Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi
Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering

Basılı/Printed ISSN: 1302-9304. Elektronik/Online ISSN: 2547-958X

Makine Öğrenmesinde Rastgele Oran ve Sıralı Küme Örneklemesi Yöntemlerinin Doğrusal Regresyon Modellerine Etkisi

Effect of Random Rate and Ranked Set Sampling Methods on Linear Regression Models in Machine Learning

Sena Aslan [1*], Tuğba Yıldız [2]

[1] Dokuz Eylül Üniversitesi Fen Bilimleri Enstitüsü İstatistik Anabilim Dalı Veri Bilimi Programı, İzmir, TÜRKİYE
[2] Dokuz Eylül Üniversitesi Fen Fakültesi İstatistik Bölümü, İzmir, TÜRKİYE
Sorumlu Yazar / Corresponding Author *: tugba.ozkal@deu.edu.tr

Geliş Tarihi / Received: 01.12.2020 *Araştırma Makalesi/Research Article*
Kabul Tarihi / Accepted: 07.06.2021 *DOI:10.21205/deufmd.2022247004*
Atıf şekli/ How to cite: ASLAN, S.,YILDIZ, T.(2022). Makine Öğrenmesinde Rastgele Oran ve Sıralı Küme Örneklemesi Yöntemlerinin Doğrusal Regresyon Modellerine Etkisi. DEUFMD, 24(70), 29-36.

Öz

Makine öğrenmesi en basit tanımıyla, insana ait özellik ve davranışları bilgisayara öğretmektir. Makine öğrenmesi algoritmaları kendilerine verilen örnek olayları inceleyerek öğrenir ve bu örnek olaylar üzerinden genelleme yapma yeteneği kazanır. Modele öğretilmek istenilenlerin öğretileceği kısım eğitim seti, ne kadar iyi öğrendiğinin test edildiği kısım ise test seti olarak adlandırılır. Makine öğrenmesi literatüründe var olan çalışmalarda, veri seti bölme işlemi kullanıcının istediği rastgele bir oranda gerçekleşmektedir. Bu çalışmada, Kaliforniya Üniversitesi'nin lisansüstü öğrenci kabul kriterleri göz önünde bulundurularak, Hindistan'daki öğrenciler için oluşturulan yüksek lisans başvuru verileri, rastgele oran yöntemi ve sıralı küme örneklemesi (SKÖ) ile bölünmüş, elde edilen eğitim setleri kullanılarak doğrusal regresyon modelleri oluşturulmuştur. Daha sonra, test setleri kullanılarak modellerin hata kareler ortalamalarının karekökleri (HKOK) üzerinden, veri seti bölme yöntemlerinin performans karşılaştırması yapılmıştır. SKÖ yöntemi ile, temel bileşenler, kısmi en küçük kareler ve ridge regresyon modelleri için tek bir durum dışında, rastgele oran yöntemine göre daha düşük hata değerlerine ulaşılmıştır. Elastic net regresyon modeli hariç, diğer doğrusal regresyon modellerinde, SKÖ yöntemi ile, rastgele oran yönteminden daha iyi sonuçlar elde edilmiştir.
Anahtar Kelimeler: *Makine Öğrenmesi, Sıralı Küme Örneklemesi, Doğrusal Regresyon Modelleri, HKOK*

Abstract

The simplest definition of machine learning is to teach human characteristics and behaviors to the computer. Machine learning algorithms learn by examining case studies given to them and gain the ability to generalize through these events. The part where the model will be taught what is wanted is called the training set, and the part where it is tested how well it is learned is called the test set. In the studies that exist in the machine learning literature, data set splitting occurs at a random rate that the user wants. In this study, considering the graduate student admission criteria of the University of California, graduate application data created for students in India were divided by both random and ranked set sampling (RSS), linear regression models were created using the obtained

training sets. Then, using the test sets, the performance comparison of the data set splitting methods was made based on the root mean square error (RMSE) of the models. With the RSS method, lower error values were obtained for principal components, partial least squares and ridge regression models compared to the random rate method except for a single case. In other linear regression models except elastic net regression model, better results were obtained with the RSS method compared to the random rate method.

Keywords: *Machine Learning , Ranked Set Sampling, Linear Regression Models, RMSE*

1. Giriş

Makine öğrenmesi, yapay zekada sayısal öğrenme ve model tanıma çalışmalarından bilgisayar biliminin bir alt dalı olarak geliştirilmiştir. Yapılan akademik çalışmalar makinelerin belirli bir aşamadan sonra verileri öğrenmek zorunda olduğunu göstermiş ve araştırmacılar bunun için çalışmışlardır. Makine öğrenmesi terimini ilk olarak Samuel [1] ortaya atmış ve dama oyunu üzerinden anlatmaya çalışmıştır. Bu terimi "Program, yazan kişinin oynayabileceğinden daha iyi bir dama oyunu oynamayı öğretecek" diyerek açıklamıştır.

Makine öğrenmesinde en iyi modeli elde etmek için model parametrelerinin doğru tahminlenmesi gerekmektedir. Aynı zamanda model daha önce görmediği veriler üzerinde iyi performans göstermelidir. Bunları sağlamak için veri seti, eğitim ve test seti olmak üzere ikiye ayrılır. Eğitim seti, modele öğretilmek istenenlerin öğretildiği settir. Test seti, modelin öğretilenleri ne kadar iyi öğrendiğini ölçmek için kullanılan settir.

Veri seti bölme işlemi, kullanıcının istediği oranda rastgele olarak yapılır. Örneğin; %80 eğitim seti, %20 test seti ya da %75 eğitim seti, %25 test seti olarak veri bölünebilir. Bu çalışmada rastgele bölünmüş veri seti ve SKÖ ile bölünmüş veri seti doğrusal regresyon modellerinde kullanılacaktır. Daha sonra iki veri seti bölme yöntemi ile elde edilen doğrusal regresyon modellerinin HKOK değerleri karşılaştırılacaktır.

Doğrusal regresyon modellerinde rastgele veri seti bölme işlemi yapılırken, bağımlı değişken üzerinden indeks üretme işlemi yapılır. Veri seti, üretilen indekslere göre istenilen oranda bölünebilir. SKÖ ile veri seti bölme işlemi yapılırken, istenilen sayıda gözlem SKÖ ile seçilir. Daha sonra seçilen gözlemler üzerinden indeks üretme işlemi yapılır. Üretilen indekslere göre veri seti bölünür.

2. Materyal ve Metot

Bu çalışmada rastgele orandan ve SKÖ yönteminden yararlanılarak veri seti, eğitim ve test seti olarak ikiye ayrılacaktır. Daha sonra her iki yöntemle elde edilen setler ile doğrusal regresyon modelleri oluşturularak elde edilen sonuçlar karşılaştırılacaktır.

2.1. Sıralı küme örneklemesi

Örneklemede amaç, kitleyi en iyi şekilde temsil edebilecek örneklemi bulabilmektir. Örneklem seçimi yapılırken maliyet, zaman ve emek faktörleri en aza indirgenerek kitle parametrelerinin en iyi şekilde tahminlenmesi amaçlanır. Basit rastgele örnekleme yöntemi, örnekleme yöntemleri arasında en eski ve en sık kullanılan yöntemlerden biridir. Basit rastgele örneklemede, sonlu büyüklükteki kitleden rastgele olarak n adet birim seçilir ve bu birimler örneklemi oluştururlar. Ancak basit rastgele örnekleme ile örneklem seçmek çok maliyetli ve zaman alıcı olabilmektedir. Bu yüzden basit rastgele örneklemeye alternatif olarak McIntyre [2] tarafından SKÖ geliştirilmiştir. SKÖ genellikle çevresel ve ekolojik alanlarda yapılan çalışmalarda kullanılır. Bu alanlarda seçilen örneklem birimlerinin ölçümlerinin yapılması zaman ve maliyet açısından oldukça zorlayıcıdır. Sonsuz büyüklükteki kitlelerde ilgilenilen özellikler görsel olarak sıralamaya uygun olduğunda, SKÖ daha düşük maliyetle çalışır.

McIntyre [2], SKÖ yöntemini ilk defa ortalama mera hasılasını tahmin etmek amacıyla yaptığı bir çalışmada kullanmıştır. Bu yöntem basit rastgele örneklemeye göre daha etkili bulunmuştur ve daha sonra Halls ve Dell [3] tarafından bir ormandaki bitkilerin ve otların ağırlıklarını tahmin etmek için kullanılmıştır. Evans [4] SKÖ'nün etkinliğini ve verimliliğini gözlemlemek için bir araştırma yapmış ve aynı zamanda basit rastgele örnekleme ile SKÖ'nün etkin olma durumlarını karşılaştırmıştır. Takahashi ve Wakimoto [5], SKÖ ile kitle ortalaması tahmini yapmış ve bulunan tahminin

yansız olduğunu göstermişlerdir. Aynı zamanda SKÖ ile bulunan kitle varyans tahmininin, basit rastgele örnekleme ile elde edilen varyans tahmininden daha küçük olduğunu bulmuşlardır. Ancak sıralama yaparken oluşabilecek hataları düşünmemişlerdir. Dell ve Clutter [6] sıralamada hata olsa da olmasa da kitle ortalaması tahmininin yansız olduğunu ve basit rastgele örnekleme yöntemine göre varyansının daha küçük çıkacağını göstermişlerdir. Martin vd. [7] bir ormandaki fundalık yığınının tahmin edilmesinde SKÖ yöntemini kullanmıştır. Araştırma sonucunda ortalama ve varyans tahminleri elde edildiğinde SKÖ'nün varyansının basit rastgele örneklemenin varyansına göre daha küçük çıktığı gözlemlenmiştir. İlk kez Stokes [8] SKÖ ile kitle varyansının bir tahmin edicisini elde etmiştir. Aynı zamanda bulduğu tahmin edicinin sıralama hatası olduğu durumlarda da yansız olduğunu göstermiştir. Ridout ve Cobby [9] kümede iyi bir sıralama yapılmaması, küme içi değişkenliğin kümeler arasındaki değişkenlikten fazla olması ve hesaplanan değerlerin asimetrik dağılımı gibi etkenlerin SKÖ üzerindeki etkisini göstermişlerdir. Stokes ve Sager [10] SKÖ yöntemini tüketici harcamalarıyla ilgili yaptıkları araştırmada kullanmışlar ve bu araştırmaları tüketici fiyat endeksinin belirlenmesinde etkili olmuştur. Patil, Sinha ve Tallie [11] bir şirketin gaz hatlarının oluşturduğu toprak kirliliğini SKÖ yöntemini kullanarak incelemişler, ancak araştırmadaki veriler çarpık dağıldığı için normal dağılımların yanı sıra normal olmayan dağılımları da araştırmışlardır. Bu çalışmaları sonucunda, verilerin dağılımıyla ilgili ön bilgiye sahipken örneklem seçme sorununa değinmişlerdir.

SKÖ'de örneklem seçimi iki bölümde yapılır. Örneklem seçiminin ilk bölümünde, her bir küme ilgilenilen değişkene göre küçükten büyüğe doğru sıralanır. Sıralama işlemi, kişinin tecrübeleri,görsel ölçüm veya yardımcı bir değişken yardımıyla yapılabilir. Örneklem seçiminin ikinci bölümünde, ilgili değişken bakımından birinci kümeden ilk birim, ikinci kümeden ikinci birim ve bu şekilde m'inci kümeden m'inci birim seçilir. Böylelikle m^2 birimden m birim ölçülmüş olur. Araştırmacı sıralamada hata oluşmaması için küme sayısını (m), 2, 3, 4 veya 5 olarak belirlemelidir. Patil, Sinha ve Taillie [12], m değeri 5'ten büyük seçilirse sıralamada zorluk çıkabileceğini öne sürmüşlerdir. Yeteri kadar birim elde edebilmek için işlemler r kez tekrarlanabilir. Kitleden seçilen m^2r birimden ölçülen mr birim, sıralı küme örneklemini oluşturur.

2.2. Çoklu doğrusal regresyon

Regresyon analizi, Montgomery, Peck ve Vining [13] tarafından, değişkenler arasındaki ilişkiyi açıklamak ve modellemek için kullanılan istatistiksel bir yöntem olarak açıklanmıştır. Regresyon modellerinde, bağımlı (açıklanan) değişken ve bağımsız (açıklayıcı) değişken ya da değişkenler bulunur. Çoklu doğrusal regresyon, doğrusal bir regresyon modelinde, bağımsız değişken sayısının birden fazla olduğu durumdur. Bağımsız değişkenlere karşılık gelen y_i değerleri ile yapılan örnekleme sonucu elde edilen değerler çoklu regresyon verisini oluştururlar. Bağımlı değişkendeki değişimi etkileyecek tüm etkenlerin aynı denklem içinde birlikte incelenmesidir.

$$Y_{\mathrm{i}} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_j X_{ij} + \varepsilon_i \qquad (1)$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y \qquad (2)$$

Çoklu doğrusal regresyonun amaçları; bağımlı değişkendeki değişimi açıklayabilmek, değişkenlerin diğer bağımsız değişkenlerin etkisi olmaksızın, bağımlı değişkene etkilerini araştırmak ve bağımlı değişkene ilişkin ortalama ya da tahmin değerlerini bulmaktır.

2.3. Temel bileşenler regresyonu

Gerçek hayat verilerinin çoğunda bağımsız değişkenler arasında yüksek korelasyon (ilişki) vardır ve benzer bilgileri içermektedirler. Bu durum, çoklu bağlantı sorununa sebep olmaktadır. Montgomery, Peck ve Vining [13], çoklu bağlantı problemi olan bir modelde, regresyon katsayılarına ve bağımsız değişkenlerin anlamlılığına dair yanlış tahminler elde edilebileceğini belirtmişlerdir. Temel bileşenler regresyonu (TBR), çoklu bağlantı ve çok boyutluluk (değişken sayısının gözlem sayısından büyük olması) problemlerine çözüm sunmaktadır. Bursa [14], TBR'nin asıl amacının, çoklu bağlantıya sebep olan boyutları yok etmek olduğunu belirtmiştir. Bu durum, küçük özdeğerlere sahip boyutların modelden çıkartılmasıyla sağlanır.

TBR'de ilk önce, bağımsız değişkenlere temel bileşenler analizi (TBA) uygulanır. TBA

sonucunda, bağımsız değişkenlerin boyutu indirgenir ve ortaya çıkan bileşenler birbirinden bağımsız değişkenler olur. Daha sonra, bu bağımsız değişkenler ile regresyon modeli kurulur.

2.4. Kısmi en küçük kareler regresyonu

Kısmi en küçük kareler regresyonu (KEKKR), bağımsız değişkenlerin ilişkili olması durumunda, model tahmininde yeni bir yöntem olarak Wold [15] tarafından geliştirilmiştir. Değişkenlerin daha az sayıda ve aralarında çoklu bağlantı problemi olmayan bileşenlere indirgenip regresyon modeli kurulmasıdır. Bulut [16], KEKKR'yi TBR'den ayıran temel farkın; TBR'nin bileşen seçiminde sadece bağımsız değişkenleri kullanırken, KEKKR'nin, bağımsız değişkenlerin yanı sıra bağımlı değişkeni de kullanması olduğunu belirtmiştir.

Bulut [16], KEKKR modelinde seçilecek ideal bileşen sayısının, genellikle en küçük hata kareler ortalamasının karekökü değerini veren bileşen sayısı olarak ya da çapraz doğrulama süreciyle kestirilen tahmin hatasını minimize ederek seçildiğini ifade etmiştir.

2.5. Ridge regresyon

Çoklu bağlantı problemi, regresyon katsayılarının tahminlerinin gerçek değerlerden uzaklaşmasına sebep olur. Hoerl ve Kennard [17], bu probleme çözüm olarak ridge regresyonu geliştirmişler ve korelasyon matrisinin köşegen elemanlarına pozitif bir sayı ekleyerek tahmin varyanslarının küçüldüğünü göstermişlerdir.

Ridge regresyonda amaç, hata kareler toplamını minimize eden katsayıları bulmaktır.

$$HKT = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{P}\beta_j x_{ij}\right)^2 \qquad (3)$$

$$SSE_{L_2} = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2 + \lambda\sum_{j=1}^{P}\beta_j^2 \qquad (4)$$

Her bir katsayının hata kareler toplamındaki rolüne göre λ, β katsayılarına ceza dağıtır. $\lambda \geq 0$ ceza parametresidir. Gupta [18], λ değerinin, modelin esnekliğinin ne kadar cezalandırılacağı ile ilgili olduğunu belirtmiştir. Model esnekliği, modelin fonksiyonel yapısının ne kadar iyi temsil edildiğini gösterir. Gupta [18] yazısında, λ değeri büyüdükçe, verideki önemli özelliklerin kaybedilmediğini ve varyans değerinin azaldığını, ancak, belirli bir değerden sonra modelin önemli özelliklerini kaybetmeye başladığını ifade etmiştir. Bu yüzden λ değerinin seçimi önemlidir. Bu seçimi yaparken k-katlı çapraz doğrulama yöntemi kullanılır. Öncelikle, λ için belirli değerleri içeren bir küme seçilir. Daha sonra her bir λ değeri için çapraz doğrulama test hatası hesaplanır. En küçük hata değerini veren λ değeri seçilerek nihai model elde edilir.

Ridge regresyonda, tüm değişkenler ile model kurulur, ilgisiz değişkenler modelden çıkarılmaz, katsayıları sıfıra yaklaştırılır.

2.6. Lasso regresyon

Lasso regresyon, Tibshirani [19] tarafından, doğrusal regresyon modellerinde karşılaşılan çoklu bağlantı sorununa çözüm sunmak ve ridge regresyonun, tüm değişkenleri modelde bırakma dezavantajını gidermek amacıyla önerilmiştir. Lasso regresyonda amaç hata kareler toplamını minimize eden katsayıları bulmaktır.

$$HKT = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{P}\beta_j x_{ij}\right)^2 \qquad (5)$$

$$SSE_{L_1} = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2 + \lambda\sum_{j=1}^{P}\left|\beta_J\right| \qquad (6)$$

Ridge regresyondan farkı, otomatik değişken seçimi yapmasıdır. Lasso regresyon katsayıları sıfıra yaklaştırır. Ancak λ yeteri kadar büyük olduğunda anlamsız katsayıları sıfır yapar. Böylece değişken seçimi yapılmış olur. λ değeri seçimi, k-katlı çapraz doğrulama yöntemi kullanılarak ridge regresyonda belirtildiği gibi yapılır.

Frank ve Friedman [20], bazı durumlarda lasso regresyonun bazı durumlarda da ridge regresyonun daha iyi tahmin sonuçları verdiğini belirtmişlerdir. Bu yüzden, ridge ve lasso regresyonun net olarak birbirlerine üstünlüğü söz konusu değildir.

2.7. Elastik net regresyon

Elastik net regresyon, ridge ve lasso regreyon yöntemlerinden yola çıkılarak, Zou ve Hastie [21] tarafından geliştirilmiştir.

$$SSE_{Enet} = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2 + \lambda_1\sum_{i=1}^{p}\beta_j^2 + \lambda_2\sum_{i=1}^{p}\left|\beta_j\right| \qquad (7)$$

Ridge regresyonla aynı şekilde cezalandırma yapar. Her bir katsayının hata kareler

toplamındaki rolüne göre λ, β katsayılarına ceza dağıtır**.** Değişken seçimini ise Lasso regresyon yöntemindeki gibi yapar. Anlamsız değişkenlerin katsayılarını sıfır yapar, böylece otomatik değişken seçimi yapılmış olur.

λ_1, λ_2 değerlerinin seçimi, k-katlı çapraz doğrulama yöntemi ile yapılır. λ_1, λ_2 için, belirli değerleri içeren bir küme seçilir ve her bir değerin çapraz doğrulama test hataları hesaplanır. En düşük hata değerini veren λ_1, λ_2 değerleri seçilerek elastik net regresyon modeli oluşturulur.

2.8. K-katlı çapraz doğrulama yöntemi

K-katlı çapraz doğrulama, Bengio ve Grandvalet [22] tarafından, mevcut verileri eğitim ve test seti olarak kullanan bir bilgisayar tekniği olarak açıklanmıştır. Bu yöntemde, veri seti eğitim ve test seti olarak ikiye ayrılır. Eğitim seti kendi içinde parçalara ayrılır. Ayrılan bu parçalardan birisi kenara ayrılır, kalan parçalarla model kurulur. Kurulan model, kenara ayrılan parça ile test edilir. Bu işlem bütün parçalara uygulanır. Yani eğitim setinin içinde ayrı bir doğrulama (validation) işlemi gerçekleştirilir. Doğrulanan modelin test hatası elde edilmiş olur. Elde edilen hata eğitim hatası ya da doğrulama hatası olarak isimlendirilebilir. Bütün parçalara bu işlem uygulandıktan sonra eğitim setinde bir model çıkar. Bu modelin test seti ile doğrulanması yapılır. Yani; test setinin bağımsız değişkenleri kullanılarak bağımlı değişken tahmin edilir.

2.9 Model performans değerlendirme ölçütü

Rastgele oran ve SKÖ yöntemi kullanılarak bölünen veri seti ile regresyon modelleri oluşturulmuştur. Hangi veri seti bölme yönteminin daha iyi performans gösterdiğini belirlemek için başarı ölçütü olarak HKOK kullanılmıştır.

Hata kareler ortalamasının karekökü:

HKOK, Chai ve Draxler [23] tarafından, model performansını ölçmek için kullanılan istatistiksel bir ölçü olarak açıklanmıştır. Ölçüm değerleri ile model tahminleri arasındaki hata oranını belirlemek için kullanılır. Tahmin hatalarının standart sapmasıdır. Elde edilen modeller içinden hangisinin daha iyi performans gösterdiğine karar vermek için modellerin HKOK değerleri karşılaştırılır. Bir modelin HKOK değerinin sıfıra yaklaşması o modelin iyi bir model olduğunu gösterir.

$$HKOK = \sqrt{\frac{\sum_{j=1}^{n} e_j^2}{n}} \qquad (8)$$

3. Bulgular

Bu çalışmada, Kaliforniya Üniversitesi'nin lisansüstü öğrenci kabul kriterleri göz önünde bulundurularak Hindistan'daki öğrenciler için oluşturulan yüksek lisans başvuru verileri kullanılmıştır. Seçilen veri seti, 400 gözlem ve 8 değişkenden oluşmaktadır. Bağımlı değişken öğrencilerin yüksek lisansa kabul olasılıklarıdır. Bağımsız değişkenler ise; GRE puanı, TOEFL puanı, üniversite derecesi, amaç mektubu, tavsiye mektubu, lisans not ortalaması, araştırma tecrübesidir. Regresyon analizleri yapılmadan önce, bağımsız değişkenler arasındaki ilişkiyi (korelasyon) gözlemlemek için korelasyon matrisi incelenmiştir. Bağımsız değişkenler arasında yüksek ilişki olduğu tespit edilmiştir. Bu durumda, veri setinde çoklu bağlantı problemi olduğu söylenebilir. Dolayısıyla, materyal ve metot bölümünde açıklanan regresyon yöntemlerinin uygulanması için uygun bir veri setidir. Veri seti bölme işlemi hem rastgele hem de SKÖ yöntemleri kullanılarak yapılmıştır. İki yöntemden elde edilen eğitim setleri ile doğrusal regresyon modelleri oluşturulmuştur. Daha sonra, oluşturulan modeller test setleri ile test edilerek HKOK değerleri elde edilmiştir. Elde edilen HKOK değerleri karşılaştırılarak, hangi veri seti bölme yönteminin daha iyi sonuç verdiği gözlemlenmiştir. Rastgele oran yöntemi ile veri setini bölerken birden fazla oran kullanılmıştır. SKÖ ile veri setini bölerken ise, rastgele oranlara karşılık gelen gözlem sayılarını sağlamak için, farklı m ve r değerleri kullanılmıştır. Her iki yöntem ile yapılan işlemler 5000 kez tekrarlanmıştır. Her tekrarda seçilen gözlemler ile doğrusal regresyon modelleri oluşturulmuştur. Elde edilen regresyon modellerinin HKOK değerleri hesaplanmıştır. Veri seti bölme yöntemleri karşılaştırılırken, 5000 tekrardan elde edilen HKOK değerlerinin ortalaması kullanılmıştır.

Bu çalışmada, doğrusal regresyon analizleri ve veri seti bölme işlemleri, RStudio programı kullanılarak yapılmıştır.

3.1. Rastgele veri seti bölme işlemi

Rastgele veri seti bölme işlemini yaparken, bağımlı değişken üzerinden indeks üretilmiştir. Bu indekse göre veri seti; %76-%24, %79-%21 ve %80-%20 oranlarında, eğitim ve test seti olacak şekilde ayrılmıştır. Her farklı oran için yapılan işlemler, 5000 kez tekrarlanmıştır. Her tekrarda elde edilen eğitim setleri ile, doğrusal regresyon modelleri oluşturulmuş, test setleri için HKOK değerleri elde edilmiştir. Her bir doğrusal regresyon modeli için elde edilen HKOK değerlerinin ortalamaları alınmıştır. Bu değerler, Tablo 2'de verilmiştir.

3.2. Sıralı küme örneklemesi ile veri seti bölme işlemi

SKÖ ile veri seti bölme işlemi gerçekleştirilirken, rastgele oranda veri seti bölme yönteminde kullanılan oranlar esas alınarak, bu oranlara karşılık gelen gözlem sayılarını elde etmek için kullanılan m ve r değerleri Tablo 1'de verilmiştir.

Tablo 1. Eğitim setindeki farklı rastgele oranlara karşılık gelen gözlem sayıları ve SKÖ'de kullanılan küme (m), tekrar sayıları (r)

Eğitim Seti	m (Küme Sayısı)	r (Tekrar Sayısı)
305 (%76)	5	61
318 (%79)	3	106
322 (%80)	2	161

Her farklı m ve r değeri için yapılan işlemler, 5000 kez tekrarlanmıştır. Her tekrarda seçilen eğitim setleri ile oluşturulan doğrusal regresyon modelleri, test setleri ile test edilerek, HKOK değerleri elde edilmiştir. Her bir doğrusal regresyon modeli için elde edilen 5000 adet HKOK değerinin ortalamaları alınmıştır. Bu değerler, Tablo 2'de yer almaktadır.

Tablo 2. Rastgele oran yöntemi ve SKÖ ile bölünen veri setleri kullanılarak oluşturulan doğrusal regresyon modellerinin, test setleri ile elde edilen HKOK değerleri

Veri Seti Bölme Yöntemi	Çoklu Doğrusal Regresyon	Temel Bileşen Regresyonu	Kısmi En Küçük Kareler Regresyonu	Ridge Regresyon	Lasso Regresyon	Elastik Net Regresyon
Rastgele Oran (%76)	0,06443717	0,06763168	0,06463873	0,06439753	0,06471603	**0,06375078**
SKÖ (m=5, r=61)	**0,06394809**	**0,067567**	**0,06415843**	**0,0643646**	**0,06451742**	0,06497169
Rastgele Oran (%79)	0,06406733	**0,06764338**	**0,06430642**	0,06440878	0,06402546	**0,06413632**
SKÖ (m=3, r=106)	**0,06389694**	0,06779521	0,06455888	**0,06382268**	**0,06386816**	0,06553345
Rastgele Oran (%80)	0,06456588	0,06779723	0,06468039	**0,06436714**	0,06460615	**0,06391357**
SKÖ (m=2, r=161)	**0,06423147**	**0,0676854**	**0,06459359**	0,06436838	**0,0641779**	0,06449959

4. Tartışma ve Sonuç

Bu çalışmada, makine öğrenmesinde modelin öğretilen bilgiyi ne kadar iyi öğrendiğini test edebilmek için, veri seti bölme işlemi rastgele oran ve SKÖ yöntemleri kullanılarak yapılmıştır. Her iki yöntemden elde edilen eğitim setleri ile doğrusal regresyon modelleri oluşturulmuştur. Daha sonra, test setleri kullanılarak modellerin HKOK değerleri üzerinden, veri seti bölme yöntemlerinin performans karşılaştırması yapılmıştır.

Veri seti %76 oranında rastgele olarak bölünmüş ve elde edilen eğitim setleri ile doğrusal regresyon modelleri oluşturulmuştur. Veri setinin %76'sına denk gelen 305 gözlem değeri SKÖ ile seçilirken, m=5, r=61 olarak

alınmıştır. Her iki yöntemle yapılan işlemler 5000 kez tekrarlanmıştır. Her tekrarda seçilen gözlemler ile doğrusal regresyon modelleri oluşturularak, test setleri için HKOK değerleri elde edilmiş ve bu değerlerin ortalaması alınmıştır. Sadece elastik net regresyon modelinde, rastgele oran yöntemi ile daha düşük hata değeri elde edilmiştir. Diğer modellerde ise, SKÖ yönteminin daha iyi sonuç verdiği görülmektedir.

Veri seti %79 oranında bölündüğünde, temel bileşenler, kısmi en küçük kareler ve elastik net regresyon modellerinde, rastgele oran yöntemi ile daha düşük hata değerleri elde edilmiştir. Diğer modellerde ise, SKÖ yönteminin daha iyi sonuç verdiği görülmektedir.

Veri seti %80 oranında bölündüğünde, elastik net ve ridge regresyon modellerinde, rastgele oran yöntemi daha düşük hata değerlerine sahiptir. Diğer modellerde ise, SKÖ yöntemi ile daha iyi sonuçlar elde edilmiştir.

Bu çalışmayla incelenen durumlarda, çoklu doğrusal ve lasso regresyon modellerinde, SKÖ yöntemi ile daha düşük hata değerleri elde edilmiştir. Temel bileşenler, kısmi en küçük kareler ve ridge regresyon modellerinde, sadece tek bir durumda rastgele oran yöntemi ile daha düşük hata değerlerine ulaşılmıştır. Elastic net regresyon modellerinde ise, bütün durumlarda, rastgele oran yöntemi kullanıldığında daha düşük hata değerleri elde edildiği gözlemlenmiştir. Sonuç olarak, makine öğrenmesi ile ilgilenen araştırmacılara, elastic net regresyon modeli hariç, diğer doğrusal regresyon modellerinde, SKÖ yöntemi rastgele oran yöntemi için daha iyi bir alternatif olarak önerilebilir.

Kaynakça

[1] Samuel, A.L. 1959. Some Studies in Machine Learning Using the Game of Checkers, IBM Journal of Research and Development, Cilt. 3, s. 211 . DOI: 10.1147/rd.33.0210

[2] McIntryre, G.A. 1952. A Method of Unbiased Selective Sampling Using Ranked Sets, Australian Journal of Agriculture Research, Cilt. 3, s. 385. DOI: 10.1071/AR9520385

[3] Halls, L.S., Dell, T.R. 1966. Trial of Ranked Set Sampling for Forage Yields, Forest Science, Vol. 12, s. 24. DOI: 10.1093/forestscience/12.1.22

[4] Evans, M.J. 1967. Application of Ranked Set Sampling to Regeneration Surveys in Areas Direct-Seeded to Longleaf Pine, Louisiana State University, School of Foresty and Wildlife Management, Masters Thesis , Baton Rouge.

[5] Takahasi K., Wakimoto K. 1968. On Unbiased Estimates of the Population Mean Based on the Sample Stratified by Means of Ordering, Annals of the Institude of Istatistical Mathematics, Cilt. 21, s. 250. DOI: 10.1007/BF02911622

[6] Dell D.R., Clutter J.L. 1972. Ranked Set Sampling Theory with Order Statistics Background, Biometrics, Cilt. 28, s. 550. DOI: 10.2307/2556166

[7] Martin, W.L., Sharik, T.L., Oderwald R.G., Smith D.W. 1980. Evaluation of Ranked Set Sampling for Estimating Shrub Phytomass in Appalachian Oak Forests, Publication Number FWS-4-80, School of Forestry and Wildlife Resources, Virginia Polytechnic Institute and State University, Blacksburg.

[8] Stokes S.L. 1980. Estimation of Variance Using Judgement Ordered Ranked Set Samples, Biometrics, Cilt. 36, s. 36. DOI: 10.2307/2530493

[9] Ridout M.S., Cobby J.M. 1987. Ranked Set Sampling with Non-random Selection of Sets and Errors in Ranking, Applied Statistics, Cilt. 36, s. 146. DOI: 10.2307/2347546

[10] Stokes S.L., Sager T.W. 1988. Characterization of a Ranked Set Sample with Application to Estimating Distribution Functions, Journal of the American Statistical Association, Cilt. 83, s. 376-377. DOI: 10.2307/2288852

[11] Patil G.P., Sinha A.K., Taillie C. 1994. Ranked Set Sampling, A Handbook of Statistics, Cilt. 12, s. 180. DOI: 10.1016/S0169-7161(05)80007-0

[12] Patil G.P., Sinha A.K., Taillie C. 1997. Ranked Set Sampling, Coherent Rankings and Size-Biased Permutations, Journal of Statistical Planning and Inference, Cilt. 63, s. 311-324. DOI: 10.1016/S0378-3758(97)00030-X

[13] Montgomery, D.C., Peck, E.A., Vining G.G. 2013. Linear Regression Analysis. 5th. John Wiley & Sons. 687s.

[14] Bursa, N. 2019. Bağımsız Bileşenler Analizi ile Çoklu Bağlantı Sorununa Bir Yaklaşım. Hacettepe Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, 124s, Ankara.

[15] Wold, H. 1985. Partial Least Squares, Encyclopedia of Statistical Sciences, Cilt. 6, s. 581-591. DOI: 10.1002/0471667196.ess1914

[16] Bulut, Y. M. 2011. Çoklu İç İlişki Durumunda Kısmi En Küçük Kareler Regresyonu ve Alternatif Yöntemlerle Karşılaştırılması. Eskişehir Osmangazi Üniversitesi, Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi, 92s, Eskişehir.

[17] Hoerl, A. E., Kennard, R. W. 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics, Cilt. 12, s. 55-67. DOI: 10.2307/1271436

[18] Gupta, P. 2017. Regularization in Machine Learning. http://www.towardsdatascience.com/regularization-in-machine-learning-76441ddcf (Erişim Tarihi: 03.04.2021)

[19] Tibshirani, R. 1996. Regression Shrinkage and Selection Via the Lasso, Journal of the Royal Statistical Society: Series B (statistical methodology), Cilt. 58, s. 267-288. DOI: 10.1111/j.2517-6161.1996.tb02080.x

[20] Frank, D. E., Friedman J. H. 1993. A Statistical View of Some Chemometrics Regression Tools,

Technometrics, Cilt. 35, s. 109-135. DOI: 10.2307/1269656

[21] Zou, H., Hastie T. 2005. Regularization and Variable Selecetion Via the Elastic Net, Journal of the Royal Statistical Society: Series B (statistical methodology), Cilt. 67, s. 301-320. DOI: 10.1111/j.1467-9868.2005.00503.x

[22] Bengio, Y., Grandvalet, Y. 2004. No Unbiased Estimator of the Variance of K-fold Cross-Validation, Journal of Machine Learning Research, Cilt. 5, s. 1089-1105.

[23] Chai, T., Draxler, R. R. 2014. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE), Geoscientific Model Development Discussions, Cilt. 7, s. 1525-1534. DOI: 10.5194/gmd-7-1247-2014