Avrupa Bilim ve Teknoloji Dergisi
Sayı 18, S. 719-727, Mart-Nisan 2020

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 18, pp. 719-727, March-April 2020

Research Article

Lojistik Regresyon ve CART Yöntemlerinin Tahmin Edici Performanslarının Yaşam Memnuniyeti Verileri için Karşılaştırılması

Arzu Yavuz[1], Özgül Vupa Çilengiroğlu[2*]

[1] Dokuz Eylül Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik Bölümü, İzmir, Türkiye (ORCID: 0000-0002-2059-2635)
[2] Dokuz Eylül Üniversitesi, Fen Fakültesi, İstatistik Bölümü, İzmir, Türkiye (ORCID: 0000-0003-0181-8376)

(İlk Geliş Tarihi 21 Şubat 2020 ve Kabul Tarihi 20 Mart 2020)

(**DOI:** 10.31590/ejosat.691215)

ATIF/REFERENCE: Vupa Çilengiroğlu, Ö. & Yavuz, A. (2020). Lojistik Regresyon ve CART Yöntemlerinin Tahmin Edici Performanslarının Yaşam Memnuniyeti Verileri için Karşılaştırılması. *Avrupa Bilim ve Teknoloji Dergisi*, (18), 719-727.

Öz

Makine öğrenimi içinde yer alan sınıflandırma ve regresyon, veri sınıflarını ortaya koyan ve değişkenler arasındaki ilişkileri modelleyen yöntemlerdir. Sınıflama ve regresyon yöntemlerinden karar ağaçları, eğitim verisini kullanarak sınıflandırma kurallarını oluşturup test verisinde bu kuralları dener ve algoritma başarısını belirler. Lojistik regresyonda kurulan model ile sınıflandırma yapılıp performanslar bulunur. Bu yöntemler, yorumunun kolay olması, büyük veri setlerine uygulabilirliği ve varsayım gerektirmemesi sebebi ile son zamanlarda birçok farklı disiplinlerde kullanılmaktadır. Yaşam memnuniyeti kavramı, günümüzde birçok farklı disiplinlerin ilgi alanına giren bir konudur. Yaşam memnuniyeti, bireyin sürdürmekte olduğu yaşamdan ne kadar zevk aldığının bir bütün olarak ele alınmasıdır. Bu çalışmanın amacı, karar ağacı yöntemlerinden olan CART ve lojistik regresyon çözümlemelerinin performanslarının Türkiye İstatistik Kurumuna ait (TÜİK) 2017 dönemini kapsayan yaşam memnuniyeti verilerini (n=8430) kullanarak yapılmasıdır. Bu amaçla yapılan çalışmada, yaşam memnuniyetini açıklayan en iyi modelin performans kriterlerine (doğruluk, duyarlılık, seçicilik, kesinlik, F-skor, ROC eğrisi R^2) bağlı olarak lojistik regresyon modeli olduğuna karar verilmiştir. Bu modelde yaşam memnuniyeti; cinsiyet, medeni durum, okul durumu, gelir, sosyal hayat, sağlık ve ulaşım değişkenleri ile açıklanmıştır.

Anahtar Kelimeler: Makine Öğrenimi, CART Algoritması, Lojistik Regresyon, Yaşam Memnuniyeti

Comparison of Predictive Performance of Logistic Regression and CART Methods for Life Satisfaction Data

Abstract

Classification and regression in machine learning are methods that reveal data classes and model the relationships between variables. Decision trees, one of the classification and regression methods, create the classification rules by using the training data, test these rules in the test data and determine the algorithm success. Classification is made with the model established in logistic regression and performances are found. These methods have been used in many different disciplines recently due to their easy interpretation, application to large data sets and no assumptions. The concept of life satisfaction is an issue of many different disciplines today. Life satisfaction is a consideration of how much the individual enjoys the life the individual lives. The purpose of this study, the performance of the CART and logistic regression analysis of the decision tree method of Turkey Statistical Institute (TSI), covering the period 2017 to life satisfaction data (n = 8430) is made using. In this study, it was decided that the best model that explains life satisfaction is the logistic regression model based on performance criteria (accuracy, sensitivity, selectivity, precision, F-score, ROC curve R2). In this model, life satisfaction; It is explained by variables of gender, marital status, school status, income, social life, health and transportation.

Keywords: Machine Learning, CART Algorithm, Logistic Regression, Life Satisfaction

[*] Sorumlu Yazar: Dokuz Eylül Üniversitesi, Fen Fakültesi, İstatistik Bölümü, İzmir, Türkiye, ORCID: 0000-0003-0181-8376, ozgul.vupa@deu.edu.tr

1. Giriş

Büyük veri kavramı ve problemleri, ilk kez Cox ve Ellsworth (1997) tarafından kullanılmıştır. Büyük veri, geleneksel veri yöntemlerinin kullanılması ile işlenmesi mümkün olmayan, farklı büyüklüklerdeki heterojen veriyi tanımlayan ve çeşitli elektronik içeriklerden oluşan bir kavram olarak tanımlanmaktadır (Gani vd., 2016). Büyük veri, "veri madenciliği" içersinde yer almaktadır. Veri madenciliği, "istatistik", "veritabanı teknolojisi", "örüntü tanıma", "makine öğrenme" ile etkileşimli ve ham verinin tek başına sunamadığı bilgiyi çıkaran, veri analizi sürecini içeren yeni bir disiplin olarak tanımlanmaktadır (Hand,1998). Veri madenciliği tanımına bilgisayar kavramı da eklenerek veri madenciliği, büyük veri setleri arasından gelecekle ilgili tahminde bulunabilecek bilgisayar programlarını kullanma işi olarak da değerlendirilmektedir (Doğan&Türkoğlu, 2007).

Veri madenciliğinin kullanım alanı çok geniştir. Veri madenciliği, pazar araştırmalarında (benzerlik saptanması, müşteri profilinin belirlenmesi, kampanya analizi, satış tahmini…), risk analizlerinde (kalite kontrol, rekabet analizi, öngörü ve sahtekarlığın saptanması, bankacılık ve sigorta analizleri, müşteri kredi risk araştırmaları…), kurum ve insan kaynakları analizlerinde (kaynakların ve kişilerin en optimal biçiminde kullanımı…), tıp araştırmalarında (hastalıkların anlık tespiti…) ve sosyo-demografik özelliklerin değerlendirilmesinde (sınıf özelliklerinin belirlenmesinde, ölçeklerin analizinde…) kullanılır.

Bir süreç çalışması olan veri madenciliğinde izlenen adımlar şu şekilde sıralanabilir. (1) Problemin tanımlanması, (2) Verilerin hazırlanması (toplama, birleştirme temizleme ve seçim), (3) Modelin kurulması ve değerlendirilmesi, (4) Modelin kullanılması ve (5) Modelin izlenmesi.

Veri madenciliği ile etkileşimli olan makine öğrenmeside çeşitli algoritmalar ve yöntemler kullanarak bu süreci kullanmaktadır. Bu algoritmalar ve yöntemler, sınıflandırma (k-en yakın komşuluk, karar ağaçları (CART: Classification and Regression Trees, CHAID,…), yapay sinir ağları, Bayesian yöntemi, rassal orman, genetik algoritma), regresyon (doğrusali lojistik, çok terimli) ve kümeleme olarak sıralanmaktadır.

Veri madenciliğinin kullanım alanı içersinde yer alan sosyo-demografik özelliklerin incelenmesinde, yaşamın genel değerlendirilmesi bulunmaktadır. Literatürde yaşamın genel değerlendirmesi "yaşam memnuniyeti" olarak incelenmektedir (Doğan&Sapmaz, 2012; Tümlü&Recepoğlu, 2011; Akın&Yalnız, 2015; Korkmaz vd., 2015; Kanbur&Özdemir, 2017). Yaşam memnuniyeti, bireylerin yaşama gösterilen duygusal tepkiler ile kendi yaşadıkları ve çevreleri ile ilgili olduğundan birçok durumdan etkilenebilmektedir. Çalışma hayatı, gelir ve sağlık durumu, eğitim seviyesi, çevre faktörleri, sosyal, fiziki ve ekonomik güvenlik bu durumlardan sayılabilmektedir (Kanbur&Özdemir, 2017). Bireylerin yaşama karşı besledikleri pozitif hisler yaşam memnuniyeti artırırken, negatif hisler yaşam memnuniyetini azaltmaktadır (Korkmaz vd., 2015). Türkiye'nin veri tabanlarının en geniş ve önemlilerinden biri olan TÜİK (Türkiye İstatistik Kurumu), yaşam memnuniyetini, kişinin sürdürmekte olduğu hayatı bir bütünlük içinde olumlu olarak değerlendirmesi anlamında kullanmaktadır (TÜİK, 2018).

Yaşam memnuniyeti kavramı literatürde temel olarak geçerlilik ve doğruluk çalışmaları ile, daha sonra tanımlayıcı istatistikler ve temel hipotez testleri ile çalışılmıştır. İleri düzeyde analizler için korelasyon ve regresyon modelleri kullanılırken, son yıllarda makine öğrenimi kavramı içersinde sınıflandırma ve karar ağaçları ile modellenmesi ile incelenmiştir.

Akın&Yalnız (2015), yaşam memnuniyeti ölçeğinin geçerliliğini ve güvenilirliğini doğrulayıcı faktör analizi, cronbach alpha katsayısı ve madde analizi ile; Dağlı&Baysal (2016), Diener ve arkadaşlarının yaşam memnuniyeti ölçeğinin Türkçe'ye uyarlanmasını Cronbach alfa ve doğrulayıcı faktör analizi yöntemleri ile yapmışlardır. Yaşam memnuniyeti ile ilgili istatistiksel dağılımların (t, z, ki-kare ve F) kullanıldığı çıkarsamalı istatistik son yıllarda birçok alanda kullanılmıştır. Saygılı vd. (2017), yaşam memnuniyetinin A ve B tipi kişilik özellikleri açısından farklılaşıp farklılaşmadığını bağımsız iki örneklem t testi ile; Bölükbaşı&Şentürk (2017), sigorta şirketinde çalışanların yaşam memnuniyeti ve bazı demografik özellikleri ilişkiyi ki-kare ilişki testi ile; Demir&Murat (2017) ve Taşlıyan vd. (2018), yaşam memnuniyeti ve bazı demografik özellikler ile bağımsız iki örneklem t testi ve ANOVA ile ve Tuncay & Fertelli (2018), yaşam aktiviteleri ile yaşam doyumu arasındaki ilişkiyi ki-kare ile incelemişlerdir.

Makine öğreniminde yer alan yöntemlerden Güler&Emeç (2006), regresyon analizi ve tanımlayıcı istatistikleri kullanarak akademik başarıda yaşam memnuniyetini, Demir (2011), korelasyon ve regresyon analizini kullanarak kimlik işlevleri ve yaşam memnuniyeti arasındaki ilişkileri, Tümlü&Recepoğlu (2013), regresyon analizi ile psikolojik dayanıklılık ve yaşam memnuniyeti arasındaki ilişkiyi, Korkmaz vd. (2015), birim kök, granger nedensellik ve regresyon analizler ile yaşam memnuniyetini incelemişlerdir.

Farklı makine öğrenim yöntemleri ile TÜİK verileri kullanılarak bireylerin yaşam memnuniyetini Gürsakal&Öngen (2008) diskriminant analizi ile; Beşel (2015), il bazlı olarak tanımlayıcı istatistiklerle; Berker (2015), zorunlu göç ile ilişkisini regresyon analizi ile; Arı&Yıldız (2016), sıralı lojistik regresyon analizi ile ve son olarak Şehribanoğlu&Diler (2018), karar ağacı algoritmalarından CART ve CHAID analizi ile incelemişlerdir.

Bu çalışmanın amacı, sınıflandırma ve karar ağacı yöntemlerinden olan CART ve lojistik regresyon çözümlemelerinin performanslarının Türkiye İstatistik Kurumuna ait (TÜİK) 2017 dönemini kapsayan yaşam memnuniyeti verilerini kullanarak karşılaştırılmasıdır. Buna bağlı olarak sonraki bölümlerde materyal ve metot, bulgular ve sonuçlar yer almaktadır.

2. Materyal ve Metot

2.1. Yöntem

Veri madenciliğinde kullanılan istatistiksel yöntemlerden en çok kullanılanları karar ağaçları, sınıflandırma modellerinde yer alırken, lojistik analizi, regresyon modelleri içinde yer almaktadır. Bu bölümde lojistik regresyon modeli, CART analizi ve performansların hesaplanması yer almaktadır.

2.1.1. Lojistik regresyon yöntemi

Basit ve çoklu doğrusal regresyonda bağımlı (sonuç) değişken sayısal ölçektedir. Ancak bağımlı değişkenin kategorik ya da sınıflayıcı&sıralayıcı ölçekte olduğu durumlarda doğrusal regresyonda parametre kestirimleri için kullanılan en küçük kareler yöntemini kullanmak bazı varsayımların (normal dağılıma uymayan hata terimi, sabit olmayan hata varyansı ve sonuç değişkeni üzerine kısıtlar) sağlanamadığından dolayı uygun olmamaktadır. Bu durumda lojistik regresyon kullanılmaktadır.

İkili lojistik regresyonda, Y_i, Bernoulli rassal değişkeninin olasılık dağılımı: $Y_i = 1$, π_i olasılıkla ve $Y_i = 0$, $1-\pi_i$ olasılıkla yazılır. Bağımsız değişkenler $X = (X_1, X_2, ...X_p)$ ile gösterildiğinde, ikili sonuç değişkeninin regresyon modeli, Y_i =0,1 olduğu durumda $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ şeklindedir. $E\{\varepsilon_i\} = 0$ olduğundan Y_i rassal değişkeninin olasılığı $E\{Y_i\} = \beta_0 + \beta_1 X_i$ yazılabilir. Y_i rassal değişkeni Bernoulli dağıldığı için, Y_i 'nin beklenen değeri, başarı olasılığı modellemiş olur.

Sonuç değişkeni ikili olduğu zaman sonuç değişkeni üzerindeki en büyük kısıt Y_i 'nin beklenen değeri üzerindedir ($0 \le E(Y_i) = \pi_i \le 1$). Bu kısıtın çözümlenmesi için lojit dönüşüm uygulanması gerekmektedir. Lojistik fonksiyonu ve bu fonksiyonu doğrusal hale getirmek için yapılan lojit dönüşüm (LF) aşağıdaki gibidir ($-\infty < LF(\pi(x_i)) < \infty$).

$$\pi(x_i) = E(Y = 1|x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

$$\ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = \ln\left[\frac{E(Y=1|x_i)}{E(Y=0|x_i)}\right] = \ln(e^{\beta_0+\beta_1 x_i}) = \beta_0 + \beta_1 x_i$$

Lojistik regresyon modelinde değişkenlere ait katsayıların kestirimi "en çok olabilirlik" yöntemi ile yapılır. En çok olabilirlik yöntemi ile kestirilen katsayıların anlamlılığı, olabilirlik fonksiyonlarına dayanan "olabilirlik oran testi, G istatistiği" veya ilgilenilen test istatistiğinin dağılımının standart normal dağılıma yaklaşımını kullanan "Wald testi" ile yapılır.

Lojistik regresyon modelinde değişkenlere ait katsayıların yorumu için odds oranından (OR)'den yararlanılır ($0 < OR < \infty$). Lojistik fonksiyonda yer alan "olmanın", "olmamaya" oranı $\pi(x_i)/(1-\pi(x_i))$ ile gösterilip "odds" olarak bilinir. OR değeri 0 ile 1 değeri arasında olursa risk faktörünün sonuç değişkeni için "koruyucu" olduğu, OR değeri 1 olursa risk faktörü ve sonuç değişkeni arasında bir fark olmadığı ve OR değeri 1'den büyük olursa risk faktörü ve sonuç değişkeni arasında bir fark olduğu ve bu farkın matematiksel olarak bir kat ile açıklanacağı şeklindedir. Ayrıca OR'a ait güven aralığı 1 değerini kapsamamalıdır.

2.1.2. Karar Ağacı ve CART Algoritması

Karar ağaçları, kurulması ve yorumlanması kolay, veri tabanı sistemleri ile kolayca elde edilebilir ve güvenilirliklerinin yüksek olması ile tercih edilen bir yöntemdir (Vahaplar, 2003). Karar ağacı yöntemini kullanarak verinin sınıflandırmasındaki ilk adım öğrenme basamağıdır. Burada önceden bilinen bir eğitim verisinde sınıflandırma algoritması kullanılarak model kurulur. Öğrenilen model, karar ağacı olarak gösterilir. İkinci adım ise eğitim verisinin karar ağacının doğruluğunu belirlemek amacıyla test edilerek kullanıldığı sınıflamadır. Eğer doğruluk kabul edilebilir oranda ise, kurallar yeni verilerin sınıflandırılması amacıyla kullanılır (Güner, 2014).

Karar ağaçlarının yapısı ağaç (kök, dallar, yapraklar) şeklindedir. Karar ağaçları, verideki tüm gözlemleri kapsayan kök ile başlar ve aşağıya doğru gittikçe veriyi alt gruplara ayıran dallara bölünür. Bu kökten dallara doğru büyüyen ağaç yapısında her boğum "düğüm" adını alır (Pehlivan, 2006). Düğümler üzerinde risk faktörlerin test işlemi yapılmakta ve test işleminin sonucu ağacın veri kaybetmeden dallara ayrılmasına neden olmaktadır. Her düğümde test ve dallara ayrılma işlemleri ardışık olarak gerçekleşmekte ve sonuç olarak ağaç sınıflar ile son bulmaktadır. Ayrılma işlemi bittikten sonra grup içindeki gözlemlerin kategorileri için oranlara bakılarak yorum yapılmaktadır (Lyn, 2000).

Karar ağacı oluşturmak için birçok algoritma kullanılmaktadır Bunlar; CHAID (Chi- Squared Automatic Interaction Detector), Exhaustive CHAID, CART (Classification and Regression Trees), ID3, C4.5, MARS (Multivariate Adaptive Regression Splines), QUEST (Quick, Unbiased, Efficient Statistical Tree), C5.0, SLIQ (Supervised Learning in Quest), SPRINT (Scalable Paralleizable Induction of Decision Trees) olarak sıralanabilir.

CART algoritması, 1984 yılında Breiman vd. tarafından önerilmiştir. Bu algoritma Morgan ve Sonquist'in AID (Automatic Interaction Detection) adlı karar ağacı algoritmasının devamı şeklindedir. Makine öğreniminin denetimli öğrenmesi içinde olan CART,

hem kategorik hem de sürekli değişkenleri kullanan sınıflandırma ve regresyon ağacı algoritmasıdır. CART algoritması üç adımdan oluşmaktadır.

a. Maksimum ağacın oluşturulması

Ağaç oluşturulması, ilgilenilen kümeyi kendinden daha homojen olan iki alt kümeye bölen bir yapıdır. Ağacın kökü, veri seti içerisindeki tüm risk faktörlerini kapsamaktadır. Ayrıca bu kök, her bir seviyede kendine özgü iki alt düğüm halinde bölünen bir ana düğüm olarak düşünülmektedir. Sonraki adımda, her alt grup bir ana grup olmaktadır. Her bölünme bir alt gruptaki tüm risk faktörlerin benzer sonuç değişkeni değerlerine sahip olacak şekilde seçilen bir açıklayıcının değeri ile tanımlanmaktadır (Kurt vd., 2008, Kıran, 2010).

En önemli risk faktörü, safsızlık (impurity) ölçütleri (Gini Diversity Index, Twoing, Chi-square, G-square) kullanılarak seçilir. Gini ölçütü her adımda en büyük veri kümesini elde eder. Böylece en iyi bölme durumu elde edilmiş olur. Ayrıca bölme işleminden sonra ilgilenmeyen kısım tek başına bırakılmış olur. Twoing ölçütü ise, Gini' ye göre daha dengeli bir yapı sunar. Bunun nedeni ise her defasında ana ve alt düğümlerin %50′ sini içermeye çalışmasıdır. Bundan dolayı Gini'ye göre daha yavaştır (Bozan, 2010, Yücel, 2017).

b. Ağaç budama

Maksimum ağaç yapıldıktan sonra bu ağaç aşırı öğrenme (overfitting) eğilimi göstermektedir. Yani CART algoritması, herhangi bir durma kuralı olmaksızın sürekli olarak bölünerek büyümektedir. Diğer modelleme yöntemlerinde olduğu gibi ağacın karmaşıklığı ve tahmin gücü arasında denge sağlamak için budama işlemi gerekmektedir. Artık yeni bir bölünmenin gerçekleşmeyeceği durumda, bu sefer, uçtan köke doğru budama işlemi başlatılır. Budama işlemi esnasında maksimum ağaçtan türetilen bir seri daha küçük alt ağaçlar arka arkaya gelen uç dallardan elde edilmekte böylece farklı alt ağaçlar en uygun olanla karşılaştırılmaktadır (Kıran, 2010; Sezer, 2010).

c. Optimum ağacın seçimi

Elde edilmiş alt ağaçlar arasından optimal olan seçilmek zorundadır. Bu seçim işlemi tahmin hatasının değerlendirilmesi üzerine kuruludur. Tahmin hatası ise çapraz geçerlilik testi kullanarak değerlendirilmektedir (Kıran, 2010). Optimal karar ağacı, her ağaç budama sonrasında, seçilen bir test verisi kullanılarak belirlenmeye çalışılır.

2.1.3. Performansların Karşılaştırılması

Bu çalışmada model performans değerlendirme ölçümü için ikili sınıflandırmaya dayalı performans değerlendirme ölçütleri kullanılmıştır. Bu ölçütler doğruluk (accuracy), duyarlılık (sensitivity), seçicilik (specificity), kesinlik (precision) ve F-skor olarak belirlenmiştir. Ölçütler için kullanılacak sınıflandırma ve ölçütlerin hesaplanması Tablo 1'de verilmiştir.

Tablo 1. Performanslar ve Hesaplamaları

		Gerçek Model		Performanslar	
		1	**0**	**Duyarlılık**	DP/(DP+YN)
Kestirim	1	DP	YP	**Seçicilik**	DN/(DN+YP)
Modeli	0	YN	DN	**Kesinlik**	DP/(DP+YP)
DP: Doğru Pozitif, DN: Doğru Negatif, YP: Yanlış Pozitif, YN: Yanlış Negatif				**Doğruluk**	(DP+DN)/(DP+YP+YN+DN)
				F-skor	2*(Kesinlik*Duyarlılık)/(Kesinlik+Duyarlılık)

"Duyarlılık", gözlenen modelin pozitif (1) olması durumunda kestirilen modelin de pozitif (1) olması, "seçicilik" ise gözlenen modelin negatif (0) olması durumunda kestirilen modelin de negatif (0) olması olasılığıdır. Kurulan çapraz tabloda duyarlılık ve seçicilik olasılıklarının aynı anda yüksek olması beklenir. ROC eğrisi ile değişik kesim noktalarında modelin duyarlılığının, modelin (1- seçicilik) oranına karşı noktalaması elde edilir. Her sınıflandırma işleminde yapıldığı gibi yöntemler, duyarlılık ve seçicilik arasındaki dengeyi kurmakla uğraşmaktadır. ROC eğrisi altında kalan alan AUC(area under curve) olarak tanımlanabilir ve bu AUC modelin pozitiflerle negatifleri ayırt edebilme başarısının en iyi göstergesi olarak kabul edilir. Bu alan 1 olduğunda pozitifler mükemmel bir şekilde negatiflerden ayrılmış demektir.

3. Araştırma Sonuçları ve Tartışma

Çalışmada kullanılan veriler, Türkiye İstatistik Kurumu'nun (TÜİK) 2017 "yaşam memnuniyeti" veri tabanından elde edilmiştir. "Yaşam memnuniyeti" veri tabanında 8430 (M=6408, %76; MD=2022, %24) kişiye ait yaşam memnuniyet skorları sınıflayıcı ölçekte verilmiştir (M: Yaşamdan Memnunum (1); MD: Yaşamdan Memnun Değilim (0)). Bu ölçek için 239 adet bağımsız değişken (risk faktörü) bulunmaktadır. Bu değişkenler; cinsiyet, yaş, çalışma durumu, çalışılan yer, işteki sorun, medeni durum ve memnuniyet ölçekleri (sağlık, evlilik, gelir, sosyal hayat, ulaşım, eğitim, konut, semt, iş, kazanç, kişisel bakım, trafikte geçirilen zaman, su, yeşil alan, akraba, arkadaş ve komşu ilişkileri, asayiş, adli durum, hastane, belediye, polis, kamu hizmetleri ve ülke durumu vb…) gibi başlıklarda toplanmaktadır. Yaşam memnuniyeti ile ilişkili olduğu düşünülen değişkenlerin bazılarının benzer ve birbiri ile ilişkili olmasından dolayı değişken sayısı korelasyon matrisine bakılarak azaltılmıştır. Bu matrise göre çalışmaya "cinsiyet", "medeni durum", "çalışma durumu", "sektör", "okul durumu", "yaş" kategorik değişkenleri ile "sağlık", "evlilik", "gelir", "sosyal hayat" ve "ulaşım" memnuniyetlerine ait ölçek değişkenlerinin alınmasına karar verilmiştir. Veri setinin istatistiksel özellikleri, veri analizi ve ön işleme için önemli olduğundan, seçilen kategorik değişkenlerin frekansları ve memnuniyet değişkeni ile ilişki durumunu gösteren ki-kare p değerleri Tablo 2'de verilmiştir. Tablo 3'de ise likert ölçeğindeki memnuniyet ölçeklerinin frekansları ve ki-kare p değerleri gösterilmiştir. Yapılan tüm analizler R programlama dili ile gerçekleştirilmiştir.

Tablo 2. Bağımsız Değişkenler için Tanımlayıcı İstatistikler ve Ki-Kare p-değerleri

	M n(%)	MD n(%)	Toplam n(%)	p değeri		M n(%)	MD n(%)	Toplam n(%)	p değeri
Cinsiyet					Okul				
Erkek	2983	954	3937(46.7)	0.034	1: Gitmiyor	816	412	1228(14.6)	0.000
Kadın	3492	1001	4493(53.3)		2: İlkokul	2047	765	2812(33.4)	
Medeni Durum					3: Ortaokul	838	300	1138(13.5)	
1:Evli Değil	1299	404	1703(20.2)	0.009	4: Lise	1326	324	1650(19.5)	
2:Evli	4720	1456	6176(73.3)		5: Önlisans	393	91	484(5.7)	
3:Dul	389	162	551(6.5)		6: Lisans	858	117	975(11.6)	
Çalışma Drumu					7: Yl.-Dr.	130	13	143(1.7)	
Çalışıyor	2920	810	3686(43.7)	0.000	Yaş				
Çalışmıyor	3488	1212	4744(56.3)		1: (-. 25)	1025	280	1305(15.5)	0.004
Sektör					2: (26.35)	1191	336	1527(18.1)	
Özel	2373	719	3092(82.9)	0.000	3: (36.45)	1262	387	1649(19.6)	
Kamu	547	93	640(17.1)		4: (46.55)	1061	380	1441(17.1)	
					5: (56.65)	852	310	1162(13.8)	
					6: (66. -)	1017	329	1346(16.0)	

Ki-kare ilişki testine göre, %95 güvenle tüm değişkenler ile yaşam memnuniyeti arasında istatistiksel olarak ilişki bulunmuştur (p değerleri<α=0.05, H_0: değişkenler arasında ilişki yoktur).

Tablo 3. Memnuniyet Ölçekleri için Frekanslar ve Ki-Kare p-değerleri

Ölçekler	S: Sağlık (M-MD)	G: Gelir (M-MD)	SH: Sosyal Hayat (M-MD)	U: Ulaşım (M-MD)
5	374-95	121-25	115-37	172-45
4	4363-1083	3172-642	3396-764	4513-1380
3	1038-361	1391-301	1380-368	879-241
2	571-405	1514-841	1359-721	721-274
1	62-78	210-213	158-132	123-82
p-değeri	0.000	0.000	0.000	0.000

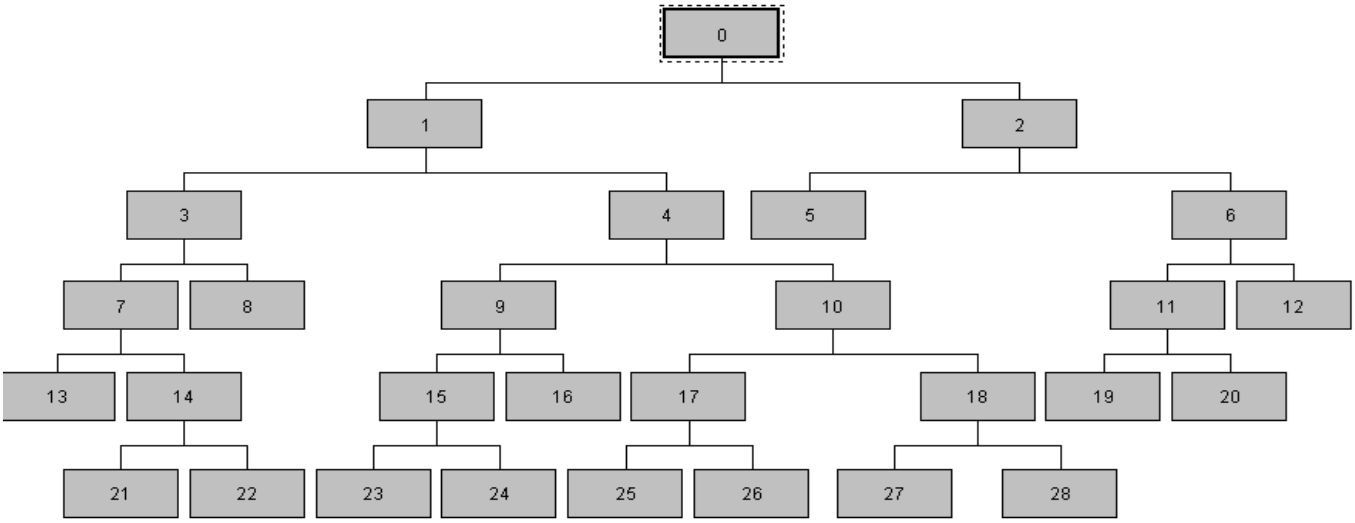
Tablo 3'e göre, yaşam memnuniyeti ile sağlık (S), gelir (G), sosyal hayat (SH) ve ulaşım (U) ölçekleri arasında %95 güvenle ilişki olduğu tespit edilmiştir (p değerleri<0.05)(5: Çok Memnun(ÇM), 4: Memnun(M), 3: Orta(O), 2: Memnun Değil(MD), 1: Hiç Memnun Değil(HMD)).

Yaşam memnuniyeti verileri, makine öğreniminin denetimli öğrenmesi içersinde olan regresyon adımı için geriye doğru lojistik regresyon modeli ile anlamlı bulunan değişkenlerle kurulmuştur. Verilerin %80'i eğitim ve %20'si test verisi olacak şeklinde kurulan lojistik regresyon modelinde, kadınların erkeklere göre yaşamdan memnun olma şansı, %18 daha fazla olarak elde edilmiştir (OR=1.18). Okul durumlarında, ilkokuldan(2) başlayarak, YL&Dr(7)'a kadar tüm kategorilerin, okula gitmeyenlere göre daha çok memnun oldukları bulunmuştur. YL&Dr'sı olanların okula gitmeyenlere göre yaşamdan memnun olma şansı yaklaşık olarak 3.39 kat daha fazla tespit edilmiştir. Evli ya da boşanmış olanların, evlenmeyenlere göre yaşamdan memnun olma şansı sırasıyla %31 ve %21 daha fazladır. Sağlık memnuniyetinde, M olanların, HMD'e göre yaşam memnuniyetinin olma şansı yaklaşık olarak 2.75 kat daha fazla olarak elde edilmiştir. Gelir memnuniyetinde, MD olanların, HMD'e göre yaşam memnuniyeti'nin olma şansı yaklaşık olarak %50 daha fazla olarak bulunmuştur. Sosyal hizmet memnuniyetinde, ÇM ile MD'in OR değerlerine ait güven aralığı 1'i içerdiğinden yorumlama yapılamaz. Ancak sosyal hizmet memnuniyetinde, M olanların, HMD'e göre yaşamdan memnun olma şansı %96 daha fazla olarak görülmüştür. Ulaşım memnuniyetinde, ÇM olanların, HMD'e göre yaşam memnuniyeti'nin olma şansı yaklaşık olarak 2.82 kat daha fazla tespit edilmiştir. Bu odds değerlerinin hepsinin 1'i içermemesi de yorumlara destekleyici olarak ifade edilmiştir.

Yaşam memnuniyeti verileri, makine öğreniminin denetimli öğrenmesi içersinde olan sınıflama ve regresyon ağaçları algoritmalarından CART ile maksimum ağaç oluşturulup, budama algoritması ile en uygun ağaç yapısı oluşturulmuştur. Veri setinin %80'i eğitim %20'si test verisi olarak ayrılan lojistik regresyon model için kullandığımız veri seti CART algoritmasında Gini ayırma kriteri ile kullanılmıştır. Şekil 1'de bu ağacın ilk bölünmesi gösterilmiştir. Tablo 5'de ise bu ağaca ait terminal ve alt düğümler, sonuç değişkenleri, dallar ve bu düğümlere ait yorumlar verilmiştir. Düğümler 0'dan 28'e kadar numaralandırılmıştır. Gini indeksi kullanılarak yapılan CART analizi ile bağımsız değişkenlere ait hem alt düzeyler belirlenmekte hem de alt düzeyler arasındaki ilişkiler elde edilmiştir. Yaşam memnuniyeti üzerinde en önemli değişkenin "gelir" olduğu CART ağacı ile bulunmuştur. Gelir değişkeni iki alt düğüme (1. alt düğüm= HMD,MD; 2. alt düğüm=O, M, ÇM) bölünmüştür. 1. alt düğümde yaşamdan memnun olanlarının oranının %61.6, memnun olmayanların ise %38.4 olduğu tespit edilmiştir. Gelir değişkeninde 1. alt düğüm, iki alt düğüme (3. alt düğüm= HMD, M, O; 4. alt düğüm=M, ÇM) daha bölünmüştür. 4. alt düğümde yaşamdan memnun olanlarının oranının %69.0, memnun olmayanların ise %31.0 olduğu tespit edilmiştir. CART analizinde bu işlem 28. Alt düğüme kadar devam etmiştir.

Tablo 4. Lojistik Regresyon Modeli (%80 Eğitim Verisi): Wald Ki-Kare p-Değeri ve Odds Oranı Güven Aralıkları

Değişkenler	p-değeri	OR(GA)	Değişkenler	p-değeri	OR(GA)
Cinsiyet	0.009	1.18 (1.042-1.336)	Gelir	0.000	
Medeni Durum	0.005		5: ÇM	0.000	2.79 (1.601-4.877)
Evli	0.001	1.31 (1.111-1.534)	4: M	0.000	3.03 (2.317-3.964)
Dul	0.182	1.21 (0.914-1.608)	3: O	0.000	3.15 (2.379-4.178)
Okul	0.000		2: MD	0.002	1.50 (1.158-1.943)
2: İlkokul	0.001	1.34 (1.123-1.609)	Sosyal Yaşam	0.000	
3: Ortaokul	0.002	1.41 (1.131-1.771)	5: ÇM	0.627	1.14 (0.665-1.968)
4: Lise	0.000	2.04 (1.637-2.547)	4: M	0.000	1.96 (1.420-2.693)
5: Önlisans	0.000	2.19 (1.593-3.006)	3: O	0.000	2.01 (1.443-2.789)
6: Lisans	0.000	3.09 (2.353-4.071)	2: MD	0.063	1.35 (0.984-1.851)
7: Yl.-Dr.	0.000	3.39 (1.802-6.408)	Ulaşım	0.000	
Sağlık	0.000		5: ÇM	0.000	2.82 (1.661-4.778)
5: ÇM	0.000	2.39 (1.466-3.882)	4: M	0.000	2.07 (1.464-2.928)
4: M	0.000	2.75 (1.829-4.147)	3: O	0.000	2.61 (1.790-3.807)
3: O	0.000	2.17 (1.423-3.308)	2: MD	0.004	1.73 (1.193-2.521)
2: MD	0.117	1.40 (0.918-2.146)	Sabit	0.000	0.08
Referans: Erkek, Evli değil ve memnuniyet ölçeklerinde HMD					
Ölçekler: 5: Çok Memnun(ÇM), 4: Memnun(M), 3: Orta(O), 2: Memnun Değil(MD), 1: Hiç Memnun Değil(HMD)					



Şekil 1. CART Analizi (%80 Eğitim Verisi) Karar Ağacı Numaralandırılması

Cart analizi sonucunda yaşam memnuniyeti üzerine en etkili değişkenin karar ağacına göre "gelir" olduğu bulunmuştur. Ancak CART ağacında görülmeyen önem seviyeleri için "normalleştirilmiş önemlilik (normalized importance)" değerleri ile önem sıraları Tablo 6'da gösterilmiştir. Bu değerleri göre"gelir" değişkenini sırasıyla "sosyal hayat(%86)" ve "sağlık durumu(%73.4)" değişkeninin takip ettiği görülmüştür.

Tablo 6. CART Analizi Normalleştirilmiş Önemlilik Değerleri

Bağımsız Değişkenler	Önem Değerleri	Normalleştirilmiş Önemlilik Değerleri (%)
Gelir	0.023	100.0
Sosyal Hayat	0.020	86.0
Sağlık	0.017	73.4
Eğitim	0.007	28.1
Ulaşım	0.002	9.8
Cinsiyet	0.001	4.0
Medeni Durum	0.001	3.1

Veri setini eğitim ve test set olarak ayırmamızın amacı, olası aşırı uyma'dan (overfitting) kaçınmak ve modelin daha önceden görmediği veri seti üzerinde nasıl performans gösterdiğini anlamak içindir. Ancak modelin dağılımdan kaynaklı bazı hatalar olabilir. Bu çalışmada TÜİK verilerindeki hataları minimum seviyeye indirmek için "k–katlamalı çapraz geçerlilik (k-fold cross validation)" yöntemi kullanılmıştır. Bunun için eğitim verisi, rasgele 10 parçaya bölünmüş ve 9 parça eğitim için 1 parçada test verisi için kullanılmıştır. Bu işlem 10 kez tekrarlanmıştır. Her tekrardan elde edilen değerlerin ortalaması ile sınıflandırıcımızdan gelen performanslar değerlendirilmiştir.

Tablo 5. CART Analizi (%80 Eğitim Verisi)

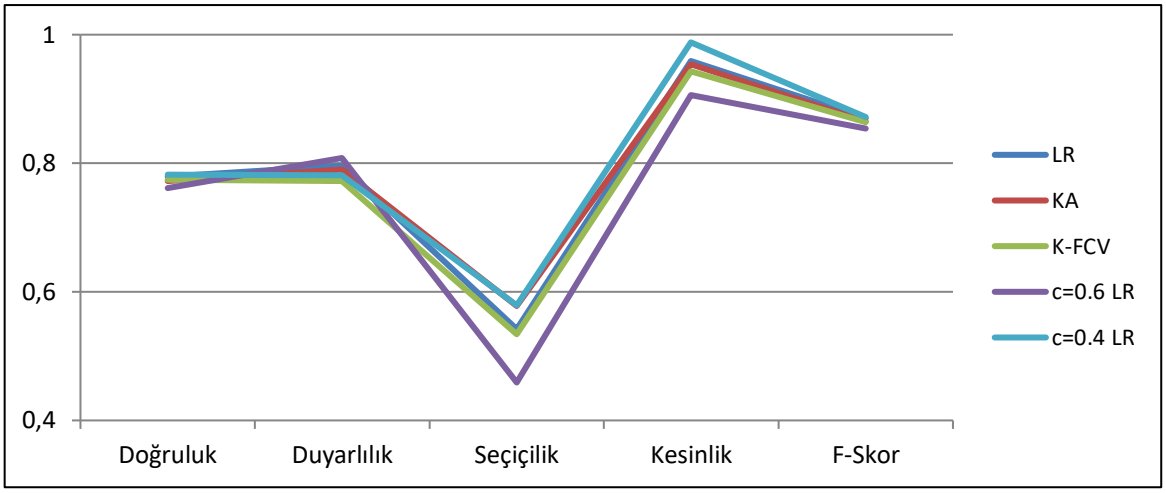
			Memnun(M)		Memnun Değil(MD)	
Düğüm	Değişken	Sınıflar	n	%	n	%
0			5080	75.7	1631	24.3
1	Gelir	MD, HMD	1370	61.6	855	38.4
2	Gelir	ÇM, M, O	3710	82.7	776	17.3
3	Sağlık	O, MD, HMD	444	50.3	438	49.7
4	Sağlık	M, ÇM	926	69.0	417	31.0
5	Okul	İlkokul, Ortaokul, okula gitmedi	2029	78.1	569	21.9
6	Okul	Önlisans, Lise, Lisans, YL-Dr	1681	89.0	207	11.0
7	Sosyal Hayat	MD, HMD	220	41.0	316	59.0
8	Sosyal Hayat	ÇM, M, O	224	64.7	122	35.3
9	Ulaşım	MD, HMD	152	57.8	111	42.2
10	Ulaşım	ÇM, M, O	774	71.7	306	28.3
11	Sosyal Hayat	M, O, MD, HMD	1631	89.6	190	10.4
12	Sosyal Hayat	ÇM	50	74.6	17	25.4
13	Gelir	MD, HMD	36	28.1	92	71.9
14	Gelir	ÇM, M, O	184	45.1	224	54.9
15	Gender	Erkek	70	49.3	72	50.7
16	Gender	Kadın	82	67.8	39	32.2
17	Sosyal Hayat	MD, HMD	329	66.1	169	33.9
18	Sosyal Hayat	ÇM, M, O	445	76.5	137	23.5
19	Okul	Önlisans, Lise	981	88.0	134	12.0
20	Okul	Lisans, YL-Dr	650	92.1	56	7.9
21	Ulaşım	MD, HMD	28	34.1	54	65.9
22	Ulaşım	ÇM, M, O	156	47.9	170	52.1
23	Okul	Lise, Lisans, Okula gitmedi	23	38.3	37	61.7
24	Okul	Önlisans, İlkokul, Ortaokul, Yl-Dr	47	57.3	35	42.7
25	Gender	Erkek	167	62.1	102	37.9
26	Gender	Kadın	162	70.7	67	29.3
27	Okul	İlkokul, Ortaokul, Okula gitmedi.	266	71.7	105	28.3
28	Okul	Önlisans, Lise, Lisans, YL-Dr	179	84.8	32	15.2

Kurulan bir lojistik regresyon modelinin sonuçlarını özetlemek için en iyi yol sınıflandırma tablosu oluşturmaktır. Bu tablo, sonuç değeri Y'nin düzeyleri ile kestirilen lojistik olasılıklar tarafından üretilen ikili bir değişkenin çarpraz sınıflandırmasıyla elde edilir. Türetilen bu ikili bağımsız değişkeni elde etmek ini c kesim noktası belirlenir ve kestirilen her bir olasılık değeri c ile karşılaştırılır. Eğer kestirilen olasılık, c değerini geçerse, türetilen ikili değişken 1'e eşit olur, diğer durumlarda 0'a eşittir. c'nin en yaygın kullanılan değeri 0.5'dir. Buna bağlı olarak lojistik regresyonda değiştirilen yeni c değerleri, CART ve k–katlamalı çapraz geçerlilik yöntemleri için performans değerlendirmeleri Tablo 7'de verilmiştir.

Tablo 7. Modellerin Algoritmalarına Göre Performans Değerlendirmesi

Algoritma	Doğruluk	Duyarlılık	Seçicilik	Kesinlik	F-Skor	R^2
Lojistik Regresyon (c=0.5)	0.779	0.796	0.552	0.961	0.870	0.671
Karar Ağacı	0.780	0.790	0.578	0.964	0.875	0.543
k-Katlamalı Çağraz Doğrulama	0.774	0.772	0.534	0.943	0.864	0.544
Lojistik Regresyon (c=0.6)	0.761	0.808	0.459	0.906	0.854	0.643
Lojistik Regresyon (c=0.4)	0.782	0.781	0.579	0.988	0.872	0.675

Yaşam Memnuniyeti TUİK veri seti için doğruluk değerleri incelendiğinde c değeri 0.4 iken yapılan lojistik regresyon modeli en yüksek değere sahiptir. Duyarlılık değerleri olarak karşılaştırıldığında c değeri 0.6 olan lojistik regresyon modeli en yüksek sonuç vermiştir. Seçiçilik ve kesinlik değerlerine bakıldığında c değeri 0.4 olarak belirlenen lojistik regresyon modeli en yüksek sonuç vermiştir. F-skoru için en yüksek sonuç veren algoritma karar ağacıdır. Bu algoritmaların çalıştırılması sonucunda elde edilen performans değerleri karşılaştırmalı olarak Şekil 2'de gösterilmiştir.



Şekil 2 Modellerin Performans Değerlendirilmesi

4. Sonuç

Bilgi özellikle son yıllarda her alanda ve herkes için çok önemli bir kavram haline gelmiştir. Ancak çok fazla verinin olması ve bu verilerin kullanılabilir bilgiye dönüştürülmesi iyi yönetilmesiyle mümkün olabilir. Bilgi haline dönüştürülmeyen ham verinin bulunduğu alana katkısı olmaz. Teknolojik gelişmelerle birlikte varolan verilerin kullanılması, çözümlenmesi ve yorumlanması alanda yer alan şirket ya da kişiler için önem oluşturmaktadır. Bununla birlikte yapılan veri madenciliği çalışmalarında verinin hazırlanma aşamasının süreç içerisinde en çok zaman alan kısım olması, veri kalitesi ve bütünlüğünün önemini göstermektedir. Verilerin yapılarındaki bozukluk, farklı kullanıcıların farklı biçimlerde ya da eksik veri girişi, tutarlı olmayan ya da işlevsel olmayan veri yapıları, veri kalitesini ve bütünlüğünü bozan unsurlar olarak gösterilebilir. Bunların tamamının veri ve verinin bilgiye dönüşüm süreciyle ilgili eksikliklerden kaynaklandığını söylenebilir.

Her alanın kendine özgü bir veri yapısı bulunmaktadır. Birçok alanda da ilgilenilen değişkenler kategorik yapıda kendini göstermektedir. Bugüne kadar kategorik verilerin sınıflandırılmasında ve çözümlenmesinde daha çok kümeleme, diskriminant ve lojistik regresyon analizi gibi çok değişkenli istatistiksel yöntemlerden yararlanılmıştır. Ancak teknolojik gelişmelerin varlığında veri büyüklüğünün artması ile bu yöntemlere göre daha yeni ve popüler olan karar ağacı algoritmaları kullanılmaya başlanmıştır.

Uygulamaya konu olan veri kümesi 2018 yılı TÜİK "Yaşam memnuniyeti" anketine katılan kişilerdir. Bu amaçla literatürde yapılan çalışmalarda TÜİK ve yaşam memnuniyeti birlikte incelendiğinde, Gürsakal&Öngen (2008) 2007 yılı TÜİK yaşam memnuniyetini inceledikleri çalışmada diskriminant analizini kullanarak cinsiyet, kır-kent durumu, sağlık, konut, gelir, akraba ve komşu ilişkileri değişkenlerini anlamlı bulmuşlardır. Beşel (2015), 2013 TÜİK verilerini kullanarak yaşam memnuniyetlerini tanımlayıcı istatistiklerle incelemiş ve mutlu&mutsuz illeri sosyal ve siyasal açıdan karşılaştırmıştır. Arı&Yıldız (2016), sıralı lojistik regresyon yöntemi ile 2014 yılına ait TÜİK yaşam memnuniyetini incelemişlerdir. Kısmi orantısal oran modeline göre cinsiyet, medeni durum, çalışma durumu, sağlık, gelir ve arkadaş değişkenlerini anlamlı bulmuşlardır. Son olarak Şehribanoğlu&Diler (2018), karar ağacı algoritmalarından CART ve CHAID analizini kullanarak 2013 yılına ait TÜİK yaşam memnuniyetinii incelemişlerdir. Bu çalışmada CHAID analizi ile gelir, CART analizi ile gelir, umut, sağlık, evlilik ve sosyal güvenlik değişkenlerinin anlamlı olduklarını bulmuşlardır. Ancak model performanslarını karşılaştırmayıp, modele giren değişken sayısına göre CHID algoritmasının daha iyi olduğunu ifade etmişlerdir.

Yapılan bu çalışmada 2018 yılı TÜİK "Yaşam memnuniyeti" anketine katılan 8430 kişiye ait yaşam memnuniyeti sınıflayıcı ölçekte elde edilmiştir. Bu ölçek için veri hazırlama aşamasında yürütülen işlemlerle modellere 7 değişken alınmıştır. İlk olarak modeldeki risk faktörleri için tahmin edilen odds oranları yardımıyla yorumlamanın yapıldığı lojistik regresyon analizi uygulanmıştır. c=0.5 kesim noktasına göre yapılan lojistik regresyon analizinde cinsiyet, medeni durum, okul durumu, gelir, sosyal hayat, sağlık ve ulaşım değişkenlerinin anlamlı bulunduğu tespit edilmiştir. Bu modele göre kadınların, evli veya dul olanların, okuyanların ve ölçekte HMD kategorisinin dışındaki diğer kategorilerin çoğunun yaşamdan memnun oldukları sonucu elde edilmiştir. CART analizi uygulamasında yaşam memnuniyeti üzerine en etkili değişkenin karar ağacına göre gelir olduğu bulunmuştur. Ayrıca normalleştirilmiş önemlilik değerlerine göre gelirden sonra önemli bulunan değişkenlerin sosyal hayat ve sağlık durumu olduğu saptanmıştır. TÜİK verilerindeki hataları minimum seviyeye indirmek için k–katlamalı çapraz geçerlilik yöntemi de kullanılmıştır. Sınıflandırma ve karar ağaçları içersinde yer alan lojistik regresyon, CART ve k–katlamalı çapraz geçerlilik yöntemleri için yapılan performans değerlendirilmesinde, c=0.4 kesim noktası olan lojistik modeli tercih edilmelidir.

Bu çalışma ile TÜİK verilerinde yer alan demografik ve ölçek verilerinin çözümlenmesi ve yorumlanması için makine öğreniminde yer alan algortimaların kullanımı gösterilmiştir. Çalışmada kullanılan sınıflandırma ve karar ağaçları yöntemleri ile aynı özellikteki veriler kullanılarak yapılabilecek performans karşılaştırılmaları belirlenmiştir. Ayrıca aynı türde yeni veriler ortaya çıktığında bu verilerin hangi sınıfta yer alması gerektiğine ilişkin ileriye yönelik tahminler makine öğrenimi yöntemleri ile kolaylıkla yapılabilecektir.

Kaynakça

Akın, A. &. Yalnız, A. (2015). Yaşam Memnuniyeti Ölçeği Türkçe Formu: Geçerlilik ve Güvenilirlik Çalışması, Elektronik Sosyal Bilimler Dergisi, 4(54): 95-102.

Arı, E. & Yıldız, Z. (2016). Bireylerin Yaşam Memnuniyetini Etkileyen Faktörlerin Sıralı Lojistik Regresyon Analizi ile İncelenmesi, Uluslararası Sosyal Araştırmalar Dergisi, 9(42): 1362-1374.

Berker, A. (2015). Zorunlu Göçün Yol Açtığı Refah Kaybının İncelenmesi: Yaşam Memnuniyeti Yaklaşımı, Türkiye ve Orta Doğu Amme İdaresi Enstitüsü, 1-30.

Beşel, F. (2015). 2013 Yılı Yaşam Memnuniyeti Araştırma Sonuçlarının İl Bazlı Ekonomik, Sosyal ve Siyasi Analizi, Karabük Ünv. Sosyal Bilimler Enstitüsü Dergisi, 5(2): 227-236.

Bozan, F. (2010). www.farukbozan.com/2010/01/cartclassification-and-regression-tree/. Son Erişim Tarihi: 18 Şubat 2020.

Bölükbaşı A. & Şentürk, Ö. (2017). Sigorta Sektöründe Çalışanların Yaşam Memnuniyeti Üzerine Bir Araştırma, Finansal Araştırmalar ve Çalışmalar Dergisi, 9(17): 2017.

Cox, M. & Ellsworth, D. (1997). Application-Controlled Demand Paging for Out-of-core Visualization, Proceedings of the 8th Conference on Visualization, IEEE.

Dağlı, A. & Baysal, N. (2016). Yaşam Doyumu Ölçeğinin Türkçe'ye Uyarlanması: Geçerlilik ve Güvenilirlik Çalışması, Elektronik Sosyal Bilimler Dergisi, 15(59): 1250-1262.

Demir, İ. (2011). Gençlerde Yaşam Doyumu ile Kimlik İşlevleri Arasındaki İlişkilerin İncelenmesi, Elektronik Sosyal Bilimler Dergisi, 10(38): 99-113.

Demir, R. & Murat, M. (2017). Öğretmen Adaylarının Mutluluk, İyimserlik, Yaşam Anlamı ve Yaşam Doyumlarının İncelenmesi, Uluslararası Toplum Araştırmaları Dergisi, 7(13): 347-378.

Doğan, Ş. & Türkoğlu, İ. (2007). Hypothyroidi and Hyperthyroidi Detection from Thyroid Hormone Parameters by Using Decision Trees, Doğu Anadolu Bölgesi Araştırmaları Dergisi, 5(2): 163-169.

Doğan, T. & Sapmaz, F. (2012). Oxford Mutluluk Ölçeği Türkçe Formunun Psikometrik Özelliklerinin Üniversite Öğrencilerinde İncelenmesi, Düşünen Adam Psikiyatri ve Nörolojik Bilimler Dergisi, 25: 297-304.

Gahi, Y., Guennoun, M., Mouftah, H. T. (2016). Big Data Analytics: Security and Privacy Challenges. 2016 IEEE Symposium on Computers and Communication (ISCC), Messina, Italy, 953.

Güler, B.K. & Emeç, H. Yaşam Memnuniyeti ve Akademik Başarıda İyimserlik Etkisi, Dokuz Eylül Ünv. İktisadi ve İdari Bilimler Fakültesi Dergisi, 21(2): 129-149.

Güner, Z.B. (2014). Veri Madenciliğinde CART ve Lojistik Regresyon Analizinin Yeri: İlaç Provizyon Sistemi Verileri Üzerinde Örnek Bir Uygulama, Sosyal Güvence Dergisi, 6: 53-99.

Gürsakal, S. & Öngen, K.B. (2008). 2007 Yaşam Memnuniyeti Anketinin İstatistiksel Yöntemler İle Analizi, Uludağ Ünv. İİBF Dergisi, XXVII(1): 1-14.

Hand, D.J. (1998). Data Mining: Statistics and More? The American Statistician, 52(2): 112-118.

Kanbur, E. & Özdemir, B. (2017). Yaşam Memnuniyeti ve Öncülleri: Karadeniz Bölgesi İncelemesi, Al-Farabi Uluslararası Sosyal Bilimler Dergisi, 1(1): 147-157.

Kıran, N. , Srinath, M., Sharma, R. (2010). A Level Set Method-Based Derivation Of Differential Equation For Developable Surfaces, International Electronic Journal of Geometry 3: 11-15

Kurt, I., Ture, M., Kurum, A. T., (2008). Comparing Performances of Logistic Regression, Classification and Regression Tree, and Neural Networks for Predicting Coronary Artery Disease, Expert Systems with Applications, 34: 366–374.

Korkmaz, M., Germir, H.N., Gürkan, A. (2015). Yaşam Memnuniyeti Üzerine Etkili Olan Sosyodemografik Bileşenler Üzerine Bir Analiz, Uluslararası Hakemli Psikiyatri ve Psikoloji Araştırmaları Dergisi, 3(2): 78-111.

Lyn, T. C. (2000). A Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Consumer, International Journal of Forecasting, 16(2): 149–172.

Pehlivan, G. (2006). Chaid Analizi ve Bir Uygulama, Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, İstanbul.

Saygılı M., Onay, Ö.A., Ayhan, M. (2017). Kişilik Özellikleri Bağlamında Yaşam Memnuniyeti Üzerine Bir Araştırma, Yorum-Yönetim-Yöntem Uluslararası Yönetim-Ekonomi ve Felsefe Dergisi, 5(2): 61-72.

Sezer, E. A., Bozkır, A. S., Yağız, S., Gökçeoğlu C. (2010). Karar Ağacı Derinliğinin CART Algoritmasında Kestirim Kapasitesine Etkisi: Bir Tünel Açma Makinesinin İlerleme Hızı Üzerinde Uygulama, Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu, Kayseri.

Şehiribanoğlu, S. & Diler, S. (2018). 2013 Yılı Yaşam Memnuniyet Araştırmasının CART ve CHAID Algoritmaları ile İncelenmesi, The Journal of Academic Social Science, 6(67): 132-141.

Taşlıyan M., Hırlak, B, Güler, B., Gündoğdu, E. (2018). İnternet Bağımlılığı, Yaşam Doyumu ve Bazı Demografik Değişkenler Arasındaki İlişkiler, Osmaniye Korkut Ata Üniversitesi, İİBF Dergisi, 2(2):166-179.

Tuncay, F.Ö. & Fertelli, T.K. (2018). Yaşlılarda Bilişsel İşlevlerin Günlük Yaşam Aktiviteleri ve Yaşam Doyumu ile İlişkisi, DEÜ Tıp Fakültesi Dergisi, 32(3): 183-190.

TÜİK (Türkiye İstatistik Kurumu). Yaşam Memnuniyeti Araştırması, 2018.

Tümlü, G.Ü. & Recepoğlu, E. (2013). Üniversite Akademik Personelinin Psikolojik Dayanıklılık ve Yaşam Doyumu Arasındaki İlişki, Yükseköğretim ve Bilim Dergisi, 3(3): 205-213.

Vahaplar, A. (2003). Bir Coğrafi Veri Madenciliği Uygulaması, Yüksek Lisans Tezi, Ege Üniversitesi Fen Bilimleri Enstitüsü, İzmir.

Yücel, Y. B. (2017). Yaşam memnuniyetini etkileyen faktörlerin sınıflama ve regresyon ağacı ile belirlenmesi, Yüksek Lisans Tezi, İstanbul Ticaret Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik, İstanbul.