklı yapılarda algoritmalarla yapılabilmektedir.

İşletme verileri, sınıflandırma analizi için bir araya getirildiğinde analizi olumsuz etkileyebilecek birçok handikap barındırabilmektedir. İşletme tarafından kullanılan ilişkisel veri tabanları ve veri ambarları gibi kaynaklardan analiz için oluşturulan veri kümesinde hedef değişken ile alakasız veya hedef değişkeni açıklayıcılığı çok düşük değişkenler de bulunabilmektedir. Bu açıklayıcı olmayan değişkenlerin ayıklanması amacı ile yine bir veri madenciliği uygulaması olan değişken seçimi algoritmaları kullanılmaktadır. Gereksiz değişkenlerin elenmesi ile sınıflandırma modelinin başarısını arttırmak, ortaya çıkan modelin

daha sade ve anlaşılabilir olmasını sağlamak ve sınıflandırma analizi için gerekli olan analiz süresini kısaltmak mümkündür.

Ancak geniş bir veri kümesinde bulunan birçok değişkenden hangilerinin seçilmesi gerektiğine karar vermek değişken seçimi algoritmaları için de her zaman çok kolay olmamaktadır. Değişken uzayının her bir alt kümesinin ideal değişken alt kümesi olma ihtimali bulunmaktadır ve geniş veri kümelerinde bu uzay çok büyük olabilmektedir. Bu kadar fazla alternatifi değerlendirmeye dayalı yöntemler bulunmasına karşın zaman ve kaynak kısıtları nedeniyle çoğu durumda kullanışlı değildirler. Bu aşamada, sorunun bir optimizasyon problemine dönüşmesi ve sezgisel (heuristic) ya da rastgele arama yöntemlerine dayalı algoritmaların kullanılması gündeme gelmektedir. Bu algoritmaların temel prensibi, değişken uzayının sınırlı bir bölümü üzerinde değerlendirme yapılarak hedef değişken üzerinde açıklayıcılık düzeyi en yüksek değişken alt kümesinin sınıflandırma modeli için önerilmesidir.

Bu çalışmanın ikinci bölümünde değişken alt kümesi seçimi yaklaşımları hakkında bilgi verilecektir. Üçüncü bölümde ise değişken alt kümesi seçiminde kullanılan modeller anlatılacaktır. Son olarak dördüncü bölümde ise, bir bankacılık müşteri veri kümesi üzerinde gerçekleştirilmiş farklı değişken alt kümesi seçimi uygulamaları karşılaştırılmalı olarak anlatılacaktır.

2. DEĞİŞKEN ALT KÜMESİ SEÇİMİ

Daha fazla sayıda açıklayıcı değişken kullanılması teoride daha yüksek düzeyde sınıflandırma başarısı elde etmeyi mümkün kılsa bile pratikte öğrenme sürecini yavaşlatacağı gibi alakasız değişkenlerin dikkate alınması nedeniyle sınıflandırıcının öğrenme verilerine aşırı uyum göstermesine de neden olabilir (Yu & Liu 2004).

Veri madenciliği süreçlerinde değişken seçiminin nedenleri şu üç başlıkta özetlenebilir (Liu & Motoda 1998b);

- **Hesaplama Süresi (Computing Time)**: Veri madenciliği süreci öncesinde gereksiz değişkenlerin ayıklanması, veri madenciliği sürecindeki işlem süresini, işlemci ve bellek kullanımını düşürecektir.
- **Öngörü Doğruluğu (Predictive Accuracy):** Alakasız değişkenlerin, oluşturulan süreci yanlış yönlendirmesi söz konusu olabilmektedir. Sadece ilgili değişkenlerin sürece dahil edilmesi ile modelin öngörü doğruluğunun artması beklenir.

- **Sonuçların Sunumu:** Sadece ilgili değişkenlerden oluşturulacak bir model, daha sade sonuçlar üreterek daha anlaşılır bir öğrenme sürecinin gerçekleşmesini sağlayacaktır.

Bu üç değerlendirme kriterinin aynı anda sağlanabilmesi ideal durumdur. Ancak gerçek veri ile çalışılırken her zaman ideal çözüm elde edilemeyebilir. Çalışmanın amacı ve öncelikleri göz önünde bulundurularak bu kriterlerin bazılarından belirli ölçülerde feragat edilebilir (Kantardzic 2011).

Sıralanmış bir değişken listesi oluşturan tipteki algoritmalar, bir alt kümeyi vermek yerine gerekli değişkenleri diğerleri ile karşılaştırmaktadırlar. Bu şekilde bir kullanım akılcı olmayabilir ve en küçük değişken alt kümesinin elde edilmesinin amaçlandığı çalışmalar için önerilmez (Guyon & Elisseeff 2003).

Değişkenleri sadece sıralamaktan ibaret olan algoritmalar yerine daha farklı algoritmalara ihtiyaç duyulabilmektedir. Bu noktada en küçük alt küme algoritmaları kullanılmaktadır. Bu algoritmalar, ilgili değişkenleri barındıran en küçük değişken alt kümesini üretir (Liu & Motoda 1998b).

k, bir veri kümesindeki değişkenlerin sayısı ise 2^k adet değişken alt kümesinin söz konusu olacağını biliyoruz. Arama yönleri, değişken alt kümesinin oluşturulması ile yakından ilgilidir. Arama yönleri genel olarak üç şekilde gruplanabilir;

(1) İleri Yönlü Arama: Boş bir değişken alt kümesi ile başlar. Her seferinde belirli bir kritere göre en iyi olduğuna karar verilen tek bir değişken kümeye eklenir.

(2) Geri Yönlü Arama: Tüm değişkenleri içeren küme ile başlar. Her seferinde belirli bir kritere göre en az öneme sahip olan tek bir değişken kümeden çıkarılır.

(3) İki Yönlü Arama: Her iki yönde başlar ve eş zamanlı olarak devam eder. Bir yöndeki arama en iyi değişken alt kümesini bulduğu ya da her iki arama da sürecin ortasına geldiği zaman sona erer (Liu & Motoda 1998a).

Arama yöntemleri, tüm değişken uzayından iyi bir değişken alt kümesi bulana kadar gezinir. Yöntemin başarısı, değişken alt kümesi değerlendirme algoritması tarafından belirlenir.

Sıralayıcı arama (Ranker) yöntemleri, bir değişken alt kümesi üretmek yerine değişkenlerin sıralaması şeklinde bir çıktı verdiklerinden dolayı önceleri bir arama yöntemi olarak kabul

edilmezdi. Bu yöntemler, her bir değişken için ayrı ayrı skor üreterek önem sırasına göre sıralanmasını sağlarlar.

Tam arama (Complete Search), boş kümeden başlayarak tüm değişkenleri içeren kümeye kadar tüm değişken kümelerini taranarak en iyi değişken alt kümesinin bulunmasını sağlar. Diğer algoritmalara nazaran daha fazla kaynak ve zaman kullanarak ideal çözümü bulabilmektedir. Sezgisel (heuristic) yöntemler ise daha hızlı çözüm üretirler ancak tam arama ile elde edilebilecek olan ideal çözümü elde etmeyi garanti etmezler. Daha kısa zamanda ve daha az kaynak kullanarak ideal çözüme en yakın çözüme ulaşmak isteniyorsa sezgisel arama kullanılabilir.

Sıralı arama (rank search), kendi içerisinde bir adet sıralayıcı (ranker) değişken değerlendirme algoritması seçilir. Bu seçilen algoritma değişkenler için bir sıralama oluşturur. Oluşturulan bu sıralama, seçilen bir değişken alt kümesi seçme algoritması tarafından kullanılarak sıralamadaki en iyi, akabinde ikinci en iyi gelecek şekilde ilerleyen bir arama süreci izlenir ve sürecin sonunda değişken alt kümesi seçim algoritmasının kriterlerine göre bir değişken alt kümesi seçilir (Witten, Frank & Hall 2011).

Rastgele arama, önceki arama yöntemlerinin tamamı aynı sırayı takip ederek çalışan ve ilk çalıştırmanın akabinde tekrar çalıştığında ilk sonuç ile aynı sonucu üreten yöntemlerdir. Rastgele arama (Nondeterministic Search) ise belirli bir sırayı takip etmeksizin çalışır ve her çalıştırılışında farklı bir sonuç üretir (Dash & Liu 1997). Bu çalışma prensibinden dolayı iyi bir sonuç üretebilir ancak en iyi sonucu ıskalama ihtimali kabul edilerek uygulanır (Liu & Motoda 1998b).

Sınıflandırma algoritmalarında kullanılmak üzere değişken alt kümesi seçme yöntemleri filtreleyici ve sarmalayıcı olmak üzere iki farklı grupta ele alınmıştır. Sarmalayıcı yöntemler değerlendirme ölçüsü olarak sınıflandırma hata oranını kullanırken, filtreleyici yöntemler ise dolaylı yollardan ölçüm yapmayı sağlayan, sınıflandırma doğruluğundan farklı bir değerlendirme ölçüsü kullanmaktadırlar (Blum & Langley 1997).

Son şekli ile ölçü değerlendirme fonksiyonları aşağıdaki gibi beş farklı başlık altında incelenmektedir;

- Mesafe (Distance)
- Bilgi (Information)

- Korelasyon (Correlation)
- Tutarlılık (Consistency)
- Sınıflandırma Hata Oranı (Prediction error rate)

Başeğmez ve arkadaşları, yumurtalık kanseri tespitinde gen seçimi üzerine, literatürde bulunan farklı değişken alt kümesi seçim yaklaşımlarından faydalanarak hibrit denemede bulunmuşlardır (Başeğmez et al. 2021).

Haldorai ve arkadaşları, kentsel sürdürülebilirlik için, kanonik korelasyon bazlı değişken seçimine dayalı yapay sinir ağları sınıflandırma çalışması gerçekleştirmiştirler. Bu yaklaşımlarında, Mark Hall tarafından ortaya atılan geleneksel korelasyon bazlı değişken seçimi yaklaşımından faydalanmamışlardır (Haldorai et al. 2021).

Mansour ve arkadaşları, COVID-19 tespitinde, korelasyon ve Naive Bayes tabanlı bir sınıflandırma yaklaşımını, genetik algoritma ile çalışan bir sarmalayıcı değişken seçimi çalışması olarak kullanmışlardır (Mansour et al. 2022).

Omuya ve arkadaşları, temel bileşenler analizi (PCA) ve bilgi kazanımından faydalanarak ortaya koydukları değişken seçimi yaklaşımı, literatürde bulunan diğer algoritmaların performansları ile karşılaştırmışlardır. Bu karşılaştırmalarda, ortaya koydukları yaklaşım ile süre ve sınıflandırma başarısı açısından iyi sonuçlar elde ettiklerini belirtmişlerdir (Omuya et al. 2021).

SaiSindhu ve Shyam, bulut bilişim ortamında DoS saldırı tespiti için meta-sezgisel algoritma tabanlı değişken seçimi ve yinelemeli yapay sinir ağları kullanarak bir sınıflandırma çalışması gerçekleştirmişlerdir. KDD cup 99 veri kümesi üzerinde dört farklı yaklaşımı uygulamış ve sonuçlarını karşılaştırmışlardır (SaiSindhu & Shyam 2021).

Sun ve arkadaşları, çok etiketli sınıflandırma problemleri için, bulanık komşuluk kaba kümelerine (Multilabel Fuzzy Neighborhood Rough Sets) ve maksimum ilişki minimum yedekliliğe (Maximum Relevance Minimum Redundancy) dayalı yeni bir değişken seçim yöntemi önermiştir. Önerdikleri yöntemi 20 farklı veri kümesi üzerinde uygulamışlardır ve sonuçlarını farklı yaklaşımlarla elde edilen sonuçlarla karşılaşmışlardır (Sun et al. 2021).

Song ve arkadaşları, Yüksek boyutlu veriler için korelasyon güdümlü kümelemeye ve parçacık sürüsü optimizasyonuna dayalı bir değişken seçimi yaklaşımı önermişlerdir. Bu

yaklaşımlarını, 18 farklı veri kümesi üzerinde uygulamışlar ve sonuçlarını karşılaştırmalı olarak incelemişlerdir (Song et al. 2021).

Thaseen ve arkadaşları, saldırı tespitinde korelasyon bazlı bir değişken seçimi ve yapay sinir ağları kullanmışlardır. Çok sınıflı bir veri üzerinde gerçekleştirdikleri değişken seçimi çalışmalarında, seçilmiş değişken alt kümesi ve değişken seçimi öncesindeki veri ile gerçekleştirdikleri sınıflandırma çalışmalarına ait sonuçları, sınıf etiketleri bazında performans karşılaştırması ile yorumlamışlardır (Thaseen et al. 2021).

Sun ve arkadaşları, Çok etiketli sınıflandırma çalışmalarında karşılaşılan sorunların üstesinden gelmek için, Fisher skorunu ve çok etiketli komşuluk kaba kümelerini (multilabel neighborhood rough sets) kullanan yeni bir değişken seçim yöntemi önermişlerdir (Sun et al. 2021).

Wang ve arkadaşları, seçilmiş değişken sayısının minimize edilmesi ve sınıflandırma başarısının maksimize edilmesini amaçlayan çok hedefli optimizasyon yaklaşımını karınca kolonisi algoritması ile birlikte kullanarak bir yaklaşım geliştirmişlerdir. 13 ayrı veri kümesi üzerinde test ettikleri yaklaşımlarının ve literatürdeki diğer yaklaşımlarla performans karşılaştırmalarını raporlamışlardır (Wang et al. 2022).

3. DEĞİŞKEN SEÇİM MODELLERİ

Değişken seçim algoritmaları genel olarak sarmalayıcı ve filtreleyici olarak iki ana grupta incelenmektedir.

Değişken seçim algoritmalarında performans ölçüsü olarak sınıflandırıcının doğruluğunun (accuracy) kullanılması basit bir yaklaşımdır. Bu durumda mümkün olan en yüksek sınıflandırıcı doğruluğunu elde etmeye yönelik bir sınıflandırıcı oluşturulmalıdır ve sınıflandırıcı için en uygun değişkenler seçilmelidir. Bu yöntem “sarmalayıcı model” olarak adlandırılır (Liu & Motoda 1998a). Sarmalayıcı yaklaşım, Kohavi ve John tarafından ortaya atılmıştır (Kohavi & John 1997).

Sarmalayıcı modellerdeki iş yükünün fazlalığı ve uzun süre gereksinimi nedeniyle dolaylı performans ölçümleri üzerinde çalışmaların yapılmasına yönelinmiştir. Genelde uzaklık, bilgi, korelasyon gibi ölçüler değişken seçimi için kullanılmıştır. Bu modeller “filtreleyici model” olarak adlandırılır. Bir filtreleyici model iki safhadan oluşmaktadır;

(1) Sınıflandırıcıdan bağımsız olarak bilgi, uzaklık, korelasyon gibi bir ölçü ile değişken seçimi yapılan safha,

(2) Öğrenme verisinden (training data) öğrenilen sınıflandırıcının, seçilen değişkenler ile birlikte sınama kümesi üzerinde uygulandığı safha.

Ayrıca filtreleyici model bazı niteliklere sahiptir;

- Sınıflandırıcının belirli bir eğilimine dayanmak yerine verinin özünde bulunan özelliklere dayanır. Değişkenler farklı sınıflandırma tekniklerinde kullanılmak üzere seçilebilir.
- Bilgi, uzaklık, korelasyon, tutarlılık ölçüleri genel olarak sınıflandırıcı doğruluğu ölçüsüne nazaran daha az çaba gerektirir. Böylece filtreleyici yöntemler daha hızlı şekilde değişken alt kümesi oluşturabilirler.
- Ölçülerin sadeliğinden ve zaman tasarrufu sayesinde sınıflandırıcının yapabileceğinden daha geniş bir veri kümesine hakim olabilir. Bu durumda sınıflandırıcı geniş bir veri kümesinden direk öğrenilmemiş, boyut (dimension) yönünden bakıldığında da indirgenmiş bir veri kümesinden öğrenilmiş olur. Bunlarla birlikte, bir filtreleyici model tarafından seçilen değişkenlerin, öğrenme algoritmasının verinin eğilimlerinin tamamından faydalanmasına engel olması gibi bir tehlike söz konusudur (Liu & Motoda 1998b).

Büyük veri kümeleri ile çalışılırken daha az kaynak gereksiniminden dolayı filtreleyici yöntemlerin kullanılması daha avantajlı olmasına karşın, sarmalayıcı yöntemler ile sınıflandırma doğruluğunu yükseltme hedefine ulaşmak daha garantidir (Das 2001).

Filtreleyici modeller, elde edilebilecek en iyi değişken alt kümesini bulamama riskini barındırırken, sarmalayıcı yöntemler ise öğrenme kümesine aşırı bağımlı (over fitting) kalma riskini barındırırlar (Ladha & Deepa 2011).

Başlıca filtreleyici değişken seçimi yöntemleri aşağıda ifade edilmiştir.

Bilgi Kazanımı Bazlı Değişken Seçimi

Bilgi kazanımı bazlı değişken seçimi (IG), basit ve hızlı olduğundan oldukça tercih edilen bir değişken seçim algoritmasıdır. Bu algoritma, ölçü olarak entropi ölçüsünü kullanır. Her bir değişken için, değişkenin eklenmesinden önce ve sonra hesaplanan entropi değerlerinin farkı ile hesaplanan değere bilgi kazanımı adı verilir (Abraham, Simha & Iyengar 2009).

A değişkenler kümesini, C ise hedef değişkenin aldığı değerlerin kümesini göstermek üzere (1) nolu denklem ve (2) nolu denklem ile hesaplanan ve H ile ifade edilen entropi değerleri kullanılarak (3) nolu denklem ile IG ile gösterilmiş olan bilgi kazanımı değeri hesaplanır (Shahbaz et al. 2016, Hall & Holmes 2000).

$$H(C) = -\sum_{c \in C} P(c)\ \log_2 P(c) \tag{1}$$

$$H(C|A) = -\sum_{a \in A} P(a) \sum_{c \in C} P(c|a)\ \log_2 P(c|a) \tag{2}$$

$$IG_i = H(C) - H(C|A_i) \tag{3}$$

Bilgi kazanımı bazlı değişken seçimi sıralayıcı değişken seçim yöntemi olarak kullanılabilir (Alhaj et al. 2016).

Kazanç Oranı Bazlı Değişken Seçimi

Kazanç oranı değişken seçimi (GR), bilgi kazanımının bir başka versiyonu olarak ifade edilebilir. Kazanılan bilginin sınıfın entropi değerine oranı olarak hesaplanmaktadır. (3) numaralı denkleme alternatif olarak,

$$GR_i = \frac{H(C) - H(C|A_i)}{H(C|A_i)} \tag{4}$$

şeklinde ifade edilir (Abraham et al. 2009, Karegowda, Manjunath & Jayaram 2010, Priyadarsini, Valarmathi & Sivakumari 2011).

Simetrik Belirsizlik Bazlı Değişken Seçimi

Simetrik belirsizlik (SU), değişkenlerin sınıfa göre entropi değerlerini kullanarak bir ölçüm yapar. Simetrik belirsizlik hesaplaması için;

$$SU(X, C) = 2 \times \left[\frac{H(X) + H(C) - H(X, C)}{H(X) + H(C)}\right] \tag{5}$$

denklemi kullanılarak her değişken için bir değer hesaplanır ve değişkenler bu değere göre sıralanır. Bu yöntem bir değişken sıralama (ranker) yöntemi olarak kullanılmış olur (Hall & Holmes 2003, Ali & Shahzad 2012).

Çalışma öncesinde sürekli değişkenler için kesikleştirme yapılarak simetrik belirsizlik (symmetrical uncertainty) hesaplamasında Eşitlik (5)'den faydalanılır (Hall 2000).

Bir değişken için simetrik belirsizlik değerinin yüksek olması, bu değişken ile hedef değişken arasında güçlü bir ilişkinin bulunduğunu gösterir ve bu değişkenin sınıflandırmada kullanılmasının faydalı olduğu anlamına gelir (Piroonratana et al. 2010).

Korelasyon Bazlı Değişken Seçimi

Bir değişken ile hedef değişken arasındaki korelasyon, değişkenin sınıf ile ilgili olduğunu ortaya koymaya yetecek kadar yüksekse ve bu değişken sınıf ile ilgili diğer değişkenler tarafından öngörülemez durumdaysa sınıflandırma amacı için iyi bir değişken olarak kabul edilebilir (Yu & Liu 2003).

Korelasyon bazlı değişken seçimini (CFS) ortaya atan Mark Hall, yaklaşımını şu cümle ile özetlemektedir;

"İyi bir değişken alt kümesinin barındırdığı değişkenler, sınıf ile yüksek korelasyona sahip iken kendi aralarında korelasyon bulunmayanlardır (Hall 1999)."

Korelasyon bazlı değişken seçimi sarmalayıcı değişken seçimi ile karşılaştırıldığında oldukça kısa sürede ve etkili sonuçlar üretebilmektedir (Hall & Smith 1997).

Korelasyon bazlı değişken seçimi, korelasyon bazlı bir değerlendirme fonksiyonuna göre değişkenleri sıralayan basit bir algoritmadır. Bu fonksiyon, hedef değişken ile yüksek korelasyona sahip iken kendi aralarında korelasyona sahip olmayan değişkenleri seçmeye eğilimlidir. Alakasız değişkenler hedef değişken ile düşük korelasyona sahip olacaklarından göz ardı edilmelidirler.

Değişken alt kümesi S, değişken alt kümesinin eleman sayısı k, değişken ile sınıf arasındaki korelasyon $\bar{r}_{cf}$, değişkenler arası korelasyonların aritmetik ortalaması $\bar{r}_{ff}$ olmak üzere, korelasyon bazlı değişken seçimine ait alt küme değerlendirme fonksiyonu;

$$M_S = \frac{k\,\bar{r}_{cf}}{\sqrt{k + k(k-1)\bar{r}_{ff}}} \qquad (6)$$

şeklindedir. M_S değeri korelasyon bazlı değişken seçiminin çekirdeğini oluşturur ve değişken alt kümeleri uzayında bulunan tüm olası değişken alt kümelerini sıralamak için kullanılır (Hall 1999, Sun et al. 2016).

Korelasyon bazlı değişken seçiminde, değişkenlerin kendi aralarındaki ve sınıf ile arasındaki korelasyon hesaplanır ve bu korelasyon hesaplarından yararlanılarak hesaplanan M_S değeri ile arama yönteminin değişken alt kümesi uzayına aktarılır. Arama yöntemi, en yüksek M_S

değerine sahip olan alt kümeyi öğrenme kümesi ve sınama kümesi ile birlikte en son değerlendirme aşamasına gönderir (Hall 1999, Jungjit 2016).

Tutarlılık Bazlı Değişken Seçimi

Bu değişken seçim yöntemi, değişkenlerin hedef değişken ile tutarlılıklarının (consistency) göz önünde bulundurulması ile oluşturulur. Bu değişken seçim yaklaşımında, başlangıç verisinin içerisindeki alt kümelerden güçlü değişken kombinasyonları aranır. Bu yaklaşım, hedef değişken ile yüksek tutarlılıkta ve küçük kümeler oluşturmaya eğilimlidir. Değişken alt kümelerine ait tutarlılık değeri,

Bir değişken alt kümesi S, benzersiz değişken alt küme sayısı j, verinin sayısı N, veri kümesi satır indisi i, değişkenin i satırında aldığı değerin gerçekleşme sayısı D_i, değişkenin i satırında aldığı değerin en geniş sınıftaki kardinalitesi M_i olmak üzere,

$$C_S = 1 - \frac{\sum_{i=0}^{j} |D_i| - |M_i|}{N} \qquad (7)$$

şeklinde elde edilir (Hall & Holmes 2003, Pedrycz, Succi & Sillitti 2016, Abraham et al. 2009).

4. SINIFLANDIRMA AMAÇLI DEĞİŞKEN ALT KÜMESİ SEÇİMİ

Bu bölümde ulusal bir bankanın veri tabanından elde edilmiş bir veri kümesi üzerinde sınıflandırma çalışması için değişken seçme uygulamaları yapılarak ve değişken seçimi ile oluşturulan değişken alt kümeleri ile sınıflandırma çalışmaları yapılarak sonuçları karşılaştırılmıştır. Bu şekilde değişken alt kümesi seçme çalışmasının sınıflandırma işlemi üzerindeki etkileri gözlemlenmeye çalışılmıştır ve farklı yöntemlerin performansları karşılaştırılmıştır. Çalışma, 1.8 GHz Dual-Core Intel Core i5 işlemci ve 4 GB 1600 MHz DDR3 bellek barındıran MacBook Air (13-inch, Mid 2012) kullanılarak gerçekleştirilmiştir.

Çalışmada Weka paket programından faydalanılarak kullanılan farklı yöntemler Tablo 1 de gösterilmiştir. Weka içerisinden sarmalayıcı yaklaşım modeli için WrapperSubsetEval, korelasyon bazlı değişken seçimi yaklaşımı için CfsSubSetEval, Tutarlılık bazlı değişken seçimi için ConsistencySubsetEval, kazanç oranı bazlı değişken seçimi için GainRatioAttributeEval, bilgi kazanımı bazlı değişken seçimi için InfoGainAttributeEval, simetrik belirsizlik bazlı değişken seçimi için SymmetricalUncertAttributeEval algoritmaları kullanılmıştır.

Tablo 1. Kullanılan Değişken Seçim Yöntemleri

	Değişken alt kümesi seçim yöntemi	Arama Yöntemi	Arama Biçimi
Yöntem-1	WrapperSubsetEval	BestFirst	İleri yönlü
Yöntem-2	WrapperSubsetEval	BestFirst	Geri yönlü
Yöntem-3	CfsSubSetEval	BestFirst	İleri Yönlü
Yöntem-4	CfsSubSetEval	RankSearch	GainRatioAttributeEval
Yöntem-5	CfsSubSetEval	RankSearch	InfoGainAttributeEval
Yöntem-6	CfsSubSetEval	RankSearch	SymmetricalUncertAttributeEval
Yöntem-7	ConsistencySubsetEval	BestFirst	İleri Yönlü

Yine Tablo 1 ile gösterilen yöntemlerde, sezgisel arama yöntemi için ileri veya geri yönlü olarak BestFirst, sıralı arama için RankSearch algoritmaları Weka içerisinden seçilmiştir.

Elde edilen yedi farklı değişken alt kümeleri ve tüm değişkenleri içeren veri kümesi için ayrı ayrı Weka'da bulunan C4.5 karar ağacı algoritmasının bir versiyonu olan J48 algoritması ile sınıflandırma çalışması yapılarak sekiz farklı karar ağacı oluşturmuştur. Bu oluşturulan karar ağaçlarına ait detaylar Tablo 2 ile gösterilmiştir.

Tablo 2. Karar Ağaçlarının Karşılaştırılması

	Değişken Sayısı	Ağaçtaki dal sayısı	Ağaç büyüklüğü	Model oluşturma Süresi(sn)	Doğru sınıflandırma oranı	Kappa	Ortalama mutlak hata (MAE)	Hataların ortalama karekökü (RMSE)	Mutlak bağıl hata (RAE)	Bağıl hataların karekökü (RRSE)	Doğru Sınıflandırılan Kayıt Sayısı	Hata matrisi
Seçim Öncesi Veri	100	184	252	0.52	%79.6026	0.5921	0.2596	0.406	%51.9139	%81.2069	1963	a b 945 288 | a = 0 215 1018 | b = 1
Yöntem-1	12	15	29	0.08	%82.11	0.6423	0.283	0.3797	%56.6042	%75.9316	2025	a b 934 299 | a = 0 142 1091 | b = 1
Yöntem-2	57	131	195	0.22	%80.6569	0.6131	0.2605	0.3908	%52.1023	%78.165	1989	a b 965 268 | a = 0 209 1024 | b = 1
Yöntem-3	11	48	60	0.03	%81.3058	0.6261	0.2743	0.3795	%54.8627	%75.8901	2006	a b 912 321 | a = 0 140 1093 | b = 1
Yöntem-4	9	13	18	0.03	%80.9813	0.6196	0.2913	0.3855	%58.2607	%77.0952	1997	a b 898 335 | a = 0 134 1099 | b = 1
Yöntem-5	12	50	64	0.05	%81.2247	0.6245	0.2807	0.3849	%56.1418	%76.9853	2003	a b 928 305 | a = 0 158 1075 | b = 1
Yöntem-6	9	27	39	0.03	%81.4274	0.6285	0.2785	0.379	%55.7008	%75.8096	2008	a b 931 302 | a = 0 156 1077 | b = 1
Yöntem-7	21	142	188	0.09	%79.9676	0.5994	0.2783	0.4019	%55.65	%80.3899	1972	a b 922 311 | a = 0 183 1050 | b = 1

Tablo 2 üzerinde gerçekleştirilen tüm karar ağacı uygulamalarının sonuçları yan yana getirilerek karşılaştırma imkanı oluşturulmuştur. Tablo 2 incelenerek en iyi ve en kötü sonuçları üreten algoritmalar Tablo 3 ile gösterilmiştir. Tablo 2 üzerindeki sonuçlara göre yöntemlerin sıralamaları ise Tablo 4 ile gösterilmiştir. Bu tablolara ait detaylar değerlendirme bölümünde incelenmiştir.

Tablo 3. Yöntemlerin Karşılaştırılması

	Seçim öncesi	En iyi sonuç		En kötü sonuç	
Değişken Sayısı	100	9	Yöntem-4 , Yöntem-6	57	Yöntem-2
Karar ağacındaki dal sayısı	184	13	Yöntem-4	131	Yöntem-2
Model oluşturma Süresi(Saniye)	0.52	0.03	Yöntem-3,Yöntem-4,Yöntem-6	0.22	Yöntem-2
Doğru sınıflandırma yüzdesi (%)	%79.6	%82.1	Yöntem-1	%79.96	Yöntem-7
Kappa	0.5921	0.64	Yöntem-1	0.5994	Yöntem-7
Hata matrisi a=0 isabetli sayısı	945	965	Yöntem-2	898	Yöntem-4
Hata matrisi b=1 isabetli sayısı	1018	1099	Yöntem-4	1024	Yöntem-2

Tablo 4. Yöntemlerin Sıralanması

	Değişken Sayısı	Ağaçtaki dal sayısı	Ağaç büyüklüğü	Model oluşturma Süresi(sn)	Doğru sınıflandırma oranı	Kappa	Doğru Sınıflandırılan a=0	Doğru Sınıflandırılan b=1	Ortalama	Medyan
Yöntem-4	1	1	1	1	5	5	7	1	2,43	1
Yöntem-1	4	2	2	5	1	1	2	3	2,71	2
Yöntem-6	1	3	3	1	2	2	3	4	2,43	3
Yöntem-3	3	4	4	1	3	3	6	2	3,29	3
Yöntem-5	4	5	5	4	4	4	4	5	4,43	4
Yöntem-7	5	6	6	6	7	7	5	6	5,86	6
Yöntem-2	6	7	7	7	6	6	1	7	5,86	7

5. DEĞERLENDİRME

Karar ağaçları incelendiğinde, değişken seçim algoritmaları yardımı ile tüm değişkenleri içeren veri kümesinde bulunan 100 değişkenin dokuz ile 57 arasında değişkene indirgenebildiği görülmektedir. Yöntem-2 ile seçilen 57 değişken dışındaki yöntemler ise dokuz ile 21 arasında değişken seçmişlerdir. Yöntem-2 algoritmasının, değişkenlerin evrensel kümesinden değişkenleri azaltarak boş kümeye doğru bir arama süreci izlemesi, sınıflandırma doğruluğunu en yüksek yapan bir lokal maksimum değerinde aramasını kesmesi nedeni ile 100 değişkenden başlayarak 57 değişkene kadar indirgeme yaptığı ve tahmin tutarlılığını yüzde 80.6569 düzeyine getirdiğinde aramayı kestiği görülmektedir.

Oluşan karar ağaçlarının yaprak sayıları incelendiğinde, tüm değişkenlerin analize dahil edildiği durumda 184 yaprak oluşurken, oluşturulan değişken alt kümelerinin katıldığı analizler sonucunda oluşan ağaçların yaprak sayılarının 13-131 bandında olduğu görülmektedir. Değişken seçim algoritmalarının amaçlarından bir diğeri olan analizi daha anlaşılır hale getirme konusunda ise yine tüm algoritmaların başarılı olduğu söylenebilir.

Tüm değişkenleri içeren veri kümesinin sınıflandırma işlemine girdiği durumda 0.52 saniyede sonuç alınabilirken değişken sayısının azaltıldığı diğer tüm durumlarda bu süre 0.03-0.22 bandında yer almaktadır. Değişken seçimi algoritmalarının amaçlarından biri olan analiz süresini düşürme konusunda tüm yöntemlerin ürettikleri değişken alt kümeleri ile yapılan analizlerde, sürenin azaltılabildiği gözlemlenmektedir.

Sınıflandırma başarısının ise tüm değişkenler ile oluşturulan modelde, yüzde 79.6026 düzeyinde iken değişken seçme algoritmalarının oluşturduğu alt kümeler ile oluşturulan modellerde yüzde 79.9611-82.11 bandında bulunduğu gözlemlenmektedir. Değişken seçim algoritmaları kullanılırken, ortaya çıkacak modelin başarısını iyileştirmek veya düşük düzeyde kayıp ile daha sade bir model ortaya koymak olduğu için belirlenen tüm değişken alt kümelerinin bu düzeyi yakaladığı söylenebilir.

Kappa istatistiği incelendiğinde oluşturulan modeller arasında çok belirgin farklar oluştuğu söylenememektedir. Tüm algoritmaların seçtiği değişken alt kümelerinin ortaya koydukları modeller üzerinden hemen hemen aynı düzeylerde kappa istatistiği hesaplandığı görülmektedir.

Hata matrislerinde hedef değişkenin değerinin bir olma durumunun model tarafından doğru öngörülen adetler incelendiğinde, tüm değişkenler ile oluşturulan model için 945 iken

değişken alt kümeleri ile oluşturulan modellerde 965-898 bandında doğru öngörü sayıları gözlemlenmektedir. Hedef değişkenin sıfır olduğu durum için tüm değişkenleri içeren veri kümesinde doğru öngörülen sayı 1018 iken oluşturulan değişken alt kümeleri ile oluşturulan modellerde 1024-1099 bandında doğru öngörüldüğü gözlemlenmektedir. Yapılan çalışmada hedef değişken için bir ve sıfır değerini alan eşit sayıda örnek ile çalışıldığı göz önünde bulundurulursa değişken alt kümesi seçim algoritmalarının, hedef değişkenin değerinin sıfır olarak öngörülmesine meyilli modellerin oluşmasını sağladığı söylenebilir.

Tüm yöntemler ile gerçekleştirilen uygulamalar incelendiğinde değişken seçimi yöntemi olarak korelasyon bazlı değişken seçimi yönteminin, arama yöntemi olarak ise kazanç oranı bazlı değişken seçimini kullanan sıralayıcı arama yönteminin uygulamasının yapıldığı yöntem-4, diğer uygulamalar ile karşılaştırıldığında en kısa sürede, en az değişken ile, en sade modelin oluşturulmasını sağlamıştır. Toplam analiz süresinin fazla artmasının istenmediği durumlar için bu yöntemin oldukça makul olduğu söylenebilir.

Sınıflandırma doğruluğunun arttırılmasının amaçlanması halinde ise ileri yönlü BestFirst arama yöntemini ile birlikte sarmalayıcı modeli kullanan yöntem-1 ‘in, uygulanan yöntemler arasındaki en iyi yöntem olduğu görülmektedir.

Sonuç olarak uygulanan tüm yöntemlerin sınıflandırma doğruluğunu arttırdığı, sınıflandırma için gerekli süreyi ve oluşturulan modelin sadeleştirilmesi konusunda başarılı olduğu görülmektedir. Analiz için ayrılan süre ve kaynakların sınırlı olması durumunda kısa sürede kabul edilebilir bir iyileştirme sağladığı gözlemlenen değişken seçimi için filtreleyici modellerden korelasyon bazlı değişken seçimi uygulaması ile birlikte kazanç oranı bazlı sıralayıcı arama tercih edilmesi makul görünmektedir. Kaynakların ve analiz için ayrılan sürenin sınırlarının geniş olduğu durumlar için sarmalayıcı yöntemlerin ileri yönlü BestFirst arama yöntemi ile kullanılmasının makul olduğu söylenebilir.

6. SONUÇ

Yöntem-1 ile yapılan uygulamada diğer yöntemlere görece daha yüksek doğruluk elde etmiştir. Analiz süresinin ön plana çıkması durumunda ise, yöntem-3, yöntem-4, yöntem-6 en kısa sürede sonuç üreten yöntemler olmuşlardır. Yöntem-4 en sade modeli oluşturmuştur. Göz önünde bulundurulan değerlendirme kriterlerinin eşit ağırlıklara sahip oldukları varsayılarak yapılan karşılaştırmaların medyan değeri göz önüne alındığında Yöntem-4 ön plana çıkmıştır. Ancak gerçek bir uygulamada, değerlendirme kriterlerinin ağırlıkları uygulayıcı tarafından belirlenmesi gerektiğinden daha farklı bir yöntem de ön plana çıkabilir.

Gelecek çalışmalarda, değişken alt kümesi seçimi ile birlikte boyut indirgeme (dimension reduction) yaklaşımlarının birlikte uygulanması ile daha yüksek performans elde edilmesine çalışılabilir. Bu amaçla, doğrusal boyut indirgeme yaklaşımı olarak temel bileşenler analizi (PCA) ve doğrusal olmayan manifold öğrenme yaklaşımları kullanılabilir.

KAYNAKÇA

Abraham, R., J. B. Simha & S. S. Iyengar (2009) Effective Discretization and Hybrid feature selection using Naïve Bayesian classifier for Medical datamining. *International Journal of Computational Intelligence Research,* 5**,** 116-129.

Alhaj, T. A., M. M. Siraj, A. Zainal, H. T. Elshoush & F. Elhaj (2016) Feature Selection Using Information Gain for Improved Structural-Based Alert Correlation. *PloS one,* 11**,** e0166017.

Ali, S. I. & W. Shahzad. 2012. A feature subset selection method based on symmetric uncertainty and ant colony optimization. In *Emerging Technologies (ICET), 2012 International Conference on*, 1-6. IEEE.

Başeğmez, H., Sezer, E., & Erol, Ç. S. (2021). Optimization for Gene Selection and Cancer Classification. Multidisciplinary Digital Publishing Institute Proceedings, 74(1), 21.

Blum, A. L. & P. Langley (1997) Selection of relevant features and examples in machine learning. *Artificial intelligence,* 97**,** 245-271.

Das, S. 2001. Filters, wrappers and a boosting-based hybrid for feature selection. In *ICML*, 74-81.

Dash, M. & H. Liu (1997) Feature selection for classification. *Intelligent data analysis,* 1**,** 131-156.

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

Haldorai, A., & Ramu, A. (2021). Canonical correlation analysis based hyper basis feedforward neural network classification for urban sustainability. Neural Processing Letters, 53(4), 2385-2401.

Hall, M. 2000. Correlation Based Feature Selection for Discrete and Numeric Class Machine Learning. In *Proc. 17th Int'l. Conf. Machine Learning*.

Hall, M. A. 1999. Correlation-based feature selection for machine learning. The University of Waikato.

Hall, M. A. & Holmes, G. (2000) Benchmarking attribute selection techniques for data mining.

Hall, M. A., & Holmes, G. (2003) Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data engineering,* 15**,** 1437-1447.

Hall, M. A. & L. A. Smith (1997) Feature subset selection: a correlation based filter approach.

Jungjit, S. 2016. New Multi-Label Correlation-Based Feature Selection Methods for Multi-Label Classification and Application in Bioinformatics. University of Kent.

Kantardzic, M. 2011. *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons.

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

Kohavi, R. & G. H. John (1997) Wrappers for feature subset selection. *Artificial intelligence,* 97**,** 273-324.

Ladha, L. & T. Deepa (2011) Feature selection methods and algorithms. *International journal on computer science and engineering,* 3**,** 1787-1797.

Liu, H. & H. Motoda. 1998a. *Feature extraction, construction and selection: A data mining perspective*. Springer Science & Business Media.

Liu, H. & H. Motoda. 1998b. *Feature selection for knowledge discovery and data mining*. Springer Science & Business Media.

Mansour, N. A., Saleh, A. I., Badawy, M., & Ali, H. A. (2022). Accurate detection of Covid-19 patients based on Feature Correlated Naïve Bayes (FCNB) classification strategy. Journal of ambient intelligence and humanized computing, 13(1), 41-73.

Omuya, E. O., Okeyo, G. O., & Kimwele, M. W. (2021). Feature selection for classification using principal component analysis and information gain. Expert Systems with Applications, 174, 114765.

Pedrycz, W., G. Succi & A. Sillitti. 2016. *Computational Intelligence and Quantitative Software Engineering*. Springer.

Piroonratana, T., W. Wongseree, T. Usavanarong, A. Assawamakin, C. Limwongse & N. Chaiyaratana. 2010. Identification of Ancestry Informative Markers from Chromosome-Wide Single Nucleotide Polymorphisms Using Symmetrical Uncertainty Ranking. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, 2448-2451. IEEE.

Priyadarsini, R. P., M. Valarmathi & S. Sivakumari (2011) Gain ratio based feature selection method for privacy preservation. *ICTACT J. Soft Comput,* 1**,** 201-205.

SaiSindhuTheja, R., & Shyam, G. K. (2021). An efficient metaheuristic algorithm based feature selection and recurrent neural network for DoS attack detection in cloud computing environment. Applied Soft Computing, 100, 106997.

Shahbaz, M. B., X. Wang, A. Behnad & J. Samarabandu. 2016. On efficiency enhancement of the correlation-based feature selection for intrusion detection systems. In *Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2016 IEEE 7th Annual*, 1-7. IEEE.

Song, X. F., Zhang, Y., Gong, D. W., & Gao, X. Z. (2021). A fast hybrid feature selection based on correlation-guided clustering and particle swarm optimization for high-dimensional data. IEEE Transactions on Cybernetics.

Sumaiya Thaseen, I., Saira Banu, J., Lavanya, K., Rukunuddin Ghalib, M., & Abhishek, K. (2021). An integrated intrusion detection system using correlation-based attribute selection and artificial neural network. Transactions on Emerging Telecommunications Technologies, 32(2), e4014.

Sun, L., Wang, T., Ding, W., Xu, J., & Lin, Y. (2021). Feature selection using Fisher score and multilabel neighborhood rough sets for multilabel classification. Information Sciences, 578, 887-912.

Sun, L., Yin, T., Ding, W., Qian, Y., & Xu, J. (2021). Feature selection with missing labels using multilabel fuzzy neighborhood rough sets and maximum relevance minimum redundancy. IEEE Transactions on Fuzzy Systems, 30(5), 1197-1211.

Sun, Y., F. Wang, B. Wang, Q. Chen, N. Engerer & Z. Mi (2016) Correlation Feature Selection and Mutual Information Theory Based Quantitative Research on Meteorological Impact Factors of Module Temperature for Solar Photovoltaic Systems. *Energies,* 10**,** 7.

Wang, Z., Gao, S., Zhou, M., Sato, S., Cheng, J., & Wang, J. (2022). Information-Theory-based Nondominated Sorting Ant Colony Optimization for Multiobjective Feature Selection in Classification. IEEE Transactions on Cybernetics.

Witten, I. H., E. Frank & M. A. Hall (2011) Data Mining: Practical Machine Learning Tools and Techniques.

Yu, L. & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, 856-863.

Yu, L. & Liu, H. (2004) Efficient feature selection via analysis of relevance and redundancy. *Journal of machine learning research,* 5**,** 1205-1224.