Çukurova Üniversitesi
Mühendislik Fakültesi Dergisi
Çukurova University
Journal of the Faculty of Engineering
ISSN: 2757-9255 CİLT/VOLUME: 41 SAYI/ISSUE: 1 MART/MARCH 2026

Büyük Veri ile Türkiye'de Hane İnternet Erişim Durumunun Makine Öğrenmesi ve Açıklanabilir Yapay Zekâ ile Analizi

Tolga TUNÇEL [1,a], **Emre KILINÇ** [2,b]

[1]*Ankara Üniversitesi, Bilişim Teknolojileri Meslek Yüksekokulu, Büyük Veri Analistliği Programı, Ankara Türkiye*
[2]*Ağrı İbrahim Çeçen Üniversitesi, Patnos Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, Ağrı, Türkiye*

[a]***ORCID**: 0000-0002-7427-2071; [b]**ORCID**: 0000-0002-5250-9322*

Makale Bilgileri
Geliş : 02.08.2025
Kabul : 30.01.2026
DOI: 10.21605/cukurovaumfd.1755138

Sorumlu Yazar
Tolga TUNÇEL
tolgatuncel@ankara.edu.tr

Anahtar Kelimeler
Büyük veri analizi
XAI
SHAP
LIME
Açıklanabilir yapay zekâ

***Atıf şekli:** TUNÇEL, T., KILINÇ, E., (2026). Büyük Veri ile Türkiye'de Hane İnternet Erişim Durumunun Makine Öğrenmesi ve Açıklanabilir Yapay Zekâ ile Analizi. Çukurova Üniversitesi, Mühendislik Fakültesi Dergisi, 41(1), 89-100.*

ÖZ

Bu çalışmada, Türkiye İstatistik Kurumu (TÜİK) tarafından 2024 yılında gerçekleştirilen Hanehalkı Bilişim Teknolojileri Kullanım Araştırması (HBTKA) mikro verileri kullanılarak, hanehalkı internet erişim durumunu belirleyen temel faktörler makine öğrenmesi ve açıklanabilir yapay zekâ (AYZ) yaklaşımlarıyla analiz edilmiştir. CatBoost, LightGBM, Lojistik Regresyon, Rastgele Orman ve XGBoost modelleri ile yüksek sınıflandırma başarımı elde edilmiş; model çıktıları SHapley Additive exPlanations (SHAP) ve Local Interpretable Model-agnostic Explanations (LIME) yöntemleri kullanılarak hem küresel hem de yerel düzeyde yorumlanmıştır. Elde edilen bulgular, hane aylık gelir düzeyi, hane birey sayısı ve dijital cihaz sahipliğinin internet erişimi üzerinde belirleyici rol oynadığını ortaya koymaktadır. Bölgesel değişkenlerin etkisi sınırlı kalmıştır. Sonuçlar, Türkiye'de dijital uçurumun azaltılmasında ekonomik kapasite ve dijital donanım odaklı politika müdahalelerinin kritik önem taşıdığını göstermektedir.

Analysis of Household Internet Access Status in Turkey with Big Data Using Machine Learning and Explainable Artificial Intelligence

Article Info
Received : 02.08.2025
Accepted : 30.01.2026
DOI: 10.21605/cukurovaumfd.1755138

Corresponding Author
Tolga TUNÇEL
tolgatuncel@ankara.edu.tr

Keywords
Big data analysis
XAI
SHAP
LIME
Explainable artificial intelligence

***How to cite:** TUNÇEL, T., KILINÇ, E., (2026). Analysis of Household Internet Access Status in Turkey with Big Data Using Machine Learning and Explainable Artificial Intelligence. Cukurova University, Journal of the Faculty of Engineering, 41(1), 89-100.*

ABSTRACT

This study investigates the key determinants of household internet access in Turkey using microdata from the 2024 Household Information Technologies Usage Survey conducted by the Turkish Statistical Institute (TurkStat). Advanced machine learning models, including CatBoost, LightGBM, Logistic Regression, Random Forest, and XGBoost, were employed to achieve high classification performance. Model predictions were interpreted at both global and local levels using explainable artificial intelligence (XAI) techniques, namely SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). The results indicate that monthly household income, household size, and digital device ownership are the most influential factors affecting internet access status. Regional variables exhibit a comparatively limited effect. These findings demonstrate that household internet access in Turkey is primarily driven by economic capacity and digital equipment availability, highlighting the importance of income- and device-oriented policy interventions to effectively reduce the digital divide.

1. GİRİŞ

Geçmiş yıllarda lüks bir araç olarak görülen internet, günümüzde lüks bir araçtan temel bir ihtiyaca dönüşmüştür. Eğitim, iletişim, sağlık ve kamu hizmetleri gibi temel alanlar başta olmak üzere birçok farklı alanda aktif olarak kullanılmaktadır. İnternet erişimi giderek yaygınlaşsa da evrensel değildir. Özellikle gelişmemiş ve gelişmekte olan ülkelerde birçok hanenin internet bağlantısı bulunmamaktadır. Genellikle dijital uçurum olarak isimlendirilen bu olgu, sosyoekonomik, coğrafi, gelir gibi birçok parametreye bağlıdır. Bu kapsamda, dijital uçurumun makine öğrenmesi yöntemleriyle analiz edildiği çalışmalarda bulunmaktadır. Örneğin Indo-Myanmar bölgesi için internet erişiminin dijital okuryazarlığa etkisi makine öğrenmesi yöntemleri ile incelenmiştir [1]. Çalışmada, Indo-Myanmar bölgesinde 342 katılımcıdan toplanan mikro veriler farklı makine öğrenmesi yöntemleri ile analiz edilmiş ve internet erişimi ile dijital okuryazarlığın birbirini desteklediği gözlemlenmiştir. Nepal'de bulunan 15000 haneyi içeren bir veri seti üzerinden dijital eşitsizliği incelemek için Rastgele Orman modeli kullanılmıştır [2]. Dijital eşitsizliğin hane düzeyinde bilgi ve iletişim teknolojilerine erişim olarak değerlendirilen bu çalışmada, elde edilen sonuçlara göre bilgi ve iletişim teknolojilerine erişimin en güçlü belirleyicileri olarak zenginlik ve eğitim düzeyi olduğu ortaya çıkmıştır. Kore'de yaşlı bireylerin dijital uçurum durumunu incelemek için makine öğrenmesi yöntemleri kullanılmıştır [3]. Dijital uçurumu belirlemede en önemli değişkenlerin, yaş, eğitim, internet kullanım düzeyi, dijital teknolojileri öğrenme konusundaki eğilim ve çevrimiçi sosyal bağlılık olduğu görülmüştür. Dijital uçurum literatürüne ek olarak, internet erişiminin sonuç değişkenleri üzerindeki etkisini inceleyen çalışmalar da mevcuttur. Covid-19 pandemisi öncesi ve sonrasında hanelerin internet kullanım talebindeki değişimi ve bu talebin etkili parametreleri analiz edilmiştir [4]. Parametrelerin tamamı anlamlı olsa da özellikle hane geliri ve hane büyüklüğünün internet kullanım olasılığını artırdığı, yaş parametresinin yüksek olmasının kullanım olasılığını azalttığı gözlemlenmiştir. İnternet erişim olasılığı düşük olan bireylerin, internet erişimi ve kullanımından elde edilen bilgi kazanımının fayda sağladığı gözlemlenmiştir [5].

İnternet erişimiyle ilgili faktörleri araştırmak için geleneksel istatistiksel yaklaşımlar yaygın olarak kullanılmaktadır. Bu yaklaşımlar yorumlanabilir olsa da, genellikle karmaşık, doğrusal olmayan yapıları ve yüksek boyutlu veri kümelerini ele almada sınırlıdır. Buna karşılık makine öğrenmesi, bu tür karmaşıklıkları modellemek ve verilerdeki karmaşık ilişkileri ortaya çıkarmak için güçlü araçlar sunar. Karmaşıklık dışında farklı alanlarda da aktif olarak kullanılmaktadır [6,7]. Ancak makine öğrenmesi araçları, kara kutu gibi çalıştıkları için yüksek doğruluk sağlasa da yapılan tahminlerin neden yapıldığı konusunda belirsizlik öne çıkmaktadır. Bu belirsizlik özellikle yorumlanabilirliğin önemli olduğu konular için bir engeldir.

Yorumlanabilirlik ve tahmin gücü arasındaki bu boşluğu kapatmak için AYZ alanı ortaya çıkmıştır. AYZ, makine öğrenmesi modellerini performanslarından ödün vermeden belirsizlikleri kaldırmak için farklı yöntem ve araçlar sunmaktadır. Bu sayede kullanıcılar için birçok şey daha anlaşılır hale gelecektir. Finansal risk yönetimi için yapay zekâ, doğal dil işleme ve AYZ birleştirilmiştir [8]. Siber zorbalık içeren yorumları belirlemek için önerilen modelin yorumlanabilmesi için AYZ kullanılmıştır [9]. Parkinson hastalarındaki bilişsel gerilemeye etki eden bilişsel ve bilişsel olmayan değişkenler tahmin edilmiştir [10]. Bilişsel gerilemeyi en çok etki eden değişkenleri anlamak için AYZ kullanılmıştır. Perioperatif majör istenmeyen kardiyovasküler olayları tahmin etmek için yorumlanabilir bir makine öğrenmesi modeli geliştirilmiştir [11]. Her bir değişkenin modelin tahminlerine katkısını ölçmek için AYZ kullanılmıştır. Eski konutlarda enerji kullanımını tahmin etmek için enerji simülasyonu ve makine öğrenmesini entegre eden yeni bir yaklaşım önerilmiştir [12]. Enerji tüketimindeki en etkili değişkenleri belirlemek için AYZ kullanılmıştır. İran genelinde, yüzey toprak neminin mekânsal etkenlerini ortaya çıkarmak için AYZ kullanmıştır [13]. Lityum iyon pillerinin ömürlerini uzatmak için doğru şarj durumunu tahmin etmek için içerisinde AYZ'nin de bulunduğu makine öğrenmesine dayalı yeni bir yöntem önerilmiştir [14]. Bu araçlar arasında SHAP ve LIME, model tahminlerini, hem model genelinde hem de yerel düzeyde açıklayabilmesiyle oldukça yaygın kullanılmaktadır. SHAP değerleri, işbirlikçi oyun teorisine dayalı olarak her bir özelliğe katkı puanları sunarken, LIME, belirli gözlemler etrafında orijinal modelin karar sınırını, yaklaşık olarak belirlemek için yerel vekil modeller oluşturmaktadır.

Mevcut çalışmalar dijital uçurumu çoğunlukla belirli alt gruplar üzerinden ele almaktadır. İnternet erişimini ya dolaylı bir değişken olarak kullanmakta ya da dijital beceri, çevrimiçi öğrenme talebi veya çevresel bilgi düzeyi gibi farklı sonuç değişkenlerine odaklanmaktadır. Buna karşın Türkiye özelinde, resmi büyük ölçekli HBTKA verilerini kullanarak hane internet erişiminin doğrudan bağımlı değişken olarak

sınıflandırıldığı, birden fazla makine öğrenmesi algoritması ile analiz edilerek karşılaştırıldığı ve model kararlarının SHAP ve LIME gibi AYZ araçları ile ayrıntılı olarak açıklandığı bir çalışmaya rastlanmamaktadır. Literatürdeki bu eksiklik, hem dijital uçurumun ülke düzeyinde kanıta dayalı biçimde ortaya konmasını hem de politika yapıcılar için hangi parametrelerin kritik öneme sahip olduğunun şeffaf bir şekilde belirlenmesini zorlaştırmaktadır.

Bu çalışmada, söz konusu boşluğu doldurmak için, büyük veri olarak değerlendirilen, TÜİK tarafından toplanan HBTKA veri seti üzerinden hane halkı düzeyinde internet erişiminin temel belirleyicilerini bulmak ve bu belirleyicileri şeffaf bir şekilde açıklamak için makine öğrenmesi ve AYZ kullanılmıştır. Tahminleme yapmak için Rastgele Orman, XGBoost, LightGBM, CatBoost ve Lojistik Regresyon, açıklanabilirliği için SHAP ve LIME kullanılmıştır.

2. YÖNTEM

Bu bölümde, hanehalkı internet erişimini analiz etmek için kullanılan veri seti, makine öğrenimi yöntemleri ve modelden bağımsız açıklanabilirlik için kullanılan yöntemler özetlenmiştir.

2.1. Veri Seti ve Ön İşleme

Bu çalışmada kullanılan veri seti 519207 talep numarasına istinaden TÜİK tarafından sağlanmıştır [15]. Veri seti, 12159 haneye ait, teknolojik cihazların sahipliği, hane halkı gelir durumları, internet erişim, bölgesel tanımlayıcılar ve hedef değişken olan internet erişimi gibi parametreleri içermektedir. Çizelge 1'de veri kümesinde bulunan parametreler ve açıklamaları gösterilmiştir.

Çizelge 1. HBTKA veri setinin parametreleri ve parametre açıklamaları

Parametreler	Açıklamalar
BT_BILG_MASAUSTU	Hanede masaüstü bilgisayar var mı?
BT_BILG_DIZUSTU	Hanede dizüstü bilgisayar var m?
BT_TABLET	Hanede tablet var mı?
BT_TEL_CEP	Hanede cep telefonu var mı?
BT_AKILLITV	Hanede akıllı televizyon var mı?
BT_AKILLISAAT	Hanede akıllı saat var mı?
BT_OYUNKONSOL	Hanede oyun konsolu var mı?
BT_DIGER_CIHAZ	Hanede diğer internet kullanılan cihazlar var mı?
HANE_AYLIK_GELİR_GRP_5	Hanenin aylık toplam net geliri nedir?
HHB	Hanede bulunan kişi sayısı kaçtır?
IBBS_1	İstatistiki bölge birimlerinin sınıflandırılması
HANE_FAKTOR	Ağırlık katsayısı
HANE_INTERNET_ERISIM_DURUMU	Hanenin internet erişim durumu nedir?
REFERANS_YIL	Anketin uygulandığı yıl
BULTEN_NO	Rastgele üretilmiş hane ve fert soru formuna ilişkin anahtar

Veri seti üzerinden yapılacak eğitimlerden önce veri setine ön bir işlem uygulanmıştır. Başlangıçta hedef değişken olan "HANE_INTERNET_ERISIM_DURUMU" parametresinin değerleri incelendiğinde, "erişim var", "erişim yok" ve "bilinmiyor" olmak üzere üç farklı kategori içerdiği belirlenmiştir. Bu değişkende yalnızca 9 gözlem "bilinmiyor" kategorisinde yer aldığından, ikili sınıflandırma yapısını korumak ve belirsiz etiketli örneklerin modele dâhil edilmesini önlemek amacıyla söz konusu gözlemler veri setinden çıkarılmıştır. Böylece analizler toplam 12150 gözlem üzerinden yürütülmüştür. Ayrıca, modelleme sürecine doğrudan katkı sağlamayan ve yalnızca zaman ile anahtar bilgisi içeren "REFERANS_YIL" ve "BULTEN_NO" değişkenleri veri setinden çıkarılmıştır. Hanede masaüstü bilgisayar, dizüstü bilgisayar, tablet, cep telefonu, akıllı televizyon, akıllı saat, oyun konsolu ve diğer dijital cihazların varlığını gösteren "BT_BILG_MASAUSTU", "BT_BILG_DIZUSTU", "BT_TABLET", "BT_TEL_CEP", "BT_AKILLITV", "BT_AKILLISAAT", "BT_OYUNKONSOL" ve "BT_DIGER_CIHAZ" değişkenleri, ham veri setinde "1 = var" ve "2 = yok" biçiminde kodlanmışken, bu çalışma kapsamında ikili gösterime dönüştürülerek "1 = var" ve "0 = yok" şeklinde yeniden kodlanmıştır. İstatistiki Bölge Birimleri Sınıflandırmasını temsil eden "IBBS_1" değişkeni ise makine öğrenmesi modellerine uygunluk sağlamak amacıyla tekil gösterge (one-hot encoding) yöntemi kullanılarak yeni

sütunlara ayrılmış; ilgili hanenin bulunduğu bölge için değer "1", diğer bölgeler için "0" olarak atanmıştır. Bu ön işlemler sonucunda analize hazır nihai veri kümesi oluşturulmuş ve veri seti, modelleme aşamasında rastgele olarak %80 eğitim ve %20 test alt kümelerine ayrılmıştır.

2.2. Makine Öğrenmesi Modelleri

İnternet erişim durumunu modellemek için hem açıklanabilir modellerle uyumlu olması hem de güçlü modeller olması nedeniyle Rastgele Orman, XGBoost, LightGBM, CatBoost ve Lojistik Regresyon modelleri seçilmiştir.

2.2.1. Rastgele Orman

Rastgele orman, hem sınıflandırıcı hem de regresyon için kullanılan bir topluluk öğrenme yöntemidir [16,17]. Önyükleme ve paketleme olarak isimlendirilen bir işlemle birlikle birden fazla karar ağacı oluşturarak çalışır. Karar ağaçları, rastgele örnekleme yoluyla oluşturulur ve ağaçtaki her düğümde, rastgele seçilen özellik alt kümleri tekrar bölünür. Bu rastgelelik, ağaçlar arasındaki çeşitliliği artırırken aralarındaki ilişkiyi azaltır. Bu sayede ormanın genel performansı artmış olur. Rastgele Orman için oluşturulan karar ağaçları topluluğunun matematiksel modeli Eşitlik (1)'deki gibi ifade edilebilir.

$$T_1(x), T_1(x), \dots, T_B(x) \tag{1}$$

Regresyon uygulamasında ise tek tek ağaçların tahminlerinin ortalamasına göre belirlenir ve Eşitlik (2)'de gösterilmiştir:

$$\hat{\gamma} = \frac{1}{B}\sum_{i=1}^{B} T_i(x) \tag{2}$$

Burada $\hat{\gamma}$, bir gözlem vektörü x için Rastgele Orman modelinin nihai tahminini göstermektedir. $T_i(x)$ değeri, x girdisi için i'inci ağacın tahminini gösterir. B, ormandaki toplam ağaç sayısıdır. Sınıflandırmanın gerekli olduğu durumlarda, sınıf tahmini ağaçlar arasındaki çoğunluk oya göre belirlenir. Rastgele Ormanın temel avantajlarından bir tanesi de yüksek boyutlu verileri ve büyük veri kümelerini etkili bir şekilde işleyebilmesidir. Esnekliği sayesinde hiperparametre ayarlarına karşı yüksek toleransa sahip olması, kullanım kolaylığı ve fazla ayar gerektirmemesi gibi nedenlerden dolayı birçok makine öğrenimi probleminde oldukça fazla kullanılmaktadır.

2.2.2. XGBoost

XGBoost, Gradient Boosting Machine (GBM) algoritmasının farklı düzenlemeler ile optimize edilerek oluşturulan güçlü bir yöntem olarak ortaya çıkmıştır. Yüksek tahmin gücü, gereksiz verileri düzeltebilmesi, hızlı işlem yapabilmesi gibi yeteneklere sahip olması nedeniyle diğer yöntemlerden ayrışabilmektedir. İlk kez kullanıldığı çalışmada döneminin popüler yöntemlerinden 10 kat daha hızlı çalıştığı gözlemlenmiştir [18]. XGBoost matematiksel olarak modellendiğinde Eşitlik (3)'deki gibidir.

$$f_1(x), f_2(x), \dots, f_k(x) \tag{3}$$

Burada $f_k(x)$, x girdisi için k'ıncı karar ağacının çıktısıdır ve her f_k bir karar ağacıdır.

XGBoost modelinin toplam tahmini Eşitlik (4):

$$\hat{\gamma}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \epsilon\ F \tag{4}$$

Burada $\hat{\gamma}_i$, x_i özellik vektörüne sahip i. gözlem için modelin toplam tahminini göstermektedir. $f_k(x)$, x girdisi için k'ıncı karar ağacının çıktısıdır ve her f_k bir karar ağacıdır. K, modelde yer alan toplam ağaç sayısını gösterirken F ise tüm karar ağaçlarının ait olduğu karar ağaçları kümesini ifade eder.
XGBoost kayıp fonksiyonu Eşitlik (5)'deki gibidir.

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{\gamma}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{5}$$

Burada $l(y_i, \hat{y}_i)$: Gerçek değer y_i ile model tahmini $\hat{y}_i$ arasındaki kayıp fonksiyonudur. $\Omega(f_k)$ ise k'ıncı ağacın karmaşıklığını ve düzenliliğini ölçen ceza fonksiyonudur. n' toplam gözlem sayısını ifade eder. Yeni ağaç ekleme Eşitlik (6)'da verilmiştir.

$$\hat{y}_i^{\,t} = \hat{y}_i^{\,t-1} + f_t(x_i) \tag{6}$$

Burada her adımda güncellenen model verilmektedir. $\hat{y}_i^{\,t}$, t. iterasyon sonunda, x_i gözlemi için modelin güncel tahmini gösterirken, $\hat{y}_i^{\,t-1}$, bir önceki iterasyondaki tahmini ifade eder. f_t, t. iterasyonda eklenen yeni karar ağacını ve $f_t(x_i)$ bu ağacın x_i üzerindeki katkısını göstermektedir.

Büyük veri setlerinde çalışmak için optimize edilen XGBoost, Rastgele Orman'dan farklı olarak hem sapmayı hem de varyansı en aza indirmek için ağaçları sıralı olarak oluşturur. Optimizasyon içinse ikinci dereceden Taylor yaklaşımını kullanır. Aşırı öğrenmeyi engellemek ve modelin performansını artırmak için ağaç budama, sütun örnekleme ve L1/L2 düzenlemesi gibi teknikleri kullanmaktadır. XGBoost, bu çalışmada kullanıldığı gibi yapılandırılmıştır ve tablo veri setleri için oldukça uygundur.

2.2.3. LightGBM

Microsoft tarafından geliştirilen LightGBM, makine öğrenmesi teknikleri ile gösterdiği başarıları sayesinde veri bilimi ve uzaktan algılama araştırmacılarının dikkatini çekmiştir [19,20]. LightGBM, eğitimleri hızlandırmak ve bellek kullanımlarını azaltmak için histogram tabanlı bir gradyan artırma çerçevesidir. Karar ağaçlarında ki seviyesel büyümenin aksine, ağaçları yaprak bazlı büyüterek daha düşük kayıp ve daha hızlı yakınsama sağlamaktadır. Yaprak tabanlı büyüme Eşitlik (7)'de verilmiştir.

$$Gain = \frac{1}{2}\left(\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right) - \gamma \tag{7}$$

Burada, G_L ve G_R sol ve sağ dalların toplam gradientini temsil ederken H_L ve H_R toplam hessianı temsil etmektedir. λ, L2 regularizasyonu γ, ceza parametresidir.

Büyük veri setlerini işleme verimliliğini artıran gradyan tabanlı tek taraflı örnekleme ve ayrıcalıklı öznitelik destekleme gibi yenilikler sunmaktadır. Gradyan tabanlı tek taraflı örnekleme sayesinde verinin tamamını ele almadan veriden ürettiği alt veri kümesini kullanır. Ayrıcalıklı öznitelik destekleme ile seyrek öznitelikleri yoğun özniteliklere dönüştürerek işlem karmaşıklığını azaltır [19-21].

2.2.4. CatBoost

CatBoost, kategorik özellikleri işlemek için özel olarak tasarlanmış bir gradyan güçlendirme algoritmasıdır. Doğruluğu ve hesaplama verimliliğini artıran sıralı güçlendirme ve habersiz karar ağaçlarını kullanır. Ardışık bir şekilde ağaçları birleştirerek, bir önceki ağacın hatalarını gidererek modelin hata oranını azaltır ve performans artışına neden olmaktadır [22]. Kategorik özelliklerin işlenmesi Eşitlik (8)'deki gibidir.

$$Encoded(x_{i,j}) = \frac{\sum_{k<i}[x_{k,j} = x_{i,j}] \,.\, y_k + a \,.\, p}{\sum_{k<i}[x_{k,j} = x_{i,j}] + a} \tag{8}$$

Burada, $x_{i,j}$ için i. gözlemdeki j. kategorik özelliği temsil ederken y_k hedef değeri temsil etmektedir. $a,\ p$ ise sıfır bölme ve yumuşatma için kullanılan sabitlerdir.

Model, kategorik değişkenleri kodlamak için permütasyon odaklı bir şema kullanarak sızıntıları azaltır. Bu sayede değişkenlerin yoğunlukta olduğu veri setlerinde avantajlıdır. Ek olarak aşırı öğrenme sorunlarına karşı güçlü olması ve daha az hiperparametre ayarı gerektirmesi sebebi ile kullanımı kolaydır [23].

2.2.5. Lojistik Regresyon

Lojistik Regresyon, özellikle sınıflandırma problemleri için yorumlanabilirliğinin güçlü olması nedeniyle temel modellerden bir tanesidir. Girdi özelliklerinin doğrusal kombinasyonuna lojistik sigmoid fonksiyonu uygulayarak sınıf üyeliğinin olasılığını modeller. Olasılık tahmini Eşitlik (9) ve Eşitlik (10)'daki gibi ifade edilebilir.

$$P(y = 1|x) = \sigma(z) = \frac{1}{1+e^{-z}} \quad (9)$$

$$z = B_0 + B_1x_1 + B_2x_2 + \cdots + B_px_p \quad (10)$$

Burada, $P(y = 1|x)$ için x özellikleri verildiğinde $y = 1$ olasılığı ifade edilirken, $\sigma(z)$ sigmoid fonksiyonu, B'ler modelin katsayıları, x'ler bağımsız değişkenler olarak temsil edilir.

Basit bir yapıda olmasına rağmen, özellikler ve hedefin olasılığı arasındaki ilişki doğrusalsa oldukça iyi performans gösterir. Lineer ilişkilere dayalı tahminleri genellikle hızlı yapması ve yorumlanabilir sonuçlar üretmesine rağmen karmaşık verilerde doğruluğu sınırlı kalabilmektedir [24].

2.3. Değerlendirme Metrikleri

Çalışma kapsamında kullanılan makine öğrenmesi modellerinin performansını değerlendirmek için, çeşitli istatistiksel metrikler kullanılmıştır. Bu metrikler, modelin ne kadar doğru ve tutarlı bir şekilde çalıştığını anlamak için kritik öneme sahiptir. Çalışmada kullanılan performans metrikleri kesinlik (precision), duyarlılık (recall), F1-skoru, doğruluk (accuracy) gibi temel ölçütlerdir.

2.3.1. Keskinlik

Keskinlik, modelin pozitif olarak tahmin ettiği durumların, gerçekten de pozitif olanların oranını ölçer. Yüksek bir keskinlik değeri, modelin yanlış pozitif sayısının düşük olduğunu gösterir, yani model, olmayan bir şeyi varmış gibi işaret eder. Eşitlik (11)'de keskinlik hesabı gösterilmiştir.

$$Keskinlik = \frac{TP}{TP+FP} \quad (11)$$

TP, doğru pozitifi yani tahmin edilen değerin doğru olduğunu temsil ederken, FP, yanlış pozitifi temsil eder. Yanlış pozitif ise yanlış tahmin değeridir. Keskinlik metriği, modelin ne kadar "hassas" olduğunu yansıtır. Özellikle, yasaklı faaliyetlerin tespiti gibi kritik uygulamalarda yüksek keskinlik, önemsiz hatalardan kaçınmak için önemlidir.

2.3.2. Duyarlılık

Duyarlılık, gerçekte pozitif olan örneklerin ne kadarının model tarafından doğru olarak pozitif olarak işaretlendiğini ölçer. Yüksek bir duyarlılık değeri, modelin gerçek pozitifleri kaçırmadığını, yani az sayıda yanlış negatif hata yaptığını gösterir. Eşitlik (12)'de duyarlılık hesabı gösterilmiştir.

$$Duyarlılık = \frac{TP}{TP+FN} \quad (12)$$

2'de TP, doğru pozitifi belirtirken FN, yanlış negatifi gösterir.

2.3.3. F1-Skoru

F1-skoru, keskinlik ve duyarlılığın harmonik ortalamasıdır ve iki metriği dengeli bir şekilde birleştirir. Bu metrik, özellikle veri seti dengesiz olduğunda yani bir sınıf diğerlerinden çok daha az ya da çok daha fazla örneğe sahip olduğunda modelin genel performansını değerlendirmek için kullanılır. Eşitlik (13)'te F1-skor hesabı gösterilmiştir.

$$F_1 - skor = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall} \quad (13)$$

F1-skoru, her iki metriği eşit ağırlıkta hesaba katarak, modelin genel doğruluk ve güvenilirlik dengesini sunar.

2.3.4. Doğruluk

Doğruluk, modelin doğru olarak sınıflandırdığı örneklerin, toplam örnek sayısına oranıdır ve Eşitlik (14)'te gösterilmiştir.

$$Doğruluk = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

Modelin genel başarısını ölçen bir metriktir.

2.4. Açıklanabilir Yapay Zekâ Teknikleri

İnternet erişim durumunun analizi için kullanılan makine öğrenmesi yöntemleri güçlü tahminler yapsa da kara kutu gibi davranmasından dolayı yorumlanabilirliği zordur. Bu nedenle, bu çalışmada internet erişim durumunu etkileyen faktörlerin tahmin gerekçelerini anlayabilmek için SHAP ve LIME AYZ araçları kullanılmıştır.

2.4.1. SHapley Additive exPlanations

SHAP, her bir özelliğe, tüm olası özellik alt kümelerindeki katkıya bağlı olarak bir önem değeri atayarak tahminleri yorumlayan ve işbirlikçi oyun teorisine dayanan bir AYZ yöntemidir. Bu yöntem, doğruluk, tutarlılık, esneklik ve eksiklik gibi temel teorik özellikleri karşılamaktadır. Bu sayede, modelin hangi değişkene ne kadar dikkat ettiği bilinebilmekte ve tahmin süreçleri daha anlamlı ve daha açıklanabilir hale gelmektedir [25]. Ayrıca yerel ve küresel açıklamalarda kullanılabilmektedir.

2.4.2. Local Interpretable Model-Agnostic Explanations

LIME, literatürdeki birçok makine öğrenmesi yöntemleri ile çalışabilen ve model tahminlerini yorumlayarak modelin açıklanabilmesini sağlayan görsel bir tekniktir [26]. Temel olarak, açıklanmak istenen örneğin etrafında rastgele örnekler üretir ve bu örnekleri orijinal modele sunarak tahminler alır. Daha sonra bu yerel veri üzerinde basit ve yorumlanabilir açıklamalar üretir. Bu sayede, modelin kararında, hangi özelliklerin ne ölçüde etkili olduğu anlaşılır bir hale gelir. Önemli avantajlarından bir tanesi modelden bağımsız olarak çalışabilmesidir. Hem doğrusal hem de doğrusal olmayan sınıflandırma veya regresyon fark etmeksizin tüm modeller için uygulanabilir.

3. ARAŞTIRMA BULGULARI

Bu bölümde, çalışma kapsamında kullanılan Rastgele Orman, LightGBM, Lojistik Regresyon, CatBoost ve XGBoost makine öğrenmesi yöntemleri ile sınıflandırma performansları, SHAP ve LIME ile açıklanabilirlik analizleri ve örnek veri noktaları üzerindeki çıkarımları detaylı bir şekilde sunulmuştur. Bulgular, yalnızca model tabanlı verilmemiş, görsel analizlerle desteklenmiştir.

3.1. Sınıflandırma Modellerinin Performans Karşılaştırılması

Çalışma kapsamında kullanılan Rastgele Orman, LightGBM, Lojistik Regresyon, CatBoost ve XGBoost sınıflandırma modelleri, veri setinin %80 eğitim %20 test bölünmesi ile değerlendirilmiştir. Her model için doğruluk, f1-skoru, keskinlik ve duyarlılık performans metrikleri elde edilmiştir. Çizelge 2'de bu sonuçlar gösterilmiştir.

Çizelge 2. Makine öğrenmesi sınıflandırma modellerinin performans karşılaştırması

Model	Sınıf	Keskinlik	Duyarlılık	F1-Skor	Doğruluk	ROC-AUC
Rastgele Orman	1	0.99	0.82	0.90	0.82	0.936 ± 0.009
	0	0.21	0.91	0.34		
XGBoost	1	0.99	0.82	0.90	0.82	0.936 ± 0.011
	0	0.20	0.89	0.33		
LightGBM	1	0.99	0.81	0.89	0.82	0.932 ± 0.010
	0	0.20	0.91	0.33		
CatBoost	1	0.99	0.81	0.89	0.82	0.934 ± 0.010
	0	0.20	0.91	0.33		
Lojistik Regresyon	1	0.99	0.80	0.89	0.81	0.936 ± 0.011
	0	0.19	0.80	0.89		

Her model için her iki sınıf içinde, Sınıf 1 internet erişimi olan haneleri Sınıf 0 internet erişimi olmayan haneleri, doğruluk, f1-skoru, keskinlik ve duyarlılık değerleri ayrı ayrı verilmiştir. Tüm modellerin genel doğruluk oranları %81-82 arasındadır. Sınıf 1 için keskinlik ve duyarlılık değerleri yüksek olsa da Sınıf 0 için keskinlik değeri düşük olduğu gözlemlenmiştir. Bu durum, Sınıf 0'a ait gözlemler üzerine yapılan tahminlerde modellerin yeterince hassas olmadığına işaret etmektedir. Ancak, Sınıf 0 gözlemlerinin çoğu doğru şekilde sınıflandırılabilmiştir. Bu düşük keskinlik, veri setindeki dengesizliği ve internet erişimi olan hanelerin sayısının fazla olmasından kaynaklanmaktadır. Sınıf dengesizliklerinin modellenmesi için gelecekte SMOTE (Synthetic Minority Over-sampling Technique) gibi veri artırma yöntemlerinin veya sınıf ağırlıklarının yeniden düzenlenmesi gibi stratejilerin kullanılması faydalı olabilir. Çizelge 2 genel olarak incelendiğinde, modellerin performans metrik değerlerinin birbirine yakın olduğu görülmüştür. Bu nedenle aşırı öğrenme riski ayrıca incelenmiştir. Aşırı öğrenme eğilimi, 5 katlı katmanlı çapraz doğrulama ile eğitim ve test kümeleri üzerindeki performans farkları incelenerek kontrol edilmiştir. Bu kapsamda, özellikle dengesiz sınıflar için ayırt ediciliği ile öne çıkan ROC (Receiver Operating Characteristic) – AUC (Area Under the Curve) değerleri hesaplanmıştır. Hesaplanan ROC–AUC değerlerinin tamamı 0.93'ün üzerinde olduğu görülmüştür. Rastgele Orman, Lojistik Regresyon ve XGBoost 0.936 ile en yüksek ayırt ediciliğe ulaşırken, CatBoost ve LightGBM yalnızca çok küçük bir farkla geride kalmıştır. Ek olarak, PR (Precision–Recall) – AUC, F1-macro (Macro-averaged F1 Score) ve Dengeli Doğruluk (Balanced Accuracy) metrikleri de eğitim ve test kümeleri üzerinde değerlendirilmiştir. Elde edilen sonuçlar incelendiğinde, bu metriklerdeki farkların %1–%5 aralığında olduğu görülmüştür. Sonuçlar belirgin bir aşırı öğrenme olmadığını göstermiştir. Sonuçların birbirine yakın çıkma sebebi olarak veri setinde, internet erişimi olan hanelerin olmayan hanelere göre sayısının fazla olması öngörülmektedir.

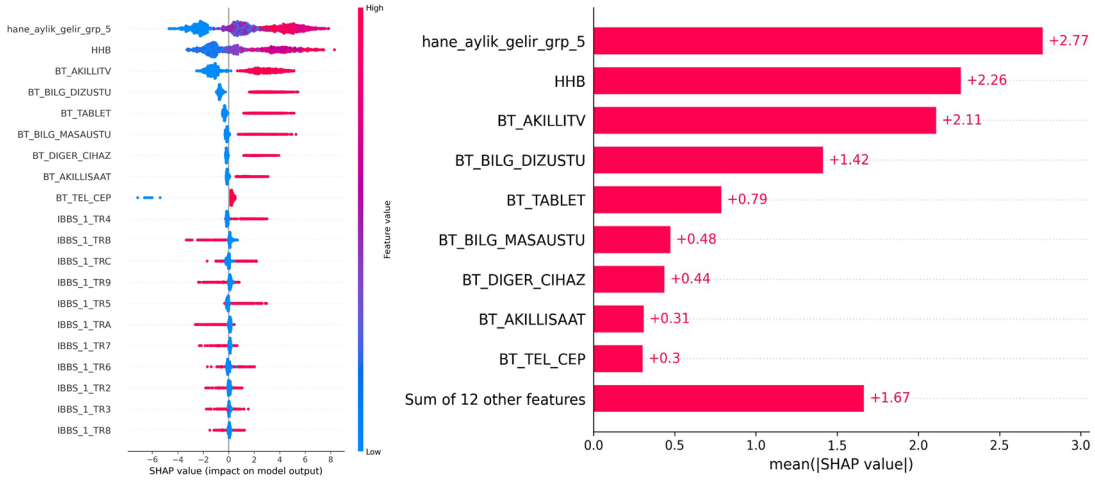
3.2. SHAP Tabanlı Özellik Etki Analizi

Çalışma kapsamında kullanılan Rastgele Orman, XGBoost, LightGBM, CatBoost ve Lojistik Regresyon gibi makine öğrenmesi modellerinin karar verme mekanizmalarını açıklayabilmek amacıyla bu çalışmada AYZ tekniklerinden biri olan SHAP yöntemi kullanılmıştır. SHAP ile kara kutu olarak görülen makine öğrenmesi modellerinin tahminlerinde hangi bağımsız değişkenlerin ne kadar etkili olduğu belirlenmiştir. Çalışmada kullanılan Rastgele Orman, XGBoost, LightGBM, CatBoost ve Lojistik Regresyon olmak üzere beş farklı makine öğrenmesi modeli için SHAP temelli özellik etki analizi gerçekleştirilmiştir. Her bir model için en etkili 10 değişken belirlenmiş ve Çizelge 3'te karşılaştırmalı olarak sunulmuştur.

Çizelge 3. Model temelli ilk 10 önemli değişkenin SHAP değerleri

Model	İlk 10 Değişken: SHAP değeri	
CatBoost	1. HANE_AYLIK_GELİR_GRP_5: **2.77** 2. HHB: **2.26** 3. BT_AKILLITV: **2.11** 4. BT_BILG_DIZUSTU: **1.42** 5. BT_TABLET: **0.79**	6. BT_BILG_MASAUSTU: **0.48** 7. BT_DIGER_CIHAZ: **0.44** 8. BT_AKILLISAAT: **0.31** 9. BT_TEL_CEP: **0.30** 10. IBBS_1_TR4: **0.23**
LightGBM	1. HANE_AYLIK_GELİR_GRP_5: **2,46** 2. HHB: **2,06** 3. BT_AKILLITV: **1,71** 4. BT_BILG_DIZUSTU: **1,47** 5. BT_TABLET: **0,73**	6. BT_BILG_MASAUSTU: **0,46** 7. BT_DIGER_CIHAZ: **0,42** 8. IBBS_1_TR5: **0,32** 9. BT_AKILLISAAT: **0,29** 10. IBBS_1_TR4: **0,26**
Lojistik Reg.	1. BT_BILG_DIZUSTU: **1,85** 2. BT_TABLET: **1,34** 3. BT_AKILLITV: **0,98** 4. BT_BILG_MASAUSTU: **0,96** 5. HHB: **0,88**	6. HANE_AYLIK_GELİR_GRP_5: **0,87** 7. BT_DIGER_CIHAZ: **0,79** 8. BT_AKILLISAAT: **0,63** 9. IBBS_1_TR2: **0,26** 10. IBBS_1_TR7: **0,24**
Rastgele Orman	1. HANE_AYLIK_GELİR_GRP_5: **0,13** 2. BT_AKILLITV: **0,12** 3. HHB: **0,11** 4. BT_BILG_DIZUSTU: **0,07** 5. BT_TABLET: **0,04**	6. BT_BILG_MASAUSTU: **0,02** 7. BT_AKILLISAAT: **0,02** 8. BT_DIGER_CIHAZ: **0,02** 9. BT_TEL_CEP: **0,01** 10. IBBS_1_TR5: **0,01**
XGBoost	1. HANE_AYLIK_GELİR_GRP_5: **1,49** 2. BT_AKILLITV: **1,01** 3. HHB: **0,99** 4. BT_BILG_DIZUSTU: **0,8** 5. BT_TABLET: **0,36**	6. IBBS_1_TR9: **0,24** 7. IBBS_1_TRB: **0,24** 8. BT_BILG_MASAUSTU: **0,21** 9. IBBS_1_TRA: **0,18** 10. IBBS_1_TR7: **0,17**

Çizelge 3 incelendiğinde, farklı modeller olmalarına rağmen belirleyici değişkenlerin büyük oranda örtüştüğü görülmektedir. Özellikle "HANE_AYLIK_GELİR_GRP_5" değişkeninin, Lojistik Regresyon modelinin dışında kalan tüm modellerde SHAP değeri en yüksek olduğu görülmüş ve hanehalkı internet erişim durumunu belirleyen en etkili faktör olduğu belirlenmiştir. Bu değişken, beş gelir grubunu temsil etmektedir. 1. ve 2. gelir grupları daha düşük, 4. ve 5. gelir grupları ise daha yüksek aylık gelir düzeylerini ifade etmektedir. Lojistik Regresyon modelinde ise en önemli değişkenin "BT_BILG_DIZUSTU" olarak öne çıktığı görülmüştür. Bu durumun, Lojistik Regresyon'un doğrusal ve parametre-temelli yapısı nedeniyle, hedef değişkenle güçlü ve doğrudan doğrusal ilişki gösteren değişkenlere daha yüksek ağırlık atamasından kaynaklandığı düşünülmektedir. Ayrıca, hane geliri ile dizüstü bilgisayar sahipliği arasındaki güçlü ilişki, gelir etkisinin bu modelde dolaylı olarak "BT_BILG_DIZUSTU" değişkeni üzerinden temsil edilmesine yol açmış olabilir. Bununla birlikte, "HANE_AYLIK_GELİR_GRP_5" değişkeni Lojistik Regresyon modelinde de üst sıralarda yer almakta ve gelir düzeyinin tüm modellerde tutarlı biçimde kritik bir belirleyici olduğunu göstermektedir. Bu durumu sırasıyla hane büyüklüğünü temsil eden "HHB" ve dijital cihaz sahipliği ile ilgili olan "BT_AKILLITV", "BT_BILG_DIZUSTU", "BT_TABLET", "BT_BILG_MASAUSTU" değişkenleri takip etmektedir. Tüm modellerde ekonomik ve teknolojik cihaz sahipliğinin ön planda olması, internet erişim durumunun bu parametrelerle güçlü bir ilişki içerisinde olduğunu göstermektedir. Ayrıca, "IBBS_1_TR4", "IBBS_1_TR5" gibi bölgesel değişkenlerin de ilk 10 arasında yer aldığı görülse de SHAP değerlerinin diğer değişkenlere göre nispeten düşük kaldığı görülmüştür. Örnek olarak Şekil 1'de CatBoost modeli için değişkenlerin etkileri verilmiştir.

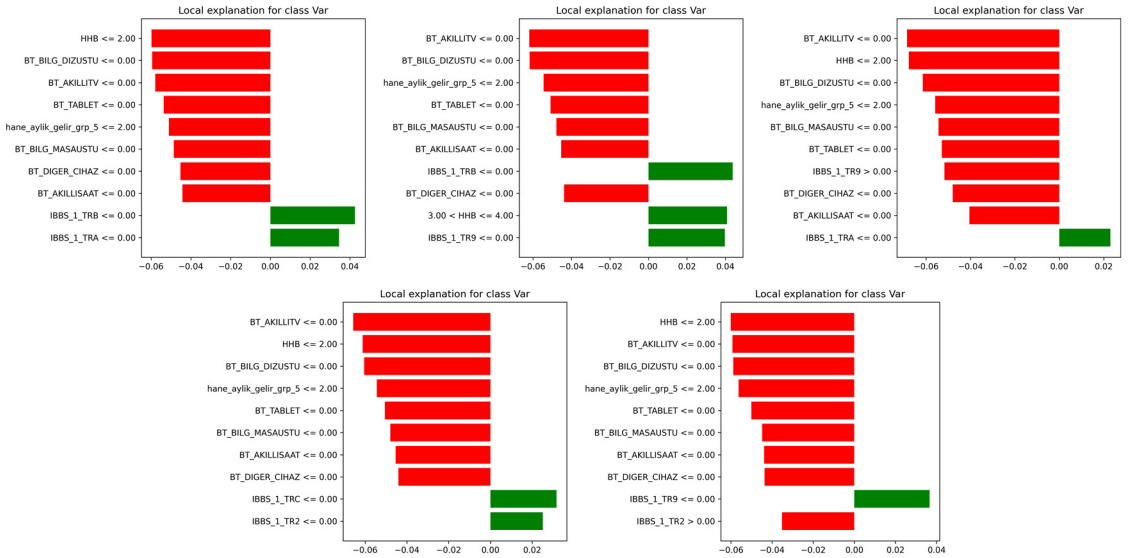


Şekil 1. CatBoost modeli için değişkenlerin SHAP grafikleri

Şekil 1 incelendiğinde hem değişkenlerin genel olarak model üzerindeki etkisi hem de her bir gözlem için bu etkilerin dağılımı ayrıntılı bir şekilde göstermektedir. Bar grafiği incelediğinde, CatBoost modeli için "HANE_AYLIK_GELİR_GRP_5" değişkeninin hane internet durumu tahmininde en yüksek etkiye sahip olduğu görülmüştür. Bunu sırasıyla "HHB" ve "BT_AKILLITV" gibi dijital cihaz sahipliklerine ait değişkenler takip etmektedir. Bu değişkenlere göre kalan değişkenliklerin etkileri daha azdır. Değişken etki özeti grafiği incelendiğinde sadece değişkenlerin SHAP değerlerini değil, aynı zamanda bu değişkenlerin gözlemler üzerindeki olumlu veya olumsuz olarak ne kadar değişkenlik gösterdiği görülmektedir. Özellikle "HANE_AYLIK_GELİR_GRP_5" parametresine bakıldığında, yüksek gelire sahip hanelerde SHAP değerlerinin pozitif olduğu ve modelin bu gözlemleri hanenin internet erişim olasılığını yüksek olarak tahmin ettiği görülmektedir. Benzer şekilde "HBB", "BT_AKILLITV", "BT_BILG_DIZUSTU" gibi parametrelere bakıldığında bu değişkenlerin model tahmininde olumlu katkı sağladığı görülmektedir. Her iki grafik birlikte değerlendirildiğinde, hanehalkı internet erişim durumunun, ekonomik, hanehalkı birey sayısı ve teknolojik aletlere sahipliğin merkez parametreler olduğu görülmüştür.

3.3. LIME ile Yerel Açıklama Analizi

Bu çalışmada kullanılan makine öğrenmesi modelleri ile elde edilen tahminlerin nasıl şekillendiğini değerlendirmek için kullanılan bir diğer AYZ tekniği olan LIME kullanılmıştır. LIME, her bir gözlem için modelin kararını hangi değişkenlerin nasıl etkilediğini görselleştirir. Bu kapsamda, CatBoost modeli için veri seti içinden rastgele seçilen beş farklı hane için yapılan LIME analizleri Şekil 2'de sunulmuştur.



Şekil 2. CatBoost modeli için seçilen beş farklı hanede LIME ile elde edilen yerel açıklama grafiği

Şekil 2 incelendiğinde kırmızı çubukların ilgili gözlemde değişkenin aldığı değerlerin modelin hanehalkı internet erişim durumunu tahminine olumsuz yönde katkı sağladığını, yeşil çubukların ise olumlu yönde katkı sağladığını göstermektedir. Örneğin "HHB <= 2.00" ifadesi kırmızı ile gösterilmiştir. Bu gözlem, hanehalkının 2 kişi veya daha az olmasının, modelin, hanehalkı internet erişim durum tahminini azalttığını belirtmektedir. Bu, insan sayısının az olduğu hanelerde internet erişim durumu olasılığının daha düşük olduğunu, 2 den fazla kişinin bulunduğu hanelerin modelde, tahmini artırıcı etkiye sahip olduğunu belirtir. Benzer şekilde, "HANE_AYLIK_GELİR_GRP_5 <= 2.00" şeklinde gösterilen kırmızı bar, düşük gelire sahip olmanın internet erişimine sahip olma olasılığını düşürdüğünü göstermektedir. "BT_AKILLITV <= 0.00", "BT_BILG_DIZUSTU <= 0.00" gibi değişkenlerin kırmızı olması, akıllı tv veya dizüstü bilgisayar sahibi olmayan hanelerde modelin, olumsuz tahmin ürettiğini, teknolojik cihaz sahipliğinin ise olumlu katkı sağladığını göstermektedir. Genel olarak incelendiğinde, hane büyüklüğü, dijital cihaz sahipliği ve hane aylık geliri değişkenlerinin model tahmininde baskın olduğu görülmüştür. Ek olarak, bazı gözlemlerde bölgesel değişkenlerin tahminler üzerinde anlamlı etkileri olduğu gözlemlenmiştir. Haneler için bu parametrelerin yada katkılarının farklı olması, modelin tahminlerinde bireysel koşulları dikkate aldığını ve esnek bir şekilde farklı değişkenlere ağırlık verdiğini göstermektedir.

4. SONUÇLAR

Bu çalışmada, beş farklı makine öğrenmesi yöntemi ve AYZ teknikleri ile hanehalkı bilişim teknolojileri kullanımında internet erişim durumunu etkileyen faktörler kapsamlı bir şekilde incelenmiştir. Elde edilen sonuçlar incelendiğinde, CatBoost, XGBoost, LightGBM, Rastgele Orman ve Lojistik Regresyon modelleri yüksek ve birbirine yakın sonuçlar vermiştir. Özellikle dengesiz sınıf yapısını daha sağlıklı değerlendirebilmek için hesaplanan ROC–AUC değerlerinin tamamının 0.93'ün üzerinde olması, modellerin internet erişimi olan ve olmayan haneleri ayırt etme konusunda güçlü bir sınıflandırma başarısına ulaştığını göstermektedir.

Açıklanabilirlik sonuçları incelendiğinde, hem SHAP ile küresel hem de LIME ile yerel düzeyde hanehalkı internet erişim durumunun en etkili belirleyicisinin, hane gelirini yansıtan "HANE_AYLIK_GELİR_GRP_5" olduğu görülmüştür. Bunu hane büyüklüğünü yansıtan "HHB" ve dijital cihaz sahipliğini temsil eden "BT_AKILLITV", "BT_BILG_DIZUSTU", "BT_TABLET", "BT_BILG_MASAUSTU" gibi değişkenler takip etmektedir. Bu sonuçlar, internet erişiminin yalnızca teknik altyapı meselesi değil, aynı zamanda ekonomik kapasite ve dijital donanım sahipliğiyle yakından ilişkili olduğunu ortaya koymaktadır. Bazı modellerde bölgesel değişkenlerin de ilk 10 değişken arasında yer alması, mekânsal farklılıkların tamamen önemsiz olmadığını, ancak gelir ve cihaz sahipliği gibi değişkenlerin gerisinde kaldığını göstermektedir.

Elde edilen bulgular, politika ve uygulama açısından da önemli çıkarımlara sahiptir. Düşük gelir gruplarındaki ve sınırlı dijital cihaz sahipliğine sahip hanelerde internet erişim olasılığının daha düşük olması, destek programlarının bu gruplara öncelik verecek şekilde tasarlanması gerektiğini göstermektedir. Hanehalkı internet erişim oranlarını artırmak için yalnızca altyapı yatırımları değil, aynı zamanda gelir düzeyi düşük hanelere yönelik hedeflenmiş cihaz destekleri ve dijital kapsayıcılığı artırmaya yönelik sosyal programlar planlanabilir. Bu bulgular belirli sınırlılıklar çerçevesinde değerlendirilmelidir. Çalışma yalnızca TÜİK tarafından yayımlanan 2024 yılı HBTKA verisine dayanmaktadır ve bu nedenle sonuçlar tek bir seneye ait durumu yansıtmaktadır. Hedef değişkenin dengesiz dağılımı nedeniyle internet erişimi olmayan haneler için keskinlik değerleri görece düşük kalmıştır. Bu çalışmada sınıf dengesizliğini gidermeye yönelik SMOTE gibi ileri dengeleme yöntemleri yalnızca öneri düzeyinde tartışılmış, ancak uygulanmamıştır. Gelecek çalışmalarda birden fazla yıla ait HBTKA verisinin kullanılmasıyla zaman içindeki değişimin incelenmesi, veri dengeleme tekniklerinin modele entegre edilmesi, dijital uçurumun daha ayrıntılı ve genellenebilir biçimde ortaya konulmasına katkı sağlayabilir. Çalışma, resmi HBTKA veriseti üzerinde gelişmiş makine öğrenmesi ve AYZ yöntemlerinin birlikte kullanılmasının, dijital uçurumun yapısal belirleyicilerini sayısal ve şeffaf bir biçimde ortaya koymak için güçlü ve tekrarlanabilir bir çerçeve sunduğunu da göstermektedir.

5. KAYNAKLAR

1. Lalmuanawma, S., Laltlankimi, H., Bolchhim, L., Lalmalsawma, R., Lalnunthari, Dawngliani, M.S. & Rosangliana, D. (2025). Exploring the digital divide: A statistics and machine learning analysis of internet penetration and IT literacy in the Indo-Myanmar region. *One Village, Collective Perspective: A Collaborative Research Endeavor*, 1-19.
2. Regmi, N. (2023). A random forest-based analysis of household survey data to infer insights on digital inequality. *Iran Journal of Computer Science, 6*(4), 333-344.
3. Park, J.R. & Feng, Y. (2023). Trajectory tracking of changes digital divide prediction factors in the elderly through machine learning. *PloS one, 18*(2), 1-20.
4. Mhlanga, D. & Dunga, H.M. (2023). Demand for internet services before and during the Covid-19 pandemic: what lessons are we learning in South Africa? *International Journal of Research in Business and Social Science, 12*(7), 626-640.
5. Cheng, X. (2023). Estimating heterogeneous effects of internet use on environmental knowledge: Taking population heterogeneity into consideration. *PloS one*, *18*(7), 1-23.
6. Alpsalaz, F., (2025). Detection of imbalance faults in industrial machines by means of frequency-based feature extraction using machine learning and deep learning approaches. *Cukurova University, Journal of the Faculty of Engineering, 40*(3), 581-592.
7. Tolun, Ö.C., Zor, K. ve Tutsoy, Ö. (2025). Akıllı şebekelerde elektrik hırsızlığı tespiti için gelişmiş makine öğrenmesi yöntemlerinin uygulanması. *Çukurova Üniversitesi, Mühendislik Fakültesi Dergisi, 40*(3), 627-641.
8. Hussain, N.Y., Babalola, F.I., Kokogho, E. & Odio, P.E. (2025). A robust model for integrating artificial intelligence into financial risk management: Addressing compliance, accuracy, and scalability issues. *International Journal of Research and Innovation in Social Science, 9*(2), 3651-3668.
9. Checkin, A., Pleshakova, E. & Gataullin, S. (2025). A hybrid KAN-BiLSTM transformer with multi-domain dynamic attention model for cybersecurity. *Technologies, 13*(6), 223.
10. Colautti, L., Casella, M., Robba, M., Marocco, D., Ponticorvo, M., Iannello, P., Antonietti, A., Marra, C. & for the CPP Integrated Parkinson's Database. (2025). Predicting cognitive decline in Parkinson's disease using artificial neural networks: An explainable AI approach. *Brain Sciences, 15*(8), 782.
11. Golubovic, M., Peric, V., Stosic, M., Stojiljkovic, V., Zivic, S., Kamenov, A., Milic, D., Dinic, V., Stojanovic, D. & Lazarevic, M. (2025). Predicting major adverse cardiovascular events after cardiac surgery using combined clinical, laboratory, and echocardiographic parameters: A machine learning approach. *Medicina, 61*(8), 1323.
12. Fan, C., Liu, R. & Liao, Y. (2025). Archetype identification and energy consumption prediction for old residential buildings based on multi-source datasets. *Buildings, 15*(14), 2573.
13. Nikraftar, Z., Parizi, E., Saber, M., Boueshagh, M., Tavakoli, M., Esmaeili Mahmoudabadi, A., Ekradi, M. H., Mbuvha, R. & Hosseini, S.M. (2025). An interpretable machine learning framework for unraveling the dynamics of surface soil moisture drivers. *Remote Sensing, 17*(14), 2505.

14. Hu, P., Li, C. Y. & Lee, C.C. (2025). SOC estimation of lithium-ion batteries utilizing EIS technology with SHAP–ASO–LightGBM. *Batteries, 11*(7), 272.
15. Türkiye İstatistik Kurumu, (2024). *Hanehalkı bilişim teknolojileri kullanım araştırması.* (Özel talep ile elde edilmiştir.)
16. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.
17. Liaw, A. & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18-22.
18. Chen, T. & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 785-794.
19. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T.Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *In Advances in Neural Information Processing Systems (NIPS 2017),* 3146-3154.
20. Ustuner, M. & Balik Sanli, F. (2019). Polarimetric target decompositions and light gradient boosting machine for crop classification: A comparative evaluation. *ISPRS International Journal of Geo-Information, 8*(2), 97.
21. Li, W., Ding, S., Chen, Y., Wang, H. & Yang, S. (2019). Transfer learning-based default prediction model for consumer credit in China. *The Journal of Supercomputing, 75*(2), 862-884.
22. Patrous, Z.S. (2018). Evaluating XGBoost for user classification by using behavioral features extracted from smartphone sensors. *Master's thesis* KTH Royal Institute of Technology, School of Computer Science and Communication, Sweden.
23. Ibrahim, A.A., Ridwan, L.R., Muhammed, M.M., Abdulaziz, R.O. & Saheed, G.A. (2020). Comparison of the catboost classifier with other machine learning methods. *International Journal of Advanced Computer Science and Applications, 11*(11), 738-748.
24. Salem A.M., Subhi Al-Batah, M., Alqaraleh, M., Abuashour, A. & Hamadah Bader, A.F. (2023). Early diagnosis of fiabetes: a comparison of machine learning methods. *International Journal of Online and Biomedical Engineering (iJOE), 19*(15), 144-165.
25. Sevimli Deniz, S. (2025). Shap analizi ile sosyal medya bağımlılığı dinamiklerinin açıklanması: temel etkenler ve önleyici stratejiler. *Firat University Journal of Social Sciences, 35*(2), 643-657.
26. Ribeiro, M.T., Singh, S. & Guestrin, C. (2016). Why should I trust you? *Explaining the predictions of any classifier Marco. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016),* 1135-1144.