MAKÜFEBED

ISSN Online: 1309-2243
http://dergipark.gov.tr/makufebed
https://doi.org/10.29048/makufebed.650463

Mehmet Akif Ersoy Üniversitesi Fen Bilimleri Enstitüsü Dergisi 11(1): 86-94 (2020)
The Journal of Graduate School of Natural and Applied Sciences of Mehmet Akif Ersoy University 11(1): 86-94 (2020)

Araştırma Makalesi / Research Paper

Sağlık Harcamalarının Tahmininde Karar Ağacının Kullanımı

Ersan OKATAN[1]*, Ali Hakan IŞIK[2]

[1]Burdur Mehmet Akif Ersoy Üniversitesi, Gölhisar Uygulamalı Bilimler Yüksekokulu, Burdur
[2]Burdur Mehmet Akif Ersoy Üniversitesi, Mühendislik Mimarlık Fakültesi, Burdur

Geliş Tarihi (Received): 24.11.2019, Kabul Tarihi (Accepted): 17.05.2020
Sorumlu Yazar (Corresponding author): ersanokatan@mehmetakif.edu.tr*
+90 248 2137601 *+90 248 2137610*

ÖZ

Sağlık harcamalarının önceden tahmin edilebilmesi gerek genel bütçe planlamasında gerekse sigortacılık sektöründe hizmet veren kurumların müşterilerine doğru fiyatlandırmayı yapabilmesinde büyük öneme sahiptir. Özellikle sigorta şirketlerinin rekabetçi fiyat teklifleri sunabilmesi ve karlılığını arttırabilmesi için doğru analizler yapması gerekmektedir. Bu çalışmada veri madenciliği yöntemlerinden biri olan karar ağacı kullanılarak sağlık harcaması tahmini yapılmış ve sonuçlar analiz edilmiştir. Açık erişimli Kaggle veri bilimi depolama platformundan alınan veri kümesindeki yaş, cinsiyet, çocuk sayısı, vücut kitle indeksi, sigara kullanma ve bölge bilgileri karar ağacının giriş değerlerini oluşturmaktadır. Sağlık harcaması ise bu değerlere bağlı olarak çıkış değerimizdir. Bu verilerden yaralanarak yapılan tahmin işleminde kullanılan karar ağacı yöntemi üzerinde analizler yapılmıştır. Elde edilen performans sonuçlarının sağlık alanında planlama yapıcılar, sigortacılık alanında hizmet veren kuruluşlar ile bu alanlardaki araştırmacılara yol gösterici olacağı düşünülmektedir.

Anahtar Kelimeler: Karar ağacı, sağlık harcaması, tahmin, veri madenciliği

Utilization of Decision Tree in Prediction of Health Care Costs

ABSTRACT

Prediction of health care cost has a big importance for general budget planning and accurate pricing of institutions which are in insurance sector. In particular, insurance companies need to make accurate analysis for competitive bidding and for increasing profitability. In this study, decision tree which is one of the data mining methods is used to make prediction of health care cost and results are analyzed. The values age, sex, number of child, bmi, region, smoker which taken from the data set given in open access Kaggle data mining data storage platform is input attributes. Health care cost is the label attribute depends on these attributes. Analysis of the decision tree method was performed in this prediction which is made by using these values. Performance results will hope to be helpful for planners on health budget, the insurance companies and researchers on those areas.

Keywords: Decision tree, health care cost, prediction, data mining

GİRİŞ

Türkiye'de kişi başına düşen sağlık harcaması her yıl Gayri Safi Yurtiçi Hasıla (GSYİH)'nın %4-5'i oranında gerçekleşmektedir (Tablo 1). Dünyada ise bu oran ortalama olarak %5-6 arasındadır (Tablo 2).

Ersan OKATAN, https://orcid.org/0000-0001-6511-3450
Ali Hakan IŞIK, https://orcid.org/0000-0003-3561-9375

Tablo 1. Türkiye'de yıllara göre sağlık harcamalarının GSYİH'ya oranı (Atalan, 2018)

Yıllar	KB-SH (GSYİH'NİN %)
2010	5,05
2011	4,69
2012	4,48
2013	4,40
2014	4,35
2015	4,14
2016	4,33
2017	4,20

Tablo 2. Dünya sağlık örgütü verilerine göre Dünya'da sağlık harcamalarının GSYİH'ya oranı (URL-1, 2018)

Yıllar	2017	2016	2015	2014	2013	2012	2011	2010
(WHO) Global	6,3%	6,3%	6,3%	6,2%	6,0%	5,9%	5,9%	6,0%

Sağlık harcamalarına ayrılan kaynak oranı ne kadar yüksekse ülkelerin gelişmişlik oranının o kadar yüksek olduğu düşünülmektedir (Yıldırım ve ark., 2018). Bu kadar yüksek oranlarda kaynak ayrılan sağlık harcamalarının planlanması, koordine edilmesi büyük önem arz etmektedir. Sağlık harcamalarına etki eden faktörlerin belirlenmesi ve bu faktörlerin kullanılması ile ortaya çıkacak tahminler planlayıcılar açısından yararlı olacaktır.

Sağlık harcamalarının tahmin edilmesi sadece ulusal düzeyde plan yapıcılar değil aynı zamanda sigorta şirketleri açısından da çok önemlidir. Sigorta şirketleri özel sağlık sigortası vb. hizmetlerle finans ve sağlık sektörünün önemli aktörlerinden biridir. Sigorta şirketleri açısından yapılacak tahminler, yeni sigorta müşterilerindeki riskin tahmin edilmesi ile birlikte sigorta suistimallerinin tespit edilmesinde de yararlı olacaktır (Akpınar, 2018). Örneğin bir sigorta firması çalışanların bilgileri ile yapacağı doğru bir tahmin sayesinde rakip firmalardan daha avantajlı bir fiyat teklifi sunabilecektir. Duncan ve ark. (2016) yaptıkları çalışmada sağlık reformu ile sigortalı sayısının artırılmaya çalışıldığından, sigorta şirketleri arasındaki rekabeti arttırarak kaynak yaratılmasının hedeflendiğinden bahsetmektedir. Sağlık harcamalarının tahmin edilmesinin faydalı bir diğer yanı ise harcamalarla ilgili değişkenlerin ortaya çıkarılması ve bu değişkenler dikkate alınarak koruyucu hekimlik çalışmalarının daha etkin yapılabilmesidir (Wang ve ark., 2018).

Sağlık harcaması tahmininde yapılmış çeşitli çalışmalar bulunmaktadır. Sushmita ve ark. (2015) çalışmasında sağlık harcaması tahmini için daha önceki harcamalar kullanılmış, M5 karar ağacı, karar ağacı ve rassal orman algoritmaları ile geçerliliği yüksek sonuçlar alınmıştır. Duncan ve ark. (2016) aralarında karar ağacının da bulunduğu modelleme yöntemlerini sağlık harcaması tahmininde kullanmış ve bu yöntemlerin bilinen doğrusal yöntemlerden çok daha etkili olduğu sonucuna ulaşmıştır. Morid ve ark. (2017) çalışmalarında birçok farklı modelleme yöntemini kullanmış ve karşılaştırmıştır. Karar ağacını içeren algoritmalardan biri olan Gradyan Arttırma en iyi sonuca ulaşırken, diğer karar ağacı algoritmaları doğrusal regresyona yakın veya daha düşük sonuçlar ortaya çıkarmıştır.

Bu çalışma 4 bölümden oluşmaktadır. Birinci bölümde Sağlık harcamalarının veri madenciliği yöntemleri ile tahmin edilmesinin neden gerekli olduğu anlatılmıştır. İkinci bölümde kullanılan regresyon yöntemi ve uygulama yapılan veri kümesi ile ilgili bilgiler verilmiştir. Üçüncü bölümde kullanılan yöntem ile elde edilen değerler ve analizleri verilmiş, dördüncü bölümde ise çalışmanın sonucu değerlendirilmiştir.

MATERYAL VE YÖNTEM

Veri Madenciliği, pek çok analiz aracı kullanımıyla veri içerisinde örüntü ve ilişkileri keşfederek, bunları geçerli tahminler yapmak için kullanan bir süreçtir. Veri Madenciliğinin amacı, geçmiş faaliyetlerin analizini temel alarak gelecekteki davranışların tahminine yönelik karar-verme modelleri yaratmaktır (Koyuncugil ve Özgülbaş, 2009). Bu çalışmada sağlık harcamalarının tahmini için veri madenciliği yöntemlerinden Karar Ağacı kullanılmıştır.

Karar ağacı veri madenciliğinde sıklıkla kullanılan yöntemlerden biridir. Sınıflandırma ve regresyonda kullanılabilir. Bir karar ağacı sınıflandırma amacıyla kullanıldığı zaman, genellikle "sınıflandırma ağacı", regresyon amacıyla kullanıldığında ise "regresyon ağacı" olarak adlandırılır (Rokach ve Maimon, 2008). T adlı bir karar ağacının n boyutlu bir özellik uzayında uygulandığını düşünelim (özellik uzayı tipik olarak R^n dir). Özellik vektörlerinin $x \in Z^n$ olduğunu varsayalım. Her v_k dahili noktası $f_k(x)=1\{x_{ik}<t_k\}$ ikili fonksiyonu ile ilişkilendirilmiştir. Burada $i_k \in [n]$, $x \in Z^n$ vektörü için indeks, t_k ise eşik değeridir. Her yaprak düğümü bir z değerine sahiptir. Kök noktasından

başlanarak her bir dahili v_k noktası için $f_k(x)$ değerleri bulunur. $f_k(x)$ değerinin 0 veya 1 oluşuna göre ağacın sol ya da sağ dalı seçilir. Bu işlem l adet yaprak düğümü bulunana kadar devam eder. T(x), z_l yaprak düğümünün çıkış değeridir.

Bir karar ağacının derinliği kökten dala en uzak mesafenin uzunluğudur. Ağacın i katmanı kökten i uzaklıktaki tüm noktalardır. Eğer $0 \leq i \leq d$ ise olmak üzere i katmanı 2^i noktaya sahipse d derinlikteki bir ağaç tamamlanmış demektir (Wu ve ark., 2016).

Normalizasyon, veri madenciliği ve makine öğrenmesi uygulamalarında sıklıkla kullanılan bir ön işleme yöntemidir ve bazı yöntemlerin başarısı için vazgeçilmezdir (Nayak ve ark.,2014). Çeşitli Normalizasyon yöntemleri ve bunların oluşturulan modele etkisi üzerinde yapılmış çalışmalar bulunmaktadır (Al Shalabi ve Shaaban, 2006; Luor, 2015). Bu çalışmada Normalizasyon yöntemi olarak aralık ölçeklendirme (min-max normalization) kullanılmıştır.

Karar ağacında kullanılan önemli yöntemlerden biri budamadır. Karar ağacı çok büyürse istenmeyen ve anlamsız sonuçlar üretebilir buna aşırı öğrenme denir. Aşırı öğrenme sorunundan kurtulmak için budama yöntemi kullanılır (Patel ve Upadhyay, 2012). Ön budama ve sonradan budama olarak iki farklı şekilde yapılır. Ön budamada ağaç oluşurken belirli kriterlerle ağacın dallarının fazla büyümesi önlenir, sonradan budamada ise ağaç oluşur ve ardından bazı dallar kesilerek ağaç küçültülür. Bu çalışmada ön budama kullanılmıştır.

Veri madenciliğinde regresyon sonucu oluşan modellerin performansını değerlendirmek için gerçek değer ile tahmin edilen değer arasındaki farktan yola çıkılarak çeşitli hesaplamalar yapılır. Bunlardan bazıları Açıklayıcılık Katsayısı (R^2), Ortalama Hata Kare Kökü (Root Mean Square Error-RMSE), Ortalama Mutlak Hatadır (Mean Absolute Error-MAE). $p_{u,i}$ tahmin edilen değer, $r_{u,i}$ gerçek değer olmak üzere RMSE değeri için kullanılan formül Eşitlik (1)'de verilmiştir.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{\{u,i\}}\left(p_{u,i} - r_{u,i}\right)^2}{N}} \qquad (1)$$

Tahmin değerleri ile gerçek değerler arasındaki farkların ortalamasına dayanan hesaplama yöntemi olan RMSE farkı öne çıkaran bir yapıya sahiptir. Ancak bu çalışmada farklı aralıklarda normalizasyonlar yapıldığı için ortaya çıkan sonuçlar farklı büyüklüklerde olacaktır bu nedenle tahmin performansının doğrudan RMSE değerleri üzerinden karşılaştırılması mümkün değildir. Örneğin; aynı karar ağacı parametreleri ile normalleştirilmiş verilerde RMSE değeri 13506,8 olurken 0-1 arası normalleştirilen değerlerde 0,26, 0-100 arası normalleştirilen değerlerde 25,91 olarak gerçekleşmiştir. Bu nedenle RMSE değerleri normalleştirilerek kullanılmıştır. NRMSE olarak adlandırılan bu değer RMSE değerinin gerçek değerlerin ortalamasına bölünmesi, gerçek değerlerin en büyüğü ile en küçüğü arasındaki farka bölünmesi veya standart sapma değerine bölünmesi gibi çeşitli şekillerde hesaplanabilir (URL-2, 2019; Shcherbakov ve ark., 2013). Bu çalışmada NRMSE değeri sağlık harcaması değerlerinin ortalamasına bölünerek bulunmuştur. NRMSE değeri kullanılarak farklı büyüklükte değerlerin tahmin edilmesi ile ortaya çıkan performans değerleri birbirleri ile karşılaştırılabilmiştir. Burada değişken değerlerinin normalleştirilmesi ile RMSE değerinin normalleştirilmesi birbirinden bağımsız işlemlerdir.

Veri Kümesi ve Analizi

Bu çalışmada Kaggle.com'dan *Kişisel Sağlık Harcaması Veri Kümesi* kullanılmıştır (URL-3, 2018). Veri kümesi Amerika Birleşik Devletleri'ndeki belirli sayıdaki kişinin sigorta verilerinden alınan yıllık sağlık harcamalarını ve kişisel bilgilerini içermektedir (Lantz, 2013; URL-4, 2019). Tablo 3'te veri kümesinde bulunan değişkenler ile ilgili bilgiler verilmiştir. Sağlık harcaması, karar ağacı yöntemi ile tahmin etmeye çalıştığımız hedef değişkendir.

Tablo 3. Örnek veriler

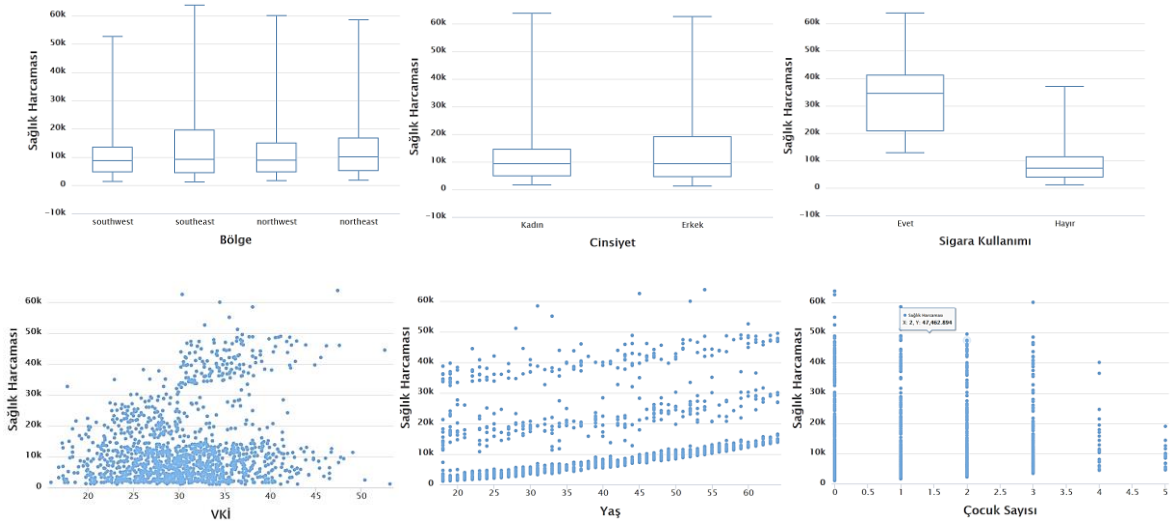
Cinsiyet	**Sigortalı kişinin cinsiyeti**
Sigara Kullanımı	Sigara kullanımı
Bölge	Kişinin ülkesinde yaşadığı bölge bilgisi
Yaş	Kişinin yaşı
VKİ	Vücut kitle indeksi
Çocuk Sayısı	Sigortalının bakmakla yükümlü olduğu çocuk sayısı
Sağlık Harcaması	Sigorta şirketi tarafından faturalandırılan bireysel tıbbi masraflar (yıllık)

Tablo 4'te veri kümesindeki veriler için özet bilgiler gösterilmiştir. Kayıt sayısı 1400'dür ve veri kümesinde kayıp veri bulunmamaktadır.

Tablo 4. Veri kümesi için özet bilgiler

Değişken	**İstatistikler**		
Cinsiyet	Erkek (662)	Kadın (676)	
Sigara Kullanımı	Evet (274)	Hayır (1064)	
Bölge	Southeast (364) Northwest (325)	Southwest (325) Northeast (324)	
Yaş	Min (18)	Mak (64)	Ortalama (39,2)
VKİ	Min (15,96)	Mak (53,13)	Ortalama (30,66)
Çocuk Sayısı	Min (0)	Mak (5)	Ortalama (1,10)
Sağlık Harcaması	Min (1121,87)	Mak (63770,43)	Ortalama (13270,42)

Değişkenlerin sağlık harcaması ile ilişkisini gösteren grafikler Şekil 1'de verilmiştir. Elde edilen grafiklere göre bölge ve cinsiyete göre önemli farklar oluşmadığı görülmektedir. Sigara kullananlar ile kullanmayanlar arasında önemli bir farklılaşma görülmektedir, sigara kullananların harcama tutarı çok daha yüksektir. VKİ ve yaş artışı ile harcama miktarının arttığı, çocuk sayısı arttıkça ise harcama miktarının düştüğü söylenebilir.



Şekil 1. Değişkenlere göre sağlık harcaması grafikleri

Tablo 5'te bağımsız değişkenler için korelasyon matrisi verilmiştir. Görüldüğü gibi değişkenler arasındaki ilişki oldukça zayıftır ve değişkenlerin birbirinden bağımsız olduğu söylenebilir. Değişkenler sağlık harcaması hedef değişkeni üzerinde belirli bir etkiye sahiptir.

Verilerin karar ağacı yöntemi ile modellenebilmesi için sınıfsal verilerin sayısal verilere dönüştürülmesi gerekli-dir. Tablo 6'da sayısallaştırılmış veri örnekleri gösterilmiştir.

Tablo 5. Bağımsız değişkenler için korelasyon matrisi

Değişken	Cinsiyet	Sigara Kullanımı	Bölge	Yaş	VKİ	Çocuk Sayısı	Sağlık Harcaması
Cinsiyet	1	-0,076	-0,005	-0,021	0,046	0,017	0,057
Sigara Kullanımı	-0,076	1	-0,002	0,025	-0,004	-0,008	-0,787
Bölge	-0,005	-0,002	1	-0,002	-0,158	-0,017	0,006
Yaş	-0,021	0,025	-0,002	1	0,109	0,042	0,299
VKİ	0,046	-0,004	-0,158	0,109	1	0,013	0,198
Çocuk Sayısı	0,017	-0,008	-0,017	0,042	0,013	1	0,068
Sağlık Harcaması	0,057	-0,787	0,006	0,299	0,198	0,068	1

Tablo 6. Sayısallaştırılmış veri örnekleri

Cinsiyet	Sigara Kullanımı	Bölge	Yaş	VKİ	Çocuk Sayısı	Sağlık Harcaması
0	0	0	19	27,9	0	16884,924
1	1	1	18	33,77	1	1725,5523
1	1	1	28	33	3	4449,462
1	1	2	33	22,705	0	21984,47061
1	1	2	32	28,88	0	3866,8552
0	1	1	31	25,74	0	3756,6216

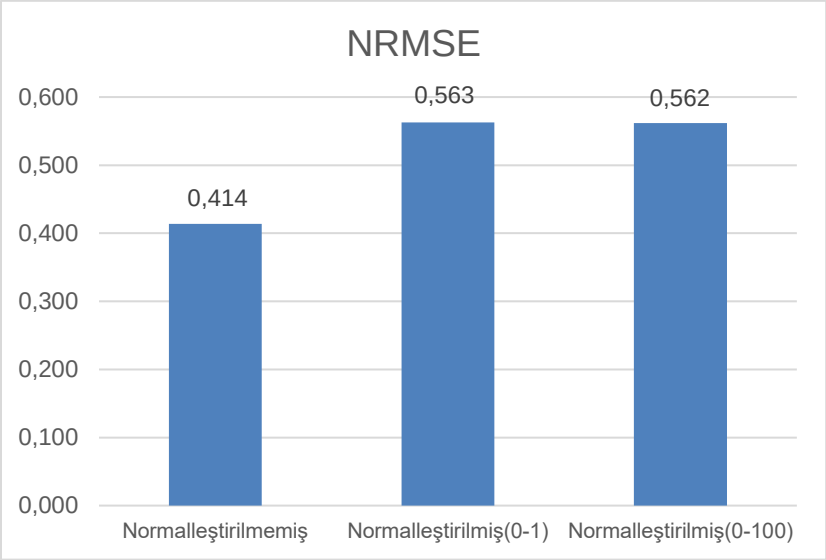
Çalışmada veri kümesinden tabakalı örnekleme ile %30 oranında kısım başarım ölçümü için ayrılmış, geri kalan %70 lik kısım ile model oluşturulmuştur. Veri kümesindeki nominal değerler sayısal değerlere dönüştürülmüştür.

BULGULAR VE TARTIŞMA

Çalışmada RapidMiner adlı veri madenciliği programı kullanılmıştır. Programda karar ağacı algoritması olarak C4.5 veya CART benzeri bir algoritma kullanılmaktadır (Moghimipour ve Ebrahimpour, 2014). Karar ağacı yönteminde farklı metriklerle elde edilen sonuçların NRMSE değerleri elde edilerek karşılaştırılmıştır. NRMSE değerinin düşük olması daha iyi sonucu ifade etmektedir.

Normalizasyonun Tahmin Performansına Etkisi

Veri kümesi kullanılırken normalizasyon için üç farklı yöntem tercih edilmiştir. Değerler normalize edilmemiş, 0-1 değer aralığında ve 0-100 aralığında normalize edilmiştir. 3 farklı yöntemle yapılan tahminlerden elde edilen ortalama NRMSE değerlerinin normalizasyon yöntemine göre ortalaması alındığında ortaya çıkan grafik Şekil 2'de gösterilmiştir. En iyi sonuç normalleştirilmemiş değerlerle elde edilmiştir (0,414). Normalizasyon sonucu elde edilen ortalama NRMSE değerleri 0-1 aralığında normalleştirilen değerler için (0,563), 0-100 aralığında normalleştirilen değerler için (0,562) olarak gerçekleşmiştir. Buna göre normalleştirilmeyen değer ile elde edilen tahmin sonuçlarının daha iyi olduğu söylenebilir.



Şekil 2. Normalizasyon yöntemine göre ortalama NRMSE değerleri

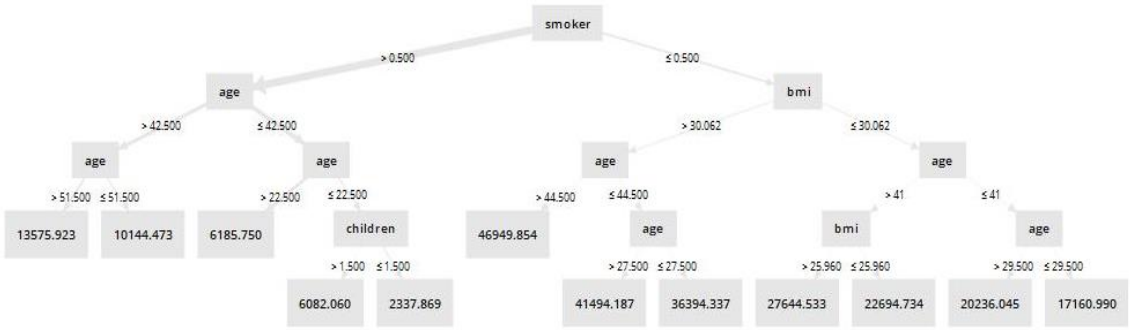
Karar Ağacı Yöntemi ile Yapılan Tahminde Metriklerin Karşılaştırılması

Tablo 7'de normalleştirilmemiş değerler ile elde edilen sonuçlara bakıldığında karar ağacı yönteminde maksimum derinliğin 5 olduğu durumun 10 ve 20 olduğu durumlara göre daha iyi sonuç verdiği görülmektedir. Ön budama yapılmaması durumunda performans daha düşüktür. Ön budama yapıldığında minimum kazanç değeri daha yüksek iken (0,1) düşük olan duruma göre (0,01) NRMSE değeri daha düşüktür ve tahmin performansı daha iyidir.

Tablo 7. Karar ağacı regresyonunda farklı parametre değerleri sonucu NRMSE değerleri

En Çok Derinlik	En Az Kazanç	Normalizasyon	NRMSE
5	0,1	Normalize edilmemiş	0,368
5	0,01	Normalize edilmemiş	0,371
10	0,1	Normalize edilmemiş	0,375
10	0,01	Normalize edilmemiş	0,425
20	0,01	Normalize edilmemiş	0,431
20	Ön budama yok	Normalize edilmemiş	0,512

Şekil 3'te en iyi NRMSE değerine sahip (0,368) karar ağacı modeli verilmiştir. Bu modelde en çok derinlik 5, en az kazanç 0,1 olarak kullanılmış ve değerler normalleştirilmemiştir.



Şekil 3. En iyi NRMSE değerine sahip karar ağacı modeli

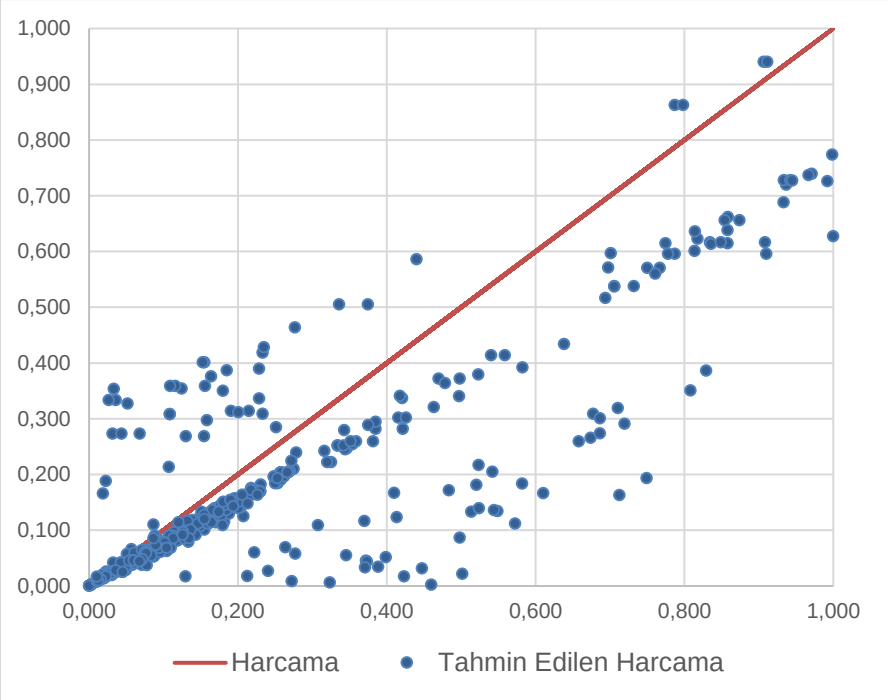
Yapılan Tahminlerde Sonuçların Karşılaştırılması

Tablo 8'de sıralı olarak verilen performans değerlerine bakıldığında normalleştirilmemiş verilerin öne çıktığı ve ilk 6 performansta yer aldığı görülmektedir. Normalizasyon yapılıp yapılmadığına bakılmaksızın en büyük derinlik değerinin 5 ve en küçük kazancın 0,1 olarak belirlendiği durumlarda en iyi sonucun elde edildiği görülmektedir. Budama için kullanılan 0,01 en küçük kazanç değeri diğer parametreler değişmeden de NRMSE değerini arttırmakta ve performansı düşürmektedir.

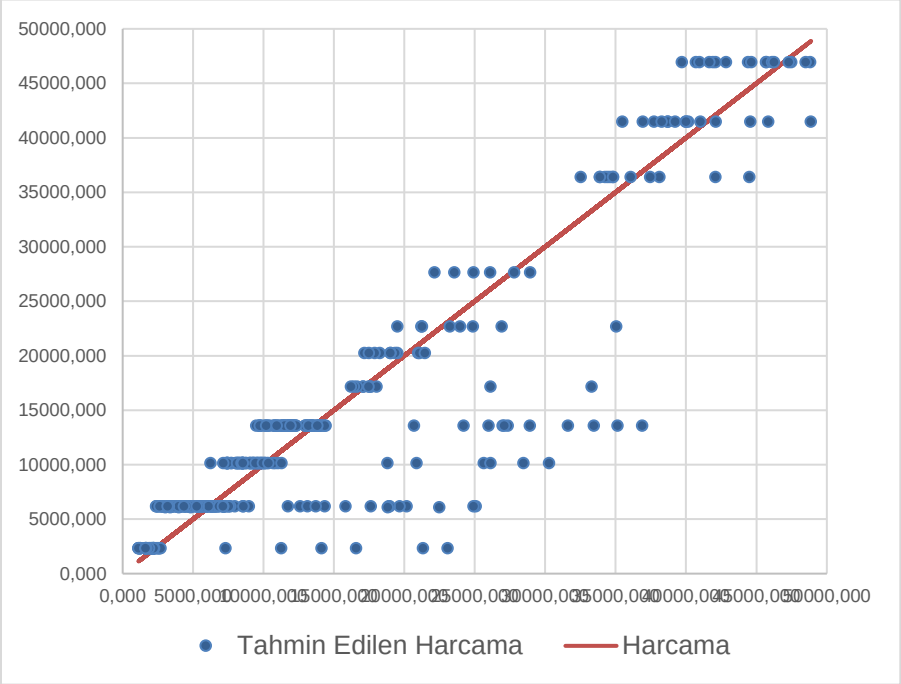
Şekil 4 ve Şekil 5'te en yüksek NRMSE değerine sahip (0,603) ve en düşük NRMSE değerine sahip (0,368) modellerin kullanılması sonucu ortaya çıkan kestirim değerleri ile gerçek değerler arasındaki durumu gösteren iki grafik verilmiştir. İkinci grafiğin doğrusal sonuca daha çok yaklaştığı görülmektedir. NRMSE değerine göre elde edilen en iyi sonuç olan ikinci grafikte küçük değerlerde tahmin performansının daha iyi olduğu, orta ve yüksek değerlerde aynı sonucun elde edilemediği anlaşılmaktadır.

Tablo 8. NRMSE değerleri

RMSE	NRMSE	En Büyük Derinlik	En Küçük Kazanç	Normalizasyon
4968,3	0,368	5	0,1	Normalize edilmemiş
5006,1	0,371	5	0,01	Normalize edilmemiş
5064,1	0,375	10	0,1	Normalize edilmemiş
5739,5	0,425	10	0,01	Normalize edilmemiş
5824,9	0,431	20	0,01	Normalize edilmemiş
6913,1	0,512	20	Ön budama yok	Normalize edilmemiş
0,140	0,542	5	0,1	Normalize edilmiş (0-1)
14,03	0,542	5	0,1	Normalize edilmiş (0-100)
0,141	0,543	5	0,01	Normalize edilmiş (0-1)
14,07	0,543	5	0,01	Normalize edilmiş (0-100)
0,142	0,549	10	0,1	Normalize edilmiş (0-1)
14,25	0,550	10	0,1	Normalize edilmiş (0-100)
0,147	0,567	10	0,01	Normalize edilmiş (0-1)
14,72	0,568	10	0,01	Normalize edilmiş (0-100)
0,148	0,572	20	0,01	Normalize edilmiş (0-1)
14,83	0,572	20	0,1	Normalize edilmiş (0-100)
15,46	0,597	20	Ön budama yok	Normalize edilmiş (0-100)
0,156	0,603	20	Ön budama yok	Normalize edilmiş (0-1)



Şekil 4. En yüksek NRMSE değerine sahip kestirim için gerçek-tahmin değerleri grafiği



Şekil 5. En düşük NRMSE değerine sahip kestirim için gerçek-tahmin değerleri grafiği

Daha önce aynı veriler ile bazı çalışmalar yapılmıştır (URL-3, 2018; Shinde ve Raut, 2018). Kullanılan performans ölçüm parametreleri farklı olduğundan tüm çalışmalarla karşılaştırma yapılamamaktadır. Ancak yapılan bazı çalışmalarda ortaya çıkan RMSE değerleri incelendiğinde iki farklı çalışmada doğrusal regresyon ile elde edilmiş 5641,95 ve 5291,53 değerlerinin bu çalışmada elde edilmiş en iyi RMSE değeri olan 4968,3 değerine göre daha düşük olduğu görülmüştür. Bu değerlere göre bu çalışmada karar ağacı ile yapılan modellemenin daha başarılı olduğu söylenebilir.

SONUÇ

Bu çalışmada bireylere ait verilerden yola çıkılarak karar ağacı yöntemi ile kişiye ait sağlık harcaması tahmin edilmeye çalışılmış ve ortaya çıkan sonuçlar karşılaştırılmıştır. Öğrenme için ayrılan veriler ile karar ağacı modeli oluşturulmuş, kalan %30'luk veri ile test yapılmıştır. Ortaya çıkan sonuçlar ve gerçek değerler ile NRMSE değerleri elde edilmiştir. Çalışma sonucunda farklı metrikler ile yapılan denemelerde normalizasyon yapılmaması durumunda daha iyi performans elde edildiği görülmüştür. Değerlendirme yapılırken farklı metrikler ile elde edilen sonuçların ortalaması alınarak karşılaştırma yapılmıştır. Elde edilen sonuçlarla normalleştirme işleminin her zaman performansı arttırmadığı söylenebilir. Elde edilen sonuçlarda ön budama yapılması ve en çok derinlik değerinin düşmesi ile performansın arttığı gözlemlenmiştir. Budama yapılması aşırı öğrenme sonucu ortaya çıkacak anormallikleri azaltmış, modelin performansını olumlu yönde etkilemiştir. Budama için kullanılan metriklerde en küçük kazanç sınırının yüksek tutulması daha yüksek performans değeri sağlamıştır, bu çok küçük kazanç değerlerinin aşırı budamaya neden olduğunu göstermektedir. Budama için kullanılan en çok derinlik değerinin yüksek olmasının performansı düşürdüğü ve yetersiz budamaya neden olduğu görülmektedir.

Çalışmanın veri madenciliği yöntemleri kullanarak sağlık harcamaları veya diğer verilerin tahmini konusunda yapılacak araştırmalara Normalizasyon, yöntem ve parametre seçimi konularında yol gösterici olacağı düşünülmektedir.

BİLGİLENDİRME

Bu çalışma, 2nd International Health Sciences and Life Congress (IHSLC 2019)' te sözlü bildiri olarak sunulmuş ve özet olarak bildiri kitapçığında yayınlanmıştır (Okatan ve Işık, 2019).

KAYNAKLAR

Akpınar, Ö. (2018). Sigorta Sektöründe Veri Madenciliği Ve Kullanım Alanları. *Dumlupınar Üniversitesi Sosyal Bilimler Dergisi* 57: 103-119.

Al Shalabi, L., Shaaban. Z. (2006). Normalization as a preprocessing engine for data mining and the approach of preference matrix. *International Conference on Dependability of Computer Systems*, May 24-28, 2006, Szklarska Poreba, Poland, Proceedings, 207-214.

Atalan, A. (2018). Türkiye Sağlık Ekonomisi İçin İstatistiksel Çok Amaçlı Optimizasyon Modelinin Uygulanması. *İşletme Ekonomi ve Yönetim Araştırmaları Dergisi*, 1(1): 34-51.

Duncan, I., Loginov, M., Ludkovski, M. (2016). Testing Alternative Regression Frameworks for Predictive Modeling of Health Care Costs. *North American Actuarial Journal*, 20 (1): 65-87.

Koyuncugil, A., Özgülbaş, N. (2009). Veri madenciliği: Tıp ve sağlık hizmetlerinde kullanımı ve uygulamaları. *Bilişim Teknolojileri Dergisi* 2(2): 21-32.

Lantz, B., (2013). *Machine Learning with R*. Packt Publishing, England.

Luor, D. C. (2015). A comparative assessment of data standardization on support vector machine for classification problems. *Intelligent Data Analysis,* 19(3): 529-546.

Moghimipour, I., Ebrahimpour, M. (2014). Comparing Decision Tree Method Over Three Data Mining Software*. International Journal of Statistics and Probability,* 3(3): 147-156.

Morid, M. A., Kawamoto, K., Ault, T., Dorius, J., Abdelrahman, S. (2017). Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation. *AMIA Annual Symposium Proceedings*, Nov 6-8, 2017, Washington, USA, 1312-1321.

Nayak, S.C., Misra, B.B., Behera, H.S. (2014). Impact of Data Normalization on Stock Index Forecasting. *International Journal of Computer Information Systems and Industrial Management Applications* 6: 257-269.

Okatan, E., Işık, A.H. (2019). Karar Ağacı Yöntemi İle Sağlık Harcaması Tahmini. *2nd International Health Science And Life Congress*, April 24-27, 2019, Burdur, Turkey, Abstract Book, 500.

Patel, N., Upadhyay, S. (2012). Study of various decision tree pruning methods with their empirical comparison in WEKA. *International Journal of Computer Applications* 60(12): 20-25.

Rokach, L., Maimon, O. Z. (2008). *Data Mining With Decision Trees: Theory And Applications*. World Scientific Publishing Co. Pte. Ltd, Singapore.

Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A., & Kamaev, V. A. E. (2013). A survey of forecast error measures. *World Applied Sciences Journal*, 24(24): 171-176.

Shinde, A., Raut, P. (2018). Comparative Study of Regression Models and Deep Learning Models for Insurance Cost Prediction. *International Conference on Intelligent Systems Design and Applications*, Dec 6-8, 2018, Vellore, India, Conference proceedings, 1102-1111.

Sushmita, S., Newman, S., Marquardt, J., Ram, P., Prasad, V., Cock, M. D., & Teredesai, A. (2015). Population cost prediction on public healthcare datasets. *5th International Conference on Digital Health*, May, 2015, Florence, Italy, Proceedings, 87-94.

URL-1. (2018). http://apps.who.int/gho/data/view.main.GHEDC-HEGDPSHA2011REGv?lang=en. WHO (World Health Organization) Global Health Observatory Data Repository. (Erişim Tarihi: 10.09.2019)

URL-2. (2019). https://www.marinedatascience.co/blog/2019/01/07/normalizing-the-rmse/ (Erişim Tarihi: 15.09.2019)

URL-3. (2018). https://www.kaggle.com/mirichoi0218/insurance (Erişim Tarihi: 15.05.2019).

URL-4. (2019). https://gist.github.com/meperezcuello/82a9f1c1c473d6585e750ad2e3c05a41 (Erişim Tarihi: 01.05.2020)

Wang, Y., Kung, L., Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change* 126: 3-13.

Wu, D. J., Feng, T., Naehrig, M., Lauter, K. (2016). Privately evaluating decision trees and random forests*. Proceedings on Privacy Enhancing Technologies* 2016(4): 335-355.

Yıldırım, Z., Kekeç, H. M., Polat, A. (2018). Türkiye'de Sağlık Harcamaları Ve Finansmanının Yıllar İtibariyle Analizi. *Gazi Üniversitesi Sosyal Bilimler Dergisi* 5(14): 550-563.