Fırat Üniversitesi Fen Bil. Dergisi
34(1), 31-37, 2022

*Araştırma Makalesi*

# Siber Saldırılar için Rastgele Orman Algoritması Kullanılarak Öznitelik Seçimi

**Abdulkadir BİLEN[1*], Ahmet Bedri ÖZER[2]**
[1] Emniyet Genel Müdürlüğü, Ankara, Türkiye
[2] Bilgisayar Mühendisliği, Mühendislik Fakültesi, Fırat Üniversitesi, Elazığ, Türkiye
[*1] abdulkadir.bilen82@gmail.com, [2] bozer@firat.edu.tr

**(Geliş/Received: 02/11/2021; Kabul/Accepted: 29/11/2021)**

**Öz:** Veri boyutlarındaki artışla birlikte araştırmacılar analiz aşamasını daha kolay hale getirmek için çeşitli yöntemlere ihtiyaç duymuşlardır. Veri boyutunu indirgemek ve analiz doğruluğu artırmak önem arz etmektedir. Veri analiz edilirken gereksiz alanlarla uğraşmamak ve daha az girdi ile daha doğru sonuç çıkarmak gerekmektedir. Öznitelik seçimi veri analizindeki en önemli aşamalardan birisidir. Öznitelik seçerken çeşitli makine öğrenmesi yöntemleri kullanılmaktadır. Çalışmada Tek Değişkenli Öznitelik seçimi, Özyinelemeli Öznitelik Eleme, Ağaç Tabanlı Öznitelik Seçimi ve Temel Bileşen Analizi yöntemleri kullanılmıştır. Bu yöntemlerle veri setindeki 13 öznitelik içinden en önemli olanları tespit edilmiştir. En önemli 6, 5 ve 4 öznitelik ayrı ayrı girdi olarak Rastgele Orman algoritması ile siber saldırı yöntemi tahmini yapılmıştır. Öznitelik sayısı 4'e indirgendiğinde en yüksek doğruluk oranı olan %97.24 elde edilmiştir. Bu oran, tahmin yapılırken öznitelik seçiminde ilişkili özniteliklerin kullanılmasının boyut ve hız açısından önemli olduğunu göstermiştir. Elde edilen sonuçlarla birlikte öznitelik seçiminin veri üzerindeki önemi bir kez daha ortaya konulmuştur.

**Anahtar kelimeler:** Öznitelik Seçimi, Rastgele Orman, Makine Öğrenmesi, Veri Analizi

## Feature Selection Using Random Forest Algorithm to Cyber Attack

**Abstract:** With the increase in data sizes, researchers needed various methods to make the analysis process easier. It is important to reduce the data size and increase the analysis accuracy. When analyzing data, it is necessary not to deal with unnecessary fields and to produce more accurate results with less input. It is one of the most important first steps in feature selection and data analysis. Various machine learning methods are used for feature selection. Univariate Feature Selection, Recursive Feature Elimination, Tree-Based Feature Selection and Principal Component Analysis methods were used in the study. With these methods, the most important ones among the 13 features in the data set were determined. The most important 6, 5, and 4 attributes were separately input, and the cyber-attack method was predicted with the Random Forest algorithm. When the number of features was reduced to 4, the highest accuracy rate of 97.24% was obtained. It has been concluded that the inclusion of related features in the estimation is important in terms of size and speed in this ratio feature selection. With the results obtained, the importance of feature selection on the data has been demonstrated once again.

**Key words:** Feature Selection, Random Forest, Machine Learning, Data Analysis

## 1. Giriş

Verilerin boyutları arttıkça makine öğrenmesine dayalı yöntemlerin başarısı ve hızı önemli ölçüde etkilenmektedir. Öznitelik seçimi, istatistik, örüntü tanıma, veri madenciliği gibi birçok alanda zorlu bir araştırma konusu olmuştur. Veri ön işleme stratejisi olan öznitelik seçimi, daha basit ve anlaşılır modeller oluşturmada, daha anlaşılabilir veri hazırlamada ve sonuç performansını artırmada faydalar sağlamaktadır [1]. Öznitelik seçimi, gürültülü, ilgisiz ve gereksiz özellikleri ortadan kaldırmak için görüntü tanımada [2], kısaltmalar, yanlış yazımlar ve eş anlamlı kelimeler gibi dildeki zorlukları gidermek için metin madenciliğinde [3], görüntüyü sınıflandırma ve bölümleme için görüntü analizinde [4], işlem maliyetini azaltmak, depolamayı en aza indirmek ve test verilerinin daha iyi anlaşılmasını sağlam için izinsiz giriş tespitinde [5], veri analizini ve yorumlanmasını önemli ölçüde azaltmak için [6] uzun yıllardır metodolojide ve pratikte kullanılmaktadır.

Veri işlemede, aşırı öğrenmeyi önlemek ve sonuçların genelleştirilmesini değerlendirmek için eğitim, doğrulama ve test kümelerine ayrılan verilerin çapraz doğrulaması yapılmaktadır. Öznitelik seçimi modelin önemli bir ön adımıdır çünkü değerlendirmesi kötü yapılan ve alakasız veriyi içeren öznitelik seçimi modelin başarısını ciddi şekilde etkilemektedir [7]. Girdi boyutundaki azalma, ya öğrenme hızını ve model karmaşıklığını azaltarak ya da genelleştirme kapasitesini ve sınıflandırma doğruluğunu artırarak performansı geliştirebilmektedir. Uygun

[*] Sorumlu yazar: abdulkadir.bilen82@gmail.com. Yazarların ORCID Numarası: [1] 0000-0003-2359-8829, [2] 0000-0002-8005-7386

özniteliklerin seçilmesi aynı zamanda ölçüm maliyetini de azaltmaktadır. Bu gibi avantajlarından dolayı öznitelik seçimi, çoğunlukla sınıflandırma ve regresyon gibi gerçek dünya problemlerinde aktif olarak kullanılmaktadır [8].

Etiket bilgilerin kullanılabilirliği açısından öznitelik seçimi tekniği, denetimli yöntemler, yarı denetimli yöntemler ve denetimsiz yöntemler olmak üzere üç ana kategoriye ayrılmaktadır. Arama stratejilerine dayalı olarak öznitelik seçimi ise, filtre yöntemleri, sarmalayıcı yöntemler ve gömülü yöntemler olarak sınıflandırılmaktadır [9]. Denetimli öznitelik seçimi, sınıflardan ve regresyon hedeflerinden veri örneklerini ayırt ederek sınıf etiketlerinin korelasyonuyla değerlendirilmektedir. Denetimsiz öznitelik seçimi, genellikle kümeleme görevlerinde veri benzerliği, ayırt edici bilgi ve yeni veri oluşturma hatası gibi kriterlere göre özniteliğini önemini değerlendirmektedir. Sıralama yöntemlerine dayalı olan filtre yöntemleri herhangi bir öğrenme yöntemine dayanmadığı için daha verimlidir [10]. Sarmalayıcı yöntemler, özniteliğinin ilgi düzeyini değerlendirmek için önceden tanımlanmış modelin öğrenme performansını kullanmaktadır. Gömülü yöntemler, model öğrenme ile öznitelik seçimini gömerek filtre ve sarmalayıcı yöntemler arasında değişim çözümü sağlamaktadır [10]. Öznitelik seçimi on yıllardır uygulanan bir makine öğrenmesi alanı olmasına rağmen büyük veri ve yeni senaryolardan dolayı hala gündemdedir. Verinin yalnızca büyük hacimli veya akan veri olması değil aynı zamanda dengesiz sınıflara, belirsiz etiketlere ve durağan olmayan dağılımlara sahip olması da öznitelik seçimini kullanmayı gerektirmektedir [11].

Öznitelik seçimi yapılırken literatürde çeşitli makine öğrenmesi yöntemleri kullanılmaktadır. Bu yöntemlerden Rastgele Orman (RO) algoritması yaygın olarak seçilmektedir. Kumar ve Shaikh tarafından RO sınıflandırıcısı ile birlikte, Relief öznitelik seçim, Rastgele orman seçim, Recursive öznitelik seçim ve Boruta öznitelik seçim algoritmaları kullanılmıştır. Araştırılan bu yöntemlerle tüm öznitelik seçim yaklaşımlarında öğrenme algoritmasının performansının arttığı gözlemlenmiştir. Çalışmanın sonucunda ise Rastgele ormana dayalı Boruta öznitelik seçim yöntemi en iyi sonucu vermiş olsa da dört öznitelik seçiminde de 9 özellik kümesi üretilmiştir [12]. Yadav ve Pal tarafından yapılan çalışmada 14 özniteliğe sahip kalp hastalığı ile ilgili veriler kullanılmıştır. Sınıflandırma doğruluğu, kesinliği ve hassasiyeti analiz edildikten sonra Pearson korelasyonu ve Rastgele orman topluğuna sahip Lasso düzenleme öznitelik seçim yöntemlerinde en iyi sonuç alınmıştır [13].

Hasan ve arkadaşları tarafından girdi özniteliklerini azaltarak izinsiz giriş tespit performansını artırmaya yönelik saldırı tespit sistemi için rastgele orman modeli önerilmiştir. 42 öznitelik ve 25 öznitelik kullanılarak yapılan bu çalışmada rastgele orman algoritmasıyla 25 özniteliğin işleme süresi daha kısa bulunmuştur. Önerilen yaklaşımda sınıflandırma için yararlı olan önemli ve alakalı öznitelikleri seçtiği ve bununda süreyi azaltmanın yanı sıra doğruluğu da artırdığı gözlemlenmiştir [14]. Li ve arkadaşları tarafından sınıflandırma doğruluğunu ispatlamak ve eğitim süresini azaltmak için rastgele orman algoritması tabanlı otomatik kodlayıcı saldırı tespit sistemi önerilmiştir. Öznitelik seçimi, öznitelik gruplandırması ve eğitim kümesinden oluşan bu yöntemle oluşturulan modelin algılama süresini azalttığı, tahmin doğruluğunu da artırdığı görülmüştür [15]. Yapılan araştırmalarda öznitelik seçimin makine öğrenmesi yöntemlerine fayda sağladığı ve özellikle rastgele orman algoritmasının başarılı olduğu tespit edilmiştir. Çalışmanın ilk bölümünde öznitelik seçiminin literatür ve hangi alanlarda uygulandığı, ikinci bölümde öznitelik seçimi yöntemlerinin neler olduğu, üçüncü bölümde örnek veri setinde öznitelik seçimin başarı durumu, son bölümde ise sonuçlar verilmiştir.

## 2. Öznitelik seçimi

Kümeleme ve sınıflandırma yöntemlerinin performansı, veri boyutlarının artmasından önemli derecede etkilenmektedir. Yüksek boyutlu verinin model oluşturma süresi ve analizi zordur. Öznitelik seçimi büyük veri kümesinden özniteliklerin alt kümesini belirlemek ve gereksiz özellikleri kaldırarak modeli daha doğru hale getirmek için kullanılan bir ön işlem sürecidir. Öznitelik seçme yöntemleri, bir arama tekniğine ve alt kümelerin performans değerlendirmesine dayanmaktadır. Seçilen özniteliklerin sayısını azaltmak veya ortadan kaldırmak ve çıktı doğruluğu performansını artırmak gibi iki önemli amacı gerçekleştirmelidir [16]. Öznitelik seçiminde denetimli, denetimsiz ve bu ikisi arasında yer alan yarı denetimli yöntemler bulunmaktadır. Öznitelik seçiminde bir model çıkarmak için kullanılan tümevarım öğrenme yöntemi arasındaki ilişkiye göre üç ana yaklaşım bulunmaktadır. Verinin genel karakterine bağlı olan ve tümevarım algoritmasından bağımsız olan filtreler, özniteliklerin altkümesini değerlendirmek için sınıflandırıcıyla elde edilen tahmini kullanan sarmalayıcılar ve eğitim sürecinde öznitelik seçimi gerçekleştiren ve öğrenme makinelerine özgü olan gömülü yöntemlerdir [17].

### 2.1. Denetimli

Denetimli öznitelik seçimi genellikle sınıflandırma problemlerine yönelik olup öznitelik ile sınıf etiketi arasındaki ilişkiyi veya korelasyonu temel ilke olarak kullanmaktadır. Özniteliklerin önemi, uyguluk ölçütleri ile değerlendirilmektedir. Bir öznitelik kümesi $X = (x_1, x_2, \ldots, x_3)$ ve sınıf etiketi C'ye sahip belirli bir $D = (X, C)$ veri kümesi için, denetimli model sınıflandırma doğruluğunu maksimuma çıkaran $\hat{S}(|\hat{S}| = \hat{k})$ optimal öznitelik altkümesini bulmayı amaçlamaktadır [18].

### 2.2. Denetimsiz

Denetimsiz öznitelik seçim yöntemleri, verinin doğan sınıflandırmasını kapsamayı ve kümeleme veya değerlendirme kriterlerine bağlı öznitelik altkümesini bularak kümeleme doğruluğunu artırmayı amaçlamaktadır. Küme algoritmalarına bağlı olup olmadıklarına göre denetimsiz filtre veya sarmalayıcı olabilmektedir. Bu yöntem özniteliklerin kümeleme yeteneğine göre en önemli özniteliklerden optimal özniteliğin altkümesini aramaktadır. Öznitelik araması, yeni seçilen özniteliğin mevcut kümeleme sonuçlarını değiştiremeyene kadar devam etmektedir. Bu yöntemin amacı öznitelik seçiminin geçerliliğinin ve tahmin doğruluğunu geliştirmektir [18].

### 2.3. Yarı denetimli

$D = (D_l, D_u)$ veri kümesi göz önüne alındığında, $D_l$ sınıf etiketlerine sahip örnek kümesidir ve $D_u$, sınıf etiketi olmayan örnek kümesidir. Yarı denetimli öznitelik seçimi, $D_l$ tarafından eğitilen öğrenme modelinin performansını iyileştirmek için $D_u$'yu kullanmaktadır. Başlıca filtre modeli olan yarı denetimli öznitelik seçim yöntemleri, yarı denetimli öğrenmede kilit rol onamaktadır. Puan işlevleri yarı denetimli öznitelik seçim yöntemlerinin çoğunda uygulanmaktadır ve varyans puanı, Laplacian puanı, Fisher puanı ve kısıtlama puanı olmak üzere dört kategoride puanlama yapmaktadır [18].

### 2.4. Filtreleme

Filtre öznitelik seçim yaklaşımında, özniteliğin uygunluğu veri setinin istatistiksel ve gerçek karakteri aracılığıyla değerlendirilmektedir. Bu özelliklere dayanarak uygun bir öznitelik makine öğrenmesi ve veri madenciliğinde seçilmektedir. Filtre metodu öznitelik seçimi için temel kriter olarak sıralama tekniklerini kullanmaktadır. Uygun sıralama kriteri öznitelikleri puanladıktan sonra kullanılmaktadır. Filtre yöntemi daha az alakalı özellikleri filtrelemek için sıralamadan önce kullanılmaktadır. Benzersiz bir öznitelik veriyle ilgili alakalı bilgiyi içermektedir [19].

### 2.5. Sarmalayıcı

Sarmalayıcı yöntemde, öznitelikleri altkümesi oluşturulmakta ve belli sınıflandırıcıların altındaki değerlendirmeye alınmaktadır. Genellikle bu yöntemde öznitelik altkümesini değerlendirmek için belirli bir amaç fonksiyonu kullanılmaktadır. $2^N$ alt kümelerini değerlendirirken NP zor bir problem haline geldiğinden optimal altkümeler, altkümeleri sezgisel olarak bulmak için arama algoritmalarıyla araştırılmaktadır. Birçok arama algoritması ilişkili amaç fonksiyonunu minimize veya maksimize ederek öznitelik altkümesini bulmak için kullanılmaktadır. Sarmalayıcı yöntemler, belli sınıflandırıcılar ile ilişkili olduğundan onun hesaplama karmaşıklığı genellikle diğer iki kategorideki öznitelik seçim yöntemlerinden daha yüksektir [19].

### 2.6. Gömülü

Gömülü yöntemde, optimizasyon problemleri iz düşüm alt uzayında gömülü öznitelikleri sıralamak için kullanılmaktadır. Bu sıralama prosedürüne bağlı olarak, öznitelikler özel uygulamalar için seçilmektedir. bu yöntem farklı öznitelik alt kümelerini yeniden sınıflandırmak için harcanan zamanı azaltmayı amaçlamaktadır. Temel amaç öğrenme sürecinin bir parçası olarak öznitelik seçimini dahil etmektir [19].

### 2.7. Rastgele Orman Algoritması

RO, sınıflandırma ve regresyon görevleri için çok popüler ve oldukça hassas olan ve model toplama fikrine dayalı bir öğrenme algoritmasıdır. RO algoritmasının arkasındaki ana fikir, her ağacın bir sınıf için oy kullandığı

ve ormanın, ormandaki tüm ağaçlar arasından en çok oyu alan sınıflandırmayı seçen değiştirme yöntemiyle eğitim verilerinin rastgele örneklerinden çok sayıda tarafsız karar ağacını üretmektir. Yöntemin en önemli avantajlarından biri, RO'ların sınıf tahminine ilişkin her özniteliğin etkisini öğrenmek için özniteliklerin önem puanını ölçebilmesidir. Fakat yüksek boyutlu problem için, özniteliklerin sayısı çok büyük olabilmektedir bu da öznitelik önem puanlarının manuel olarak araştırılmasını ve sınıflandırma için en uygun özniteliklerin seçimini çok zorlaştırmaktadır. Bu açıdan önem puanına bağlı otomatik öznitelik seçim prosedürü ilgili, etkili ve ayırt edici özniteliklerin seçilmesine yol açmaktadır. RO, rastgele seçilen karar ağaçlarına dayanan topluluk öğrenicisidir, topluluktaki her ağaç orijinal eğitim verisinin değiştirilmesiyle orijinal verinin rastgele bir örneğinden oluşturulmaktadır. Karar ağaçlarını oluşturmak ve her ağaçtaki son sınıfı belirlemek içi gini indeksi kullanılmaktadır. Dolayısıyla $\upsilon$ düğümündeki gini indeksi $Gini\ (\upsilon), \upsilon$'nin saflığını ölçmektedir. Denklem 1'deki formülle ifade edilmektedir [20].

$$Gini\ (\upsilon) = \sum_{i=1}^{I} f_i(1 - f_i) \tag{1}$$

Burada $f_i, \upsilon$ düğümünde kaydedilen i sınıfının kesridir. Yine $\upsilon$ ağaç düğümünü bölmek için $X_i$ özniteliğinin gini bilgi kazancı Denklem 2'deki formülde tanımlanmıştır.

$$gain\ (X_i, \upsilon) = Gini\ (X_i, \upsilon)\ -\ (W_L\ Gini\ (X_i,\ \upsilon^L)\ +\ (W_R\ Gini\ (X_i,\ \upsilon^R)) \tag{2}$$

Burada $Gini\ (X_i, \upsilon), \upsilon$ düğümündeki kirliliktir. $W_L$ ve $W_R$, sırasıyla $\upsilon$ düğümünün sol ve sağ alt düğümleridir. . $W_L$ ve $W_R$, sol ve sağ alt düğüme atanan örneklerin kesridir. Son olarak, kirlilikteki azalmayı maksimize eden öznitelik ayırma özniteliği olarak kullanılmaktadır. $X_i$ özniteliği için önem puanı Denklem 3'teki gibi $gain\ (X_i, \upsilon)$'den hesaplanmaktadır.

$$Imp_i = \frac{1}{n_{tree}} \sum_{k \in S_{X_i}} gain\ (X_i, \upsilon) \tag{3}$$

Burada $n_{tree}$, RO'daki veya topluluk boyutundaki ağaçların sayısıdır ve $k \in S_{X_i}$, bölünmüş düğümler kümesidir. Önem skorunun normalleştirilmesi Denklem 4'deki gibidir.

$$Imp_{norm} = \frac{Imp_i}{Imp_{max}} \tag{4}$$

Burada $Imp_i$, RO'dan $X_i$'nin önem puanını temsil etmektedir. $Imp_{max}$, maksimum önemi temsil etmektedir ve normalleştirilmiş önem puanı $0 \leq Imp_{norm} \geq 1$ arasında yer almaktadır [20].

## 3. Bulgular ve Tartışma

### 3.1. Veri öznitelikleri

Çalışmada Elazığ ilinde meydana gelen siber suç verileri kullanılmıştır. Veri setinde 14 farklı öznitelik bulunmaktadır. Meydana gelen olayla ilgili olan bu öznitelikler mağdurun cinsiyeti, yaşı, yaş aralığı, eğitimi, mesleği, geliri, medeni hali, suçun türü, suçun yılı, saldırının amacı, saldırının zararı, saldırının yöntemi, failin bilinme durumu ve failin yakalanma durumundan oluşmaktadır.

### 3.2. Rastgele Orman Yöntemi ile Tahmin

Veri seti öncelikle %80 eğitim ve %20 test verisi olarak ayrılmıştır. Mağdurun cinsiyeti, yaşı, yaş aralığı, eğitimi, mesleği, geliri, medeni hali, suçun türü, saldırının amacı, saldırının zararı, failin bilinme durumu ve failin yakalanma durumu öznitelikleri girdi olarak verilmiş, saldırı yöntemi tahmin edilmiştir. Veri setindeki 13 öznitelik girdi olarak verildiğinde, saldırı yönteminin tahmin edilmesinde rastgele orman sınıflandırması ile %95.03 doğru tahmin yapılmıştır.

Öznitelikler, Tek Değişkenli Öznitelik seçimi, Özyinelemeli Öznitelik Eleme, Ağaç Tabanlı Öznitelik Seçimi ve Temel Bileşen Analizi yöntemlerine göre öznitelik seçimi yapılmıştır. En güçlü öznitelikleri tespit etmek için yapılan Tek Değişkenli Öznitelik seçiminde en yüksek puan alan k özelliği dışındaki tüm özellikleri kaldıran

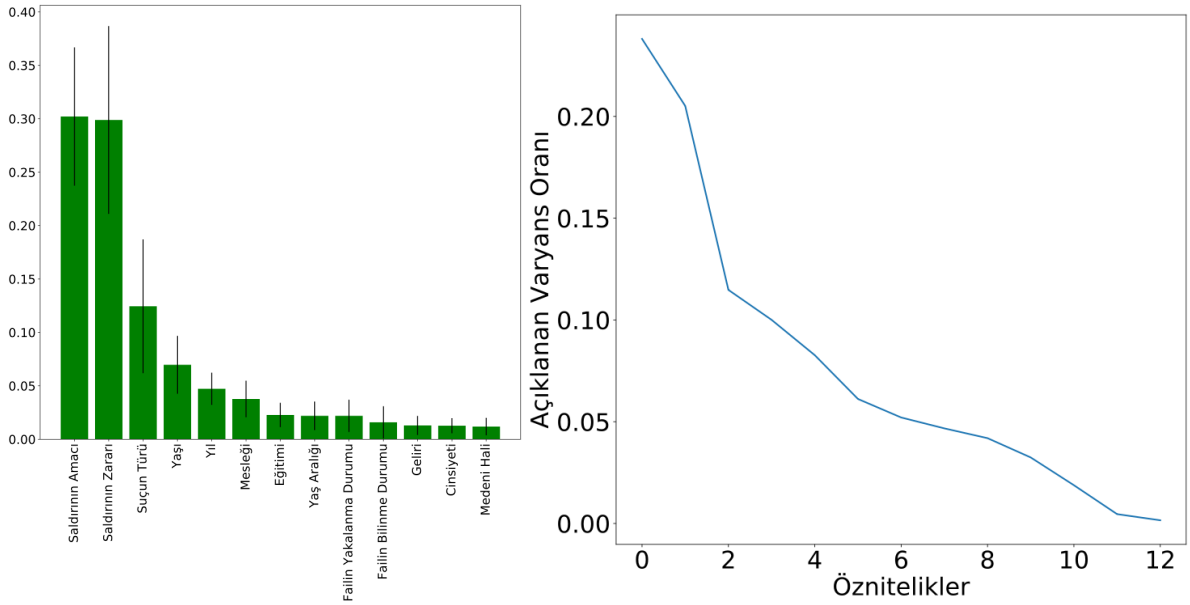
SelectKBest yöntemi kullanılmıştır ve özniteliklerin puanları Tablo3.1'de verilmiştir. Saldırının amacı, mağdurun yaşı, suçun türü, saldırının zararı ve mağdurun mesleği öznitelikleri 100 puanın üzerinde puanlanmıştır.

**Tablo 3.1.** Model sonuçları

| **Saldırının Amacı** | **Mağdurun Yaşı** | **Suçun Türü** | **Saldırının Zararı** | **Mağdurun Mesleği** | **Suçun Yılı** | **Yaş Aralığı** |
|---|---|---|---|---|---|---|
| 442 | 379 | 293 | 222 | 138 | 82 | 58 |
| **Failin Bilinme Durumu** | **Mağdurun Cinsiyeti** | **Mağdurun Medeni Hali** | **Mağdurun Geliri** | **Failin Yakalanma Durumu** | **Mağdurun Eğitimi** | |
| 31 | 27 | 16 | 15 | 14 | 12 | |

Özyinelemeli öznitelik eleme (Recursive feature elimination-RFE) yönteminde, bir modele uyan ve belirlenen öznitelik sayısına ulaşılana kadar en zayıf özniteliği (veya öznitelikleri) ortadan kaldıran bir öznitelik seçme yöntemidir. Bu yöntemde Saldırının amacı, saldırının türü, saldırının zararı, mağdurun yaşı, mağdurun mesleği ve suçun yılı en güçlü 6 öznitelik olarak elde edilmiştir.

Ağaç tabanlı öznitelik önem sıralaması Şekil 3.2A'da görüldüğü üzere yine aynı öznitelikler sıralamada önde yer almıştır. Temel Bileşen Analizinde yüksek miktar ve boyuttaki verilerin yüksek ilişkili olanları belirlemektedir ve verideki temel öznitelikleri tespit ederek daha az sayıya indirgemektedir. Bu analize göre saldırının amacı, suçun türü, saldırının zararı ve mağdurun yaşı öznitelikleri ön plana çıkmaktadır ve grafik Şekil 3.2B'de verilmiştir. Her iki şekilde de öznitelik açısından önemli olanlar grafiklerde ön sırada yer almıştır. Saldırı yöntemi ile saldırının amacı, zararı ve türü arasında diğer özniteliklere göre önemli derecede ilişki bulunmaktadır.

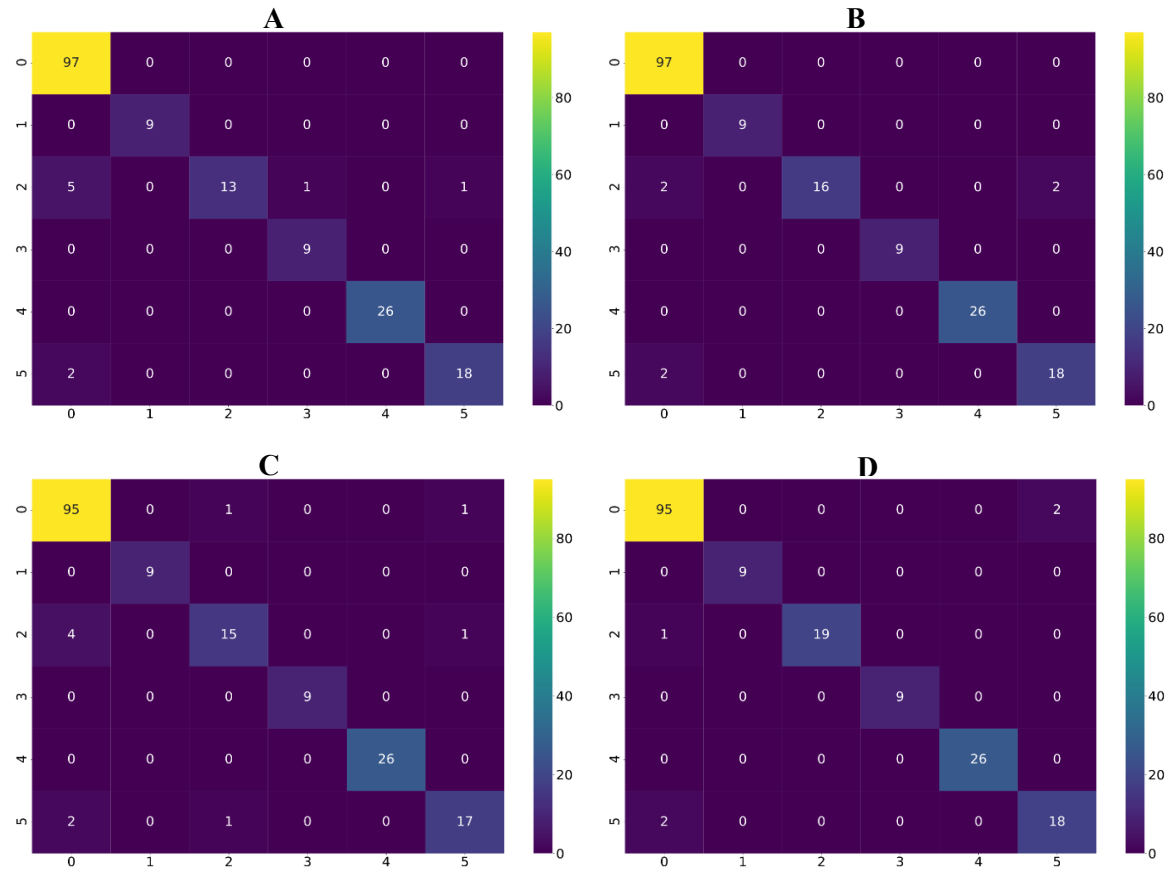



**Şekil 3.1. A)** Öznitelik önem sıralaması **B)** Temel bileşen analizi sıralaması

**Tablo 3.2.** Model sonuçları

| **RO Tahmini** | **Eğitim Doğruluğu (%)** | **Eğitim Süresi (Sn)** | **Test Doğruluğu (%)** | **Test Süresi (Sn)** |
|---|---|---|---|---|
| **13 Öznitelik** | 99.58 | 0.06249159 | 95.03 | 0.01563142 |
| **6 Öznitelik** | 99.30 | 0.04688907 | 96.69 | 0.01561856 |
| **5 Öznitelik** | 98.75 | 0.04687858 | 94.48 | 0.01559711 |
| **4 Öznitelik** | 99.00 | 0.04686189 | 97.24 | 0.01558018 |

Dört farklı öznitelik seçme yöntemi ile öznitelik sayıları azaltılmıştır. RO algoritması ile saldırı yöntemi tahmin edilmiştir. Tablo 3.2'de gösterildiği gibi toplam 13 öznitelik girdi olarak verildiğinde doğruluk oranı %95.03; ilk 6 öznitelik için %96.69; ilk 5 öznitelik için %94.48; ve ilk 4 öznitelik için %97.24 olarak elde edilmiştir. Saldırının amacı, suçun türü, saldırının zararı ve mağdurun yaşı girdi olarak verildiğinde saldırı yöntemini en yüksek oranda tahmin edilmiştir. Yine eğitim süresi ve test süresi açısından değerlendirildiğinde aralarındaki süre miktarları çok küçük olsa bile öznitelik sayısı azaldıkça sürelerinde azaldığı tespit edilmiştir. Bu değerlerde öznitelik seçiminin önemini bir kez daha ortaya koymuştur. Şekil 3.1 (A, B, C, D'deki) hata matrisine bakıldığında hata oranlarının oldukça düşük olduğu ve tahmin oranlarının birbirine yakın olduğu görülmektedir.




**Şekil 3.2. A)** 13 Öznitelikle yapılan tahminin hata matrisi **B)** 6 Öznitelikle yapılan tahminin hata matrisi **C)** 5 Öznitelikle yapılan tahminin hata matrisi **D)** 4 Öznitelikle yapılan tahminin hata matrisi (Burada 0 = Hack Araçları veya Zararlı Yazılım Kullanarak, 1 = Kart Kopyalama, Üretme Cihazlarını Kullanarak, 2 = Phising (oltalama) Saldırısı Kullanarak, 3= Sahte Alışveriş Sitesi Oluşturarak, 4= Sosyal Medyadaki Herkese Açık Verilerini Alarak, 5= Sosyal Mühendislik Kullanarak olarak sayısallaştırılmıştır.)

## 4. Sonuçlar

Öznitelik seçimi veri boyutu arttıkça daha da önem kazanmaktadır. Çok yüksek boyutlu veri ile çalışılırken eğitim süreside veri boyuna göre artmaktadır. Eğitim süresini azaltmaya ve doğruluk oranlarını da artırmaya yarayan öznitelik seçiminin faydaları literatür çalışmalarından da görülmektedir. Öznitelik seçiminde kullanılan yöntemler arasında bazı farklılıklar olsa da genellikle aynı öznitelikler ön plana çıkmıştır. Çalışmada kullanılan veri setindeki en önemli öznitelikler saldırının amacı, suçun türü, saldırının zararı ve mağdurun yaşı tespit edilmiştir. RO algoritması ile seçilen öznitelik yönteminin başarılı olduğu görülmüştür. RO algoritması ile yapılan tahminde en önemli 4 öznitelikle %97.24 doğruluk oranıyla en başarılı tahmin yapılmıştır. Yapılan uygulama ile

öznitelik seçimin hayati öneme sahip olduğu ve doğruluk oranını %2 civarında artırdığı tespit edilmiştir. Birçok sınıflandırma probleminde öznitelik seçim yöntemlerinin kullanılabileceği değerlendirilmektedir.

## Kaynaklar


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

[2] Zhang, Y., Wang, X., Cai, Z., Zhou, Y., & Philip, S. Y. (2021, July). Tensor-Based Unsupervised Multi-View Feature Selection for Image Recognition. In 2021 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.

[3] Hossny, A. H., Mitchell, L., Lothian, N., & Osborne, G. (2020). Feature selection methods for event detection in Twitter: a text mining approach. Social Network Analysis and Mining, 10(1), 1-15.

[4] Bolón-Canedo, V., & Remeseiro, B. (2020). Feature selection in image analysis: a survey. Artificial Intelligence Review, 53(4), 2905-2931.

[5] Alazzam, H., Sharieh, A., & Sabri, K. E. (2020). A feature selection algorithm for intrusion detection system based on pigeon inspired optimizer. Expert systems with applications, 148, 113249.

[6] Liang, S., Ma, A., Yang, S., Wang, Y., & Ma, Q. (2018). A review of matched-pairs feature selection methods for gene expression data analysis. Computational and structural biotechnology journal, 16, 88-97.

[7] Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: Introduction and review. Journal of biomedical informatics, 85, 189-203.

[8] Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. Computers in biology and medicine, 112, 103375.

[9] Miao, J., & Niu, L. (2016). A survey on feature selection. Procedia Computer Science, 91, 919-926.

[10] Li, J., & Liu, H. (2017). Challenges of feature selection for big data analytics. IEEE Intelligent Systems, 32(2), 9-15.

[11] Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2016). Feature selection for high-dimensional data. Progress in Artificial Intelligence, 5(2), 65-75.

[12] Kumar, S. S., & Shaikh, T. (2017, September). Empirical evaluation of the performance of feature selection approaches on random forest. In 2017 international conference on computer and applications (ICCA) (pp. 227-231). IEEE.

[13] Yadav, D. C., & Pal, S. A. U. R. A. B. H. (2020). Prediction of heart disease using feature selection and random forest ensemble method. International Journal of Pharmaceutical Research, 12(4), 56-66.

[14] Hasan, M. A. M., Nasser, M., Ahmad, S., & Molla, K. I. (2016). Feature selection for intrusion detection using random forest. Journal of information security, 7(3), 129-140.

[15] Li, X., Chen, W., Zhang, Q., & Wu, L. (2020). Building auto-encoder intrusion detection system based on random forest feature selection. Computers & Security, 95, 101851.

[16] El-Hasnony, I. M., Barakat, S. I., Elhoseny, M., & Mostafa, R. R. (2020). Improved feature selection model for big data analytics. IEEE Access, 8, 66989-67004.

[17] Bolón-Canedo, V., & Alonso-Betanzos, A. (2019). Ensembles for feature selection: A review and future trends. Information Fusion, 52, 1-12.

[18] Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. Neurocomputing, 300, 70-79.

[19] Zhang, R., Nie, F., Li, X., & Wei, X. (2019). Feature selection with multi-view data: A survey. Information Fusion, 50, 158-167.

[20] Uddin, M. T., & Uddiny, M. A. (2015, May). A guided random forest based feature selection approach for activity recognition. In 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT) (pp. 1-6). IEEE.