Fırat Üniversitesi Müh. Bil. Dergisi
34(1), 115-121, 2022

Araştırma Makalesi
https://doi.org/10.35234/fumbd.973038

Regresyon Yöntemlerine Dayalı Suç Tespit Analizi Karşılaştırması Elazığ İli Örneği

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: 27/07/2021; Kabul/Accepted: 10/10/2021)

Öz: Ülkelerin ve toplumların önce gelen sorunlarından biri olan suçu önlemek, devletin ilk görevleri arasındadır. Bu suçların önemli bir türü siber suçtur. Siber suçlarla mücadele edebilmek için öncelikle bu suçun nasıl gerçekleştiğini ve yöntemini bilmek gerekmektedir. Siber saldırıları önceden tahmin etmek kişilerin ve kurumların uğrayacağı zararları azaltacaktır. Bu tahminleri yapabilmek için lineer regresyon, polinom regresyon, ridge regresyon ve lasso regresyon yöntemlerinden oluşan dört farklı model uygulanmıştır. Elazığ ilinde işlenen siber suçların öznitelikleri çıkarılmış ve bu dört modele dayalı tahminler yapılmıştır. Ortalama mutlak hata (MAE), ortalama kare hatası (MSE), kök ortalama kare hatası (RMSE) ve R Square değerlendirme kriterlerine göre modeller karşılaştırılmıştır. Yapılan uygulama neticesinde 0.79 doğruluk oranıyla kendi içinde en iyi yöntem polinom regresyon sonuç vermiştir. Diğer yöntemlerin başarı oranı çok düşük sonuç vermiştir. Elde edilen sonuçlar suç analizine ve suçla mücadeleye bir ön adım olacaktır.

Anahtar kelimeler: Yapay Zekâ, Regresyon, Suç Analizi, Siber Suç.

Comparison of Crime Detection Analysis Based on Regression Methods The Case of Elazığ

Abstract: Preventing crime, which is one of the foremost problems of countries and societies, is among the first duties of the state. An important type of these crimes is cybercrime. To fight against cybercrimes, it is necessary to know how this crime took place and its method. Predicting cyber-attacks will reduce the damage to individuals and institutions. To make these estimations, four different models consisting of linear regression, polynomial regression, ridge regression and lasso regression methods have been applied. Attributes of cybercrimes committed in Elazığ province were extracted and predictions were made based on these four models. Models were compared according to mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE) and R Square evaluation criteria. As a result of the application, polynomial regression was the best method with an accuracy rate of 0.79. The success rate of other methods gave extremely low results. The results obtained will be a preliminary step towards crime analysis and the fight against crime.

Key words: Artificial Intelligence, Regression, Crime Analysis, Cybercrime.

1. Giriş

Suç toplumların ve ülkelerin önde gelen problemlerinden biridir. Suçun işlenmesi nüfus, eğitim, maddi durum ve işsizlik gibi faktörlere göre değişebilmektedir [1]. Suç oranlarının ve çeşitliliğinin artması suç örüntüsünü değiştirmekte ve suçun önlenmesine ilişkin analiz çalışmalarını zorlaştırmaktadır [2]. Suçla ilgili tüm veriler bazen düzenli veri tabanlarında bazen sosyal medyada bazen de diğer depolama birimlerinde tutulmaktadır. Bu verileri toplamak ve sonrasında anlam çıkarmak için analiz etmek oldukça zorlu bir süreçtir. Suç analizi yapılırken istatistiksel yaklaşım, uzman bilgi yaklaşımı, veri madenciliği teknikleri, kümeleme, birliktelik kuralı madenciliği ve makine öğrenimi gibi yöntemler bulunmaktadır [3]. Suç analizi, suçu tahmin etmek veya suç kayıtlarına göre suçlu grupları oluşturmak için gerçekleştirilmektedir. Elde edilen veriler çeşitli öz işleme süreçlerinden geçirildikten soran metin içeriği, suç faktörleri, suç özellikleri, suçun coğrafi konumu vb. şeklinde olabilecek öznitelikler çıkarılmaktadır. Daha sonra bu yöntemlerle tahmin ya da analiz işlemi yapılmaktadır [4]. Suçlar dolandırıcılık tespiti, trafik şiddeti, şiddet suçu, cinsel suç, siber suç gibi bazı kategorilere ayrılmaktadır. Bu suçlar içerinde siber suç analizi kolluk kuvvetleri için önemli bir sorumluluktur [5]. Yine yüksek doğrulukta bir suç tahmini yapabilmek için suçun doğasını anlamak önemlidir [6]. Matlhare ve arkadaşları tarafından yapılan çalışmada Botswana üniversitesindeki gençlerin siber suçların farkında olduğunu fakat bu farkındalığın yetersiz olduğu ortaya koyulmuştur. Yine siber suçların düşük düzeyde tespit edilmesinden kaynaklı olarak bu suçlarla mücadele ederken kanun koyucular ve kamu-özel sektör iş birliğinin önemi vurgulanmıştır [7].

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

Suçu analiz ederken ve çeşitli tahminler yapılırken makine öğrenmesi yöntemlerin başarılı olduğu görülmüştür. Bhuriya ve arkadaşları tarafından borsa yatırımcılarına yardımcı olmak amacıyla 5 farklı regresyon yöntemi kullanarak hisse senedi fiyatlarını tahmin etmişlerdir ve lineer regresyon yöntemi en başarılı olarak bulunmuştur [8]. Obagbuwa ve Abidoye tarafından Kaggle isimli web sitesinden Güney Afrika'da işlenen 27 farklı kategorideki suç verisi elde edilmiştir. Doğrusal regresyon yöntemi kullanılarak suç tahminine dayalı analiz yapılmıştır ve Güney Afrika makamlarının ve güvenlik kurumlarının suç eğilimleri hakkında fikir sahibi olmaları sağlanmıştır [9]. Awal ve arkadaşları tarafından Bangladeş polisinin web sitesinden alınan veriler lineer regresyon yöntemi ile model oluşturularak eğitilmiştir. Soygun, cinayet, kadın ve çocuk şiddeti, adam kaçırma, hırsızlık ve diğer suçlarla ilgili tahminler yapılmıştır, çalışmanın sonucunda nüfus artışıyla suçların da arttığı gözlemlenmiştir. Suç eğilimlerini tahmin etme, önleme veya çözme konusunda kolluk birimlerine yardımcı olmak amaçlanmıştır [10].

Siber güvenlik, yük tahmininde araştırmaya yeni bir boyut kazandırmıştır ve Luo ve arkadaşları tarafından yapılan çalışmada geçmiş verilere kötü niyetli bir şekilde yanlışlık dayatma saldırısı ele alınmıştır. Yük tahmini yapmak için üç sağlam regresyon modeli önerilmiştir. Yapılan deneyler neticesinde karşılaştırılan modeller arasında en iyi yöntem sağlam regresyon olmuştur. Siber uzaydaki diğer veri bütünlüğü saldırı türleri altında yük tahmini için yeni teori ve metodolojilerin araştırılmasına yol açabilecektir [11]. Qian ve arkadaşları tarafından örüntü sınıflandırması için yeni bir ikili sağlam regresyon modeli önerilmiştir. LFW yüz görüntüsü, FRGC yüz görüntüsü, CUHK yüz çizimi, PolyU Palm, NUST-RF yüz görüntüsü ve Caltech 101 olmak üzere altı kamuya açık veri tabanı üzerinde kapsamlı deneyler gerçekleştirmişlerdir ve önerilen modelin son teknoloji regresyon tabanlı sınıflandırma yöntemlerine göre daha başarılı olduğu görülmüştür [12].

Kibria ve Banik tarafından çapılan çalışmada çoklu doğrusal bağlantı probleminin çözümü için beş ridge tahmincisi için kapsamlı bir araştırma yapılmıştır. Simülasyonların ve sayısal örneklerin sonuçlarına dayanılarak tahmin karşılaştırmaları yapılmıştır [13]. Pereira ve arkadaşları tarafından 2010-2012 yıllarında konaklama endüstrisine ait 401 iflas eden ve 2032 iflas etmeyen firmadan oluşan bir veri seti ile çalışma yapılmıştır. Şirket iflasını tahmin etmek için ampirik modeller geliştirmek için birçok nicel yöntem ve farklı değişken seçim teknikleri kullanılmıştır. SPSS'de uygulanan kademeli yöntemlere kıyasla, ridge ve lasso regresyon modellerinin eğitim setinde daha ağır ağırlıkla görünen bağımlı değişken kategorisini tercih etme eğiliminde olduğunu göstermiştir [14].

Wang ve arkadaşları tarafından lasso regresyon algoritmasına bağlı gemi yakıt tüketimini tahmin eden bir çerçeve önerilmiştir. Gemilerin operasyonel verileri ve hava durumu verileri kullanılmıştır. Geleneksel yöntemlerden daha iyi performans gösteren lasso regresyon aynı zamanda yorumlanabilirlik, genelleme yeteneği ve sayısal kararlılık gibi özelliklere sahiptir [15]. Reid ve arkadaşları tarafından yapılan çalışmada çeşitli varyans tahmincileri gözden geçirilerek yapılan simülasyon neticesinde geniş bir seyreklik ve sinyal gücü ayarları aralığında iyi bir performans göstermiştir [16]. Alves ve arkadaşları tarafından kentsel metrikleri kullanarak regresyon yöntemine dayalı bir suç analiz tahmini yapılmıştır [17].

Yapılan çalışmalarda özellikle suç analizinde ve diğer çalışmalarda regresyon yöntemlerinin başarısı görülmüştür. Suçu analiz ederken diğer makine öğrenmesi yöntemleri de ciddi başarı gösterdiğinden daha önce aynı verilerle yapılan çalışma ile [18] karşılaştırılmıştır. Çalışmada özellikle regresyon yöntemlerinin analizdeki başarısının önceki çalışma ile karşılaştırılması amaçlanmaktadır. Yine en iyi performans gösteren regresyon yöntemini belirlemektir.

İlk bölümde daha önce yapılan çalışmalar incelenmiştir, ikinci bölümde kullanılan yöntem detayları ve veri seti tanıtılmıştır. Üçüncü bölümde sonuçlar tartışılmış ve son bölümde çalışmanın sonuçları verilmiştir.

2. Materyal ve Yöntem

Çalışmada Elazığ ilinde işlenen siber suç dataları veri seti olarak kullanılmıştır. Veri seti elde edilirken tüm siber suç detayları incelenmiş ve içerisinden gerekli olmayan alanlar çeşitli veri bilimi yöntemiyle temizlenmiştir. Veri setinde suç, cinsiyet, yaş, gelir, meslek, medeni durum, eğitim, saldırı şekli, saldırı zararı ve saldırı yöntemi öznitelikleri kullanılmıştır. Tüm algoritmalar için verinin %80'i eğitim, %20'si test için ayrılmıştır. Regresyon yöntemlerinin kullanılmasının sebebi, yapılandırılmış ve yapılandırılmamış birçok veri desenini tanıması, suç analizinde başarılı olması, karmaşık veriler arasındaki ilişkileri ortaya çıkarmasıdır.

2.1. Lineer Regresyon (Linear Regression)

Lineer regresyon, karşılıklı bağımlılığa sahip iki rastgele değişken arasındaki doğrusal ilişkinin ölçümüdür. Regresyon analizi, bir veya daha fazla yanıt değişkeni ile tahmin ediciler arasındaki ilişkiyi keşfetme yöntemidir.

y ile gösterilen bağımlı değişkenler, açıklanan değişkenler, tahmin değerleri veya gerilemeler olarak adlandırılabilmektedir. $x_1, x_2, \ldots x_p$ ile gösterilen değişenler açıklayıcı değişken, kontrol değişkeni ya da regresör olarak adlandırılmaktadır. Basit doğrusal regresyon, iki değişken arasındaki doğrusal ilişkiyi modellemek içindir. Bunlardan biri bağımlı değişken y, diğeri ise bağımsız değişken x'tir. Örneğin, basit doğrusal regresyon, kas gücü (y) ile yağsız vücut kütlesi (x) arasındaki ilişkiyi modelleyebilmektedir. Basit regresyon modeli genellikle Denklem 1'deki biçimde yazılmaktadır [19].

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad (1)$$

Burada y bağımlı değişkendir, β_0 y kesişimidir, β_1 regresyon çizgisinin gradyanı veya eğimidir, x bağımsız değişkendir ve $\varepsilon,$ rastgele hatadır. Basit lineer regresyonda genellikle $\varepsilon,$ hatasının $E(\varepsilon,) = 0$ ve sabit bir varyans $Var(\varepsilon,) = \sigma_2$ ile normal olarak dağıldığı varsayılmaktadır. Bir bağımlı değişkeni ve birden fazla bağımsız değişkeni olan doğrusal bir regresyon modeli çoklu doğrusal regresyondur. Çoklu doğrusal regresyonda, yanıt değişkeninin model parametrelerinin doğrusal bir fonksiyonu olduğu ve modelde birden fazla bağımsız değişken olduğun varsayılmaktadır. Çoklu doğrusal regresyon modelinin genel formülü Denklem 2'deki gibidir [19].

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon, \quad (2)$$

Burada y bağımlı değişken, $\beta_0, \beta_1, \ldots \beta_p$ regresyon katsayıları ve $x_1, x_2, \ldots x_p$ modeldeki bağımsız değişkenlerdir. Klasik regresyon ayarında genellikle hata terimi $\varepsilon,$'nin $E(\varepsilon,) = 0$ ve sabit bir varyans $Var(\varepsilon,) = \sigma_2$ ile normal dağılımı takip ettiği varsayılmaktadır.

Basit doğrusal regresyon, bir bağımlı değişken ile bir bağımsız değişken arasındaki doğrusal ilişkiyi araştırırken, çoklu doğrusal regresyon, bir bağımlı değişken ile birden fazla bağımsız değişken arasındaki doğrusal ilişkiye odaklanmaktadır. Çoklu doğrusal regresyon, ortak doğrusallık, varyans artırıcı, regresyon teşhisinin grafiksel gösterimi ve regresyon aykırı değerinin ve etkili gözlemin tespiti gibi basit doğrusal regresyondan daha fazla konuyu içermektedir [19].

2.2. Polinom Regresyon (Polynomial Regression)

Polinom regresyon, yalnızca bağımsız bir X değişkeni ile çoklu regresyonun özel bir durumudur. Tek değişkenli polinom regresyon modeli aşağıdaki formüldeki gibi ifade edilmektedir. Burada k polinomun derecesidir. Polinomun derecesi modelin sırasıdır. Etkin şekilde bu, $X_1 = X$, $X_2 = X^2$, $X_3 = X^3$ değişkenleri ile çoklu bir modele sahip olmakla aynıdır ve formülü Denklem 3'te verilmiştir [20].

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 {x_i}^2 + \beta_3 {x_i}^3 \ldots + \beta_k {x_i}^k + \varepsilon_i \, , for \; i = 1,2,\ldots,n \quad (3)$$

2.3. Ridge Regresyon

Ridge regresyon, bağımlı bir değişken ile bazı açıklayıcı değerler arasındaki doğrusal bir ilişkiyi modelleyen istatistiksel bir yöntemdir. Öneri sistemleri gibi birçok öğrenme algoritmasında önemli bir rol oynayan bir yapı taşıdır. Bu kullanıcı profillerini geri bildirim yoluyla öğrenmektedir ve kullanıcının belirgin bir sorgu yapmasına gerek kalmadan ilgi alanında yer almaktadır. Regresyon tekniği, toplanan bir dizi veriyi analiz etmektedir ve öğenin kullanıcı ile ne kadar ilgili olduğunu belirlemek için bunları kompakt formda özetlemektedir. Diğer geleneksel makine öğrenimi algoritmalarına benzer şekilde ridge regresyon modeli oluştururken verilerin düz metin formunda olması gerekmektedir. Bu, çevrimiçi hizmetle meşgul olan kullanıcının regresyon için hizmet sağlayıcıyla sahip olduğu verilerini paylaşması gerektiği anlamına gelmektedir. Ancak paylaşılan veriler kişisel bilgileri içeriyorsa, kullanıcı bunu yapmayı reddedebilmektedir [21].

$1 \leq u \leq N$ için, girdi dizisi $x_u = (x_{u,1}, x_{u,2}, \ldots x_{u,d}) \in \mathbb{R}^2$ ve buna karşılık gelen $y_u \in \mathbb{R}$ çıktı olarak verilmektedir. Çoklu doğrusal regresyon problemi, $w = (w_1, w_2, \ldots w_d) \in \mathbb{R}^d$'yi $y = Xw$ olarak öğrenmektedir, burada $X = \left[x_{u,j}\right]_{Nxd}$ ve $y = [y_u]_{Nx1}$ olarak ifade edilir. $y = Xw$'yi karşılayan w parametresi mevcut olmayabilir. Bu nedenle, ridge regresyon yöntemi, Denklem 4'teki amaç fonksiyonunu $E: \mathbb{R}^2 \longrightarrow \mathbb{R}$ en aza indirerek w'nin en yakınlığını tahmin etmektedir [21].

$$E(w) := \|y - Xw\|^2 + \alpha \|w\|^2 \quad (4)$$

Pozitif α için, modellerin fazla takılmasını önlemek için $\alpha\|w\|^2$düzenlileştirme terimi kullanılır. W.r.t'nin türevi alındığında; w, formülün minimizasyonu doğrusal sistemi çözerek hesaplanmaktadır. $Aw = b$, burada $A = X^T\mathrm{X} + \alpha I$ ve $b = X^T\mathrm{y}$ olarak ifade edilmektedir. $\alpha \geq 0$ olduğunda, A matrisinin simetrik ve pozitif tanımlı olduğuna dikkat etmek gerekmektedir [21].

2.4. LASSO Regresyon

Regresyon modelleri genellikle aşırı risk öngörerek özellikle düşük performans gösterme eğilimindedir. Bu sorunu çözmek için düzenlileştirme yapan lasso regresyon uygun bir tercihtir. Öznitelik seçimini otomatik olarak yapar ve çıktı olarak da ayrık bir model vermektedir. Tahmin hatasını en aza indiren modeli oluşturan değişkenleri ve karşılık gelen regresyon katsayılarını belirlemeyi amaçlamaktadır. Lasso regresyonu bazı ayarlarda standart yöntemlerden daha iyi performans gösterdiği gözlemlenmiştir. Tek tek değişkenlerin katkısının tahmin ve yorumunun doğruluğuna değil, en iyi kombine tahmine odaklandığı için regresyon katsayılarının bağımsız risk faktörleri açısından güvenilir bir şekilde yorumlanamamasıdır [22].

2.5. Değerlendirme Metrikleri

Ortalama kare hatası (Mean Squared Error-MSE) , rastgele hata teriminin σ^2 varyansının tarafsız bir tahminidir ve Denklem 5 ile tanımlanmaktadır. Burada y_i gözlemlenen değerlerdir ve $\hat{y}_i$, i'nci durum için Y bağımlı değişkenin uygun değerleridir. Ortalama kare hatası, ortalamanın serbestlik derecesine bölünerek yapıldığı ortalama kare hatası olduğu için, MSE, regresyonun verilere ne kadar iyi uyduğunun bir ölçüsüdür. MSE'nin karekökü, rastgele hata teriminin standart sapması σ'nın bir tahmincisidir. Kök ortalama kare hatası (Root Mean Squared Error-RMSE) $RMSE = \sqrt{MSE}$, σ'nın tarafsız tahmincisi değildir, ancak yine de iyi bir tahmin edicidir. MSE ve RMSE, regresyondaki hataların boyutunun ölçüleridir ve regresyon uyumunun açıklanan bileşeni hakkında bir işaret vermemektedir [20].

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(k-1)} \tag{5}$$

Ortalama mutlak hata (Mean Absolute Error-MAE), göreceli performansı ölçtüğü için farklı öğeler veya ürünler arasındaki tahminlerin doğruluğunu karşılaştırmak için en kullanışlı ölçümdür. Nicel tahmin yöntemlerinde yaygın şekilde kullanılan bir doğruluk ölçümüdür. Denklem 6'da tanımlanmaktadır. MAE hesaplanan değeri 0,1'den az ise, mükemmel doğrulukta tahmin, 0,1 – 0,2 arasında iyi derecede tahmin, 0,2 – 0,5 arasında kabul edilebilir tahmin ve 0,5'ten fazla yanlış tahmin olarak yorumlanmaktadır [20].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{6}$$

Çoklu regresyonun R^2'si, determinasyon katsayısının Denklem 7'de tanımlandığı gibi basit regresyona benzerdir. Burada $\bar{y}$, Y değişkeninin aritmetik ortalamasıdır. R^2, açıklayıcı değişken X tarafından açıklanan Y yanıt değişkenindeki varyasyon yüzdesini ölçer. Dolayısıyla, regresyon modelinin verilere ne kadar iyi uyduğunun önemli bir ölçüsüdür. R^2'nin değeri her zaman sıfır ile bir arasındadır ($0 \leq R^2 \leq 1$). 0,9 veya üzeri bir R^2 değeri çok iyidir, 0,8'in üzerindeki bir değer iyidir ve 0,6 veya üzeri bir değer bazı uygulamalarda tatmin edici olabilmektedir, ancak bu gibi durumlarda tahmindeki hataların nispeten yüksek olabileceği gerçeğinin farkında olmak gerekmektedir. R^2 değeri 0,5 veya altında olduğunda, regresyon verilerdeki varyasyonun yalnızca %50 veya daha azını açıklar bu nedenle tahmin zayıf olmaktadır [20].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

3. Bulgular ve Tartışma

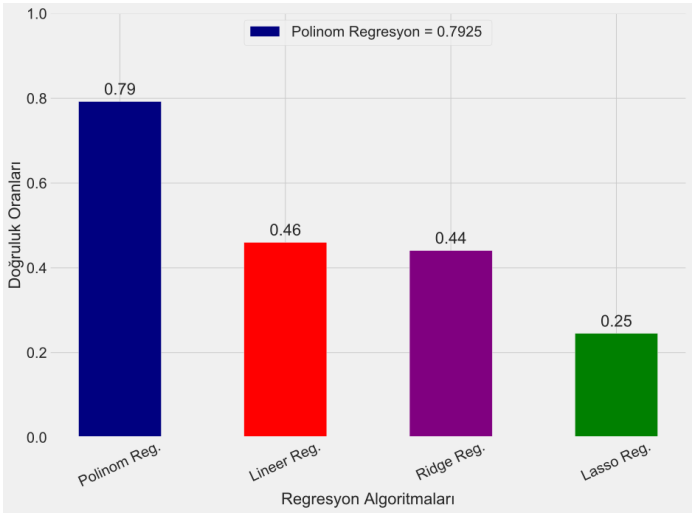
Çalışmada Pyhton 3.7 programıyla lineer regresyon, polinom regresyon, ridge regresyon, lasso regresyon yöntemlerini kullanarak dört farklı modelde saldırı yöntemi tahmin edilmiştir. Eğitim aşamasında öznitelik olarak suç, cinsiyet, yaş aralığı, gelir, meslek, medeni hal, eğitim, saldırı şekli, saldırı amacı ve fail durumu kullanılmıştır.

Model sonuçları, Ortalama mutlak hata (MAE), Ortalama kare hatası (MSE), Kök ortalama kare hatası (RMSE), R Square kullanılarak değerlendirilmiştir ve sonuçlar Tablo 1'de gösterilmiştir.

Tablo 1. Model sonuçları

	Model	MAE	MSE	RMSE	R Square	Cross Validation
1	**Lineer Regresyon**	1,2159	2,2475	1,4991	0,4602	-0,8280
2	**Polinom Regresyon**	0,6563	0,8640	0,9295	0,7925	0,0000
3	**Ridge Regresyon**	1,2622	2,3274	1,5255	0,4410	-0,8254
4	**Lasso Regresyon**	1,5699	3,1425	1,7727	0,2452	-0,3489

Regresyon modellerinde R square açısından lineer, ridge ve lasso regresyonların başarı oranı çok düşük olmakla birlikte en başarılısı polinom regresyon olmuştur. Ancak polinom regresyonunda başarı oranı da tatmin edici değildir. MAE, MSE ve RMSE değerleri açısından bakıldığında 0'a en yakın sonuçlar polinom regresyon tarafından elde edilmiştir. Diğer regresyon yöntemlerinde başarı oranlarının düşük olduğu ve hata oranının yüksek olduğu gözlemlenmiştir. R Square karşılaştırması Şekil 1'de grafiksel olarak gösterilmiştir.

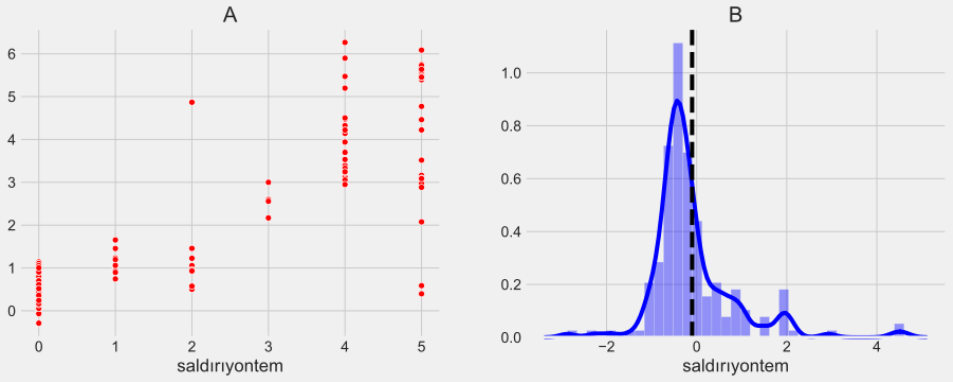


Şekil 1. Algoritmaların doğruluk karşılaştırması

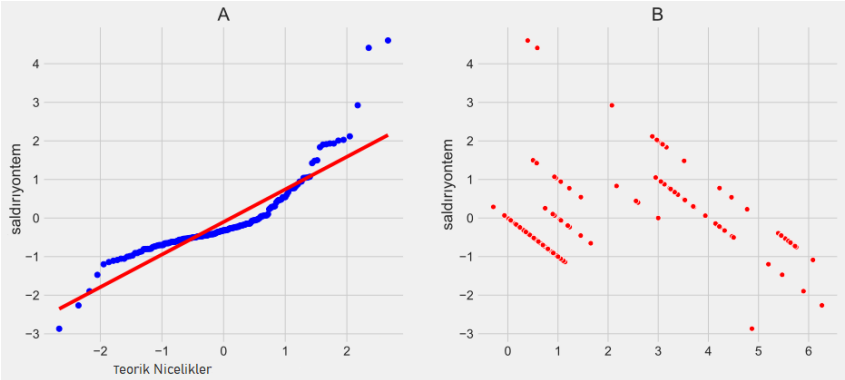
Regresyonda bağımlı ve bağımsız değişken arasındaki ilişki doğrusal olmaktadır, bu da gerçek değer ile tahmin edilen değer arasındaki farktan anlaşılmaktadır. Artık hata grafiği normal dağıtılmalı ve olabildiğince 0'a yakın olmalıdır. Tüm değişkenlerin çok değişkenli normal olması gerekmektedir ve Q-Q grafiği ile kontrol edilmektedir. Varyans enflasyon faktörü (Variance Inflation Factor-VIF) bağımsız değişkenler arasındaki kolerasyonu ve bu kolerasyonun gücünü tanımlamaktadır. Eşvaryanslılık durumunda artıklar regresyon çizgisi boyunca eşit olması gerekmektedir. Şekil 2 (A) da gösterildiği gibi polinom regresyon modeli gerçek ve tahmin verileri nispeten başarılı olduğunu yeterince tatmin edici olmadığını göstermiştir. Şekil 2 (B) de görüldüğü üzere artık grafiği sağa çarpıktır.

Şekillerde saldırı yöntemi türleri "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" şeklinde ifade edilmektedir. Şekil 3 (A) daki Q-Q grafiğinde 2'den büyük değerler artış eğilimindedir. Şekil 3 (B) de değişen varyans sergilediğinden belli noktadan sonra hata artmaktadır. VIF değeri 5'ten küçük olduğu için çoklu bağlantı zayıf olarak tespit edilmiştir.

Daha önce yapılan çalışma [18] ile karşılaştırıldığında Lineer, Polinom, Lasso ve Ridge Regresyon modellerinin daha başarısız olduğu yeni modelin yeterince başarı elde edemediği görülmüştür.



Şekil 2. A) Doğrusallık Kontrolü (Gerçek && Tahmin Değerleri) **B)** Artık Normalliği Kontrolü & Ortalama Artık Hata



Şekil 3. A) Çok Değişkenli Normallik Kontrolü (Q-Q Grafiği) **B)** Eşvaryanslık Kontrolü (Artık && Tahmin)

4. Sonuçlar

Araştırmadaki temel amaç suç istatistiğinden elde edilen siber saldırıları analiz ederek, saldırı yönteminin ne olacağını ve regresyon yöntemlerinin başarı oranını tespit etmekti. Suç analizinde lineer regresyon, polinom regresyon, ridge regresyon ve lasso regresyon olmak üzere dört farklı tahmin yöntemi kullanılmıştır. Uygulanan modelde lineer, ridge ve lasso regresyon doğruluk oranlarının çok düşük olduğundan başarısız olduğu tespit edilmiştir. Polinom regresyon yönteminde ise 0.79 R Square doğruluk oranıyla 4 yöntem arasında en başarılı yöntem olduğu görülmüş olsa da daha önce benzer yapılan çalışmaya [18] göre başarı oranı düşük kalmıştır. Yapılan çalışmada polinom regresyon modelinin geliştirilmesi halinde suç analizi ve tahminlerde kullanılabileceği değerlendirilmektedir. Gelecek çalışmalar için hibrit yöntemler geliştirilerek suç ve suçlu analizinde kullanılabilecektir.

Teşekkür

Bu çalışmadaki veriler Elazığ Valiliği ve Elazığ İl Emniyet Müdürlüğünden alınan izin neticesinde kullanılmıştır ve vermiş oldukları izin ve destekler için teşekkür ederiz.

Kaynaklar

[1] Kim, S., Joshi, P., Kalsi, P. S., & Taheri, P. Crime analysis through machine learning. In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) pp. 415-420. IEEE.

[2] Yadav, S., Timbadia, M., Yadav, A., Vishwakarma, R., & Yadav, N. (2017, April). Crime pattern detection, analysis & prediction. In 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA) (Vol. 1, pp. 225-230). IEEE.

[3] Sujatha, R., S., Ezhilmaran, A. A Comparatıve Study On Predıctıon Of Crıme Patterns, International Journal of Pharmacy and Technology 2016; 8(4):5104-5117
[4] David, H., & Suruliandi, A. (2017). Survey On Crime Analysıs And Predıctıon Usıng Data Mınıng Technıques. ICTACT journal on soft computing, 7(3).
[5] Prabakaran, S., & Mitra, S. (2018, April). Survey of analysis of crime detection techniques using data mining and machine learning. In Journal of Physics: Conference Series (Vol. 1000, No. 1, p. 012046). IOP Publishing.
[6] Ingilevich, V., & Ivanov, S. (2018). Crime rate prediction in the urban environment using social factors. Procedia Computer Science, 136, 472-478.
[7] Matlhare, B., Faimau, G., & Sechele, L. Rısk Perception And Knowledge Of Cybercrıme And Its Preventıve Strategıes Among Youth At The Unıversıty Of Botswana.
[8] Bhuriya, D., Kaushal, G., Sharma, A., & Singh, U. Stock market predication using a linear regression. In 2017 international conference of electronics, communication and aerospace technology (ICECA) 2017; Vol. 2, pp. 510-513.
[9] Obagbuwa, I. C., & Abidoye, A. P. South Africa Crime Visualization, Trends Analysis, and Prediction Using Machine Learning Linear Regression Technique. Applied Computational Intelligence and Soft Computing, 2021.
[10] Awal, M. A., Rabbi, J., Hossain, S. I., & Hashem, M. M. A. Using linear regression to forecast future trends in crime of Bangladesh. In 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV) pp. 333-338.
[11] Luo, J., Hong, T., & Fang, S. C. Robust regression models for load forecasting. IEEE Transactions on Smart Grid, 2008; 10(5), 5397-5404.
[12] Qian, J., Zhu, S., Wong, W. K., Zhang, H., Lai, Z., & Yang, J. Dual robust regression for pattern classification. Information Sciences, 2021; 546, 1014-1029.
[13] Kibria, B. M., & Banik, S. Some ridge regression estimators and their performances, 2020.
[14] Pereira, J. M., Basto, M., & da Silva, A. F. The logistic lasso and ridge regression in predicting corporate failure. Procedia Economics and Finance, 2016; 39, 634-641.
[15] Wang, S., Ji, B., Zhao, J., Liu, W., & Xu, T. Predicting ship fuel consumption based on LASSO regression. Transportation Research Part D: Transport and Environment, 2018; 65, 817-824.
[16] Reid, S., Tibshirani, R., & Friedman, J. A study of error variance estimation in lasso regression. Statistica Sinica, 2016; 35-67.
[17] Alves, L. G., Ribeiro, H. V., & Rodrigues, F. A. Crime prediction through urban metrics and statistical learning. Physica A: Statistical Mechanics and its Applications, 2018;505, 435-443.
[18] Bilen, A., & Özer, A. B. Cyber-attack method and perpetrator prediction using machine learning algorithms. PeerJ Computer Science, 2021; 7, e475.
[19] Yan, X., & Su, X. G. Linear regression analysis. Theory and Computing, 2003.
[20] Ostertagová, E. Modelling using polynomial regression. Procedia Engineering, 2012; 48, 500-506.
[21] Chen, Y. R., Rezapour, A., & Tzeng, W. G. Privacy-preserving ridge regression on distributed data. Information Sciences, 2018; 451, 34-49.
[22] Ranstam, J., & Cook, J. A. LASSO regression. Journal of British Surgery, 2018; 105(10), 1348-1348.