Uludağ Üniversitesi Mühendislik Fakültesi Dergisi, Cilt 22, Sayı 2, 2017 ARAŞTIRMA

DOI: 10.17482/uumfd.338805

SAĞLIK HARCAMASININ TAHMİNİNDE MAKİNE ÖĞRENMESİ REGRESYON YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Songül ÇINAROĞLU*

Alınma: 07.03.2016; düzeltme: 18.05.2017; kabul: 19.08.2017

Öz: Farklı veri setleri üzerinde yapılan uygulamalar sonucunda modellenmesi zor olan değişkenlerin varlığında klasik regresyon yöntemlerine alternatif olarak makine öğrenmesi regresyon yöntemlerinin kullanımı tavsiye edilmektedir. Sağlık harcaması modellenmesi zor olan bir değişken olup, literatürde makine öğrenmesi regresyon yöntemleri karşılaştırılarak bu değişkenin modellendiği bir çalışmaya rastlanmamıştır. Bu çalışmada kişi başı sağlık harcamasının tahmini amacıyla bir çoklu regresyon modeli oluşturulmuştur. Farklı hiperparametre değerleri belirlendiğinde elde edilen Lasso Regresyon, Rastgele Ağaç Regresyonu ile Destek Vektör Makinesi Regresyon performans sonuçları karşılaştırılmıştır. Çalışmada hiperparametre değeri olarak Lasso Regresyon için lamda (λ) değeri, Rastgele Ağaç Regresyonu için ağaç sayısı, Destek Vektör Regresyonu için epsilon (ε) değeri esas alınmıştır. Sonuçlar 5 ile 50 arasında değişen "k" parça çapraz geçerlilik uygulanarak performe edildiğinde makine öğrenmesi regresyon yöntemlerine ait performans sonuçlarının R^2, RMSE ve MAE değerleri bakımından istatistiksel olarak anlamlı farklılıklar gösterdiği ($p<0.001$) tespit edilmiştir. Tahmin performanslarına ait yüzey ve çubuk grafikleri ile istatistiksel test sonuçları incelendiğinde farklı hiperparametre değerlerine göre Rastgele Ağaç Regresyonun ($R^2 > 0.7500$, RMSE ≤ 0.6000 ve MAE ≤ 0.4000) daha iyi tahmin sonuçlarına sahip olduğu belirlenmiştir. Çalışma sonuçlarının, sağlık harcamasının modellendiği araştırmalar için makine öğrenmesi regresyon yöntemleri kullanıldığında en uygun hiperparametre değerlerinin belirlenmesi konusunda katkı sağlaması beklenmektedir.

Anahtar Kelimeler: Makine Öğrenmesi, Lasso Regresyon, Rastgele Ağaç Regresyonu, Destek Vektör Regresyonu, Sağlık Harcaması

Comparison of Machine Learning Regression Methods to Predict Health Expenditures

Abstract: As a result of experimental studies on different datasets, it is recommended to use machine learning regression methods as an alternative to classical regression methods in the existence of variables which are difficult to model. Health expenditure is an indicator which is difficult to model and there is no study in the literature about modelling health expenditure comparing machine learning regression methods. In this study a multiple regression model was conducted to predict health expenditure per capita. Performance results of Lasso Regression, Random Forest Regression and Support Vector Machine Regression compared when different hyperparameter values were determined. Lambda (λ) value for Lasso Regression, number of trees for Random Forest Regression, epsilon (ε) value for Support Vector Regression was determined as hyperparameter values. Study results performed by using "k" fold cross validation changed from 5 to 50, indicate the difference between machine learning results in terms of R^2, RMSE and MAE values that are statistically significant ($p<0.001$). Surface and bar plots and statistical test results of prediction performances show that Random Forest Regression ($R^2 > 0.7500$, RMSE ≤ 0.6000 ve MAE ≤ 0.4000) has better prediction performance according to different hyperparameter values. It is hoped that study results make contribution to studies about determining optimal hyperparameter values for machine learning regression methods for studies about modelling health expenditures.

Keywords: Machine Learning, Lasso Regression, Random Forest Regression, Support Vector Regression, Health Expenditure

* Hacettepe Üniversitesi, IIBF, Sağlık Yönetimi Bölümü, Beytepe, Ankara,
İletişim Yazarı: Songül ÇINAROĞLU (cinaroglus@hacettepe.edu.tr)

1. GİRİŞ

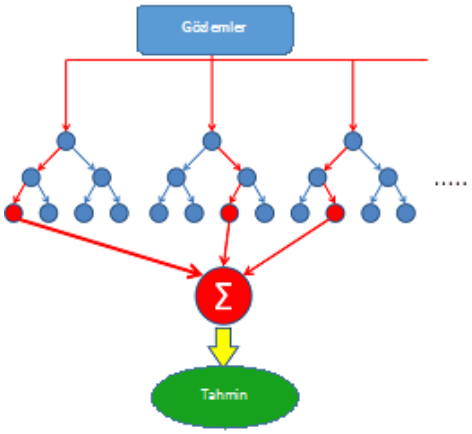
Sağlık harcamaları herhangi bir ülkede toplumun sağlık seviyesini gösteren, ülke genelinde büyüme ve kalkınmanın en temel göstergesidir (Gupta ve Mitra 2004). Bu nedenle dünya genelinde herkesin daha kaliteli sağlık hizmetlerine erişebilmesi için uluslararası kuruluşlar tarafından sağlık alanına ciddi miktarda kaynak ayrılmakta ve yatırım yapılmaktadır. Gelişmekte olan ülkeler ile başta Afrika ülkeleri olmak üzere gelişmemiş toplumları hedef alan bu politikalar sayesinde sağlığın küresel anlamda gelişmesi ve küresel kalkınma ve büyümenin sağlanması hedeflenmektedir. Bu konudaki çalışmalarda öncü kuruluşlar arasında Dünya Bankası ile Dünya Sağlık Örgütü bulunmakla birlikte bu kuruluşlar tarafından düzenli aralıklarla farklı ülkelerde sağlık sistemi performansını değerlendirmeye yönelik araştırma raporları yayınlamaktadırlar. Yayınlanan bu raporlar ve yapılan incelemelerde sağlık sisteminin temel dinamikleri, sağlıkta performansı belirleyen ve etkileyen unsurlar, sağlığı geliştirici politika ve müdahaleler üzerinde durulmaktadır. Bu politika ve öneriler sağlık sistemlerinin iyileştirilmesi için politika belirleyicilere yardımcı olan, karar vermeye yardımcı kılavuz niteliği taşımaktadır (WHO 2000). Bu sayede farklı ülkelerin sağlık sistemlerinin kapsamlı olarak incelenmesi, örnek ülke uygulamalarının model olarak alınması mümkün olabilmektedir. Küresel sağlık otoritelerinin sağlık sistemini geliştirmeye yönelik olarak belirledikleri alanlar arasında ön plana çıkan bir konu ise sağlık harcamalarıdır (Frenk 2010).

Sağlık harcamalarının seviyesi herhangi bir toplumda sağlığa yapılan yatırımlar ve ayrılan kaynakların bir belirleyicisi olarak toplum genelinde kalkınmanın bir anahtarıdır. Bu nedenle gelişmiş, gelişmekte olan ve gelişmemiş ülkelerde sağlık harcamalarının belirleyicilerini ortaya koymaya yönelik pek çok çalışmanın yapıldığı görülmektedir. Bu çalışmalarda sağlık harcamalarını belirleyen temel değişkenlerin başta gelir seviyesi olmak üzere toplam nüfus ile nüfusun yaş dağılımı, sağlık hizmetlerinin sunumu ve bu hizmetlere erişim seviyesi olduğu görülmektedir (Martin ve diğ., 2011). Bu değişkenler içerisinde gelir seviyesi gelişmişlik düzeyinin bir göstergesi olarak ön plana çıkmaktadır (Sinha ve diğ., 2016).

Sağlık harcamalarını belirlemeye yönelik araştırma modellerinde çoğunlukla regresyon modelleri kullanılmakla birlikte ülke genelinde sağlık harcamalarını belirleyen faktörleri tespit etmeye yönelik farklı yaklaşım ve modellemelerin geliştirildiği görülmektedir (Mihaylova ve diğ., 2011). Sağlık harcamalarının konu edinildiği modelleme çalışmalarında ön plana çıkan bir konu sağlık harcaması dağılımının normal dağılım özelliği göstermeyip aşırı derecede sağa çarpık olmasıdır. Sağlık harcamalarının modellendiği çalışmalarda model performansını bozucu bir etki yapan bu sorun ile başa çıkabilmek için farklı dönüşüm yaklaşımları uygulanmakta, farklı yaklaşımlar denenerek en uygun yaklaşımın tercih edilmesi tavsiye edilmektedir (Manning 1998). Bu yaklaşımlar içerisinde ön plana çıkan ve en fazla kullanılan dönüşüm logaritmik dönüşüm (Jones ve diğ., 2007; Basu ve diğ., 2004) olmakla birlikte Manning (2006) tarafından tavsiye edilen bir dönüşüm türü ise Box-Cox dönüşümüdür (Manning 2006). Box-Cox dönüşümü 1964 yılında Box-Cox tarafından geliştirilmiş (Box ve Cox 1964) bir yaklaşım olup normal dağılım özelliği göstermeyen verilere uygulanarak normalliğin sağlanmasına katkı sağlamaktadır. Manning (2006) tarafından belirtildiği üzere bu yaklaşımın sağlık harcamalarının modellendiği çalışmalarda da kullanımı, model performansının yükselmesine katkı sağlamaktadır.

Sağlık harcamalarını tahmin etmeye yönelik modelleme çalışmalarında daha çok regresyon modellerinin tercih edildiği bilinmekte olup bu modellerde sağlık harcaması değişkeni bağımlı değişken olup, bu değişkeni tahmin etmeye yönelik olarak bağımsız değişkenlerin kullanıldığı çoklu doğrusal regresyon modelleri kurulmaktadır (Mihaylova ve diğ., 2011). Örüntü tanıma, makine öğrenmesi ve veri madenciliği teknikleri sağlık ekonomisi alanında uzman araştırmacılar için nispeten yeni alanlar olmakla birlikte bu tekniklerin kullanımı sayesinde yalnızca sağlık ekonomisi alanında değil, sağlık ile ilgili klinik, politika ve uygulama seviyelerinde de pek çok soruna ilişkin çözüm önerisinde bulunmak mümkün olabilmektedir

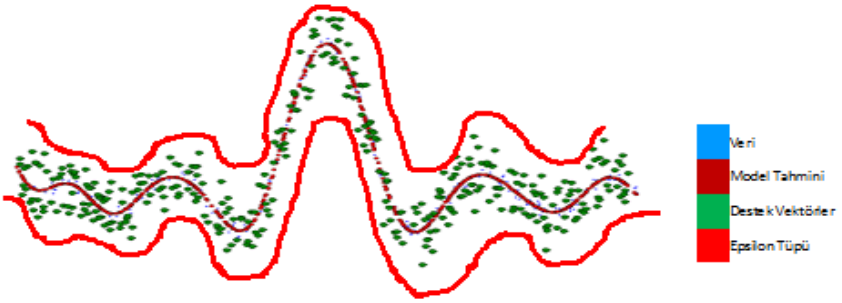
(Crown 2015). Klasik istatistiksel yaklaşımlara alternatif niteliği taşıyan bu yaklaşımların kullanılması sayesinde sağlık harcamalarının tahminine yönelik oluşturulacak modellerden daha yüksek model performansının elde edilmesi mümkün olabilmektedir. Bu alternatif yaklaşımlar arasında sınıflama ve tahmin amacıyla kullanılan Lasso Regresyon, Rastgele Ağaç (Random Forest) Regresyonu ile Destek Vektör Regresyonu ön plana çıkmaktadır. Rastgele Ağaç Regresyonu Classification and Regression Trees (CART) algoritmasından yararlanılan, maksimum seviyede benzer alt sınıflar oluşturma ilkesine dayanan, bölünme kriteri olarak Gini katsayısının kullanıldığı bir yaklaşımdır. Yöntemin temel amacı, "ağaç oluşturma" aşamasında oluşturulabilecek en fazla sayıda alt ağacı belirlemektedir. Temel amaç ise alt ağaçlar arasında bağımlı değişken ile ciddi ölçüde ilişkili olan ağaçları seçmektir. Bu model hızlı ve iyi performansa sahip bir model olması nedeniyle tavsiye edilen bir modeldir (Rodriguez ve diğ., 2015). Random Forest yönteminde CART algoritması kullanılarak ağaçlar oluşturulur ve oluşturulan ağaçlar budanmaz. CART algoritması veri setinin hangi değişkenden başlayarak dallara ayrılacağına "bilgi kazancını" kullanarak karar verir. Bu yöntemde oluşturulacak her bir ağaç için veri setinden bootstrap yöntemi ile örneklem seçilir ve seçilen verilerin 2/3'ü ağaç oluşturmak için kullanılarak bir sınıflama yapılır. Bu sınıflama bir "oy" alır ve bu algoritma "orman" içindeki tüm ağaçlardan en çok oy alanı seçerek onun sınıflamasını kullanır. Yapılan bu sınıflamada en düşük hata oranına sahip olan ağacın daha iyi bir sınıflayıcı olduğuna karar verilir. Bu algoritmanın hızlı olması genel olarak sağlık ve özellikle de genetik alanında yaygınlıkla kullanılmasına imkân vermiştir (Coşgun ve Karaağaoğlu 2011). Zaman içerisinde yüksek tahmin gücüne sahip olması nedeniyle uygulama alanı artarak, sosyal bilimler alanında, ekonomide gelir ve harcamaların modellenmesinde yararlanılan bir algoritma haline gelmiştir (Einav ve Levin 2014). Şekil 1'de ağaç oluşturma süreci görülen Rastgele Ağaç regresyonu ile çok sayıda ağaç türetilerek en optimal tahmin sonucunu belirlemek mümkün olabilmektedir.



Şekil 1:
Rastgele Ağaç Regresyonu Yönteminde Ağaçların Oluşumu
Kaynak: *Hastie T. Tibshirani R. Friedman J. (2009) "Random Forests", p.587-613. The Elements of Statistical Learning Data Mining, Inference and Prediction.*

Doğrusallaştırılamayan modellerin doğrusallaştırılmasında faydalı bir diğer makine öğrenmesi yaklaşımı ise Destek Vektör Makinesi (Support Vector Machine-SVM) regresyonudur. Destek Vektör Makineleri ile ilgili çalışmalar daha çok sınıflama amacıyla uygulanmış olup, bu yöntemin regresyon için uyarlaması Vapnik (1997) tarafından yapılmıştır. Destek Vektör Makinesi yönteminde doğrusal olarak sınıflanabilen verileri birbirinden ayırt etmek için olası pek çok doğrusal fonksiyon arasından, marjini en büyük olan belirlenmektedir. Sınıflandırılacak örnekler doğrusal bir düzlemle ayrıştırılabilecek düzeyde olmadığında, bu yöntemde belirlenen bir Kernel fonksiyonu yardımı ile daha yüksek boyutlu bir uzaya aktarılması mümkün olmaktadır. Bu şekilde marjini en yüksek olan hiper düzlemler bulunur.

Sonuç olarak veriler bu ayırt edici hiper düzleme göre sınıflara atanır (Coşgun ve Karaağaoğlu 2011; Yılmaz 2016). Şekil 2'de görülen Destek Vektör Regresyonunda bir grup veriyi uzayda en fazla epsilon kadar hata ile tahmin eden, mümkün olan en doğrusal fonksiyonu bulmak amaçlanmaktadır. Burada bahsedilen epsilon değeri regresyon modelinin duyarlılığını belirlemektedir. Destek Vektör Regresyonu ile en fazla ± epsilon aralığında kalan alan epsilon, epsilon dışında kalan noktalar ise "destek vektörler" olarak isimlendirilmektedir. Bu yöntemde temel amaç model tahminini epsilon tüpü aralığına sığdırmak ve bu sayede en yüksek regresyon modeli performansına ulaşmaktır (Kazem ve diğ., 2013).



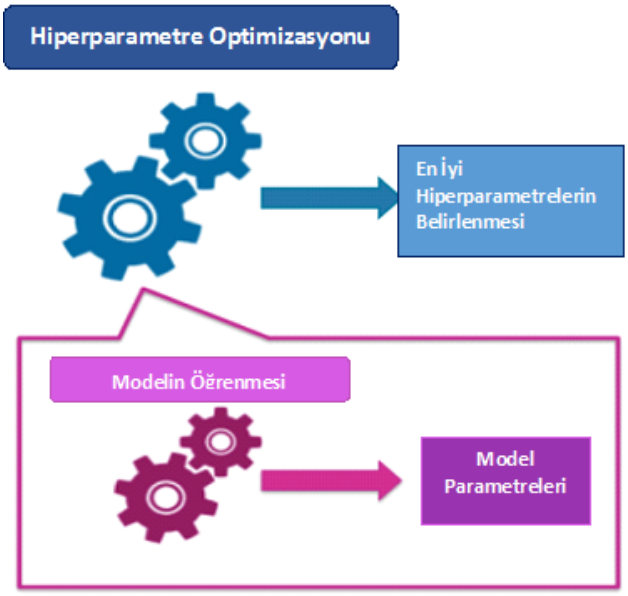
Şekil 2:
Destek Vektör Regresyonu
Kaynak: *Cristianini N. Shawe-Taylor J. (2000) An Introduction to Support Vector Machines and other Kernel Based Learning Methods, Cambridge University Press, p.93-122.*

Neden-sonuç ilişkisi bulmaya yönelik araştırmalarda, özellikle yüksek boyutlu veriler ile karşılaşılması durumunda standart regresyon analizi ile elde edilen tahminlerden istenilen performansın elde edilememesi durumu ile karşılaşılmaktadır. Bu problemin çözümü için cezalı regresyon yöntemleri (penalized regression) içerisinde yer alan Lasso Regresyon yöntemi gibi tekniklerden faydalanılmaktadır (Elasan ve diğ. 2016). Lasso Regresyon tahminleri dengelediği için bir tür "büzülme modeli (Shrinkage)" olarak isimlendirilmektedir (Frank ve Friedman 1993). Lasso regresyonda ceza ölçütü olarak "Manhattan Uzaklığı" kullanılmaktadır. Bu yöntemde ceza parametresi, katsayıların mutlak değerlerinin toplamıdır. Lasso'da düşük lamda (λ) parametre değerlerinde en küçük kareler tekniği ile benzer sonuçlar elde edilmekte iken, lamda (λ) parametresi yükseldikçe Shrinkage devreye girmekte ve tahmin performansı yükselmektedir (Jaggi 2014). Lasso regresyon yöntemini performans bakımından diğer regresyon yöntemleri ile karşılaştırmalı olarak inceleyen araştırmalar sonucunda gözlem sayısının açıklayıcı değişken sayısından daha yüksek olması durumunda bu yöntemin daha yüksek performans sergilediği bulunmuştur (Tibshirani 1996).

Literatürde Lasso Regresyon modeli yüksek boyutlu verilerin analizinde tahmin performansının yükselmesine katkı sağlayan yöntemlerden birisi olarak ön plana çıkmaktadır. Rastgele Ağaç Regresyonu ile Destek Vektör Regresyonu ise Lasso'ya göre sağlık ekonomisi alanında daha yeni uygulama alanı bulmuş konular arasında sayılmaktadır. Ekonomik analizlerde karşılaşılan temel güçlük basit analiz varsayımlarının sağlanmaması ve bu nedenle analiz öncesi hazırlık aşamalarına ihtiyaç duyulmasıdır. Bu nedenle cezalı regresyon yöntemlerinin kullanılması tavsiye edilmektedir (Belloni ve diğ. 2012). Özellikle aykırı gözlemler ile karşılaşılması durumunda Rastgele Ağaç algoritmasının etkili bir yöntem olduğu vurgulanmaktadır (Gislason ve diğ. 2006). Brieman (2001) Rastgele Ağaç algoritmasının tahmin doğruluğu bakımından oldukça popüler bir yöntem olduğunu belirtmiştir. Suthaharan (2016) ise Destek Vektör Makinesinin sınıflama ve regresyon performansı bakımından diğer pek çok yönteme göre üstün performans sergilediğini belirtmiştir. Hassan ve diğ. (2013) tarafından yapılan ve biyoproses modellemesinde Lasso Regresyon, Rastgele Ağaç Regresyonu ve Klasik Çoklu Regresyon modellerinin karşılaştırmalı olarak incelendiği bir çalışmada ise

Lasso ve Rastgele Ağaç Regresyonu yöntemlerinin klasik regresyon yöntemlerine göre daha yüksek performans sergiledikleri ortaya konulmuştur (Hassan ve diğ. 2013).

Literatürde Lasso Regresyon, Rastgele Ağaç Regresyonu ile Destek Vektör Regresyonu yaklaşımlarına ait performans sonuçlarının optimizasyonu için kullanılabilecek farklı parametre optimizasyon teknikleri bulunmaktadır. Hiperparametre optimizasyon teknikleri olarak isimlendirilen bu teknikler kullanılarak makine öğrenmesi yöntemleri için optimal performans sonuçlarının elde edilmesi mümkün olmakta, bu sayede karşılaştırmalar yapmak imkanı doğmaktadır (Duan ve diğ., 2003). Şekil 3'de görüldüğü üzere hiperparametre optimizasyonu, makine öğrenmesi yolu ile en uygun model parametrelerinin belirlendiği ve en iyi hiperparametre değerlerinin tespit edildiği bir süreçtir.



Şekil 3:
Makine Öğrenmesi Yöntemleri İçin Hiperparametre Optimizasyonu
Kaynak: *Zheng A. (2015) Evaluating Machine Learning Models A Beginner's Guide to Key Concepts and Pitfalls, O'Reilly.*

Bergstra ve Bengio (2012) tarafından belirtildiği üzere hiperparametre optimizasyon teknikleri arasında grid arama yöntemleri ön plana çıkmaktadır. Bu teknikler içerisinde model performansına ek katkı sağlayan bir hiperparametre değeri ise "k" parça çapraz geçerliliktir (Tsamardinos ve diğ., 2015). Çapraz geçerlilikte veri seti rastgele "k" parçaya ayrılmakta, her seferinde 1 parçası dışarıda bırakılarak "k-1" olarak ifade edilen ve verinin kalan kısmını temsil eden parça ile model bulunarak dışarıda bırakılan parça ile test işlemi gerçekleştirilmektedir. Bu işlem tüm parçaların tek tek dışarıda tutulması sağlanana kadar devam etmekte, sonuçta ortalama bir performans skoru elde edilmektedir. Hem sınıflama hem de regresyon yöntemlerinde, ayrıca farklı büyüklükteki veri setlerinde kullanışlı olduğu görülen bu teknik sayesinde model sonuçlarının optimizasyonu mümkün olmaktadır (Kohavi 1995).

Lasso Regresyonda performans ölçüsü olarak lamda (λ) parametresi belirlenmekte ve daha yüksek lamda (λ) değerlerinin model performansının optimizasyonuna katkı sağladığı belirtilmektedir (Jaggi 2014). Farklı algoritmalar (ID.3, C4.5, CART) kullanılarak geliştirilen karar ağaçları için hiperparametre değeri olarak ağacın derinliği (tree depth) ve yaprak sayısı (number of leaves) esas alınabilmektedir. Çok sayıda karar ağacının bir araya gelmesi ile elde edilen Rastgele Ağaç yöntemi için optimal ağaç sayısının belirlenmesi ise bir tür düzenleyici hiperparametre (regularization hyperparameter) optimizasyonu olarak isimlendirilmektedir. Farklı fonksiyonlar (doğrusal, kernel, polinomial, radyal, sigmoid) kullanılarak performe edilen

ve doğrusallaştırılamayan modellerde iyi performans sergilediği görülen Destek Vektör Makinesi regresyon performansının optimizasyonu amacıyla kernel parametresi (γ), ceza faktörü (C), epsilon değeri (ε) gibi hiperparametrelere ait değerlerdeki değişim incelenebilmektedir (Bergstra ve Bengio 2012). Bu hiperparametrelerden birisi olan epsilon değeri (ε) için bu değerin yükselmesi Destek Vektör Regresyonda tahmin doğruluğunu düşürmekte, destek vektörlerin sayısında azalmaya neden olmakta, veriye daha fazla basınç uygulayarak model performansını olumsuz olarak etkilemektedir. Literatürde optimal epsilon (ε) parametre değerinin ne olması gerektiği konusunda kesin kabul gören bir yaklaşım bulunmamakla brilikte Mattera ve Haykin (1999) tarafından büyük ya da gürültülü veriler için büyük epsilon (ε) parametre değerlerinin kullanılabileceği belirtilmektedir.

Makine öğrenmesi regresyon yöntemlerinin performans değerlendirmelerinde kullanılan temel performans göstergeleri arasında Açıklayıcılık Katsayısı (R^2), Ortalama Hata Kare Kökü (Root Mean Square Error-RMSE), Ortalama Mutlak Hata (Mean Absolute Error-MAE) bulunmaktadır. Bu performans ölçüleri içerisinde Açıklayıcılık Katsayısının (R^2) modelin veriye uyumunu gösteren bir ölçü olduğu ve 0.70 ve üzerindeki değerlerin daha iyi performans sonuçlarına işaret ettiği bilinmektedir (Alpar 2011). RMSE ve MAE ise birer hata ölçüsü olduklarından dolayı düşük değerlerin daha iyi performansa işaret ettiği ölçülerdir (Wang ve Xu 2004). Örneğin RMSE sıfıra eşit olması durumunda mükemmel performansa işaret etmektedir.

Sağlık harcaması değişkeni aşırı derecede sağa çarpıklık özelliği göstermekte olup, bu değişkenin modellendiği çalışmalarda aşırı uyum sorunu ile karşılaşılmaktadır. Bu durum model performansını olumsuz olarak etkilemektedir. Daha yüksek model performansı elde edebilmek amacıyla literatürde sağlık harcamalarının modellendiği çalışmalar için makine öğrenmesi ve veri madenciliği tekniklerinin kullanılmasının sağlayacağı avantajlar üzerinde durulmaktadır (Collins 2016). Bu nedenle bu çalışmada sağlık harcaması değişkeninin tahminine yönelik olarak makine öğrenmesi regresyon yöntemleri kullanılarak bir regresyon modeli oluşturulmuştur. Oluşturulan regresyon modelinin performansı farklı hiperparametre optimizasyon yöntemleri kullanılarak incelenmiş, performans sonuçları karşılaştırılmıştır.

2. YÖNTEM

2.1. Veri Seti

Araştırmada kullanılan veri seti 2013 yılı itibariyle Dünya Bankası'na üye olan 214 ülkeye aittir. Oluşturulan regresyon modelinde bağımlı değişken olarak kişi başı toplam sağlık harcaması değişkeni seçilmiş olup, bağımsız değişkenler; gelir grubu, coğrafi bölge, 65 yaş üzeri toplam nüfus, doğuşta beklenen yaşam süresi ile toplam nüfus değişkenleridir. Tüm veriler 2013 yılına ait olup, çalışmada kullanılan değişkenler içerisinde bağımsız ve kategorik formda bulunan değişkenler arasında; gelir grubu (Dünya Bankası tarafından belirlenen ülke gelir grupları $) ve coğrafi bölge (Dünya Bankası tarafından belirlenen coğrafi bölge grupları) bulunmaktadır. Çalışmada bağımsız değişken olarak kullanılan ve sayısal formda bulunan değişkenler; 65 yaş ve üzeri nüfus (toplam nüfus içerisinde 65 yaş ve üzeri nüfus yüzdesi (%)), doğuşta beklenen yaşam süresi (ülke genelinde toplam nüfus için doğuşta beklenen yaşam süresi (yıl)), toplam nüfus (ülke genelinde toplam nüfus) olarak belirlenmiştir. Son olarak çalışmada bağımlı değişken olarak yer alan ve sayısal formda bulunan bir değişken olan kişi başı sağlık harcaması değişkeni özel ve kamu sağlık harcaması toplamının toplam nüfusa oranını ($) göstermektedir.

2.2. Verilerin Analizi

Verilerin analizine tanımlayıcı bilgilerin elde edilmesi ile başlanmıştır. Sayısal ve kategorik formda bulunan değişkenlere ait tanımlayıcı bilgilerin sunulmasının ardından çoklu regresyon modellerinin temel varsayımı olan bağımsız değişkenler arasındaki ilişkinin çoklu bağlantı sorunu bakımından incelenmesi gerçekleştirilmiştir (Alpar 2011). Bilindiği gibi çoklu regresyon modellerinde bağımsız değişkenler arasında çoklu bağlantı sorunun varlığı aşırı uyum sorununu

beraberinde getirmekte, doğrusallıktan ayrılmaya neden olmakta ve model performansını bozucu bir etki yaratmaktadır (Hawkins 2004). Sağlık harcamalarının incelendiği regresyon modellerinde vurgulanan bir diğer nokta ise sağlık harcaması değişkeninin aşırı derecede sağa çarpıklık özelliği ile başa çıkabilmek için uygun ayarlama yöntemleri kullanılarak dağılımın normal dağılıma yakın hale getirilmesinin sağlanmasıdır. Bu amaçla kullanılan yaklaşımlardan birisi Box-Cox dönüşümüdür (Manning 2006). Bu çalışmada kişi başı sağlık harcaması değişkeninin tahmin edilmesi amacıyla oluşturulan modelde bağımlı değişken olarak kullanılan kişi başı sağlık harcaması değişkenine Box-Cox dönüşümü uygulanmıştır. Dönüşüm sonrasında normal dağılıma yakın hale geldiği görülen kişi başı sağlık harcaması değişkeninin tahminine yönelik olarak kullanılacak makine öğrenmesi yaklaşımları; Lasso Regresyon, Rastgele Ağaç Regresyonu ile Destek Vektör Regresyonu olarak belirlenmiştir. Bu üç yönteme ait performans sonuçları farklı hiperparametre değerleri belirlenerek optimize edilmiştir. Şekil 4'de veri analizinin aşamaları özetlenmiştir.

Şekil 4:
Veri Analizinin Aşamaları

Verilerin analizi için bir veri madenciliği yazılım programı olan Orange kullanılmıştır. Aynı zamanda farklı makine öğrenmesi yöntemleri arasındaki performans farklılıklarını ortaya koyabilmek amacıyla çapraz geçerlilikte farklı "k" değerleri belirlenmiş, hiperparemetre değerleri arasındaki karşılaştırmalar bu farklı "k" değerlerine göre kaydedilen sonuçlar üzerinden yapılmıştır. Farklı makine öğrenmesi yöntemlerinin hiperparametre optimizasyonu kullanılarak elde edilen performans sonuçları arasındaki farklılıklar Kruskall Wallis Varyans Analizi ile Anova (F) testleri kullanılarak incelenmiştir.

3. BULGULAR

3.1. Tanımlayıcı Bilgiler

Bu çalışmada yer verilen bağımlı ve bağımsız değişkenlere ait tanımlayıcı bilgiler incelendiğinde, sağlık harcaması değişkeninin tahminine yönelik olarak oluşturulan modelde kullanılan bağımlı değişken olan kişi başı sağlık harcaması ortalaması 1021.23 (±1769.73), bağımsız değişken olan doğuşta beklenen yaşam süresi ortalama 71.25 (±8.23), 65 yaş ve üzeri nüfus yüzdesi ortalama 7.86 (±5.36), ülke genelinde toplam nüfus ortalaması 33.416.716,63 (±131.968.340,41)'dir. Çalışmada kullanılan ve kategorik formda bulunan değişkenler arasında yer alan gelir grubu değişkeni için Dünya Bankası tarafından yapılan ülke gelir grubu sınıflandırması esas alınmıştır. Buna göre 1.045$ ve daha az gelire sahip olan ve düşük gelir grubu olarak yer alan 31 (%14.5) ülke, 1.046$≤ Düşük Orta Gelir (DOG) <4.125$ aralığında bulunan ve yüksek orta gelir grubunda olarak sınıflandırılan 51 (%23,8) ülke, 4.125$≤ Yüksek Orta Gelir (YOG) <12.746$ aralığında bulunan ve yüksek orta gelir grubunda bulunan 53

(%24.8), son olarak 12.746$ ve üzerinde gelire sahip olan ve yüksek gelir grubunda olduğu belirtilen 79 (%36.9) ülke bulunmaktadır. Bu tanımlayıcı bilgilere göre ülkelerin en fazla yüksek gelir grubunda bulundukları söylenebilmektedir. Çalışmada kullanılan ve kategorik formda bulunan bir diğer değişken olan coğrafi bölge değişkenine ait tanımlayıcı bilgiler incelendiğinde ise Doğu Asya ve Pasifik'te yer alan 36 (%16.8), Avrupa ve Merkezi Asya'da 57 (%26.6), Latin Amerika ve Karayipler'de 41 (%19.2), Orta ve Kuzey Afrika'da 21 (%9.8), Kuzey Amerika'da 3 (%1.4), Güney Asya'da 8 (%3.7) ve Sahra Altı Afrika'da yer alan 48 (%22.4) ülke olduğu görülmektedir. Bu bilgilere göre Dünya Bankası'na üye olan ülkelerin daha çok Avrupa ve Merkezi Asya'da yer aldıkları söylenebilmektedir.

3.2. Bağımsız Değişkenler Arasındaki İlişkinin Çoklu Bağlantı Sorunu Bakımından İncelenmesi

Bu çalışmada bağımsız değişkenler arasındaki ilişkinin çoklu bağlantı sorununun varlığı bakımından incelenmesinde, ele alınan değişkenler arasında çarpık dağılıma sahip değişkenler bulunduğundan dolayı, bir tür nonparametrik ilişki katsayısı olan Spearman korelasyon katsayısı (r_s) kullanılmıştır. Buna göre değişkenler arasındaki ilişkide çoklu bağlantı sorununa işaret edecek derecede yüksek korelasyona (r_s<0.70) rastlanmadığı görülmüş, seçilen tüm bağımsız değişkenlerin modele katılmasında bir engel olmadığı kanaatine varılmıştır.

3.3. Sağlık Harcaması Değişkenine Ait Dağılımın Normalleştirilmesi

Sağlık harcaması değişkenine ait dağılımın normal dağılıma yakın hale getirilmesi için bu değişene ait dağılıma Box-Cox dönüşümü uygulanmıştır. Box-Cox dönüşümü değişen dağılımlarının normal dağılıma uygun olmaması durumunda sıklıkla kullanılan, varyansın bağımlı değişkenin değerlerine paralel olarak arttığı durumlarda pozitif değerli değişkenler için tercih edilen logaritmik dönüşümün özel bir türü olarak tarif edilmektedir (Manning 2006). Uygulanan dönüşüm sonrasında Box-Cox dönüşümünün kişi başı sağlık harcaması dağılımını normal dağılıma yaklaştırdığı gözlemlenmiştir.

3.4. Makine Öğrenmesi Regresyon Yöntemleri İçin Hiperparametrelerin Belirlenmesi

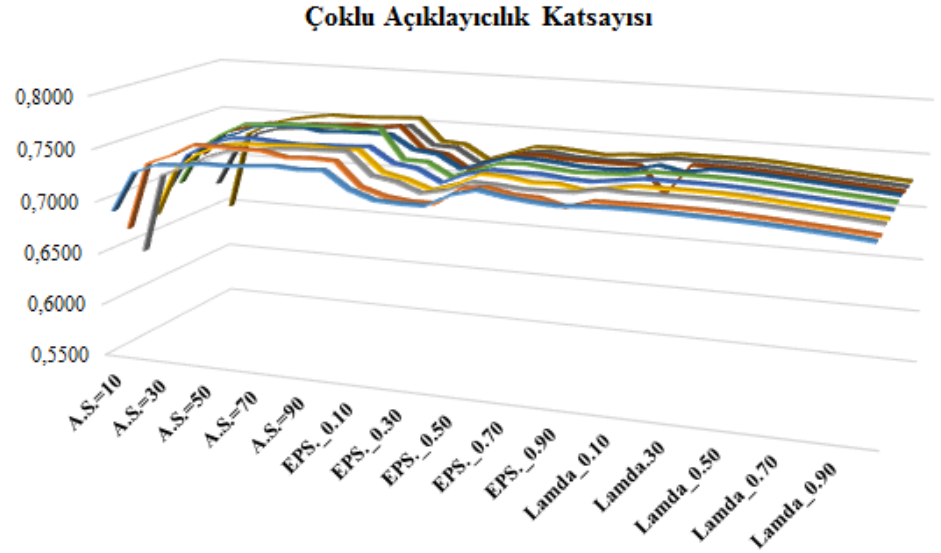
Bu çalışmada makine öğrenmesi yöntemleri arasında sayılan ve bağımlı değişkene ilişkin değerlerin tahmininde klasik regresyon yaklaşımları karşısında iyi bir alternatif olduğu savunulan Lasso Regresyon, Rastgele Ağaç Regresyonu ile Destek Vektör Makinesi Regresyon yöntemleri (Witten ve Frank 2005) kullanılmıştır. Hiperparametre değerlerinin belirlenmesinde Lasso Regresyon için lamda (λ) değeri (Jaggi 2014), Rastgele Ağaç Regresyonu için bir düzenleyici hiperparametre (regularization hyperparameter) türü olarak kabul edilen ağaç sayısı (Liaw ve Wiener 2002), Destek Vektör Makinesi regresyon yöntemi için ise epsilon değeri (ε), kullanılmıştır (Cherkassy ve Ma 2004; Schölkopf ve Smola 2002; Kavaklıoğlu 2011; Jaggi 2014). Ek 1, 2 ve 3'de bu çalışmada bağımlı değişken olarak kullanılan kişi başı sağlık harcaması değişkenine Box-Cox dönüşümü uygulandıktan sonra Lasso Regresyon, Rastgele Ağaç Regresyonu ve Destek Vektör Makinesi Regresyon yöntemleri için farklı hiperparametre değerlerinin belirlenmesi durumunda elde edilen performans sonuçlarına yer verilmiştir. Bu incelemede makine öğrenmesi yöntemleri için genel bir hiperparametre ölçüsü olarak kabul edilen "k" parça çapraz geçerlilik uygulanmıştır. Buna göre 5 ile 50 arasında değişmek üzere farklı sayılarda "k" parametreleri belirlenmiş olup, performans sonuçları sunulmuştur. Sonuçlar R^2, RMSE ve MAE değerleri üzerinden değerlendirilmiştir.

3.5. Makine Öğrenmesi Regresyon Yöntemlerinin Farklı Hiperparametrelere Göre Belirlenen Tahmin Performanslarının Grafiksel Olarak İncelenmesi

Bu çalışmada bağımlı değişken olarak yer alan sağlık harcaması değişkenine Box-Cox dönüşümünün uygulanması ve farklı "k" parametrelerinin belirlenmesi durumunda Lasso Regresyon, Rastgele Ağaç Regresyonu ve Destek Vektör Makinesi Regresyon yöntemlerine ait farklı performans sonuçları Tablo 1'de yüzey grafikleri üzerinde gösterilmiştir. Buna göre

makine öğrenmesi regresyon yöntemlerinin performans sonuçları çoklu açıklayıcılık katsayısı (R^2) bakımından değerlendirildiğinde ve bu değerin yüksek olması durumunun daha iyi performansa işaret ettiği göz önünde bulundurulduğunda, Tablo 1'de "k" parametresinin yükselmesine paralel olarak, daha yüksek bir lamda (λ) değerinin belirlenmesi durumunda (λ>0.70), daha yüksek açıklayıcılık katsayısı değerlerinin (R^2=0.7200-0.7400) elde edildiği görülmektedir. Diğer taraftan Rastgele Ağaç Regresyonu yönteminde daha fazla sayıda ağaç türetilmesi durumunda (Ağaç sayısı>40) daha yüksek açıklayıcılık katsayısı (R^2) değerlerinin (R^2=0.7000-0.8000) elde edildiği görülmektedir. Destek Vektör Makinesi Regresyon yöntemi için değerlendirildiğinde ise daha yüksek epsilon değerlerinde (ε>0.50) daha yüksek açıklayıcılık katsayısı değerlerinin (R^2=0.7300-0.7400) elde edildiği görülmektedir. Başka bir deyişle parametre değerlerinin yükselmesi durumunda daha iyi tahmin performanslarına ulaşılması durumu dikkat çekmektedir. Tablo 1'de yüzey grafiklerinde regresyon modeli performans ölçüsü olarak kullanılan diğer ölçüler ise ortalama hata karekökü olarak bilinen Root Mean Squared Error (RMSE) ile ortalama mutlak hata olarak bilinen Mean Absolute Error (MAE)'dır. RMSE ve MAE birer hata ölçüsü olduklarından dolayı bu değerlerin görece daha düşük olması daha iyi performans sonuçlarına işaret etmektedir. Bu açıdan makine öğrenmesi regresyon sonuçları incelendiğinde Rastgele Ağaç Regresyonu yöntemi kullanılarak elde edilen RMSE ve MAE değerlerinin, diğer yöntemlere göre daha düşük olduğu görülmektedir. Bunun yanı sıra Rastgele Ağaç Regresyonu için ağaç sayısındaki artışın performans sonuçlarında çok belirgin bir farklılık ortaya koymadığı görülmektedir.

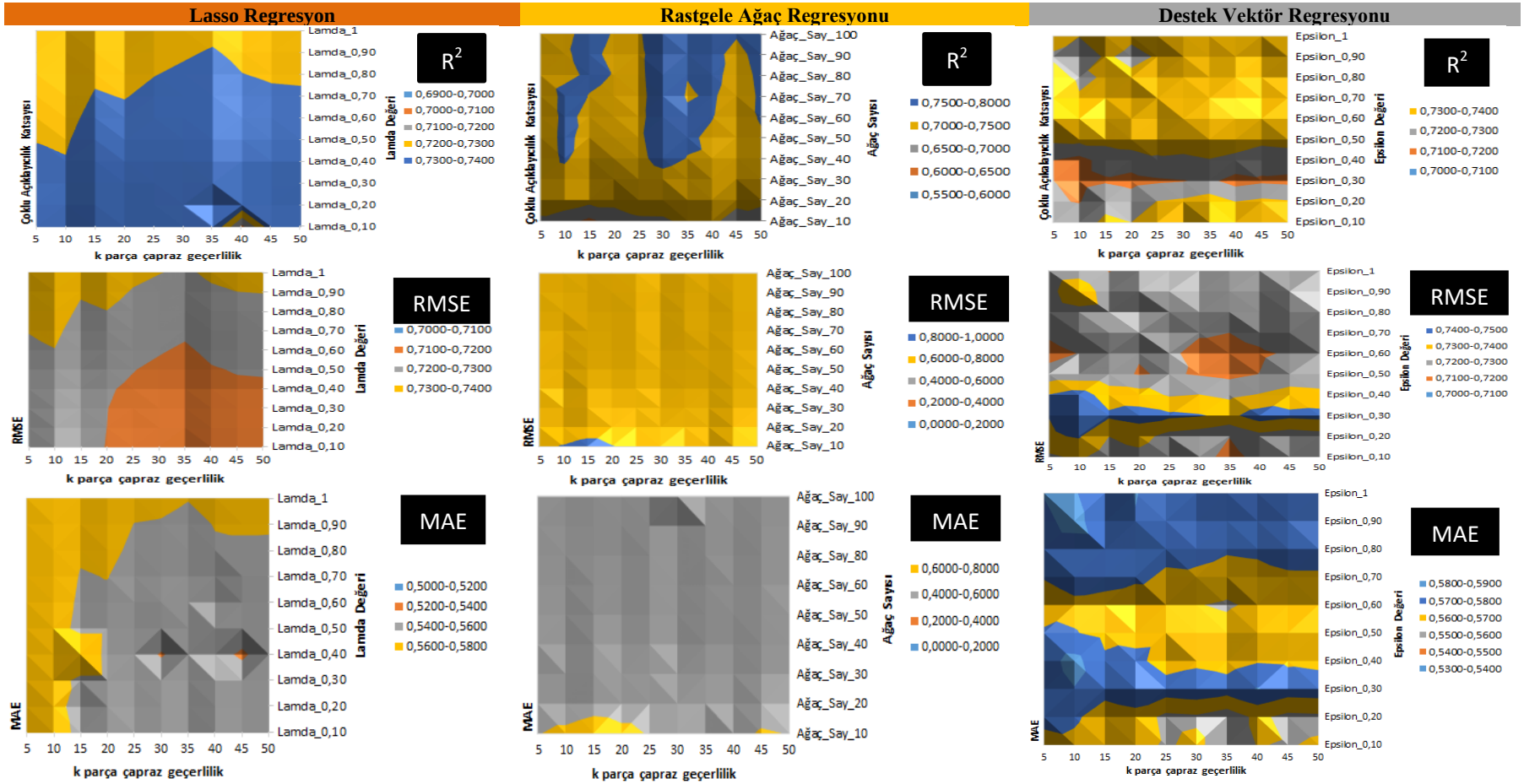
Grafik 1'de Lasso Regresyon yönteminde Lamda değeri 0.10 ile 1 arasında belirlendiğinde, Rastgele Ağaç Regresyonu yönteminde 10 ile 100 arasında değişen sayıda ağaç türetildiğinde, Destek Vektör Regresyonda ise 0.10 ile 1 arasında değişmek üzere farklı epsilon parametreleri belirlendiğinde çoklu açıklayıcılık katsayısı (R^2) bakımından elde edilen sonuçlar karşılaştırmalı olarak çubuk grafiği üzerinde gösterilmiştir. Buna göre daha yüksek R^2 değerlerinin daha iyi performansa işaret ettiği göz önünde bulundurulduğunda, Rastgele Ağaç Regresyonu yönteminden elde edilen R^2 değerlerinin diğer yöntemler ile karşılaştırıldığında nispeten daha yüksek olduğu ve R^2 bakımından Rastgele Ağaç Regresyonu yönteminin kişi başı sağlık harcamasını tahmin etmede daha iyi olduğu söylenebilmektedir.



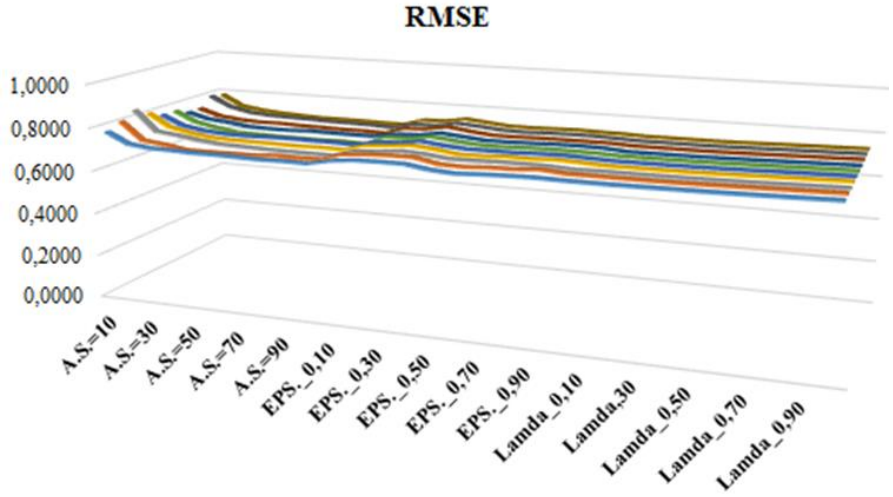
Grafik 1:
Farklı Hiperparametre Değerlerinin Belirlenmesi Durumunda Lasso Regresyon, Rastgele Ağaç Regresyonu ile Destek Vektör Regresyon Performanslarının Çoklu Açıklayıcılık Katsayısı (R^2) Bakımından Karşılaştırılması

Kısaltmalar: A.S.: Ağaç Sayısı, **EPS.:** Epsilon değeri

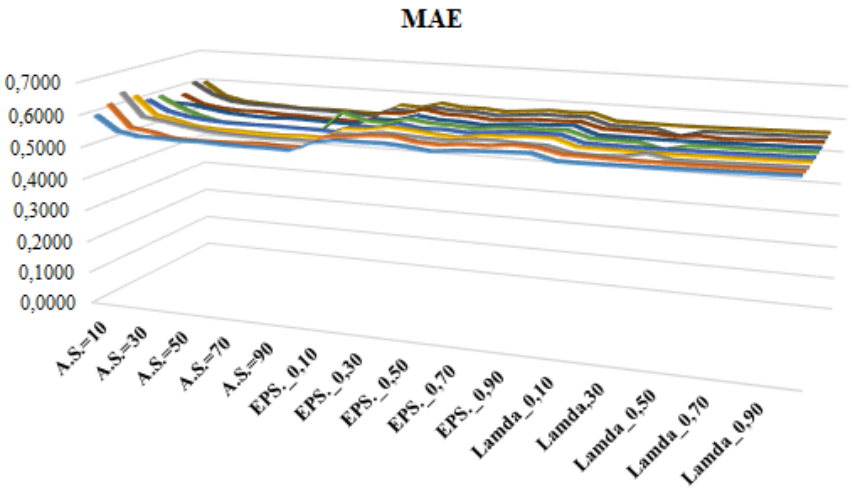
Tablo 1. Lasso Regresyon, Rastgele Ağaç Regresyonu ile Destek Vektör Regresyonu Performanslarının Çoklu Açıklayıcılık Katsayısı (R^2), RMSE ve MAE Bakımından İncelenmesi



Grafik 2 ve 3'de sırasıyla Lasso Regresyon yönteminde Lamda değeri 0.10 ile 1 arasında belirlendiğinde, Rastgele Ağaç yönteminde 10 ile 100 arasında ağaç türetildiğinde, Destek Vektör Regresyonunda ise 0.10 ile 1 arasında değişmek üzere farklı epsilon parametreleri belirlendiğinde elde edilen RMSE ve MAE değerlerine ait performans sonuçları görülebilmektedir. Buna göre daha düşük RMSE ve MAE değerlerinin daha iyi performansa işaret ettiği göz önünde bulundurulduğunda, Rastgele Ağaç Regresyonuna ait performans sonuçlarının RMSE ve MAE bakımından nispeten daha düşük değerlere işaret ettiği, başka bir ifade ile daha iyi performans sergilediği söylenebilmektedir.



Grafik 2:
Farklı Hiperparametre Değerlerinin Belirlenmesi Durumunda Lasso Regresyon, Rastgele Ağaç Regresyonu ile Destek Vektör Regresyon Performanslarının RMSE Bakımından Karşılaştırılması
Kısaltmalar: RMSE: Root Mean Square Error, **A.S.:** Ağaç Sayısı, **EPS.:** Epsilon değeri



Grafik 3:
Farklı Hiperparametre Değerlerinin Belirlenmesi Durumunda Lasso Regresyon, Rastgele Ağaç Regresyonu ile Destek Vektör Regresyon Performanslarının MAE Bakımından Karşılaştırılması
Kısaltmalar: MAE: Mean Absolute Error, **A.S.:** Ağaç Sayısı, **EPS.:** Epsilon değeri

3.6. Farklı Hiperparametre Değerlerine Göre Makine Öğrenmesi Regresyon Yöntemlerinin Tahmin Performansları Arasındaki Farkın İstatistiksel Olarak Test Edilmesi

Çalışmanın bu bölümünde farklı makine öğrenmesi yöntemleri için farklı hiperparametre değerlerinin belirlenmesi durumunda bunun her bir yöntem için ayrı ayrı regresyon modeli performans ölçüleri bakımından istatistiksel olarak anlamlı bir farklılığa işaret edip etmediği test edilmiştir. Bu amaçla Lasso Regresyon, Rastgele Ağaç Regresyonu ile Destek Vektör Regresyon yöntemleri için farklı hiperparametre değerlerinden elde edilen regresyon performans ölçüleri arasında istatistiksel olarak anlamlı bir farklılık bulunup bulunmadığı bir nonparametrik test türü olan Kruskall Wallis Varyans analizi ile test edilmiştir. Bir sonraki aşamada ise farklı hiperparametre değerleri kullanılarak Lasso Regresyon, Rastgele Ağaç Regresyonu ile Destek Vektör Regresyonu yöntemlerinden elde edilen genel performans sonuçları arasındaki farklılık, Anova (F) testi kullanılarak incelenmiştir.

Tablo 2'de Lasso Regresyon, Rastgele Ağaç Regresyonu ile Destek Vektör Regresyonu yöntemleri kullanıldığında farklı hiperparametre değerlerinden elde edilen tanımlayıcı bilgilere yer verilmiştir. Buna göre çoklu açıklayıcılık katsayısı (R^2) için daha büyük değerlerin daha iyi performans sonucu anlamına geldiği göz önünde bulundurulduğunda; Lasso Regresyon için lamda (λ) değeri 0.30 olarak belirlendiğinde 0.7360 (min. 0.7306, mak. 0.7393) ile R^2 için en büyük ortanca değer elde edilmektedir. Bunun yanı sıra, Rastgele Ağaç Regresyonu yönteminde 100 ağaç türetildiğinde 0.7509 (min.0.7447, mak.0.7615), Destek Vektör Regresyonda epsilon parametresi 0.60 olarak belirlendiğinde 0.7348 (min.0.7322, mak. 0.7378) ile en büyük ortanca değer elde edilmektedir. Tanımlayıcı bilgiler ortalama hata karekökü (RMSE) ve ortalama mutlak hata (MAE) bakımından incelendiğinde ve bu performans ölçülerine ilişkin daha düşük değerlerin daha iyi performans anlamına geldiği göz önünde bulundurulduğunda; Lasso Regresyon için lamda (λ) değeri 0.10 olarak belirlendiğinde 0.7185 (min. 0.7135, mak. 7265), Rastgele Ağaç Regresyonu yöntemine göre 100 ağaç türetildiğinde 0.6981 (min.0.6832, mak.0.7068), Destek Vektör Regresyonda epsilon parametresi 0.60 olarak belirlendiğinde 0.7204 (min. 0.7163, mak. 0.7239) RMSE için en küçük ortanca değerler elde edilmektedir. Elde edilen sonuçlar ortalama mutlak hata (MAE) bakımından incelendiğinde ise Lasso Regresyon için lamda (λ) değeri 0.40 olarak belirlendiğinde 0.5539 (min.0.5360, mak. 0.5710), Rastgele Ağaç Regresyonunda 80 ağaç türetildiğinde 0.5310 (min.0.5212, mak.0.5387), Destek Vektör Regresyonda epsilon parametresi 0.20 olarak belirlendiğinde 0.5603 (min.0.5541, mak. 0.5751) en küçük ortanca değerler elde edilmektedir (Tablo 2). Tablo 3'de Lasso Regresyon, Rastgele Ağaç Regresyonu ve Destek Vektör Regresyon için farklı hiperparametre değerleri kullanılarak elde edilen regresyon performans sonuçları incelendiğinde Kruskall Wallis varyans analizi sonuçlarına göre üç yöntem için de R^2, RMSE ve MAE değerleri arasındaki farklılığın istatistiksel olarak anlamlı olduğu görülmektedir. Bu farklılık her bir performans ölçüsü için ayrı ayrı ele alındığında ise Lasso Regresyon yönteminde 0.10 ile 1 arasında değişmek üzere farklı lamda (λ) değerleri belirlendiğinde elde edilen R^2 değerlerine ait sıra ortalamaları arasındaki farkın istatistiksel olarak anlamlı olduğu (X^2=62.751, $p<0.001$), Rastgele Ağaç yönteminde 10 ile 100 arasında değişmek üzere farklı sayılarda ağaç türetildiğinde elde edilen R^2 değerlerine ait sıra ortalamaları arasındaki farkın istatistiksel olarak anlamlı olduğu (X^2=65.543, $p<0.001$) görülmektedir. Benzer şekilde Destek Vektör Regresyon bakımından farklı hiperparametre değerleri arasında R^2 bakımından gözlemlenen farklılığın (X^2=66.667, $p<0.001$) da anlamlı olduğu görülmektedir.

Tablo 2. Lasso Regresyon, Rastgele Ağaç Regresyonu ve Destek Vektör Regresyonu İçin Farklı Hiperparametre Değerlerine Ait Tanımlayıcı Bilgiler

R^2											
Lasso Regresyon	**Min.**	**Mak.**	**Ort.**	**Rastgele Ağaç Regresyonu**	**Min.**	**Mak.**	**Ort.**	**Destek Vektör Regresyonu**	**Min.**	**Mak.**	**Ort.**
Lamda _0.10	0.7073	0.7397	**0.7356**	**A.S._10**	0.6398	0.7125	**0.6835**	**Eps._0.10**	0.7263	0.7374	**0.7324**
Lamda _0.20	0.7307	0.7380	**0.7355**	**A.S._20**	0.7157	0.7359	**0.7278**	**Eps._0.20**	0.7194	0.7349	**0.7301**
Lamda _0.30	0.7306	0.7393	**0.7360**	**A.S._30**	0.7258	0.7442	**0.7385**	**Eps._0.30**	0.7163	0.7221	**0.7185**
Lamda _0.40	0.7303	0.7389	**0.7359**	**A.S._40**	0.7373	0.7557	**0.7450**	**Eps._0.40**	0.7167	0.7295	**0.7249**
Lamda _0.50	0.7293	0.7377	**0.7346**	**A.S._50**	0.7409	0.7573	**0.7476**	**Eps._0.50**	0.7320	0.7370	**0.7338**
Lamda _0.60	0.7279	0.7359	**0.7328**	**A.S._60**	0.7415	0.7589	**0.7491**	**Eps._0.60**	0.7322	0.7378	**0.7348**
Lamda _0.70	0.7263	0.7341	**0.7309**	**A.S._70**	0.7433	0.7594	**0.7493**	**Eps._0.70**	0.7301	0.7351	**0.7328**
Lamda _0.80	0.7245	0.7322	**0.7292**	**A.S._80**	0.7450	0.7602	**0.7501**	**Eps._0.80**	0.7294	0.7333	**0.7314**
Lamda _0.90	0.7228	0.7305	**0.7276**	**A.S._90**	0.7435	0.7586	**0.7501**	**Eps._0.90**	0.7232	0.7332	**0.7309**
Lamda _1	0.7210	0.7286	**0.7258**	**A.S._100**	0.7447	0.7615	**0.7509**	**Eps._1**	0.7304	0.7355	**0.7332**
RMSE											
Lasso Regresyon	**Min.**	**Mak.**	**Ort.**	**Rastgele Ağaç Regresyonu**	**Min.**	**Mak.**	**Ort.**	**Destek Vektör Regresyonu**	**Min.**	**Mak.**	**Ort.**
Lamda _0.10	0.7135	0.7265	**0.7185**	**A.S._10**	0.7501	0.8396	**0.7818**	**Eps._0.10**	0.7169	0.7319	**0.7237**
Lamda _0.20	0.7141	0.7260	**0.7187**	**A.S._20**	0.7189	0.7459	**0.7298**	**Eps._0.20**	0.7203	0.7410	**0.7267**
Lamda _0.30	0.7143	0.7260	**0.7187**	**A.S._30**	0.7069	0.7326	**0.7153**	**Eps._0.30**	0.7375	0.7451	**0.7422**
Lamda _0.40	0.7148	0.7265	**0.7188**	**A.S._40**	0.6915	0.7171	**0.7064**	**Eps._0.40**	0.7276	0.7446	**0.7336**
Lamda _0.50	0.7165	0.7278	**0.7206**	**A.S._50**	0.6892	0.7121	**0.7027**	**Eps._0.50**	0.7174	0.7242	**0.7217**
Lamda _0.60	0.7189	0.7298	**0.7231**	**A.S._60**	0.6870	0.7113	**0.7007**	**Eps._0.60**	0.7163	0.7239	**0.7204**
Lamda _0.70	0.7214	0.7319	**0.7256**	**A.S._70**	0.6862	0.7088	**0.7004**	**Eps._0.70**	0.7200	0.7267	**0.7231**
Lamda _0.80	0.7240	0.7343	**0.7280**	**A.S._80**	0.6851	0.7064	**0.6993**	**Eps._0.80**	0.7225	0.7277	**0.7250**
Lamda _0.90	0.7263	0.7365	**0.7301**	**A.S._90**	0.6875	0.7085	**0.6993**	**Eps._0.90**	0.7225	0.7360	**0.7257**
Lamda _1	0.7288	0.7389	**0.7324**	**A.S._100**	0.6832	0.7068	**0.6981**	**Eps._1**	0.7195	0.7264	**0.7225**
MAE											
Lasso Regresyon	**Min.**	**Mak.**	**Ort.**	**Rastgele Ağaç Regresyonu**	**Min.**	**Mak.**	**Ort.**	**Destek Vektör Regresyonu**	**Min.**	**Mak.**	**Ort.**
Lamda _0.10	0.5523	0.5665	**0.5561**	**A.S._10**	0.5615	0.6433	**0.5969**	**Eps._0.10**	0.5543	0.5668	**0.5606**
Lamda _0.20	0.5519	0.5656	**0.5548**	**A.S._20**	0.5486	0.5690	**0.5606**	**Eps._0.20**	0.5541	0.5751	**0.5603**
Lamda _0.30	0.5516	0.5646	**0.5548**	**A.S._30**	0.5386	0.5605	**0.5442**	**Eps._0.30**	0.5746	0.5806	**0.5757**
Lamda _0.40	0.5360	0.5710	**0.5539**	**A.S._40**	0.5237	0.5488	**0.5368**	**Eps._0.40**	0.5646	0.5832	**0.5688**
Lamda _0.50	0.5526	0.5664	**0.5556**	**A.S._50**	0.5231	0.5426	**0.5355**	**Eps._0.50**	0.5641	0.5717	**0.5675**
Lamda _0.60	0.5517	0.5678	**0.5567**	**A.S._60**	0.5216	0.5412	**0.5317**	**Eps._0.60**	0.5586	0.5676	**0.5621**
Lamda _0.70	0.5548	0.5694	**0.5576**	**A.S._70**	0.5222	0.5394	**0.5319**	**Eps._0.70**	0.5657	0.5750	**0.5685**
Lamda _0.80	0.5568	0.5713	**0.5592**	**A.S._80**	0.5212	0.5387	**0.5310**	**Eps._0.80**	0.5712	0.5757	**0.5748**
Lamda _0.90	0.5581	0.5731	**0.5606**	**A.S._90**	0.5220	0.5370	**0.5317**	**Eps._0.90**	0.5721	0.5837	**0.5737**
Lamda _1	0.5603	0.5753	**0.5624**	**A.S._100**	0.5244	0.5840	**0.5325**	**Eps._1**	0.5723	0.5810	**0.5763**

Kısaltmalar: R^2: Açıklayıcılık Katsayısı, **RMSE**: Root Mean Square Error, **MAE:** Mean Absolute Error, **A.S.:** Ağaç Sayısı, **Eps.:** Epsilon, **Min.:** Minimum, **Mak.:** Maksimum, **Ort.:** Ortanca

Bir diğer performans ölçüsü olan RMSE açısından sırasıyla Lasso Regresyon yönteminde 0.10 ile 1 arasında değişmek üzere farklı lamda (λ) değerleri belirlendiğinde elde edilen R^2 değerlerine ait sıra ortalamaları arasındaki farkın istatistiksel olarak anlamlı olduğu (X^2=67.947, p<0.001), Rastgele Ağaç Regresyonunda 10 ile 100 arasında değişmek üzere farklı ağaç sayıları belirlendiğinde elde edilen RMSE değerlerine ait sıra ortalamaları arasındaki farkın istatistiksel olarak anlamlı olduğu (X^2=65.265, p<0.001), Destek Vektör Regresyonda 0.10 ile 1 arasında değişmek üzere farklı epsilon parametreleri belirlendiğinde elde edilen RMSE değerlerine ait

sıra ortalamaları arasındaki farkın istatistiksel olarak anlamlı olduğu (X^2=66.679, p<0.001) söylenebilmektedir. Diğer taraftan, başka bir performans ölçüsü olan MAE bakımından sırasıyla Lasso Regresyon yönteminde 0.10 ile 1 arasında değişmek üzere farklı lamda (λ) değerleri belirlendiğinde elde edilen MAE değerlerine ait sıra ortalamaları arasındaki farkın istatistiksel olarak anlamlı olduğu (X^2=29.687, p<0.001), Rastgele Ağaç Regresyonunda 10 ile 100 arasında değişmek üzere farklı sayılarda ağaç türetildiğinde elde edilen MAE değerlerine ait sıra ortalamaları arasındaki farkın anlamlı olduğu (X^2=61.593, p<0.001), Destek Vektör Regresyonda 0.10 ile 1 arasında değişmek üzere farklı epsilon parametreleri belirlendiğinde elde edilen MAE değerlerine ait sıra ortalamaları arasındaki farkın istatistiksel olarak anlamlı olduğu (X^2= 73.750, p<0.001) görülmektedir (Tablo 3).

Tablo 3. Lasso Regresyon, Rastgele Ağaç Regresyonu ve Destek Vektör Regresyonu İçin Farklı Hiperparametre Değerleri Arasında Regresyon Performans Sonuçları Bakımından Görülen Farklılıklar

R^2														
Lasso Regresyon	**N**	**Sıra Ort.**	X^2	**p**	**Rastgele Ağaç Regresyonu**	**N**	**Sıra Ort.**	X^2	**p**	**Destek Vektör Regresyonu**	**N**	**Sıra Ort.**	X^2	**p**
Lamda (λ)_0.10	10	67.85			**A.S._10**	10	5.50			**Eps._0.10**	10	56.10		
Lamda (λ)_0.20	10	74.25			**A.S._20**	10	16.55			**Eps._0.20**	10	36.95		
Lamda (λ)_0.30	10	75.55			**A.S._30**	10	27.45			**Eps._0.30**	10	6.95		
Lamda (λ)_0.40	10	72.20			**A.S._40**	10	52.10			**Eps._0.40**	10	18.45		
Lamda (λ)_0.50	10	64.10			**A.S._50**	10	59.95			**Eps._0.50**	10	77.00		
Lamda (λ)_0.60	10	51.95	**62.751**	**<0.001**	**A.S._60**	10	64.25	**65.543**	**<0.001**	**Eps._0.60**	10	86.65	**66.667**	**<0.001**
Lamda (λ)_0.70	10	40.35			**A.S._70**	10	66.40			**Eps._0.70**	10	62.60		
Lamda (λ)_0.80	10	28.80			**A.S._80**	10	68.60			**Eps._0.80**	10	48.20		
Lamda (λ)_0.90	10	18.95			**A.S._90**	10	69.60			**Eps._0.90**	10	44.90		
Lamda (λ)_1	10	11			**A.S._100**	10	74.60			**Eps._1**	10	67.20		
RMSE														
Lasso Regresyon	**N**	**Sıra Ort.**	X^2	**p**	**Rastgele Ağaç Regresyonu**	**N**	**Sıra Ort.**	X^2	**p**	**Destek Vektör Regresyonu**	**N**	**Sıra Ort.**	X^2	**p**
Lamda (λ)_0.10	10	25.50			**A.S._10**	10	95.50			**Eps._0.10**	10	44.95		
Lamda (λ)_0.20	10	25.40			**A.S._20**	10	84.40			**Eps._0.20**	10	63.95		
Lamda (λ)_0.30	10	26.75			**A.S._30**	10	73.20			**Eps._0.30**	10	94.10		
Lamda (λ)_0.40	10	30.25			**A.S._40**	10	49.20			**Eps._0.40**	10	82.50		
Lamda (λ)_0.50	10	38.05			**A.S._50**	10	41.10			**Eps._0.50**	10	23.95		
Lamda (λ)_0.60	10	50.20	**67.947**	**<0.001**	**A.S._60**	10	36.65	**65.265**	**<0.001**	**Eps._0.60**	10	14.35	**66.679**	**<0.001**
Lamda (λ)_0.70	10	61.60			**A.S._70**	10	34.70			**Eps._0.70**	10	38.50		
Lamda (λ)_0.80	10	73.40			**A.S._80**	10	32.40			**Eps._0.80**	10	53.00		
Lamda (λ)_0.90	10	82.90			**A.S._90**	10	31.50			**Eps._0.90**	10	56.00		
Lamda (λ)_1	10	90.95			**A.S._100**	10	26.35			**Eps._1**	10	33.70		
MAE														
Lasso Regresyon	**N**	**Sıra Ort.**	X^2	**p**	**Rastgele Ağaç Regresyonu**	**N**	**Sıra Ort.**	X^2	**p**	**Destek Vektör Regresyonu**	**N**	**Sıra Ort.**	X^2	**p**
Lamda (λ)_0.10	10	42.60			**A.S._10**	10	94.80			**Eps._0.10**	10	15.30		
Lamda (λ)_0.20	10	35.50			**A.S._20**	10	83.80			**Eps._0.20**	10	22.05		
Lamda (λ)_0.30	10	34.60			**A.S._30**	10	71.80			**Eps._0.30**	10	82.10		
Lamda (λ)_0.40	10	34.45			**A.S._40**	10	49.35			**Eps._0.40**	10	52.05		
Lamda (λ)_0.50	10	40.10			**A.S._50**	10	40.80			**Eps._0.50**	10	38.95		
Lamda (λ)_0.60	10	45.65	**29.687**	**<0.001**	**A.S._60**	10	33.55	**61.593**	**<0.001**	**Eps._0.60**	10	19.20	**73.750**	**<0.001**
Lamda (λ)_0.70	10	54.45			**A.S._70**	10	34.45			**Eps._0.70**	10	48.55		
Lamda (λ)_0.80	10	65.70			**A.S._80**	10	31.65			**Eps._0.80**	10	68.50		
Lamda (λ)_0.90	10	72.50			**A.S._90**	10	30.45			**Eps._0.90**	10	73.05		
Lamda (λ)_1	10	79.45			**A.S._100**	10	34.35			**Eps._1**	10	85.25		

Kısaltmalar: A.S.: Ağaç Sayısı, **Eps.:** Epsilon, **Min.:** Minimum, **Mak.:** Maksimum, **Ort:** Ortalama, **X^2:** Kruskall-Wallis Varyans analizi Ki-kare değeri

Üç farklı regresyon yöntemi için farklı hiperparametre değerleri kullanılarak elde edilen performans sonuçları birlikte değerlendirildiğinde bu üç yönteme göre R^2, RMSE ve MAE değerlerinden elde edilen performans sonuçlarına ait istatistiksel farklılıklar Tablo 4'de sunulmuştur. Buna göre Lasso Regresyon yönteminde 0.10 ile 1 arasında değişmek üzere farklı lamda (λ) değerleri belirlendiğinde, Rastgele Ağaç Regresyonu yönteminde 10 ile 100 arasında değişmek üzere farklı sayılarda ağaç türetildiğinde ve Destek Vektör Regresyonda 0.10 ile 1 arasında değişmek üzere farklı epsilon parametre değerleri belirlendiğinde elde edilen R^2 ($F=11.301$, $p<0.001$), RMSE ($F=15.765$, $p<0.001$) ve MAE ($F=80.804$, $p<0.001$) değerlerine ait ortalamalar arasındaki farklılığın istatistiksel olarak anlamlı olduğu görülmektedir.

Tablo 4. Farklı Hiperparametre Değerlerine Göre Belirlenen Lasso Regresyon, Rastgele Ağaç Regresyonu ve Destek Vektör Regresyonu Performans Sonuçlarının Karşılaştırılması

Performans Ölçüsü	Regresyon Yöntemi	N	Ort.	F	p
R^2	**Lasso Regresyon (λ = 0.10-1)**	100	0.7316	11.301	**<0.001**
	Rastgele Ağaç Regresyonu (Ağaç Sayısı = 10-100)	100	0.7387		
	Destek Vektör Regresyonu (ε = 0.10-1)	100	0.7300		
RMSE	**Lasso Regresyon (λ = 0.10-1)**	100	0.7240	15.765	**<0.001**
	Rastgele Ağaç Regresyonu (Ağaç Sayısı = 10-100)	100	0.7139		
	Destek Vektör Regresyonu (ε = 0.10-1)	100	0.7268		
MAE	**Lasso Regresyon (λ = 0.10-1)**	100	0.5588	80.804	**<0.001**
	Rastgele Ağaç Regresyonu (Ağaç Sayısı = 10-100)	100	0.5438		
	Destek Vektör Regresyonu (ε = 0.10-1)	100	0.5697		

Kısaltmalar: R^2: Çoklu Açıklayıcılık Katsayısı, **RMSE:** Root Mean Square Error (Ortalama hata karekökü), **MAE:** Mean Absolute error (Ortalama mutlak hata), **λ:** Lamda, **ε:** Epsilon, **Ort.:** Ortalama, **F:** Anova testi

4. SONUÇ

Literatürde regresyona dayanan modelleme çalışmalarında veri setinin tür ve büyüklüğüne bağlı olarak düşük tahmin performansının elde edilmesi durumu ile sıklıkla karşı karşıya kalınmaktadır. Bu sorunun çözümü için kullanılan değişkenlerin ve veri setinin genel durumuna en uygun ayarlama yöntemleri ile incelenmesi tavsiye edilmektedir. Bu sayede daha yüksek tahmin performanslarının elde edilmesi mümkün olabilmektedir.

Model performansının iyileştirilmesine yönelik olarak uygulanabilecek, klasik regresyon yöntemlerine alternatif nitelik taşıyan yaklaşımlar arasında makine öğrenmesi regresyon yöntemleri ön plana çıkmaktadır. Bu yöntemler arasında Lasso Regresyon, Rastgele Ağaç Regresyonu ile Destek Vektör Regresyon yöntemleri sayılabilmektedir. Veri setinden öğrenme gerçekleştiren, farklı hiperparametre değerleri kullanılarak sonuçların optimizasyonuna imkân veren bu yöntemler sayesinde en iyi performans sonucunun araştırılması mümkün olmaktadır. Bu amaçla yararlanılabilecek hiperparametre değerleri arasından Lasso Regresyon için lamda (λ) değeri, Rastgele Ağaç yöntemi için bir tür düzenleyici hiperparametre (regularization hyperparameter) değeri olarak kabul edilen ağaç sayısı sayılabilmekte iken, Destek Vektör regresyon için epsilon (ε) değeri tahmin doğruluğunun bir ölçüsü olarak kabul edilmektedir.

Bu çalışmada kişi başı sağlık harcamasının tahmini amacıyla Lasso Regresyon, Rastgele Ağaç Regresyonu ile Destek Vektör Regresyonu yöntemleri kullanılarak çoklu regresyon modelleri oluşturulmuştur. Lasso Regresyon için 0.10 ile 1 arasında değişmek üzere farklı lamda (λ) değerleri belirlenmiş, Rastgele Ağaç Regresyonu için 10 ile 100 arasında değişmek üzere farklı sayıda ağaç türetilmiş, Destek Vektör regresyon için ise 0.10 ile 1 arasında değişen farklı epsilon parametreleri belirlenerek elde edilen performans sonuçları R^2, RMSE ve MAE değerleri üzerinden değerlendirilmiştir. Kullanılan her üç yöntem için de hiperparametre değerlerinin yükselmesi kişi başı sağlık harcamalarının tahmininde daha yüksek model

performansının elde edilmesi anlamına gelmektedir. Çalışmada ayrıca makine öğrenmesi yöntemlerine ait performans sonuçlarının karşılaştırılması amacıyla literatürde sıklıkla kullanılan bir hiperparametre olan "k" parça çapraz gerçerlilik uygulanmıştır. Çalışma sonucunda Lasso Regresyonda farklı lamda değerleri belirlendiğinde, Rastgele Ağaç Regresyonunda farklı sayılarda ağaç türetildiğinde ve Destek Vektör Regresyonda farklı epsilon parametreleri belirlendiğinde R^2, RMSE ve MAE değerlerinden elde edilen performans sonuçları arasındaki farklılığın istatistiksel olarak anlamlı olduğu bulunmuştur. Çalışma sonucunda kişi başı sağlık harcamalarının tahmininde Rastgele Ağaç Regresyonu yöntemine ait tahmin performanslarının diğer yöntemlere göre daha iyi olduğu gözlemlenmiştir. Bu bulgulardan yola çıkılarak sağlık harcamalarının tahminine yönelik modelleme çalışmaları için Rastgele Ağaç Regresyonu yönteminin iyi bir alternatif olduğu savunulabilecektir. Çalışma sonuçlarının makine öğrenmesi yöntemleri kullanılarak sağlık harcamalarının modellendiği araştırmalar için en uygun hiperparametre değerlerinin belirlenmesi konusunda literatüre katkı sağlaması ümit edilmektedir.

KAYNAKLAR

1. Alpar R. (2011) *Uygulamalı çok değişkenli istatistiksel yöntemler*, Detay Yayıncılık, Ankara, 415-620.

2. Basu, A., Manning, W.G. ve Mullahy, J. (2004). Comparing alternative model: log and cox proportional hazard? *Health Economics*, 13(8), 749-765. doi: 10.1002/hec.852.

3. Belloni, A., Chernozhukov, V., Hansen, C. (2012) Inference for high-dimensional sparse econometric models. https://arxiv.org/abs/1201.0220. doi: 10.1017/CBO9781139060035.008. Erişim Tarihi: 01.01.2016.

4. Bergstra, J. ve Bengio, Y. (2012) Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281-305. http://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf. Erişim Tarihi: 01.02.2016.

5. Box, G.E.P. ve Cox, D.R. (1964) An analysis of transformations, *Journal of the Royal Statistical Society*, 26(2), 211-252. doi: 10.1.1.321.3819.

6. Brieman, L. (2001) Random forests, *Machine Learning*, 45, 5-32. doi: 10.1023%2FA%3A1010933404324.

7. Cherkassky, V. ve Ma, Y. (2004) Practical selection of SVM parameters and noise estimation for SVM regression, *Neural Networks*, 17(1), 113-126. doi:10.1016/S0893-6080(03)00169-2.

8. Cosgun E., Karaağaoğlu E. (2011). Veri madenciliği yöntemleriyle mikrodizilim gen ifade analizi, *Hacettepe Tıp Dergisi*, 42, 180-189. http://docplayer.biz.tr/3432783-Veri-madencili-i-yontemleriyle-mikrodizilim-gen-ifade-analizi.html. Erişim Tarihi: 01.02.2016.

9. Collins, B. (2016) Big data and health economics: strengths, weaknesses, opportunities and threats, *Pharmacoeconomics*, 34(2), 101-106. doi: 10.1007/s40273-015-0306-7.

10. Cristianini, N. ve Shawe-Taylor, J. (2000). *An introduction to support vector machines and other Kernel based learning methods*, Cambridge University Press, UK, 93-122.

11. Crown, W.H. (2015) Potential application of machine learning in health outcomes research and statistical cautions, Value in Health, 18(2), 137-140. doi: 10.1016/j.jval.2014.12.005.

12. Duan, K., Keerthi, S.S., Poo, A.N. (2003) Evaluation of simple performance measures for tuning SVM hyperparameters, Neurocomputing, 51, 41-59. doi.org/10.1016/S0925-2312(02)00601-X.

13. Einav, L., Levin, J.D. (2014) The data revolution and economic analysis. *NBER/Innovation Policy and the Economy*, 14(1): 1-24. doi: 10.3386/w19035.

14. Elasan, S., Keskin, S., Arı E. (2016) İlişkili bileşen regresyonu: DNA hasarını belirleme modeli üzerinde uygulanması, *Türkiye Klinikleri Biyoistatistik Dergisi*, 8(1): 45-52. doi: 10.5336/biostatic.2015-48311.

15. Frank, I.E., Friedman J.H. (1993) A statistical view of some chemometrics regression tools. *Technometrics*, 35, 109-148. doi: 10.2307/1269656.

16. Frenk, J. (2010) The global health system: strengthening national health systems as the next step for global progress, *PLOS Medicine*, 7(1), 1-3. doi: 10.1371/journal.pmed.1000089.

17. Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R. (2006) Random Forest for land cover classification. *Pattern Recognition Letters*, 27(4), 294-300. doi: 10.1016/j.patrec.2005.08.011.

18. Gupta, I., Mitra, A. (2004) Economic growth, health and poverty: an exploratory study for India, *Development Policy Review*, 22(2), 193-206. doi: 10.1111/j.1467-7679.2004.00245.x.

19. Hassan, S.S., Farhan, M., Mangayil, R., Huttunen, H., Aho, T. (2013) Bioprocess data mining using regularized regression and random forests, *BMC System Biology*, 7(1):1-7. doi: 10.1186/1752-0509-7-S1-S5.

20. Hastie, T., Tibshirani, R. ve Friedman, J. (2009) *Random Forest.* The elements of statistical learning data mining, Inference and Prediction. Springer Series in Statistics, 587-613.

21. Hawkins, D.M. (2004) The problem of overfitting, *Journal of Chemical Information and Modeling*, 44(1), 1-12. doi: 10.1021/ci0342472.

22. Jaggi, M. (2014) An equivalence between the lasso and support vector machines, https://arxiv.org/pdf/1303.1152.pdf, Erişim Tarihi: 16.5.2017. arXiv:1303.1152v2.

23. Jones, A.M., Rice, N., d'Uva, T.B. ve Balai, S. (2007) *Applied health economics*, Routledge, Taylor & Francis, London and New York, 280-319.

24. Kavaklıoğlu, K. (2011) Modeling and prediction of Turkey's electricity consumption using support vector regression, *Applied Energy*, 88(1), 368-375. doi: 10.1016/j.apenergy.2010.07.021.

25. Kazem, A., Sharifi, E., Hussain, F.K., Saberi, M. ve Hussain, O.K. (2013) Support vector regression with chaos-based firefly algorithm for stock market price forecasting. *Applied Soft Computing*, 13(2), 947-958. doi: 10.1016/j.asoc.2012.09.024.

26. Kohavi, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection, *International Joint Conference on Artificial Intelligence (IJCAI'95),* vol.2, 1137-1143.

27. Liaw, A., Wiener, M. (2002) Classification and regression by random forest, *R News*, vol.2/3, 18-22. http://www.bios.unc.edu/~dzeng/BIOS740/randomforest.pdf. Erişim Tarihi: 01.01.2016.

28. Manning, W. (2006) *Dealing with skewed data on costs and expenditures*, Jones A.M. (2006) The Elgar Companion to Health Economics, Second Edition, Edward Elgar Publishing, Inc. Massachusetts, USA, p.439-446.

29. Manning, W.G. (1998) The logged dependent variable, heteroscedasticity, and the retransformation problem, *Journal of Health Economics*, 17(3), 283-295. doi: 10.1016/S0167-6296(98)00025-3.

30. Martin, M.J.J., Gonzalez, M.P.L.A. ve Garcia, M.D.C. (2011) Review of the literature on the determinants of healthcare expenditure, *Applied Economics*, 43(1), 19-46. doi: 10.1080/00036841003689754.

31. Mattera, D. ve Haykin, S. (1999) *Support vector machines for dynamic reconstruction of a chaotic system*, Ed. Schöl B. Burges C.J.C. Smola A.J. (1999) Advances in Kernel Methods, Massachusetts Institute of Technology (MIT), 211-239.

32. Mihaylova, B., Briggs, A., O'Hagan, A. ve Thompson, S.G. (2011) Review of statistical methods for analysing healthcare resources and costs, *Health Economics*, 20(8), 897-916. doi: 10.1002/hec.1653.

33. Rodriguez, J.J., Diez-Pastor, J.F., Gonzalez A.A. ve Garcia-Osorio, C. (2015) *An experimental study on combining binarization techniques and ensemble methods of decision trees*, Multiple Classifier Systems 12th International Workshop, MCS 2015, Günzburg, Germany, June 29-July 1 2015 Proceedings, Springer.

34. Schölkopf, B., Smola, A.J. (2002) *Learning with kernels. Support vector machines, regularization, optimization, and beyond*, The MIT Press, Cambridge, Massachusetts, London, England.

35. Sinha, R.K., Chatterjee, K., Nair, N. ve Tripathy, P.K. (2016) Determinants of out-of-pocket and catastrophic health expenditure: a cross sectional study, *British Journal of Medicine & Medical Research*, 11(8), 1-11. doi : 10.9734/BJMMR/2016/21470.

36. Suthaharan, S. (2016) Support vector machine. Machine learning models and algorithms for big data classification, *Integrated Series in Information Systems*, vol.36, 207-235.

37. Tibshirani, R. (1996) Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society*, 58(1): 267-288. doi: 10.1111/j.1467-9868.2011.00771.x.

38. Tsamardinos, I., Rakhshani, A. ve Lagani, V. (2015). Performance-estimation properties of cross-validation-based protocols with simultaneous hyper-parameter optimization, *International Journal of Artificial Intelligence Tools*, 24(5), 1-30. http://www.mensxmachina.org/wp-content/uploads/2014/03/SETN-2014-Model Selection.pdf. Erişim Tarihi: 01.02.2016.

39. Vapnik, V., Golowich, S.E. ve Smola, A. (1997). *Support vector method for function approximation, regression estimation and signal processing*, In M. Mozer, M. Jordan and T. Petshe, editors, Advances in Neural Information Processing Systems, 9. Cambridge MA. 1997. MIT Press. 281-287.

40. Wang, W. ve Xu, Z. (2004). A heuristic training for support vector regression, *Neurocomputing*, 61, 259-275. doi: 10.1016/j.neucom.2003.11.012.

41. WHO (World Health Organization) *The World Health Report 2000: Improving health systems: improving performance*, The World Health Organization.

42. Witten, I.H. ve Frank, E. (2005) *Data mining practical machine learning tools and techniques*, Second Edition, Morgan Kaufmann Publications, Elsevier, San Francisco, USA.

43. Yılmaz, E. (2016). Kardiotokogram verisinden fetal iyilik halinin belirlenmesi için bir karar destek sistemi, *Uludag University Journal of The Faculty of Engineering*, 21(2):331-340. doi: 10.17482/uumfd.278033.

44. Zheng, A. (2015) *Evaluating machine learning models a beginner's guide to key concepts and pitfalls*, O'Reilly, USA.

Ek 1. Lasso Regresyonu İçin Farklı Hiperparametre Değerlerine Ait Sonuçlar

Hiperparametre: Lamda (λ)	Performans Ölçüsü	"k" parça çapraz geçerlilik									
		k=5	k=10	k=15	k=20	k=25	k=30	k=35	k=40	k=45	k=50
Lamda (λ) = 0.10	R^2	0.7319	0.7303	0.7343	0.7352	0.7367	0.7382	0.7397	0.7073	0.7365	0.7361
	RMSE	0.7243	0.7265	0.7211	0.7199	0.7179	0.7158	0.7135	0.7174	0.7185	0.7186
	MAE	0.5629	0.5665	0.5565	0.5586	0.5557	0.5553	0.5523	0.5571	0.5551	0.5550
Lamda (λ) = 0.20	R^2	0.7320	0.7307	0.7346	0.7351	0.7366	0.7380	0.7349	0.7370	0.7362	0.7360
	RMSE	0.7242	0.7260	0.7206	0.7200	0.7179	0.7161	0.7141	0.7175	0.7186	0.7188
	MAE	0.5631	0.5656	0.5538	0.5581	0.5547	0.5549	0.5519	0.5562	0.5544	0.5543
Lamda (λ) = 0.30	R^2	0.7314	0.7306	0.7345	0.7350	0.7365	0.7378	0.7393	0.7369	0.7361	0.7360
	RMSE	0.7250	0.7260	0.7208	0.7201	0.7182	0.7163	0.7143	0.7175	0.7186	0.7188
	MAE	0.5639	0.5646	0.5574	0.5573	0.5544	0.5542	0.5516	0.5553	0.5539	0.5537
Lamda (λ) = 0.40	R^2	0.7306	0.7303	0.7343	0.7346	0.7360	0.7375	0.7389	0.7368	0.7360	0.7359
	RMSE	0.7261	0.7265	0.7211	0.7207	0.7188	0.7167	0.7148	0.7177	0.7187	0.7189
	MAE	0.5651	0.5641	0.5710	0.5569	0.5544	0.5380	0.5516	0.5490	0.5360	0.5534
Lamda (λ) = 0.50	R^2	0.7299	0.7293	0.7336	0.7334	0.7349	0.7363	0.7377	0.7356	0.7347	0.7346
	RMSE	0.7271	0.7278	0.7221	0.7224	0.7203	0.7184	0.7165	0.7194	0.7205	0.7207
	MAE	0.5664	0.5644	0.5574	0.5574	0.5553	0.5545	0.5526	0.5559	0.5547	0.5546
Lamda (λ) = 0.60	R^2	0.7289	0.7279	0.7322	0.7316	0.7333	0.7346	0.7359	0.7339	0.7329	0.7327
	RMSE	0.7284	0.7298	0.7240	0.7247	0.7225	0.7207	0.7189	0.7217	0.7230	0.7233
	MAE	0.5678	0.5649	0.5585	0.5587	0.5565	0.5570	0.5539	0.5517	0.5560	0.5559
Lamda (λ) = 0.70	R^2	0.7275	0.7263	0.7306	0.7297	0.7315	0.7328	0.7341	0.7319	0.7310	0.7308
	RMSE	0.7303	0.7319	0.7260	0.7273	0.7249	0.7231	0.7214	0.7243	0.7255	0.7258
	MAE	0.5694	0.5662	0.5595	0.5602	0.5570	0.5568	0.5554	0.5548	0.5577	0.5576
Lamda (λ) = 0.80	R^2	0.7259	0.7245	0.7288	0.7277	0.7298	0.731	0.7322	0.7301	0.7293	0.7291
	RMSE	0.7325	0.7343	0.7285	0.7299	0.7272	0.7255	0.7240	0.7267	0.7279	0.7281
	MAE	0.5713	0.5681	0.5611	0.5617	0.5587	0.5582	0.5568	0.5592	0.5592	0.5591
Lamda (λ) = 0.90	R^2	0.7241	0.7228	0.727	0.7258	0.7281	0.7293	0.7305	0.7286	0.7277	0.7275
	RMSE	0.7348	0.7365	0.7310	0.7325	0.7295	0.7278	0.7263	0.7288	0.7300	0.7302
	MAE	0.5731	0.5694	0.563	0.5629	0.5599	0.5596	0.5581	0.5603	0.5606	0.5606
Lamda (λ) = 1	R^2	0.7220	0.7210	0.7250	0.7238	0.7263	0.7275	0.7286	0.7266	0.7259	0.7258
	RMSE	0.7376	0.7389	0.7336	0.7351	0.7319	0.7303	0.7288	0.7315	0.7324	0.7325
	MAE	0.5753	0.5706	0.5648	0.5648	0.5618	0.5614	0.5603	0.5625	0.5624	0.5624

Ek 2. Rastgele Ağaç Regresyonu İçin Farklı Hiperparametre Değerlerine Ait Sonuçlar

		"k" parça çapraz geçerlilik									
Hiperparametre: Ağaç Sayısı	**Performans Ölçüsü**	**k=5**	**k=10**	**k=15**	**k=20**	**k=25**	**k=30**	**k=35**	**k=40**	**k=45**	**k=50**
Rastgele Ağaç Regresyonu Ağaç Sayısı = 10	R^2	0.6906	0.6678	0.6398	0.6698	0.6916	0.6924	0.7120	0.7125	0.6765	0.6470
	RMSE	0.7782	0.8063	0.8396	0.8039	0.7768	0.7758	0.7508	0.7501	0.7957	0.7855
	MAE	0.5963	0.6191	0.6433	0.6199	0.5952	0.5952	0.5615	0.5775	0.6065	0.5976
Rastgele Ağaç Regresyonu Ağaç Sayısı = 20	R^2	0.7286	0.7324	0.7165	0.717	0.7271	0.7285	0.7326	0.7359	0.7157	0.7258
	RMSE	0.7288	0.7236	0.7449	0.7442	0.7309	0.7288	0.7235	0.7189	0.7459	0.7325
	MAE	0.5518	0.5489	0.5690	0.5622	0.5629	0.5631	0.5591	0.5486	0.5689	0.5541
Rastgele Ağaç Regresyonu Ağaç Sayısı = 30	R^2	0.7375	0.7413	0.7258	0.7334	0.7395	0.7442	0.7399	0.7427	0.7321	0.7375
	RMSE	0.7168	0.7115	0.7326	0.7223	0.7139	0.7076	0.7134	0.7069	0.7241	0.7167
	MAE	0.5397	0.5403	0.5605	0.5505	0.5492	0.5396	0.5481	0.5387	0.5495	0.5386
Rastgele Ağaç Regresyonu Ağaç Sayısı = 40	R^2	0.7388	0.7532	0.7373	0.7439	0.7461	0.7557	0.7496	0.7485	0.7387	0.744
	RMSE	0.7149	0.6949	0.7171	0.7079	0.7050	0.6915	0.7000	0.7015	0.7151	0.7079
	MAE	0.5398	0.5257	0.5488	0.5351	0.5416	0.5237	0.5385	0.5349	0.5438	0.5345
Rastgele Ağaç Regresyonu Ağaç Sayısı = 50	R^2	0.7417	0.7536	0.7449	0.747	0.7476	0.7573	0.7521	0.7477	0.7409	0.748
	RMSE	0.7109	0.6944	0.7066	0.7036	0.7028	0.6892	0.6966	0.7027	0.7121	0.7023
	MAE	0.5371	0.5236	0.5391	0.5308	0.5378	0.5231	0.5349	0.5362	0.5426	0.5317
Rastgele Ağaç Regresyonu Ağaç Sayısı = 60	R^2	0.7415	0.7529	0.7472	0.7476	0.746	0.7589	0.7524	0.7506	0.7423	0.7515
	RMSE	0.7113	0.6954	0.7034	0.7029	0.7050	0.6870	0.6961	0.6986	0.7101	0.6973
	MAE	0.5386	0.5246	0.5346	0.5303	0.5376	0.5216	0.5322	0.5313	0.5412	0.5272
Rastgele Ağaç Regresyonu Ağaç Sayısı = 70	R^2	0.7433	0.7527	0.7498	0.7486	0.7450	0.7594	0.7489	0.7513	0.7454	0.7513
	RMSE	0.7088	0.6957	0.6998	0.7014	0.7064	0.6862	0.7010	0.6976	0.7059	0.6977
	MAE	0.5354	0.5264	0.5328	0.5285	0.5394	0.5222	0.5363	0.5310	0.5392	0.5285
Rastgele Ağaç Regresyonu Ağaç Sayısı = 80	R^2	0.7450	0.748	0.7513	0.7498	0.746	0.7602	0.7504	0.7531	0.7461	0.7520
	RMSE	0.7064	0.7023	0.6977	0.6997	0.7051	0.6851	0.6990	0.6951	0.7049	0.6966
	MAE	0.5350	0.5305	0.5315	0.5274	0.5378	0.5212	0.5349	0.5274	0.5387	0.5294
Rastgele Ağaç Regresyonu Ağaç Sayısı = 90	R^2	0.7435	0.7493	0.7517	0.7484	0.7466	0.7586	0.7510	0.7526	0.7474	0.7533
	RMSE	0.7085	0.7005	0.6970	0.7017	0.7042	0.6875	0.6981	0.6958	0.7031	0.6948
	MAE	0.5360	0.5282	0.5329	0.5306	0.5370	0.5220	0.5332	0.5269	0.5369	0.5275
Rastgele Ağaç Regresyonu Ağaç Sayısı = 100	R^2	0.7447	0.7484	0.7524	0.7496	0.7480	0.7615	0.7509	0.7551	0.7510	0.7545
	RMSE	0.7068	0.7017	0.6962	0.7000	0.7022	0.6832	0.6982	0.6923	0.6981	0.6932
	MAE	0.5342	0.5291	0.5328	0.5302	0.5349	0.5840	0.5323	0.5244	0.5329	0.5271

Ek 3. Destek Vektör Regresyonu İçin Farklı Hiperparametre Değerlerine Ait Sonuçlar

Hiperparametre:		"k" parça çapraz geçerlilik									
Epsilon Değeri	**Performans Ölçüsü**	**k=5**	**k=10**	**k=15**	**k=20**	**k=25**	**k=30**	**k=35**	**k=40**	**k=45**	**k=50**
Epsilon Değeri (ε) = 0.10	R^2	0.7291	0.7263	0.7316	0.7300	0.7338	0.7324	0.7374	0.733	0.7329	0.7324
	RMSE	0.7281	0.7319	0.7247	0.7269	0.7218	0.7237	0.7169	0.7229	0.7230	0.7237
	MAE	0.5636	0.5668	0.5603	0.5633	0.5589	0.5623	0.5543	0.5603	0.5595	0.5610
Epsilon Değeri (ε) = 0.20	R^2	0.7212	0.7194	0.7276	0.7249	0.7303	0.7319	0.7349	0.7300	0.7321	0.7309
	RMSE	0.7386	0.741	0.7301	0.7338	0.7265	0.7243	0.7203	0.7269	0.7240	0.7257
	MAE	0.5751	0.5721	0.5635	0.566	0.5574	0.5585	0.5541	0.5622	0.5573	0.5580
Epsilon Değeri (ε) = 0.30	R^2	0.7201	0.7167	0.719	0.717	0.7186	0.7196	0.7221	0.7163	0.7184	0.7176
	RMSE	0.7402	0.7446	0.7416	0.7441	0.7421	0.7408	0.7375	0.7451	0.7424	0.7434
	MAE	0.5747	0.5797	0.5746	0.5783	0.5754	0.5772	0.5747	0.5806	0.5753	0.5760
Epsilon Değeri (ε) = 0.40	R^2	0.7200	0.7167	0.7236	0.7243	0.7285	0.7275	0.7295	0.7273	0.7246	0.7253
	RMSE	0.7402	0.7446	0.7354	0.7346	0.7289	0.7302	0.7276	0.7306	0.7341	0.7332
	MAE	0.5762	0.5832	0.5716	0.5733	0.5646	0.5675	0.5657	0.5692	0.5685	0.5671
Epsilon Değeri (ε) = 0.50	R^2	0.732	0.7327	0.7329	0.7356	0.7344	0.7357	0.737	0.735	0.7323	0.7332
	RMSE	0.7242	0.7232	0.723	0.7194	0.721	0.7192	0.7174	0.7201	0.7238	0.7225
	MAE	0.5717	0.5705	0.5674	0.566	0.5667	0.5655	0.5641	0.5692	0.5687	0.5676
Epsilon Değeri (ε) = 0.60	R^2	0.7378	0.7349	0.7322	0.7347	0.7341	0.7369	0.7369	0.7359	0.7345	0.7339
	RMSE	0.7163	0.7203	0.7239	0.7206	0.7214	0.7176	0.7176	0.7190	0.7208	0.7217
	MAE	0.5636	0.5676	0.5657	0.5609	0.5618	0.5605	0.5586	0.5623	0.562	0.5622
Epsilon Değeri (ε) = 0.70	R^2	0.7338	0.7303	0.7301	0.7311	0.7341	0.7345	0.7351	0.7336	0.7315	0.732
	RMSE	0.7218	0.7266	0.7267	0.7255	0.7214	0.7208	0.7200	0.7221	0.7249	0.7242
	MAE	0.5707	0.575	0.5746	0.573	0.5679	0.5686	0.5657	0.5664	0.5684	0.5684
Epsilon Değeri (ε) = 0.80	R^2	0.7313	0.7294	0.7303	0.7319	0.7315	0.7326	0.7327	0.7333	0.7305	0.7301
	RMSE	0.7251	0.7277	0.7265	0.7244	0.7249	0.7234	0.7232	0.7225	0.7263	0.7267
	MAE	0.5748	0.5752	0.5757	0.5749	0.5739	0.5724	0.5712	0.5721	0.575	0.5754
Epsilon Değeri (ε) = 0.90	R^2	0.7296	0.7296	0.7296	0.7296	0.7296	0.7296	0.7296	0.7296	0.7296	0.7296
	RMSE	0.7275	0.7275	0.7275	0.7275	0.7275	0.7275	0.7275	0.7275	0.7275	0.7275
	MAE	0.5776	0.5776	0.5776	0.5776	0.5776	0.5776	0.5776	0.5776	0.5776	0.5776
Epsilon Değeri (ε) = 1	R^2	0.7312	0.7312	0.7312	0.7312	0.7312	0.7312	0.7312	0.7312	0.7312	0.7312
	RMSE	0.7252	0.7252	0.7252	0.7252	0.7252	0.7252	0.7252	0.7252	0.7252	0.7252
	MAE	0.5810	0.5810	0.5810	0.5810	0.5810	0.5810	0.5810	0.5810	0.5810	0.5810