Comparison of M5-Prime and Linear Regression Methods in Various Relationship Types between Variables

Hüseyin YILDIZ; Bolu Provincial Directorate of National Education, Measurement and Evaluation Center, https://orcid.org/0000-0003-2387-263X

Alperen YANDI, Bolu Abant İzzet Baysal University, Department of Educational Sciences, https://orcid.org/0000-0002-1612-4249

İnönü University
Journal of the Faculty of Education
Vol 22, No 1, 2021
pp. 744-771
DOI: 10.17679/inuefd.758378

Article type:
Research article

Abstract

In social sciences, the slope coefficient of the relationship between two variables may vary in different value ranges of the independent variable. For example, while the relationship between two variables has a low slope in each range of values, it may be that the slope increases or decreases after a point. In this study, it is aimed to compare the M5-Prime and linear regression methods in cases where the slope coefficient of the relationship between two variables varies in different value ranges of the independent variable. In this direction, data were produced for four different conditions in which the variance slope coefficient of the relationship between variables was diversified. 3000 replications were made for each data set condition. R programming language was used in the production of data sets. The correlation coefficient (r), explained variance ratio (r^2), mean absolute error, RMSE, and relative absolute error values obtained from the analyzes performed by two different methods were examined. All the analyzes were carried out in the RWeka package of the R programming language. According to the analyzes results, M5-Prime and linear regression methods gave equivalent results in the first of four different types of data set conditions. The results show that results of both methods are equivalent in the data set conditions where the slope coefficient of the relationship between the two variables is constant. In the other three types of data set conditions where the slope coefficient varied, it was seen that the M5-Prime method gave more appropriate results. In this context, researchers can be suggested to use the M5-Prime algorithm in cases where the slope coefficients of the relationships between variables vary.

Keywords: *Linear Regression, M5-Prime, the various relationships among the variables, regression trees.*

Received : 26.06.2020
Accepted : 27.04.2021

Suggested Citation

Yıldız, H. & Yandı A. (2021). Comparison of M5-prime and linear regression methods in various relationship types between variables, *Inonu University Journal of the Faculty of Education*, *22*(1), 744-771. DOI: 10.17679/inuefd.758378

EXTENDED ABSTRACT

Introduction

The relationship means the link between at least two variable structures such as events, people, or features. It is very important to determine the relationship between two variables. The reason for this is that there can be an opportunity to treat in two variables, which are measured according to the direction and the slope coefficient of the relationship. Researchers in the fields of social science focus on explaining the relationships among psychological structures quite frequently. Improvements can be made in the social environment and education environments by planning to be made according to the direction and the slope coefficient of the relationships between the variables. The next step of the relationship examinations is trying to predict the variables. This process is the prediction of the others through one or more of the psychological structures that are related to each other. It is used with regression models designed by researchers in the prediction process. Regression models differ according to the form of the relationship between the variables to be examined. Linear regression models are preferred when the relationship between independent and dependent variables included in the study is linear, and non-linear regression models are preferred if not linear. It is seen that linear regression analysis is frequently preferred in predictive studies on psychological structures in the social sciences (Altan & Eldeklioğlu, 2019; Bostancı & Tosun, 2019; Doğan & Şirin, 2019). The relationship between the variables included in the linear regression analysis must be able to settle on a line. The differences in the direction and the slope coefficient of the relationship between the variables included in the analysis may lead to error in the explained variance ratio calculated for between variables. One of the methods suggested as an alternative to linear regression models is the M5-Prime algorithm, which is performed based on the decision trees, which are classification models, and linear regression models. The M5-Prime algorithm is an improved form of the M5 algorithm developed by Quinlan (1992). The M5-Prime algorithm performs as a combination of

decision trees and simple linear regression analysis. M5-Prime algorithm is a very powerful technique for simple linear regression method and other non-classification methods (neural networks etc.) in predicting the relationships between variables and explaining variance rates (Diaz et al., 2014). In this study, it is aimed to compare the results obtained with M5-Prime algorithm and simple linear regression method. As a result of the study, to facilitate the use of the M5-Prime algorithm in different studies, the researchers were provided with information about the parameters obtained with the M5-Prime algorithm. Thus, it is aimed to spread the use of this algorithm in predictive studies in the field of social sciences and to inform the researchers about the usage process.

Purpose

The aim of this study is to compare the M5-Prime and linear regression methods in various types of relationships between variables that may be encountered in social sciences. In addition, the research can be evaluated in the basic research class as it aims to compare two different theoretical methods. Since the direction and the slope coefficient of the relationships between the variables may vary, it is important to compare the methods made to explain the relationships more accurately. This research has been carried out for this situation. In addition, this research involves the introduction of a method that is uncommon and complex. Therefore, it can be claimed that it is important because it is thought to provide information to other researchers.

Method

Since the data sets used in the research is produced in computer environment, it is a simulation study. Simulation studies can be described as an experimental, correlational, or descriptive model (Erkuş, 2013). In this research two different method were performed on the data sets and the parameters values were examined for comparing the differences between these two methods. From this aspect, it can be claimed that research is a descriptive study.

The data sets used in this study were produced with the codes written in the R programming language. A total of 12000 data sets were produced. 3000 replications were made for each of the four different data set conditions, where the direction and the slope coefficient of the relationship between variables varied. Analysis of linear regression and M5-Prime methods were performed using the RWeka package of the R programming language. Correlation coefficients, explained variance ratio, mean absolute error, RMSE and relative absolute error parameters were used to evaluate the success of the models created by linear regression and M5-Prime methods.

Findings

Within the scope of the research, the correlation coefficient, explained variance ratio, mean absolute error, RMSE, relative absolute error parameters were calculated to evaluate the results of linear regression and M5-Prime analysis performed with four different data set conditions. Accordingly, the R values obtained with the M5-Prime method for four data set condition are 0.948, 0.984, 0.960 and 0.802, respectively. R values obtained by linear regression method were calculated as 0.948, 0.927, 0.843 and 0.262, respectively. R2 values obtained using the M5-Prime method are 0.900, 0.968, 0.923, 0.643, and R2 values calculated with the linear regression method are 0.900, 0.860, 0.712, 0.073, respectively. In addition, the mean absolute error values obtained with the M5-Prime method are 12.014, 10.063, 12.871, 11.862, and the mean absolute error values were reached because of the linear regression methods are 12.023, 20.682, 23.088, and 18,327. As a result of M5-Prime analysis, RMSE values of 13.988, 11.849, 15.359 and 14.105 are reached for four data set condition, while RMSE values obtained because of linear regression analysis are 13.998, 24.858, 28.953 and 22.150. Finally, the relative absolute error values resulting from M5-Prime analysis are 31.55%, 17.27%, 31.97% and 62.20%, while the relative absolute error values resulting from linear regression analysis are 31.57%, 35.48%, 57.24% and 95.96%.

Discussion & Conclusion

The studies comparing the regression trees algorithms, which are the members of the M5-Prime algorithm and linear regression models in the literature are examined. Considering the data sets features used in this study and current studies, it is seen that the results obtained in this study are consistent with the results of the study in the literature (Gonzalez-Sanchez, Frausto-Solis, & Ojeda-Bustamante, 2014; King, Rice & Vaughan, 2018; Shafiullah, Simson, Thompson, Wolfs & Ali, 2008). In this study, it was determined that the models put forward with the M5-Prime algorithm in three of the four data sets representing four different situations were more successful in explaining the relationship between variables. It has been observed that M5-Prime algorithm gives better results than linear regression analysis in terms of correlation coefficient, explained variance rate, average absolute error, RMSE and relative absolute error parameters. For the data set in which the relationship between the variables did not vary the direction and the slope coefficient at different levels of the independent variable, it was observed that the same results obtained with both methods. According to results, it can be suggested to use the M5-Prime algorithm in studies aiming to explain these relationships considering the situation that the relationships between variables in the social sciences may vary.

Değişkenler Arası Farklı İlişki Tiplerinde M5-Prime ve Doğrusal Regresyon Yöntemlerinin Karşılaştırması

Hüseyin YILDIZ; Bolu İl Milli Eğitim Müdürlüğü, Ölçme ve Değerlendirme Merkezi, https://orcid.org/0000-0003-2387-263X

Alperen YANDI, Bolu Abant İzzet Baysal Üniversitesi, Eğitim Bilimleri Bölümü, https://orcid.org/0000-0002-1612-4249

Öz

Sosyal bilimlerde iki değişken arasındaki ilişkinin eğim katsayısı, bağımsız değişkenin farklı değer aralıklarında değişiklik gösterebilmektedir. Örneğin, iki değişken arasındaki ilişki, belirli bir değer aralığında düşük bir eğime sahip seyrederken, bir noktadan sonra eğimin artma veya azalma göstermesi durumu söz konusu olabilir. Bu çalışmada iki değişken arasındaki ilişkiye ait eğim katsayısının, bağımsız değişkenin farklı değer aralıklarında değişiklik gösterdiği durumlarda, M5-Prime ve doğrusal regresyon yöntemlerinin karşılaştırılması amaçlanmıştır. Bu doğrultuda değişkenler arası ilişkiye ait eğim katsayı değişimlerinin çeşitlendirildiği dört farklı veri seti koşulu altında veriler üretilmiştir. Her bir veri seti koşulu için 3000 replikasyon yapılmıştır. Veri setlerinin üretiminde R programlama dili kullanılmıştır. İki farklı yöntemle yapılan analizlerde elde edilen korelasyon katsayısı (R), açıklanan varyans oranı (R^2), ortalama mutlak hata, RMSE, göreceli mutlak hata değerleri incelenmiştir. Analizlerin tümü R programlama dilinin RWeka paketinde gerçekleştirilmiştir. Analiz sonuçlarına göre dört farklı tipteki veri seti koşulundan ilkinde M5-Prime ve doğrusal regresyon yöntemleri eş değer sonuçlar vermiştir. Elde edilen sonuçlar, iki değişken arasındaki ilişkiye ait eğim katsayısının sabit seyrettiği veri seti koşulunda her iki yöntemin eş değer sonuçlar verdiğini göstermektedir. Eğim katsayısının farklılaştığı diğer üç tip veri seti koşulunda ise M5-Prime yönteminin daha uygun sonuçlar verdiği görülmüştür. Bu bağlamda araştırmacılara, değişkenler arası ilişkilere ait eğim katsayılarının farklılaştığı durumlarda M5-Prime algoritmasını kullanması önerilebilir.

Anahtar Kelimeler: *Doğrusal Regresyon, M5-Prime, değişkenler arası farklı tipteki ilişkiler, regresyon ağaçları.*

İnönü Üniversitesi
Eğitim Fakültesi Dergisi
Cilt 22, Sayı 1, 2021
ss. 744-771
DOI: 10.17679/inuefd.758378

Makale türü:
Araştırma makalesi

Gönderim Tarihi : 26.06.2020
Kabul Tarihi : 27.04.2021

Önerilen Atıf

Yıldız, H. & Yandı A. (2021). Değişkenler arası farklı ilişki tiplerinde M5-prime ve doğrusal regresyon yöntemlerinin karşılaştırılması, *İnönü Üniversitesi Eğitim Fakültesi Dergisi*, *22*(1), 744-771. DOI: 10.17679/inuefd.758378

Değişkenler Arası Farklı İlişki Tiplerinde M5-Prime ve Doğrusal Regresyon Yöntemlerinin Karşılaştırması

İlişki, en az iki durum, olay, kişi veya özellik gibi değişken yapılar arasındaki bağlantı anlamı taşımaktadır. İki değişken arasındaki ilişkinin belirlenmesi oldukça önemlidir. Bunun nedeni, değişkenler arası ilişkilere göre ölçülen değişkenlere farklı şekilde müdahale olanağı ortaya çıkabilecek olmasıdır. Bir değişken üzerinden diğer bir değişkendeki değişimi açıklama olanağı ilişkilerin tanımlamaları üzerinden kurulan matematiksel modellerle ile mümkün olmaktadır (Chatterjee & Hadi, 2015; Sykes, 1993).

Sosyal bilimlerde, araştırmacılar psikolojik yapılar arası ilişkilerin açıklanması üzerinde oldukça sık bir şekilde durmaktadır. Farklı yapıdaki değişkenler arasında ilişkinin incelenmesi, bu değişkenlere ait durumların düzenlenebilmesine zemin hazırlamaktadır. Değişkenler arasındaki ilişkilere göre yapılacak planlamalarla sosyal çevre ve eğitim ortamlarında iyileştirmeler gerçekleştirilebilir.

İlişki incelemelerinin bir sonraki adımı ise değişkenlere ilişkin öngörüde bulunmaya çalışmaktadır. Bu işlem ilişkili olan psikolojik özelliklerin, biri veya birkaçı aracılığı ile diğerlerinin tahmin edilmesidir. Tahmin etme sürecinde araştırmacılar tarafından kurgulanan regresyon modelleri ile kullanılmaktadır (Alpar, 2017; Civelek & Durukan, 2012). Regresyon modelleri bir veya birkaç değişkene ilişkin elde edilen verilerden yararlanılarak gelecekteki durumlara ilişkin birey veya grup bazında tahminler elde edilmesi amacındadır. Regresyon modellerinde değişkenler açıklayıcı (yordayıcı) ve açıklanan (yordanan) rolleri üstlenmektedir. Açıklayıcı roldeki değişkenler, açıklanan roldeki değişkendeki değişimin açıklanması için kullanılmaktadır. Bu değişkenlerin sırasıyla bağımsız ve bağımlı değişken olarak nitelendirilme durumu da söz konusudur (Field, 2009).

Regresyon modelleri, incelenecek olan değişkenlerin arasındaki ilişkinin biçimine göre farklılaşmaktadır. İncelemeye dâhil edilen bağımsız ve bağımlı değişkenler arasındaki ilişkinin doğrusal olması halinde doğrusal (linear) regresyon modelleri, doğrusal olmaması halinde ise doğrusal olmayan (non-linear) regresyon modelleri tercih edilmektedir.

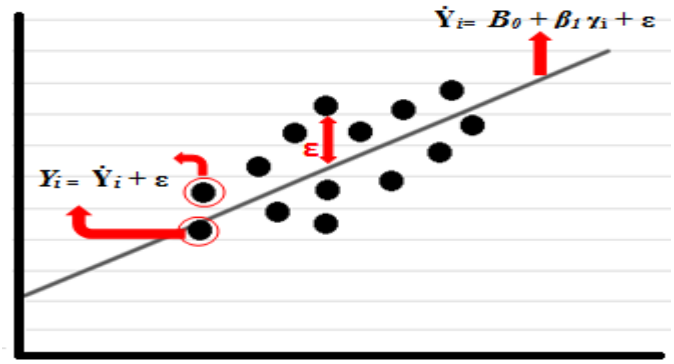
Sosyal bilimler alanında psikolojik yapılara ilişkin gerçekleştirilen yordama incelemelerinde de doğrusal regresyon analizinin sıklıkla tercih edildiği görülmektedir (Altan & Eldeklioğlu, 2019; Doğan & Şirin, 2019; Bostancı & Tosun, 2019). Buna ek olarak Google arama motoru üzerinde son beş yılda, bilimsel içerik olması kaydıyla, doğrusal regresyon analizine ilişkin arama sayısı 1209 olarak rapor edilirken, doğrusal olmayan regresyona ilişkin arama sayısı ise kayda değer olmadığı şeklinde rapor edilmiştir. Doğrusal ve doğrusal olmayan regresyon analizine ilişkin bu bulgular araştırmacıların sıklıkla doğrusal regresyon modellerini kullandığına işaret etmektedir. Bu durum önemli bir soruyu beraberinde getirmektedir: Değişkenler sürekli yapıda olduğunda basit doğrusal regresyon analizine dahil edilen değişkenler arasındaki ilişki, bağımsız değişkenin tüm değerler aralıkları için doğrusal mıdır? Bir başka ifadeyle analize bağımsız ve bağımlı pozisyonunda dahil edilen değişkenlerin birlikte değişimleri, değişkenlerin her değer aralığında aynı eğime mi sahiptir? Bu soruların cevabının net bir biçimde verilmemesi durumunda doğrusal regresyon analizinden elde edilen sonuçların doğruluğu tartışmalı hale gelecektir. Nitekim alanyazın incelendiğinde değişkenler arası doğrusal olmayan ilişkilerin tespit edildiğine yönelik çalışmalar olduğu görülmektedir (Harma, 2008; Dumludağ, Gökdemir & Giray, 2016; Güvercin, 2018).

Doğrusal regresyon, analize dahil edilen değişkenler arasındaki ilişkinin, değişkenlere ilişkin her değer aralığında doğrusal olduğu varsaymaktadır. Basit doğrusal regresyon analizi denklemi eşitlik 1'deki gibidir.

Eşitlik 1.$\dot{Y}i = B0 + \beta 1\ \chi i + \varepsilon i$

Burada β0 ve β1 regresyon modelinin bilinmeyenleridir. Ẏi parametresi doğrusal model tarafından yordanan Y değeridir. Xi parametresi i. birey için yordayıcı değişken değeridir. Denklemde yer alan β0 regresyon doğrusunun y eksenini kestiği noktayı gösterir ve kesim noktası veya sabit gibi adlandırılır. Eğer değişkeninin dağılım aralığı sıfırı içeriyorsa, β0 katsayısı 0'a eşit olduğunda y'nin alacağı ortalama değeri verir (Chatterjee & Simonoff, 2013). β1 ise regresyon katsayısıdır. Bağımsız değişkende bir birimlik değişme olduğunda, bağımlı

değişkende meydana gelecek ortalama değişim miktarını gösterir (Leech, Barrett & Morgan, 2005). ɛi parametresi ise terimine karşılık gelir. Hata terimi gerçek Yi değeri ile regresyon denklemi ile hesaplanan Ẏi değeri arasındaki farka eşittir. Bu bilinmeyenlere parametre ya da genel olarak regresyon katsayıları denir (Tabachnick & Fidell, 2007). Eşitlik 1'deki bu denklem için çizilen grafik ise Şekil 1'de sunulmuştur.



Şekil 1 *Doğrusal Regresyon Model Oluşturma Süreci*

Korelasyon temelli bir analiz olan doğrusal regresyon analizinde sürekli yapıda olan bir veya daha fazla değişkenin, bir bağımlı değişkendeki değişimi açıklayabilme oranına dair hesaplamalar yapılmaktadır. Doğrusal regresyon analizi, bağımsız ve bağımlı değişkenler arasındaki en iyi açıklayan doğruyu bulmayı amaçlamaktadır (Alpar, 2017). Değişkenler arası ilişkileri açıklamak için yönü ve eğimi sabit olan doğruları kullanması, doğrusal regresyon analizi modellerinin en önemli sınırlılığı olarak ileri sürülebilir. Analize dahil edilen değişkenler arasındaki ilişkinin yön ve eğim katsayısındaki farklılaşmalar, bağımsız değişken(ler)in bağımlı değişkendeki değişimi açıklanma oranın olduğundan farklı hesaplanmasına yol açabilir.

Korelasyon temelli analizlerde değişkenler arası ilişkinin doğrusal olmaması durumunda (üstel, polinominal, lojistik vb.) değişkenler arasındaki bağlantının mükemmel şekilde açıklanması mümkün olmamaktadır (Onwuegbuzie & Daniel 2002). Bu durumun yarattığı sınırlılıklar farklı çalışmalarda değerlendirilmiştir (Anscombe 1973; Eugene & Johnston 1996;

Rodgers & Nicewander 1988; Lee 1992). Doğrusal olmayan ilişkilerin açıklanmasında kullanılabilecek alternatif modeller üzerinde çalışmalar gerçekleştirilmiştir (Breiman & Friedman, 1985; Cox, 1981; Nguyen ve diğ., 2014; Wang & Witten, 1996). Doğrusal modellere sunulan alternatif yöntemlerle, bu modellere ilişkin sınırlılıkların ortadan kaldırılarak, sürekli değişkenler arasındaki ilişkilerin daha doğru bir biçimde tanımlanması hedeflenmektedir.

M5 karar ağacı modeli, sürekli değişkenler için tahminde bulunabilme imkanı tanıyan, tahmin sürecinde doğrusal regresyon işlevlerini kullanan bir karar ağacı algoritmasıdır (Kisi, Shiri, & Demir, 2017). M5-Prime algoritması ise, Quinlan (1992) tarafından geliştirilmiş olan M5 algoritmasının geliştirilmiş bir formudur (Behnood ve diğ., 2017). Wang ve Witten (1996) tarafından geliştirilen bu algoritmada, M5 algoritmasının sayısal veri tiplerine uyarlaması, kayıp verilerle baş edebilir hale gelmesi ve Breiman ve diğerleri (1984) tarafından ortaya konulan tekniklerin birleştirilmesi söz konusudur. Sınıflama ve Karar Ağaçları (Classification And Regression Trees (CART) olarak bilinen bu teknikler, büyük veri setlerinin analizinde, bir başka ifadeyle veri madenciliği sürecinde kullanılan yenilikçi teknikler olarak nitelendirilmektedir. Bireylerden alınan yanıt türüne göre isimlendirilen bu teknikler, yanıtlar sayısal olduğunda regresyon ağacı olarak; yanıtlar kategorik olduğunda ise sınıflandırma ağacı olarak ele alınır (Siciliano & Mola, 2000).

Sınıflandırma algoritmaları içerisinde en sık kullanılan algoritmalardan birisi akış şemasına benzer bir yapıya sahip, aşağıya doğru dallanan (branch) karar ağaçlarıdır (Gupta, Malviya & Singh, 2012; Silahtaroğlu, 2016). Karar ağacı algoritmaları genel olarak, elde edilen veri seti bir bütün olarak düşünüldüğünde, bu bütünü oluşturan kendi içinde homojen küçük parçalar tespit edilmesi işlemi olarak düşünülebilir (Behnood ve diğ., 2017). Belirli kavramların bilinmesi karar ağacı algoritmalarının anlaşılmasını kolaylaştıracaktır. Bu kavramlardan ilki düğümdür (node). Düğüm kavramı, ağaca benzer dallanan bir yapıda alt dalları olan noktalardır (Han, Kamber & Pei, 2012; Quinlan, 1986). Düğüm noktalarında, hesaplanan kesme değerlerine göre noktanın alt bölümünde dallanma söz konusudur. Karar ağaçlarında üç tür düğüm

bulunabilir. Bunlardan ilki sınıflandırmaya başlanan ve ağacın en üst noktasında yer alan düğüme, kök düğümü (root node) adı verilmektedir (Han, Kamber & Pei, 2012). İkinci düğüm türü ise dahili düğümlerdir (internal node). Şans düğümleri olarak da adlandırılan bu düğüm türü bulunduğu noktada mevcut olan olası seçeneklerden birini temsil etmektedir (Song & Lu, 2015). Son düğüm türü ise uç düğüm veya yaprak düğüm (leaf node) olarak isimlendirilmektedir. Sınıfları ve sınıf dağılımlarını temsil eden (Gupta, Malviya & Singh, 2012) bu düğümler, ağaç yapısı içerisinde altında dallanma olmayan, noktalardır. Yapraklar ağaçlar için sınıflandırmanın sonlandığı noktalar –sınıflar- olarak nitelendirilebilir (Song & Lu, 2015). Karar ağacı oluşturma sürecinde değinilebilecek kavramlardan sonuncusu budama işlemidir. Karar ağacı oluşum sürecinde veri içerisindeki uç değerlerin, oluşturulan ağaç dallarında görünmesi durumu olabilir. Bir başka ifadeyle oluşturulan karar ağacı, fazladan, gereksiz görülebilecek dallara ayrılabilir. Budama işlemi, oluşturulan bir karar ağacında, ortaya çıkan sınıflandırmaya bir katkısı olmayan ve sonucu etkilemeyen, bu tür dalların ağaçtan çıkarılması işlemidir (Han, Kamber & Pei, 2012; Song & Lu, 2015). Karar ağaçları oluşturulurken analize dahil edilen veri setinin bir kısmı ağacın oluşturulmasında, bir kısmı ise ağacın amaçlanan şekilde çalışıp çalışmadığının kontrol edilmesi için kullanılır. Ağacın oluşturulması sürecinde ilk olarak kök oluşturulur. Ardından tüm veriler sırayla ağaca uygulanarak uygun olan (üyeleri ile benzerlik gösterdiği) yaprağa (sınıfa) yerleştirilir (Silahtaroğlu, 2016). Karar ağaçlarının kullanılmasında farklı algoritmalar mevcuttur. Bu algoritmaların göz önüne aldığı parametreler farklılaştığı için farklı yapıda ağaçlar üretilmesi söz konusu olabilir. Karar ağacı oluşturma algoritmalarından biri de M5-Prime tekniğinde de kullanılan Quinlan (1993) tarafından geliştirilen C4.5 algoritmasıdır.

ID3 algoritmasını gelişmiş bir versiyonu olan C4.5 algoritması (Weka programında J48 olarak isimlendirilir.) sınıflama sürecinde belirli kavramlardan yararlanır. Sayısal verilerde kullanılabilme, kayıp verileri analize dahil etmeme ve kök, düğüm ve yaprak belirleme süreçleri ile ID3'ten farklılaşan C4.5 algoritması, sınıflama sürecinde ise ID3'te olduğu gibi entropi

kavramından yararlanır (Lestari, 2020; Xiaoliang, ve diğ., 2009). Entropi bir veri grubundaki düzensizlik veya farklılaşma olarak tanımlanabilir (Dunham, 2003). Bir veri grubunun entropisi 0-1 aralığında değer alır. Entropinin 0 olması veri grubunda hiç düzensizlik olmadığı anlamına gelir. Örneğin cinsiyetlerin tahmin edildiği bir veri grubunda herkesin kadın olması entropinin 0 olarak hesaplanmasına yol açar. Entropinin 1'e eşit olması veri grubundaki her bir bireyin tahmin edilmek istenen değişken açısından ayrı kategorilere ait olması durumu olarak düşünülebilir. Entropi değerini hesaplamak için aşağıdaki formül kullanılır.

Eşitlik 2. $H(S) = \sum_{c \in C} - p(c) \log_2 p(c)$

H(S)= S veri setine ait entropi değerini,

C = Tahmin edilecek değişkenin kategorileri,

p(c) = S veri setinde C kategorisine ait veri oranını temsil eder.

C4.5 algoritması ile ağaçlandırma sürecinde analize dahil edilen değişkenlerden her biri için hesaplanan entropi değerleri kullanılarak, karar ağacındaki kök düğümler, dahili ve yaprak düğümler belirlenir. Bu işlem sürecinde karar ağacı algoritmalarının temel amacı her zaman korunarak ilerlenir. Karar ağaçların temel amacı, veri setini oluşturan kendi içinde homojen küçük sınıflar oluşturmaktadır (Behnood ve diğ., 2017). Bu bağlamda düşünüldüğünde bir karar ağacında yukarıda belirtilen etmenler belirlenirken, kökten yapraklara ilerlerken entropinin azalması esas alınır. Bir başka ifadeyle algoritma ağaç dallanmaya başladıkça kendi içinde homojen küçük sınıflar oluşturmayı amaçlar. Kök noktasının ve düğüm noktalarının belirlenmesindeki bir diğer önemli parametreler ise entropi farkları üzerinden hesaplanan bilgi kazanım düzeyi ve bilgi kazanım oranıdır.

C4.5 algoritmasında karar ağacının oluşturulmasında kullanılan parametre bilgi kazanımı oranıdır (information gain ratio). ID3'te bilgi kazanım kavramı olarak karşımıza çıkan bu parametre, ID3'ün çok sayıda benzersiz ölçüme sahip değişkenlerde sınırlı işlem olması nedeniyle, C4.5 algoritmasında bilgi kazanım oranı olarak karşımıza çıkmaktadır. ID3 çok sayıda benzersiz ölçüm içeren veri setlerinde kullanılması durumunda belirlenen entropi değeri düşük

çıkacak ve bu nedenle bilgi kazanımın kullanılması, karar ağacında çok fazla dallanmaya neden olacağında analiz sürecinde kullanışsızlığa yol açacaktır. C4.5 algoritmasında bilgi kazanımı doğrudan kullanılmazken; bilgi kazanım oranının hesaplanmasında sürecine dahil edilir. Bilgi kazanım oranı, bilgi kazanımının, ayırma bilgisine olan oranını ifade etmektedir. Bilgi kazanımı, S veri setinin dallara ayrılmadan önceki ve sonraki durumlardaki entropi miktarları arasındaki fark olarak tanımlanabilir. Bilgi kazanımı değerini hesaplamak için kullanılan formül aşağıda sunulmuştur.

Eşitlik 3. $IG(A,S) = H(S) - \sum_{\text{teT}} - p(t)H(t)$

IG(A,S) = S veri setindeki A değişkenine ait bilgi kazanımı değerini,

t = A değişkeninin her bir alt kategorisini,

p(t) = A değişkenindeki t kategorinde yer alan veri sayısının toplam veri sayısına oranını,

H(t) = t kategorisine ait entropi değeri

Ayırma bilgisi ise;

Eşitlik 4. $A(D;S) = H(\frac{|D1|}{|D|}, \frac{|D2|}{|D|}, \frac{|D3|}{|D|}, \ldots \frac{|Dt|}{|D|}$

Eşitlik 5. H(Dt)=∑(p(t) log(1/p(t))

formülleri ile hesaplanmaktadır. Bilgi kazanım ve ayırma bilgisi formüllerinden hareketle ise bilgi kazanım oranı formülü ise şu şekilde ortaya çıkmaktadır:

Eşitlik 6. $\text{Bilgi kazanım oranı} = \frac{H(\text{S}) - \sum p(t)H(t)}{H(\frac{|D1|}{|D|}, \frac{|D2|}{|D|}, \frac{|D3|}{|D|} \ldots \frac{|Dt|}{|D|})}$

Bilgi kazanım oranı, analize dahil edilen tüm değişkenler için hesaplanır. Ardından en düşük bilgi kazanım oranına sahip değişken entropiyi yani düzensizliği en çok azaltacak değişken olacağı için ağacın kökü olarak seçilir. Ağacın kökündeki dallanmadan sonra, oluşturulacak olan düğümler içinde aynı işlemler tekrarlanarak, düğümlerin belirlenmesi işlemi gerçekleştirilir (Hssina ve diğ., 2014). Böylece entropisi en olacak şekilde ağaca katkı sağlayan homojen sınıflar oluşturulur.

M5-Prime, parçalı doğrusal fonksiyonlara dayalı bir sınıflandırma üreten bir model regresyon ağacı yapıcısı şeklinde tanımlanabilir (Wang & Witten, 1996). M5-Prime algoritmasının çalışma süreci basitçe iki adım olarak düşünülebilir. Bu yönteminin üç temel özelliği şu şekildedir (Diaz ve diğ., 2017): (1) Regresyon ağaçları C4.5 algoritması kullanılarak oluşturulması. (2) Ardından her bir yaprak düğümündeki değer doğrusal bir regresyon fonksiyonu kullanılarak tahmin edilmesi. (3) Her düğümde, alt ağaçta meydana gelen özniteliklerin yalnızca bir alt kümesini kullanılması. Böylece doğrusal regresyondan farklı olarak tüm veri setin için tek bir regresyon denklemi ile açıklama işlemi yapılmak yerine, veri setini oluşturan daha küçük entropiye sahip yapraklar için farklı farklı regresyon denklemleri ile açıklama işlemi gerçekleştirilir. Bu durum basit doğrusal regresyon denklemi ile bir bağımsız değişkenin, bir bağımlı değişkende açıkladığı varyans oranının daha hassas şekilde kontrol edilip gerçekte açıklanabilecek varyans oranına daha yakın sonuçlar elde edilmesini sağlamaktadır.

Değişkenler arası ilişkilerin tahmin edilmesinde ve varyans oranlarının açıklanma işlemlerinde M5-Prime algoritması, basit doğrusal regresyon yöntemine ve sınıflama temelli olmayan diğer yöntemlere (Sinir ağları vb.) oldukça güçlü bir tekniktir (Gonzalez-Sanchez, Frausto-Solis & Ojeda-Bustamante, 2014). Basit doğrusal regresyon yöntemi, açıklanan varyans oranı bağlamında sınırlı iken; sinir ağları gibi yöntemler ise yorumlanabilirlik noktasında sınırlı kalmaktadır. Son beş yıl için yapılan alanyazın incelemelerinde M5-Prime algoritmasının genellikle, tarım, işletme, ticari amaçlı tahminleme yapılan çalışmalarda kullanıldığı belirlenmiştir (Bhardwaj & Bangia, 2020; Chen & He, 2019; King, Rice & Vaughan, 2018; Melesse ve diğ., 2020). Araştırmacılar tarafından bu algoritma kullanılarak sosyal bilimler alanında yapılmış olan bir çalışmaya rastlanmamıştır. Bu algoritmanın psikolojik yapılar arası ilişkilerin incelenmesinin sıklıkla yapıldığı sosyal bilim alanındaki araştırmacılar tarafından kullanılmasının sağlanması için bu çalışma önemli görülmektedir. Buna ek olarak M5-Prime algoritmasının uygulama sürecinin ve farklılıklarının tanıtılması, değişkenler arası ilişkilerin

daha doğru şekilde tespit edilerek yorumlanması için önem arz etmektedir. Bu algoritmanın kullanılabilirliği ve elde edilen bulguların değerlendirilmesine yönelik sunulacak bilgiler özellikle bu algoritmanın sınırlı sayıda çalışmada kullanıldığı alanlarda çalışan araştırmacılara önemli bir kaynak oluşturabilir.

Yapılan bu çalışmada değişkenler arasındaki ilişkilerin betimlenmesi, yordama tahminleri gibi işlemlerde kullanılmak üzere geliştirilen M5-Prime algoritması ve basit doğrusal regresyon yöntemi ile elde edilen sonuçların karşılaştırılması amaçlanmaktadır. Bu amaç doğrultusunda araştırmacılar tarafından üretilen farklı özellikteki veri setleri üzerinde analizler gerçekleştirilmiştir. Her iki yöntemle elde edilen parametre değerleri yorumlanarak iki yöntem arasındaki farklılıklara değinilmiştir. Buna ek olarak farklı çalışmalarda M5-Prime algoritmasının kullanım sürecinin kolaylaştırılması için araştırmacılara, M5-Prime algoritması ile elde edilen parametrelere ilişkin bilgiler sunulmuştur. Böylece sosyal bilimler alanında yapılan yordama amaçlı çalışmalarda bu algoritmanın kullanımının yaygınlaşması ve araştırmacıların kullanım sürecine ilişkin bilgilendirilmesi hedeflenmektedir.

Yöntem

Bu bölümde yapılan bu araştırmanın modeli, verilerin üretilmesi ve üretilen verilerin özellikleri, veri analizi için kullanılan M5-Prime algoritması ve basit doğrusal regresyon analizine ait parametrelerin yorumlanması ve analiz sürecine ilişkin bilgiler verilmiştir.

Araştırma Modeli

Yapılan bu araştırmada, araştırmacılar tarafından üretilmiş olan farklı özellikteki veriler üzerinde M5-Prime algoritması ve basit doğrusal regresyon yöntemi ile analizler gerçekleştirilmiştir. Bu nedenle araştırma farklı özelliklere sahip üretilen verilerle yürütüldüğü için simülasyon çalışması niteliğindedir. Simulasyon çalışmaları deneysel, korelasyonel veya betimsel model olarak nitelendirilebilir (Erkuş, 2013). Yapılan bu araştırmada iki farklı yöntem, üretilen veriler üzerinde işe koşularak elde edilen parametre değerleri incelenerek aralarındaki

farklar ortaya konulmuştur. Bu yönüyle araştırmanın betimsel bir çalışma olduğu ileri sürülebilir.

Verilerin Üretilmesi

Araştırma kapsamında 4 farklı veri seti koşulu üretilmiştir. 4 farklı veri seti üretim koşulu için 3000'er replikasyon yapılmıştır. Toplam 12000 adet veri seti üzerinden analizler gerçekleştirilmiştir. Her bir veri seti koşulu, kullanılan yöntemlerin farklarının net bir biçimde ortaya konulması için farklı özellik taşımaktadır. Üretilen veri setlerinde değişkenler arasındaki ilişkiye ait eğim katsayısının, bağımsız değişkenin farklı değer aralıklarında değişiklik gösterdiği durumlar örneklenmeye çalışılmıştır. Bu veri setlerinin özellikleri Tablo 1'deki gibidir.

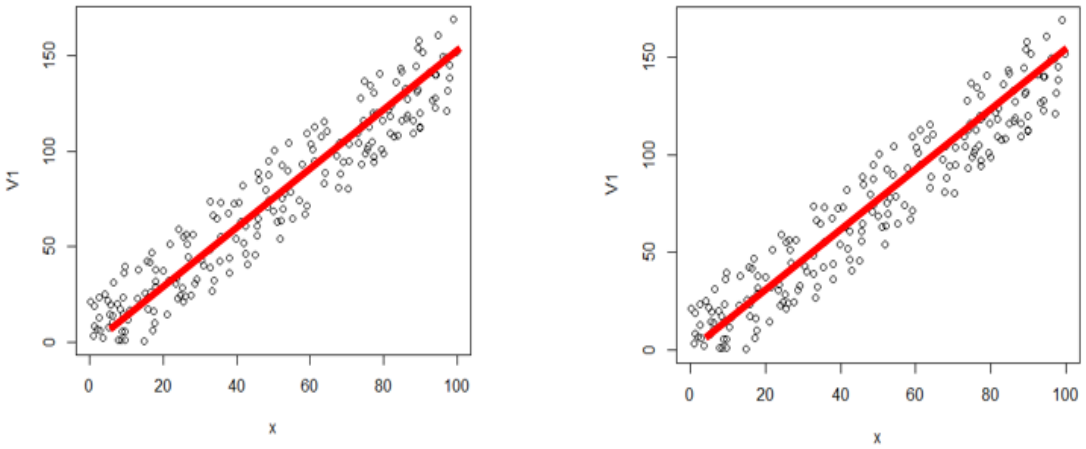
Tablo 1

Üretilen Verilerde Bağımlı Bağımsız Değişken Arasındaki İlişkinin Özellikleri

Veri seti koşulu	İlişkinin yönü	İlişkiye ait eğim katsayısı
Veri seti koşulu 1	Sabit	Sabit
Veri seti koşulu 2	Sabit	Değişken
Veri seti koşulu 3	Değişken	Değişken
Veri seti koşulu 4	Değişken	Değişken

Tablo 1'de verilen özellikler incelendiğinde, doğrusal regresyon ile M5-Prime arasındaki farkın ortaya konulabilmesi için bağımlı-bağımsız değişkenler arasındaki ilişkinin yönü ve eğim katsayısında farklılıklar oluşturulmuştur. Buna göre iki yöntem arasındaki farklılıkların en net bir biçimde ortaya konulabilmesi için öncelikli olarak ilişkinin yönü ve eğim katsayısı sabit şekilde alınmıştır. Bu özelliğe sahip olan veri seti koşulu 1 ile doğrusal regresyon ve M5-Prime arasındaki farkın en az olabileceği durumun örneklenmesi amaçlanmıştır. Veri seti koşulu 1 için iki yönteme ait saçılım grafiği örnekleri Şekil 2'de sunulmuştur.

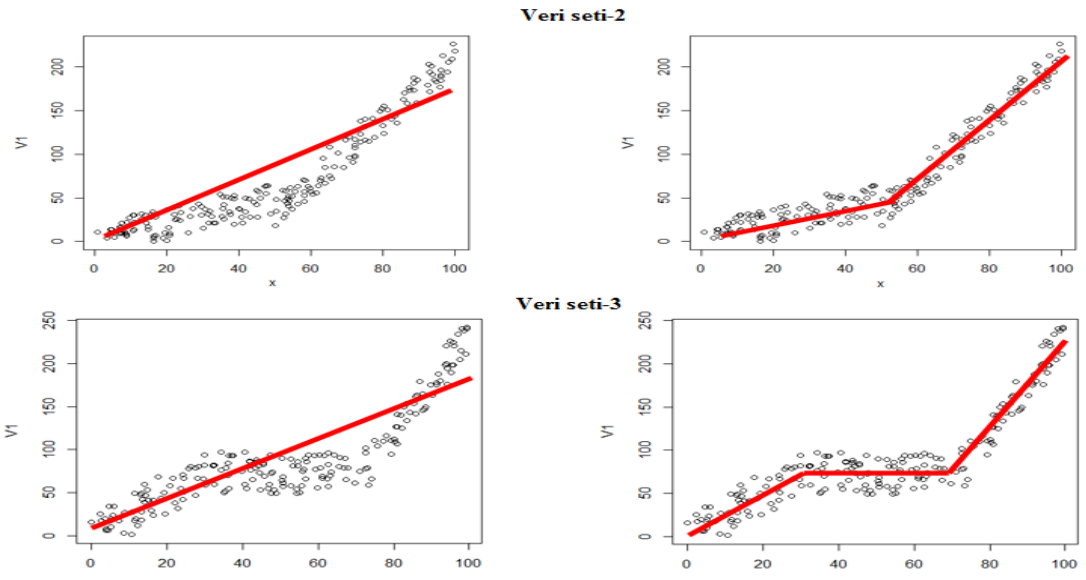
Doğrusal regresyon M5-prime



Şekil 2 *Veri Seti Koşulu 1 İçin İki Yöntem Kapsamında Çizilen Saçılım Grafikleri*

Veri seti koşullarından 2 ve 3 için ise ilişkinin eğim katsayısında farklılaşmaya neden olacak şekilde değişimlemeler yapılarak, gerçek durumlarda ortaya çıkması muhtemel olan ilişki durumlarının örneklenmesi amaçlanmıştır. Bu iki veri seti koşulunda bağımlı, bağımsız değişken arasındaki ilişkinin yönü her değer aralığında aynı iken eğim katsayısında ise farklılaşmalar söz konusudur. Veri seti koşullarından 2 ve 3 için iki yönteme ait saçılım grafiği örnekleri Şekil 3'de sunulmuştur.

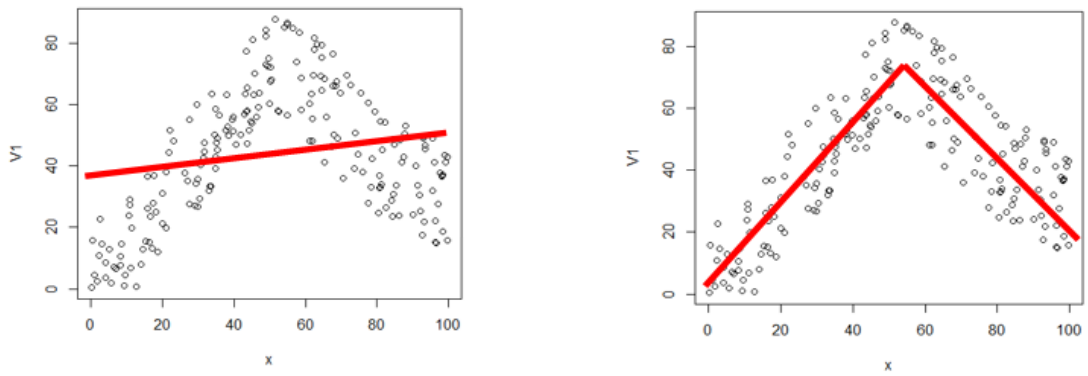
Doğrusal regresyon M5-prime



Şekil 3 *Veri Seti Koşullarından 2 Ve 3 İçin İki Yöntem Kapsamında Çizilen Saçılım Grafikleri*

Son olarak veri seti koşulu 4'te ise iki değişken arasındaki ilişki, bağımsız değişkenin farklı değer aralıklarında farklı yön ve eğim katsayısına sahip olacak şekilde tasarım yapılmıştır. Bu koşul için iki yönteme ait saçılım grafiği örnekleri Şekil 4'te sunulmuştur.

Doğrusal regresyon M5-prime



Şekil 4 *Veri Seti Koşulu 4 İçin İki Yöntem Kapsamında Çizilen Saçılım Grafikleri*

Şekil 2 3 ve 4'te çalışma kapsamında karşılaştırılan iki yöntemin farklı özellikteki veri seti koşulları için oluşturacağı olası regresyon doğruları gösterilmiştir. Buna göre doğrusal regresyon tek bir doğru ile değişkenler arasındaki ilişkiyi açıklamaya çalışırken; M5-Prime ise bağımsız değişken için oluşturduğu farklı aralıklardaki farklı yön ve güçteki ilişkileri açıklamak için birden fazla regresyon doğrusu oluşturabilmektedir. İki yöntem arasındaki farklılığın en önemli kaynağının oluşturulan regresyon modelleri olduğu ileri sürülebilir.

Grafikleri ve özellikleri sunulan bu veri setleri R programla dilinde araştırmacılar tarafından yazılan kodlarla üretilmiştir.

Verilerin Analizi

Verilerin analizi sürecinde üretilen R programlama dilinde hazırlanmış olan RWeka version 0.4-42 paketi kullanılmıştır. M5-Prime analizleri için bu paketin M5P fonksiyonundan yararlanılmıştır. Doğrusal regresyon analizleri için ise aynı paketin LinearRegression fonksiyonu tercih edilmiştir. RWeka paketi, Hornik ve diğerleri (2009) tarafından geliştirilmiş olup, 2020 yılında versiyon güncellemesi yapılmıştır.

Verilerin analizi ile elde edilen sonuçlarda doğrusal regresyon ve M5-Prime yöntemlerinin oluşturdukları modellerin başarılarının değerlendirilmesi açısından beş farklı parametre yorumlanmıştır. Bu parametreler, R (korelasyon), R2 (açıklanan varyans), ortalama mutlak hata (OMH-Mean Absolute Error), hataların ortalama karekökü (RMSE-Root Mean Square Error), göreceli mutlak hata (GMH- Relative Absolute Errror) şeklindedir.

R (korelasyon): Gözlenen değerler ve kestirilen değerler arasındaki korelasyon değeridir.

R2 (açıklanan varyans): R değerinin karesi alınarak elde edilen bu parametre, gözlenen değerlerin, kestirilen değerlerde açıkladığı varyans miktarıdır.

Ortalama mutlak hata (OMH-Mean Absolute Error): Gözlenen değerlerle, kestirilen değerler arasındaki farkların mutlak değerlerinin ortalamasıdır. Bir kestirimdeki ortalama hata miktarını temsil eder. Ortalama mutlak hata değeri aşağıdaki eşitlik yardımı ile hesaplanabilir.

Eşitlik 7. $\frac{\sum |Gözlenen\ değer - Kestirilen\ Değer|}{Gözlem\ Sayısı}$

Hataların ortalama karekökü (RMSE-Root Mean Square Error): Model parametrelerinin evren kovaryansları ile ne derece uyumlu olduğunu gösteren uyum indeksidir (Byrne, 1998). RMSE değeri aşağıdaki eşitlik yardımı ile hesaplanabilir.

n, gözlem sayısı olmak üzere;

Eşitlik 8. $\sqrt{\frac{\sum(Gözlenen\ değer - Kestirilen\ değer)}{n}}$

Göreceli mutlak hata (GMH- Relative Absolute Errror): Göreceli mutlak hata, ortalama mutlak hata göstergesindeki şans başarısı faktörünü ortadan kaldırmak için geliştirilmiş bir değerlendirme sistemidir (Armstrong ve Collopy, 1992). GMH hesaplamada kullanılan formül aşağıda paylaşılmıştır.

Eşitlik 9. $GMH = \frac{\text{Ortalama mutlak hata (OMH)}}{\text{İlkel tahmincinin ortalama mutlak hata miktarı}}$

Yukarıdaki eşitlikte ilkel (primitive) tahmincinin ortalama mutlak hata miktarı en ilkel kestirim sisteminde elde edilecek hata oranını gösterir. Sayısal değişkenler için yapılan

tahminler ilkel tahminci tüm verilere ortalama değeri atamaktadır. Örneğin, 100 veriden oluşan bir değişken ile bir tahmin modeli oluşturmak isteyen bir araştırmacı için en ilkel tahmin yöntemi bu 100 verinin ortalama değerinin tahmin değeri olarak atanmasıdır. Bu işlemin amacı oluşturulan modelin hata oranını, ilkel tahmicinin hata oranı ile karşılaştırarak modelin başarısını daha net ortaya koymaktır. GMH değeri azaldıkça modelin başarısının arttığı yorumu yapılabilir (Cichosz, 2015). GMH değeri 100% olarak hesaplanırsa oluşturulan modelin tamamen anlamsız, ilkel tahminciden farksız olduğu yorumu yapılmaktadır. Genel olarak parametrelerin yorumlanma sürecinde R ve R2 değerlerinin yüksek olması; OMH, RMSE ve GMH değerlerinin ise düşük olması modelin başarılı olduğunu göstermektedir.

Etik Kurul Kararı

Araştırma da herhangi bir şekilde insan gruplarına uygulama yapılmadığından ve klinik bir inceleme süreci gerçekleştirilmediğinden etik kurul izni alınmamıştır. Araştırmada kullanılan veriler R programlama dili kullanılarak araştırmacılar tarafından üretilmiştir.

Bulgular

Bu bölümde üretilen verilere ait betimsel istatistikler, doğrusal regresyon ve M5-Prime yöntemleri ile yapılan analiz sonuçlarına ait bulgular sunulmuştur. İlk olarak verilerin betimsel istatistiklerine ait sonuçlar rapor edilmiştir. Ardından Veri seti koşulu 1-2-3-4 için elde edilen sonuçlar sırasıyla açıklanmıştır.

Betimsel İstatistiklere İlişkin Bulgular ve Yorum

Bu kısımda 4 farklı veri seti koşulu için ayrı ayrı üretilen 3000’er adet veri setinde yer alan bağımlı (V1) ve bağımsız (x) değişkenlere ait ortalama ve standart sapma değerlerine verilmiştir. Bu değerler Tablo 2’de açıklanmıştır.

Tablo 2.

Veri Setlerine Ait Betimsel İstatistikler

Veri seti koşulu	Ortalama V1	Standart Sapma V1	Ortalama x	Standart Sapma x
Veri Seti Koşulu 1	45.562	23.111	49.960	28.846
Veri Seti Koşulu 2	67.406	66.949	50.035	28.872
Veri Seti Koşulu 3	75.779	44.524	50.054	28.889
Veri Seti Koşulu 4	83.580	54.343	50.040	28.840

Elde edilen betimsel istatistiklerde veri setlerindeki bağımlı değişken için hesaplanan aritmetik ortalama değerleri 45.562 – 83.580 arasında; bağımsız değişken için ise 49.960 - 50.040 arasında değişiklik göstermektedir. Standart sapma değerleri ise bağımlı değişken için 23.111 – 66.949 arasında iken; bağımsız değişken için ise 28.840 – 28.846 arasında yer almaktadır.

Doğrusal Regresyon ve M5-Prime Analizlerine İlişkin Bulgular ve Yorum

Çalışma kapsamında yapılan analizlerde elde edilen parametrelere ilişkin sonuçlar her iki yöntem içinde verilmiştir. Elde edilen sonuçlar Tablo 3'te yer almaktadır. Elde edilen sonuçların her biri için elde edilen farkın anlamlılıkları karşılaştırılmıştır.

Tablo 3.

Doğrusal Regresyon ve M5-Prime Yöntemleri Parametre Sonuçları

Veri seti koşulu	R		R2		Ortalama Mutlak Hata		RMSE		Göreceli Mutlak Hata (%)	
	M5P	LR	M5P	LR	M5P	LR	M5P	LR	M5P	LR
Veri Seti K. 1	0.948	0.948	0.900	0.900	12.014	12.023	13.988	13.998	31.55	31.57
Veri Seti K. 2	0.984	0.927	0.968	0.860	10.063	20.682	11.849	24.858	17.27	35.48
Veri Seti K. 3	0.960	0.843	0.923	0.712	12.871	23.088	15.359	28.953	31.97	57.24
Veri Seti K. 4	0.802	0.262	0.643	0.073	11.862	18.327	14.105	22.150	62.20	95.96

Tablo 3'te yer alan sonuçlar incelendiğinde;

Veri seti koşulu 1 için M5-Prime yöntemi ve doğrusal regresyon ile elde edilen 3000 adet R ve R2 değerlerinin aynı olduğu gözlemlenmiştir. Sırasıyla bu değerler 0.948 ve 0.900 olarak hesaplanmıştır. OMH, RMSE ve GMH için ise değerler arasında önemli bir fark olmadığı görülmüştür. Bu durumun sebebi olarak Veri seti koşulu 1'de bağımlı ve bağımsız değişken arasındaki ilişkinin yön veya eğim katsayısının değişmiyor olması gösterilebilir. Bu koşul altında M5-Prime yöntemi, doğrusal regresyonda olduğu gibi modeli oluşturmak için tek bir doğrusal denklem üretmiştir. Dolayısıyla iki yöntem arasındaki sonuçlar farklılaşmamıştır.

Veri seti koşulu 2 için M5-Prime yöntemi ile elde edilen R değeri 0.984 iken doğrusal regresyonla elde edilen değer ise 0.927 şeklindedir. Her iki yöntemle elde edilen R2 değerleri ise sırasıyla 0.968 ve 0.860'dır. Ortalama mutlak hata miktarları incelendiği ise M5-Prime için 10.063; doğrusal regresyon için ise 20.682 değerleri elde edildiği görülmektedir. RMSE değerleri, her iki yöntem için sırasıyla 11.849 ve 24.858 olarak belirlenmiştir. Son olarak göreceli mutlak hata miktarı M5-Prime için %17.27 iken doğrusal regresyon için %35.48'dir. Veri seti koşulu 2 için bağımlı ve bağımsız değişken arasındaki ilişkinin yönü değişmese de bağımsız değişkenin farklı değer aralıkları için ilişkinin eğim katsayısı değişmektedir. Bu sebeple, M5-Prime yöntemi için ilişkinin eğim katsayısının farklı olduğu iki aralık için en az iki farklı regresyon modeli oluşturmuştur. Ancak doğrusal regresyon ise yine tüm veri seti koşulları için tek bir regresyon modeli üretmiştir. Bu durum M5-Prime yönteminin, doğrusal regresyon yöntemine göre daha başarılı kestirimler yapmasını sağlamıştır. İki yöntemle de elde edilen 3000 adet R2 ve RMSE değerleri ortalaması arasındaki farkın 0,05 düzeyinde anlamlı olduğu belirlenmiştir.

Veri seti koşulu 3 için M5-Prime yöntemi ile elde edilen R değeri 0.960; doğrusal regresyonla elde edilen değer ise 0.843 olarak hesaplanmıştır. Her iki yöntemle elde edilen R2 değerleri sırasıyla 0.923 ve 0.712 şeklindedir. Ortalama mutlak hata miktarları ise M5-Prime için 12.871; doğrusal regresyon için ise 23.088 olarak belirlenmiştir. RMSE değerlerine, her iki

yöntem için sırasıyla 15.359 ve 28.953 olarak ulaşılmıştır. Son olarak göreceli mutlak hata miktarı M5-Prime için %31.97 iken doğrusal regresyon için %57.21'dir. Veri seti koşulu 3 için bağımlı ve bağımsız değişken arasındaki ilişkinin eğim katsayısı iki kez değişiklik göstermektedir. Bu sebeple, M5-Prime yöntemi ilişkinin eğim katsayısın farklı olduğu üç bağımsız değişken değer aralığında, en az üç farklı regresyon modeli oluşturmuştur. Ancak doğrusal regresyon ise yine tüm veri seti koşulları için tek bir regresyon modeli üzerinde işlem yapmıştır. Bu durum M5-Prime yönteminin, doğrusal regresyon yöntemine kıyasla daha başarılı kestirimler yapmasını sağlamıştır. Veri seti koşulu 2 ile karşılaştırıldığında, ilişkin eğim katsayısında daha fazla farklılaşmanın olması iki yöntem arasındaki başarı farkının artmasına yol açmıştır. Buna göre veri seti koşulu 3'te iki yöntem arasındaki model başarısı farkının arttığı gözlemlenmiştir. İki yöntemle de elde edilen 3000 adet R2 ve RMSE değerleri ortalaması arasındaki farkın 0,05 düzeyinde anlamlı olduğu belirlenmiştir.

Veri seti koşulu 4 için M5-Prime yöntemi ile R değeri 0.802; doğrusal regresyonla ise 0.262 olarak hesaplanmıştır. Her iki yöntemle elde edilen R2 değerleri ise sırasıyla 0.643 ve 0.073'dür. Ortalama mutlak hata miktarları, M5-Prime için 11.862; doğrusal regresyon için ise 18.327 olarak ortaya konulmuştur. RMSE değerleri, M5-prime için 14.105 ve doğrusal regresyon için ise 22.150 olarak rapor edilmiştir. Son olarak göreceli mutlak hata miktarı M5-Prime için %62.20 iken doğrusal regresyon için %95.96'dır. Veri seti koşulu 4 için bağımlı ve bağımsız değişken arasındaki ilişkinin hem yönü hem eğim katsayısı değişiklik göstermektedir. İki değişken arasındaki ilişkinin yönü ve eğim katsayısı bağımsız değişkenin farklı değer aralıklarında değişiklik göstermektedir. Bu koşul altında M5-Prime bağımsız değişkenin iki farklı aralığı için en az iki regresyon modeli oluştururken; doğrusal regresyon ise bu ilişkiyi tek bir doğru ile açıklamaya çalışmıştır. Bu sonuçlarla birlikte, Şekil 3'te görüldüğü gibi bir noktada yön değiştirmiş olan bir ilişkinin tek bir doğru denklemi ile açıklanmasının çok zor olduğu ileri sürülebilir. Veri seti koşulu 4 doğrusal regresyon yöntemi için açıklaması en güç ilişkiyi temsil etmektedir. Elde edilen sonuçlar karşılaştırıldığında iki yöntem arasındaki başarı farkının en üst

düzeye çıktığı veri setinin, veri seti koşulu 4 olduğu görülmektedir. Bu veri seti koşulu için de iki yöntemle de elde edilen 3000 adet R2 ve RMSE değerleri arasındaki farkın 0,05 düzeyinde anlamlı olduğu belirlenmiştir.

Tartışma, Sonuç ve Öneriler

Alanyazında yapılan M5-Prime algoritmasının üyesi olduğu regresyon ağaçları algoritmaları ile doğrusal regresyon modellerinin karşılaştırıldığı çalışmalar incelenmiştir. Buna göre, Shafiullah ve diğerleri (2008) yapmış oldukları çalışmada, değişkenler arası ilişkileri incelemeyi amaçlamışlar ve bu doğrultuda dokuz farklı yöntemi karşılaştırmışlardır. Bu yöntemlerden ikisi doğrusal regresyon ve M5-Prime algoritmasıdır. Elde edilen sonuçlara göre iki yöntemi eş değer sonuçlar verdiği rapor edilmiştir. İnceleme yapılan beş veri setinden ilkinde bu iki yöntemin tahmin edilen ve gerçek değerler arası korelasyonu 1,00 olacak şekilde kestirimde bulunduğu belirtilmiştir. Rapor edilen sonuçların, yapılan bu çalışma kapsamında incelenen veri seti koşulu 1'e ait sonuçlarla örtüştüğü görülmüştür.

Gonzalez-Sanchez, Frausto-Solis ve Ojeda-Bustamante (2014) ve King, Rice ve Vaughan (2018) gerçekleştirdikleri çalışmalarda, farklı yöntemler kullanarak değişkenler arası ilişki incelemeleri gerçekleştirmiştir. Her iki çalışmada elde edilen sonuçlar, M5-Prime algoritması aracılığıyla ulaşılan RMSE değerlerinin; doğrusal regresyon analizi ile ulaşılana göre daha düşük olduğunu göstermektedir. Bu bulgular ışığında M5-Prime algoritmasının değişkenler arası ilişkiyi açıklama konusunda doğrusal regresyona kıyasla daha başarılı olduğu gözlemlenmektedir. Bu sonuç yapılan bu çalışmada veri 2 – 3 ve 4 için elde edilen analiz sonuçlarını destekler niteliktedir.

Yapılan bu araştırmada iki sürekli değişken arasındaki ilişkinin yönü ve eğim katsayısının değiştiği veri setlerinde basit doğrusal regresyon ve M5-Prime algoritması karşılaştırılmıştır. Kullanılan veri setlerinde ilişkinin yönü ve eğim katsayısının değişme durumu, bağımsız değişkenin farklı değer aralıklarında, değişkenler arasındaki ilişkiye ait doğrunun yön ve eğiminin farklılaşması olarak ele alınmıştır. Bu bağlamda araştırma kapsamında, sosyal

bilimlerde karşılaşılması olası durumları örneklemek üzere dört farklı özelliğe sahip olan, değişkeneler arası ilişkinin yön ve eğim katsayısının, bağımsız değişkenin belirli düzeylerinde değiştiği veri setleri üretilmiştir. Sosyal bilimlerde değişkenler arası ilişkilerin sabit bir yön ve eğim katsayısına sahip olmasının beklenmesi her zaman mümkün olmamaktadır. Bu tür durumlarda değişkenlerin birbiri üzerindeki etkilerinin ve aralarındaki ilişkilerin incelenmesi için uygun modellemelerin oluşturulması gerekmektedir. Değişkenler arası ilişkiler için uygun modellerin tercih edilmesi araştırmacıların değişkenler arasındaki ilişkiyi daha doğru şekilde açıklamasına katkıda bulunabilir.

Yapılan bu araştırmada dört farklı durumu temsilen eden veri setlerinden üçünde, M5-Prime algoritması ile ortaya konulan modellerin değişkenler arası ilişkinin açıklanmasında daha başarılı olduğu belirlenmiştir. Korelasyon katsayısı, açıklanan varyans oranı, ortalama mutlak hata, RMSE ve göreceli mutlak hata parametreleri bakımından M5-Prime algoritmasının, doğrusal regresyon analizine nazaran daha iyi sonuçlar verdiği görülmüştür. Değişkenler arasındaki ilişkinin; bağımsız değişkenin farklı değer aralıklarında yön ve eğim katsayısı değiştirmediği veri setlerinde ise her iki yönteminde aynı sonuçları verdiği gözlenmiştir. Söz konusu veri seti koşulu için değişkenler arası ilişkinin yönü ve eğim katsayısı değişmediği için M5-Prime algoritması ilişkinin açıklanması sürecinde tek bir doğru denklemi kullanmıştır. Bu nedenle de ilişkiyi her tür veri seti koşulunda tek bir doğru denklemi ile açıklamaya çalışan doğrusal regresyonla aynı sonuçları vermiştir. Doğrusal regresyon modellerine en uygun olan değişkenler arası ilişkinin yönü ve eğim katsayısının aynı kaldığı veri setlerinde dahi, M5-Prime algoritmasının doğrusal regresyon modelleri ile aynı sonuçları verdiği bulgusuna ulaşılmıştır. Buna göre sosyal bilimlerdeki değişkenler arası ilişkilerin farklılaşabileceği durumu göz önüne alınarak; bu ilişkileri açıklamayı amaçlayan çalışmalarda M5-Prime algoritması kullanılması önerilebilir.

Çıkar Çatışması Bildirimi

Yazar(lar), bu makalenin araştırılması, yazarlığı ve / veya yayınlanmasına ilişkin herhangi bir potansiyel çıkar çatışması beyan etmemiştir.

Destek/Finansman Bilgileri

Yazar(lar), bu makalenin araştırılması, yazarlığı ve / veya yayınlanması için herhangi bir finansal destek almamıştır.

Etik Kurul Kararı

Araştırma da herhangi bir şekilde insan gruplarına uygulama yapılmadığından, klinik bir inceleme süreci gerçekleştirilmediğinden etik kurul izni alınmamıştır. Araştırmada kullanılan veriler R programlama dili kullanılarak araştırmacılar tarafından üretilmiştir.

Kaynakça/References

Alpar, R. (2017). *Uygulamalı çok değişkenli istatistiksel yöntemler*. Detay Yayıncılık.

Altan, T., & Eldeleklioğlu, J. (2019). Lise öğrencilerinde siber zorbalığın yordayıcısı olarak siber mağduriyet ve duygusal zekâ. *Elementary Education Online, 18*(4), 2147-2156. https://doi.org/10.17051/ilkonline.2019.639360

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician 27*, 17–21. https://amstat.tandfonline.com/doi/abs/10.1080/00031305.1973.10478966?journalCode=utas20#.YIbjipAzbIU

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons, *International Journal of Forecasting, 8,* 69-80. https://doi.org/10.1016/0169-2070(92)90008-W

Behnood, A., Behnood, V., Gharehveran, M. M., & Alyamac, K. E. (2017). Prediction of the compressive strength of normal and high-performance concretes using M5P model tree algorithm. *Construction and Building Materials*, *142*, 199-207. https://doi.org/10.1016/j.conbuildmat.2017.03.061

Bhardwaj R., & Bangia A. (2020). Assessment of stock prices variation using ıntelligent machine learning techniques for the prediction of BSE. In D. Dutta & B. Mahanty (Eds.) *Numerical optimization in engineering and sciences. Advances in ıntelligent systems and computing*, (Vol 979, pp. 159-166). Springer, Singapore. https://doi.org/10.1007/978-981-15-3215-3_15

Bostancı, A. B., & Tosun, A. (2019). Okulların DNA profilleri ile öğretmenlerin örgütsel vatandaşlık düzeyleri arasındaki ilişki. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 19*(3), 1115-1127. https://doi.org/10.17240/aibuefd.2019.19.49440-541193

Breiman, L., & Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association 80,* 580–598. https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1985.10478157#.YIbkYJAzaUk

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS and SIMPLIS: Basic concepts, applications and programming*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Chatterjee, S., & Simonoff J. S. (2013). *Handbook of regression analysis*. Canada: John & Sons, Inc., Hoboken.

Chatterjee, S., & Hadi, A. S. (2015). *Regression analysis by example*. John Wiley & Sons.

Chen, E., & He, X. J. (2019). Crude oil price prediction with decision tree based regression approach. *Journal of International Technology and Information Management, 27*(4), 2-16. https://scholarworks.lib.csusb.edu/jitim/vol27/iss4/1/

Cichosz, P. (2015). *Data mining algorithms: explained using R*. John Wiley & Sons Incorporated.

Civelek, M., & Durukan, M. (2012). *İstatistiksel analiz, istatiksel bilgi kullanıcıları için el kitabı.* Ankara: Nobel Akademi.

Cox, D. R. (1981). Statistical analysis of time series: Some recent developments. *Scandinavian Journal of Statistics 8,* 93–115. https://www.jstor.org/stable/4615819?seq=1#metadata_info_tab_contents

Díaz, I., Mazza, S. M., Álvarez, E. F. C., Giménez, L. I., & Gaiad, J. E. (2017). Machine learning applied to the prediction of citrus production. *Spanish journal of agricultural research*, *15*(2), 1-12. https://dialnet.unirioja.es/servlet/articulo?codigo=6334815

Doğan, K., & Şirin, H. D. (2019). Yetişkin güvensiz bağlanma boyutlarının eş tükenmişliğini yordama gücü: Üniversite akademik personeli örneği. *Selçuk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 42*, 23-32. http://dergisosyalbil.selcuk.edu.tr/susbed/article/view/1554/1224

Dumludağ, D., Gökdemir, O., & Giray, S. (2016). Income comparison, collectivism and life satisfaction in Turkey. *Quality & Quantity*, *50*(3), 955-980. https://doi.org/10.1007/s11135-015-0185-1

Dunham, M.H. (2003). *Data mining introductory and advanced topics*. Upper Saddle River, NJ: Pearson Education, Inc.

Eugene, Y. K., & Johnston, R. G. (1996). The ineffectiveness of the correlation coefficient for ımage comparisons, *Technical Report LAUR-96-2474*. http://citeseerx.ist.psu.edu/viewdoc/citations?doi=10.1.1.47.399

Erkuş, A. (2013). *Davranış bilimleri için bilimsel araştırma süreci*. Seçkin Yayıncılık.

Güvercin, D. (2018). Terörizmin, eğitimde cinsiyet eşitsizliği üzerine etkisi: Türkiye üzerine il bazında uygulamalı çalışma. *Journal of Yaşar University*, *13*(51), 281-292. https://dergipark.org.tr/en/pub/jyasar/issue/39015/378651

González Sánchez, A., Frausto Solís, J., & Ojeda Bustamante, W. (2014). Predictive ability of machine learning methods for massive crop yield prediction. *Spanish Journal of Agricultural Research, 12*(2), 313-328. http://dx.doi.org/10.5424/sjar/2014122-4439

Gupta, D. L., Malviya, A. K., & Singh, S. (2012). Performance analysis of classification tree learning algorithms. *International Journal of Computer Applications*, *55*(6), 39-44. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.244.9217&rep=rep1&type=pdf

Han, J., Kamber, M., & Pei, J. (2011). Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems*, *5*(4), 83-124.

Harma, M. (2008). *The impact of parental control and marital conflict on adolescents' self-regulation and adjustment* [Unpublished master's thesis]. Middle East Technical University, Ankara, Türkiye.

Hornik, K., Buchta, C., & Zeileis, A. (2009). Open-source machine learning: R meets Weka. *Computational statistics, 24*(2), 225-232. https://doi.org/10.1007/s00180-008-0119-7

Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A comparative study of decision tree ID3 and C4. 5. *International Journal of Advanced Computer Science and Applications, 4*(2), 13-19. İlgili makaleye ait indirme linki için tıklayınız.

King, B. E., Rice, J., & Vaughan, J. (2018). Using machine learning to predict national hockey league average home game attendance. *Journal of Prediction Markets, 12*(1), 85-98. doi: https://doi.org/10.5750/jpm.v12i2.1608

Kisi, O., Shiri, J., & Demir, V. (2017). Hydrological time series forecasting using three different heuristic regression techniques. In P. Samui, S. Sekhar, & V. E. Balas (Eds.) *Handbook of Neural Computation* (pp. 45-65). Academic Press. https://doi.org/10.1016/C2016-0-01217-2

Lee, J. (1992). A coutionary note on the use of the correlation-coefficient. *British Journal of Industrial Medicine 49*, 526–527. https://doi.org/10.1136/oem.49.7.526-a

Leech, N.L., Barrett, K. C., & Morgan, G.A. (2005). *Spss for intermediate statistics: Use and interpretation*. London: Lawrence Erkbaum Associates Inc.

Lestari, A. (2020). Increasing accuracy of C4. 5 algorithm using information gain ratio and adaboost for classification of chronic kidney disease. *Journal of Soft Computing Exploration, 1*(1), 32-38. https://doi.org/10.52465/joscex.v1i1.6

Melesse, A. M., Khosravi, K., Tiefenbacher, J. P., Heddam, S., Kim, S., Mosavi, A., & Pham, B. T. (2020). River water salinity prediction using hybrid machine learning models. *Water, 12*(10), 2951. https://doi.org/10.3390/w12102951

Nguyen, H. V., Muller, E., Vreeken, J., Keller, F., & Böhm, F. (2013). CMI: An information-theoretic contrast measure for enhancing subspace cluster and outlier detection. In J. Ghosh , Z. Obradovic , J. Dy , Z.H. Zhou , C. Kamath, & S. Parthasarathy (Eds.) *Proceedings of the 2013 SIAM International Conference on Data Mining, Austin, TX*, *USA 2–4 May 2013,* (pp. 198–206). https://epubs.siam.org/doi/abs/10.1137/1.9781611972832.22

Onwuegbuzie, A. J., & Daniel, L. G. (2002). Uses and misuses of the correlation coefficient. *Research in the Schools, 9*, 73–90. https://eric.ed.gov/?id=ED437399

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81-106. https://link.springer.com/article/10.1007/BF00116251

Quinlan, J. R. (1992). Learning with continuous classes. In Adam, & Sterling (Eds.) *Proceedings Australian Joint Conference on Artificial Intelligence, Hobart, Australia 16-18 November 1992*, (pp.343-348.). https://sci2s.ugr.es/keel/pdf/algorithm/congreso/1992-Quinlan-AI.pdf

Quinlan, J. R. (1993). *C4. 5: Programs for Machine Learning.* Morgan Kaufmann.

Rodgers, J. L., & Nicewander. W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician, 42*, 59–66. https://doi.org/10.1080/00031305.1988.10475524

Shafiullah, G. M., Simson, S., Thompson, A., Wolfs, P. J., & Ali, A. B. M. S. (2008). Forecasting vertical acceleration railway wagons-A comparative study. *In 4th International Conference on Data Mining (DMIN'08), 14-17 July 2008*, Las Vegas, NV, USA. https://researchrepository.murdoch.edu.au/id/eprint/31856/

Siciliano, R., & Mola, F. (2000). Multivariate data analysis and modeling through classification and regression trees. *Computational Statistics & Data Analysis, 32*(3-4), 285-301. https://doi.org/10.1016/S0167-9473(99)00082-1

Silahtaroğlu, G. (2016). *Veri madenciliği* (3. Basım). Papatya Yayınları.

Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130. https://doi.org/10.11919/j.issn.1002-0829.215044

Sykes, A. O. (1993). *An Introduction to Regression Analysis* [Working paper]. Coase-Sandor Institute for Law & Economics, Chicago, USA.

Tabachnick, B.G. & Fidell L.S. (2007). *Using Multivariate Statistics*. (Fifth Editition). ABD: Pearson Education.

Wang, Y., & Witten, I. H. (1996). Induction of model trees for predicting continuous classes. (Working paper 96/23). Hamilton, New Zealand: University of Waikato, Department of Computer Science. https://researchcommons.waikato.ac.nz/bitstream/handle/10289/1183/uow-cs-wp-1996-23.pdf?sequence=1&isAllowed=y

Xiaoliang, Z., Hongcan, Y., Jian, W., & Shangzhuo, W. (2009). Research and application of the improved algorithm C4. 5 on decision tree*. International Conference on Test and Measurement*, *2*, 184-187.

İletişim/Correspondence

Uzman Hüseyin YILDIZ, huseyinyildiz35@gmail.com

Dr. Öğr. Üyesi Alperen YANDI, alperenyandi@gmail.com