©2025 Published in 11th International Symposium on Intelligent Technologies in Engineering and Science 16-18 May 2025 (ISITES2025 Diyarbakır - Turkey)

Veriden Habersiz Regresyon Modelleri Eğitilmesi ve Karşılaştırılması

[1*]Didem CİVELEK , [1]Tolga BÜYÜKTANIR, [1]Burak AKTÜRK
[1]Agra Fintach Yazılım Çözümleri A.Ş., İstanbul, Türkiye

Özet

Gelişen teknoloji ile beraber kişisel verilerin korunması, gizlilik ve güvenlik konuları oldukça önemli bir noktaya gelmiştir. Firmalar kendi geliştirmedikleri ve başka firmalardan alacakları çözümlerde veri gizliliği konusunda büyük endişeler yaşamaktadır. Özellikle finansal veriye dokunan firmalar bu konuda ekstra hassasiyet göstermektedir. Kaygıların giderilmesi için şifrelenmiş verilerden analitikler çıkarılması önemli hale gelmiştir. Bu noktadan yola çıkılarak, çalışma kapsamında homomorfik şifrelenmiş veriler üzerinde yapılan analitik çalışmaların şifresiz veriler ile yapılan sonuçlarla karşılaştırılması hedeflenmiştir. Bu çalışmada, makine öğrenmesi algoritmaları kullanılarak bir veri kümesi üzerinde tahminleme çalışmaları gerçekleştirilmiştir. Ayrıca, aynı veri kümesi Zama AI Concrete ML kütüphanesi ile homomorfik şifrelenmiş ve makine öğrenmesi modelleri şifreli veriler üzerinde çalıştırılmıştır. Makine öğrenmesi modellerinin homomorfik şifreli veriler ile eğitilmesi eğitim süresini arttırsa da, elde edilen sonuçlar şifresiz verilerle yapılan analizlerle büyük ölçüde uyumludur. Bu durum, gizliliği koruyan makine öğrenmesi tekniklerinin uygulanabilirliğini göstermektedir.

Anahtar Kelimeler: Homomorfik şifreleme, şifreli veri ile analitik, veriden habersiz öğrenme.

Abstract

With advancing technology, protecting personal data, privacy, and security has become a critical concern. Companies face significant concerns about data privacy when using solutions developed by external providers. This issue is particularly important for businesses handling financial data, which requires extra caution. Extracting analytics from encrypted data has become essential to address these concerns. In this study, machine learning algorithms were applied to a dataset for predictive analysis. Additionally, the same dataset was homomorphically encrypted using the Zama AI Concrete ML library, and machine learning models were trained on encrypted data. While training machine learning models on homomorphically encrypted data increased training time, the results remained highly comparable to those obtained from unencrypted data. This demonstrates the feasibility of privacy-preserving machine learning techniques.

*Sorumlu Yazar: Adres: Agra Fintech Yazılım Çözümleri A.Ş., Acarkent Mahallesi, 56. Sokak, B761, Beykoz, Türkiye. E-mail adres: didem.civelek@agrafintech.com, Telefon: +5378247179

1. Giriş

Veri odaklı sistemler geliştirmek ve bu veriler üzerinden çıkarım yapmak, günümüzün rekabetçi şartlarında önemli bir konudur. Yapay zeka ve makine öğrenmesi tabanlı modellerin başarısı, kullanılan veri miktarı ve kalitesine bağlıdır. Verilerin hassasiyetinin artmasıyla birlikte, gizlilik odaklı hesaplama tekniklerine duyulan ihtiyaç da giderek artmaktadır. Özellikle finansal ve kişisel veriler gibi yüksek derecede gizlilik gerektiren durumlarda, bu verileri doğrudan işlemek veya üçüncü taraflarla paylaşmak çoğu zaman mümkün olmamaktadır. Bu durum, söz konusu verileri şirket dışına çıkararak işlemek ya da üçüncü taraf hizmetlere sunmak isteyen kurumlar için engel teşkil etmektedir. Gelişen sistemlere ayak uydurmak ve müşteriyi anlayabilme ihtiyacı bu veriler üzerinde analizler yapılmasını zorunlu kılmaktadır. Bu noktada şifreli veriler üzerinden analiz yapabilecek yöntemler geliştirilme ihtiyacı ortaya çıkmıştır. Şifreli veriler üzerinde doğrudan makine öğrenmesi algoritmalarını çalıştırabilmeyi mümkün kılan Tam Homomorfik Şifreleme (FHE) yaklaşımları önem kazanmaktadır [1]. Geliştirilen bazı sistemler sadece şifreleme işlemleri yaparken bazı sistemler hem şifreleme hem de şifrelenmiş verilerden makine öğrenmesi algoritmalarının çalıştırılmasına olanak sağlamaktadır [2]. Tam homomorfik şifreleme ile; verilerin şifresi çözülmeden belirli matematiksel işlemler yapılabilmekte, makine öğrenmesi modellerinin eğitimi ya da çıkarım aşaması gerçekleştirilebilmektedir. Bu yöntemler veri sahiplerinin, gizlilik ve güvenlik endişelerini asgari düzeye indirebilirken; hizmet sağlayıcı da modeli eğitmek veya tahmin üretmek için gerekli işlemleri şifreli veriler üzerinde gerçekleştirebilmektedir.
Finansal veriden analizler yapabilmek için bu sistemlerin kullanılması, sistemlerin performans ölçümlerinin yapılması ve çözüm doğruluklarının test edilmesi amacıyla bu çalışma gerçekleştirilmiştir. Verilerin hassasiyetinin artmasıyla birlikte, gizlilik odaklı hesaplama tekniklerine duyulan ihtiyaç da giderek artmaktadır. Son yıllarda bu şifreli veriler üzerinden makine öğrenmesi yapılarak çeşitli çalışmalar gerçekleştirilmiştir.

Stoian ve arkadaşları tarafından gerçekleştirilen bir çalışmada, iki farklı bilgisayarlı görme görevi için derin sinir ağı (DNN) mimarileri sunulmuş ve Concrete-ML kütüphanesi kullanılarak test edilmiştir [3]. Frery ve arkadaşlarının diğer bir çalışması, karar ağaçları, rastgele ormanlar ve gradient destekli ağaçlar gibi makine öğrenmesi modellerine tam homomorfik şifreleme uygulanmış ve bu yöntemlerin şifrelenmiş tablo verileri üzerinde gizliliği koruyarak yüksek doğrulukla çalıştığı gözlemlenmiştir [4]. Attaullah ve arkadaşları tarafından yürütülen bir çalışma, XGBoost, Rastgele Orman ve Karar Sınıflandırıcı modellerinin tam homomorfik şifreleme kullanılarak akıllı ev veri setleri üzerindeki performanslarını incelemiştir [5]. Podschwadt ve arkadaşları tarafından yapılan bir çalışmada derin öğrenme mimarilerinde tam homomorfik şifreleme kullanımını ele almıştır [6]. Behera ve arkadaşları tarafından gerçekleştirilen bir çalışma, tam homomorfik şifreleme kullanarak doğrusal regresyon algoritmalarının şifrelenmiş veriler üzerinde nasıl çalıştığını incelemiştir [7]. Ameur ve arkadaşları, büyük veri teknolojileri ve homomorfik şifreleme entegrasyonuna odaklanarak gizlilik koruma teknolojilerinin makine öğrenmesi ile nasıl optimize edilebileceğini araştırmıştır [8]. Chillotti ve arkadaşları, homomorfik şifreleme kullanarak derin sinir ağlarının güvenli bir şekilde çıkarım yapmasını sağlayan yeni bir yaklaşım sunmuştur [9]. Li ve arkadaşları, homomorfik şifreleme tabanlı güvenli ortak makine öğrenmesi sistemi olan HomoPAI'yi önermiş ve paralel homomorfik şifreleme hesaplamalarının verimliliğini göstermiştir [10].

Yapılan çalışmalar, tam homomorfik şifreleme'nin gizlilik gerektiren pek çok senaryoda makine öğrenmesi yöntemleriyle birlikte başarıyla uygulanabildiğini ortaya koymaktadır. İşlem hızı ve depolama maliyeti gibi konularda çalışmalar devam etmektedir. Bu alandaki teknolojik ilerlemeler ve paralel hesaplama yaklaşımlarıyla birlikte tam homomorfik şifreleme tabanlı makine öğrenmesi yöntemlerinin giderek yaygınlaşacağı, özellikle finans, sağlık gibi paylaşılmaması gereken veri barındıran sektörlerde önemli sorunlara çözüm sunacağı düşünülmektedir.

2. Araştırma Yöntemleri

Bu çalışmada veri mahremiyeti probleminin ortadan kaldırılması, makine öğrenmesi modellerinin bu veriler ile eğitilmesi ve açık veriler ile eğitilmiş modellerle kıyas edilerek modellerin etkinliklerinin kıyaslanması sağlanmıştır.

2.1. Veri Kümesi

Bu çalışmada iki veri kümesi kullanılmıştır.

Birinci veri kümesi, 2015-2023 yılları arasındaki Dünya Mutluluk Endeksi ve ülkelerin çeşitli ekonomik göstergelerini içermektedir. Bu veri kümesi, ülkelerin mutluluk skorlarını belirleyen faktörlerle birlikte enflasyon verilerini bir araya getirerek kapsamlı bir analiz imkanı sunmaktadır. Veri kümesinde her bir ülke ve yıl için mutluluk sıralaması ve mutluluk skoru yer almakta olup, ekonomik ve sosyal göstergeler detaylı şekilde sunulmaktadır. Bu göstergeler arasında kişi başına düşen gayri safi yurtiçi hasıla (GSYH), sosyal destek, doğumda beklenen sağlıklı yaşam süresi, hayat seçimleri yapma özgürlüğü, cömertlik düzeyi ve yolsuzluk algısı gibi bireylerin yaşam kalitesini etkileyen değişkenler bulunmaktadır. Ayrıca, ülkelerin ekonomik durumu ve enflasyon eğilimlerini anlamaya yönelik çeşitli makroekonomik göstergeler veri kümesine dahil edilmiştir. Bunlar arasında enerji tüketici fiyat enflasyonu, gıda tüketici fiyat enflasyonu, gayri safi yurt içi hasıla deflatörü büyüme oranı, manşet tüketici fiyat enflasyonu, resmi çekirdek tüketici fiyat enflasyonu ve üretici fiyat enflasyonu gibi kritik ekonomik göstergeler bulunmaktadır. Veri kümesinde her ülkenin coğrafi bölgesi (kıta) bilgisi de yer almaktadır. Bu sayede, bölgesel farklılıklar analiz edilerek mutluluk düzeyi ile ekonomik göstergeler arasındaki ilişkiler kıtasal bazda da değerlendirilebilmektedir.

İkinci veri kümesi, 2000-2015 yılları arasındaki dünya genelinde yaşam beklentisi verileri ile ülkelerin çeşitli sağlık ve ekonomik göstergelerini içermektedir. Bu veri kümesi, ülkelerin yaşam süresini belirleyen faktörlerle birlikte sağlık harcamaları, ölüm oranları ve makroekonomik verileri bir araya getirmektedir. Veri kümesinde her bir ülke ve yıl için yaşam beklentisi değerleri bulunmakta olup, sağlık, ekonomik ve sosyal göstergeler yer almaktadır. Bu göstergeler arasında erişkin ölüm oranı, bebek ölüm oranı, kişi başına düşen gayri safi yurtiçi hasıla, sağlık harcamaları, aşılama oranları ve eğitim düzeyi gibi bireylerin yaşam süresini etkileyen değişkenler bulunmaktadır. Ülkelerin sağlık durumu ve ekonomik kalkınma eğilimlerini anlamaya yönelik çeşitli makroekonomik göstergeler veri kümesine dahil edilmiştir. Bunlar arasında sağlık harcamalarının GSYH içindeki payı, toplam sağlık harcamaları, aşılama oranları (Hepatit B, Polio, Difteri), erişkin ve çocukluk çağı ölüm oranları, beklenen eğitim süresi, HIV/AIDS oranları, nüfus ve demografik göstergeler ile Beden Kitle İndeksi (BMI) gibi kritik değişkenler yer almaktadır.

Veri kümesinde her ülkenin gelişmişlik durumu (gelişmiş, gelişmekte olan) da bulunmaktadır.

2.2. Veri Önişleme

Veri kümesinde yer alan eksik veriler üzerinde detaylı ön işleme adımları uygulanmıştır. Birinci veri kümesinde, özellikle enerji enflasyonu, gıda enflasyonu ve çekirdek enflasyon gibi sütunlarda eksik gözlemler tespit edilerek, bu değerler ilgili değişkenlerin ortanca (medyan) değerleri ile doldurulmuştur. Ortanca kullanımı, uç değerlerin etkisini minimize ederek daha dengeli bir veri dağılımı sağlamaktadır. İkinci veri kümesinde ise GSYH, nüfus, alkol tüketimi, toplam sağlık harcamaları ve aşılama oranları gibi sütunlardaki eksik gözlemler belirlenerek benzer şekilde ortanca değerler kullanılmıştır. Eksik verilerin yoğun olduğu bazı değişkenlerde, verilerin genel dağılımını bozmamak adına en uygun doldurma yöntemleri uygulanmış ve veri bütünlüğü korunarak analizlerin güvenilirliği artırılmıştır.
İki veri kümesinde de eksik verilerin doldurulmasının ardından, veri kümelerindeki tüm sayısal değişkenler standart bir ölçüye getirilmiştir. StandardScaler yöntemi kullanılarak tüm sayısal değişkenler ortalaması 0 ve standart sapması 1 olacak şekilde ölçeklendirilmiştir. Bu işlem, değişkenler arasında ölçek farklılıklarını ortadan kaldırarak makine öğrenmesi modellerinin daha sağlıklı ve kararlı çalışmasını sağlamaktadır. Böylece, tüm değişkenlerin eşit katkıda bulunabileceği adil bir veri yapısı oluşturulmuştur.

2.3. Modelleme

Birinci veri kümesinde Yolsuzluk Algısı etiket olarak belirlenmiş ve veri setindeki diğer değişkenler öznitelik olarak kullanılmıştır. İkinci veri kümesinde Yaşam Beklentisi etiket olarak belirlenmiş ve veri setindeki diğer değişkenler öznitelik olarak kullanılmıştır. İlk aşamada, şifrelenmemiş veriler üzerinde çeşitli makine öğrenmesi algoritmaları kullanılarak tahminleme çalışmaları gerçekleştirilmiştir. Bu kapsamda, Random Forest, Linear Regression, Decision Tree ve XGBoost algoritmaları uygulanmış ve modellerin performansı değerlendirilmiştir.

Daha sonraki aşamada, Concrete ML kütüphanesi kullanılarak veri seti tam homomorfik şifreleme (Fully Homomorphic Encryption - FHE) yöntemiyle şifrelenmiştir. Şifreleme işleminin ardından, Concrete ML aracında yer alan aynı makine öğrenmesi algoritmaları şifrelenmiş veriler üzerinde çalıştırılmış ve modellerin performansı analiz edilmiştir. Bu süreçte, verilerin mahremiyetini koruyarak doğrudan şifreli haliyle işlenmesi sağlanmış, böylece gizliliği koruyan makine öğrenmesi yöntemlerinin uygulanabilirliği değerlendirilmiştir.

2.4. Performans Değerlendirme

Her bir algoritmanın tahmin performansı, Ortalama Kare Hatası (MSE) ve R^2 skoru kullanılarak değerlendirilmiş, ayrıca her modelin tahmin süresi hesaplanmıştır. Bu ölçümler, modellerin doğruluk seviyelerini ve işlem verimliliğini karşılaştırmak amacıyla analiz edilmiştir. Elde edilen sonuçlar, şifrelenmiş ve şifrelenmemiş veriler üzerinde çalışan modellerin performans farklarını incelemek için kullanılmıştır.

3. Deneysel Sonuçlar ve Bulgular

Bu bölümde, şifrelenmemiş ve şifreli iki veri kümesi üzerinde gerçekleştirilen makine öğrenmesi tahmin çalışmalarının sonuçları karşılaştırılmaktadır. Çalışmada, Concrete ML ve scikit-learn kütüphaneleri kullanılmış olup, tüm deneyler docker konteyner ortamında Python 3.10 sürümü ile gerçekleştirilmiştir.

Makine öğrenmesi modellerinin performansını değerlendirmek amacıyla Ortalama Kare Hatası (MSE) ve R^2 skoru kullanılmış, ayrıca her modelin tahmin süresi hesaplanmıştır. Elde edilen sonuçlar, şifrelenmiş ve şifrelenmemiş veriler üzerinde çalışan modellerin doğruluklarını ve işlem sürelerini karşılaştırarak şifreleme yönteminin etkisini analiz etmeyi amaçlamaktadır.

Tablo 1'de birinci veri seti ile şifrelenmemiş veriler üzerinde gerçekleştirilen tahmin çalışmaları sonucunda elde edilen performans ölçütlerini içermektedir.

Tablo 1. Birinci Veri Kümesi Şifrelenmemiş Veri Tahmin Sonuçları

	MSE	R^2
Random Forest	2,74	64,6
Linear Regression	5,8	25,2
Decision Tree	5,85	24,6
XGBoost	3,55	54,2

Tablo 2'de ikinci veri kümesi ile şifrelenmemiş veriler üzerinde gerçekleştirilen tahmin çalışmaları sonucunda elde edilen performans ölçütlerini içermektedir.

Tablo 2. İkinci Veri Kümesi Şifrelenmemiş Veri Tahmin Sonuçları

	MSE	R^2
Random Forest	6,11	91,4
Linear Regression	13,54	80,9
Decision Tree	8,60	87,9
XGBoost	4,54	93,6

Şifrelenmiş birinci veri kümesi üzerinde gerçekleştirilen tahminleme çalışmaları sonucunda elde edilen performans değerleri Tablo 3'te gösterilmektedir.

Tablo 3. Birinci Veri Kümesi Şifrelenmiş Veri Tahmin Sonuçları

	MSE	R^2
Random Forest	3,02	61,1
Linear Regression	5,81	25,0
Decision Tree	4,45	42,6
XGBoost	4,33	44,1

Şifrelenmiş ikinci veri kümesi üzerinde gerçekleştirilen tahminleme çalışmaları sonucunda elde edilen performans değerleri Tablo 4'te gösterilmektedir.

Tablo 4. İkinci Veri Kümesi Şifrelenmiş Veri Tahmin Sonuçları

	MSE	R^2
Random Forest	6,37	91,0
Linear Regression	13,77	80,6
Decision Tree	6,56	90,8
XGBoost	6,20	91,3

Birinci veri kümesi şifrelenmemiş ve şifrelenmiş veri seti üzerinde gerçekleştirilen tahminleme çalışmalarının çözüm sürelerine ilişkin performans değerleri Tablo 5'te gösterilmektedir.

Tablo 5. Birinci Veri Kümesi Şifreli ve Şifresiz Veri ile İşlem Süreleri Karşılaştırması

	Şifresiz Veri Çözüm Süresi (Saniye)	Şifreli Veri Çözüm Süresi Saniye)
Random Forest	0,0758	4698
Linear Regression	0,0129	0,84
Decision Tree	0,0049	2679
XGBoost	0,0908	3801

İkinci veri kümesi şifrelenmemiş ve şifrelenmiş veri seti üzerinde gerçekleştirilen tahminleme çalışmalarının çözüm sürelerine ilişkin performans değerleri Tablo 6'da gösterilmektedir.

Tablo 6. İkinci Veri Kümesi Şifreli ve Şifresiz Veri ile İşlem Süreleri Karşılaştırması

	Şifresiz Veri Çözüm Süresi (Saniye)	Şifreli Veri Çözüm Süresi Saniye)
Random Forest	0,2707	29061
Linear Regression	0,0341	1,7497
Decision Tree	0,0309	27229
XGBoost	0,378	18175

İki veri kümesinde de şifrelenmiş ve şifrelenmemiş veri üzerinde gerçekleştirilen tahminler incelendiğinde, Random Forest ve XGBoost modellerinin şifreli veriler üzerinde de yüksek doğruluk sağladığı görülmektedir. Birinci veri kümesinde Random Forest modelinin, MSE değerinde düşüklük ve R^2 skorunda yükseklik ile güçlü bir performans sergilemiştir, ikinci veri kümesinde ise XGBoost modelinin sonuçları ön plana çıkmaktadır.

Elde edilen sonuçlara göre, şifreleme işlemi nedeniyle model tahmin süreleri önemli ölçüde artmıştır. Şifrelenmiş veriler üzerinde gerçekleştirilen işlemler saatler sürebilirken, aynı işlemler şifresiz verilerle saniyeler içinde tamamlanabilmiştir. Bu durum, homomorfik şifreleme yönteminin güvenlik avantajlarına rağmen, işlem süresi açısından ek maliyet getirdiğini ortaya koymaktadır.

4. Sonuç

Bu çalışmada, tam homomorfik şifreleme yöntemi kullanılarak makine öğrenmesi algoritmalarının şifreli veriler üzerinde çalıştırılması sağlanmış ve elde edilen sonuçlar şifresiz veri ile elde edilen sonuçlarla karşılaştırılmıştır. Elde edilen bulgular, şifreleme sonrası tahmin performansının büyük ölçüde korunabildiğini, ancak işlem süresinde önemli bir artış yaşandığını göstermektedir. Özellikle Random Forest ve XGBoost algoritmalarının, şifreli verilerle yapılan tahminlerde de yüksek doğruluk sağladığı gözlemlenmiştir. Ancak, homomorfik şifreleme nedeniyle işlem süresinin büyük ölçüde arttığı ve bu durumun uygulama maliyetlerini artırabileceği anlaşılmıştır. Sonuç olarak, bu çalışma homomorfik şifreleme ile makine öğrenmesi modellerinin veri gizliliğini koruyarak başarılı tahminler üretebildiğini göstermektedir. Pratik kullanım açısından işlem süresini azaltmaya yönelik optimizasyon tekniklerinin geliştirilmesi gerekmektedir.

Referanslar

[1] C. Gentry, Fully homomorphic encryption using ideal lattices, in Proceedings of the forty-first annual ACM symposium on Theory of computing, 2009, pp. 169–178.

[2] T. Graepel, K. Lauter, and M. Naehrig, Ml confidential: Machine learning on encrypted data, in International conference on information security and cryptology. Springer, 2012, pp. 1–21.

[3] A. Stoian, J. Frery, R. Bredehoft, L. Montero, C. Kherfallah, and B. Chevallier-Mames, Deep neural networks for encrypted inference with tfhe, in International Symposium on Cyber Security, Cryptology, and Machine Learning. Springer, 2023, pp. 493–500.

[4] J. Frery, A. Stoian, R. Bredehoft, L. Montero, C. Kherfallah, B. Chevallier-Mames, and A. Meyre, Privacy-preserving tree-based inference with fully homomorphic encryption, Cryptology ePrint Archive, 2023.

[5] H. Attaullah, S. Sanaullah, and T. Jungeblut, Analyzing machine learning models for activity recognition using homomorphically encrypted real-world smart home datasets: A case study, Applied Sciences, vol. 14, no. 19, p. 9047, 2024.

[6] R. Podschwadt, D. Takabi, P. Hu, M. H. Rafiei, and Z. Cai, A survey of deep learning

architectures for privacy-preserving machine learning with fully homomorphic encryption, IEEE Access, vol. 10, pp. 117 477–117 500, 2022.

[7] S. Behera and J. R. Prathuri, Application of homomorphic encryption in machine learning, in 2020 2nd PhD Colloquium on Ethically Driven Innovation and Technology for Society (PhD EDITS). IEEE, November 2020, pp. 1–2.

[8] Y. Ameur, S. Bouzefrane, and V. Audigier, Application of homomorphic encryption in machine learning, in Emerging Trends in Cybersecurity Applications. Cham: Springer International Publishing, 2022, pp. 391–410.

[9] I. Chillotti, M. Joye, and P. Paillier, New challenges for fully homomorphic encryption, in Privacy-preserving machine learning (PPMLPriML 2020) NeurIPS 2020 workshop, 2020.

[10] Q. Li, Z. Huang, W. J. Lu, C. Hong, H. Qu, H. He, and W. Zhang, Homopai: A secure collaborative machine learning platform based on homomorphic encryption, in 2020 IEEE 36th International Conference on Data Engineering (ICDE). IEEE, April 2020, pp. 1713–1717.