All Sciences Proceedings

1st International Conference on Modern and Advanced Research

July 29-31, 2023 : Konya, Turkey

All Sciences Proceedings
http://as-proceeding.com/

https://as-proceeding.com/index.php/icmar

İstatistiksel ve Geometrik İlişkiye Dayalı Yeni Bir Sentetik Veri Üretme Yaklaşımının Geliştirilmesi

Fatma Akalın[1*]

[1]Bilişim Sistemleri Mühendisliği, / Bilgisayar ve Bilişim Bilimleri Fakültesi, Sakarya Üniversite, Sakarya,Türkiye

[*](fatmaakalin@sakarya.edu.tr)

Özet **–** Makine öğrenmesi, verileri analiz etmek ve değerlendirmek için kullanılan bir çerçevedir. Bu çerçeve ile belirlenen görevlerin makineler tarafından gerçekleştirilmesi hedeflenir. Bu hususta makine öğrenmesi yaklaşımının sahip olduğu potansiyelin açığa çıkarmak için veri önemli bir kriterdir ve kategoriler arası dengeli, kaliteli ve yeterli veri kümesi vasıtasıyla başarılı çıkarımlar yapmak mümkündür. Ancak yasal sınırlamalar, etik kurallar, maliyet parametresi ve yetersiz veri temini makine öğrenmesinin başarısını olumsuz etkileyen engellerdir. Bu engelleri aşmak için sentetik veri üretme, gerçek dünya problemlerinde bir çözüm noktası sunar. Bununla birlikte sentetik veri üretimine ilişkin standart bir çerçeve yoktur. Bu çalışmada istatistiksel ve geometrik temele dayanan yeni bir sentetik veri üretme yaklaşımı önerilmiştir. Böylece tanımlanabilir, hassas ve kritik bilgilerin kullanılması engellenecek ve gizlilik korunacaktır. Pahalı bir süreç sunabilen veri etiketleme ve veri toplama aşamalarında düşük maliyetli bir çözüm geliştirilecektir. Ek olarak sentetik veri üretme yaklaşımı ile artan veri sayısı ile doğru orantılı olarak eğitim başarısı iyileştirilecektir. Bu doğrultuda sunulan çalışmada yeni bir sentetik veri üretme yaklaşımı önerilmiştir. Dört ayrı kategoriye sahip lenfoma veri kümesine uygulanan sentetik veri üretme yaklaşımı ile veri sayısı iki katına çıkarılmıştır. Ardından yapay zeka ve bulanık mantık yöntemlerinin birlikte kullanıldığı ANFIS yöntemi ile orijinal ve zenginleştirilmiş veri kümeleri sınıflandırılmıştır. Sınıflandırma sonucunda orijinal ve zenginleştirilmiş veri kümeleri için sırasıyla %45 ve %75 doğruluk oranları elde edilmiştir. Çalışma sonucunda orijinal verilerin dinamiğinde üretilen sentetik veriler ile artan veri çeşitliliğinin doğruluk oranında sunduğu başarı, önerilen yaklaşımın gelecekte bir karar destek sistemi olarak kullanılmasına işaret etmektedir.

Anahtar Kelimeler – Makine Öğrenmesi, İstatistiksel İlişki, Geometrik Analiz, ANFIS, Sentetik Veri Üretme

I. GİRİŞ

Makine öğrenmesi, makineler tarafından görevlerin öğrenilmesi ve gerçekleştirilmesini sağlayan algoritmaların yer aldığı bir çerçevedir[1]. Akıllı bilgisayar sistemlerindeki görevlerin özerk olarak gerçekleştirilmesini sağlar ve endüstriyel inovasyonun sınırlarını zorlar. Aynı zamanda devasa bilgi içeren veri yığınlarını yönetir ve analiz eder. Özellikle bilgi işleme, simülasyon ve modelleme gibi birçok eylemi gerçekleştirmek makine öğrenmesi algoritmaları için önemli bir işlevselliktir.

Makine öğrenmesi teknolojisinin geliştirilmesinde ve uygulanmasında veriler kritik bir öneme sahiptir. Çünkü makine öğrenmesi, verileri kullanarak analizler yapar. Bununla birlikte veriler, kalitesiz olabilir. Verilerin kullanılması yasal düzenlemeler nedeniyle mümkün olmayabilir. Bazı sektörlerde üretilen veri sayısı kısıtlı olabilir ya da veri toplamanın ve etiketlemenin pahalı olmasından dolayı yüksek maliyet yeterli veri

içeren veri kümesinin oluşturulmasına engel olabilir. Bu problemler için sentetik veri üretme bir çözüm noktası oluşturmaktadır[2].

Sentetik veri üretme aşaması ile birlikte verilerin eğitim için yeterli bir sayıya ulaşması, her bir örneğe ilişkin kategori bilgisinin dengeli bir orana yakınlaşması ve hassas veriler için mahremiyetin korunması yapay verilerin üretilmesi için en temel nedenlerdir. Sentetik veri üretimine ilişkin literatür çalışmaları derlenmiş ve aşağıda sunulmuştur. Bu doğrultuda [3] çalışmasında bilgisayarlı görü endüstriyel kontrol sistemi geliştirmek için sentetik veri üretme süreci geliştirilmiştir. [4] çalışmasında gizlilik sorunları nedeniyle az sayıda veritabanı örneği bulunan dövme veri kümelerinden yarı sentetik görüntülerin oluşturulması için denetimsiz üreteç yaklaşımı önerilmiştir. [5] çalışmasında rüzgar hızı ve rüzgar gücü tahminlerinin oluşturulmasında gerçekçi sentetik rüzgar verileri üretmek için yeni bir model sunulmuştur. [6] çalışmasında kamuya açık verilerin kullanılma riskinden dolayı genel kullanıma yönelik sentetik veriler üretmek için uzay-zamansal veri analizi yöntemlerinin kullanımı önerilmiştir. [7] çalışmasında salgın hastalığının başlangıç zaman diliminin tahmin edilmesine ilişkin modellerin eğitilmesinde sentetik verilerin oluşturulmasına yönelik bir yöntem geliştirilmiştir. Güvenilir kararlar almak ve dengeli veri kategorilerine sahip olan veri kümesi oluşturmak için [8] çalışmasında genetik arama algoritması kullanarak sentetik veri üretmek için yeni bir yaklaşım sunulmuştur. [9] çalışmasında hematolojik malignitelerdeki istatistiksel, klinik ve genomik özellikler arasındaki ilişkiler vasıtasıyla oluşturulan modellerdeki güvenli tahminleri arttırmak amacıyla MDS/AML sentetik kohortları oluşturulmuştur. [10] çalışmasında down sendromlu bireylerin erken tanısında dengesiz ve mahrem olan veri problemi için SMOTE tabanlı ve GAN tabanlı yöntemler ile tahmin gücünün iyileştirilmesi hedeflenmiştir. [11] çalışmasında yasal düzenlemelerden ve etik ihlallerden muaf olmak amacıyla sentetik elektronik sağlık kayıtları oluşturmak için sentetik hastaların yaşam sürelerini simüle eden Synthea isimli yazılım paketi geliştirilmiştir. Önerilen yapının hastalık ve tedavi modülleri ile genişletilebileceği ifade edilmiştir.

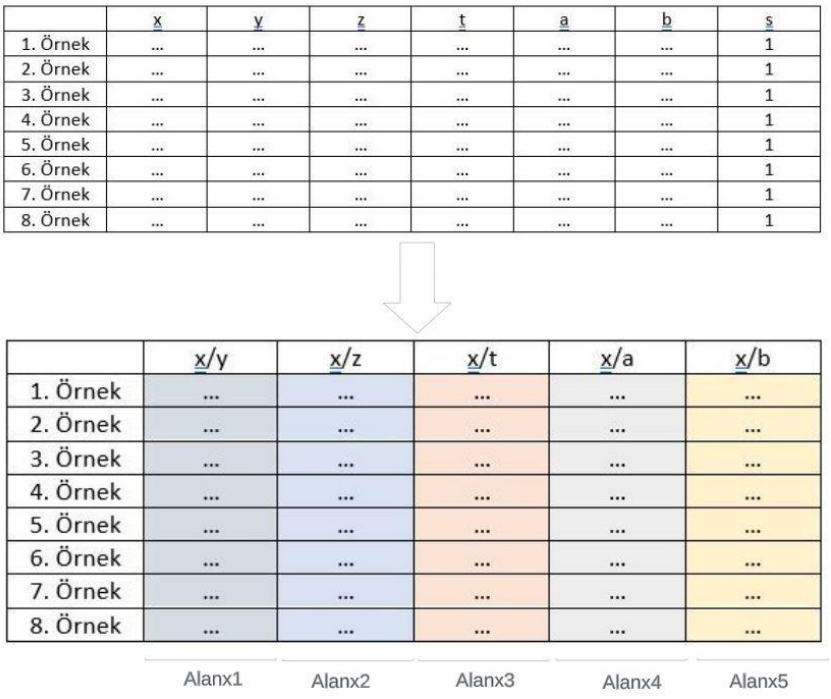
Literatürdeki makalelere kıyasla bu çalışma geometrik ve istatistiksel özellikler vasıtasıyla çıkarım yaparak orijinal veri kümelerine benzer veriler üretmeyi amaçlamaktadır. Bu hususta önerilen yöntemin tüm detayları metodoloji başlığı altında açıklanmıştır.

II. MATERYAL VE YÖNTEM

Sentetik veri üretimi, sağlık, biyometri, enerji tüketimi gibi farklı birçok alanda başarısını ispatlayan bir yaklaşımdır [12]. Fakat sentetik verilerin üretimine ilişkin halen standart bir çerçeve yoktur[2]. Bu nedenle yeni ve güncel yaklaşımların geliştirilmesine yönelik çalışmalar devam etmektedir.

Bu çalışmada NCBI-GEO veri kümesi kullanılmıştır. Bu veri kümesindeki GSE17920 serisi için GSM447610-GSM447739 aralığındaki idlere sahip olan veriler [13] sitesinden tedarik edilmiştir. Veriler, Hodgkin lenfoma hastalarından elde edilen tanısal biyopsi örneklerinin değerlerini içermektedir. Toplam 130 veri içeren veri kümesinin 31 örneğinde eksik özellikler mevcuttur[14]. Eksik özelliklere sahip veriler arasında istatistiksel ve geometrik ilişkinin kurulması başarılı çıkarımların yapılması ve güçlü desenlerin keşfedilmesinde bir engel oluşturduğu için bu çalışmada eksik veri içermeyen 99 veri kullanılarak sentetik veri üretme süreci inşa edilmiştir.

Lenfoma veri kümesi vasıtasıyla sentetik veri üretme süreci için inşa edilen algoritmanın ilk aşamasında her bir kategori sınıfı gruplandırılmaktadır. İkinci aşamada her bir örneğin özellikleri oranlanarak yeni bir matris elde edilmektedir. Üçüncü aşamada matlab programı vasıtasıyla polyarea fonksiyonu kullanılarak her bir sütun için alan değerlerine ulaşılmakta ve dördüncü aşamada üçgende benzerlik oranı yaklaşımı kullanılarak girilen x değeri vasıtasıyla diğer değişkenlerin bulunması sağlanmaktadır. Bu hiyerarşinin genel çerçevesi Şekil 1’ de sunulmuştur.



Şekil 1. Önerilen hiyerarşinin görsel tasviri

Şekil 1'de, veri kümesinde kategori bilgisi 1 olarak nitelendirilen tüm örnekler arasında bir oranlama yapıldıktan sonra bu oran hesabı kullanılarak her bir özelliğe ilişkin alan bilgisi hesaplanır. Ardından üçgenlerin benzerliği kuralından esinlenerek x/y=Alanx1/ToplamAlan, x/z=Alanx2/ToplamAlan, x/t=Alanx3/ToplamAlan, x/a=Alanx4/ToplamAlan, x/b=Alanx5/ToplamAlan bağlantıları kurulur. Bağlantılar sayesinde girilen x değeri ile y,z,t,a ve b değerleri için atamalar yapılır. Fakat bu hesaplamalar benzer üçgenler için kullanılan bir yaklaşım olduğu için diğer adımlarda atamaları yapılan değerlerin gerçek değere daha yakın bir çıktı üretmesi hedeflenir. Bu nedenle bu değerlere aşağıda verilen maddeler sırasıyla uygulanır.

1- Her bir özelliğin yer aldığı sütunlar için ortalama değer bulunduktan sonra atamaları yapılan y,z,t,a ve b değerleri için karşılaştırma sağlanır. Örneğin tahmin edilen y değeri veri kümesindeki y değerlerinin ortalamasından küçükse, bu iki değer birbirinden çıkarıldıktan sonra mutlak değeri alınır. Aksi durumda mutlak değeri alınmaz. Bu işlemler sonucunda sapma değeri elde edilir ve z,t,a ve b sütunları için aynı işlemler tekrar edilir.
2- Atamaları yapılan y,z,t,a ve b değerleri her bir özelliğin yer aldığı matristeki özelliklerin ortalama değerleri ile karşılaştırılır. Örneğin atanan y değeri ortalaması hesaplanan y değerinden büyük ise ataması yapılan y değeri, uyarlanabiliry katsayısı ile çarpılan y değerinden çıkarılır. Aksi durumda ataması yapılan y değeri, uyarlanabiliry katsayısı ile çarpılan y değeri ile toplanır. Bu denklemde yer alan uyarlanabiliry katsayısı tahmin edilen y değerinin ortalaması hesaplanan y değerine bölünmesi ile elde edilir. Tüm bu işlemler z,t,a ve b sütunları için de yapılır.
3- İkinci madde ile elde edilen y,z,t,a ve b değerleri için karşılaştırma yapılır. Örneğin güncellenen y değeri her bir özelliğin yer aldığı sütunlar için ortalama değeri bulunan y özelliği ile kıyaslanır. Öyleki güncel y değerinin ortalama y değerinden büyük olması durumunda güncel y değerinden ilk maddedeki y değeri için elde edilen sapma değerinden çıkarılması sağlanır. Aksi

durumda toplama işlemi yapılır. Tüm bu işlemler z,t,a ve b sütunları için de yapılır.

4- İlk madde de elde edilen tüm sapma değerlerinin ortalaması alındıktan sonra y,z,t,a ve b değerleri için elde edilen sapma değerleri ile karşılaştırılır. Örneğin y değeri için elde edilen sapma değeri, ortalama sapmadan büyük ise yeni cevap üçüncü maddede elde edilen sonuçtur. Aksi durumda ise yeni cevap 2. maddede elde edilen sonuçtur. Tüm bu işlemler z,t,a ve b sütunları için de yapılır.

1,2,3 ve 4. maddede verilen işlemler tüm kategoriler için ayrı ayrı uygulanır. Böylece hedef örneğe ilişkin orijinal veri kümesinde belirlenen aralıklarda istenilen özellik değerleri kullanılarak diğer özelliklerin sentetik değerleri üretilir. Matematiksel ve geometrik bir temele dayanan bu çalışmada orijinal veri kümesinde belirlenen aralıklarda olmak şartıyla her bir özellik (x,y,z,t,a ve b) için yeni örnekler üretilmiştir. Ardından üretilen sentetik verilerin kalitesi ANFIS sınıflandırma yapısı kullanılarak analiz edilmiştir.

ANFIS, yapay sinir ağı ile bulanık mantık algoritmasının birlikte kullanılmasını sağlayan yapay sinir ağı tekniğidir. Girdi çıktı çiftlerini kullanarak uygun kuralların öğrenilmesi ve sonuç parametrelerinin ayarlanabilmesini sağlar. Rastgele doğrusal fonksiyon içeren bir sonuç kısmı sunar[15].

Orijinal veri kümesindeki verilere ek olarak orijinal veri kümesinde yer alan veriler arasındaki istatistiksel ve geometrik ilişki kullanılarak üretilen sentetik verilerin orijinal veri kümesine eklenmesi ile oluşturulan güncel veriler üzerinde ANFIS yaklaşımı vasıtasıyla elde edilen performans bulgular ve tartışma başlığı altında verilmiştir.

III. BULGULAR VE TARTIŞMA

Sentetik veri üretimi, yasal düzenlemeler, mahremiyet, maliyet ya da yetersiz veri temini problemlerinde bir çözüm noktası oluşturmak için başarısı ispat edilen bir yaklaşımdır. Standart bir çerçevesi yoktur. Fakat veriler arasındaki ilişkilerin, desenlerin ve kritik noktaların keşfedilmesinde artan veri sayısı ile doğru orantılı olarak eğitim başarısı iyileşmektedir. Bu potansiyel noktasından yola çıkarak orijinal veri kümesi ve orijinal veri kümesine eklenen sentetik veriler ile oluşturulan güncel veri kümesi vasıtasıyla ANFIS yöntemi vasıtasıyla bir sınıflandırma yapılmıştır. Eğitim ve test veri kümesi üzerinde sınıflandırma sonucunda ulaşılan performans ölçütleri Tablo 1 ve Tablo 2'de verilmiştir.

Tablo 1. Veri Kümesi 1 için eğitim ve test verileri üzerinde elde edilen performans değerleri

	Stage 1			Stage 2			Stage 3			Stage 4		
Veri Kümesi1	Ks.	Dy.	F S.	Ks.	Dy.	F S.	Ks.	Dy.	F S.	Ks.	Dy.	F S.
TRAIN	1	0.66	0.79	0.94	1	0.96	1	0.84	0.91	0.88	1	0.93
TEST	0.25	0.2	0.22	0.36	0.8	0.49	0.5	0.2	0.28	1	0.6	0.75

Tablo 2. Veri Kümesi 2 için eğitim ve test verileri üzerinde elde edilen performans değerleri

	Stage 1			Stage 2			Stage 3			Stage 4		
Veri Kümesi2	Ks.	Dy.	F S.	Ks.	Dy.	F S.	Ks.	Dy.	F S.	Ks.	Dy.	F S.
TRAIN	0.83	0.83	0.83	0.84	0.87	0.85	0.8	0.59	0.67	0.74	0.88	0.80
TEST	1	0.8	0.88	0.56	1	0.71	0.66	0.6	0.62	0.84	1	0.91

Veri Kümesi 1, stage etiketi ile 4 ayrı kategoriye sahip orijinal lenfoma veri kümesidir. Veri Kümesi 2, stage etiketi ile 4 ayrı kategoriye sahip zenginleştirilmiş lenfoma veri kümesidir.

Bu tablolarda eğitim ve test kümesi olarak ayrılan orijinal ve zenginleştirilmiş veri kümesinde her bir kategori için yapılan doğru tahminlerin aynı kategori kapsamında yapılan yanlış tahminler ile birlikte değerlendirilmesini sağlayan kesinlik değerinin, her bir kategori için yapılan doğru tahminlerin aynı kategoriye ait tüm tahminlere oranı olan duyarlılık değerinin ve Kesinlik ile duyarlılık değerlerinin harmonik ortalaması olan ve kategoriler arasında yapılan dengesiz tahminlerin

değerlendirilmesini sağlayan F skorunun 1'e yakın bir sonuç üretmesi uygun bir veri kümesi ile doğru seçilen sınıflandırma yaklaşımının başarılı bir kombinasyonuna işaret etmektedir. Zenginleştirilmiş veri kümesi için güvenilir sonuçlar üreten bu kriterlere ek olarak orijinal veri kümesindeki eğitim verileri ile eğitilen modelin test veri kümesi üzerinde ürettiği %45 başarı oranı ile zenginleştirilmiş veri kümesindeki eğitim verileri ile eğitilen modelin test veri kümesi üzerinde ürettiği %75 başarı oranı sentetik verilerin orijinal verilere benzer bir dinamikte oluştuğunu göstermektedir. Aksi durumda orijinal veri kümesine aykırı üretilen veri yığınları ile sınıflandırma aşamasında yetersiz performans sağlanacaktır.

Bu çalışmada veriler arasında istatistiksel ve geometrik ilişkinin kurulması ile orijinal verilerin dinamiğine benzer yapay veriler üretilmiştir. Önerilen sentetik veri üretme yaklaşımının literatüre katkı sağlaması hedeflenmektedir.

IV. SONUÇLAR

Bu çalışmada istatistiksel ve geometrik temele dayanan yeni bir sentetik veri üretme yaklaşımı önerilmiştir. Bu yaklaşım ile tanımlanabilir bilgiler içeren verilerin kullanılması engellenecek ve gizlilik korunacaktır. Yasal sınırlamalara takılmadan veri analizi yapılacaktır. Veri etiketleme ve veri toplamanın maliyetli olduğu durumlar için maliyet etkin bir çözüm noktası geliştirilecektir. Veriler arasında önemli bir potansiyele sahip olan bilgilerin modele öğretilmesi için veri çeşitliliği sağlanacaktır. Orijinal verilerin dinamiğinde üretilen sentetik veriler ile artan veri sayısı ile doğru orantılı olarak eğitim başarısı iyileşecektir. Bu nedenle önerilen yaklaşımın gelecekte karar destek sistemi olarak kullanılması hedeflenmektedir.

KAYNAKLAR

[1] D. A. Hashimoto, T. M. Ward, and O. R. Meireles, 'The Role of Artificial Intelligence in Surgery', *Adv. Surg.*, pp. 1–13, 2020, doi: 10.1016/j.yasu.2020.05.010.

[2] Y. Lu, M. Shen, H. Wang, and W. Wei, 'Machine Learning for Synthetic Data Generation : A Review', *arXiv*, vol. 14, no. 8, pp. 1–18, 2021.

[3] I. Reutov, D. Moskvin, A. Voronova, and M. Venediktov, 'Generating Synthetic Data To Solve Industrial Control Problems By Modeling A Belt Conveyor', *Procedia Comput. Sci.*, vol. 212, pp. 264–274, 2022, doi: 10.1016/j.procs.2022.11.010.

[4] L. J. Gonzalez-soler, C. Rathgeb, and D. Fischer, 'Semi-synthetic Data Generation for Tattoo Segmentation', *2023 11th Int. Work. Biometrics Forensics*, pp. 1–6, doi: 10.1109/IWBF57495.2023.10157837.

[5] A. Naimo, 'A Novel Approach to Generate Synthetic Wind Data', *Procedia - Soc. Behav. Sci.*, vol. 108, pp. 187–196, 2014, doi: 10.1016/j.sbspro.2013.12.830.

[6] H. Quick and L. A. Waller, 'Using spatiotemporal models to generate synthetic data for public use', *Spat. Spatiotemporal. Epidemiol.*, vol. 27, pp. 37–45, 2018, doi: 10.1016/j.sste.2018.08.004.

[7] A. L. Buczak, S. Babin, and L. Moniz, 'Data-driven approach for creating synthetic electronic medical records', *BMC Med. Inform. Decis. Mak.*, pp. 1–28, 2010.

[8] P. Thogarchety and K. Das, 'Synthetic Data Generation Using Genetic Algorithm', *2023 2nd Int. Conf. Innov. Technol.*, pp. 1–6, 2023, doi: 10.1109/INOCON57975.2023.10101072.

[9] S. D. Amico, D. D. Olio, C. Sala, L. D. Olio, E. Sauta, and E. All., 'Synthetic Data Generation by Artificial Intelligence to Accelerate Research and Precision Medicine in Hematology', *Clin. cancer informatics*, pp. 1–22, 2023, doi: 10.1200/CCI.23.00021.

[10] T. K. Tran et al., 'Evaluation of Synthetic Data Generating Methods in Down Syndrome Prediction During the First Trimester Screening in Vietnam', *2022 Int. Conf. Multimed. Anal. Pattern Recognition, MAPR 2022 - Proc.*, 2022, doi: 10.1109/MAPR56351.2022.9924885.

[11] J. Walonoski et al., 'Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record', *J. Am. Med. Informatics Assoc.*, vol. 25, no. 3, pp. 230–238, 2018, doi: 10.1093/jamia/ocx079.

[12] M. Hernandez, G. Epelde, A. Alberdi, R. Cilla, and D. Rankin, 'Synthetic data generation for tabular health records: A systematic review', *Neurocomputing*, vol. 493, pp. 28–45, 2022, doi: 10.1016/j.neucom.2022.04.053.

[13] 'NCBI Gene Expression Omnibus'. https://www.ncbi.nlm.nih.gov/geo.

[14] F. AKALIN, M. F. ORHAN, and M. BUYUKAVCİ, 'A Decision Support System For Detecting Stage In

Hodgkin Lymphoma Patients Using Artificial Neural Network and Optimization Algorithms', *Sak. Univ. J. Comput. Inf. Sci.,* vol. 5, no. 3, 2022, doi: 10.35377/saucis...1210786.

[15] F. Akalın, 'Sayısal haritalama teknikleri kullanılarak DNA dizilimleri üzerinden lösemi hastalığının temel türlerinin yapay zeka tabanlı algoritmalar ile sınıflandırılması', *Doktora Tezi,* Sakarya Üniversitesi, Sakarya, Mart. 2023.