İŞLETME ARAŞTIRMALARI DERGİSİ
JOURNAL OF BUSINESS RESEARCH-TURK
2021, 13(1), 450-463
https://doi.org/10.20491/isarder.2021.1144

Otel Oda Fiyatlarını Açıklamada Makine Öğrenmesi Algoritmalarının Kıyaslanması
(Comparison of Machine Learning Algorithms in Explaining of Hotel Room Prices)

Yılmaz AĞCA [a]
[a] Tokat Gaziosmanpaşa Üniversitesi, Niksar Sosyal Bilimler Meslek Yüksekokulu, Tokat, Türkiye. agca.yilmaz@gmail.com

MAKALE BİLGİSİ

Anahtar Kelimeler:
Makine öğrenmesi
Otel oda fiyatları
Hedonik fiyatlandırma
Web madenciliği

Gönderilme Tarihi 24 Ekim 2020
Revizyon Tarihi 16 Şubat 2021
Kabul Tarihi 10 Mart 2021

Makale Kategorisi:
Araştırma Makalesi

ÖZET

Amaç – Çalışmada, makine öğrenmesi algoritmalarından bazılarının, web madenciliği ile elde edilen büyük veri kullanılarak, analiz edilmesi ve otel oda fiyatlarını açıklama performanslarının test edilmesi amaçlanmaktadır. Böylelikle otel oda fiyatlarını en doğru açıklayan modelin belirlenmesidir.

Yöntem – Verinin elde edilmesinde web madenciliği/web kazıma yöntemi kullanılmıştır. Hedef web sitesi geliştirilen bir algoritma yardımıyla yaklaşık altı ay boyunca taranmış ve elde edilen 6558 konaklama tesisine ait veriler, analizlerde kullanılmıştır. Araştırmanın ikinci kısmı istatistiksel analizlerden ve makine öğrenmesi algoritmalarının uygulanmasından oluşmaktadır. Analizlerin yapılması ve algoritmaların uygulanması için Python programlama dili kullanılmıştır. Bu dile ait kütüphaneler, pandas, numpy veri işleme için, seaborn, matplotlib grafikler ve görselleştirme için, scikit-learn ise makine öğrenmesi algoritmaları için kullanılmıştır. Analizlerden sonra veri için en uygun olduğu düşünülen lojistik regresyon ile bir model oluşturulmuştur.

Bulgular – Karşılaştırılan Rassal Orman ve Karar Ağacı algoritmalarının her ikisinin de yaklaşık %99,89 oranında veri setini açıkladığı dolayısıyla ağaç/dallanmaların başarı ile gerçekleştiği görülmektedir. KNN algoritması ise en yüksek performansı üç kümeli bir sınıflandırma ile %62,12 oranında gerçekleştirmiştir. Doğrusal sınıflandırma yöntemini kullanan Lojistik Regresyon, Olasılıksal Dereceli Azalma ve Destek Vektör Makineleri algoritmalarından en yüksek skoru %39,13 ile lojistik regresyon yöntemi elde etmiştir. Lojistik regresyon ile oluşturulan modelde, konukların tesise verdikleri puan, tesisin bölgede diğer tesisler arasındaki sırası, tesisin türü ve bulunduğu şehir istatistiki olarak anlamlı bulunmuştur ($p<0,05$).

Tartışma – Araştırma sonucunda, elde edilen makine öğrenmesi performans skorlarının yanı sıra, Türkiye'de otel oda fiyatları hakkında önemli bilgiler elde edilmiştir. Oluşturulan regresyon modeli ile 44 bağımsız değişkenden hangilerinin otel oda fiyatlarını açıklamada anlamlı olduğu ortaya konulmuştur.

ARTICLE INFO

Keywords:
Machine learning
Hotel room price
Hedonic price
Web mining

Received 24 October 2020
Revised 16 February 2021
Accepted 10 March 2021

Article Classification:
Research Article

ABSTRACT

Purpose: The aim of the study is to analyze some of the main machine learning algorithms using big data obtained by web mining and to test the performance of these algorithms to explain hotel room prices. Thus, it is the determination of the model that best explains the hotel room prices.

Design/methodology/approach – Web mining/scraping method was used to obtain research data. The target website was scanned for about six months with the help of an algorithm, and the data obtained from 6558 accommodation facilities were used in analysis. The second part of the research consists of statistical analysis and comparison of machine learning algorithms. Python programming language was used for analysis and implementation of algorithms. Pandas, numpy libraries for data processing; seaborn, matplotlib for graphics and visualization; scikit-learn is used to run machine learning algorithms. After the analysis, a model was created by logistic regression, which was thought to be the most suitable for the data.

Results: It is seen that the compared Random Forest and Decision Tree algorithms both explain the data set at a rate of approximately 99.89%, so the tree/branching has been successful. The KNN algorithm achieved the highest performance with a classification of three clusters at 62.12%. Logistic Regression, Stochastic Gradient Decent and Support Vector Machines using the linear classification method obtained the highest score with 39.13% logistic regression method. In the model created by logistic regression, the score given to the hotel by the guests, the rank of the hotel among other hotels in the region, the type of the hotel and the city in which it is located were found to be statistically significant ($p<0.05$).

Discussion: As a result of the research, machine learning algorithms were compared using the hotel room price data that obtained from web mining/scraping. It was also found important information about the hotel room rates in Turkey. A regression model has revealed which of the 44 independent variables are significant in explaining the hotel room price.

Önerilen Atıf/ Suggested Citation
Ağca, Y. (2021). Otel Oda Fiyatlarını Açıklamada Makine Öğrenmesi Algoritmalarının Kıyaslanması, *İşletme Araştırmaları Dergisi*, 13 (1), 450-463.

1. GİRİŞ

Bir ürünün fiyatının, o ürünü meydana getiren ürün ve hizmetlerin fiyatlarının birleşiminden meydana geldiği yaklaşımına *hedonik fiyatlandırma* denilmektedir (Monson, 2009:62; Löch ve Axhausen, 2010:39; Yayar ve Gül, 2014:87; Andersson, 2008:231; Castro, Ferreira ve Ferreira, 2016:691). Hedonik fiyatlandırma yaklaşımına iyi örneklerden biri olarak *turizm ürünü* gösterilebilir. Turizm ürünü veya turistik ürün, birçok ürün ve hizmetin birleşiminden meydana gelmekte, tüketiciye bir paket halinde sunulmaktadır. Örneğin bir tatil paketi, konaklama, transferler, rehberlik ve ulaşım gibi hizmetlerin birleşiminden oluşmaktadır. Bu alt ürünlerin de her biri başka hizmet ve ürünlerden oluşabilmektedir. Örneğin konaklama hizmeti, oda, kahvaltı, resepsiyon, konsiyerj gibi hizmetlerden oluşmaktadır. Bu örnekleri hem turizm sektöründe hem de diğer sektörlerde daha da çoğaltmak mümkündür.

Bir ürünün fiyatını meydana getiren diğer ürünlerin fiyattaki payını belirleme çoğu sektörde önemli bir araştırma konusudur. Bu oranlar hem tüketici hem de üretici açısından arz ve talebin oluşmasında etkilidir. Örnek vermek gerekirse, bir tüketici bir elektronik cihaz satın alırken o cihazın sahip olduğu özellikler, talep üzerinde etkilidir. Benzer şekilde üretici de piyasaya sürdüğü cihazların özelliklerine göre fiyatlandırma yapabilmektedir. Her ne kadar bu fiyatlandırma yaklaşımı basit bir mantığa dayansa da bir ürünün fiyatında o ürünü oluşturan parçaların oranını belirlemek çoğu kez kolay olmamaktadır. Çoğu üründe bu duruma oranların belirgin bir yapıya sahip olmaması neden olmaktadır. Fakat yine de bu yöntem yaygın kullanılan bir fiyatlandırma modelidir.

Ürün fiyatını meydana getiren ürün ve hizmetlerin oranını belirlemek için kullanılan en eski ve bilinen yöntemlerden biri regresyondur. Duruma göre lineer veya lojistik regresyon ile oluşturulan model, bağımlı fiyat değişkeni üzerinde bağımsız değişkenlerin oranını verebilmektedir. Regresyon modelleri güçlü ve yaygın olmakla birlikte uygulaması zor değildir. Yıllardır bu modeller pek çok ürün ve sektörde kullanılmaktadır. Gelişen bilgi ve iletişim sektörü ile birlikte regresyona alternatif yöntemler ortaya çıkmıştır. Bu yöntemlerin birçoğu geçmişte düşük işlem gücüne sahip bilgisayarlar ile mümkün olmayan veya zor olan yöntemlerdir.

20.yy'ın ikinci yarısından sonra insan hayatına giren bilgisayarlar; sosyal, kültürel ve ekonomik yönden, kısacası her alanda büyük gelişmelere neden olmuştur. İlk başlarda sadece hükümetler ve büyük kuruluşlar tarafından kullanılan bu cihazlar, günümüzde neredeyse bütün elektrikli aletlerin içine girmiştir. Üretildikleri ilk zamanlarda boyut olarak büyük ve ağırlık olarak tonları bulan bu cihazlar, günümüzde mikroişlemci teknolojisinin gelişimi ile gram olarak ölçülecek hale gelmiştir. Zamanla bilgisayarların boyutları küçülüp maliyetleri azalırken, işlem güçleri ise artmıştır. İşlem gücünün artması, *veri madenciliği, makine öğrenmesi, yapay zekâ* gibi kavramların ortaya çıkmasına neden olmuştur. Günümüzde bu teknolojiler, kendi kendine giden otomobillerden, fotoğraflardaki nesneleri tanıyan algoritmalara kadar pek çok alanda kullanılmaktadır.

Bilgisayarların insan hayatı üzerinde bu kadar etkili olmasındaki bir diğer faktör *internettir*. Basitçe bilgisayarların birbirlerine bağlanması ile oluşturulan bu küresel ağ sayesinde dünyanın herhangi bir yerinden başka bir yerine veri iletilebilmektedir. Saniyeler içerisinde iletilen büyük miktardaki veri sayesinde, görüntülü haberleşmeden, bir kütüphanedeki kitaba erişmeye kadar pek çok iş bu ağ üzerinden yapılabilmektedir.

İnternetin insan hayatına etkisi her alanda kendini göstermiştir. Bu etkilerden biri de alışverişte olmuştur. Araştırmanın veri kaynağını da oluşturan internet üzerinden alışveriş, günümüzde çoğu kişi için sıradan bir şey haline gelmiştir. İnsanlar perakende dükkanlarına gitmek yerine internet erişimi olan cihazlar vasıtasıyla, ödemeyi kredi kartı gibi alternatif yöntemlerle yaparak istedikleri ürünleri sipariş etmektedir. Örneğin bir tatil paketi satın almak isteyen tüketici, perakendeci bir seyahat acentasına gitmek yerine alternatifi olan *online seyahat acentası* web sitesinden aynı işlemi yapabilmektedir.

Perakendeci seyahat acentelerinde kataloglardan yapılan tatil/paket tur, konaklama tesisi pazarlaması, online seyahat acentelerinde web siteleri üzerine, yani internete taşınmıştır. Bu web siteleri, acente kataloglarında olması mümkün olmayan tesislere ait pek çok bilgiyi üzerinde barındırdığından, tüketici açısından büyük kolaylık ve avantaj sağlamaktadır (Keskinkılıç, Ağca ve Karaman, 2016:445). Bu nedenle de günümüzde online seyahat acenteleri pazar paylarını hızla artırmaktadır.

İnternet, günlük kullanıcılara sağladığı kadar araştırmacılara da büyük kolaylıklar sağlamaktadır. İnternet, üzerinde bulundurduğu veri ile günümüzdeki en büyük veri tabanlarından biridir. Binlerce işletmenin veri tabanlarından yayınlanmış veriler ve milyarlarca kişiye ait çeşitli bilgiler, e-ticaret siteleri, bloglar, sosyal ağlar, video yayın siteleri, haber siteleri üzerinde tamamen kamuya açık bir şekilde bulunmaktadır. Dağınık halde olan ve bilgi çıkarımında bulunulması sadece insan çabası ile güç olan bu veri, sistemli hale getirildiği zaman çok değerli olabilmektedir. Dağınık halde bulunan bu verinin çeşitli yöntemler ile bilgi çıkarımında kullanılabilir hale getirilmesine *web madenciliği* denilmektedir. Bu çalışmada veri elde etmek için web madenciliği yöntemi kullanılmıştır.

Bahsedildiği üzere internet muazzam büyüklükte bir veri tabanıdır. Fakat bu veri tek bir yapı üzerinde veya tam standart halde bulunmaz. Web sitelerinde genellikle ilişkisel veri tabanlarında kayıtlı olan verilerin sadece istenilen kısımları kullanıcılarla paylaşılmaktadır. Buna rağmen bu veriler bile araştırmacılara önemli çıkarımlarda bulunabilecek imkanlar sağlar. Web madenciliği olarak adlandırılan bu bilgi çıkarsama kavramı tek bir yöntem veya yazılım değildir. Sonucu, sistemli veri elde etme olan birçok yöntem ve algoritmadan oluşan yöntemler bütünüdür (Ağca, 2019:55).

Bir araştırmada problemi çözüme kavuşturacak doğru ve yeterli miktarda veriyi elde etmek çoğu zaman araştırmanın önemli bir aşamasını oluşturmaktadır. Bu çalışmada kullanılan web madenciliği, klasik veri elde etme yöntemleri olan *anket* veya *görüşme* gibi yöntemlere elde edilmesi çok zor olan veya olmayan miktar ve nitelikte veri elde etmeyi olanaklı kılmıştır.

Araştırmada, hedef online seyahat acentesi sitesi üzerinde yer alan konaklama tesislerine ait nitelik ve fiyat bilgisi, bir sezon boyunca geliştirilen bir *web madenciliği algoritması* ile elde edilmiştir. Elde edilen veriler çeşitli işlemlerden geçirilerek *makine öğrenmesi algoritmalarına* hazır hale getirilmiş (Dontha, 2018; García, Luengo ve Herrera, 2015:39; Han, Kamber ve Pei, 2012:5-7) ve kullanılmıştır. Bu algoritma ile 6558 konaklama tesisine ait 44 özellik/nitelik bilgisi ve 19 haftalık fiyat bilgisi elde edilmiş ve analizlerde kullanılmıştır. Çalışmada, makine öğrenmesi algoritmalarından bazıları, elde edilen büyük veri kullanılarak işletilmiş ve otel oda fiyatlarını açıklama performansları test edilmiştir. Buradaki amaç bool[1] (boolean) tipinde olan otellerin bazı özelliklere sahip olup olmadığını ifade eden, bağımsız değişkenler ile, bağımlı fiyat değişkenini en iyi açıklayan makine öğrenmesi algoritmasını ortaya koymaktır.

Bu çalışma, birçok açıdan, geçmişte hem makine öğrenmesi algoritmalarının performansını ölçmeyi amaçlayan çalışmalara (Williams, Zander ve Armitage, 2006:7; Isuhuaylas, Hirata, Santos ve Torobeo, 2018:782; Noi ve Kappas, 2017:1; Tomiazzi, Pereira, Judai, Antunes ve Favareto, 2019:6481) hem de otel oda fiyatlarını etkileyen değişkenleri araştıran çalışmalara (Carvell ve Herrin, 1990:27; Bull, 1994:10; Taylor, 1995:169; Papatheodorou, 2002:133; Aguiló, Alegre ve Sard, 2003:255; Espinet, Saez, Coenders ve Fluvià, 2003) katkı sağlamayı ve boşlukları doldurmayı amaçlamaktadır. Araştırma verisi elde edilirken kullanılan web madenciliği, yeni, alternatif ve verimli bir veri elde etme yöntemidir. Çalışmanın, bu haliyle Türkiye'de konaklama tesisleri üzerine yapılan en kapsamlı nitel çalışmalardan biri olması hedeflenmektedir. Bu çalışma, uluslararası yazında otel oda fiyatlarının makine öğrenmesi algoritmalarıyla analizi (Razavi ve Israeli, 2019:2149) konusunda da çok az rastlanılan çalışmadan biri olacaktır.

2. KAVRAMSAL ÇERÇEVE

Araştırmada oda fiyatları analiz edilirken makine öğrenmesi algoritmaları kullanılmıştır. *Makine öğrenmesi* mevcut verinin karakteristik özelliklerinin belirlenmesi vasıtasıyla bilgi çıkarımında bulunulması ve bu *öğrenme süreci* sayesinde yeni gelen veri üzerinde tahminde bulunulması temeline dayanmaktadır. Makine öğrenmesi son yıllarda popülerleşen kavramlardan biridir. Makine öğrenmesinde kullanılan bazı yöntemler on yıllardır bilinmesine rağmen, kavramın yaygınlaşması ve yerleşmesi için bazı teknolojik gelişmelerin yaşanması gerekmiştir.

2.1. Gelişen Bilişim Teknolojisi ve Büyük Veri Kavramı

Her nesne, her olay ve süreç belirli miktarda veri üretmektedir. Örneğin bir kişinin evinden işine gitmesi veri üreten bir süreçtir. Fakat bu verinin belirli bir amaç için kullanılabilmesi için öncelikle depolanabilmesi

[1] Doğru (True) veya Yanlış (False), Sıfır (0) veya Bir (1) değerlerini alan değişken tipidir (Lutz, 2009, s. 250). Araştırmadaki veri setinde bir özelliğin olması veya olmaması şeklinde yer almaktadır.

gerekmektedir. Büyük miktardaki verinin depolanabilmesi, çoğu kişinin tahmin edebileceği üzere depolama birimlerinin kapasitesinin artması ve depolama maliyetinin azalması ile mümkün olmuştur. Daha önce depolanmayan çoğu veri depolama donanımlarındaki bu gelişmelerle birlikte depolanmaya başlanmıştır. Makine öğrenmesine giden süreç sadece verinin depolanabilmesi ile ilgili değildir. Kullanılan mobil teknolojiler, internet ve sensorler muazzam büyüklükte bir verinin üretilmesine imkân tanımıştır. Öyle ki, örnekte verilen olayda kişinin işe giderken üzerinde taşıdığı akıllı telefon veya saat, sahip olduğu sensorler aracılığıyla sürekli veri üretmektedir.

Bu teknolojik gelişmeler olurken bilişim alanında çalışanlar ve bilim adamları tarafından verilerin bilgi çıkarımında bulunmak veya çeşitli tahminlerde bulunabilmek için kullanılabileceği fikri gelişmiştir. Daha önce de vurgulandığı gibi veri analizinde kullanılan yöntemlerin bazıları daha önce de bilinmekte ve kullanılmakta idi fakat analiz edilen verinin büyüklüğü bir darboğaz oluşturuyordu. Bunun için de teknolojinin belirli bir olgunluğa erişmesi gerekmekteydi. Bu da bilgisayarların işlem güçlerinin gelişmesi ile meydana geldi. Bu gelişmeye de bir örnek vermek gerekirse 1946'da bir süper bilgisayar olan ENIAC'ın işlem gücü yaklaşık 500 FLOPS[2] iken, günümüzün süper bilgisayarları yaklaşık 93 petaFLOPS hızında işlem yapabilmektedir. Kıyaslamanın daha iyi anlaşılması için şöyle bir açıklama yapılabilir. Günümüzde birkaç yüz gram ağırlığındaki standart bir cep telefonu teorik olarak yaklaşık 200 gigaFLOP işlem yapabilmektedir (Frumusanu, 2018) ve bunu yapmak için gerekli olan enerjiyi ise üzerinde taşıdığı yaklaşık 3000mAh kapasiteli bir pilden almaktadır. Sonuç olarak bilgisayarlar, internet, dijital veri, sosyal ağlar, bloglar, mobil aygıtlar, sensorler bir araya geldiğinde büyük veri kavramını meydana getirmiş (Ağca, 2019:38-54), büyük veri kullanılarak makine öğrenmesi algoritmalar geliştirilmiş ve çeşitli amaçlarla kullanılmaya başlanmıştır.

Makine öğrenmesi algoritmaları; arama motorları, görüntü işleme, el yazısı ve ses tanıma, yapay zekâ, otonom sürüş, tıbbi teşhis ve burada yer verilmeyen onlarca başka alanda kullanılmaktadır. Teknolojik gelişmelerin sağladığı imkanlar ile ortaya çıkan ve gelişen makine öğrenmesi, diğer birçok alanda olduğu gibi, bilimsel araştırmacıların da ilgisini çekmektedir. Öngörüler, bu teknolojinin kullanımının daha da artacağı yönündedir (Maksymenko, 2020; Rose, 2020).

3. LİTERATÜR İNCELEMESİ

Bu algoritmalar, turizm sektöründe de diğer iş alanlarındaki gibi, çeşitli araştırmalar için kullanılmaktadır. Bunlara örnek vermek gerekirse; hangi değişkenlerin rezervasyon iptallerinde etkili olduğu (Antonio, Almeida ve Nunes, 2017:1049; Sánchez-Medina ve C-Sánchez, 2020:1), coğrafi bilgi sistemi (GIS/Geographic Information System) verisi kullanılarak otel konumlarının analizi (Yang, Tang, Luo ve Law, 2015:14), müşteri/konuk yorumlarının analiz edilmesi (Ku, Chang, Wang, Chen ve Hsiao, 2019:5268), internet arama motoru verisi kullanılarak turist gelişlerinin hesaplanması (Sun, Wei, Tsui ve Wang, 2019:1) gibi bir çok farklı türde çalışma makine öğrenmesi imkanları ile turizm verisi üzerinde yapılmıştır.

Otel oda fiyatını etkileyen faktörlerin belirlenmesine yönelik çeşitli çalışmalar yapılmıştır. Bu çalışmaların ortak noktası araştırma verisini anket, katalog taraması gibi klasik yöntemlerle yapmaları, dolayısıyla sınırlı bir örneklem üzerinde çalışmaları ve araştırmayı geniş bir coğrafyaya genişletememeleri olmuştur (Carvell ve Herrin, 1990:27; Bull, 1994:10; Taylor, 1995:169; Papatheodorou, 2002:133; Aguiló, Alegre ve Sard, 2003:255; Espinet, Saez, Coenders ve Fluvià, 2003). Bu çalışmalardan farklı olarak web madenciliğini kullanarak otel oda fiyatlarını etkileyen faktörleri araştıran Razavi ve Israeli (2019), araştırmada kullandıkları veriyi web madenciliği ile Trivago sitesinden elde etmişlerdir. Araştırmacılar web crawler yardımı ile ABD'deki Manhattan bölgesinden 309 otele ait fiyat verisi alıp analiz etmiştir. Bu çalışmada da *makine öğrenmesi* algoritmaları kıyaslanmış ve veriyi açıklamadaki performansları test edilmiştir. Razavi ve Israeli (2019), veriyi Regresyon (OLS), Karar Ağaçları (CART, Random Forest ve Gradient Boosting Machines), Destek Vektör Makineleri (SVMs) ve Yapay Sinir Ağları (NNs) algoritmaları ile analiz etmişlerdir. Elde edilen skorlar 0,32-0,46 aralığında olmuştur. En düşük skor 0,32 puanla Regresyon modeli ile elde edilirken, en yüksek skor ise 0,46 puanla Gradient Boosting Machines'e aittir. Çalışmada ortalama oda fiyatı değişkenini açıklamak için

[2] FLOPS: mikroişlemcilerin hızlarını ifade eden ölçü birimlerinden biridir. Kısaltma, saniyede kayan noktalı sayı işlemi (floating-point operations number) ifade eder. gigaFLOPS 10^9 FLOPS iken, petaFLOPS 10^{15} FLOPS'dur (Lebanon ve El-Geish, 2018, s. 11).

sadece üç değişken olan, otellerin yıldızları, konukların verdiği puanlar ve konukların yorum sayıları kullanılmıştır (Razavi ve Israeli, 2019:2149).

Shehhi ve Karathanasopoulos (2020) yaptıkları çalışmada STR'den (Smith Travel Research) elde ettikleri veriyi kullanarak SARIMA ve ANFIS gibi modellerin performanslarını test etmişlerdir. Elde ettikleri sonuçlar, otel oda fiyatlarını açıklamakta, makine öğrenmesi algoritmalarının geleneksel istatistiki yöntemlere benzer sonuçlar verdiği yönündedir (Shehhi ve Karathanasopoulos, 2020:40).

4. YÖNTEM

Çalışmada kullanılan yöntemler kendi içlerinde çeşitli basamaklara ayrılsa da bütün olarak iki kısma ayrılabilir. İlk kısım verinin toplanması, ikinci kısım verinin analiz edilmesinden oluşmaktadır. Bu aşamalar detaylı olarak şu şekilde yürütülmüştür.

Verinin elde edilmesinde web madenciliği kullanılmıştır. Web madenciliği webi büyük bir veri kaynağı olarak ele alarak, bu ağ üzerindeki web sayfalarından bilgi derlemeyi olanaklı hale getiren yöntemler bütünüdür (Ağca, 2019:55). Web madenciliği için hedef web sitesi seçilirken araştırmada kullanılacak uygun ve yeterli miktarda veriyi içermesi, verinin elde edilmesine engel olmaması (bot, örümcek ve robot engelleyici) ve kaynak kodlarının sistemli olması, kriterlerini taşımasına dikkat edilmiştir. Hedef olarak belirlenen; veri tabanı geniş ve çok ziyaret alan siteler üzerinde crawler yazılımları ile pilot çalışma yapılmış, yukarıda verilen kriterlere göre en uygun olan *tripadvisor.com* online seyahat acentası/sitesi tercih edilmiştir. Veriler bu siteden 01.05.2018-30.09.2018 tarihleri arasında haftada bir kez olarak şekilde elde edilmiştir. Fiyatlar tek oda bir gecelik iki kişi olarak alınmıştır. Bu tarihler arasında fiyatlar değişkenlik göstermektedir. Bu nedenle analizlerde kullanılan fiyatlar 19 haftanın ortalaması olarak alınmıştır.

Verinin elde edileceği site belirlendikten sonra sitenin kaynak kodları incelenerek gerekli olan *XPath* ve *CSS* sorguları oluşturulmuştur. Web kazıma için açık kaynak kodlu *Knime Analytics Platform v3.6* yazılımı tercih edilmiş fakat algoritmanın belirli yerlerinde *Python* programlama dilinden ve bu dilin *BeautifulSoup* ve *Scrapy* gibi kütüphanelerinden faydalanılmıştır. Veri temizleme, düzenleme ve analizlere hazırlama aşamasında *Microsoft Excel v.2013* daha yüksek işlem gücü gerektiren uygulamalar için *Google Docs Cloud* servisi kullanılmıştır.

Veriler temizlendikten sonraki haliyle, bağımlı *Otel Oda Fiyat* değişkeni sürekli değişken formundadır. Otellerin sunduğu hizmetleri ve donanımları tutan bağımsız değişkenler ise web sitesinden alındığı halinden dönüştürülerek, hizmetin veya donanımın olması "1" olmaması "0" şeklinde yeniden kodlanmıştır. Bilineceği üzere çoğu makine algoritması için bağımlı veya bağımsız değişkenin *kategorik* formda olsa dahi, *numerik* olması gerekmektedir. Bu nedenle bu türden bir kodlama işlemi yapılmıştır. Daha sonraki aşamalarda yapılacak olan istatistiksel analize ve makine öğrenmesi algoritmasına göre *bağımlı değişken* sürekli veya sıralı kategorik hale getirilmiştir.

Yöntemin ikinci kısmı istatistiksel analizlerden ve makine öğrenmesi algoritmalarının uygulanmasından oluşmaktadır. Analizler yapılması ve algoritmaların uygulanması için Python programlama dili kullanılmıştır. Bu dile ait kütüphaneler, *pandas, numpy* veri işleme için, *seaborn*, *matplotlib* grafikler ve görselleştirme için, *scikit-learn* ise makine öğrenmesi algoritmaları için kullanılmıştır.

Eksik değerlere sahip *Puan* ve *Yorum* değişkenleri tamamlanırken değişkenin aritmetik ortalaması kullanılmıştır.

5. BULGULAR

Sonuçların değerlendirilmesine katkı sağlamak için *ortalama oda fiyatı* değişkeni çeşitli değişkenlere göre gruplandırılmış, çıktılar tablolaştırılmıştır. Veri çözümleme ve temizleme aşamalarından sonra yapılan bu işlemde, toplam 6558 konaklama tesisine ait ortalama oda fiyatı değişkeni *tesis tipine* göre gruplandırılmış ve çıktılar *Tablo 1*'de verilmiştir.

Tablo 1. Ortalama Oda Fiyatlarının Tesis Türlerine Göre İstatistikleri

Tesis Türü	*N*	*Ortalama**	*S.S*
Otel	3057	403,86	382,05
Belirtilmemiş	927	561,02	594,12
Pansiyon	1823	311,22	354,06
Özel Konaklamalar	473	374,94	412,07
Apart	148	368,41	285,17
Lodge	86	274,19	232,19
Villa	21	490,12	307,82
Kamping	3	357,99	140,41
Hostel	20	220,94	77,97
Toplam	**6558**	**395,44**	**417,31**

*01.05.2018-30.09.2018 tarihleri arasında her tesise ait bir gece iki kişi konaklama ücreti haftada bir kez olacak şekilde alınmış ve ortalaması hesaplanmıştır.

Bu tarihler arasında Türkiye Cumhuriyeti Merkez Bankası'na (TCMB) ait USD döviz kuru ortalaması 6,82 TL'dir. Bu durumda verinin elde edildiği dönemde Türkiye'deki ortalama oda fiyatı yaklaşık 57,98 $'dır.

MB'nin sağladığı ve Türkiye İstatistik Kurumu'nun (TÜİK) tüketici fiyat endeksine (TÜFE) göre hazırladığı *enflasyon hesaplayıcısına* göre fiyatların alındığı ayların ortalaması olan Temmuz ayı ile araştırmanın hazırlandığı Ağustos 2020 arasındaki enflasyon katsayısı 1,2961'dir. Buna göre 395,44 TL'nin Ağustos 2020 karşılığı yaklaşık 512,53 TL'dir (TCMB, 2021).

Tablo 1 değerlendirildiğinde web madenciliği yapılan *tripadvisor.com* online seyahat acentasında toplam 6558 konaklama tesisinin yer aldığı görülmektedir. Bu tesislerin yarısına yakınının kendilerini *otel* olarak nitelendirdikleri, *apart, lodge, villa, kamping* ve *hostel* gibi tesis tiplerinin görece az olduğu görülmektedir. Bu istatistiklere göre Türkiye'deki konaklama tesislerinin 2018 yılı verilerine göre Mayıs-Eylül ayları dahil olmak üzere ortalama oda fiyatlarının *iki kişi* bir gece 395,44 TL olduğu görülmektedir.

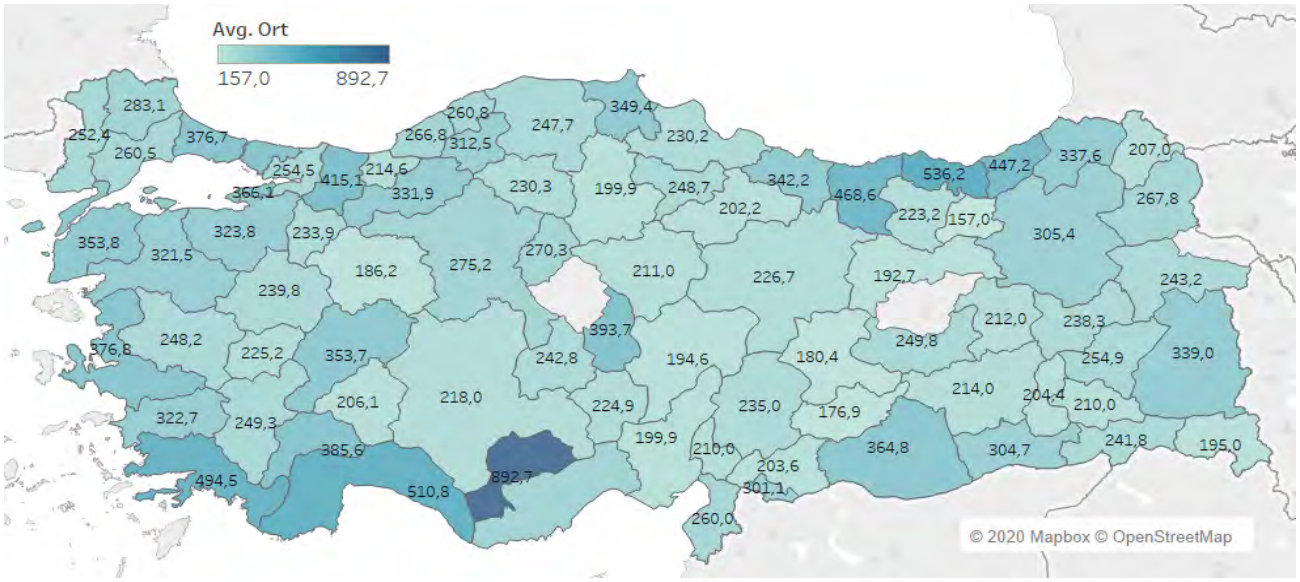
Sonuçların değerlendirilmesine katkı sağlamak için *ortalama oda fiyatı* değişkeni çeşitli değişkenlere göre gruplandırılmış, çıktılar tablolaştırılmıştır. İlk olarak ortalama oda fiyatı en yüksek *on il Tablo 2*'de verilmektedir.

Tablo 2. Ortalama Oda Fiyatı En Yüksek Olan On İl

No	*Şehir*	*N*	*Ortalama Fiyat*	*S.S.*	*Min.*	*Mak.*
1	Karaman	4	892,71	1359,01	184,00	2930,50
2	Trabzon	113	536,19	370,05	158,00	2664,00
3	Antalya	1171	510,78	579,97	60,50	9012,60
4	Muğla	948	494,51	645,94	98,25	12279,88
5	Giresun	18	468,57	817,25	155,00	3721,00
6	Rize	34	447,18	206,74	186,70	815,50
7	Sakarya	43	415,05	459,08	150,00	2869,10
8	Nevşehir	278	393,66	332,18	71,50	3373,00
9	Burdur	5	385,62	163,26	135,00	555,20
10	İzmir	547	376,76	314,18	80,00	4482,14
	Toplam	**6558**	**395,44**	**417,31**	**12279,88**	**60,50**

Tablo 2 incelendiğinde ortalama oda fiyatı açısından en yüksek fiyata sahip ilin *Karaman* (Ort.=892,71; S.S.=1359,01) olduğu görülmektedir. Fakat dikkat edileceği üzere bu ildeki konaklama tesisi sayısı 4'tür. Ayrıca standart sapmanın (S.S.=1359,01) diğer illere nazaran daha yüksek olduğu görülmektedir. Kıyı turizmi açısından önemli illerden olan Antalya (Ort.=510,78; S.S.=597,97) ve Muğla'da (Ort.=494,51; S.S.=645,94) ise ortalama fiyat farkının yaklaşık 16 TL olduğu görülmektedir.

İllere göre ortalama oda fiyatı değişkeni Türkiye haritasında yüksek fiyat koyu düşük fiyat açık olacak biçimiyle *Şekil 1*'de verilmektedir.



Şekil 1. İllere Göre Ortalama Oda Fiyatı

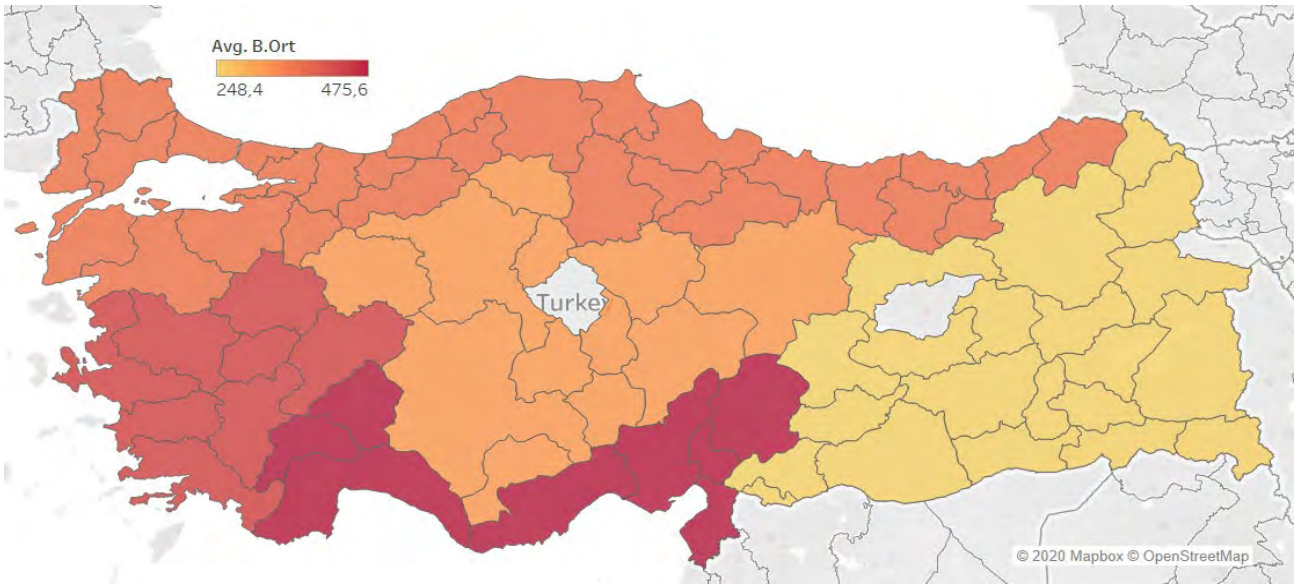
Şekil 1 incelendiğinde çok belirgin olmamakla birlikte kıyı illerinde veya turistik olarak nitelendirilebilen (Nevşehir, Urfa, Sinop vb.) illerde fiyatların biraz daha yüksek olduğu görülmektedir.

Ortalama oda fiyatı değişkeni Türkiye'deki coğrafi bölgelere göre de gruplandırılmıştır. İl merkezinin bulunduğu bölgeye göre ortalama oda fiyatı değişkenine ait bazı istatistikler Tablo 3'de verilmektedir.

Tablo 3. Bölgelere Göre Ortalama Oda Fiyatı

No	*Bölge*	*N*	*Ortalama Fiyat*	*S.S.*	*Min.*	*Mak.*
1	Akdeniz	1352	475,63	548,24	60,50	9012,60
2	Doğu Anadolu	104	249,04	163,19	89,20	1333,33
3	Ege	1765	427,69	515,83	80,00	12279,88
4	G.doğu Anadolu	113	248,44	230,87	96,00	2272,25
5	İç Anadolu	628	315,18	275,47	71,50	3373,00
6	Karadeniz	439	366,47	322,17	72,50	3721,00
7	Marmara	2157	362,79	257,47	84,33	4301,71
	Toplam	**6558**	**395,44**	**417,31**	**12279,88**	**60,50**

Tablo 3 incelendiğinde ortalama oda fiyatı en yüksek bölgenin Akdeniz (Ort: 475,63, S.S: 548,24), ortalama oda fiyatı en düşük olan bölgelerin ise birbirine yakın fiyatlarla Doğu Anadolu (Ort: 249,04, S.S: 163,19) ve Güneydoğu Anadolu (Ort: 248,44, S.S: 230,87) olduğu görülmektedir. Bu tablo görselleştirildiğinde ortaya *Şekil 2* çıkmaktadır.



Şekil 2. Bölgelere Göre Ortalama Oda Fiyatı

Şekil 2 incelendiğinde kitle turizminin geliştiği Akdeniz ve Ege bölgelerinde ortalama fiyatların diğer bölgelere oranla daha yüksek olduğu görülmektedir.

Ortalama oda fiyatı bağımlı değişkeni konaklama tesislerinin konuklara sunduğu hizmetleri ve tesislerin sahip olduğu çeşitli donanımları ifade eden 44 bağımsız değişken ile birlikte yaygın kullanılan bazı makine öğrenmesi algoritmaları ile analiz edilmiştir. Algoritmaların ilgili veri üzerindeki başarı oranları Tablo 4'de verilmektedir.

Tablo 4. Kullanılan Makine Öğrenmesi Algoritmalarının Araştırma Verisini Açıklama Oranları

No	*Model*	*Skor*
1	Rassal Orman (Random Forest)	99,45
2	Karar Ağacı (Decision Tree)	99,45
3	KNN	61,63
4	Lojistik Regresyon (Logistic Regression)	39,13
5	Olasılıksal Dereceli Azalma (Stochastic Gradient Decent)	33,87
6	Destek Vektör Makineleri (Support Vector Machines)	33,89
7	Perceptron	19,08
8	Naive Bayes	8,14

Tablo 4 incelendiğinde algoritmaların birbirlerinden oldukça farklı skorlar elde ettiği görülmektedir. Burada en yüksek skorların Rassal/Rastgele Orman ve Karar Ağacı algoritmalarına ait olduğu görülmektedir. Her iki algoritmada da neredeyse %100 yakın bir sonuç çıkmıştır. Bu iki algoritma Karar Ağacı olarak adlandırılan sınıflandırma algoritmalarının türevleridir (Shalev-Shwartz ve Ben-David, 2014:217). Bu algoritmalar, ağacın tamamına yakınını oluşturabildiğinden, yüksek skorlar elde edilmiştir (Rassal Orman: 99,45; Karar Ağacı: 99,45). Karar ağaçları yaygın kullanılan veri madenciliği yöntemlerinden biridir. Karar ağaçları algoritmaları bir veri setini bir ağacın dallarına benzer şekilde sınıflandırmak için kullanılmaktadır. Daldaki her *düğüm*

(node) verinin sınıflandırıldığı noktayı ifade eder. İlk sınıflandırma kökten başlar, ilerledikçe sınıflandırma sayısı artar. Karar ağaçları görselleştirilebildiği gibi *mantıksal ifadelerle* de veriyi açıklayabilmektedir. Çalışmada yapılan karar ağacı sonucunda belli başlı bazı ifadeler aşağıda verilmektedir.

- Eğer tesisin tipi *otel* ise ortalama oda fiyatı %79,4 oranında *206,27-1263,01 TL* aralığında olacaktır.
- Eğer tesisin tipi otel, konukların verdiği ortalama puan 5 ise ortalama oda fiyatı %84,6 oranında *206,27-1263,01 TL* aralığında olacaktır.
- Eğer tesisin tipi otel, konukların verdiği ortalama puan 4,5 üzeri ve tesis *tenis kortuna* sahipse, ortalama oda fiyatı %100 oranında *206,27 TL*'den *yüksek* olacaktır.
- Eğer tesisin tipi otel, konukların verdiği ortalama puan 4,5 üzeri, tesis *tenis kortuna* sahip değil ve spaya sahipse, ortalama oda fiyatı %74,7 oranında *317,52 TL*'den *yüksek* olacaktır.
- Eğer tesisin tipi pansiyon, konukların verdiği ortalama puan 5 ise ve tesis klimaya sahipse ortalama oda fiyatı %57,2 oranında *317,52-3373,01 TL* aralığında olacaktır.

Yapılan karar ağacı analizi sonucunda yukarıda verilen mantıksal ifadelerin daha fazlası algoritma ile elde edilmiştir. Bazı değişkenler eklenip çıkartılarak yeni ifadeler de elde etmek mümkündür. Bu çalışma makine öğrenmesi algoritmalarının performansını ölçmeyi ve kıyaslamayı amaçladığından çıktıların sadece bir kısmına yer verilmiştir.

Doğrusal sınıflandırma (Lineer Classification) yöntemini kullanan Lojistik Regresyon, Olasılıksal Dereceli Azalma ve Destek Vektör Makineleri algoritmalarından en yüksek skoru lojistik regresyon (39,13) elde etmiştir.

Olasılıksal Dereceli Azalma, bir kayıp fonksiyonu altında uyum iyiliğini (goodness of fit) adım adım ölçüp kaybı en aza indirerek model oluşturmaktadır (Bilogur, 2018). Fakat yine de Lojistik Regresyonun skoru daha yüksek çıkmıştır. Yapay Sinir Ağları'nın bir örneği olan Perceptron'ın skoru bu algoritmalardan daha düşüktür (19,08).

KNN (k Nearest Neighborhood) önceden verilen küme sayısına göre (k) bağımlı değişkeni benzerliklerine göre sınıflandıran bir algoritmadır. Burada varsayılan olarak verilen *Minkowski Mesafesi* kullanılmıştır. KNN algoritmasında ön tanımlı olarak *üç* küme belirtilmiştir. Algoritmanın veri setindeki konaklama tesislerini %61,63 oranında üç küme olarak sınıflandırdığı görülmüştür.

Belirtildiği üzere KNN algoritmasında verinin sınıflandırılacağı küme sayısı önceden verilmelidir. Şart olmamasına rağmen genellikle küme sayısı 3, 5, 7 gibi tek sayı olarak verilmektedir. Analizde en doğru küme sayısını bulmak için farklı küme sayıları bir döngü yardımıyla denenmiştir. Buna göre performans küme sayısı üç olduğunda 61.63, beş olduğunda 55.01, yedi olduğunda 51.73 ve dokuz olduğunda 49,37 olarak belirlenmiştir. Çıkan bu sonuçlardan hareketle, en doğru sonuca ulaşan algoritmanın belirlenebilmesi için öncelikle veri setine uygun olan ve yüksek doğruluk oranına sahip Lojistik Regresyon seçilmiştir. Sonraki süreçte lojistik regresyonun en doğru varyasyonunun belirlenmesi amaçlanmaktadır. Burada karar ağacı ve KNN algoritmaları araştırmanın amacından bir nebze farklı olduğundan tercih edilmemiştir.

Model için tercih edilen lojistik regresyon lineer regresyonun sınıflandırma amacı ile kullanılan bir türüdür. Makine öğrenmesi algoritmaları içerisinde sıklıkla kullanılmaktadır. Bilindiği üzere lineer regresyonda y bağımlı değişken ve x1'den xn'e bağımsız/açıklayıcı değişkenler olmak üzere;

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta nX$$

eşitliği kullanılmaktadır. Lojistik regresyonda eğriyi ifade etmek için *Sigmoid Fonksiyonu* kullanılmaktadır. Bu durumda ortaya aşağıdaki eşitlik çıkmaktadır.

$$p = 1/\,1 + e^{-(\beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta nXn)}$$

Tablo 5'de ortalama oda fiyatı ile aralarında en yüksek korelasyona sahip on değişken verilmektedir.

Tablo 5. Ortalama Oda Fiyatı ile En Yüksek Korelasyona Sahip Değişkenler

No	*Değişken*	*Ortalama Oda Fiyatı*
	Ortalama Oda Fiyatı	1,00
1	Yorum	0,35
2	Puan	0,23
3	Spa	0,21
4	Kapalı Havuz	0,20
5	Spor Salonu/Fitness Merkezi	0,20
6	Tenis Kortu	0,19
7	Konferans Salonu	0,16
8	Isıtmalı Havuz	0,16
9	Deniz Kıyısı	0,15
10	Ziyafet Salonu	0,14

Yapılan korelasyon analizi sonucuna göre *ortalama oda fiyatı* ile en yüksek korelasyona sahip değişken *konukların yorum sayısı* olarak belirlenmiştir (r=0,35). Buna göre yorum sayısı artıkça tesislerin oda fiyatları da artmaktadır. Sıralamada ikinci sırada konukların tesislere verdikleri puan yer almaktadır. Buradan konukların ortalama oda fiyatları üzerinde tesisin sahip olduğu donanım veya hizmetlerden daha etkili olduğu sonucu çıkarılabilir. Tesise ait bir hizmet olan *spa* ortalama oda fiyatı ile korelasyonda üçüncü sırada yer almaktadır. Spa hizmeti 6558 tesisin sadece 651'nde bulunmaktadır.

Regresyon modellerinde açıklayıcı (bağımsız/explanatory) değişkenlerin mümkün olduğunca az olması tercih edilmektedir. Veri setinde daha önce de belirtildiği gibi 44 tane açıklayıcı değişken bulunmaktadır. Bunlardan bağımlı değişken üzerinde en etkili olanları belirlemek için bir sonraki aşamada *geri eleme (backward elimination)* yöntemi kullanılmıştır. Geri eleme yönteminden önce, *kukla değişken tuzağından (dummy variable trap)* kaçınmak için (Lewis-Beck, 1995:60-61; Crown, 1998:67) aralarında yüksek korelasyon olan ve benzer özellikteki *Havuz/AçıkHavuz, Restoran/Bar Oturma Salonu, Konferans Salonu/Toplantı Odaları, Kuru Temizleme/Çamaşır Yıkama Hizmeti, Mini Bar/Odada Buzdolabı* değişkenleri birleştirilmiştir. Analiz sonucunda anlamlılık p değeri 0,05 üzerinde olan değişkenlerin anlamlı olmadığı kabul edilmiş ve modelden çıkarılmıştır. Modele ilişkin analiz sonucu şekilde verilmektedir.

```
                          Results: Logit
================================================================================
Model:                    Logit                 Pseudo R-squared:      -12.261
Dependent Variable:       Ort                   AIC:                   592.9709
Date:                     2020-09-14 07:17      BIC:                   864.5085
No. Observations:         6558                  Log-Likelihood:        -256.49
Df Model:                 39                    LL-Null:               -19.341
Df Residuals:             6518                  LLR p-value:           1.0000
Converged:                1.0000                Scale:                 1.0000
No. Iterations:           8.0000
--------------------------------------------------------------------------------
                              Coef.  Std.Err.     z     P>|z|    [0.025  0.975]
--------------------------------------------------------------------------------
Puan                         -0.6467   0.0447 -14.4766 0.0000 -0.7343 -0.5592
Sirasi                       -0.0017   0.0004  -4.1778 0.0000 -0.0024 -0.0009
Yorum                         0.0003   0.0002   1.7307 0.0835 -0.0000  0.0006
Tesis_Tipi                   -0.2491   0.0665  -3.7464 0.0002 -0.3794 -0.1188
Sehir                        -0.0372   0.0082  -4.5398 0.0000 -0.0532 -0.0211
================================================================================
```

Şekil 3. Lojistik Regresyon Modeline İlişkin Çıktı Sonucu

Değişken birleştirme işlemi yapıldıktan sonra geriye kalan 39 değişken, geri eleme (backward elimination) yöntemi ile analiz edilmiştir. Bu değişkenlerden sadece 4 tanesinin ortalama oda fiyatı üzerinde etkisi olduğu belirlenmiştir. Bu değişkenler şekilden de görüldüğü üzere konuk puanları, bulunduğu bölgedeki beğeni sırası, tesis tipi ve bulunduğu şehirdir. Konukların tesis hakkında yaptıkları yorum sayısı 0,05 p değerine yakın olsa da eşik değerin üzerinde olduğundan anlamlı çıkmamıştır.

6. SONUÇ ve TARTIŞMA

Bu çalışmada makine öğrenmesi algoritmalarından hangisi/hangilerinin otellerde oda fiyatına etki eden faktörleri en doğru şekilde ifade edebileceği araştırılmıştır. Bu amaçla; makine öğrenmesi algoritmalarından bazıları, araştırma için web madenciliği yoluyla elde edilen *büyük veri* kullanılarak kıyaslanmış ve otel oda fiyatlarını açıklama performansları test edilmiştir.

Karar ağacı türündeki Rassal Orman, Karar Ağacı, kümeleme türündeki KNN ve sınıflandırma türündeki Lojistik Regresyon, Olasılıksal Dereceli Azalma, Destek Vektör Makineleri, Perceptron, Naive Bayes algoritmaları Türkiye'de hizmet vermekte olan 6558 konaklama tesisine ait bir sezonluk fiyat üzerinde test edilmiştir. Veri seti, tesisin online seyahat acentesinin sunmuş olduğu, gecelik iki kişilik oda fiyatının, bir sezon boyuncaki (yaklaşık 180 gün) ortalamasından oluşan ortalama oda fiyatı bağımlı/açıklanan değişken ile yine tesisin sahip olduğu çeşitli hizmetler ve özelliklerden oluşan 44 bağımsız/açıklayıcı değişkenden oluşmaktadır.

Karşılaştırılan Rassal/Rastgele Orman ve Karar Ağacı algoritmalarının her ikisinin de yaklaşık %99,89 oranında veri setini açıkladığı dolayısıyla ağaç/dallanmaların başarı ile gerçekleştirdiği görülmektedir. KNN algoritması ise en yüksek performansı üç kümeli bir sınıflandırma ile %62,12 oranında gerçekleştirmiştir.

Makine öğrenmesi algoritmalarından lineer sınıflandırma yöntemini kullanan Lojistik Regresyon, Olasılıksal Dereceli Azalma ve Destek Vektör Makineleri algoritmalarından en yüksek skoru lojistik regresyon (39,13) yöntemi elde etmiştir. Çalışmada bu sınıflandırma yönteminde daha da detaya inilerek regresyon modelinde hangi değişkenlerin ortalama oda fiyatını açıklamada anlamlı olduğunu belirlemek için veri üzerinde geri eleme (backward elimination) yöntemi uygulanmıştır. Analiz sonucuna göre modelde, konukların tesise verdikleri *puan*, tesisin bölgede diğer tesisler arasındaki *sırası*, *tesisin türü* ve bulunduğu *şehir* istatistiki olarak anlamlı bulunmuştur ($p<0,05$). Analize giderken lojistik regresyon için yapılan ön hazırlıklara, bulgular kısmında yer verilmiştir.

Razavi ve Israeli (2019) Ağustos'un ilk haftasına ait New York'un Manhattan bölgesindeki toplam 309 tesisin fiyat verisi üzerinde benzer bir çalışma yapmış, bağımsız değişken olarak *tesislerin yıldızları*, *konukların verdikleri ortalama puanlar* ve *tesislere ait yorumların ortalamasını* kullanmışlardır. Çalışmada en iyi regresyon modelini belirlemek için aşamalı regresyon modeli (stepwise regression) kullanılmış, en yüksek skor olarak R^2 0,39 bulunmuştur. Bu sonuç bu çalışmaya çok yakındır (%39,13). Bu araştırmada da yukarıda belirtildiği gibi *konukların verdikleri puanlar* anlamlı çıkmıştır. Razavi ve Israeli (2019) makine öğrenmesi modeli olarak regresyon (sıradan en küçük kareler, OLS), karar ağaçları (decision trees, CART ve RF), destek vektör makineleri (support vector machines) ve yapay sinir ağları (neural networks) kullanmışlardır. Bu çalışma ile Razavi ve Israeli (2019) çalışması arasındaki en büyük farkı *kullanılan veri miktarı* ve *değişken sayısı* oluşturmaktadır. Bu çalışmada bir ülkenin online seyahat acentesine kayıtlı bütün konaklama tesisleri ve bunlara ait 44 değişken açıklayıcı kullanılırken, Razavi ve Israeli (2019) sadece belirli bir bölgeye ait veriyi ve üç değişkeni kullanmışlardır. Ayrıca bu çalışmada bir sezona ait ortalama oda fiyatı veri olarak kullanılırken, Razavi ve Israeli'nin (2019) çalışmalarında Ağustos'un ilk haftasına ait fiyat kullanılmıştır. Razavi ve Israeli'nin (2019) çalışmaları dışında yazındaki diğer çalışmaların turist gelişleri ve zamana bağlı fiyat tahminine yönelik olduğu görülmektedir. Bu çalışma dışında turizm konusunda makine öğrenmesi algoritmaları genellikle turizm talebinin belirlenmesi (Law ve Au, 1999; Law, 2000; Burger, Dohnal, Kathrada ve Law, 2001; Cho, 2003), turist gelişlerinin hesaplanması (Hadavandi, Ghanbari, Shahanaghi ve Abbasian-Naghneh, 2011; Akın, 2015) ve ziyaretçi sayılarının belirlenmesi (Pattie ve Snyder, 1996) gibi amaçlarla kullanılmıştır.

Bu çalışmada belli başlı bazı makine öğrenmesi algoritmaları bir online seyahat acentesine ait veriler kullanılarak performans bazlı test edilmiştir. Analizde *regresyon* ve *naive bayes* gibi klasik yöntemlerin yanı

sıra karar ağaçları gibi daha güncel yöntemler de kullanılmıştır. Sonuçlar, *regresyon* gibi analizlerin hala güçlü sınıflandırma ve makine öğrenmesi algoritmalarından olduğunu ortaya koymuştur. Araştırmada yapılan regresyon analizinde daha doğru bir modeli ortaya koymak için yapılan geri eleme yönteminde hangi bağımsız değişkenlerin ortalama oda fiyatı değişkenini açıklamada anlamlı olduğu belirlenmiştir. Elde edilen betimsel istatistikler Türkiye'deki konaklama tesislerinin fiyat dağılımları hakkında daha önce saptanmamış önemli bilgiler vermektedir.

Araştırmada bazı kısıtlılıkların açıklanması faydalı olacaktır. Öncelikle en büyük problem online seyahat acentesinin konaklama tesisine sağladığı veri formunun doğru doldurulmasından ortaya çıkmaktadır. Örneğin Türkiye'deki konaklama tesislerinin büyük çoğunluğunda klima olmasına rağmen bazı lüks tesislerin klimayı önemsemedikleri/müşterinin önemsemeyeceğini düşündükleri için "*yok*" olarak forma girmeleri veride bozulmalara neden olmaktadır. Verinin temizleme aşamasında her ne kadar bu tür şeylerin düzeltilmesi için yoğun çaba sarf edilse de yine de problem teşkil etmektedir.

KAYNAKÇA

Aguiló, E., Alegre, J. ve Sard, M. (2003). Examining the Market Structure of The German and UK Tour Operating Industries Through An Analysis of Package Holiday Prices. *Tourism Economics, 9*(3), 255-278.

Ağca, Y. (2019). *Çevrimiçi Seyahat Acentalarında Oda Fiyatlarına Etki Eden Faktörlerin Araştırılması* (Yayınlanmamış Doktora Tezi). Atatürk Üniversitesi, Sosyal Bilimler Enstitüsü, Erzurum.

Akın, M. (2015). A novel approach to model selection in tourism demand modeling. *Tourism Management*, 64-72. doi:https://doi.org/10.1016/j.tourman.2014.11.004

Andersson, D. E. (2008). Hotel attributes and hedonic prices: an analysis of internet-based transactions in Singapore's market for hotel rooms. *Ann Reg Sci*, 229-240. doi:10.1007/s00168-008-0265-4

Antonio, N., Almeida, A. d. ve Nunes, L. (2017). Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model. *16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (s. 1049-1054). Cancun, Mexico: IEEE. doi:10.1109/ICMLA.2017.00-11

Bilogur, A. (2018, 04 28). *Support vector machines and stoch gradient descent.* https://www.kaggle.com/residentmario/support-vector-machines-and-stoch-gradient-descent adresinden alındı

Bull, A. O. (1994). Pricing A Motel's Location. *International Journal of Contemporary Hospitality Management, 6*(6), 10-15.

Burger, C., Dohnal, M., Kathrada, M. ve Law, R. (2001). A practitioner's guide to time-series methods for tourism demand forecasting – a case study of Durban, South Africa. *Tourism Management*, 403-409.

Carvell, S. A. ve Herrin, W. E. (1990). Pricing in the Hospitality Industry: An Implicit Market Approach. *Hospitality Review*, 27-37. http://scholarship.sha.cornell.edu/articles/194/ adresinden alındı

Castro, C., Ferreira, F. A. ve Ferreira, F. (2016). Trends in hotel pricing Identifying guest value hotel attributes using the cases of Lisbon and Porto. *Worldwide Hospitality and Tourism Themes, 8*(6), 691-698. doi:10.1108/WHATT-09-2016-0047

Cho, V. (2003). A comparison of three different approaches to tourist arrival forecasting. *Tourism Management*, 323-330. doi:https://doi.org/10.1016/S0261-5177(02)00068-7

Crown, W. H. (1998). *Statistical Models for the Social and Behavioral Sciences: Multiple Regression and Limited-dependent Variable Models.* London, UK: Greenwood Publishing Group.

Dontha, R. (2018, 12 19). *Digital Transformation*. 05 04, 2019 tarihinde Data Mining Steps: https://digitaltransformationpro.com/data-mining-steps/ adresinden alındı

Espinet, J. M., Saez, M., Coenders, G. ve Fluvià, M. (2003). Effect on Prices of the Attributes of Holiday Hotels: A Hedonic Prices Approach. *Tourism Economics, 9*(2), 1-13.

Frumusanu, A. (2018, 03 26). *www.anandtech.com*. 04 28, 2019 tarihinde The Samsung Galaxy S9 and S9+ Review: Exynos and Snapdragon at 960fps: https://www.anandtech.com/show/12520/the-galaxy-s9-review/6 adresinden alındı

García, S., Luengo, J. ve Herrera, F. (2015). *Data Preprocessing in Data Mining.* Cham, Switzerland: Springer International Publishing.

Hadavandi, E., Ghanbari, A., Shahanaghi, K. ve Abbasian-Naghneh, S. (2011). Tourist arrival forecasting by evolutionary fuzzy systems. *Tourism Management*, 1196-1203. doi:https://doi.org/10.1016/j.tourman.2010.09.015

Han, J., Kamber, M. ve Pei, J. (2012). *Data Mining: Concepts and Techniques* (3 b.). Waltham: Morgan Kaufmann Publishers.

http://www3.tcmb.gov.tr/enflasyoncalc/enflasyonyeni.php, erişim tarihi 14.02.2021

https://evds2.tcmb.gov.tr/index.php, erişim tarihi 14.02.2021

Isuhuaylas, L. A., Hirata, Y., Santos, L. C. ve Torobeo, N. S. (2018). Natural Forest Mapping in the Andes (Peru): A Comparison of the Performance of Machine-Learning Algorithms. *Remote Sensing, 10*(5), 782. doi:http://dx.doi.org/10.3390/rs10050782

Keskinkılıç, M., Ağca, Y. ve Karaman, E. (2016). İnternet ve Bilgi Sistemleri Kullanımının Turizm Dağıtım Kanallarına Etkisi Üzerine Bir Uygulama. *İşletme Araştırmaları Dergisi, 8*(4), 445-472. doi:10.20491/isarder.2016.227

Ku, C. H., Chang, Y.-C., Wang, Y., Chen, C.-H. ve Hsiao, S.-H. (2019). Artificial Intelligence and Visual Analytics: A Deep-Learning Approach to Analyze Hotel Reviews & Responses. *52nd Hawaii International Conference on System Sciences* (s. 5268-5277). Honolulu, US: University of Hawaii at Manoa. doi:10.24251/HICSS.2019.634

Law, R. (2000). Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting. *Tourism Management*, 331-340. doi:https://doi.org/10.1016/S0261-5177(99)00067-9

Law, R. ve Au, N. (1999). A neural network model to forecast Japanese demand for travel to Hong Kong. *Tourism Management*, 89-97. doi:https://doi.org/10.1016/S0261-5177(98)00094-6

Lebanon, G. ve El-Geish, M. (2018). *Computing with Data: An Introduction to the Data Industry.* Cham, Switzerland: Springer.

Lewis-Beck, M. (1995). *Data Analysis: An Introduction.* Thousand Oak, US: SAGE Publications.

Löch, M. ve Axhausen, K. W. (2010). Modeling Hedonic Residential Rents for Land Use and Transport Simulation While Considering Spatial Effects. *The Journal of Transport and Land Use, 3*(2), 39-63. doi:10.1598/jtlu.v3i2.117

Lutz, M. (2009). *Learning Python: Powerful Object-Oriented Programming* (4 b.). Sebastopol, US.: O'Reilly Media, Inc.

Maksymenko, S. (2020). *10 AI And Machine Learning Trends To Impact Business In 2020.* mobidev.biz: https://mobidev.biz/blog/future-ai-machine-learning-trends-to-impact-business adresinden alındı

Monson, M. (2009). Valuation Using Hedonic Pricing Models. *Cornell Real Estate Review, 7*, 62-73.

Noi, P. T. ve Kappas, M. (2017). Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors, 18*(1), 1-20. doi:http://dx.doi.org/10.3390/s18010018

Papatheodorou, A. (2002). Exploring Competitiveness in Mediterranean Resorts. *Tourism Economics, 8*(2), 133-150.

Pattie, D. C. ve Snyder, J. (1996). Using a neural network to forecast visitor behavior. *Annals of Tourism Research*, 151-164. doi:https://doi.org/10.1016/0160-7383(95)00052-6

Razavi, R. ve Israeli, A. A. (2019). Determinants of online hotel room prices: comparing supply-side and demand-side decisions. *International Journal of Contemporary Hospitality Management, 31*(5), 2149-2168.

Rose, S. (2020, 03 21). *What is the Future of Machine Learning?* codeburst.io: https://codeburst.io/what-is-the-future-of-machine-learning-f93749833645 adresinden alındı

Sánchez-Medina, A. J. ve C-Sánchez, E. (2020). Using machine learning and big data for efficient forecasting of hotel booking cancellations. *International Journal of Hospitality Management, 89*, 1-9. doi:10.1016/j.ijhm.2020.102546

Shalev-Shwartz, S. ve Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms* (1 b.). New York, US: Cambridge University Press.

Shehhi, M. A. ve Karathanasopoulos, A. (2020). Forecasting hotel room prices in selected GCC cities using deep learning. *Journal of Hospitality and Tourism Management, 42*, 40-50.

Sun, S., Wei, Y., Tsui, K.-L. ve Wang, S. (2019). Forecasting tourist arrivals with machine learning and internet search index. *Tourism Management, 70*, 1-10. doi:10.1016/j.tourman.2018.07.010

Taylor, P. (1995). Measuring Changes in the Relative Competitiveness of Package Tour Destinations. *Tourism Economics, 1*(2), 169-182.

Tomiazzi, J. S., Pereira, D. R., Judai, M. A., Antunes, P. A. ve Favareto, A. P. (2019). Performance of machine-learning algorithms to pattern recognition and classification of hearing impairment in Brazilian farmers exposed to pesticide and/or cigarette smoke. *Environmental Science and Pollution Research, 26*, 6481–6491. doi:https://doi.org/10.1007/s11356-018-04106-w

Williams, N., Zander, S. ve Armitage, G. (2006). A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. *ACM SIGCOMM Computer Communication Review, 36*(5), 7-15. doi:https://doi.org/10.1145/1163593.1163596

Yang, Y., Tang, J., Luo, H. ve Law, R. (2015). Hotel location evaluation: A combination of machine learning tools and web GIS. *International Journal of Hospitality Management, 47*, 14-24.

Yayar, R. ve Gül, D. (2014). Mersin Kent Merkezinde Konut Piyasası Fiyatlarının Hedonik Tahmini. *Anadolu Üniversitesi Sosyal Bilimler Dergisi, 14*(3), 87-99.