Pamukkale Üniversitesi
Sosyal Bilimler Enstitüsü Dergisi
Pamukkale University Journal of Social Sciences Institute
ISSN1308-2922 EISSN2147-6985

Article Info/Makale Bilgisi
√Received/Geliş: 21.02.2017 √Accepted/Kabul:20.04.2017
DOİ: 10.5505/pausbed.2017.82712

GENETİK ALGORİTMA İLE DOĞRUSAL REGRESYONDA TAHMİN AMAÇLI MODEL SEÇİMİ

Muhsin ÖZDEMİR*

Özet

Farklı sayıda değişken içeren regresyon modellerinden seçim yapmak için Genetik Algoritmalar (GA) olarak adlandırılan sezgisel yaklaşıma dayanan bir prosedür önerilmektedir. GA'nın kromozomları ikili sayısı dizi yerine, uzunluğu (p) kullanıcı tarafından belirlenen ve değişken setlerini temsil eden tamsayı dizisi olarak kodlanmıştır. GA, kromozomları sıralamak için kromozomundaki değişkenlerle elde edilen regresyon modellerinin 20 tane Bootstrap örneklemindeki RMSE (tahmin hatalarının karelerinin ortalaması) değerlerinin ortalamasından oluşan bir değerlendirme fonksiyonu kullanmaktadır. GA, farklı değişken sayılarıyla değerlendirme fonksiyonunu en aza indirgemek için çalıştırılır. GA tarafından seçilen setler nihai olarak en iyi değişken alt setini belirlemek için tek gözlemli çapraz geçerlilik yöntemi ile değerlendirilmektedir. Önerilen GA, UCI veri deposundan alınan Topluluklar ve Suç veri setine uygulanmıştır. GA, farklı sayılarda (p) değişken seçmek için kullanılmış ve 30 değişken (p = 30) içeren alt set, tek gözlemli çapraz geçerlilik kriterine göre en iyi alt set olarak bulunmuştur. Önerilen prosedür mevcut değişken seçim yöntemleri ile karşılaştırılmış ve daha iyi performans göstermiştir.

Anahtar Kelimeler: *Değişken seçimi, Genetik algoritma, Model seçimi, Regresyon modeli seçimi, Özellik seçimi.*

PREDICTIVE MODEL SELECTION IN LINEAR REGRESSION BY GENETIC ALGORITHMS

Abstract

A procedure based on a heuristic approach called Genetic Algorithms (GA) is proposed for selecting regression models constructed by different size of independent variables. Instead of binary representation, the chromosomes are encoded as user-defined size (p) of integer arrays which represent variable subsets. The GA uses an evaluation function which consists of an average fitness (residual mean square error) of the regression model (chromosome) fitted in to 20 bootstrap samples in order to rank the chromosomes. GA runs for different size of variable subset in order to minimize the fitness function. The subsets determined by GA are finally evaluated by leave-one-out-cross-validation in order to decide the best variable subset. The proposed GA is applied to Communities and Crime dataset taken from UCI dataset repository. The GA is used to select different number of variables and the variable subset containing 30 variables (p=30) is found as the best variable subset based on leave-one-out-cross-validation score. The proposed procedure was compared with available feature selection methods and showed better performance.

Keywords: variable selection, Genetic algorithms, Model selection, Regression model selection, Feature selection

*Doç.Dr., Adnan Menderes Üniversitesi, Söke İşletme Fakültesi, Yönetim Blişim Sistemleri Bölümü, AYDIN.
e-posta: mozdemir@adu. edu.tr

1.GİRİŞ

Bilgi teknolojilerindeki gelişmeler ve yenilikler veri toplamayı ve depolamayı oldukça kolaylaştırmıştır. Veri toplama ve depolama maliyetleri teknolojik gelişmelere paralel olarak da düşmektedir. Ancak elektronik olarak depolanan bu verilerin insanlar tarafından kolaylıkla anlaşılabilir bilgi haline getirilmesi bilgi teknolojilerindeki gelişmeye paralel olarak gelişememektedir. Tahmin amaçlı regresyon analizi uygulamalarında çok sayıda bağımsız değişkenlerin bulunması durumu kaçınılmaz olmaktadır. Değişken gözlem altındaki bir işlemin ya da bir şeyin bireysel olarak ölçülebilen bir özelliği olarak tanımlanmaktadır. Değişken seçimi için farklı yöntemler geliştirilmiştir. Hangi yöntemin kullanılacağına karar vermek için genellikle yöntemin tahmin gücüne bakılmaktadır. Değişken seçme yöntemlerinin amacı mümkün olan en az sayıda değişken ile veriyi daha anlaşılır hale getirerek tahmin gücü yüksek modeller elde etmektir.

Değişken seçimi birçok sınıflama ve regresyon problemlerinde yaygın olarak ortaya çıkmaktadır. Burada genellikle faydalı değişken sayısı en aza düşürülürken aynı zamanda modelin tahmin gücünün de daha yükseğe çıkartılması amaçlanmaktadır. Bu açıdan bakıldığında değişken seçimi çok amaçlı optimizasyon probleminin özel bir durumu olarak karşımıza çıkmaktadır.

Prensipte değişkenlerin hepsi (ilgisiz, faydasız ya da ayırt edici olmayan değişkenler de dahil) bağımlı değişken ile bağımsız değişkenler arasındaki fonksiyonel ilişkinin modellenmesinde kullanılabilmektedir. Ancak uygulamada modelde kullanılan ilgisiz ve faydasız değişkenler® birçok sorunlara yol açabilmektedir (Kewley vd., 1998). Bunlar;

- İlgisiz özellikler problemin boyutunu arttıracağından daha fazla gözlem değeri gerektirir.
- İlgisiz özellikler elde edilen modelin genelleme yapamamasına neden olabilir.
- İlgisiz özellikler problemin boyutunu arttıracağından daha fazla hesaplama zamanı gerektirir.
- Fazla sayıdaki değişken elde edilen modelin tahmin etmesini zorlaştırır.
- Fazla sayıdaki değişken elde edilen modelin anlaşılmasını ve yorumlanmasını zorlaştırır.

Değişkenlerin hepsini regresyon modeline dahil etmek yerine bu değişkenlerden bir kısmını kullanmanın amacı (Miller, 1984);

- regresyon katsayılarını daha düşük standart hata ile tahmin yapmak (özellikle bağımsız değişkenler arasındaki korelasyonun yüksek olması durumunda),
- toplanan verideki değişkenlerinin sayısının azaltılmasıyla daha az maliyete katlanarak tahmin yapmak,
- ayırt edici bilgi içermeyen değişkenlerin elenmesiyle daha doğru tahmin yapmak,
- çok değişkenli veri setini daha kısa ve özlü bir şekilde tanımlamak olabilir.

Yukarıdaki amaçların hepsi doğal olarak birbirleriyle tamamen uyum içerisinde olmamakla birlikte en yaygın amaç tahmin yapmaktır. Amacın tahmin olması durumunda regresyon katsayıları birinci amaç olmaktan çıkmakta ve iyi bir şekilde tahmin edilemeyen regresyon katsayıları bazen kabul edilebilir tahmin sonuçları verebilmektedir.

Bu çalışmada amaç çok sayıda değişken içeren bir veri setindeki en az sayıda bağımsız değişkeni kullanarak tahmin gücü yüksek regresyon modelleri elde etmektir. Bu amacı gerçekleştirmek için Genetik Algoritmalar (GA) olarak adlandırılan sezgisel yaklaşıma dayanan bir değişken seçim prosedürü önerilmektedir.

2.MODEL SEÇİMİ

Bir değişkenler ya da özellikler kümesinden bir alt set (küme) seçme problemi doğrusal modellerde yoğun bir şekilde dikkate alınmıştır. Değişken seçimi model seçimi probleminin özel bir durumu olarak da düşünülebilmektedir. Diğer bir ifade ile burada her bir model farklı değişken setine karşılık gelmektedir.

Bağımlı bir değişkene (Y) karşı k tane bağımsız değişken (X) ve bunlar arasındaki doğrusal ilişki Y=Xβ+ε denklemi ile ifade edilebilirse bu modele çoklu doğrusal regresyon modeli denir. Burada Y (nx1) boyutunda bağımlı değişkene ait n adet gözlem değeri içeren bir vektör; X (nxk) boyutunda ve her sırası k adet bağımsız değişkene ait n adet gözlem değeri içeren bir matris; β (kx1) boyutunda bilinmeyen ancak tahmin edilecek parametre (kısmi regresyon katsayıları) vektörü ve ε ise ortalaması (beklenen değeri) sıfır olan tesadüfi (rassal) bir vektördür. Eğer regresyon modelinde bütün kısmi regresyon katsayılarının sıfır olması durumda sabit sayıya uydurulur (sabit model) ise bu durumda X matrisine bütün değerleri 1 olan bir sütun ilave edilir. Kısmi regresyon katsayılarının (β) en küçük kareler tahminleri $X'X\beta = X'Y$ normal denklemlerinin çözümü ile bulunur. β katsayılarının tahmin edicilerini b ile gösterirsek $b = (X'X)^{-1}X'Y$ olur. Bu şekilde hesaplan b değerleri karşılık gelen β regresyon parametrelerini tahmin edebilen doğrusal tahmin ediciler arasında yansız ve varyansı en küçük olan tahmin edicilerdir (Thompson, 1978).

2.1.Değişken Seçim Kriterleri

Bir değişkenler setini içeren modeli seçmek için istatistik biliminin doğrusal modeller alanında bazı kriterler (ölçüler) önerilmiştir. Bu kriterlerden en çok bilineni tahmin hatalarının karelerinin ortalamasının karekköüdür (Root Mean Square Error). Tahmin hatalarının karelerinin ortalamasının karekökü RMSE olarak gösterilecektir. RMSE, toplam k adet değişkenden ve n adet gözlem değerinden oluşan veri setinden p tane değişken içeren bir model için $RMSE_p = \sqrt{\frac{SSE_p}{n}}$ şeklinde hesaplanır. Burada SSE_p p tane değişken içeren modelin tahmin hatalarının (farklarının) karelerinin toplamını (residual sum of square errors), n ise gözlem sayısını gösterir. $SSE_p = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ şeklinde hesaplanır. Burada y_i bağımlı değişkenin i'inci gerçek değerini, $\hat{y}_i$ ise regresyon modelinin tahmin ettiği bağımlı değişkenin i'inci tahmini değerini göstermektedir. İki regresyon modelinin karşılaştırılması durumunda RMSE değeri küçük olan model seçilir. Genel olarak modele ilave edilen değişken sayısı arttıkça RMSE değerinin düşeceği de unutulmamalıdır.

Çoklu Belirlilik Katsayısı (R^2) bağımsız değişkenlerin, bağımlı bir değişkenin davranışını açıklamadaki başarı ölçüsünü veren betimleyici bir istatistiktir ve p adet bağımsız değişken içeren bir model için $R_p^2 = 1 - \frac{SSE_p}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$ şeklinde hesaplanır. R^2 değeri büyük olan model tercih edilir. Bir regresyon modeline ilave edilen bağımsız bir değişken, bağımlı değişkenle ilişkili olmasa dahi R^2 değerinin artmasına yol açtığından farklı değişken sayılarına sahip modellerin karşılaştırılmasında kullanılması uygun olmamaktadır. Bağımsız değişkenlerin modele etkisinin olması gerekenden fazla olmasını önlemek için Düzeltilmiş Belirlilik Katsayısı (adj.R^2) geliştirilmiştir.

Değişken seçimi regresyon bağlamında model seçim problemi olarak düşünülebilmektedir. Bilgi kriterleri ilk olarak genel model seçimi problemlerinde ortaya çıkmıştır. Akaike Bilgi Kriteri (AIC) model seçiminde modelin doğru bir şekilde uyması ve modelin basit olması (modeldeki değişken sayısının azlığı) talepleri arasındaki itilafı dengelemeye çalışmaktadır (Chatterjee ve Hadi, 2006). Akaike Bilgi Kriteri istatistiği $AIC_p = n \ln\left(\frac{SSE_p}{n}\right) + 2p$ olarak hesaplanır ve AIC_p değeri küçük olan model tercih edilir.

Akaike Bilgi Kriterine birkaç küçük değişiklikler yapılarak yeni kriterler önerilmiştir. Bu kriterlerden en popüler olanı Bayes Bilgi Kriteri (BIC) dir (Chatterjee ve Hadi, 2006). Bayes Bilgi Kriteri istatistiği $BIC_p = n \ln\left(\frac{SSE_p}{n}\right) + p(\ln n)$ olarak hesaplanır ve BIC_p değeri küçük olan model

tercih edilir. Bayes Bilgi Kriteri modeldeki değişken sayısını daha şiddetli cezalandırmakta ve AIC kriterine göre daha az sayıda değişken içeren modelleri seçmektedir.

Modelin veri değerlerine ezbere oldukça uygun bir şekilde uyması durumunu önlemek için Akaike Bilgi Kriterine değişiklik yapılarak Düzeltilmiş Akaike Bilgi Kriteri (AICC) önerilmiştir (Chatterjee ve Hadi, 2006). Düzeltilmiş Akaike Bilgi Kriteri istatistiği $AICC_p = AIC_p + \frac{2(p+2)(p+3)}{n-p-3}$ olarak hesaplanır ve $AICC_i$ değeri küçük olan model tercih edilir. Bu kriterdeki düzeltme veri setindeki veri sayısı (n) küçük ve değişken sayısı (p) büyük olduğunda büyük olmaktadır.

Doğrusal regresyon modellerinde değişken seçimi için sık kullanılan diğer kriter ise Mallow Cp istatistiğidir (Mallows, 1973). Bir değişken setinin oluşturduğu regresyon modelinden elde edilen tahmin değerleri yanlı ya da taraflı olmaktadır. Bu durum tahminin öngörü değerlerinin varyansından ve modele dahil edilmeyen değişkenlerin yokluğundan kaynaklanmaktadır. Mallow Cp istatistiği $C_p = \frac{SSE_p}{\hat{\sigma}^2} + 2p - n$ olarak ifade edilir. Burada $\hat{\sigma}^2$ veri setindeki k adet değişkenin tümünün oluşturduğu regresyon modelinden hesaplanan rassal hataların varyansının tahminini gösterir ve $\hat{\sigma}^2 = \frac{SSE_k}{n-k}$ şeklinde hesaplanır. Regresyon modelinde herhangi bir yanlılık bulunmaması durumda Cp istatistiğinin beklenen değeri p olacaktır.

2.2.Değişken Seçim Yöntemleri

Regresyonda değişken seçimi için yaygın olarak kullanılan üç yöntem (ileriye doğru seçim, geriye doğru eleme ve adımsal regresyon) ortaklaşa adımsal yöntemler olarak adlandırılmaktadır (Ruengvirayudh, Brooks, 2016). Bu yöntemler bir regresyon modeline seçilen kritere göre ya bir değişken ilave ederek ya da bir değişken çıkararak adım adım işlemden geçirmektedirler. Adımsal yöntemler değişken seçimi yaparken daha az sayıda regresyon modelleri araştırmakla birlikte nadiren en iyi modeli seçebilmektedir. Bu yöntemlere alternatif olası bütün regresyon modellerini bir kritere (R^2, Cp, AICC gibi) göre değerlendirerek en iyisini seçmektir.

Toplam k adet değişkenden oluşan bir veri setinde elde edilebilecek olası regresyon modeli sayısı 2^k adettir. Regresyon modellerini hesaplamak için yapılması gereken matematiksel hesaplama sayısı değişken sayısına bağlı olarak üstel artacağından gerekli hesaplamalar için hızlı ve verimli çalışan bir algoritmaya sahip olmak oldukça önemlidir. Ancak değişken sayısı k çok büyük olduğunda olası bütün regresyon modellerini değerlendirmek bilgisayar hesaplama zamanı açısından katlanılabilir olamamaktadır.

SPSS programı $k \leq 20$ olması durumunda bütün regresyon modellerini ele alınan bir kritere (R^2, ASE ya da AICC) göre değerlendirerek en iyi modeli seçebilmektedir.

Leaps programı etkin bir dal ve sınır algoritması ile olası bütün alt setleri araştırarak doğrusal regresyonda bağımlı değişkeni en iyi tahmin eden bağımsız değişkenlerin alt setini (kümesini) ele alınan bir kritere göre (R^2, Cp gibi) her bir model büyüklüğünde (değişken sayısında) en iyi modeli bulmaktadır. Leaps programı $k \leq 31$ olması durumunda regresyon problemlerinde olası bütün regresyon modellerini belli bir kritere göre değerlendirerek en iyi regresyon modelini seçmektedir (Lumley, 2009). Dolayasıyla değişken sayısı artması durumunda Genetik algoritma gibi sezgisel yöntemlerin kullanılması kaçınılmaz olmaktadır.

2.3.Evrimsel Algoritmalar

Evrimsel algoritmalar, doğadaki evrim prensiplerinden esinlenilerek geliştirilmiş stokastik optimizasyon yöntemleri olup kullandıkları arama yöntemleri en iyinin hayatta kalması ve genetik kalıtım gibi doğal olguları taklit ederek modellerler (Fogel, 1997). Evrimsel algoritmalar, klasik yöntemlerin uygulanmasının olası olmadığı (türevin olmaması gibi) ya da tatmin edici sonuç elde edilmesinin zor olduğu (yerel optimaların bulunması gibi) problemlere uygulanabilmektedir. Evrimsel algoritmalar çözülmek istenen probleme ait potansiyel çözümlerin oluşturduğu bir topluluk ile çalışmaktadırlar. Topluluğu oluşturan her bir potansiyel çözüm birey olarak adlandırılmaktadır. Her birey üzerinde çalışılan problemin yapısına uygun bir şekilde kodlanmış potansiyel bir çözümdür ve kromozom olarak adlandırılmaktadır. Kodlanan

her bir bireye amaç fonksiyonu tarafından uygunluk ölçüsü verilmektedir. Bundan dolayı amaç fonksiyonu bireylerin içinde yaşadığı çevre olarak düşünülebilir. Uygunluk ölçüsü yüksek olan bireylerin seçilmesi eğilimi, amaç fonksiyonunun kullanılmasını gerektirmektedir. Kodlanmış bireyler genetik işlemlerin uygulanmasından dolayı değişime uğrayarak ve evrimleşerek yeni bir topluluk oluşturmaktadırlar. Genetik işlemler kromozomların yapısını değiştirerek problemin çözümüne ilişkin olası bütün çözümlerin oluşturduğu kümenin araştırılması sağlamakta ve iyi bir çözümün bulunmasına yol açmaktadırlar.

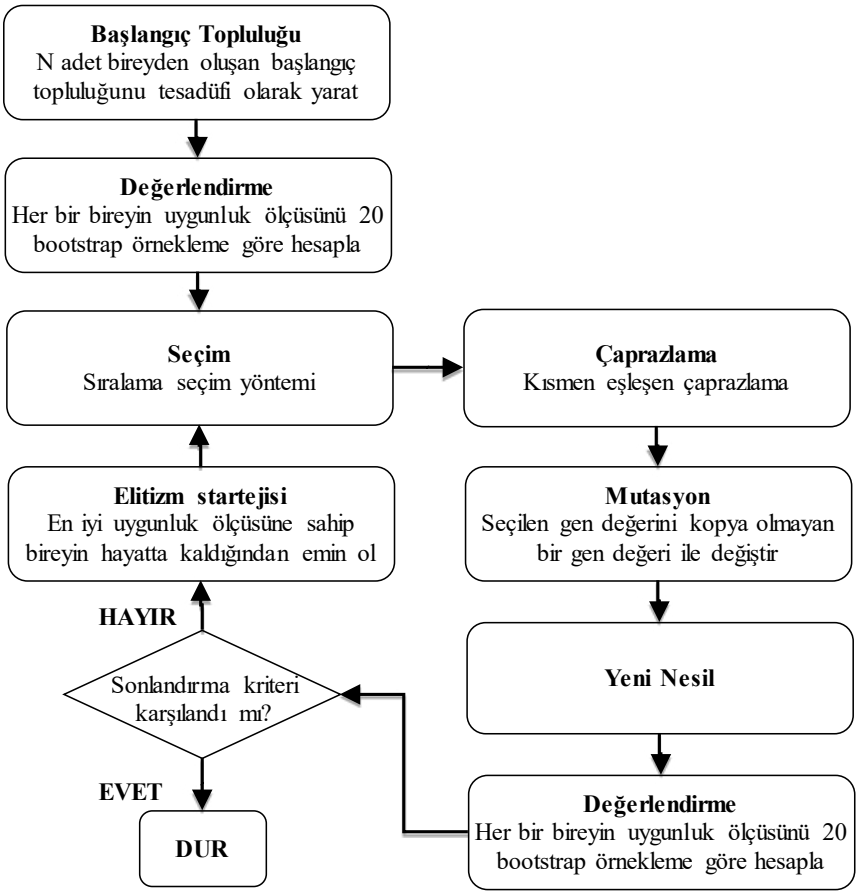
Genel olarak, mutasyon ve çaprazlama olmak üzere iki tip genetik işlem bulunmaktadır. Mutasyon tipi işlemler tek taraflı (eşeysiz) işlemler olup ele aldığı tek bir bireyde küçük bir değişimin yapılmasıyla yeni birey ya da bireyleri yaratmaktadır. Diğer yanda, çaprazlama tipi işlemler çok taraflı (çok eşeyli) işlemler olup yeni bireyleri iki ya da daha fazla sayıda bireylerden alınan parçaların birleştirilmesi ile yaratmaktadırlar (Field, 1995). Belli sayıda kuşak (nesil) evrimleşmesinden sonra süreç bir bitirme (sonlandırma) ölçüsüne dayanılarak sonlandırılmaktadır. Nihai olarak elde edilen kuşakta bulunan uygunluk ölçüsü en yüksek olan birey problem için bir çözüm olarak önerilmektedir. Önerilen bu çözümün optimal ya da optimal çözüme yakın bir çözüm olduğu ümit edilmektedir. Evrimsel algoritmalar Genetik algoritmalar (Genetic Algorithms), Evrimsel programlama (Evolutionary Programming), Evrim stratejileri (Evolutionary Strategies) ve Genetik programlama (Genetic Programming) olmak üzere dört alt sınıfa ayrılmaktadırlar. Her ne kadar bu evrimsel algoritmalar arasında birçok yakın benzerlikler bulunsa da aralarında son derece belirgin farklılıklar da bulunmaktadır (Michalewicz, 1996). Bu farklılıklar genellikle modellenen evrimsel olgu hiyerarşindeki düzeyle (kromozom, birey ya da tür gibi) ilgilidir. Evrimsel algoritmalardan iki ya da daha fazlasının kendilerine has özelliklerinin bir arada kullanılması ile birçok hibrit yöntemler de geliştirilmiştir.

Genetik algoritmalar, genetik ve biyolojik evrim teorisinden esinlenen stokastik optimizasyon algoritmalardan biridir (Van Rooji, Jain ve Johnson, 1996). Genetik algoritmaların arkasındaki düşünce, doğaya özgü evrimin bir simülasyonu yapılarak eldeki amaç fonksiyonunun optimize edilmesidir. Genetik algoritmalar, pratik ve güçlü bir optimizasyon ve arama yöntemi olarak ortaya çıkmıştır (Michalewicz, 1996). Literatüre bakıldığında Holland'ın genetik algoritması Basit Genetik Algoritma olarak adlandırılmakta ve ikili sayı dizi (binary string) şeklinde kodlanmış bireyler (kromozomlar) topluluğu ile çalışmaktadır (Vose, 1999).

Leardi, Boggia ve Terrile (1992) ikili sayısı dizi olarak kodlanmış GA'yı değişken seçimi problemlerine uygulamışlar ve klasik yöntemlere kıyasla daha iyi sonuçlar elde etmişlerdir. Paterlini ve Minerva (2010) regresyon problemlerinde değişken seçimi için iki tane genetik algoritma önermiştir. Önerilen birinci GA sadece regresyon modeli için değişken seçimi yaparken ikinci GA hem değişken seçimi hem de değişkenler için uygun matematiksel transformasyon seçimi yapmaktadır. Her iki yöntem adımsal değişken seçim yöntemleri ile karşılaştırılmış ve önerilen yöntemlerin daha iyi sonuç verdiği rapor edilmiştir. Kabir, Shahjahan ve Murase (2011) değişken seçimi problemi için bir yerel arama yöntemini ikili sayı dizi olarak kodlanmış genetik algoritma içine entegre ederek hibrit bir yöntem geliştirmişlerdir. Yerel arama yöntemi değişkenler arasındaki korelasyon bilgisine dayanarak yeni nesildeki değişkenler arasındaki korelasyonu azaltmayı amaçlamaktadır. Tsai, Eberle ve Chu (2013) veri madenciliğinde veri setinin analiz öncesi ön işleme aşamasının iki önemli konusu olan özellik seçimi ve gözlem seçimi için sınıflandırma problemlerinde ikili sayı dizisi olarak kodlanmış genetik algoritma kullanmışlardır. GA Waikato Environment for Knowledge Analysis (WEKA) programı ortamında çalışmakta ve uygunluk fonksiyonu olarak Bayes ağları öğrenme algoritmasını kullanmaktadır. Jung, M. ve Zscheischler (2013) değişken seçimi problemi için hibrit bir genetik algoritma önermişlerdir. Genetik algoritmanın uygunluk fonksiyonu değerlendirme sayısı hibrit bir yöntem ile azaltılmakta ve daha hızlı ve güvenilir sonuçlar elde edildiği rapor edilmektedir.

3.ÖNERİLEN GENETİK ALGORİTMA

Bu makalede önerilen Genetik Algoritmanın genel olarak çalışması şematik olarak Şekil 1'de gösterilmiştir.



Şekil 1. Önerilen Genetik Algoritmanın Akış Diyagramı

3.1.Bireylerin Kodlanması

Genetik algoritmalar, genel bir optimizasyon yöntemi olmakla birlikte bireylerin çözülmek istenen probleme uygun şekilde kodlanması gerekmektedir. Bireylerin kodlanması problemin olası çözüm kümesini temsil edebilecek ve genetik işlemlerin uygulanabilmesine olanak verecek şekilde yapılmalıdır. Değişken seçimi problemlerinde amaç belli bir kritere (R^2, Cp gibi) göre optimal değişken setinin bulunmasıdır. Geleneksel ve basit kodlama bireylerin sabit uzunlukta ikili sayı dizi (binary string) olarak kodlanmasıdır. Burada kromozomların uzunluğu regresyon modeline dahil edilebilecek tüm bağımsız değişkenlerin sayısıdır. Modele dahil edilebilecek bağımsız değişken sayısını k ile gösterelim. Elimizdeki veri setinde 10 adet bağımsız değişken olduğunu varsayarsak k=10 olacaktır. Burada bir kromozom k adet üniteden oluşan bir dizidir. Kromozomdaki her bir ünite "gen" olarak adlandırılmakta ve her bir ünite 0 ya da 1 değerini alabilmektedir. Kromozomdaki 0 sayısı karşılık gelen bağımsız değişkenin regresyon modelinde yer almadığını, 1 sayısı ise yer aldığını göstermektedir. Kromozomların ikili sayı dizisi kodlaması Şekil 2'de gösterilmiştir.

Değişkenler	1	2	3	4	5	6	7	8	9	10
Kromozom	1	1	0	1	1	1	0	1	0	1

Şekil 2. Kromozomların İkili Sayı Dizisi Kodlaması (Değişkenler 1, 2, 4, 5, 6, 8 ve 10 regresyon modelinde yer alacaktır.)

İkili sayı dizisi kodlaması değişken seçimi probleminin olası çözüm kümesini temsil etmekte yetersiz kalması ve genetik işlemlerin uygulanmasındaki güçlükler nedeniyle bazı olumsuzlukları bulunmaktadır (Özdemir, 2011). Bu olumsuzluklar ikili sayı dizisi kodlaması kullanan genetik algoritmaların prematüre yani olgunlaşmamış çözüm üretmesine neden olmaktadır. Genetik algoritmaların çeşitliliği az olan nesiller üretmesi, problemin çözüm kümesinin hızlı ve iyi bir şekilde araştırılmasını engellemekte ve kalitesi düşük çözümler bulunmasına neden olmaktadır. Bu olumsuzlukları ortadan kaldırmak için önerilen genetik algoritmada ikili sayı dizisi kodlaması yerine Tam Sayı Dizisi Kodlaması kullanılmıştır. Bu kodlamada, bireyler 1'den k'ya kadar olan tamsayıların permütasyonu olan p adet tam sayıdan oluşan bir tam sayı dizi olarak kodlanmıştır. Burada k veri setinde bulunan toplam bağımsız değişken sayısını ifade eder. Regresyon modeline dahil olacak bağımsız değişken sayısı p ise önceden belirlenmiştir. Kromozomlar p adet tam sayıdan oluşmaktadır. Kromozomların tam sayı dizi kodlaması Şekil 3'de gösterilmiştir.

2	8	5	1	10	6	4	7	3	9

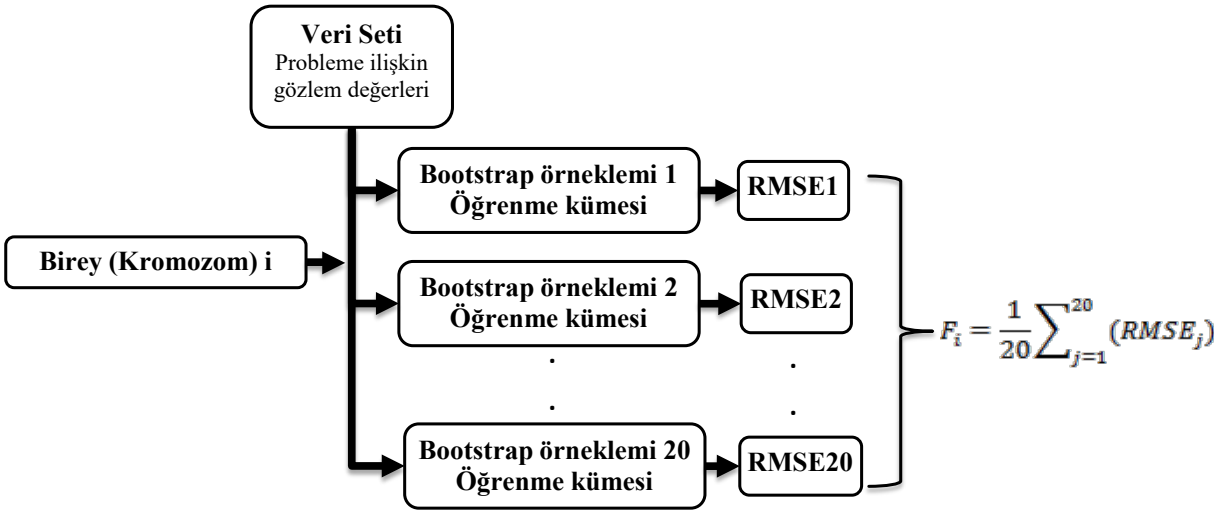
Kromozom

Şekil 3. Kromozomların Tam Sayı Dizisi Kodlaması (k=10, p=7) (Değişkenler 1, 2, 4, 5, 6, 8 ve 10 regresyon modelinde yer alacaktır.)

3.2.Değerlendirme (Amaç Fonksiyonu)

Kromozomlardaki tam sayılara karşılık gelen bağımsız değişkenler regresyon modeline dahil edilmektedir. Veri setindeki gözlem değerlerinin %90'nı yerine koymadan tesadüfi olarak seçilerek bir küme oluşturulur. Elde edilen bu örneklem yerine koymadan Bootstrap örneklemi olarak adlandırılmaktadır. Diğer bir ifadeyle, veri setindekindeki gözlem değerleri tesadüfi olarak ayrık ve bütünü kapsayıcı iki kümeye ayrılır. Birinci küme öğrenme kümesi olarak adlandırılmakta ve gözlem değerlerinin %90'nından, ikinci küme ise analiz dışı küme olarak adlandırılmakta ve gözlem değerlerinin %10'nundan oluşmaktadır. Regresyon modelinin elde edilmesinde birinci küme yani öğrenme kümesi kullanılmaktadır. İkinci kümedeki gözlem değerleri ne regresyon modelini oluşturmada ne de regresyon modelinin test edilmesinde kullanılmaktadır.

Bir bireyin uygunluk derecesi kromozomundaki değişkenlerle 20 tane Bootstrap örnekleminin öğrenme kümesindeki gözlem değerlerinden elde edilen regresyon modellerinin RMSE değerlerinin ortalaması alınarak hesaplanmaktadır. Birey i'nin uygunluk (amaç fonksiyonu) değerini F_i ile gösterirsek $F_i = \frac{1}{20}\sum_{j=1}^{20}(RMSE_j)$ olarak hesaplanır. Burada $RMSE_j$ birey i'nin j'inci Bootstrap örneklemindeki tahmin hatalarının karelerinin ortalamasının karekökü'dür. Amaç fonksiyonunun hesaplanması Şekil 4'de gösterilmiştir. Genetik algoritma yukarıda ifade edilen fonksiyonu minimum yapan bireyleri yani regresyon modellerini (değişken setlerini) bulmayı amaçlamaktadır.



Şekil 4. Birey i'nin uygunluk (amaç fonksiyonu) değeri F_i'nin hesaplanması

Amaç fonksiyonunda yukarıda açıklanan bootstrap örnekleminin kullanılmasındaki amaç; değişken setleriyle elde edilen regresyon modellerinin veri değerlerine ezbere oldukça uygun bir şekilde uymasını önlemek ve modelin geleceğe yönelik tahmin gücünü arttırmaktır. Amaç fonksiyonundaki bootstrap örneklem sayısının artması bu amacı gerçekleştirecek iyi değişkenlerin seçimine yardımcı olmakla birlikte algoritmanın çalışma süresinin artmasına yol açacaktır. Yapılan simülasyon çalışmaları 20 bootstrap örnekleminin hem değişken seçimi hem de bilgisayar hesaplama süreleri açısında uygun olduğu göstermiştir.

3.3.Genetik İşlemler

Genetik algoritmalar, problemin kombinatoriyel çözüm kümesini hızlı ve etkin bir şekilde araştırarak kaliteli bir çözüm bulabilmesi için ilgili problemin kodlamasına uygun genetik işlemlere ihtiyaç duyarlar. Bu amacı gerçekleştirmek için Tek Genli Mutasyon İşlemi ve Kısmen Eşlenen Çaprazlama İşlemi kullanılmıştır.

3.3.1.Tek Genli Mutasyon İşlemi

Mutasyon işlemi, seçilen birey üzerindeki bir geni tesadüfî olarak seçmekte ve bu genin değerini 1 ile k arasında tesadüfî olarak seçilen bir tam sayı ile değiştirmektedir. Burada k veri setinde bulunan toplam bağımsız değişken sayısıdır. Eğer mutasyona uğrayan genin değeri kromozomun diğer genleri ile aynı değere sahip değilse mutasyon işlemi gerçekleşir. Eğer mutasyona uğrayan genin kromozomun genlerinden farklı değer değil ise mutasyon işlemine belli sayıda tekrar (5 defa) edilir. Bu tekrarlarda da kromozomda bulunmayan farklı bir gen değeri bulunamaz ise kromozom mutasyona uğramadan yeni nesil topluluğuna dahil edilir.

3.3.2.Kısmen Eşlenen Çaprazlama İşlemi

Kısmen eşlenen çaprazlama işlemi, Gezgin Satıcı Problemi için geliştirilmiş bir genetik işlemdir (Goldberg ve Lingle, 1985). Bu genetik işlem tek noktalı çaprazlama işleminin özel bir şeklidir. Burada çaprazlama işlemine tabi tutulacak kromozomlar üzerinde iki nokta seçilir. Bu iki nokta arasındaki genler ebeveyn kromozomlar arasında değiştirilir. Bu değişim yapılırken bir kromozomda birbirinin aynı genler (kopya ya da klon genler) olursa özel bir düzeltme mekanizması ile bu genler düzeltilir.

3.4.Sıralama Seçim Yöntemi

Parametrik olmayan bir seçim yöntemi olan sıralama seçim yöntemi ilk olarak Baker (1985) tarafından uygulanmıştır. Bu yöntemde, topluluğu oluşturan bireyler amaç fonksiyonu tarafından verilen uygunluk ölçüsü değerine göre sıralanmakta ve bir sonraki nesilde bireylerin yer alma olasılıkları bireylerin bu sıralamadaki pozisyonlarına göre hesaplanmaktadır. Sıralama seçim yöntemi bireylere göreceli olarak hayatta kalma olasılıkları atamaktadır. Örneğin, sıralamada arka arkaya gelen bireylerin uygunluk ölçüleri arasındaki fark oldukça fazla olsa bile hayatta kalma olasılıkları bir birine yakın olabilmektedir. Bu yöntemin, bireylerin uygunluk ölçüsünü baz alan Rulet Tekerleği seçim yöntemine göre genetik algoritmaların daha iyi çözümler bulmasını sağladığı ifade edilmektedir (Whitley, 1989).

3.5.Elitizm Stratejisi

Elitizm stratejisi ilk olarak De Jong (1975) tarafından uygulanmıştır. Bu strateji mevcut topluluktaki uygunluk ölçüsü en iyi olan belli sayıdaki bireyi bir sonraki nesil topluluğuna dahil etmektedir. Bu strateji genetik algoritmanın problem için bulduğu en iyi kromozomların (çözümlerin) gelecek nesil topluluklarına aktarılmasını sağlamayı amaçlamaktadır.

3.6.Sonlandırma Kriteri

Genetik algoritma uygulamalarında karar verilmesi gereken önemli konulardan biri de evrim sürecinin ne zaman sonlandırılacağıdır. Önerilen genetik algoritma sonlandırma kriteri olarak kullanıcının belirlediği maksimum nesil sayısını kullanmaktadır. Toplam nesil sayısı belirlenen bu sayıya ulaştığında genetik algoritma durur ve mevcut nesil topluluğundaki uygunluk ölçüsü en iyi olan kromozomu problemin çözümü olarak önerir. Belirlenen maksimum nesil sayının az olması genetik algoritmanın iyi sonuçlar bulmasını engelleyecektir. Fazla olması ise gereksiz yere genetik algoritmanın çözüm bulma süresini uzatacaktır. Maksimum nesil sayısı, genetik algoritmanın değişik tesadüfî bilgisayar çekirdek sayısı kullanılarak deneme/yanılma süreci ile belirlenmektedir.

4.**UYGULAMA**

4.1.Veri Seti

Toplumlar ve Suç veri seti (Communities and Crime dataset) UCI yapay öğrenme veri bankasından alınmıştır. Veri seti Amerika Birleşik Devletleri'nde 1990 yılındaki nüfus sayımından, 1990 yılında kolluk güçlerinin yaptığı anket çalışmasından ve 1995 yılında ABD Adalet Bakanlığı tarafından yapılan araştırmadan elde edilen verilerin birleştirilmesi ile oluşturulmuştur. Veri seti 1994 yerleşim yerindeki yaşayan topluma (satır) ait suçla ilişkili olabilecek 127 bağımsız değişken (sütun) verisi ihtiva etmektedir. Burada amaç toplumdaki suç oranını tahmin etmek olarak ifade edilmiştir. Bağımsız değişkenlerden ayırt edici olmayan 5 adet değişken ve eksik veri içeren 23 adet değişken analizlere dahil edilmemiştir. Analizlerde geriye kalan 99 adet nicel bağımsız değişken kullanılmıştır. Önerilen genetik algoritma nicel değişken içeren regresyon problemlerine uygulanmaktadır.

Bu çalışmanın elde edilen sonuçlarıyla başka araştırmacıların kolay bir şekilde karşılaştırma yapabilmeleri için veri setindeki hiçbir değişken herhangi bir transformasyona ya da normalizasyona tabi tutulmamıştır.

4.2.**GA tarafından En İyi Değişken Sayısının Bulunması**

Önerilen genetik algoritmada seçilecek değişken sayısı araştırmacı (kullanıcı) tarafından önceden belirlenmesi gereken bir parametredir. Genetik algoritma 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90 ve 95 değerleri değişken sayısı olarak belirlenerek ayrı ayrı çalıştırılmaktadır. Genetik algoritmanın diğer parametreleri aşağıda verilmiştir. Bu

parametreler farklı bilgisayar çekirdek sayıları ile simülasyon yapılarak belirlenmiş ve çalışmada yapılan analizler boyunca sabit tutulmuştur.

Başlangıç topluluğu büyüklüğü = 100

Maksimum nesil sayısı = 1000

Çaprazlama işlemi olasılığı = 0,90

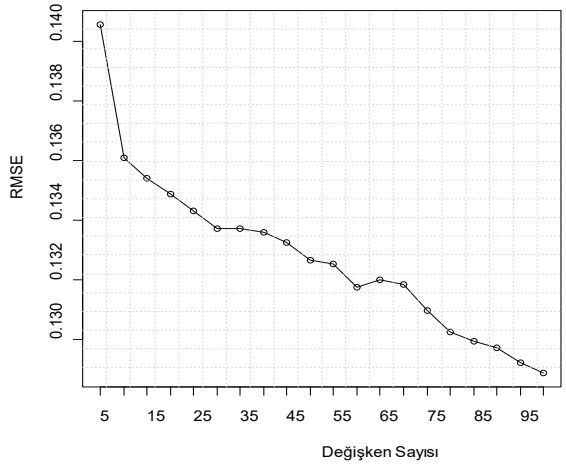
Mutasyon işlemi olasılığı = 0,02

Genetik algoritmanın bulduğu en iyi değişken setleri ve bütün gözlem değerleri kullanılarak elde edilen regresyon modellerinin RMSE değerleri Tablo 1'de verilmiştir.

Tablo 1. Genetik algoritmanın bulduğu değişken setlerinin bütün gözlem değerleri kullanılarak elde edilen regresyon modellerinin RMSE değerleri

Değişken sayısı	RMSE	Değişken sayısı	RMSE
5	0,140550	55	0,132530
10	0,136100	60	0,131765
15	0,135410	65	0,132013
20	0,134887	70	0,131858
25	0,134318	75	0,130976
30	0,133729	80	0,130278
35	0,133714	85	0,129937
40	0,133608	90	0,129737
45	0,133253	95	0,129225
50	0,132675	99	0,128902

Her ne kadar iki regresyon modelinin karşılaştırılması durumunda RMSE değeri küçük olan modelin seçilmesi söz konusu olsa da genel olarak modele ilave edilen değişken sayı arttıkça RMSE değeri de düşmektedir. Bu durum Tablo 1'deki değerlerin Şekil 5'deki grafiksel gösterimine bakıldığında kolay bir şekilde fark edilmektedir.



Şekil 5. Genetik algoritmanın bulduğu değişken setlerinin bütün gözlem değerleri kullanılarak elde edilen regresyon modellerinin RMSE değerleri

Dolasıyla regresyon modellerinin RMSE değerlerine bakıldığında veri setindeki mevcut bütün (99 adet) bağımsız değişkenlerin regresyon modeline dahil edilmesiyle en düşük RMSE değeri elde edilmektedir. Eğer amaç bağımlı değişken ile bağımsız değişkenler arasında en iyi fonksiyonel ilişkiyi bulmaksa mevcut bütün bağımsız değişkenli regresyon modeli en iyi modeldir. Hatta Yapay Sinir Ağları (Neural Networks) gibi modellerle öğrenme setinin yani modelin elde edildiği verinin RMSE değeri sıfıra oldukça yakın modeller oluşturabilir. Ancak amaç geleceğe yönelik tahmin olduğunda bu tür modellerin başarısı düşük olmaktadır. Değişken seçiminde amaç hem veri setini en iyi şekilde modelleyecek hem de geleceğe yönelik iyi tahminde bulunabilecek modelleri oluşturan değişkenleri bulmaktır. Tahmin amaçlı regresyon analizinde kısmi regresyon katsayıları öncelikli amaç olmamakta ve kötü bir şekilde tahmin edilen regresyon katsayıları bazen kabul edilebilir tahmin sonuçları verebilmektedir. Diğer bir ifade ile regresyon modelinin tahmin gücü, modelin oluşturulmasında yer almayan gözlem değerleri üzerinden yapılması daha doğru bir yaklaşım olacaktır.

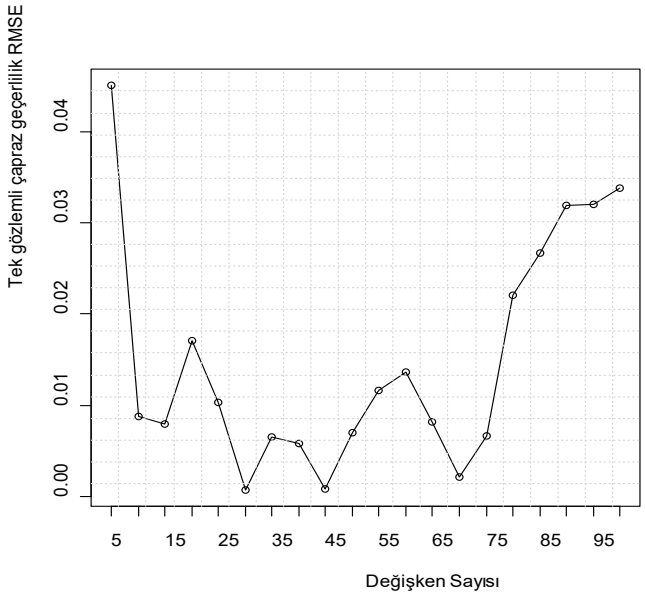
Bu amaç için çapraz geçerlilik yöntemi (crossvalidation) kullanılmıştır. Bu yöntemde veri setindeki gözlem değerleri aynı sayıda gözlem değeri içeren ayrık ve bütünü kapsayıcı alt setlere ayrılmaktadır. Çapraz geçerlilikteki alt set sayısını s, veri setindeki gözlem sayısını n ile ifade edersek, her bir alt setteki gözlem sayısı n/s olacaktır. Burada sırasıyla her bir çapraz geçerlilik alt seti dışarıda bırakılarak geri kalan s-1 alt setteki gözlem değerleri kullanılarak regresyon modeli elde edilir. Elde edilen bu regresyon modeli dışarıda bırakılan alt setteki gözlem değerlerine uygulanarak RMSE değeri hesaplanır. Bu işlem s defa uygulanarak elde edilen RMSE değerlerinin ortalaması alınır. Elde edilen ortalama RMSE değerine regresyon modelinin çapraz geçerlilik RMSE değeri denir. Çapraz geçerlilik yönteminde s sayısının n sayına eşit olma (s=n) durumu tek gözlemli çapraz geçerlilik yöntemi olarak adlandırılmaktadır. Burada sırası ile her bir gözlem değeri veri setinin dışında tutularak geriye kalan gözlem değerleri ile regresyon modeli elde edilmekte ve modelin RMSE değeri dışarıda tutulan gözlem değeri üzerinden hesaplanmaktadır. Bu işlem n defa yapılmakta ve elde edilen RMSE değerlerinin ortalaması regresyon modelinin ortalama RMSE değeri olarak ifade edilmektedir. Tahmin amaçlı model seçiminde tek gözlemli çapraz geçerlilik yöntemi aynı veri setinden elde edilen farklı sayıda değişken içeren modellerin karşılaştırılmasında kullanılabilen bir ölçüdür (Montgomery vd., 2012).

Genetik algoritmanın bulduğu en iyi değişken setleri (5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90 ve 95 değişkenli) kullanılarak elde edilen regresyon modellerinin tek gözlemli çapraz geçerlilik yöntemi ile elde edilen ortalama RMSE değerleri Tablo 2'de verilmiştir.

Tablo 2. Genetik algoritmanın seçtiği değişken setlerinin Tek Gözlemli Çapraz Geçerlilik RMSE değerleri

Değişken sayısı	Tek gözlemli çapraz geçerlilik RMSE	Değişken sayısı	Tek gözlemli çapraz geçerlilik RMSE
5	0,045020	55	0,011703
10	0,008802	60	0,013651
15	0,008015	65	0,008222
20	0,017087	70	0,002133
25	0,010299	75	0,006697
30	0,000802	80	0,022020
35	0,006567	85	0,026723
40	0,005853	90	0,031921
45	0,000870	95	0,031985
50	0,007092	99	0,033816

Tablo 2'deki değerlerin şekil 6'daki grafiksel gösterimine bakıldığında regresyon modeline dahil olan değişken sayısının artmasıyla çapraz geçerlilik RMSE değerleri dalgalanmakla birlikte genel olarak değişken sayısının artmasıyla düşüş eğilimde daha sonraları da artış eğilimde olduğu görülmektedir. Modele dahil olan bağımsız değişken sayısının artması regresyon modelinin RMSE değerini her ne kadar düşürse de modelin tahmin gücünü ifade eden çapraz geçerlilik RMSE değerini düşürmemektedir.



Şekil 6. Genetik algoritmanın seçtiği değişken setlerinin tek gözlemli çapraz geçerlilik RMSE değerleri

Şekil 6'dan görüldüğü gibi çapraz geçerlilik RMSE değeri en düşük olan regresyon modeli 30 değişken içeren regresyon modelidir. Dolayısıyla genetik algoritma tarafında seçilen 30 değişken göreceli olarak en iyi değişken seti olduğu sonucuna varılmıştır. Genetik algoritmanın bulduğu en iyi değişken setini oluşturan değişkenler Tablo 3'de verilmiştir.

Tablo 3. Genetik algoritma tarafından bulunan en iyi değişken seti

agePct65up	MedOwnCostPctIncNoMtg	PctHousOwnOcc	pctWInvInc
AsianPerCap	MedRentPctHousInc	PctIlleg	PctWorkMom
HispPerCap	NumInShelters	PctKids2Par	pctWRetire
HousVacant	NumStreet	PctPersDenseHous	PctYoungKids2Par
indianPerCap	PctEmploy	PctSameCity85	PersPerOccupHous
MalePctDivorce	PctForeignBorn	pctUrban	racepctblack
MalePctNevMarr	PctHousLess3BR	PctVacantBoarded	
MedNumBR	PctHousOccup	pctWFarmSelf	

4.3.Önerilen GA Metodunun SPSS Değişken Seçim Yöntemi ile Karşılaştırılması

Önerilen Genetik algoritma tarafından bulunan sonuçların karşılaştırılması için SPSS programı kullanılmıştır. SPSS programı değişken sayısı k'ya bağlı olarak değişken seçimi problemini $k \leq 20$ ve $k > 20$ olmak üzere iki bölüme ayırmaktadır. Eğer $k \leq 20$ ise olası bütün regresyon modellerini (En iyi alt set yöntemi) ele alınan bir kritere (R^2, ASE ya da AICC) göre değerlendirerek en iyi modeli seçmekte; $k > 20$ ise ileriye adımsal yöntem ile En iyi alt set yöntemini bir arada kullanan hibrit bir yöntem kullanmaktadır (IBM Corp., 2010).

Değişken sayısı $k > 20$ olması durumunda öncelikle ileriye adımsal yöntem ele alınan bir kritere göre uygulanarak p adet değişken seçilmektedir. Seçilen değişken sayısına bağlı olarak aşağıdaki yaklaşımlardan uygun olanı kullanılmaktadır.

- Eğer $p \leq 20$ ise En iyi alt set yöntemi uygulanır.
- Eğer $20 < p \leq 40$ ise $p - 20$ adet değişken 3. Tip kareler toplamı testi (ANOVA) ile seçilerek regresyon modeline dahil edilir. Geriye kalan 20 değişkene ise olası bütün regresyon modellerini değerlendirme yöntemi uygulanır.
- Eğer $p > 40$ ise herhangi bir değişken seçimi yapılmaz. İleriye adımsal yöntem ile seçilen değişkenlerle elde edilen regresyon modeli en iyi model varsayılır.

SPSS programı farklı sayıda bağımsız değişken içeren regresyon modellerinin karşılaştırılmasında Ayarlanmış Çoklu Belirlilik Katsayısı (adj.R^2), Aşırı Uygunluk Önleme kriteri (ASE) ya da Düzeltilmiş Akaike Bilgi kriteri (AICC) kullanmaktadır (IBM Corp., 2010). Aşırı Uygunluk Önleme kriteri (ASE) kullanıldığında, modelin veri değerlerine ezbere oldukça uygun bir şekilde uyması durumunu önlemek için modelin oluşturulacağı öğrenme setinden veri değerlerinin tesadüfi olarak yaklaşık %30'u dışarıda bırakılır. Bu set aşırı uygunluk önleme seti olarak adlandırılır. Elde edilen modelin uygunluğu bu set üzerinden RMSE değeri olarak elde edilir. Ayarlanmış Çoklu Belirlilik Katsayısı (adj.R^2) modelin öğrenme setine nasıl uyduğunu gösteren ve modeldeki değişken sayısına göre ayarlanmış bir kriterdir. Düzeltilmiş Akaike Bilgi kriteri (AICC) modelin elde edildiği öğrenme verisinin (likelihood) en çok olabilirliğine dayanan ve aşırı karmaşık (fazla sayıda bağımsız değişken içeren) modelleri cezalandırmaya ayarlanmış bir ölçüdür.

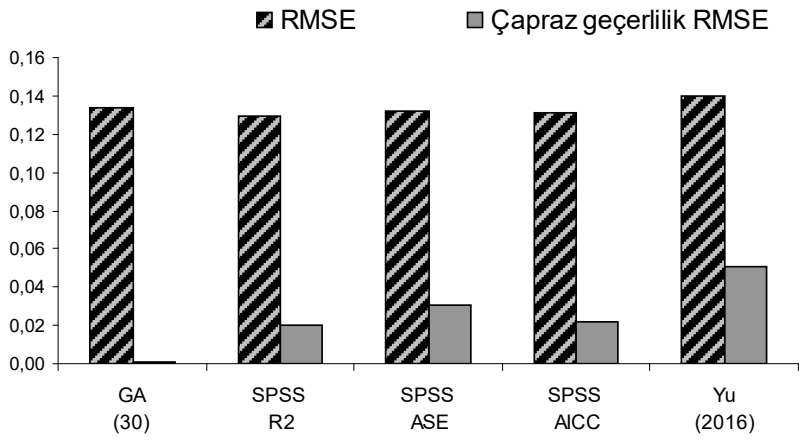
SPSS programı ile yukarıda açıklanan yöntem, üç model seçme kriteri (ASE, adj.R^2 ve AICC) kullanılarak Toplumlar ve Suç veri setine uygulamıştır. Elde edilen değişken setleri ve bu setlerle oluşturulan regresyon modellerinin RMSE değerleri Tablo 4'de verilmiştir. Elde edilen değişkenler EK 1'de listelenmiştir. Tablo 4'deki verilerin grafiksel gösterimi Şekil 7'de sunulmuştur.

Yu (2016), bir sürekli bağımlı değişkenle belirli bir fonksiyonel ilişki varsayımda bulunmayan çok sayıda bağımsız değişken içeren bir veri setinden değişken seçimi için ileriye doğru kademeli bir yöntem önermiştir. Yu (2016) tarafından önerilen değişken seçim yöntemi Topluluklar ve Suç veri setine uygulanmış ve 8 adet değişken (PctKids2Par, racePctWhite, FemalePctDiv, pctWInvInc, PctPersDenseHous, PctIlleg, HousVacant, racePctHisp) en iyi değişken alt seti olarak bulunmuştur. Yu (2016) tarafından seçilen değişken seti ile oluşturulan regresyon modelinin RMSE değerleri de karşılaştırma yapabilmek amacıyla Tablo 4'de verilmiştir.

Tablo 4. SPSS programı ile değişken seçimi ve değişken setleriyle elde edilen regresyon modellerinin RMSE değerleri

Yöntemler	Değişken sayısı	RMSE	Tek gözlemli çapraz geçerlilik RMSE
GA	30	0,1337293	0,0008018
SPSS R^2	56	0,1293886	0,0202249
SPSS ASE	56	0,1321208	0,0303482
SPSS AICC	36	0,1312888	0,0220450
Yu (2016)	8	0,1397784	0,0509670

Tablo 4'deki SPSS programı ile bulunan değişken setlerinin sonuçları genetik algoritma ile bulunan en iyi değişken setinin regresyon modeli sonuçlarıyla karşılaştırıldığında yaklaşık daha düşük RMSE değerine sahip olmakla birlikte daha yüksek çapraz geçerlilik RMSE değeri vermektedir.



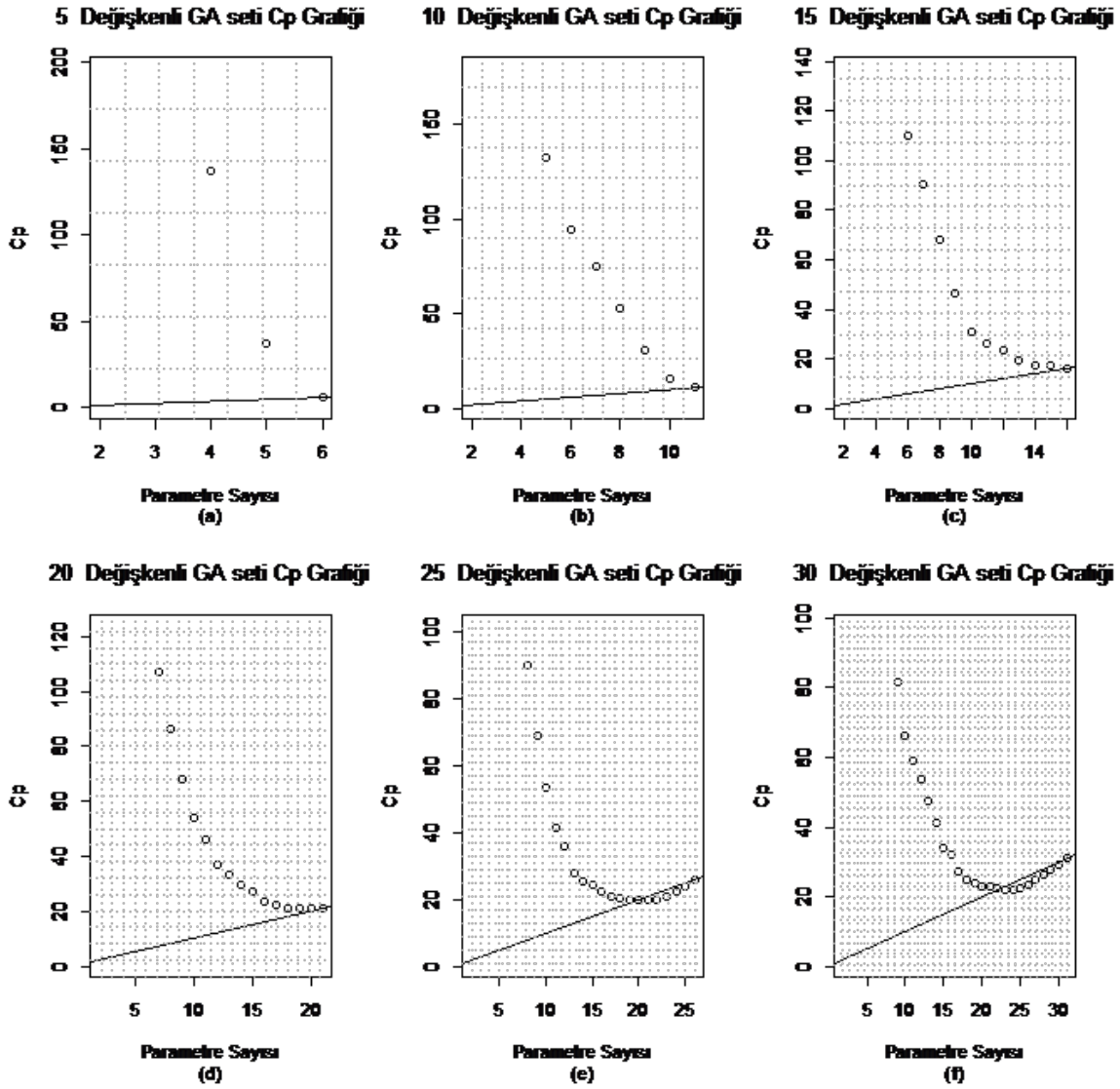
Şekil 7. SPSS programının seçtiği değişkenlerin regresyon modellerinin RMSE değerleri

Genetik algoritmanın seçtiği değişken seti ile oluşturulan regresyon modelinin, modelin oluşturulmasında kullanılmayan test verilerini daha az hata ile tahmin edebildiği görülmektedir. Genetik algoritmanın kullandığı amaç fonksiyonu regresyon modelinin veri değerlerini ezberlemesini önlemekte ve geleceğe yönelik daha tutarlı tahmin yapan modelleri (değişken setlerini) seçebilmektedir.

4.4.Genetik Algoritmanın Bulduğu Değişken Setlerinin Değerlendirilmesi

Genetik algoritma tarafından seçilen 5, 10, 15, 20, 25 ve 30 değişken içeren değişken setlerinin ayrı ayrı değerlendirmeleri R Programlama (R Core Team, 2012) ortamında regresyon alt set seçimi (Leaps) program paketi ile yapılmıştır (Lumley, 2009). Leaps program paketi etkin bir dal ve sınır algoritması ile geniş kapsamlı bir araştırma yaparak yani olası bütün alt setleri araştırarak doğrusal regresyonda bağımlı değişkeni en iyi tahmin eden bağımsız değişkenlerin alt setini (kümesini) bulmaktadır. Leaps programı ele alınan bir kritere göre (R^2, Cp gibi) her bir model büyüklüğünde (değişken sayısında) en iyi modeli bulmaktadır. Leaps programı en fazla 31 bağımsız değişken içeren regresyon problemlerinde olası bütün regresyon modellerini belli bir kritere göre değerlendirerek en iyi regresyon modelini seçmektedir.

Önerilen Genetik algoritmanın bulduğu 5, 10, 15, 20, 25 ve 30 değişken içeren değişken setlerinin değerlendirmeleri Leaps programında Cp kriterine göre yapılmıştır. Daha önce açıklandığı üzere en iyi regresyon modeli modeldeki parametre sayısının Cp değerine en yakın olduğu modeldir. Leaps programı regresyondaki sabit terimi de parametre sayısına dahil ettiğinden dolayı modeldeki bağımsız değişken sayısı parametre sayısının bir eksiğine karşılık gelmektedir. Değişken setlerinin Cp grafikleri Şekil 8'de verilmiştir. Cp grafiklerinde her bir model büyüklüğünde (değişken sayısında) en iyi regresyon modelinin (yani en düşük Cp değerine sahip modelin) Cp değeri verilmiştir. Grafikler üzerinde gösterilen doğrular, orijinden geçen ve eğimi 1 olan doğrulardır. Bu doğrular grafik üzerinde bir regresyon modelinin Cp değerinin o modelin değişken sayısına soldan sağa doğru ilk defa en yakın olduğu değişken sayısını tespit etmek için kullanılmaktadır.



Şekil 8. Genetik algoritmanın seçtiği değişken setlerinin Cp grafikleri

Şekil 8'deki Cp grafiklerinden kolay bir şekilde görüleceği üzere Leaps programı Genetik algoritma tarafından seçilen 5, 10, 15 ve 20 değişkenli setlerden Cp kriterine göre daha iyi performans gösteren herhangi bir değişken alt seti bulamamıştır. Leaps programı GA'nın seçtiği 25 ve 30 değişkenli setler için en iyi alt setleri 20 ve 22 değişkenli setler olarak tespit etmiştir.

Genetik algoritma tarafından seçilen değişken setleri SPSS programında da değerlendirilmiştir. SPSS programı 20 değişkene kadar olası bütün regresyon modellerini ele alınan bir kritere (R^2, ASE ya da AICC) göre değerlendirerek en iyi modeli seçebildiğinden dolayı GA tarafından seçilen 5, 10, 15 ve 20 değişkenli setlere değişken seçimi uygulanmıştır. Elde edilen sonuçlar Tablo 5'de sunulmuştur.

Tablo 5. Genetik algoritmanın bulduğu değişken setlerinin SPSS programı değişken seçimi sonuçları

	SPSS R^2		SPSS ASE		SPSS AICC	
Genetik Algoritma Setleri	adj.R^2	Değişken sayısı	ASE	Değişken sayısı	Bilgi Ölçütü	Değişken sayısı
5	0,635	5	0,018	5	-7813,186	5
10	0,657	10	0,018	9	-7931,392	10
15	0,659	15	0,017	12	-7941,535	15
20	0,661	20	0,017	17	-7946,777	20

SPSS programı değişken seçim yöntemi de Leaps programı gibi R^2 ve AICC kriterlerini kullanarak GA'nın seçtiği setlerden daha iyi performans gösteren herhangi bir alt setini bulamamıştır. SPSS programında değişken seçimi için ASE kriterini kullandığında ise bu kritere göre daha iyi performans gösteren alt setleri bulabilmiştir. Bu durum ASE kriterinin uygulanmasından kaynaklandığı sonucuna varılmıştır. Çünkü ASE kriteri uygulamasında veri gözlem değerleri tesadüfi olarak öğrenme ve test seti olarak ikiye bölündüğünden dolayı seçilen değişkenler ve seçilen değişken sayısı setlerdeki gözlem değerlerine bağlı olarak farklı farklı olabilmektedir.

5.**SONUÇ VE ÖNERİLER**

Genel olarak GA tarafından seçilen alt setlerin performansına bakıldığında R^2, AICC ve Cp kriterlerinin hepsi için iyi ve tutarlı sonuçlar verdiği sonucuna varılmaktadır. He ne kadar GA uygulamaları fazla miktarda bilgisayar hesaplama zamanı gerektirse de elde ettiği sonuçlar klasik ve standart yöntemlerden oldukça iyi olduğu görülmektedir. Önerilen genetik algoritmanın genetik işlemleri problemin çözüm kümesini etkin bir şekilde araştırabilmektedir. Önerilen genetik algoritmanın en önemli ve diğer yöntemlere kıyasla üstün tarafı problemin çözüm kümesini araştırma yönteminin amaç fonksiyonundan bağımsız olmasıdır. Genetik algoritmanın amaç fonksiyonu araştırmacı tarafından istenilen amaca uygun bir şekilde değiştirilebilmektedir.

GA algoritmaların kullanıcı tarafından belirlenmesi ve optimize edilmesi gereken parametreleri (mutasyon oranı, çaprazlama oranı, popülasyon büyüklüğü, sona erme kriteri vb.) bulunmaktadır. Dolayısıyla bu yöntemi kullanan araştırmacıların ya da kişilerin bu konularda yeterli bilgi ve tecrübeye sahip olmaları elde edilecek sonuçların tutarlı ve iyi olması açısından önem taşımaktadır.

Birçok araştırmacı değişken seçimi yöntemlerinin kullanımına ilişkin kaygılarını ifade etmektedir. Değişken seçimi yöntemlerine en kritik itiraz tahmin amaçlı kullanılsa bile bu yöntemlerin optimal değişken setini seçmekte başarısız olacağına yöneliktir (Ruengvirayudh, Brooks, 2016). Ancak Genetik algoritmalar, pratik ve güçlü bir optimizasyon ve arama yöntemi olarak ortaya çıkmış ve birçok alanda başarılı sonuçlar vermiştir. Bu çalışmada elde edilen sonuçlar iyi bir şekilde tasarlanmış bir genetik algoritmanın problemin çözümüne ilişkin kabul edilebilir sonuçlar verebileceğini göstermektedir. Günümüzde dijital ortamda veri toplama ve depolama oldukça kolaylaşmıştır. Dolayısıyla karar verme ve geleceğe yönelik öngörülerde bulunmak için çok sayıda değişken içeren verilerin analiz edilmesi gerekmektedir. Klasik ve standart yöntemlerin uygulanmasının zor olduğu ya da tahmin edici sonuçlar verememesi durumunda genetik algoritmalar iyi bir alternatif yöntem olarak karşımıza çıkmaktadır.

KAYNAKÇA

Baker, J. (1985). "Adaptive Selection Methods for Genetic Algorithms", Hillsdale, NJ, United States: L. Erlbaum Associates Inc..In Grefenstette , J. J. (Eds.), **The First International Conference on Genetic Algorithms and Their Applications** (p. 101-111).

Chatterjee, S., Hadi, A.S. (2006). **Regression Analysis by Example**, 4 ed. New Jesey: Wiley Series.

Communities and Crime Data Set, UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

De Jong, K. A. (1975). *Analysis of the Behavior of a Class of Genetic Adaptive Systems*, Ph.D. Thesis, Department of Computer and Communication Sciences: University of Michigan.

Field, P. (1995). *A Multary Theory for Genetic Algorithms: Unifying Binary and Nonbinary Problem Representations*, Ph.D. Thesis, Department of Computer Science. London: University of London.

Fogel, L. J. (1997). "A Retrospective View and Outlook on Evolutionary Algorithms". Berlin, Germany: Springer-Verlag.In Reusch, B. (Eds.), **Computational Intelligence: Theory and Applications**, 5th Fuzzy Days (p. 337-342).

Goldberg, D. E., & Lingle, R. (1985). "Alleles, Loci, and the Traveling Salesman Problem", Hillsdale, New Yersey, United States: Lawrence Erlbaum.In Grefenstette, J. J. **International Conference on Genetic Algorithms and Their Applications** (p. 154-159).

IBM Corp. Released 2010. IBM SPSS Statistics for Windows, Version 19.0. Armonk, NY: IBM Corp.

Jung, M., & Zscheischler, J. (2013). "A guided hybrid genetic algorithm for feature selection with expensive cost functions". **Procedia Computer Science**,18, 2337-2346.

Kabir, M. M., Shahjahan, M., & Murase, K. (2011). "A new local search based hybrid genetic algorithm for feature selection". **Neurocomputing**, 74(17), 2914-2928.

Kewley, R., Embrechts, M. J., & Breneman, C. M. (1998). "Neural Network Analysis for Data Strip Mining Problems", **Intelligent Engineering Systems through Artificial Neural Networks**, vol. 8, C. Dagli, Ed. Nashville - Missouri: ASME Press, pp. 391-396.

Leardi, R., Boggia, R., & Terrile, M. (1992). "Genetic algorithms as a strategy for feature selection". **Journal of chemometrics**, 6(5), 267-281.

Lumley, T. (2009). Leaps: regression subset selection using Fortran code by Alan Miller, R package version 2.9. http://CRAN.R-project.org/package=leaps

Mallows, C. L. (1973). "Some Comments on Cp", **Technometric**, vol. 15, pp. 661-675.

Michalewicz, Z. (1996). **Genetic Algorithms + Data Structures = Evolution Programs**, 2. ed, Springer-Verlag, New York, United States.

Miller, A. J. (1984). "Selection of Subsets of Regression Variables", **Journal of the Royal Statistical Society**. Series A (General), Vol. 147, No. 3, 389 -425.

Montgomery, D. C., Peck, E. A.,Vining, G. G. (2012). **Introduction to Linear Regression Analysis**, 5 ed., John Willey & Sons, Inc., New Jersey, United States.

Özdemir, M. (2011). "Genetik Algoritma Kullanılarak Portföy Seçimi", **İktisat İşletme ve Finans**, Cilt 26, Sayı. 299, Sayfa: 67–89. DOI: 10.3848/iif.2011.299.2831

Paterlini, S., & Minerva, T. (2010). "Regression model selection using genetic algorithms". In Proceedings of the 11th WSEAS international conference on nural networks and 11th WSEAS international conference on evolutionary computing and 11th WSEAS international conference on Fuzzy systems (pp. 19-27). World Scientific and Engineering Academy and Society (WSEAS).

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Ruengvirayudh P., Brooks, G. P. (2016)." Comparing Stepwise Regression Models to the Best-Subsets Models", **the Art of Stepwise General Linear Model Journal**, Vol. 42(1) pp. 1-14

Thompson, M. L. (1978). "Selection of Variables in Multiple Regression: Part I. A Review and Evaluation", **International Statistical Review**, Vol. 46, No. 1, pp. 1-19.

Tsai, C., Eberle, & W., Chu, C. (2013). "Genetic algorithms in feature and instance selection", **Knowledge-Based Systems** , 39, 240–247.

Van Rooji, A. J. F., Jain, L. C., & Johnson, R. P. (1996). "Neural Networks Training Using Genetic Algorithms". Series in Machine Perception and Artificial Intelligence, Vol. 26, pp.130, Singapore: World Scientific.

Vose, M. D. (2010). **The Simple Genetic Algorithm: Foundations and Theory**. Cambridge, Massachussets, United States: MIT Press.

Whitley, D. (1989). "The GENITOR Algorithm and Selection Pressure: Why Rank-based Allocation of Reproductive Trials is Best", San Mateo, CA, United States: Morgan Kaufmann. In Schaffer , J. D. (Eds.), **Third International Conference on Genetic Algorithms** (p. 116-121).

Yu, T. (2016). "Nonlinear variable selection with continuous outcome: a nonparametric incremental forward stagewise approach". **arXiv preprint arXiv:1601.05285**.

1.EK

Bağımlı değişken: ViolentCrimesPerPop

Ayırt edici olmayan değişkenler: state, county, community, communityname, fold

Eksik veri içeren değişkenler: PolicBudgPerPop, LemasGangUnitDeploy, LemasPctPolicOnPatr, PolicOperBudg, PolicCars, PolicAveOTWorked, NumKindsDrugsSeiz, OfficAssgnDrugUnits, PctPolicMinor, PctPolicAsian, PctPolicHisp, PctPolicBlack, PctPolicWhite, RacialMatchCommPol, PolicPerPop, PolicReqPerOffic, LemasTotReqPerPop, LemasTotalReq, LemasSwFTFieldPerPop, LemasSwFTFieldOps, LemasSwFTPerPop, LemasSwornFT, OtherPerCap

Ek Tablo 1. Topluluklar ve Suç Veri Setinin Analizlerde Kullanılan Değişkenleri

No	Değişken Adı	No	Değişken Adı	No	Değişken Adı
1	agePct12t21	34	PctBSorMore	67	PctSpeakEnglOnly
2	agePct12t29	35	PctEmplManu	68	PctTeen2Par
3	agePct16t24	36	PctEmploy	69	PctUnemployed
4	agePct65up	37	PctEmplProfServ	70	pctUrban
5	AsianPerCap	38	PctFam2Par	71	PctUsePubTrans
6	blackPerCap	39	PctForeignBorn	72	PctVacantBoarded
7	FemalePctDiv	40	PctHousLess3BR	73	PctVacMore6Mos
8	HispPerCap	41	PctHousNoPhone	74	pctWFarmSelf
9	householdsize	42	PctHousOccup	75	pctWInvInc
10	HousVacant	43	PctHousOwnOcc	76	PctWOFullPlumb
11	indianPerCap	44	PctIlleg	77	PctWorkMom
12	LandArea	45	PctImmigRec10	78	PctWorkMomYoungKids
13	LemasPctOfficDrugUn	46	PctImmigRec5	79	pctWPubAsst
14	MalePctDivorce	47	PctImmigRec8	80	pctWRetire
15	MalePctNevMarr	48	PctImmigRecent	81	pctWSocSec
16	medFamInc	49	PctKids2Par	82	pctWWage
17	medIncome	50	PctLargHouseFam	83	PctYoungKids2Par
18	MedNumBR	51	PctLargHouseOccup	84	perCapInc
19	MedOwnCostPctInc	52	PctLess9thGrade	85	PersPerFam
20	MedOwnCostPctIncNoMtg	53	PctNotHSGrad	86	PersPerOccupHous
21	MedRent	54	PctNotSpeakEnglWell	87	PersPerOwnOccHous
22	MedRentPctHousInc	55	PctOccupManu	88	PersPerRentOccHous
23	MedYrHousBuilt	56	PctOccupMgmtProf	89	PopDens
24	numbUrban	57	PctPersDenseHous	90	population
25	NumIlleg	58	PctPersOwnOccup	91	racePctAsian
26	NumImmig	59	PctPopUnderPov	92	racepctblack
27	NumInShelters	60	PctRecentImmig	93	racePctHisp
28	NumStreet	61	PctRecImmig10	94	racePctWhite
29	NumUnderPov	62	PctRecImmig5	95	RentHighQ
30	OwnOccHiQuart	63	PctRecImmig8	96	RentLowQ
31	OwnOccLowQuart	64	PctSameCity85	97	RentMedian
32	OwnOccMedVal	65	PctSameHouse85	98	TotalPctDiv
33	PctBornSameState	66	PctSameState85	99	whitePerCap

Ek Tablo2. Topluluklar ve Suç Veri Setinden Seçilen Değişkenler

Yöntem	Değişken sayısı	Seçilen Değişken No	RMSE	Tek gözlemli çapraz geçerlilik RMSE
GA	5	10, 14, 44, 57, 94	0,140550	0,045020
GA	10	4, 10, 14, 28, 44, 49, 57, 70, 77, 92	0,136100	0,008802
GA	15	5, 10, 14, 20, 22, 28, 39, 44, 49, 57, 70, 77, 81, 85, 92	0,135410	0,008015
GA	20	5, 8, 10, 11, 13, 14, 20, 22, 27, 28, 39, 42, 44, 57, 64, 70, 75, 77, 81, 92	0,134887	0,017087
GA	25	2, 4, 5, 8, 10, 11, 13, 14, 20, 22, 27, 28, 40, 42, 44, 49, 57, 65, 70, 74, 75, 77, 86, 89, 92	0,134318	0,010299
GA	30	4, 5, 8, 10, 11, 14, 15, 18, 20, 22, 27, 28, 36, 39, 40, 42, 43, 44, 49, 57, 64, 70, 72, 74, 75, 77, 80, 83, 86, 92	0,133729	0,000802
GA	35	4, 5, 8, 9, 10, 11, 14, 20, 22, 27, 28, 35, 36, 38, 40, 42, 43, 44, 47, 49, 57, 61, 65, 68, 70, 74, 75, 77, 82, 83, 86, 91, 92, 96, 99	0,133714	0,006567
GA	40	1, 4, 5, 8, 9, 10, 14, 15, 20, 21, 22, 27, 28, 35, 36, 37, 38, 39, 40, 42, 43, 44, 49, 56, 57, 64, 67, 70, 71, 72, 73, 74, 75, 77, 78, 81, 83, 85, 86, 92	0,133608	0,005853
GA	45	1, 4, 5, 8, 9, 10, 13, 14, 15, 20, 22, 27, 28, 32, 35, 36, 37, 38, 39, 40, 42, 43, 44, 49, 56, 57, 58, 64, 66, 67, 70, 71, 72, 73, 74, 75, 77, 78, 80, 81, 83, 85, 86, 87, 92	0,133253	0,000870
GA	50	1, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 18, 20, 22, 26, 27, 28, 35, 36, 37, 38, 39, 40, 42, 43, 44, 49, 56, 57, 58, 64, 67, 70, 71, 72, 73, 74, 75, 77, 80, 81, 82, 83, 85, 86, 87, 89, 92	0,132675	0,007092
GA	55	3, 4, 5, 6, 7, 10, 11, 12, 13, 14, 15, 20, 22, 24, 25, 26, 27, 28, 29, 30, 31, 32, 39, 40, 41, 44, 45, 46, 47, 48, 50, 51, 52, 53, 54, 55, 57, 59, 60, 61, 62, 63, 69, 71, 72, 74, 76, 79, 88, 89, 90, 91, 92, 93, 98	0,132530	0,011703
GA	60	1, 2, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 20, 21, 22, 23, 25, 26, 27, 28, 31, 32, 34, 35, 37, 39, 40, 42, 43, 44, 48, 49, 55, 56, 57, 64, 65, 66, 70, 71, 72, 73, 74, 75, 77, 78, 80, 81, 82, 83, 85, 86, 87, 89, 92, 96, 97, 98, 99	0,131765	0,013651
GA	65	1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 20, 21, 22, 25, 26, 27, 28, 30, 31, 32, 33, 35, 36, 37, 38, 39, 40, 42, 43, 44, 45, 46, 47, 48, 49, 55, 56, 57, 58, 61, 63, 64, 65, 67, 70, 71, 72, 73, 74, 75, 77, 81, 82, 83, 85, 86, 87, 89, 91, 92, 93	0,132013	0,008222
GA	70	1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21, 22, 23, 25, 26, 27, 28, 30, 34, 35, 37, 40, 42, 43, 44, 45, 46, 47, 48, 49, 55, 56, 57, 58, 60, 61, 62, 63, 64, 65, 67, 70, 74, 75, 77, 78, 80, 81, 82, 83, 84, 85, 86, 87, 89, 91, 92, 93, 96, 97, 98, 99	0,131858	0,002133
GA	75	1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 20, 21, 22, 23, 25, 26, 27, 28, 30, 32, 34, 35, 36, 37, 38, 39, 40, 42, 43, 44, 45, 46, 47, 48, 49, 55, 56, 57, 58, 60, 61, 62, 63, 64, 65, 66, 67, 70, 71, 72, 74, 75, 77, 78, 81, 82, 83, 84, 85, 86, 87, 89, 91, 92, 93, 96, 97, 98, 99	0,130976	0,006697

GA	80	1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 31, 33, 34, 35, 36, 37, 38, 39, 40, 42, 43, 44, 45, 46, 47, 48, 49, 51, 54, 55, 56, 57, 58, 61, 62, 63, 64, 65, 66, 67, 70, 71, 72, 74, 75, 77, 78, 80, 81, 82, 83, 84, 85, 86, 87, 89, 90, 91, 92, 93, 96, 97, 98, 99	0,130278	0,022020
GA	85	1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21, 22, 23, 25, 26, 27, 28, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 42, 43, 44, 45, 46, 47, 48, 49, 50, 54, 55, 56, 57, 58, 60, 61, 62, 63, 64, 65, 66, 67, 70, 71, 72, 74, 75, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 96, 97, 98, 99	0,129937	0,026723
GA	90	1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 53, 54, 55, 56, 57, 58, 60, 61, 62, 63, 64, 65, 66, 67, 70, 71, 72, 74, 75, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 96, 97, 98, 99	0,129737	0,031921
GA	95	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 96, 97, 98, 99	0,129225	0,031985
SPSS R^2	56	2, 5, 8, 10, 11, 13, 14, 15, 16, 18, 19, 20, 21, 22, 25, 26, 27, 28, 31, 32, 34, 35, 36, 39, 40, 42, 43, 44, 49, 50, 52, 54, 55, 57, 58, 59, 64, 70, 71, 72, 73, 74, 75, 77, 78, 80, 81, 82, 86, 88, 92, 93, 95, 96, 98, 99	0,1293886	0,0202249
SPSS ASE	56	1, 2, 5, 6, 7, 8, 10, 13, 14, 15, 16, 17, 19, 20, 21, 24, 27, 34, 36, 39, 41, 43, 44, 46, 48, 52, 53, 58, 59, 62, 64, 65, 67, 68, 70, 72, 74, 75, 78, 79, 80, 81, 82, 84, 85, 86, 87, 88, 90, 92, 93, 94, 96, 97, 98, 99	0,1321208	0,0303482
SPSS AICC	36	5, 8, 10, 11, 13, 14, 19, 20, 21, 22, 24, 26, 27, 28, 35, 36, 39, 44, 49, 54, 57, 59, 70, 72, 73, 74, 75, 77, 80, 82, 84, 92, 93, 96, 98, 99	0,1312888	0,0220450
Yu (2016)	8	7, 10, 44, 49, 57, 75, 93, 94	0,1397784	0,0509670