Fırat Üniversitesi Müh. Bil. Dergisi
35(2), 753-760, 2023

Araştırma Makalesi
https://doi.org/10.35234/fumbd.1333258

Genetik Algoritma Temelli Yeni Bir Sentetik Veri Üretme Yaklaşımının Geliştirilmesi

Fatma AKALIN[1*]
[1] Bilişim Sistemleri Mühendisliği, Bilgisayar ve Bilişim Bilimleri Fakültesi, Sakarya Üniversitesi, Sakarya, Türkiye
[*1] fatmaakalin@sakarya.edu.tr

(Geliş/Received: 26/07/2023; Kabul/Accepted: 11/08/2023)

Öz: Yapay zeka tabanlı çalışmalar, iş sektörlerinde karar destek sistemi oluşturmak, etkili çıktılar üretmek, sistem verimliliğini arttırmak ve maliyet etkin çözümler sunmak için büyük bir ilgi odağına sahiptir. Özellikle inovasyon sürecinin gelişmesinde, hızlanmasında ve hedef alana evrilmesinde yapay zeka tabanlı çalışmalar ile yenilikler sağlanmaktadır. Bu yeniliklerin gerçekleşmesinde veri, kritik bir anlama sahiptir. Algoritmalar vasıtasıyla eğitilen modellerin bilgisayarlar ya da özel makineler tarafından işlevselleştirilmesinde önemli bir rol oynamaktadır. Bununla birlikte yetersiz veri erişimi, yasal düzenlemeler, etik kurallar, gizlilik prosedürleri, mahremiyet, veri paylaşım kısıtı ve maliyet; verilerin sahip olduğu potansiyelin açığa çıkarılmasının önündeki engellerdir. Bu engelleri aşmak için sentetik veri üretme yaklaşımı tercih edilmektedir. Fakat sentetik veri üretme yaklaşımına ilişkin standart bir çerçeve olmadığı için yeni ve güncel yaklaşımların geliştirilmesine yönelik araştırmalar devam etmektedir. Bu çalışmada genetik algoritma temelli yeni bir sentetik veri üretme yaklaşımı önerilmiştir. Bu doğrultuda orijinal veri kümesinin dinamiğinde yapay veriler üretmek için hedef veri kümesine uyarlanan çaprazlama ve mutasyon genetik operatörleri kullanılarak veri çeşitliliği arttırılmıştır ve yeni bir nesil elde edilmiştir. Ardından üretilen bu nesildeki yapay örneklerin kategori tanımlaması, genetik algoritmanın maliyet fonksiyon bileşeni kullanılarak belirlenmiştir. Son aşamada üretilen yapay verilerin orijinal verilere benzerliğinin başarısını ölçmek için 6 farklı makine öğrenmesi sınıflandırıcısı kullanılmıştır. Zenginleştirilen veri kümesi üzerinde Destek Vektör Makinesi sınıflandırıcısı ile maksimum duyarlılık ölçütü, %100 olarak elde edilmiştir. Bu durum artan veri sayısı ile orantılı olarak eğitim başarısının pozitif yönde eğilim gösterdiğini ifade etmektedir.

Anahtar kelimeler: Sentetik veri üretimi, genetik algoritma, makine öğrenmesi sınıflandırıcıları

Development of a New Synthetic Data Generation Approach Based on Genetic Algorithm

Abstract: Artificial intelligence-based studies have a great interest in creating decision support systems in business sectors, producing effective outputs, increasing system efficiency and providing cost-effective solutions. Especially in the development of the innovation process, the acceleration of the innovation process and its evolution into the target area, innovations are provided with artificial intelligence-based studies. In the realization of these innovations, data has a critical meaning for artificial intelligence-based studies. It plays an important role in the functionalization of models trained through algorithms by computers or special machines. However, insufficient data access, legal regulations, ethical rules, confidentiality procedures, privacy, data sharing limitations and cost; are major obstacles to revealing the potential of data. To overcome these obstacles, the synthetic data generation approach is preferred. But, since there is no standard framework for the synthetic data generation approach, research on the development of new and current approaches continues. In this study, a new synthetic data generation approach based on a genetic algorithm is proposed. In this direction, data diversity has been increased and a new generation has been obtained by using the crossover and mutation genetic operators adapted to the target dataset to produce artificial data in the dynamics of the original dataset. Then, the category definition of the artificial samples in this generation was done using the cost function component of the genetic algorithm. In the last stage, 6 different machine learning classifiers were used to measure the success of the similarity of the artificial data produced to the original data. The maximum sensitivity criterion was obtained as 100% with the Support Vector Machine classifier on the enriched dataset. This indicates that educational success tends to be in the positive direction in proportion to the increasing number of data.

Key words: Synthetic data generation, genetic algorithm, machine learning classifiers

1. Giriş

Yapay zeka, insanın sahip olduğu bilişsel yeteneklerin özel makineler ya da bilgisayarlar aracılığı ile taklit edilmesini sağlayan geniş bir çerçevedir [1]. Makine öğrenmesi, sinir ağları ve derin öğrenme alt alanlarını içeren yapay zeka çerçevesi ile matematik, istatistik, psikoloji ve nörobiyoloji gibi birçok alanda aktif çalışmalar gerçekleştirilmektedir. Yapay zeka çerçevesindeki tüm alt alanlar birbiri ile ilişkilidir. Bu nedenle göreve bağlı olarak her bir alt alanda yer alan tekniklerin diğer alt alanlarda yer alan teknikler ile kullanımı mümkündür [2].

[*] Sorumlu yazar: fatmaakalin@sakarya.edu.tr. Yazarların ORCID Numarası: [1] 0000-0001-6670-915X

Makine öğrenmesi, günümüzde ciddi etki oluşturan popüler bir yapay zeka kategorisidir [3]. Algoritmalar vasıtasıyla eğitilen modellerin bilgisayarlar ya da özel makineler vasıtasıyla işlevselleştirilmesini sağlar. Makine öğrenmesi yaklaşımları ile başarılı bir eğitimin gerçekleştirilmesinde ve makinelere güçlü bir fonksiyonellik kazandırılmasında veri kritik bir öneme haizdir [4]. Çünkü yetersiz veri sayısı ile başarılı bir eğitim mümkün değildir. Bununla birlikte verilerin kalitesiz olması, yasal düzenlemeler, etik kurallar ve gizlilik prosedürleri; makine öğrenmesi yaklaşımları ile veriler arasındaki ilişkilerin, desenlerin, benzerliklerin ve farklılıkların makineler tarafından öğrenilmesinde ve sonuçların değerlendirilmesinde bir engel teşkil etmektedir. Bunun sonucunda makine öğrenmesi modellerinde yetersiz öğrenme oluşmaktadır. Yetersiz öğrenme durumuna bir çözüm üretmek için kaliteli veri toplama ve etiketleme yapılabilir. Ancak bu süreçler maliyetli bir çıktı sunduğu için [4] sentetik veri üretme yaklaşımı vasıtasıyla orijinal verilerin dinamiğinde zenginleştirilmiş veri kümesi oluşturmak tercih edilen diğer bir seçenektir.

Sentetik veriler, tanımlanabilir bilgiler içermediği için kişi mahremiyetini koruyan yapay bilgi topluluğudur. Araştırmaların hızlı yürütülmesini ve araştırma süreçlerinde pozitif bir etki oluşturulmasını sağlar. Sonuçları maksimum verim ve düşük bir maliyet ile iyileştirir. İstatistiksel simülasyon ya da hesaplamalı türetme ile üretilen yapay veriler vasıtasıyla artan veri çeşitliliği ile geniş tanımlayıcı analizler yapılır. Gerçek verilere kıyasla veri paylaşımı kolay ve hızlı bir şekilde gerçekleşir. Artan veri ile birlikte geleceğe yönelik kritik problemlere ilişkin bir öngörücü rolü üstlenir. Özellikle tıp alanında tıbbi inovasyonun hızlandırılmasını sağlayabilen bir potansiyel içerir. Ek olarak sentetik veri yaklaşımı kapsamında hedef amaç doğrultusunda kullanılan yapay verilere ait tüm ayrıntıların bilimsel yayında ve üretilen tüm içerikler içinde ifade edilmesi gerekli bir konudur [1].

Sentetik veri üretimine ilişkin [5] çalışmasında orta ölçekli Şili madenciliğindeki atık barajların kapatılmasında fiziksel stabilitenin farklı potansiyel yenilme mekanizmalarına (PFM) ilişkin tahmini için atık barajların kritik değişkenlerin analizi sağlanmıştır ve beş olası PFM'nin tahmininde makine öğrenme algoritmaları kullanılmıştır. Burada veri azlığı nedeni ile üretken çekişmeli ağlar (GAN) yöntemi kullanılarak veri sayısı arttırılmıştır. Zenginleştirilmiş veri kümeleri üzerinde makine öğrenmesi modelleri vasıtasıyla elde edilen F1 skor metriğinde %30 artış görülmüştür. [6] çalışmasında ilişkisiz arka plan görüntüleri vasıtasıyla etki alanını rastgeleleştirerek nesne algılama görevine sahip sinir ağının nesne özelliklerini öğrenmesi ve arka plandan bağımsız olması için ardışık bir düzen önerilmiştir. Sentetik veri üretme amacına sahip bu düzen vasıtasıyla koni nesne algılama uygulaması kapsamında gerçekleştirilen bu çalışmanın başarılı bir çıktı ürettiği ifade edilmiştir. Sürücü davranışını modellemek için gerekli veriye ulaşmanın zorluğundan esinlenerek [7] çalışmasında orijinal verinin istatistiki ve dağılım özelliğini kullanan synthpop kütüphanesi ile sentetik veri üretimi yapılmıştır. Ulaşılan sonuçların umut vaat edici çıktılar sunduğu ifade edilmiştir. Sentetik radyasyon verilerinin oluşturulmasının amaçlandığı [8] çalışmasında birden fazla şehir bloğunu kapsayan modüler 3-D Monte Carlo modelleri ile veri kümeleri oluşturmak için Monte Carlo simülasyonlarının kullanımı açıklanmıştır. En tehlikeli afetlerden biri olan orman yangınlarının analizi ve yönetimi, önemlidir. Fakat büyük orman yangınlarının sayısı orman yangınlarının sayısına göre oldukça az olması durumlarında makine öğrenmesi modelini eğitmek için kullanılması planlanan veri kümesi dengesizdir [9]. Bu nedenle [9] çalışmasında beş farklı sentetik veri oluşturma yöntemi değerlendirilmiştir ve sentetik verilere ilişkin kullanımın tahmin gücünde iyileştirme sağlandığı ifade edilmiştir. Su altı dünyasında ıstakoz türlerinin bolluğu ve biyoçeşitliliğini izlemek için su altı görüntülenmektedir. Fakat görüntülerin manuel değerlendirilmesi zaman alıcı bir süreç sunduğu için [10] çalışmasında ıstakoz algılama işleminin otomatikleştirilmesi hedeflenmiştir. Açıklamalı eğitim veri kümesi eksikliğinden dolayı nesne algılama modellerinde kullanmak üzere sentetik veriler üretilmiştir. Sentetik veri üretimine ilişkin SPD isimli bu yaklaşımın nihai çıktı üzerinde performans artışı sağladığı belirtilmiştir. İnsan merkezli veri toplama sürecinin maliyetli olduğunu ifade eden [11] çalışmasında GAN yaklaşımı ile parmak izi tabanlı yerelleştirmede üretilen sentetik veriler vasıtasıyla eğitim verisi toplanmıştır. Kabul edilebilir bir doğruluk elde edildiği ifade edilmiştir. [12] çalışmasında manyetik rezonans görüntülemeye dayalı incelemenin zahmetli, hataya açık ve zaman alıcı olması probleminden dolayı MRI veri analizinde beyin tümörlerini sınıflandırmak için derin öğrenme yaklaşımı kullanılmıştır. Sınıflandırma doğruluğunu iyileştirme hedefi kapsamında sunulan veri arttırma konsepti ile önerilen stratejinin etkinliği değerlendirilmiştir. Son teknoloji tekniklere kıyasla daha iyi performans elde edildiği ifade edilmiştir. Devasa verinin mevcut olduğu günümüz çağında halen küçük veri kümelerinin kullanımına ilişkin sınırlamaların aşılması ve denetimli öğrenme algoritmalarının kullanılması için [13] çalışmasında Geometrik Küçük Veri Aşırı Örnekleme Tekniği önerilmiştir. Veri oluşturma mekanizması Geometrik SMOTE yaklaşımına dayanan bu teknik ile mevcut örneklerin etrafındaki geometrik bölgelerden faydalanarak yüksek kaliteli örnekler oluşturulmuştur. Doğruluk değerinde önemli bir başarı sağlandığı ifade edilmiştir. Belgelerden finansal ve idari alanlar için kimlik doğrulaması yapmak gerekli bir konu iken imza doğrulama sistemlerinin performansını etkileyen gerçek imza sayısının yetersiz olmasından dolayı [14] çalışmasında özellik üreteci önerilmiştir. Yapay bağışıklık sistemlerinin mutasyon, klonlama ve kaynak rekabeti mekanizmaları üzerinde temellenen bu yaklaşımın

doğrulama adımında SVM sınıflandırıcısı kullanılarak iki tanımlayıcı vasıtasıyla değerlendirme sağlanmıştır. Sentetik özelliklerin etkinliğini vurgulayan bu çalışmanın geliştirilen doğrulama sistemi üzerinde bir iyileştirme sağladığı ifade edilmiştir. Ticari tarım faaliyetlerinde sürdürülebilirliğin sağlanması ve kalite standartlarının oluşturulması için [15] çalışmasında, tüketimde kullanılacak asma yapraklarının türünün tanınması hedeflenmiştir. Bu doğrultuda ESRGAN modeli vasıtasıyla yaprağa ilişkin doku özelliklerinin korunduğu bir veri kümesi elde edilmiştir. Ardından VGG19 modeli ile verilerden çıkarılan öznitelikler arasından en iyi öznitelik alt kümesinin seçimi PCA algoritması ile yapılmıştır. Son aşamada Destek Vektör Makineleri yaklaşımı kullanılarak gerçekleştirilen sınıflandırma işlemi sonucunda %96,14 doğruluk oranı elde edilmiştir.

Literatürdeki çalışmalar sentetik veri üretme yaklaşımı ile oluşturulan zenginleştirilmiş veri kümelerinde eğitim performansının orantılı olarak arttığına işaret etmektedir. Bu doğrultuda sunulan çalışmada 54 farklı özelliğin yer aldığı boşanmayı tahmin etme ölçeği kullanılarak verilen cevaplar, makine öğrenmesi sınıflandırıcıları ile sınıflandırılmıştır. Ardından genetik algoritma temelli yeni bir sentetik veri üretme yaklaşımı kullanılarak üretilen yapay veriler, orijinal veri kümesine eklenmiştir ve zenginleştirilmiş veri kümesinde yeniden bir sınıflandırma işlemi sağlanmıştır. Değerlendirme sonucunda verilerden sağlanan çıkarım performansının iyileştirildiği ve artan performans eğrisine sahip bir başarının elde edildiği görülmüştür. Böylece önerilen sentetik veri üretme yaklaşımı ile yapay zeka temelli çalışmalarda kritik bir öneme sahip olan veri sayısının bir doyum noktasına ulaşmasının sonucunda amaca uygun çıktıların üretilmesi, modelin başarılı bir şekilde eğitilmesi, yapay zeka çözümlerinin hayatın içerisine dahil edilmesi ve düşük maliyet sunan bir karar destek sisteminin inşa edilmesi için çözüm potansiyeli oluşturulmuştur.

2. Metodoloji

Sentetik veri üretimi, gerçek veri kümesindeki orijinal verilerin dinamiğinde yapay verilerin üretilmesini ifade eden bir yaklaşımdır. Gerçek dünya problemlerine ilişkin hedef veriler üzerinde başarılı analizler yapmak, kararlı çıkarımlar üretmek ve inovasyonu desteklemek amacıyla tercih edilmektedir.

Bu çalışmada genetik algoritma temelli yeni bir sentetik veri üretme yaklaşımı önerilmiştir ve önerilen yaklaşımın performansı belirli bir boyuta sahip örneklem büyüklüğü ile analiz edilmiştir. Bu çalışmada kullanılan veri kümesi, genetik algoritma ve genetik algoritma temelli sentetik veri üretme yaklaşımı aşağıda ayrıntılı olarak açıklanmıştır.

2.1. Boşanma öngörü veri kümesi

Hayat alanı kapsamında kişisel bilgi formu ve boşanmayı yordama ölçeği kullanılarak elde edilen Boşanma Öngörü Veri Kümesi (Divorce predictors dataset), UCI (the University of California Irvine Machine Learning Repository) halka açık gen bankasından [16] tedarik edilmiştir. Toplam 54 farklı öznitelikten oluşan veri kümesinde kişilerin boşanma tercihleri 2 ayrı kategoride (0 ve 1) sunulmuştur ve veri kümesinde toplam 170 farklı örnek mevcuttur.

Bu çalışmada Boşanma Öngörü Veri Kümesi kullanılarak genetik algoritma temelli sentetik veri üretme yaklaşımı vasıtasıyla orijinal veri kümesinin dinamiğinde üretilen sentetik veriler ile zenginleştirilmiş veri kümesi elde edilmiştir. Ardından sınıflandırma aşamasında orijinal ve zenginleştirilen güncel veri kümesinin %65'i eğitim ve %35'i test kümesi olarak ayrılmıştır. Son aşamada orijinal ve zenginleştirilen veri kümesi üzerinde önerilen sentetik veri üretme yaklaşımının performansı makine öğrenmesi sınıflandırıcıları ile değerlendirilmiştir.

2.2. Genetik algoritma

Genetik algoritma, 1975 yılında Holland tarafından geliştirilmiştir. Evrimsel fikirler üzerinde temellenen bu algoritma; seçim, mutasyon ve çaprazlama teknikleri vasıtasıyla çözümler üretmektedir [17]. Matematiksel işlemlerin aksine ayrık ve doğrusal olmayan işlemlerden oluşur. Temel amaç, nesilden nesile evrilme sürecinde optimum duruma yaklaşmak ya da optimum durumu elde etmektir. Aynı zamanda karmaşık problemlerin çözümü için tercih edilen bir yaklaşım sunmaktadır. Bu nedenle bilgisayar ağları, yazılım mühendisliği, görüntü işleme, konuşma tanıma, sağlık hizmetleri ve makine öğrenmesi gibi farklı birçok alanda kullanımı mevcuttur [18].

Genetik algoritmanın ilk adımında başlangıç çözümleri rastgele oluşturulur. İkinci adımda hedef problem doğrultusunda değerlendirme fonksiyonu belirlenir. Probleme özgü performansın başarısını değerlendirmek için gerçekleştirilen bu aşamadan sonra uygunluk değerlerine göre iki çözüm seçilir. Ardından bu çözümlerin genetik algoritma parametreleri vasıtasıyla çoğalması sağlanır ve her tur sonucunda yeni nesil elde edilir. Bu adım,

optimum çözümü bulma, optimuma yakın çözüm elde etme ya da sonlandırma kriterine ulaşma durumuna kadar devam eder [18]. Genetik algoritmanın genel hiyerarşisi Kaba Kod 1 kısmında sunulmuştur [18].

Kaba Kod 1. Genetik algoritmanın genel hiyerarşisi

```
1 | Başlangıç popülasyonunu oluştur
2 | Her bir bireyin uygunluk değerini değerlendir
3 | Seleksiyon işlemini gerçekleştir
4 | Çaprazlama yap
5 | Mutasyon uygula
6 | Belirlenen iterasyon sayısına ulaşıldı mı?
7 | Belirlenen iterasyon sayısına ulaşıldı ise sonlandır / Belirlenen iterasyon
  | sayısına ulaşılmadı ise ikinci maddeye geri dön
8 | Bitir
```

Kaba kod 1'de özetlendiği gibi bireyler topluluğu kullanılarak inşa edilen genetik algoritma yaklaşımında bireyler birleşir ve çocuklar oluşur. Çocuklara seleksiyon, çaprazlama ve mutasyon işlemlerinin uygulanması ve bu işlemlerin iterasyon boyunca tekrar edilmesi ile evrim gerçekleşir. Güçlü olan çocuğun hayatta kalması genetik algoritmanın temel mantığıdır [19]. Aslında kromozomların evrimini yansıtan bu süreç vasıtasıyla en çok uyum gösteren kromozomlar, organizmaları güçlü olan canlılar olarak nitelendirilir [20].
Doğal seçilimden ilham alan genetik algoritma ile en uygun çözüme ulaşılması temel hedeflerdendir [20].

2.3. Genetik algoritma yaklaşımı kullanılarak sentetik veri üretimi

Genetik algoritma, evrimden ilham alınarak yazılmıştır. Temel mantığı nesilden nesile aktarılan toplulukta kötü çözümlerin yok olması ve iyi çözümlerin varlığını devam ettirmesidir. Topluluk, kromozomlardan meydana gelen bir kümedir ve kromozomlar problem için olası çözümleri temsil eder. Kromozomların yer aldığı çözümlerde, uygunluk fonksiyonu vasıtasıyla çözümün kalitesi belirlenir. Bununla birlikte kaliteyi ve uygunluğu iyileştirmek amacıyla mutasyon ve çaprazlama uygulanan diğer işlemlerdir. Genetik işlemlerin uygulandığı bu süreç sonunda seçilen bireyler içerisinden en kaliteli çözümlerin elde edilmesi sağlanır.

Genetik algoritma her bir problem çeşidi için ortak bileşenlere sahiptir. Bu bileşenler, kodlama ve maliyet fonksiyonudur. Bu bileşenler problemin doğasına yönelik bir temsil sağlamalıdır. Problemlerin temsilinde kodlama yaklaşımı kapsamında ikili kodlama, permütasyon kodlaması, değer kodlaması ve ağaç kodlaması literatürde kullanılan yaklaşımlardır [18].

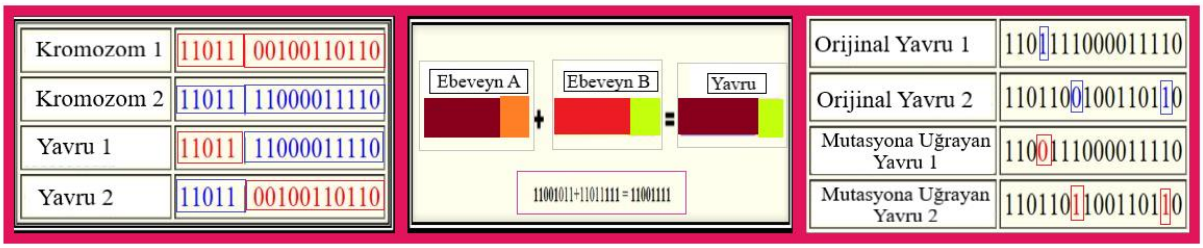
Bu çalışmada toplam 54 farklı öznitelikten oluşan boşanma öngörü veri kümesinde katılımcılar tarafından 0 ile 4 ölçeği arasında değişen bir değerlendirme yapısı kullanılarak katılımcıların değerlendirmeye katılımları sağlanmıştır ve bir öngörü oluşturulması hedeflenmiştir. Bu doğrultuda mevcut problem uzayının doğasına yönelik bir temsil sağlanması için 0 ile 4 arasında kodlama yapılmıştır ve matris oluşturulmuştur. Başlangıç popülasyonunda 0 ile 4 arasında değişen rastlantısal değerlerin üretilmesinden sonra popülasyondaki tüm kromozomların uygunluk değeri vasıtasıyla maliyet değeri hesaplanmıştır. Ardından yeni bir popülasyonu oluşturmak için seleksiyon, çaprazlama ve mutasyon işlemleri gerçekleştirilmiştir.

Seleksiyon işlemi, çiftleşme ve üreme için kromozomların seçilmesi işlemidir. Yeni yavruların üretilmesinde çaprazlamada kullanılacak ebeveynlerin belirlenmesi sağlanır [18]. Bu çalışmada seleksiyon işlemi için rastgele bir seçim yapılmıştır ve kromozom çifti seçilmiştir. Ardından yavruların oluşturulması için parça değişimi yapılmıştır. Çaprazlama ismindeki bu işlevsellik kapsamında tek noktalı çaprazlama türü seçilmiştir. Böylece ebeveynlerin belirli bir bölgesinin seçilmesi ve karşılıklı değiştirilmesi işlemi ile çeşitlilik sağlanmıştır [18].

Ek olarak, yerel optimumlardan kaçınmak ve uyarlanabilir yeni çözümler oluşturmak önemlidir. Bunun için çaprazlama işleminden sonra mutasyon gerçekleştirilmiştir. Mutasyon genetik operatörü ile kromozomdaki bir ya da daha fazla gene mutasyon işlemi uygulanmakta ve genin değiştirilmesi sağlanmaktadır [18]. Önerilen çalışmada veri kümesi uzayına uyarlanan mutasyon işlemi için aşağıda sunulan 5 madde ile çeşitlilik sağlanmıştır.

1- Mutasyon için seçilen gen değeri 4 ise 3 ile değiştir.
2- Mutasyon için seçilen gen değeri 3 ise 2 ile değiştir.
3- Mutasyon için seçilen gen değeri 2 ise 1 ile değiştir.
4- Mutasyon için seçilen gen değeri 1 ise 0 ile değiştir.
5- Mutasyon için seçilen gen değeri 0 ise 1 ile değiştir.

Genetik algoritma yapısının daha net anlaşılabilmesi için Binary (İkili) kodlama yapılan bir çözüm uzayında çaprazlama, çaprazlama yönteminin özel bir türü olan tek nokta çaprazlama ve mutasyon işlemlerinin görsel tasviri Şekil 1'de verilmiştir.



Şekil 1. Genetik algoritmada kullanılan çaprazlama, tek nokta çaprazlama ve mutasyon operatörlerinin tasviri [18]

Şekil 1'de sunulan görüntünün üç kutucuğu sırasıyla çaprazlama, tek nokta çaprazlama ve mutasyon genetik operatörlerini göstermektedir. Tüm bu işlemler çeşitliliğin sağlanması için önemli adımlardır.

Genetik algoritma temelli sentetik veri üretmek amacıyla inşa edilen bu çalışmada mevcut probleme yönelik bir temsil sağlanması için 0 ile 4 arasında yapılan kodlama ile oluşturulan çözüm uzayında iterasyon sayısı 100 olarak belirlenen sonlandırma kriteri sonucunda yeni bir nesil elde edilmiştir. Fakat yeni bir nesil oluşturma eyleminin yanı sıra elde edilen güncel neslin sahip olduğu kategorinin belirlenmesi de sentetik veri üretme aşamasının temel bir adımıdır. Bunun için maliyet fonksiyonu kullanılmıştır.

Maliyet fonksiyonu vasıtasıyla elde edilen maliyet değeri kullanılarak gerçekleştirilen değerlendirme işlevi, orijinal veri kümesinin analiz edilmesinin sonucunda performansın derecelendirilmesidir. Bu çalışmada maliyet değeri, boşanma tahminine ilişkin iki ayrı kategori için sınır değeri üreten bir parametredir. Bu doğrultuda yapılan analizler sonucunda 54 farklı özellik içeren ve 0 ile 4 arasında kodlanan ölçek için verilen puanların toplam değeri, maliyet fonksiyonu olarak nitelendirilmiştir. Aynı zamanda maliyet fonksiyonu kullanılarak orijinal veri kümesinde yapılan analizler sonucunda eşik değeri tanımlaması yapılmıştır ve eşik değeri 50 olarak belirlenmiştir. Böylece 50 eşik değerinin altında kalan maliyet fonksiyonu için kategori değeri 0 ile eşleştirilirken 50 ve üstü eşik değerine sahip maliyet fonksiyonu için kategori değeri 1 ile eşleştirilmiştir. Tüm bu işlemler sonucunda artan veri çeşitliliği ile sentetik veriler üretilmiştir.

3. Araştırma Bulguları ve Tartışma

Veri, endüstriyel inovasyonu geliştirmek için algoritmalar vasıtasıyla eğitilen modellerin bilgisayarlar ya da özel makineler tarafından işlevselleştirilmesinde önemli rol oynayan bilgi kümesidir. Fakat yetersiz veri erişimi, yasal düzenlemeler, etik kurallar, gizlilik prosedürleri, mahremiyet, veri paylaşım kısıtı ve maliyet; başarılı bir eğitimin ve güçlü bir çıkarımın gerçekleştirilmesinin önündeki engellerdir. Bu engelleri aşmak, araştırmaların hızlı yürütülmesini sağlamak ve araştırma süreçlerine pozitif etki sunmak için sentetik veri üretimi tercih edilen bir yaklaşımdır. Bununla birlikte sentetik veri üretme yaklaşımına ilişkin standart bir çerçeve yoktur [4]. Fakat yeni ve güncel yaklaşımların geliştirilmesine yönelik çalışmalar devam etmektedir.

Bu çalışmada genetik algoritma temelli yeni bir sentetik veri üretme yaklaşımı önerilmiştir. Böylece Boşanma Öngörü Veri Kümesi kapsamında 54 farklı özelliğin 0 ile 4 arasındaki bir ölçekte nitelendirildiği örneklere seleksiyon, çaprazlama ve mutasyon genetik faktörleri uygulanarak veri çeşitliliği arttırılmıştır. Ardından üretilen sentetik verilerin kategori bilgisi maliyet fonksiyonu ile tanımlanan bir eşik değeri kullanılarak belirlenmiştir ve orijinal veri kümesine sentetik veriler eklenerek zenginleştirilmiş yeni bir veri kümesi oluşturulmuştur. Son aşamada orijinal veri kümesi ve zenginleştirilmiş veri kümesi üzerinde makine öğrenmesi sınıflandırıcıları kullanılarak sınıflandırma yapılmıştır. Makine öğrenmesi sınıflandırıcıları kapsamında 170 veri içeren orijinal veri kümesinde ve 300 veri içeren zenginleştirilmiş veri kümesinde Eğitim (Eğitim K.) ve Test (Test K.) verileri için elde edilen sınıflandırma sonuçları Tablo 1-6'da verilmiştir.

Tablo 1. Orijinal ve zenginleştirilmiş veri kümeleri üzerinde Naive Bayes sınıflandırıcısı kullanılarak elde edilen performans ölçütleri

Naive Bayes Sınıflandırıcısı		Doğruluk Oranı	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
Orijinal Veri Kümesi	Eğitim K.	0.9909	1	1	1	1
	Test K.	0.95	0.9666	0.9655	0.9666	0.9666
Zenginleştirilmiş Veri Kümesi	Eğitim K.	0.9897	0.9793	1	1	0.9895
	Test K.	0.9904	0.9807	1	1	0.9902

Tablo 2. Orijinal ve zenginleştirilmiş veri kümeleri üzerinde Karar Ağacı sınıflandırıcısı kullanılarak elde edilen performans ölçütleri

Karar Ağacı Sınıflandırıcısı		Doğruluk Oranı	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
Orijinal Veri Kümesi	Eğitim K.	1	1	1	1	1
	Test K.	0.95	0.9354	0.9655	0.9666	0.9507
Zenginleştirilmiş Veri Kümesi	Eğitim K.	1	1	1	1	1
	Test K.	0.9904	0.9807	0.9811	0.9807	0.9807

Tablo 3. Orijinal ve zenginleştirilmiş veri kümeleri üzerinde Rastgele Orman sınıflandırıcısı kullanılarak elde edilen performans ölçütleri

Rastgele Orman Sınıflandırıcısı		Doğruluk Oranı	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
Orijinal Veri Kümesi	Eğitim K.	1	1	1	1	1
	Test K.	0.9666	0.9354	0.9807	1	0.9666
Zenginleştirilmiş Veri Kümesi	Eğitim K.	1	1	1	1	1
	Test K.	0.9904	0.9807	1	1	0.9902

Tablo 4. Orijinal ve zenginleştirilmiş veri kümeleri üzerinde KNN sınıflandırıcısı kullanılarak elde edilen performans ölçütleri

KNN Sınıflandırıcısı		Doğruluk Oranı	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
Orijinal Veri Kümesi	Eğitim K.	0.9818	0.9622	0.9793	1	0.9807
	Test K.	0.9666	0.9354	0.9807	1	0.9666
Zenginleştirilmiş Veri Kümesi	Eğitim K.	0.9897	0.9793	1	1	0.9895
	Test K.	0.9904	0.9807	1	1	0.9902

Tablo 5. Orijinal ve zenginleştirilmiş veri kümeleri üzerinde Lojistik Regresyon sınıflandırıcısı kullanılarak elde edilen performans ölçütleri

Lojistik Regresyon Sınıflandırıcısı		Doğruluk Oranı	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
Orijinal Veri Kümesi	Eğitim K.	1	1	1	1	1
	Test K.	0.9666	0.9354	0.9807	1	0.9666
Zenginleştirilmiş Veri Kümesi	Eğitim K.	1	0	1	1	1
	Test K.	0.9904	0.9807	1	1	0.9902

Tablo 6. Orijinal ve zenginleştirilmiş veri kümeleri üzerinde DVM sınıflandırıcısı kullanılarak elde edilen performans ölçütleri

DVM Sınıflandırıcısı		Doğruluk Oranı	Duyarlılık	Özgüllük	Kesinlik	F Ölçütü
Orijinal Veri Kümesi	Eğitim K.	1	1	1	1	1
	Test K.	0.9666	0.9354	0.9807	1	0.9666
Zenginleştirilmiş Veri Kümesi	Eğitim K.	1	1	1	1	1
	Test K.	0.9904	0.9807	1	1	0.9902

Tablo 1-6'da verilen eğitim sonuçları kapsamında Boşanma Öngörü Veri Kümesi dinamiğinde zenginleştirilen veri kümesi üzerinde makine öğrenmesi sınıflandırıcıları kullanılarak ulaşılan performans kriterleri değerlendirilmiştir. Bu kriterler doğrultusunda doğruluk oranı, iki ayrı boşanma tercihi için yapılan doğru tahminlerin tüm tahminlere oranını ifade etmektedir. Duyarlılık, boşanma durumuna ilişkin olumlu tercihler için gerçek çıktı değerleri ile bilgisayar destekli sistemler tarafından yapılan ortak tahminlerin tüm tahminlere oranını ifade etmektedir. Özgüllük, boşanma durumuna ilişkin olumsuz tercihler için gerçek çıktı değerleri ile bilgisayar

destekli sistemler tarafından yapılan ortak tahminlerin tüm tahminlere oranını ifade etmektedir. Kesinlik, iki ayrı boşanma tercihi için gerçekleştirilen tahminlerin ayırt etme gücünü ifade etmektedir. F Ölçütü, kesinlik ve duyarlılık değerlerinin harmonik ortalamasıdır. Uç değerlerin dikkate alınmasını sağlar. Bu performans ölçütlerine ilişkin matematiksel ifadeler, denklem 1-5'te verilmiştir [21].

Doğruluk Oranı = (DP+DN) / (DP+YP+DN+YN) (1)

Duyarlılık = (DP) / (DP+YN) (2)

Özgüllük = (DN) / (DN+YP) (3)

Kesinlik = (DP) / (DP+YP) (4)

F Ölçütü = (2*kesinlik*duyarlılık) / (kesinlik+duyarlılık) (5)

Doğru Pozitif (DP), Doğru Negatif (DN), Yanlış Pozitif (YP) ve Yanlış Negatif (YN) ölçütleri kullanılarak değerlendirilen Doğruluk Oranı, Duyarlılık, Özgüllük, Kesinlik ve F Ölçütü kriterlerinin %100 ya da %100'e yakın bir sonuç üretmesi önerilen yaklaşımın başarısını gösterir. Bu doğrultuda Tablo 1-6 incelendiğinde genetik algoritma temelli sentetik veri üretme yaklaşımı ile oluşturulan zenginleştirilmiş veri kümesinin sırasıyla Naive Bayes, Karar Ağacı, Rastgele Orman, K-Nearest Neighbors (KNN), Lojistik Regresyon ve Destek Vektör Makinesi (DVM) sınıflandırıcıları vasıtasıyla ulaşılan sonuçları üzerinde bir iyileştirme sağladığı görülmektedir. Zenginleştirilen veri kümesi üzerinde maksimum duyarlılık ölçütü, Destek Vektör Makinesi sınıflandırıcısı ile %100 olarak elde edilmiştir. Aslında sınıflandırma performansı veri kümesinin veri sayısı ile ilişkilidir. Çünkü orijinal veri kümesinin dinamiğinde artan veri sayısı ile orantılı olarak benzer ilişkiler, benzer desenler ve benzer çıkarımlar daha güçlü bir şekilde öğrenilmiştir. Bunun sonucunda da eğitim başarısı iyileştirilmiştir ve daha kararlı ve istikrarlı çıkarımlar yapılmıştır. Aksi durumun gerçekleşme ihtimali performans başarısında bir azalma eğilimi göstereceği için önerilen sentetik veri üretme yaklaşımının literatüre katkı sağlaması hedeflenmektedir.

4. Sonuç

Veri, yapay zeka temelli çalışmalarda kritik bir öneme sahiptir. Amaca uygun çıktıların üretilmesinde, modelin başarılı bir şekilde eğitilmesinde, yapay zeka çözümlerinin hayatın içerisine dahil edilmesinde, sorunlara verimli çözüm noktaları oluşturulmasında ve düşük maliyet sunan bir karar destek sisteminin inşa edilmesinde büyük bir potansiyel barındırmaktadır. Fakat yetersiz veri erişimi, yasal düzenlemeler, etik kurallar, gizlilik prosedürleri, mahremiyet, veri paylaşım kısıtı ve maliyet; veriye ulaşmanın önündeki engellerdir. Bununla birlikte sentetik veri üretme yaklaşımına ilişkin standart bir çerçeve de yoktur. Fakat yeni ve güncel yaklaşımların geliştirilmesine yönelik çalışmalar devam etmektedir.

Bu çalışmada genetik algoritma temelli yeni bir sentetik veri üretme yaklaşımı geliştirilmiştir. Önerilen sentetik veri üretme yaklaşımı kullanılarak orijinal veri kümesinin dinamiğinde zenginleştirilen veri kümesi üzerinde 6 farklı makine öğrenme sınıflandırıcısı vasıtasıyla önerilen yaklaşımın başarısı değerlendirilmiştir. Değerlendirme sonucunda zenginleştirilen veri kümesi üzerinde maksimum duyarlılık ölçütü, Destek Vektör Makinesi sınıflandırıcısı ile %100 olarak elde edilmiştir. Bu durum veri sayısının artarak bir doyum noktasına ulaşması sonucunda veriler arasındaki ilişki ve çıkarımın artmasına ve gerçekleştirilen sınıflandırma işlemi sonucunda performansın iyileştirilmesine işaret etmektedir. Çünkü yeterli veri sayısı, yapay zeka tabanlı yöntemlerin potansiyelini ortaya çıkarır ve endüstriyel inovasyonun sınırlarını zorlar.

Gelecekte, çok yönlü genetik operasyona sahip olan genetik algoritmanın sahip olduğu çeşitlilikten ilham alarak sürekli verilerden oluşan veri kümesi örnekleri için de çalışmanın genişletilmesi planlanmaktadır.

Kaynaklar

[1] Mavrogenis AF, Scarlat MM. Artificial intelligence publications: synthetic data, patients, and papers, Int Orthop 2023; 47:1395–1396.

[2] Hashimoto DA, Ward TM, Meireles OR. The Role of Artificial Intelligence in Surgery. Adv. Surg 2020; 54:89–101.

[3] Shah S, Gandhi D, Kothari J. Machine learning based Synthetic Data Generation using Iterative Regression Analysis. Proc. 4th Int. Conf. Electron. Commun. Aerosp. Technol ICECA 2020; pp. 1093–1100.

[4] Lu Y, Shen M, Wang H, Wei W. Machine Learning for Synthetic Data Generation : A Review. arXiv 2021; 14(8): 1–18.

[5] Pacheco F. et al. Generation of Synthetic Data for the Analysis of the Physical Stability of Tailing Dams through Artificial Intelligence. Mathematics 2022; 10(23):1–15.

[6] Belke M, Blanke P. Storms S, Herfs W. Object pose estimation in industrial environments using a synthetic data generation pipeline, Proc. - 2022 6th IEEE Int Conf Robot Comput IRC 2022; pp. 435–438.

[7] Ucuzova E, Kurtulmaz E, Gokalp Yavuz F, Karacan H, Sahin NE. Synthetic CANBUS data generation for driver behavior modeling. 29th IEEE Conf. Signal Process. Commun Appl Proc SIU 2021; pp. 28–31.

[8] Nicholson AD, Peplow DE, Ghawaly JM, Willis MJ, Archer DE. Generation of Synthetic Data for a Radiation Detection Algorithm Competition. IEEE Trans. Nucl. Sci 2020; 67(8): 1968–1975.

[9] Pérez-Porras FJ, Triviño-Tarradas P, Cima-Rodríguez C, Meroño-De-larriva JE, García-Ferrer A, Mesas-Carrascosa FJ. Machine learning methods and synthetic data generation to predict large wildfires. Sensors 2021; 21:1–19.

[10] Mahmood A, Bennamoun M, An S, Sohel F, Boussaid F, Hovey R, Kendrick G. Automatic detection of western rock lobster using synthetic data. ICES Journal of Marine Science 2020; 77(4): 1308–1317.

[11] Nabati M, Navidan H, Shahbazian R, Ghorashi SA, Windridge D. Using Synthetic Data to Enhance the Accuracy of Fingerprint-Based Localization: A Deep Learning Approach. IEEE Sensors Lett 2020; 4(4):1–4.

[12] Khan AR, Khan S, Harouni M, Abbasi R, Iqbal S, Mehmood Z. Brain tumor segmentation using K-means clustering and deep learning with synthetic data augmentation for classification. Microsc. Res. Tech. 2021; 84(7): 1389–1399.

[13] Douzas G, Lechleitner M, Bacao F. Improving the quality of predictive models in small data GSDOT: A new algorithm for generating synthetic data. PLoS One 2022; 17(4):1–15.

[14] Arab N, Nemmour H, Chibani Y. A new synthetic feature generation scheme based on artificial immune systems for robust offline signature verification. Expert Syst Appl 2023; 213.

[15] İmak A, Doğan G, Şengür A, and Ergen B. Asma Yaprağı Türünün Sınıflandırılması için Doğal ve Sentetik Verilerden Derin Öznitelikler Çıkarma, Birleştirme ve Seçmeye Dayalı Yeni Bir Yöntem. Int J Pure Appl Sci 2022; 9(1): 46–55.

[16] UCI (the University of California Irvine Machine Learning Repository), https://archive.ics.uci.edu/.

[17] Turgun FS, Zorlu H. Parçacık Filtresinin Optimizasyonu için Genetik Algoritma Tabanlı Yeni Bir Yaklaşım/A New Approach Based on Genetic Algorithm for Optimization of Particle Filter. Bozok J Eng Archit 2023; 2(1):24–33.

[18] Hassanat A, Almohammadi K, Alkafaween E, Abunawas E, Hammouri A, Prasath VBS. Choosing mutation and crossover ratios for genetic algorithms-a review with a new dynamic approach. Information 2019; 10:1–36.

[19] Altay A. Genetik Algoritma ve Bir Uygulama, Yüksek Lisans Tezi, İstanbul Teknik Üniversitesi, İstanbul, 2007.

[20] Zhou J, Huang S, Zhou T, Armaghani DJ, Qiu Y, Employing a genetic algorithm and grey wolf optimizer for optimizing RF models to evaluate soil liquefaction potential. Artificial Intelligence Review 2022; 55: 5673-5705.

[21] Akalın F, Sayısal Haritalama Teknikleri Kullanılarak DNA Dizilimleri Üzerinden Lösemi Hastalığının Temel Türlerinin Yapay Zeka Tabanlı Algoritmalar ile Sınıflandırılması, Doktora Tezi, Sakarya Üniversitesi, Sakarya, 2023.