Araştırma Makalesi

TÜİK Mikro Verileri ile Çocuklarda Cep Telefonu Sahipliğinin Tahmini: Makine Öğrenimi Modellerinin Karşılaştırmalı Performansı

Kâmil Abdullah EŞİDİR[1]

[1]Dr., Fırat Kalkınma Ajansı, Elâzığ Yatırım Destek Ofisi, abdullahesidir@yahoo.com, ORCID: 0000-0002-8106-1758

Öz: Bu çalışmada, 2022 yılı Türkiye Çocuk Araştırması mikro veri seti kullanılarak çocukların cep telefonu sahipliğini tahmin edilmiş ve çocuklarda cep telefonu sahipliğini etkileyen faktörler analiz edilmiştir. Tahminlemede RandomForest, XGBoost, Gradient Boosting ve SVM makine öğrenme modelleri kullanılmıştır. Modellerin performansı Kesinlik, Duyarlılık, F1 Skoru ve ROC AUC metrikleri ile değerlendirilmiştir. Elde edilen bulgular, çocukların yaşlarının ve internete erişim imkânlarının cep telefonu sahipliği üzerinde belirgin etkisi olduğunu göstermiştir. Makine öğrenimi modelleri, istatistiksel metrikler açısından yüksek doğruluk değerleri sağlamıştır. Çalışma, makine öğrenimi modellerinin karar alma süreçlerini geliştirdiğini ve politika yapıcılar için etkili araçlar sağladığını ortaya koymuştur. Aynı zamanda, makine öğrenimi modellerinin sosyal bilimler alanında etkili bir şekilde kullanılabileceği de gösterilmiştir. Modellerin sunduğu yüksek doğruluk oranları ile veri odaklı politika geliştirme süreçlerinin daha etkin ve verimli hale getirilebileceği anlaşılmıştır.

Anahtar Kelimeler: Makine Öğrenimi, Yönetim Bilişim Sistemleri, Veri Analizi, RandomForest, XGBoost
Jel Kodları C53, C88, C45

Atıf: Eşidir, K. A. (2025). TÜİK mikro verileri ile çocuklarda cep telefonu sahipliğinin tahmini: Makine öğrenimi modellerinin karşılaştırmalı performansı. *Fiscaoeconomia*, *9*(3), 1525-1544. https://doi.org/10.25295/fsecon.1594029

Geliş Tarihi: 05.12.2024
Kabul Tarihi: 03.05.2025

Forecasting Cell Phone Ownership among Children with TurkStat Micro Data: Comparative Performance of Machine Learning Models

Abstract: In this study, we estimate children's mobile phone ownership and analyze the factors affecting mobile phone ownership among children using the micro data set of the Turkey Child Survey for 2022. RandomForest, XGBoost, Gradient Boosting, and Support Vector Machines (SVM) machine models are used in forecasting. The performance of the models was evaluated with Precision, Recall, F1 Score and ROC AUC metrics. The findings show that children's age and access to the internet have a significant impact on cell phone ownership. Machine learning models provided high accuracy values in terms of statistical metrics. The study found that machine learning models improve decision-making processes and provide effective tools for policymakers. Simultaneously, it has been demonstrated that they can be effectively utilized in the field of social sciences. The high accuracy rates achieved by the models demonstrate that data-driven policy development processes can be enhanced to become more effective and efficient.

Keywords: Machine Learning, Management Information Systems, Data Analysis, RandomForest, XGBoost
Jel Codes: C53, C88, C45

1. Giriş

Günümüzde veri odaklı karar alma süreçleri oldukça önem kazanmıştır. Büyük veri setlerini analiz etmek, karmaşık ilişkileri açığa çıkarmak ve doğru tahmin modelleri geliştirmek, bilimsel araştırma ve politika geliştirme süreçlerinin temel unsurlarıdır. Geleneksel istatistiksel yöntemler bazı durumlar için etkin olsa da bu yöntemler büyük

ve karmaşık veri setlerini analiz etmede genellikle yetersiz kalabilmektedir (Eşidir, 2025a). Bu bağlamda makine öğrenimi modelleri, karmaşık veri yapılarından anlamlı sonuçlar çıkartabilme imkân ve kabiliyetleri ile ön plana çıkmaktadır.

Çocukların yaşam koşulları, eğitim, sağlık ve sosyal gelişimleri, toplumların gelecekleri açısından büyük önem taşımaktadır. Türkiye'nin genç nüfus yapısı dikkate alındığında; çocukların refah seviyelerinin artırılması ve etkin politikaların geliştirilmesi, ülke kalkınma hedefleri açısından kritik öneme sahiptir. 2022 Türkiye Çocuk Araştırması; Türkiye İstatistik Kurumu (TÜİK), Aile ve Sosyal Hizmetler Bakanlığı ve UNICEF iş birliği ile gerçekleştirilmiştir. Araştırma sonucu elde edilen bulgular, çocukların mevcut durumunu daha iyi anlamaya olanak tanımakta ve çocuk refahını artırmaya yönelik stratejik müdahalelere rehberlik edecek nitelikler barındırmaktadır (TÜİK, 2023).

Çalışmada, bahsedilen mikro veri seti kullanılarak çocukların cep telefonu sahipliği tahmin edilmiş ve cep telefonu sahipliğini etkileyen faktörler belirlenmiştir. Makine öğrenmesi modellerinin uygulanmasında Python programlama dili temel araç olarak kullanılmıştır. Makine öğrenimi modelleri, geleneksel modellere kıyasla daha geniş bir değişken yelpazesi ile çalışabilmekte ve karmaşık etkileşimleri modelleyebilmektedir. Bu sebeplerden ötürü, çalışmada makine öğrenimi modelleri tercih edilmiştir. RandomForest, XGBoost ve Gradient Boosting ve benzeri modeller, özellikle büyük veri setlerinde yüksek doğruluk ve güvenilirlik sağlayarak, yapılan analizlerin etkinliğini artırmaktadır. Geleneksel istatistiksel modeller; (regresyon analizleri, lojistik regresyon ve ANOVA vs.), belli koşullar altında ve sınırlı değişkenler ile etkili olabilmektedir (Zhu vd., 2016). Fakat, büyük veri setleri ve çok boyutlu veri yapıları söz konusu olduğunda, bu modeller yetersiz kalmakta ve karmaşık ilişkileri açığa çıkarmada başarısız olmaktadır (Ji, 2023). Buna karşılık makine öğrenmesi modelleri ise, büyük ve karmaşık veri kümelerinden öğrenebilme yeteneği sayesinde, geleneksel modeller ile aşılamayan birçok zorluğu kolaylıkla aşabilmektedir (Bae vd., 2021). Makine öğrenimi modellerinin seçilmesinin temel nedeni, verinin karmaşıklığını ve çeşitliliğini daha iyi yönetebilme yeteneğidir (Pakarinen vd., 2022)

2. Literatür Taraması

Son zamanlarda yapılan çalışmalar, geleneksel istatistiksel yöntemlerin kapsamlı veri kümelerindeki karmaşık ilişkileri ve gizli kalıpları ortaya çıkarmada yeterli olmayabileceğini ve daha gelişmiş teknik ve modellerin kullanılmasını gerektirdiğini göstermiştir (Speer, 2021). Makine öğrenimi algoritmaları kullanılarak, daha geniş bir değişken ve etkileşim yelpazesini dikkate alan daha doğru tahmin modelleri geliştirilebilir (Lim vd., 2024). Makine öğrenimi algoritmaları, daha geniş bir değişkenler dizisini ve bunların etkileşimlerini dikkate alarak daha kapsamlı bir yaklaşımlar sunmaktadır (Wu, 2023).

Tosunoğlu vd. (2021) yaptıkları çalışmada, makine öğrenmesi modellerinin eğitim bilimlerindeki uygulamalarını incelemiştir. 2015-2020 yılları arasında yayımlanan 184 makale üzerinde yapılan içerik analizleri neticesinde, en sık kullanılan modellerin karar ağaçları (%22,5), destek vektör makineleri (%17,8) ve Naive Bayes (%14,2) olduğu belirlenmiştir. Yapılan çalışmaların çoğu deneysel veya uygulamalı olup, lisans öğrencileriyle ve mevcut veri tabanlarından elde edilen verilerle gerçekleştirilmiştir. Yapılan araştırma, eğitimde makine öğrenmesi kullanımına dair eğilimleri ortaya çıkartarak literatüre bu konuda katkı sunmaktadır.

Eriş Dereli (2021), TÜİK'in 2019 yılında yaptığı Çocuk İşgücü Araştırmasına ait mikro verileri analiz ederek, çocuk istihdamını etkileyen faktörleri incelemiştir. Çalışma sonucunda; kız çocuklarının çalışma olasılığının erkek çocuklara göre daha düşük olduğu, eğitimine devam eden çocukların istihdam olasılığının azaldığı ve ebeveyn eğitim seviyesinin artmasının çocuk işçiliğinin azalttığını ortaya çıkarmıştır. İlave olarak, çocuk yaşının artması ve anne-babanın işsiz olmasının istihdam olasılığını artırdığı tespit edilmiştir. Çalışma, çocuk işçiliği ile mücadelede kadınların eğitim ve istihdam politikalarına yönelik öneriler ile literatüre katkılar sunmaktadır.

Paslı Yeniay (2023), makine öğrenmesi algoritmalarının çocukların davranışlarını ve kimlik kazanma süreçlerini nasıl etkilediğini araştırmıştır. İnceleme, YouTube algoritmalarının çocukların izleme alışkanlıklarını yönlendirdiğini, medya içerikleriyle tüketim ve sosyal rollerini şekillendirdiğini vurgulamıştır. Çalışma; dijital platformların çocuk gelişimi üzerindeki etkisine dikkat çekerek, ebeveynlerin medya kullanımında rehberlik sağlamasının önemini anlatmaktadır.

Akın (2023) çalışmasında, 2022 yılına ait TÜİK Türkiye Çocuk Araştırması mikro verilerini kullanmıştır. İncelemesinde çocuk sosyalleşmesinin aile içi ilişkilerin ötesinde, arkadaşlık, akrabalık ve dijital araçlar aracılığıyla çok yönlü bir süreç olduğu vurgulanmıştır. Teknolojik cihazların (cep telefonu, tablet, bilgisayar vb.) çocukların sosyalleşmesinde önemli bir rol oynadığı ve bu araçların kullanım oranlarının yaş ve eğitim seviyesi arttıkça yükseldiği tespit edilmiştir. Dijitalleşme; sosyalleşme araçlarını çeşitlendirirken, çocukların boş zaman etkinliklerinden eğitime kadar birçok alanda etkili olduğu anlaşılmıştır. Sosyalleşme süreçlerinde yaş grupları ve eğitim seviyesine göre farklılıklar gözlemlenmiştir.

Zilyas & Yılmaz (2023) çalışmalarında, ortaokul öğrencilerine yönelik bir anket verisini analiz ederek, eğitim başarılarını tahmin etmek için makine öğrenimi modelleri geliştirmiştir. Çalışmada, Türkçe notu bağımlı değişken olarak seçilmiş ve öğrenci başarısını etkileyen aile geliri, ders çalışma süresi, özel ders alımı gibi faktörler analiz edilmiştir. Rastgele Orman algoritması, %88 doğruluk oranı ile en başarılı model olarak belirlenmiştir. Çalışma, eğitimde makine öğrenimi modelleirinin başarılı bir şekilde kullanılabileceğini göstermiştir.

Erbay Mermer (2023), makine öğrenmesinin eğitim alanında bireyselleştirilmiş öğrenme deneyimleri, öğrenci performansını izleme ve öğretim materyalleri geliştirme gibi süreçlerdeki etkilerini incelemiştir. Çalışmada, denetimli ve denetimsiz öğrenme algoritmalarının yanı sıra eğitimde veri analizine dayalı tahmin ve karar verme süreçlerinin önemine vurgu yapılmıştır. Ayrıca, makine öğrenmesinin eğitimdeki etik ve gizlilik sorunlarına dikkat çekilmiş, bu alanlarda çözüm önerileri sunulmuştur.

Özdemir vd. (2024), Türkiye'de çocuk işçiliği üzerine çalışma yapmışlardır. Çocuk işçiliğinin sağlık ve sosyal yaşam koşulları üzerindeki etkilerini incelemişlerdir. Çalışma, çocuk işçiliğinin fiziksel, zihinsel ve sosyal gelişimi olumsuz yönde etkilediğini ve bu durumun eğitimden kopuş, yoksulluk döngüsünün devamı gibi sonuçlara yol açtığını ortaya koymuştur. Araştırmada, çocukların çalışma koşullarının iyileştirilmesi ve çocuk işçiliğinin önlenmesi adına sosyal politikaların etkin şekilde uygulanmasının önemi vurgulanmıştır. Bu bağlamda, çocuk refahının artırılmasına yönelik yapılacak çalışmalarda kapsamlı veri temelli yaklaşımların geliştirilmesi gerektiği ifade edilmektedir.

Yapılan çalışma, mevcut literatürde çocukların cep telefonuna sahip olmasını etkileyen faktörlerin, Türkiye bağlamında makine öğrenmesi modelleri ile kapsamlı olarak ele alınmasındaki eksikliği doldurmaktadır. Geleneksel istatistiksel modellerin karmaşık ve çok boyutlu veri yapılarını analiz etmedeki sınırlılıklarına karşın, RandomForest, XGBoost, Gradient Boosting ve SVM gibi ileri düzey makine öğrenmesi modelleri kullanılarak daha kaliteli ve güvenilir bulgular elde edilmiştir. Ayrıca yaş, cinsiyet, internet erişimi ve eğitim türü gibi farklı değişkenlerin etkileşimli olarak incelenmesi ile yapılan çalışma literatüre katkı sağlamıştır. Çalışma, makine öğrenimi modellerinin sosyal bilimlerdeki potansiyelini ve önemini vurgulamaktadır.

3. Metodoloji

Büyük veri; çeşitlilik, hız ve hacimden dolayı analizlerde çeşitli zorluklara yol açmaktadır (Suthaharan, 2014). Makine öğrenmesi veriye dayalı problemlerin çözümünde, uygun algoritmalarla modelleme yapmayı hedefleyen bir çeşit hesaplama disiplinidir (El Naqa & Murphy, 2015). Çalışmada kullanılan metodoloji, sınıflandırma performansını iyileştirmek için farklı makine öğrenimi modellerini göstermeyi amaçlamaktadır.

Çalışmada kullanılan makine öğrenimi modelleri; veri setinin yapısı ve analiz hedefleri dikkate alınarak seçilmiştir. RandomForest modeli, değişkenler arası karmaşık ilişkileri güçlü biçimde yakalayabilme ve aşırı öğrenmeye (overfitting) karşı dirençli olma özelliklerinden ötürü tercih edilmiştir. XGBoost ve Gradient Boosting modelleri ise, özellikle büyük ve dengesiz veri setlerinde yüksek tahmin performansı göstermeleri ve sınıflar arasındaki ayrımı başarılı bir şekilde sağlamalarından dolayı seçilmiştir. SVM (Destek Vektör Makineleri) modeli ise doğrusal olmayan ilişkileri kernel fonksiyonları aracılığıyla etkin biçimde yakalayabildiği için modele dahil edilmiştir. Modellerin hiperparametre optimizasyonu aşamasında Grid Search kullanılmıştır. RandomForest modelinde karar ağacı sayısı (n_estimators), maksimum derinlik (max_depth), XGBoost ve Gradient Boosting modellerinde ise öğrenme oranı (learning rate), karar ağacı sayısı ve maksimum derinlik gibi hiperparametreler optimize edilmiştir. SVM modelinde ise kernel tipi, gamma ve C parametreleri için optimizasyon gerçekleştirilmiştir. Böylelikle modellerin performansları en üst seviyeye çıkarılmıştır.

3.1. Veri Seti

Çalışmada, 2022 yılına ait Türkiye Çocuk Araştırması mikro veri seti kullanılmıştır. Veri setinin elde edilmesi için gerçekleştirilen saha araştırması, TÜİK-Aile ve Sosyal Hizmetler Bakanlığı (Çocuk Hizmetleri Genel Müdürlüğü) iş birliği ile yürütülmüştür. Mikro veri seti, Türkiye genelinde 0-17 yaş grubunda çocuk bulunan hanelerden derlenen çeşitli sosyoekonomik, eğitim ve sağlık bilgilerini içermektedir. Veri setinde çocukların cinsiyeti, yaşı, eğitim durumu, yaşam koşulları, fiziksel çevresi, sosyal yardımlar, okul kalitesi, teknoloji kullanımı, beslenme alışkanlıkları gibi çok fazla ve çok çeşitli bilgiler yer almaktadır. Mikro veri seti, çocukların yaşam kalitesini etkileyen pek çok faktörü kapsayan detaylı anket verilerinden oluşmaktadır.

TÜİK'ten elde edilen orijinal mikro veri setindeki eksik veriler temizlenmiş ve makine öğrenimi analizlerine uygun hale getirilmiştir. 9.869 gözlem içeren temizlenmiş veri setinde, biri bağımlı ve 10'u bağımsız olmak üzere toplamda 11 adet değişken bulunmaktadır. Çocuğun cep telefonunun olup olmaması (CEP_TEL) bağımlı değişken olarak belirlenmiş, diğer değişkenler ise bağımsız değişkenler olarak analizlerde değerlendirilmiştir. Tablo 1'de mikro veri setinde yer alan değişkenlerin isimleri, açıklamaları ve kodlama seçenekleri ifade edilmiştir.

Tablo 1. Çalışmada Kullanılan Mikro Veri Seti Değişkenleri, Açıklamaları ve Kodlama Seçenekleri

Değişken Adı	**Açıklama**	**Seçenekler**
CINSIYET	Çocuğun cinsiyeti	1: Erkek, 2: Kadın
YAS	Çocuğun yaşı	
OKUL_TUR	Çocuğun devam ettiği okulun türü	1: Devlet okulu, 2: Özel okul
YARDIM_SOSYAL_DESTEK	Çocuk için sosyal ve ekonomik destek alınıp alınmadığı	1: Evet, 2: Hayır
KENDI_ODA	Çocuğun kendine ait bir odası olup olmadığı	1: Evet, 2: Hayır
DERS_MASA	Çocuğun ders çalışmak için bir masası olup olmadığı	1: Evet, 2: Hayır
CAL_ORTAM	Çocuğun çalışma ortamının uygunluğu	1: Uygun, 2: Uygun değil
BILGISAYAR	Evde bilgisayar bulunup bulunmadığı	1: Evet, 2: Hayır
INTERNET	Evde internet bağlantısı olup olmadığı	1: Evet, 2: Hayır
MEYVE_YEME	Çocuğun düzenli olarak meyve tüketip tüketmediği	1: Hiç, 2: Nadiren, 3: Haftada birkaç kez, 4: Her gün, 5: Ayda bir kez, 6: Yılda birkaç kez, 7: Yılda bir kez, 8: 2 veya daha fazla yılda bir, 9: Hiçbir zaman
CEP_TEL (Bağımlı Değişken)	Çocuğun cep telefonu olup olmadığı	1: Evet, 2: Hayır

Tablo 1'de yer alan bağımsız değişkenler, çocuğun demografik özellikleri, ekonomik durum ve yaşam koşullarını kapsayan farklı kategorilerde tanımlanmıştır. Bu değişkenler, bağımlı değişken olan çocuğun cep telefonu sahipliği ile olan ilişkisini anlamak için kullanılmıştır.

3.2. Veri Ön İşleme

Veri ön işleme, veri setinin analiz ve modelleme için hazır hale getirilmesi sürecidir. Bu aşamada gerçekleştirilen işlemler, modellerin doğruluğunu ve genel performansını önemli ölçüde etkilemektedir (Gür, 2024a). Modellerin eğitilmesi sırasında ihtiyaç duyulmayan çok fazla eksik veriler içeren sütunlar, veri setinden çıkarılmıştır Hedef (bağımlı) değişken, "Evet" değeri için 0 ve "Hayır" değeri için 1 olacak şekilde ikili formata dönüştürülmüştür. Kategorik değişkenler, makine öğrenmesi algoritmalarıyla uyumlu hale getirilmek için sayısal değerlere dönüştürülmüştür. Ardından, dönüştürülen veri seti, eğitim (%80) ve test (%20) setlerine ayrılmıştır. Sayısal veriler, Standard Scaler kullanılarak ölçeklendirilmiştir. Bu işlem, özellikler arasındaki ölçek farklılıklarını gidererek modelin daha etkili öğrenmesine yardımcı olmaktadır (Zeiler & Fergus, 2014). Matematiksel olarak bu işlem Denklem 1'de gösterildiği gibi ifade edilmektedir:

$$z = \frac{(x - \mu)}{\sigma} \tag{1}$$

Burada, x ölçeklendirilecek olan özellik değerini, μ bu özelliğin ortalamasını, σ bu özelliğin standart sapmasını ve z ölçeklendirilmiş sonucu temsil etmektedir.

3.3. Model Performans Değerlendirme Kriterleri

Makine öğrenmesinde analizi gerçekleştirilen modellerin performansını objektif bir şekilde değerlendirmek için çeşitli metrikler ve yöntemler kullanılmaktadır (Gür, 2024c). Bu çalışmada, makine öğrenmesi modellerinin performansları, Kesinlik (Precision), Duyarlılık (Recall), F1 Skoru ve Doğruluk (Accuracy) metrikler ile değerlendirilmiştir. Bu metriklerin, matematiksel formülleri sırasıyla Denklem 2, 3, 4 ve 5'te gösterilmiştir. Denklemlerde, DP doğru pozitif, YP yanlış pozitif, DN doğru negatif ve YN yanlış negatiftir. Ayrıca, tahminlerin doğruluğunu görsel açıdan değerlendirebilmek amacıyla karmaşıklık matrisleri de kullanılmıştır.

$$Kesinlik\ (Precision) = \frac{DP}{DP + YP} \tag{2}$$

$$Duyarlılık\ (Recall) = \frac{DP}{DP + YN} \tag{3}$$

$$F1\ Skoru = 2 \times \frac{Kesinlik \times Duyarlılık}{Kesinlik + Duyarlılık} \tag{4}$$

$$Doğruluk\ (Accuracy) = \frac{DP + DN}{DP + DN + YP + YN} \tag{5}$$

Makine öğrenmesi modellerinin performansını değerlendirmek için kullanılan Kesinlik, Duyarlılık, F1 Skoru ve Doğruluk metrikleri, farklı senaryolarda modelin gücünü ve zayıflıklarının belirlenmesi açısından önem arz etmektedir (Gür, 2024c).

4. Makine Öğrenmesi Modelleri

Makine öğrenimi, bilgisayarlara veri analiz ederek öğrenme ve tahmin yapma yeteneği kazandırmayı hedefleyen bir bilim dalıdır (Eşidir, 2025c). Çözüm süreçlerinde farklı matematiksel ve istatistiksel modeller kullanılmakla beraber, modellerin seçimi veri kümesinin yapısına ve özelliklerine göre belirlenmektedir (Akbulut & Adem, 2023).

4.1. XGBoost (Extreme Gradient Boosting)

XGBoost, gradient boosting algoritmalarının geliştirilmiş bir versiyonudur ve yüksek performans gösteren makine modellerinden birisidir. Sınıflandırma ve regresyon problemlerinde sıkça kullanılmaktadır. Hızlı olması ve yüksek doğruluk oranları sunması, veri bilimi yarışmaları ve büyük ölçekli projelerde popüler hale gelmesini sağlamıştır. XGBoost, son yıllarda makine öğrenmesi alanında başarılı ve yaygın

kullanılan modellerden birisidir (Ma vd., 2018). Spam e-postaların sınıflandırılması, reklam eşleştirme, dolandırıcılık tespiti ve anomali tespiti benzeri birçok alanda etkili sonuçlar sunmaktadır (Chen & Guestrin, 2016). XGBoost, doğru tahminler üretmede başarısını ispatlamıştır ve alternatif makine öğrenimi modellerine kıyasla hesaplama avantajları sunmaktadır (Gür, 2024b). Özellikle büyük veri setlerinde ve karmaşık problemlerde etkili performans göstermesi, XGBoost'u popüler yapmıştır (Abar, 2020).

XGBoost, gradyan artırma algoritmasına dayalı olarak çalışan bir topluluk öğrenme modelidir. Bu teknik, bir hata fonksiyonunu en aza indirmek için bir dizi zayıf öğreniciyi sırayla birleştirmeyi ve böylece tahmin doğruluğunu artırmayı içermektedir (Chen, 2016). XGBoost, ölçeklenebilirliği ve esnekliği nedeniyle popülerlik kazanmış ve gradyan artırma kütüphanesi içinde giderek yükselen bir topluluk öğrenme yöntemi haline gelmiştir (Aslan, 2025). Torbalama ve istifleme gibi yaygın olarak kullanılan diğer topluluk yöntemlerinin yanı sıra, boosting algoritması kategorisine giren bir topluluk öğrenme algoritması olarak kabul edilmektedir (Wang vd., 2020). Extreme Gradient Boosting'in kısaltması olan XGBoost, gradient boosting ailesinin bir parçasıdır ve zayıf öğrenenleri güçlü öğrenenlere dönüştürmeyi amaçlamaktadır (Liang vd., 2020). Modelin tahmin fonksiyonu Denklem 6'da gösterilmektedir:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \qquad f_k \in F \tag{6}$$

Burada, $\hat{y}_i$ tahmin edilen değerleri, f karar ağaçlarının uzayını, k toplam ağaç sayısını ve $f_k$her bir ağacı temsil etmektedir. XGBoost, her iterasyonda hatayı minimize etmek için gradyan iniş algoritmasını kullanır ve bu sayede modelin doğruluğunu artırır. Modelin genel kayıp fonksiyonu Denklem 7'deki gibidir:

$$L(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{7}$$

Burada, L(ϕ) genel kayıp fonksiyonunu, l(y_i, $\hat{y}_i$) tahmin hatasını, $\Omega(f_k)$ karmaşıklık cezasını, y_i gerçek değeri ve $\hat{y}_i$ tahmin edilen değeri göstermektedir.

4.2. RandomForest (Rastgele Orman)

RandomForest, farklı alt veri setlerinde birden fazla karar ağacının birleşiminden oluşmaktadır. Rastgele Orman (Random Forest), sınıflandırma ve regresyon görevlerinde yaygın olarak kullanılan, topluluk öğrenme yaklaşımına dayalı bir makine öğrenimi algoritmasıdır. Bu algoritma, birden fazla karar ağacını bir araya getirerek her bir ağacın zayıflıklarını dengelemeyi hedeflemekte ve bu sayede daha güçlü ve dengeli bir model ortaya çıkmaktadır (Oguine & Oguine, 2021).

Rastgele Orman, eğitim veri setinden birçok bootstrap örneği çeker (her örnek, orijinal veri setinin rastgele seçilmiş örneklerini içermektedir). Her bir bootstrap örneği için bağımsız bir karar ağacı kurulur. Her ağaç, veri setinin farklı bir alt kümesi üzerinde eğitilir. Her düğümdeki bölünme işlemi, tüm özellikler yerine rastgele seçilen bir alt küme kullanılarak yapılmaktadır. Bu durum, modelin varyansını azaltmaya yardımcı olur ve ağaçlar arası korelasyonu düşürür (Eşidir, 2025a). Sınıflandırma için tahmin yaparken genellikle Denklem 8'de gösterildiği gibi bir ifade kullanılmaktadır:

$$y = \text{mod}\{y_1, y_2, \dots, y_n\} \tag{8}$$

Burada, $y_1, y_2, \dots, y_n$ her bir karar ağacının tahmin ettiği sınıf etiketleridir ve mod fonksiyonu en sık rastlanan etiketi belirler.

4.3. GradientBoosting (Gradyan Artırma)

Gradient Boosting, sınıflandırma ve regresyon gibi gözetimli öğrenme görevleri için kullanılan güçlü bir makine öğrenimi tekniğidir. Bu model, zayıf tahmin modellerini (genellikle karar ağaçları) ardışık bir şekilde geliştirerek modelin doğruluğunu

artırmaktadır. Geliştirilen her yeni model, önceki modellerin hatalarını düzeltmeye odaklanmaktadır, böylece her adımda performansın artması sağlanmaktadır (Ou, 2020). Gradient Boosting, veri setindeki hedef değişkenin ortalama değeriyle başlayarak basit bir model kurar ve bu model regresyon için doğrudan ortalamayı, sınıflandırma için ise log (odds) değerini kullanır (Bazilevych vd., 2023). İlk modelin hataları (artıkları), sonraki modellerin iyileştirmeye çalışacağı hedef olarak belirlenir. Her adımda, önceki modelin hatalarını en iyi düzeltecek yeni bir zayıf öğrenici (genellikle basit karar ağaçları) eklenir. Her zayıf öğrenicinin katkısı, gradient descent benzeri bir optimizasyon tekniğiyle ayarlanır, böylece modelin hata fonksiyonu azaltılır. Ağaçların çıktıları belirli bir öğrenme hızıyla ağırlıklandırılarak toplanır ve bu süreç, model yeterince iyi performans gösterene veya belirlenen iterasyon sayısına ulaşana kadar devam eder. Bu yöntem, modelin hızla ve etkin bir şekilde öğrenmesini sağlar ve genellikle düşük bir öğrenme hızı tercih edilir (Aslan, 2025).

Her bir adımda $F_m(x)$ modeli için güncellenme formülü Denklem 9'da gösterildiği gibidir:

$$F_{m+1}(x) = F_m(x) + \gamma_m h_m(x) \tag{9}$$

Formül, mevcut modelin $F_m(x)$, bir sonraki adımda eklenen zayıf öğrenici $h_m(x)$ ve bu öğrenicinin model üzerindeki etkisini belirleyen ağırlık γ_m ile nasıl güncellendiğini ifade etmektedir. γ_m değeri, hata fonksiyonunu minimize edecek şekilde Denklem 10'da gösterildiği gibi seçilmektedir:

$$\gamma_m = \arg\, min_\gamma \sum_{i=1}^{n} L\big(y_i, F_m(x_i) + \gamma h_m(x_i)\big) \tag{10}$$

Burada, L kayıp fonksiyonudur ve y_i , x_i sırasıyla i 'nci gerçek değer ve özelliklerdir.

4.4. Destek Vektör Makineleri (Support Vector Machines - SVM)

Destek Vektör Makineleri (SVM), hem sınıflandırma hem de regresyon görevlerinde kullanılan, güçlü ve esnek bir makine öğrenimi modelidir. Bu model, sınıfları en iyi şekilde ayıran bir hiper düzlem bulmaya odaklanır ve doğrusal olarak ayrılabilir verilerde oldukça etkili sonuçlar verir. Doğrusal olarak ayrılamayan veri setlerinde ise kernel fonksiyonları kullanılarak yüksek boyutlu uzaylara geçiş sağlanır ve sınıflandırma doğruluğu artırılır. SVM; metin ve görüntü tanıma, biyometrik tanımlama, finansal tahminler ve biyoinformatik gibi çeşitli alanlarda başarı ile uygulanmaktadır. Özellikle, sınıflandırma problemlerinin çözümünde etkinliği kanıtlanmıştır (Ayhan & Erdoğmuş, 2014).

Destek Vektör Regresyonu (DVR), Destek Vektör Makineleri (DVM) prensiplerini regresyon problemlerine genişleten bir makine öğrenme tekniğidir. DVR, tüm örnek noktalarının hiper düzlemden toplam sapmasını en aza indiren optimal bir hiper düzlem bulmayı amaçlamaktadır (Cao vd., 2022). DVR'de algoritma, eğitim örneklerine dayalı olarak optimum bir regresyon hiper düzlemi oluşturur ve noktaları hiper düzleme yakın tutarak ayrımı en aza indirmeye çalışır (Wen vd., 2019). SVR modeli, hiper düzlem ile eğitim verileri arasındaki marjı en üst düzeye çıkarmak için ikinci dereceden bir optimizasyon problemini çözer ve noktaları belirli bir hata toleransı dahilinde etkili bir şekilde kapsar. DVR, giriş verilerini doğrusal olmayan işlevler aracılığıyla düşük boyutlu bir uzaydan yüksek boyutlu bir özellik uzayına eşleyerek ve ardından örnek noktaların bu hiper düzleme mümkün olduğunca yakın düşmesini sağlamak için yüksek boyutlu özellik uzayında optimum bir hiper düzlem arayarak çalışır (Zhang vd., 2021). Bu süreç, eğitilmiş veri kümesine minimum mesafeyi sağlayan hiper düzlemi bulmayı ve veri örneklerine iyi uyması için hiper düzlemi optimize etmeyi içerir. DVR'nin prensibi, hiper düzlemden en uzaktaki örneklerin geometrik mesafesini en aza indiren ve veri dağılımını

etkili bir şekilde yakalayan optimum hiper düzlemi aramaktır. Modelin matematiksel ifadesi Denklem 11'de gösterilmiştir.

$$y = \sum_{i=1}^{n} (a_i - a_i^*) K(x_i, x) + b \tag{11}$$

Burada, y tahmin edilen değeri, a_i ve a_i^* Lagrange çarpanlarını, $K(x_i, x)$ çekirdek fonksiyonunu, b bias terimini ve x_i eğitim verisini temsil etmektedir. SVR, doğrusal olmayan ilişkileri modellemek için çekirdek fonksiyonları (lineer, polinomial, radyal bazlı fonksiyonlar gibi) kullanır. Çekirdek fonksiyonu $K(x_i, x)$, yüksek boyutlu uzayda veri noktaları arasındaki benzerlikleri ölçer.

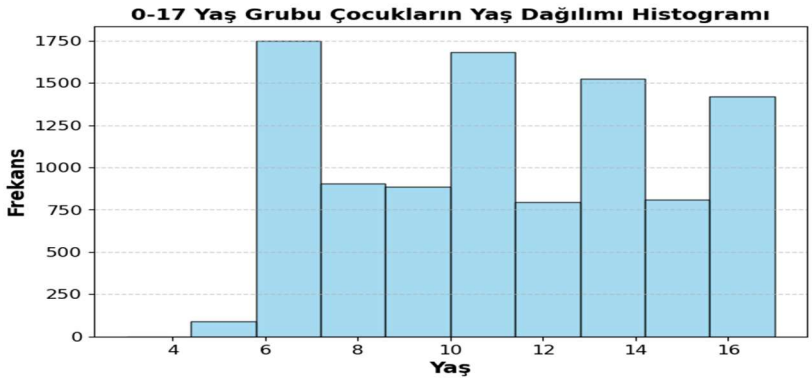
5. Yazılım ve Donanım

Yüksek performanslı donanım ve cihazlar isteyen veri analizi ve makine öğrenimi süreçlerini desteklemek için, güçlü bir bilgisayar sistemi kullanılmıştır. Kullanılan bilgisayar, Windows 11 Pro işletim sisteminin en güncel sürümünü çalıştırmakta ve 64-bit mimarili Intel64 Family 6 Model 142 Stepping 10 işlemciye sahiptir. 16 GB RAM kapasitesi geniş veri kümelerini işlemek ve modelleri eğitmek için yeterli bellek alanı barındırmaktadır.

Analizler, Python programlama dilinin en güncel sürümü olan 3.12.7 kullanılarak gerçekleştirilmiştir. Python programlama dili, bilimsel hesaplama ve veri analizi için geniş kütüphane desteği (NumPy, Pandas, Scikit-learn, TensorFlow, Keras vb.) sunmasından ötürü tercih edilmiştir. İşlemlerin kolay yapılabilmesi ve analiz süreçlerinin anlaşılır olması için Jupyter Notebook 7.2.2 ortamından yararlanılmıştır. Jupyter Notebook; kodları, grafikleri ve açıklamaları entegre olarak sunabilen bir platform olarak (Eşidir, 2025b) çalışmaların etkin olarak yürütülebilmesine olanak sağlamıştır.

6. Bulgular

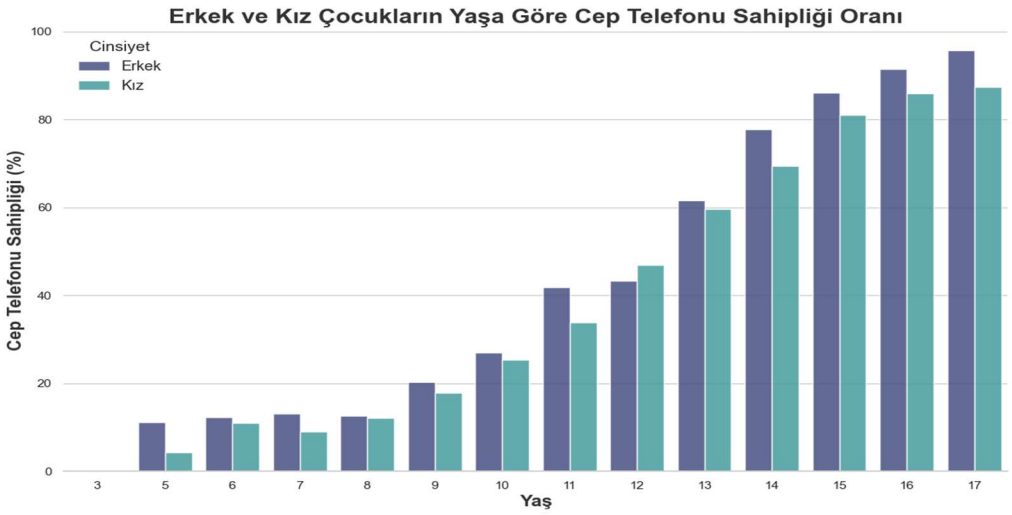
Çalışmada, 0-17 yaş grubundaki çocukların cep telefonu sahipliğini etkileyen demografik ve sosyoekonomik faktörler incelenmiştir. Analizler, yaş, cinsiyet, internet erişimi ve eğitim türü gibi değişkenlerin cep telefonu sahipliğinde önemli bir rol oynadığını ortaya koymuştur. Özellikle 6 ve 10 yaş gruplarındaki yoğunluk, çocukların yaşa bağlı teknoloji kullanım alışkanlıklarının değişimini yansıtmaktadır. Bu bağlamda, makine öğrenimi modelleri kullanılarak daha derinlemesine analizler gerçekleştirilmiştir. Rastgele Orman (Random Forest), XGBoost (Extreme Gradient Boosting), Gradyan Artırma (Gradient Boosting) ve Destek Vektör Makineleri (Support Vector Machines - SVM) modelleri, çocukların cep telefonu sahipliğini tahmin etmede etkin bir şekilde kullanılmıştır.



Şekil 1. 0-17 Yaş Grubundaki Çocukların Yaş Dağılımı Histogramı

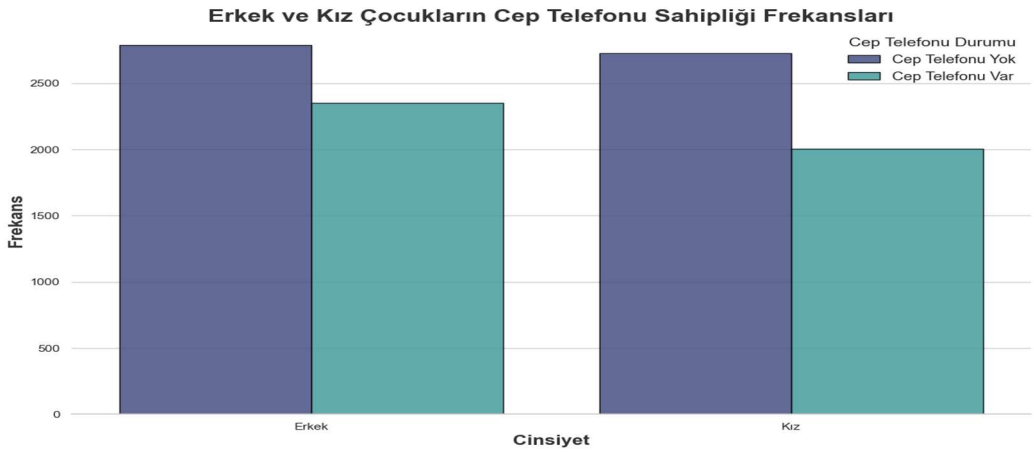
Şekil 1'de 0-17 yaş grubundaki çocukların yaş dağılımı yer almaktadır. 6 ve 10 yaş gruplarında diğer yaş gruplarına kıyasla yüksek bir frekans gözlenmektedir. Bu yaş

aralıklarında daha fazla çocuk bulunmaktadır. Eksik verilerin çıkarılmasından sonra oluşturulan histogram, yaş gruplarının homojen olmayan dağılımını net bir şekilde göstermektedir.



Şekil 2. Erkek ve Kız Çocukların Yaşa Göre Cep Telefonu Sahipliği Oranı

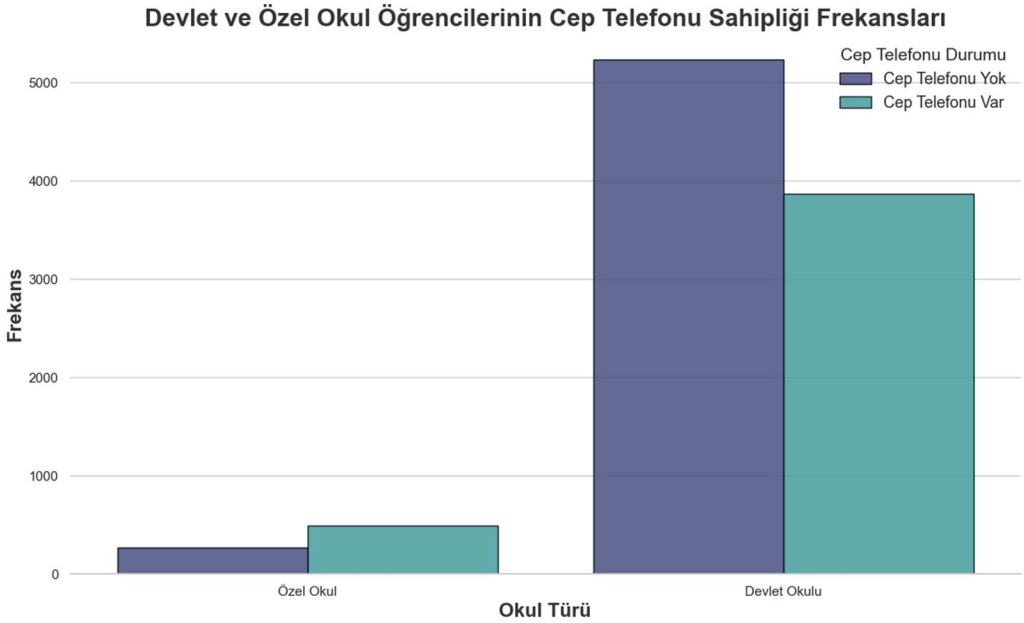
Şekil 2'de cep telefonu sahipliği oranlarının yaşla birlikte belirgin bir şekilde arttığı görülmektedir. 3-9 yaş arası çocuklarda cep telefonu sahipliği oldukça düşük seviyelerde kalırken, özellikle 13 yaş ve sonrasında bu oran hızla artmaktadır. Bu eğilim, ergenlik dönemine giren çocukların iletişim, sosyalleşme ve kişisel ihtiyaçlarının artmasıyla ilişkilidir. Bu yaş aralığında cep telefonlarının neredeyse temel bir ihtiyaç haline geldiği hem erkek hem de kız çocuklarının %80'in üzerinde sahiplik oranına ulaşmasından anlaşılmaktadır. Küçük yaş gruplarında ise sahiplik oranlarının düşüklüğü, bu yaşlardaki çocukların teknolojik cihazlara olan ihtiyacının az olmasından veya ebeveynlerin bu konuda daha kısıtlayıcı olmasından kaynaklanabilmektedir. Erkek ve kız çocuklarının cep telefonu sahipliği arasında küçük de olsa farklılıklar gözlemlenmektedir. Özellikle ergenlik dönemine girerken, erkek çocuklarının cep telefonu sahiplik oranı kızlara kıyasla biraz daha yüksektir. Ancak 14 yaş ve sonrasında bu fark giderek azalmakta ve her iki cinsiyetin de sahiplik oranı birbirine oldukça yakın seviyelere ulaşmaktadır. Bu durum, ergenlik döneminde cep telefonunun hem erkek hem de kız çocukları için önemli bir ihtiyaç haline geldiğini ve bu cihazların gençlerin günlük yaşamlarında ne kadar önemli bir yer kapladığını göstermektedir.



Şekil 3. Erkek ve Kız Çocuklarının Cep Telefonu Sahipliği Frekansları

Şekil 3'te erkek ve kız çocuklarının cep telefonu sahipliği durumları frekans bazında karşılaştırılmıştır. Bu grafikten çıkarılabilecek önemli sonuç, cinsiyetin cep telefonu

sahipliği üzerinde belirli bir rol oynadığı ve her iki cinsiyet arasında sahiplik oranlarında farklılıklar gözlemlendiğidir. Cinsiyetler arasında cep telefonu sahipliği farkı, kültürel, sosyoekonomik veya aile içindeki tercihlere bağlı olabilir. Grafik, aynı zamanda teknoloji kullanımına yönelik toplumsal eğilimleri de yansıtmaktadır.



Şekil 4. Okul Türüne Göre Cep Telefonu Sahipliği Oranı

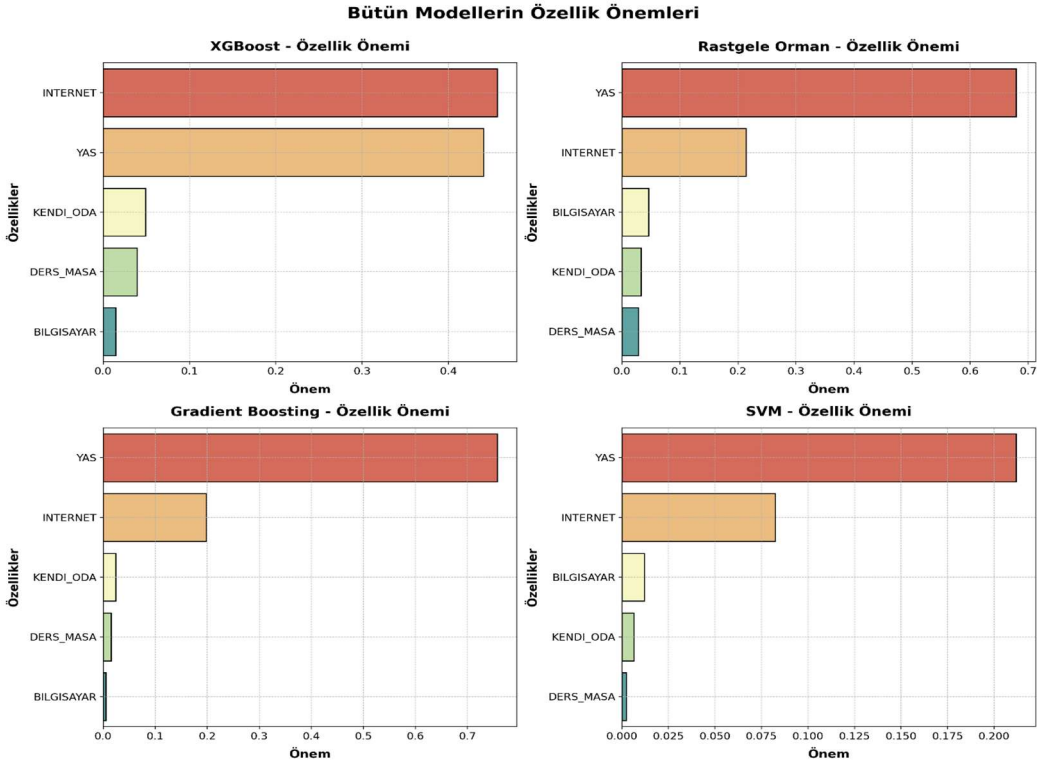
Şekil 4'te, özel okul ve devlet okulu öğrencileri arasında cep telefonu sahipliği durumu karşılaştırılmaktadır. Özel okul öğrencilerinin cep telefonu sahipliği, devlet okulu öğrencilerine kıyasla belirgin şekilde daha yüksektir. Bu durum, okul türü ile cep telefonu sahipliği arasında bir fark olduğunu göstermektedir. Bu farklılık, özel okullardaki öğrencilerin sosyoekonomik durumu ve teknolojiye erişim imkanlarının daha iyi olduğunu göstermektedir. Bu farklılık, ailelerin çocuklarına sağladığı teknolojik imkanları ifade etmektedir.

Yapılan veri bilimi analizleri sonucu elde edilen bulgular, sosyal politikaların şekillendirilmesinde kritik öneme sahip çıkarımlar ortaya koymaktadır. Yaş ve internet erişiminin cep telefonuna sahip olmadaki belirleyici etkisi, çocukların dijital araçlara erişim imkanlarındaki eşitsizliklerin giderilmesi gerektiğini göstermektedir. Özellikle devlet ve özel okullar arasında ortaya çıkan belirgin farklılık, eğitim politikalarının yalnızca eğitim kalitesine değil, aynı zamanda teknoloji erişimine yönelik stratejileri de içermesi gerektiğini ortaya koymaktadır. Bu bağlamda, düşük sosyoekonomik düzeydeki ailelerin çocuklarına yönelik teknoloji destek programlarının geliştirilmesi hem eğitim eşitliğini sağlama hem de dijital bölünmeyi azaltma açısından önemlidir. Ayrıca, yaşa bağlı olarak çocukların dijital cihazlara erişiminin kontrollü biçimde düzenlenmesi, ebeveynlerin ve eğitim kurumlarının teknolojiyi bilinçli kullanımına yönelik farkındalıklarını artıracak politikaların tasarlanmasını da gerektirmektedir. Mikro veri temelli analiz sonuçları, çocukların refahının artırılması ve dijital dünyaya erişim konusundaki eşitsizliklerin giderilmesi açısından politikacıların veri bilimi temelli müdahalelerde bulunabilmesine katkılar sunmaktadır.

Tablo 2. Makine Öğrenimi Modellerinin Performans Ölçütlerinin Karşılaştırması

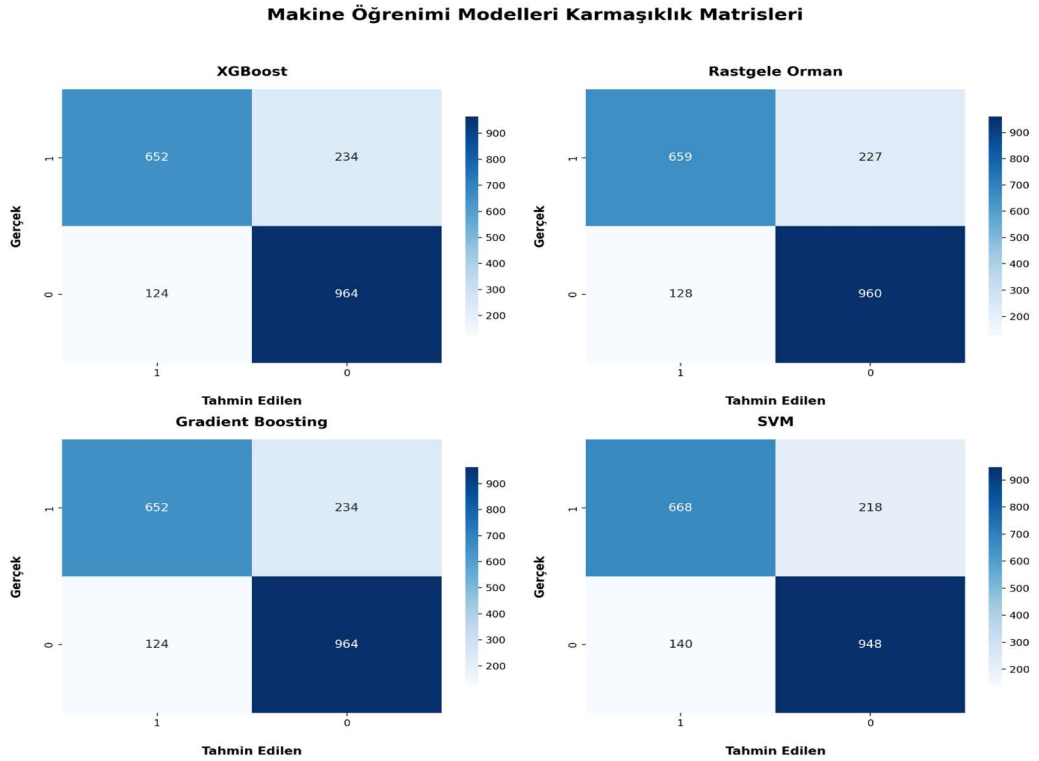
Makine Öğrenimi Modelleri	Doğruluk (Accuracy)	Kesinlik (Precision)	Duyarlılık (Recall)	F1 Skoru	ROC AUC
XGBoost	0,819	0,821	0,819	0,817	0,893
Rastgele Orman	0,820	0,822	0,820	0,819	0,889
Gradient Boosting	0,819	0,821	0,819	0,817	0,893
SVM	0,819	0,819	0,819	0,818	0,873

Tablo 2'de sunulan sonuçlara göre, makine öğrenimi modelleri arasında Rastgele Orman, doğruluk (0,820), kesinlik (0,822), duyarlılık (0,820) ve F1 skoru (0,819) metriklerinde diğer modellerden bir adım öndedir. Bu model, hem pozitif hem de negatif sınıflar arasında dengeli bir performans sergileyerek, özellikle sınıflandırma başarısına önem verilen uygulamalarda güçlü bir seçenek olarak öne çıkmaktadır. Bununla birlikte, XGBoost ve Gradient Boosting modelleri, ROC AUC metriğinde 0,893 ile en yüksek performansı göstermiştir. Bu, modellerin tahmin yeteneğini ölçen önemli bir metrik olarak dikkate alınabilir. SVM ise diğer modellerle benzer doğruluk ve F1 skorlarına sahip olmasına rağmen, ROC AUC değeri (0,873) açısından bir miktar geride kalmıştır. Sonuçlara göre, analizlerde hangi metriğin daha kritik olduğuna bağlı olarak model seçimi yapılabilir. Eğer genel sınıflandırma performansı önemliyse Rastgele Orman, daha güçlü ayrım gücü gerekiyorsa XGBoost veya Gradient Boosting tercih edilebilir.



Şekil 5. Makine Öğrenimi Modellerinin Önem Derecelerinin Karşılaştırılması

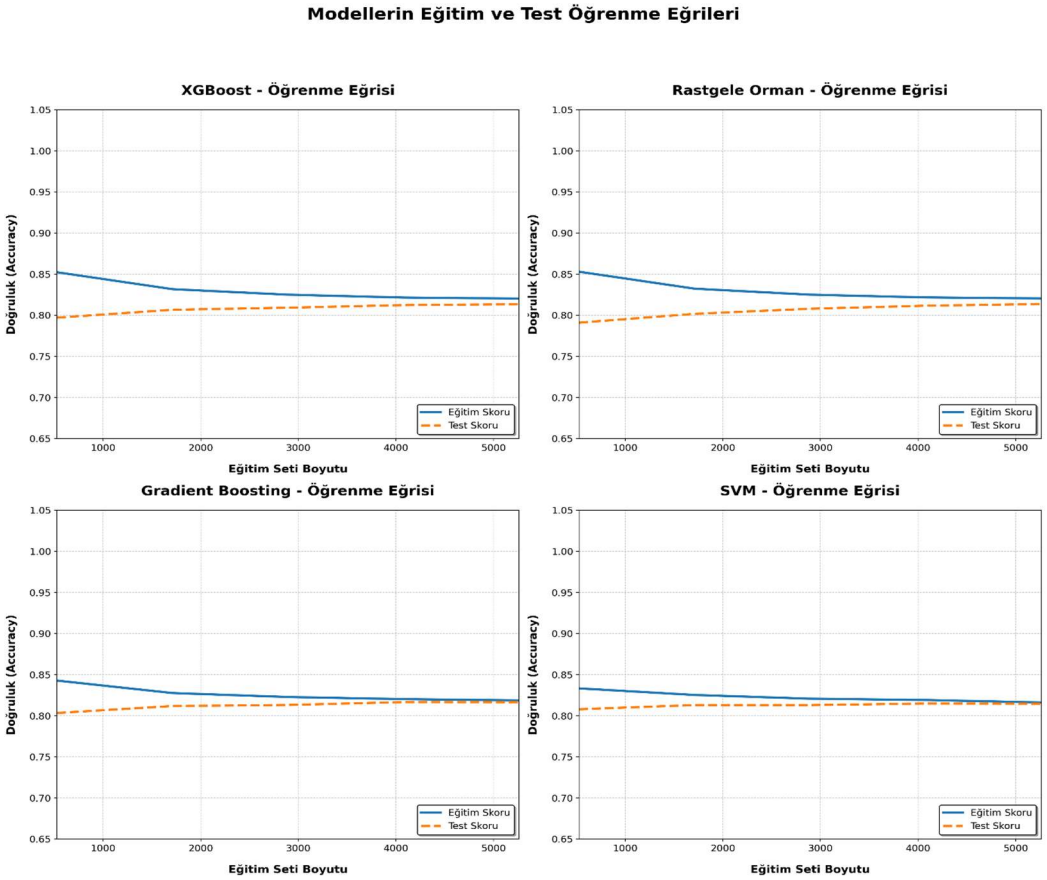
Şekil 5'te makine öğrenimi modellerinin önem derecelerinin karşılaştırılmıştır. Tüm modellerde yaş değişkeni, cep telefonu sahipliğini belirleyen en güçlü faktör olarak öne çıkmaktadır. Bu bulgu, yaşın, teknolojik araçlara erişim ve kullanım ihtiyacını etkileyen temel bir değişken olduğunu doğrulamaktadır. İnternet değişkeni ise cihaz sahipliğinde ikinci en önemli belirleyici olarak dikkat çekmektedir ve dijital erişim politikalarının çocukların teknoloji kullanımı üzerindeki etkisini vurgulamaktadır.



Şekil 6. Makine Öğrenimi Modelleri için Karmaşıklık Matrisleri

Şekil 6'da modellerin karmaşıklık matrisleri gösterilmiştir. Verilen karmaşıklık matrisleri, XGBoost, Rastgele Orman, Gradient Boosting ve SVM modellerinin cep telefonu sahipliğini tahmin etmedeki performanslarını karşılaştırmaktadır. XGBoost, Rastgele Orman ve Gradient Boosting modelleri benzer sonuçlar sergileyerek doğru sınıflandırma oranlarını yüksek tutmuştur. SVM modeli, "1" sınıfını tanımada diğer modellere göre biraz daha başarılıdır (668 gerçek pozitif), ancak bu avantajı yanlış pozitif sayısında artışa neden olmuştur. Genel olarak, XGBoost ve Gradient Boosting modelleri daha dengeli bir performans gösterirken, SVM modeli pozitif sınıfları yakalamakta daha hassastır.

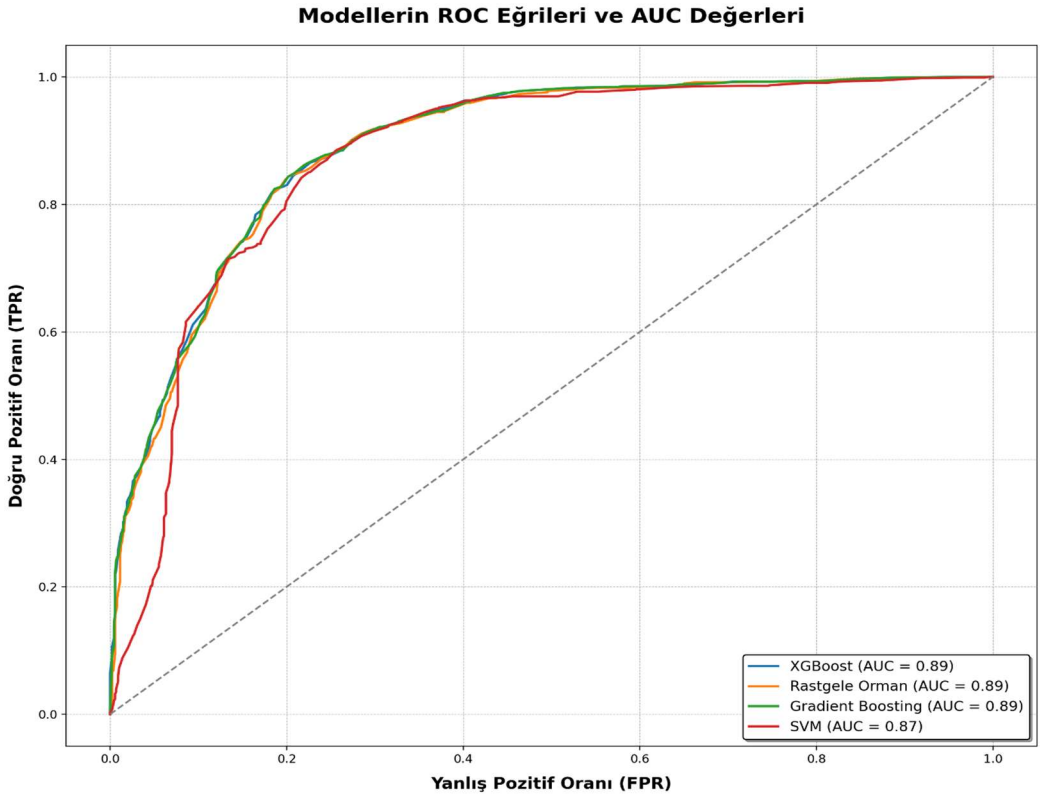
Model seçimi yapılırken, yanlış pozitif ve yanlış negatiflerin dengelenmesi isteniyorsa XGBoost veya Gradient Boosting modelleri tercih edilebilir. Pozitif sınıfın (örneğin, belirli bir risk grubunu kaçırmamak) daha iyi yakalanması önemliyse, SVM modelinin kullanılması uygun olabilir. Bu değerlendirmeler, belirli koşullara ve önceliklere göre hangi modelin daha iyi performans göstereceğine karar vermede rehberlik edebilir.



Şekil 7. Makine Öğrenimi Modellerinin Eğitim ve Test Öğrenme Eğrileri

Şekil 7'de dört farklı makine öğrenimi modelinin (XGBoost, Rastgele Orman, Gradient Boosting ve SVM) eğitim ve test seti üzerindeki öğrenme eğrileri gösterilmektedir. Tüm modellerde benzer bir eğilim gözlemlenmektedir: Eğitim seti büyüdükçe, eğitim skoru azalırken test skoru artış göstermekte ve belirli bir seviyede sabitlenmektedir. Bu durum, modellerin eğitim setine aşırı uyum sağlamaktan (overfitting) kaçındığını ve test setindeki genelleme yeteneklerinin arttığını göstermektedir. XGBoost ve Gradient Boosting modellerinde eğitim skoru başta oldukça yüksekken (yaklaşık 0,95 civarında), veri miktarı arttıkça bu skorlar test skorlarına yaklaşmakta ve belirli bir denge noktasına ulaşmaktadır.

Rastgele Orman ve SVM modelleri ise daha dengeli bir performans sergilemektedir ve her iki modelin de test skorları yaklaşık 0,80 civarında sabitlenmiştir. Bu, bu modellerin genel olarak daha düşük bir varyans gösterdiğini, yani daha kararlı olduğunu göstermektedir. Eğitim ve test skorlarının birbirine yakın olması, bu modellerin overfitting'e daha az yatkın olduğunu işaret eder. Genel olarak, tüm modellerin öğrenme eğrileri, daha fazla veri kullanıldıkça genelleme performanslarının iyileştiğini ve aşırı uyum sorunlarının azaldığını göstermektedir.



Şekil 8. Modellerin ROC Eğrileri ve AUC Değerleri

Şekil 8'de modellerin ROC eğrileri ve AUC değerleri gösterilmiştir. Bu grafikte, dört farklı makine öğrenimi modelinin (XGBoost, Rastgele Orman, Gradient Boosting ve SVM) ROC eğrileri ve AUC (Area Under Curve) değerleri gösterilmektedir. ROC eğrileri, bir modelin doğru tahmin oranını yanlış tahmin oranıyla karşılaştırarak genel performansını değerlendirir ve modelin sınıflandırma yeteneğini değerlendirmek için kullanılır. Eğrinin altındaki alan (AUC) değeri, modelin genel sınıflandırma performansını ifade eder; 1'e ne kadar yakınsa modelin performansı o kadar iyidir. Grafikte XGBoost, Rastgele Orman ve Gradient Boosting modellerinin AUC değerlerinin 0.89 olduğu görülmektedir. Bu, bu üç modelin sınıflandırma performansının oldukça benzer ve yüksek olduğunu, yani pozitif ve negatif sınıfları iyi ayırt edebildiğini gösterir. SVM modeli ise 0.87 AUC değerine sahiptir, bu da diğer modellere kıyasla biraz daha düşük bir sınıflandırma performansına sahip olduğunu gösterir. Ancak genel olarak, SVM modeli de kabul edilebilir düzeyde bir performans sergilemektedir. Sonuç olarak, XGBoost, Rastgele Orman ve Gradient Boosting modelleri daha iyi bir genel sınıflandırma performansı gösterirken, SVM bu modellerin biraz gerisinde kalmaktadır. Bu üç modelin AUC değerlerinin 0.89 olması, sınıflandırmada güvenilir sonuçlar elde edilebileceğini gösterir. Buradaki AUC değerleri, model doğruluğunun ve sınıflandırma performansının genel bir göstergesidir. AUC değeri, modelin sınıfları ayırt etme kabiliyetini gösterir ve bu değerin yüksek olması, modelin pozitif ve negatif sınıfları doğru bir şekilde sınıflandırmada başarılı olduğunu gösterir.

7. Sonuç

Çalışmada, 2022 yılı Türkiye Çocuk Araştırması mikro veri seti kullanılarak çocukların cep telefonu sahipliği tahmin edilmiş ve cep telefonu sahipliğini etkileyen faktörler makine öğrenimi modelleri ile analiz edilmiştir. Elde edilen bulgular, yaş ve internet erişimi değişkenlerinin cep telefonu edinimini önemli derecede etkilediğini göstermiştir. Makine öğrenim modelleri arasında XGBoost ve Gradient Boosting, güçlü doğruluk ve genelleme kabiliyetleri ile ön plana çıkmıştır. Analizlerde kullanılan mikro

veri seti, çocukların yaşam koşulları ile ilgili detaylı bilgiler sunmaktadır. Mikro veri setinden elde edilen bulgular ve sonuçlar, politikacılar ve yöneticiler için veri odaklı bir rehber sağlamaktadır.

Pandemi dönemiyle birlikte dijital teknolojilerin ve internetin önemi tüm dünyada çarpıcı biçimde artmıştır. İnternet erişimi, eğitimden kamu hizmetlerine kadar pek çok alanda temel bir ihtiyaç haline gelmiştir. Ancak düşük sosyoekonomik düzeydeki ailelerin çocukları ile kırsal bölgelerde yaşayan çocuklar, dijital imkânlara erişim konusunda ciddi eşitsizliklerle karşı karşıyadır. Dijital eşitsizliği gidermek ve kırsaldaki çocukların fırsatlara eşit erişimini sağlamak amacıyla bütüncül ve somut stratejiler geliştirilmelidir. Mobil operatörler ile iş birliği yapılarak nüfusu düşük yerleşim yerlerinde 4.5G kapsamasının tesisi, aradaki uçurumun kapanmasına yardımcı olacaktır. Araştırma ve analiz bulguları doğrultusunda ilgili yöneticilerin, çocukların teknoloji erişimini artırmak ve dijital eşitsizliği azaltmak için somut adımlar atmaları beklenmekte ve önerilmektedir. Bu bağlamda, düşük gelirli ailelerin çocuklarına yönelik "Teknoloji Destek Paketi" benzeri projeler geliştirilebilir. Önerilen destek paketinde; ücretsiz ya da düşük maliyetli internet erişiminin sağlanması, eğitim amaçlı tablet veya bilgisayar dağıtımı ve dijital okuryazarlık eğitimlerinin verilmesi gibi somut ve elle tutulur adımlar yer almalıdır. Ayriyeten, özel okul öğrencileri ile devlet okulu öğrencileri arasındaki teknolojiye erişim farkını azaltmak için devlet okullarında teknoloji altyapısının güçlendirilmesi ve dijital eğitim materyallerine erişimin genişletilmesi de önemlidir. Yaş gruplarına uygun teknoloji kullanma rehberlerinin oluşturulması hem ebeveynlerin hem de eğitimcilerin teknoloji kullanımını daha bilinçli yönetmesine olanak sağlayacaktır.

Elde edilen sonuçlar, daha önceden yapılmış olan çalışmalar ile uyumlu olarak, dijital eşitsizliğin çocuk refahı üzerindeki etkisine dikkati çekmektedir. Örneğin, Zilyas & Yılmaz (2023), eğitimde başarıyı etkileyen faktörler arasında sosyoekonomik durumun önemli olduğunu belirtmiştir. Benzer şekilde, bu çalışma, özel okul öğrencilerinin cep telefonu sahipliğinin devlet okulu öğrencilerine göre daha yüksek olduğunu göstermiştir. Analizler sonucu ortaya çıkan bu fark; yalnızca teknolojik erişim değil, aynı zamanda eğitim altyapısındaki eşitsizliklere de işaret etmektedir.

Çalışma sonucu, makine öğrenimi modellerinin, çocukların cep telefonu sahipliğini etkileyen faktörlerin analizi için güçlü bir araç olduğunu göstermiştir. RandomForest, XGBoost, Gradient Boosting ve SVM modellerinin performansları karşılaştırılmış ve özellikle XGBoost ile Gradient Boosting modellerinin yüksek doğruluk (%81) ve AUC değeri (%89) ile öne plana çıktığı gözlemlenmiştir. Makine modelleri, veri setindeki değişkenlerin önem sırasını belirlemede başarılı sonuçlar sunmuştur. Bu durum, makine öğrenimi modellerinin büyük veri setlerinden anlamlı sonuçlar çıkarma ve bu sonuçları karar destek süreçlerinde kullanılabilir hale getirme yeteneğini vurgulamaktadır. Modellerin farklı metriklerle değerlendirilmesi, uygulama alanına göre hangi modelin daha uygun olabileceğine dair rehberlik sağlamaktadır. Örneğin, yanlış pozitif sınıflamaların azaltılmasının kritik olduğu durumlarda, XGBoost ve Gradient Boosting gibi dengeli performans gösteren modeller tercih edilebilirken, pozitif sınıfların yakalanmasında ise SVM gibi modeller daha avantajlı olabilmektedir.

Akademisyen ve araştırmacıların ileri dönemlerde yapacakları benzer çalışmalarda, farklı makine öğrenim modelleri ile analizler yapabilirler ve farklı veri setleri ile çalışabilirler. Ayrıca, hibrit modeller ve derin öğrenme algoritmaları kullanılarak daha detaylı ve kapsamlı tahmin çalışmaları geliştirilebilir. Çalışma, makine öğrenimi modellerinin sosyal bilimlerdeki potansiyelini göstererek, çocuklar ile ilgili stratejilerin geliştirilmesine önemli katkılar sağlamaktadır.

Kaynakça

Abar, H. (2020). XGBoost ve Mars yöntemleriyle altın fiyatlarının kestirimi. *EKEV Akademi Dergisi, 83*, 427-446.

Akbulut, S., & Adem, K. (2023). Derin öğrenme ve makine öğrenmesi yöntemleri kullanılarak gelişmekte olan ülkelerin finansal enstrümanlarının etkileşimi ile BIST 100 tahmini. *Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi*, *12*(1), 52-63. https://doi.org/10.28948/ngumuh.1131191

Akın, M. H. (2023). Türkiye'de çocuk sosyalleşmesinin bazı görünümleri. *Güncel Sosyoloji, 1*(1), 1-24. https://guncelsosyoloji.com/content/5-sayilar/1-1/1-turkiyede-cocuk-sosyallesmesinin-bazi-gorunumleri/1.makale.pdf

Aslan, K. (2025). Yapay zekâ makine öğrenmesi ve veri bilimi kursu, sınıfta yapılan örnekler ve özet notlar. C ve Sistem Programcıları Derneği, İstanbul.

Ayhan, S., & Erdoğmuş, Ş. (2014). Destek vektör makineleriyle sınıflandırma problemlerinin çözümü için çekirdek fonksiyonu seçimi. *Eskişehir Osmangazi Üniversitesi İktisadi ve İdari Bilimler Dergisi,* 9(1), 175-201.

Bae, C. Y., Im, Y., Lee, J., Park, C., Kim, M., Kwon, H. U., & Kim, J. (2021). Comparison of biological age prediction models using clinical biomarkers commonly measured in clinical practice settings: AI techniques vs. traditional statistical methods. *Frontiers in Analytical Science*, *1*. https://doi.org/10.3389/frans.2021.709589

Bazilevych, K., Kyrylenko, O., Parfenyuk, Y., Krivtsov, S., Meniailov, I., Kuznietcova, V., & Chumachenko, D. (2023). Comparative analysis of the machine learning models determining COVID-19 patient risk levels. *Radioelectronic and Computer Systems*, (3), 5-17. https://doi.org/10.32620/reks.2023.3.01

Cao, T., She, D., Zhang, X., & Yang, Z. (2022). Understanding the influencing factors and mechanisms (land use changes and check dams) controlling changes in the soil organic carbon of typical loess watersheds in china. *Land Degradation &Amp; Development*, *33*(16), 3150-3162. https://doi.org/10.1002/ldr.4378

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*. https://doi.org/10.1145/2939672.2939785

El Naqa, I., & Murphy, M. J. (2015). What is machine learning?. *Machine Learning in Radiation Oncology* (3-11). Springer. https://doi.org/10.1007/978-3-319-18305-3_1

Erbay Mermer, Ş. (2023). Eğitimde dijital dönüşüm: Makine öğrenmesi. *Dijital Eğitim Araştırmaları Dergisi*, *12*(3), 135-153. https://doi.org/10.58830/ozgur.pub303.c1379

Eriş Dereli, B. (2021). Çocuk istihdamını etkileyen faktörler. *İstanbul Ticaret Üniversitesi Sosyal Bilimler Dergisi, 20*(42), 1505-1519. doi:10.46928/iticusbe.974781

Eşidir, K. A. (2025a). Makine öğrenimi modelleri ile yetişkin eğitimi analizi: Modellerin karşılaştırmalı performansı. *Elektronik Sosyal Bilimler Dergisi, 24*(2), 946-964. https://doi.org/10.17755/esosder.1589887

Eşidir, K. A. (2025b). Kanatlı sektöründe yapay zekâ uygulamaları: Etlik piliç yemi fiyat tahmini. *Journal of History School,* 74, 269-290. http://dx.doi.org/10.29228/joh.78735

Eşidir, K. A. (2025c). TÜİK mikro verileri ile çocuk işgücü tahmini: Makine öğrenimi modellerinin performans analizleri. *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, 37(1), 453-466. https://doi.org/10.35234/fumbd.1607609

Ge, X., Ding, J., Jin, X., Wang, J., Chen, X., Li, X., … & Xie, B. (2021). Estimating agricultural soil moisture content through uav-based hyperspectral images in the arid region. *Remote Sensing, 13*(8), 1562. https://doi.org/10.3390/rs13081562

Gür, Y. E. (2024a). Development and application of machine learning models in US consumer price index forecasting: Analysis of a hybrid approach. *Data Science in Finance and Economics*, *4*(4), 469-513. https://doi.org/10.3934/DSFE.2024020

Gür, Y. E. (2024b). Forecasting the euro exchange rate using deep learning algorithms and machine learning algorithms. *İstanbul Ticaret Üniversitesi Sosyal Bilimler Dergisi*, *23*(49), 1435-1456. https://doi.org/10.46928/iticusbe.1379268

Gür, Y. E. (2024c). Stock price forecasting using machine learning and deep learning algorithms: A case study for the aviation industry. *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, *36*(1), 25-34. https://doi.org/10.35234/fumbd.1357613

Ji, H. (2023). Robustness analysis on stock market prediction method. *Highlights in Business, Economics and Management*, *21*, 791-801. https://doi.org/10.54097/hbem.v21i.14763

Liang, W., Luo, S., Zhao, G., & Wu, H. (2020). Predicting hard rock pillar stability using gbdt, xgboost, and lightgbm algorithms. *Mathematics,* 8(5), 765. https://doi.org/10.3390/math8050765

Oguine, O. C., & Oguine, M. B. (2021). Comparative analysis and forecasting on the death rate of COVID-19 patients in Nigeria using random forest and multinomial Bayesian epidemiological models. *Journal of Clinical Case Studies, Reviews & Reports*, 1-7. https://doi.org/10.47363/jccsr/2021(3)182

Ou, R. (2020). Out-of-core GPU gradient boosting. https://doi.org/10.48550/arxiv.2005.09148

Özdemir, N. E., Akyiğit Albayrak, E., & Korkutan, M. (2024). Türkiye'de çocuk işçiliği: Sağlık ve sosyal yaşam koşulları bağlamında bir değerlendirme. *Sosyal Bilimler Akademi Dergisi*, 7(1), 44-66. https://doi.org/10.38004/sobad.1413138

Pakarinen, O., Karsikas, M., Reito, A., Lainiala, O., Neuvonen, P., & Eskelinen, A. (2022). Prediction model for an early revision for dislocation after primary total hip arthroplasty. *PLOS ONE*, 17(9), e0274384. https://doi.org/10.1371/journal.pone.0274384

Paslı Yeniay, K. (2023). Makine öğrenmesinin çocuk ruh sağlığına etkisi. *Akademik Sosyal Araştırmalar Dergisi*, 11(140), 269-280. https://doi.org/10.29228/ASOS.69003

Speer, A. B. (2021). Empirical attrition modelling and discrimination: Balancing validity and group differences. *Human Resource Management Journal*, 34(1), 1-19. https://doi.org/10.1111/1748-8583.12355

Suthaharan, S. (2014). Big data classification: Problems and challenges in network intrusion prediction with machine learning. *ACM SIGMETRICS Performance Evaluation Review*, *41*(4), 70-73.

Tosunoğlu, E., Yılmaz, R., Özeren, E., Sağlam, Z. (2021). Eğitimde makine öğrenmesi: Araştırmalardaki güncel eğilimler üzerine inceleme. *Ahmet Keleşoğlu Eğitim Fakültesi Dergisi, 3*(2), 178-199.

Türkiye Çocuk Araştırması Mikro Veri Seti. (2022). Yayım Tarihi: Temmuz 2023, Yayın No: 4699, Türkiye İstatistik Kurumu Bilgi Dağıtım Grup Başkanlığı. ISBN: 978-625-8368-39-0, Ankara.

Türkiye İstatistik Kurumu. (2023). *Türkiye çocuk araştırması 2022.* Aile ve Sosyal Hizmetler Bakanlığı ve UNICEF. Ankara: Türkiye İstatistik Kurumu Yayınları. ISBN: 978-625-8368-33-8. Erişim adresi: https://www.unicef.org/turkiye/media/17591/file/2022%20T%C3%BCrkiye%20%C3%87ocuk%20Ara%C5%9Ft%C4%B1rmas%C4%B1%20.pdf

Wang, L., Wang, X., Chen, A., Jin, X., & Che, H. (2020). Prediction of type 2 diabetes risk and its effect evaluation based on the xgboost model. *Healthcare, 8*(3), 247. https://doi.org/10.3390/healthcare8030247

Wen, J., Zhang, Y., Yang, G., He, Z., & Zhang, W. (2019). Path loss prediction based on machine learning methods for aircraft cabin environments. *IEEE Access, 7*, 159251-159261. https://doi.org/10.1109/access.2019.2950634

Wu, Y. (2023). Job embeddedness review: Presentation, measurement and development. *Advances in Economics, Management and Political Sciences*, *47*(1), 169-174. https://doi.org/10.54254/2754-1169/47/20230393

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *Computer Vision – ECCV 2014*, 818-833. https://doi.org/10.1007/978-3-319-10590-1_53

Zhang, Z., Xin, Q., & Li, W. (2021). Machine learning-based modeling of vegetation leaf area index and gross primary productivity across north america and comparison with a process-based model. *Journal of Advances in Modeling Earth Systems, 13*(10). https://doi.org/10.1029/2021ms002802

Zhu, X., Sawhney, R., & Upreti, G. (2016). Determinates of employee voluntary turnover and forecasting in departments: A case study. *Studies in Engineering and Technology*, *3*(1), 64-73. https://doi.org/10.11114/set.v3i1.1635

Zilyas, D., & Yılmaz, A. (2023). Makine öğrenmesi yöntemleri ile eğitim başarısının tahmini modeli. *DÜMF MD, 14*(3), 437–447, doi: 10.24012/dumf.1322273.

Çıkar Çatışması: Yoktur.

Finansal Destek: Yoktur.

Etik Onay: Yoktur.

Yazar Katkısı: Kamil Abdullah EŞİDİR (%100)

Conflict of Interest: None.

Funding: None.

Ethical Approval: None.

Author Contributions: Kamil Abdullah EŞİDİR (100%)

Forecasting Cell Phone Ownership among Children with TurkStat Micro Data: Comparative Performance of Machine Learning Models

Kamil Abdullah EŞİDİR

Extended Abstract

This study investigates the determinants of mobile phone ownership among children in Turkey by applying supervised machine learning models to nationally representative survey microdata from the 2022 Turkey Child Survey conducted by TurkStat in collaboration with the Ministry of Family and Social Services and UNICEF. As digital technologies become increasingly ubiquitous in contemporary societies, early access to mobile phones has emerged as a key factor shaping children's engagement with education, communication, and digital participation. In this context, disparities in mobile phone ownership reflect broader patterns of digital inequality and socio-economic stratification. The aim of this research is to identify the principal predictors of mobile phone ownership and to evaluate the comparative performance of various machine learning algorithms in modeling this binary outcome using complex, high-dimensional social data.

The dataset comprises 9,869 observations of children aged 0 to 17. The binary dependent variable captures mobile phone ownership status. Independent variables include age, gender, type of school (public or private), availability of internet access at home, presence of a personal room and study desk, household computer ownership, receipt of government social assistance, and the adequacy of the study environment. These variables reflect multi-dimensional aspects of children's lives, encompassing demographic characteristics, educational conditions, digital infrastructure, and overall well-being.

In order to model and predict mobile phone ownership, four machine learning algorithms were implemented: Random Forest, Extreme Gradient Boosting (XGBoost), Gradient Boosting, and Support Vector Machines (SVM). The selection of these models is based on their demonstrated capacity to manage non-linear interactions, high-dimensional feature spaces, and imbalanced class distributions—common characteristics of real-world social data. Ensemble methods such as Random Forest and XGBoost are particularly well-suited for classification tasks involving correlated explanatory variables, while SVM is known for its robust performance in separating classes within complex multidimensional feature spaces. The dataset was partitioned into training (80%) and testing (20%) sets, and standard preprocessing techniques were applied, including the imputation of missing values, label encoding for categorical variables, and normalization of numerical features using standard scaling. Hyperparameter optimization was performed via grid search cross-validation.

Model performance was evaluated using a range of classification metrics: Accuracy, Precision, Recall, F1 Score, and ROC AUC. These complementary indicators provide a comprehensive assessment of classification effectiveness, particularly in the presence of class imbalance or unequal misclassification costs. Among the tested models, the Random Forest classifier yielded the highest accuracy (0.820), indicating strong predictive reliability. Both XGBoost and Gradient Boosting achieved the highest ROC AUC scores (0.893), suggesting superior discriminatory power. Although the SVM model showed a slightly lower AUC (0.873), it performed notably well in identifying true positives—an essential quality when predicting access to socially significant resources such as mobile technologies.

Variable importance analysis revealed that age is the most influential predictor across all models, with mobile phone ownership rates increasing markedly after the age of 13. This age threshold corresponds with early adolescence, a period characterized by growing autonomy, intensified peer interaction, and heightened educational demands—often mediated through mobile technologies. The availability of internet access at home emerged as the second most impactful variable, indicating that mobile phone ownership is closely intertwined with broader digital infrastructure at the household level. Other notable predictors included school type, where students attending private schools exhibited a significantly higher probability of owning a mobile phone, reflecting socio-economic disparities in digital access. Additional factors, such as having a private room or a computer at home, contributed to model predictions but carried relatively lower weights. Gender demonstrated a modest yet statistically significant influence in some models.

This study underscores the methodological advantages of utilizing machine learning techniques in social science research. Traditional statistical approaches, such as logistic regression, are often constrained by assumptions of linearity

and independence among variables, which may not hold in high-dimensional and interrelated data structures. In contrast, machine learning models are data-driven and non-parametric, enabling the identification of latent patterns, non-linear associations, and complex variable interactions. Ensemble methods, particularly Random Forest and XGBoost, exhibited high levels of model stability and interpretability, making them especially suitable for socially relevant and policy-oriented research.

From a policy perspective, the findings highlight critical disparities in mobile phone access that are shaped by structural factors such as digital infrastructure, school type, and household socio-economic conditions. These inequalities reflect a deeper digital divide with implications for children's educational opportunities, social inclusion, and technological participation. As mobile phones become increasingly integrated into educational systems and public service delivery, a lack of access may deepen social exclusion and restrict upward mobility for disadvantaged groups.

The study's policy recommendations emphasize the necessity of targeted interventions that address both infrastructural and educational inequalities. Expanding publicly funded broadband initiatives in low-income and underserved regions can enhance baseline digital access. Incorporating mobile device provision into national social welfare programs can ensure that children from economically vulnerable households are not excluded from digital opportunities. Furthermore, strengthening ICT infrastructure in public schools and embedding digital literacy into curricula can help bridge the school-type gap in mobile ownership. National campaigns that promote age-appropriate and responsible mobile phone use can further support the safe and constructive integration of mobile technologies into children's daily lives.

This study contributes to the growing field of computational social science by demonstrating the applicability of machine learning techniques to nationally representative social survey data. The analytical rigor, combined with interpretive depth achieved through feature importance analysis and model comparison, provides a replicable framework for future studies on digital access and social equity. Moreover, the use of machine learning facilitates real-time policy monitoring and the predictive evaluation of interventions, opening new avenues for data-driven governance.

Future research could build upon these findings by incorporating longitudinal data to investigate causal mechanisms and temporal patterns in mobile technology adoption. Enriching the model with additional socio-economic variables—such as parental education, household income, and geographic indicators—would further enhance explanatory power. Comparative studies across different regions or countries could reveal contextual variations in digital inequality. Additionally, integrating qualitative methods—such as interviews with children and parents—would provide deeper insight into the sociocultural dimensions of mobile phone usage.

In addition to its empirical insights, this study offers a substantial methodological contribution by showcasing the applicability of ensemble-based machine learning algorithms in the context of nationally representative and multidimensional social datasets. Traditional parametric techniques often struggle to accommodate the complex, non-linear, and interactive nature of socio-demographic variables. In contrast, the models applied here—especially Random Forest and XGBoost—demonstrate a high degree of adaptability, predictive reliability, and interpretability, making them particularly suited for policy-relevant social science research. The incorporation of model explainability techniques, including variable importance measures and confusion matrices, enhances the transparency and accountability of the analytical process, enabling stakeholders to better understand the mechanisms underlying mobile phone ownership among children. Furthermore, the study provides a transferable analytical framework that can be replicated in similar investigations across different socio-political contexts. It sets the stage for future research to adopt more sophisticated machine learning techniques, including deep learning or hybrid modeling, particularly in the context of longitudinal and geospatial data. Ultimately, the findings underscore the growing utility of computational methods not only in predictive analytics but also in unveiling structural inequalities and informing targeted interventions within the broader agenda of digital inclusion and child welfare policy.

In conclusion, this study presents a comprehensive, data-driven investigation into mobile phone ownership among children in Turkey. It demonstrates the efficacy of machine learning models in identifying key socio-demographic and infrastructural determinants of technological access while also offering a policy-relevant framework to address digital inequalities. The applied models not only achieved high predictive performance but also yielded interpretable insights

that can inform targeted social interventions. As the digital transformation of childhood and education accelerates, ensuring equitable access to essential technologies such as mobile phones must remain a central objective of inclusive policy design. This research delivers both empirical evidence and methodological innovation to support national and international efforts in promoting digital inclusion and advancing child welfare in the digital age. Bridging the digital divide in childhood is not only about ensuring access to technology—it is about securing equitable educational and social outcomes in an increasingly digital world.