Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 14(4), 1408-1431, 2024
Journal of the Institute of Science and Technology, 14(4), 1408-1431, 2024

ISSN: 2146-0574, eISSN: 2536-4618

Bilgisayar Mühendisliği / *Computer Engineering* **DOI: 10.21597/jist.1495455**

Araştırma Makalesi / *Research Article*

Geliş tarihi / *Received*: 04.06.2024 **Kabul tarihi / *Accepted*: 12.09.2024**

Atıf İçin: Akalın, F., (2024). Değiştirilmiş Yapay Arı Kolonisi Optimizasyon Algoritması ve İstatistiksel Modelleme ile Sentetik Veri Üretimi. *Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, 14(4), 1408-1431.

To Cite: Akalın, F., (2024). A Synthetic Data Generation Approach Based on a Modified Artificial Bee Colony Optimization Algorithm and Machine Learning Methods . *Journal of the Institute of Science and Technology*, 14(4), 1408-1431.

Değiştirilmiş Yapay Arı Kolonisi Optimizasyon Algoritması ve İstatistiksel Modelleme ile Sentetik Veri Üretimi

Fatma AKALIN[1*]

Öne Çıkanlar:
- Sentetik Veri,
- Optimizasyon Algoritması,
- Makine Öğrenmesi

Anahtar Kelimeler:
- Sentetik veri üretimi
- Değiştirilmiş ABC optimizasyon algoritması,
- R-kare katsayısı,
- Polinom regresyonu,
- Karar ağacı sınıflandırıcısı

ÖZET:

Makine öğrenmesi, gerçek yaşam verilerini analiz etmede ve değerlendirmede kullanılan güçlü bir karar destek sistemidir. Bu sistem, yeni çözümler üretmeyi ve performansı iyileştirmeyi amaçlamaktadır. Bu nedenle, veri bilimi alanıyla ilişkilidir. Bu ilişki temelinde veriler vardır. Verilerden anlamlı içgörüler çıkarma etkinliği, model eğitiminin kalitesine bağlıdır. Bu performansı iyileştirmek için, veriler arasındaki kombinasyonların çeşitliliği ve veri kümesindeki toplam veri sayısı artırılmalıdır. Ancak bu konuda, yetersiz veri erişimi, yasal düzenlemeler, etik kurallar, gizlilik prosedürleri, gizlilik, veri paylaşımı kısıtlamaları ve maliyet parametreleri engellerdir. Tüm bu sorunları çözmek, işlevselliği iyileştirmek ve güçlü makine öğrenimi çıkarımları sağlamak için sentetik veri üretimi, veri bilimi alanında temel bir adımdır. Bu nedenle, bu çalışmada 3 temel aşamadan oluşan yeni bir sentetik veri üretimi yaklaşımı önerilmiştir. İlk aşamada, orijinal verilerin dağılımına benzer şekilde sentetik veri üretimi, modifiye edilmiş ABC (Yapay Arı Kolonisi) optimizasyon algoritması ile gerçekleştirilmiştir. İkinci aşamada, üretilen yapay veriler arasında regresyon yöntemleriyle analiz edilen istatistiksel değerlendirme ile bağımsız değişkenlerin kategori bilgileri belirlenmiştir. Üçüncü aşamada, üretilen yapay verilerin etkinliği ve uygulanabilirliği, makine öğrenimi sınıflandırıcıları ile değerlendirilmiştir. Değerlendirme sonucunda, önerilen sentetik veri üretim yönteminin, veri sayısının artışı ile orantılı olarak makine öğrenmesi sınıflandırıcılarının performansını artırdığı kanıtlanmıştır Maksimum performans gösteren karar ağacı algoritması, zenginleştirilmiş 5 ayrı veri kümesi üzerinde sırasıyla %100, %92.5, %100, %85, %66 başarı oranları üretmiştir.

Synthetic Data Generation with Modified Artificial Bee Colony Optimization Algorithm and Statistical Modeling

Highlights:
- Synthetic data,
- Optimization algorithm,
- Machine learning

Keywords:
- Synthetic data generation,
- Modified ABC optimization algorithm,
- R-squared coefficient,
- Polynomial regression,
- Decision tree classifier

ABSTRACT:

Machine learning is a powerful decision support system used in analyzing and evaluating real-life data. This system aims to create new solutions and improve performance. Therefore, it is related to the field of data science. There are data on the basis of this relationship The effectiveness of drawing meaningful insights from data depends on the quality of the model's training. To improve this performance, the variety of combinations among the data and the total number of data in the dataset should be increased. But in this topic, insufficient data access, legal regulations, ethical rules, confidentiality procedures, privacy, data sharing restrictions and cost parameters are obstacles. Synthetic data generation is a basic step in the field of data science in order to solve all these problems, improve functionality and provide powerful machine-learning inferences. Therefore, a new synthetic data generation approach consisting of 3 basic stages is proposed in this study. In the first stage, synthetic data production similar to the distribution of the original data was carried out with the modified ABC (Artificial Bee Colony) optimization algorithm. In the second stage, the category information of the independent variables was determined by the statistical evaluation analyzed with regression methods among the artificial data produced. In the third stage, the efficiency and applicability of the artificial data produced were evaluated with supervised machine learning classifiers. As a result of the evaluation, it has been proven that the proposed synthetic data generation approach improves the performance of machine learning classifiers in proportion to the increasing number of data. The decision tree algorithm that showed maximum performance produced success rates of 100%, 92.5%, 100%, 85%, and 66% on 5 separate enriched datasets, respectively.

[1] Fatma AKALIN (**Orcid ID:** 0000-0001-6670-915X), Sakarya University, Faculty of Computer and Information Sciences, Department of Information Systems Engineering Sakarya, Türkiye

***Sorumlu Yazar/*Corresponding Author**: Fatma AKALIN, e-mail: fatmaakalin@sakarya.edu.tr

Etik Kurul Onayı / *Ethics Committee Approval*: There is no requirement for an ethics committee.

INTRODUCTION

Machine learning is a framework for managing and analyzing data. By this framework, target tasks are performed by machines through algorithms and statistical models. In machine learning, a classification task is used to predict a discrete value output, and a regression task is used to model the relationship between continuous variables. These tasks, considered in the scope of supervised learning build a training process using labeled input data (Hashimoto, 2020; El Mrabet et al., 2021).

Recently, machine learning has been a popular field preferred for performing complex tasks (Kinaneva et al., 2021). Through this area, tasks in intelligent computer systems can be integrated into machines autonomously, adapted to many real-life problems, analyzed through real-life problems, and given answers to target problems with the inferences obtained as a result of the analysis. However, there are obstacles in appearing the potential related to the machine learning approach. Poor quality of data, wrong supply, unbalanced number of samples between categories and insufficient data input are the factors that negatively affect the success of the analysis process. For the development and implementation of the machine learning framework, solutions must first be found for these problems. Data collection and adding a description is a recommended solution. But this solution is both time-consuming and expensive. At the same time, it is not used in areas where the risk of revealing sensitive and critical information is not preferred (Lu et al., 2021).

Generating synthetic data is a fundamental step in data science and protects security and privacy. At the same time, more successful training is carried out with the rich data variety created for datasets with a small number of samples. This is an important solution for different real-world applications where insufficient data is available. Innovation is achieved through this approach, increasing functionality (Lu et al., 2021). In order to contribute to this development and remove the barriers that limit the power of machine learning, new approaches to synthetic data generation have been developed. In this context, the MC-GEN approach was proposed in the (Li et al., 2023) study. By the MC-GEN approach, which has feature-level clustering, sample-level clustering, privacy sanitizer, and generative model components, synthetic data is produced under a differential privacy guarantee to protect privacy. It has been stated that synthetic datasets produced using the MC-GEN approach with successful parameter tuning show similar performance to the original datasets. Synthetic Data Vault (SDV) system is proposed in the (Patki et al., 2016) study, which creates a model by repetition of all possible relationships between the data in the database. Through this system, data are synthesized using the sample space taken from a part of the database. However, it has been stated that the boundaries of the SVD are not always correct in a noisy parent table. In order to construct a structurally and statistically similar dataset around the privacy procedure, the DataSynthesizer tool is presented in (Ping et al., 2017) study. By this tool, which consists of DataDescriber, DataGenerator and ModelInspector modules, the data types, correlations and distributions of the attributes in the dataset are evaluated, and a data summary is produced by preprocessing to protect confidentiality. Then, samples from this summary are used and synthetic data are obtained. In the final stage, the summarization process is evaluated by the data owner. It is stated that this open-source system is used as a component in various applications. In the (Dahmen & Cook, 2019) study, which is based on simple and realistic concepts, a synthetic data generation method called SynSys is proposed. It is stated that this proposed approach for synthetic data generation with the combination of Hidden Markov Models and regression algorithms produces realistic data. In the (Li et al., 2020) study is suggested the SYNC approach. This study using state-of-the-art machine learning and statistical methods, states that the proposed data are very similar and consistent with real population behaviours. A software tool named HAPNEST has been developed for human genetics applications in

the (Wharrie et al., 2022) study which aims to create synthetic datasets using public datasets. As a result of the evaluation done with Kinship analysis and IBS analysis, it was stated that the HAPNEST software tool provides a balance for relevance and sample diversity. In the (Douzas et al., 2022) study, the Geometric Small Data Oversampling Technique was proposed to provide a successful analysis of datasets with a small number of samples. With this technique, new samples were created using the geometric regions around the existing data. It has been stated that this study, which is an open-source project, is more successful than standard synthetic data generation approaches. The (Arab et al., 2023) study aimed to authenticate using offline handwriting. The proposed approach includes mutation, cloning, and resource competition mechanisms of artificial immune systems. It has been stated that this approach provides an improvement of %8. These studies generally investigate various synthetic data generation methods to provide better performance with small data sets. For example, methods such as MC-GEN, SDV, DataSynthesizer offer different approaches to data generation. On the other hand, SynSys, SYNC, HAPNEST and Geometric Small Data Oversampling Technique suggest various methods to obtain more realistic and high-performance results in data generation. In this direction, each study makes significant contributions to synthetic data production with its own specific methods and approaches. Also, these studies in the literature indicate that the synthetic data generation approach offers a suitable solution for data analysis. Because there are strict privacy regulations such as the General Data Protection Regulation (GDPR), and these regulations present a complex and expensive process to obtain and store data. It is also not considered a reliable approach to protecting sensitive data as it contains identifiable information (Douzas et al., 2022). In addition, some sectors can not obtain enough data for a successful analysis within the scope of the target subject. As a result of this, the size of the sample can not contain enough data for a successful analysis. In this context, there is a practical rule called “a dumb algorithm with lots and lots of data beats a clever one with modest amounts of it”. This rule explains that reliable and successful performance cannot be expected from small sample-size datasets (Douzas et al., 2022). As well as the insufficient sample size, the unbalanced learning problem also negatively affects the training of machine learning models (Douzas et al., 2022). For this reason, sufficient and balanced datasets should be studied to make a reliable assessment.

Synthetic data generation is a fundamental step, especially in the field of data science. But, there is no standard framework in this scope (Li et al., 2020) and, studies related to the utility of synthetic data continue. These studies are developing in two separate directions, designing a new mechanism for generating synthetic data and, evaluating the performance of existing synthetic data generators (Dankar & Ibrahim, 2021). In addition to this situation, the insufficient number of samples in data sets creates significant difficulties in machine learning and data analysis processes. Because the sample diversity and distribution of data points required for models to learn effectively and make reliable predictions are often limited. The failure to fully reveal the relationships between the data and to clearly identify patterns weakens the generalization ability of the model. Insufficient number of samples, especially in data sets with complex relationships or rare events, is an obstacle for the model to learn the critical points and produce reliable results in real-world applications. Synthetic data generation is also critical for such scenarios. Synthetic data generation improves the performance of the model by ensuring that relationships and patterns between data are accurately revealed and contributes to more reliable results.

In this study, artificial data generation was performed for 2 and 3 category classes in datasets where the number of sample data varied between a maximum of 90 and a minimum of 14. In this direction, firstly, synthetic data showing a distribution similar to the original dataset were produced with the artificial bee colony optimization algorithm. Then, with the help of regression algorithms, the statistical features, relationships and patterns captured specific to the category among the examples in the original

data set were also discovered among the synthetic data. At the end of this discovery process, category assignment was made for synthetic data. Thus, by providing various data for training, the capacity of the model to learn different situations and patterns was increased. Additionally, generalization ability has been improved. For this reason, the fact that the enriched dataset shows more performance than the original dataset within the scope of performance criteria obtained using classification algorithms is a criterion that shows the conformity of the produced synthetic data to the original data. That is, it is success of the proposed data generation approach. In this context, as a result of the classification process performed with k nearest neighbor, logistic regression, support vector machines, decision tree and random forest classification algorithms on 5 different data sets, an improvement was achieved on the performance criteria specific to each data set. The conducted experiments show the usability of the proposed methodology within the sample and category boundaries determined in the categories of other, computer science, social science, health and medicine.

MATERIALS AND METHODS

Synthetic data generation is an appropriate approach for conditions where data collection is difficult and costly. It allows for an increase in the available data, produces a balanced distribution among the categories, improves educational success in direct proportion to the increasing amount of data, and produces a safe sample for situations with confidentiality regulations and ethical boundaries (Douzas et al., 2022). Therefore, a novel synthetic data generation approach a fundamental step of data analysis (Li et al., 2020), has been developed in this study. In the scope of this approach, the datasets used in the study, the artificial bee colony optimization algorithm, the modified artificial bee colony optimization algorithm, and regression and classification processes are explained in detail below.

Datasets

Knowledge, experience and intelligence functions are human characteristics, and the realization of these characteristics by machines is among the future goals for many sectors. However, there are some obstacles to the real-life adaptation of these goals planned to make inferences from the data. For example, some industries may not be able to generate sufficient sample sizes for successful analysis based on their target problems (Douzas et al., 2022) or obtaining and storing data for industries with strict confidentiality regulations presents a complex and expensive process. Therefore, the synthetic data generation stage is a necessary step for many fields. But, there isn't a standard framework for this stage (Li et al., 2020).

In this study, a new synthetic data generation approach has been developed. The success of this developed approach was analyzed on datasets in different subjects with insufficient sample sizes. In this direction, data sets with sample numbers varying between a maximum of 90 and a minimum of 14 were selected. In order to investigate the functionality of the proposed methodology, data defined in the categories of other, computer science, social science, health and medicine were studied. In this context, datasets were obtained from the public gene bank (UCI, 2024a), UCI (the University of California Irvine Machine Learning Repository). All datasets used in the study are explained in detail below.

Lenses public dataset (UCI, 2024b)

It is a dataset related to contact lens selection. A decision is made about the type of lens to be used by the attributes presented in the dataset. The attributes are the age of the patient, spectacle prescription, astigmatic, and tear production rate, and the decision is evaluated in the three different categories. These categories are the patient should be fitted with hard contact lenses, the patient should be fitted with soft contact lenses, the patient should not be fitted with contact lenses. There are 24 samples in this dataset.

But, the number of samples is not enough to decide the actual category type. For this reason, the distribution information of the data in the original data set and the relationship of the data with the categories were analyzed. Then, an enriched dataset was created using this information.

Covid-19 surveillance dataset (UCI, 2024c)

This dataset is Coronavirus Disease (COVID-19) Surveillance dataset. It provides guidance on the prevention and control of coronavirus disease (COVID-19) by the attributes presented in the dataset. The attributes are given as A01, A02, A03, A04, A05, A06 and A07, and the decision is evaluated in 3 different categories. These categories are PUS, PIM, and PWS. But, there is only one example for the PWS type. Since there aren't different samples for the PWS type, successful relationships are not made between the attributes of the PWS type and, successful statistical inferences are not made. For this reason, the dataset presented a total of 14 data was evaluated with 13 samples and 2 categories. But, the number of examples is insufficient to decide the actual category type. For this reason, the distribution information of the data in the original data set and the relationship of the data with the categories were analyzed. Then, an enriched dataset was created using this information.

Ballons dataset (UCI, 2024d)

It is a dataset used for cognitive psychology experiments. It consists of 4 sub-datasets. Sub-datasets contain different conditions of an experiment, but these have the same attributes. The decision is made with the attributes presented in the data set. The attributes are color, size, act and age, and the decision is evaluated in 2 separate categories. These categories are inflated = T and inflated = F. There are 16 samples in each sub-dataset of this dataset. But, the number of examples is insufficient to decide the actual category type. For this reason, the distribution information of the data in the original data set and the relationship of the data with the categories were analyzed. Then, an enriched dataset was created using this information.

Caesarian section classification dataset (UCI, 2024e)

It is a data set that provides information about the results of cesarean section on pregnant women who have the most important features of delivery problems in the medical field. It is decided whether there will be a cesarean delivery or not by attributes presented in the data set. The attributes are age, delivery number, delivery time, blood of pressure and heart problem, and the decision is evaluated in 2 separate categories. These categories are caesarian=No and caesarian=Yes. There are 80 samples in this dataset. But, the number of examples is insufficient to decide the actual category type. For this reason, the distribution information of the data in the original data set and the relationship of the data with the categories were analyzed. Then, an enriched dataset was created using this information.

Post-operative patient dataset (UCI, 2024f)

It is a data set that provides information about the surgical process of patients in the postoperative recovery area. It is decided where the patients should be sent after the surgery by the attributes presented in the dataset. Attributes are L-CORE (patient's internal temperature in C), L-SURF (patient's surface temperature in C), L-O2 (oxygen saturation in %), L-BP (last measurement of blood pressure), SURF-STBL (stability) of patient's surface temperature), CORE-STBL (stability of patient's core temperature), BP-STBL (stability of patient's blood pressure), COMFORT (patient's perceived comfort at discharge, measured as an integer between 0 and 20), and the decision is evaluated in 3 separate categories. These categories are I (patient sent to Intensive Care Unit), S (patient prepared to go home), and A (patient sent to general hospital floor). But, there is only one example for the I (patient sent to Intensive Care Unit). In addition, there are also samples with missing data. In such a case, since there aren’t different cases

for the I type, successful relationships between the attributes of this data are not made, and successful statistical analyzes are not performed. This case is also discussed in missing data. For this reason, the dataset presented as 90 with a total number of data was evaluated over 86 samples and 2 categories. But, the number of examples is insufficient to decide the actual category type. For this reason, the distribution information of the data in the original data set and the relationship of the data with the categories were analyzed. Then, an enriched dataset was created using this information.

In this study, enriched datasets in the scope of Social, Life, Computer and Other subject fields were used. Then, the success of the proposed synthetic data generation approach was tested with classification algorithms. At the classification stage, 60% and 40% of the dataset were separated as training and test datasets, respectively. At the end of the study, the success of the proposed approach was evaluated.

Artificial Bee Colony Optimization Algorithm

It is a popular optimization algorithm inspired by the foraging behaviour of honey bees (Akay et al., 2021). It was developed by Derviş Karaboğa in 2005 (Alvarado-Iniesta et al., 2013). It presents a mathematical analysis of the hierarchy between the food sources around the hive and the bees. The main goal is to find optimal solutions. This process is carried out with information sharing by employed bees, onlooker bees and scout bees in the honey bee colony (Akay et al., 2021). The most important factor in sharing information among bees is the dance area. Information about the location and quality of the food source is shared through the dance area. However, the working period of bees with self-organize is not simultaneous. All stages of the model that summarizes the working period are presented below (Karaboğa, 2020).

1-The scout bees search randomly around the hive to find food.

2- After the food sources are found, the scout bees turn into employed bees, and the discovered food is carried to the hive by the employed bees.

3- After the food sources are left to the hive by the employed bees, two different options emerge. The first option is to return the employed bee to the source. The second option is to present the information about the source to the onlooker bees in the hive with the figures made by employed bees in the dance area.

4- The figures of dance are watched by the onlooker bees, and a choice is made between the discovered sources through the inferences made from the figures.

The intelligent behaviours performed by the bees among the food sources around the hive are modelled by the ABC (Artificial Bee Colony) algorithm. Each step of the ABC algorithm is analyzed in detail below.

Production of food sources

In order to construct the mathematical model of the ABC algorithm, the search space must be created. The search space is the area that has the food resources. In this field, random places are generated as the amount defined using the parameters whose lower and upper limits are determined. The mathematical expression of this step is given in Equation 1.

$$X_{ij} = x_j^{min} + \text{rand}(0,1) \quad x_j^{max} - x_j^{min} \tag{1}$$

The i variable in Equation 1 is a parameter determined between 1 and FS. FS is the total number of food sources. The variable j is a randomly generated integer between 1 and PN. The PN variable represents the total number of parameters to be optimized. After the creation of the areas for food

sources, all the bees employed in the hive work in coordination, and the optimum solution is produced (Karaboğa, 2020).

Sending to food sources of worker bees

In the process of constructing the model of the ABC algorithm, some assumptions were made. The first assumption is that the food at each source must be supplied by an employed bee. With this assumption, the number of employed bees is accepted as equal to the total number of food. The second assumption is that the total number of employed bees must be equal to the total number of onlooker bees. In the model, food sources are solutions, and the source with the most food is the optimum solution. Each worker bee evaluates and compares the food status of its neighbouring food sources (Karaboğa, 2020). The mathematical expression of this situation is given in Equation 2.

$$V_{ij}=x_{ij}+\boldsymbol{\Phi}_{ij}(x_{ij}-x_{kj}) \quad k\neq i \tag{2}$$

The k≠i in Equation 2 shows that two different points are not equal. The V_i source is found by changing the parameter of each source expressed with X_i. The V_i source is the neighbour of X_i. The difference between the j th parameter of the X_k neighbour solution and the j th parameter of the current source is taken and multiplied by $\boldsymbol{\Phi}$. The k parameter is determined between 1 to PN, and $\boldsymbol{\Phi}$ takes a random value between -1 and +1. The decreasing difference between the x_{ij} and x_{kj} shows that the solutions are similar. As a result of resembling each other the solutions, the amount of change will adaptively decrease. Thus, the lower and upper limit values of the j^{th} parameter will be shifted (Karaboğa, 2020). This situation is explained in Equation 3.

$$v_{ij}=\begin{cases} x_j^{min}, & v_{ij} < x_j^{min} \\ v_{ij}, & x_j^{min} \leq v_{ij} \leq x_j^{max} \\ x_j^{max}, & v_{ij} > x_j^{max} \end{cases} \tag{3}$$

V_i and X_i given in Equation 3 represent new and old sources, respectively. The new and old answers obtained by implementing the ABC algorithm are given in equations 4 and 5, respectively.

$$v_i = (v_{i1}, v_{i2}, v_{i3}, \ldots, v_{iFS}) \tag{4}$$

$$x_i = (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{iFS}) \tag{5}$$

The foraging process of the bees employed in the hive is provided through iteration cycles. At the end of each cycle, the "counter of failure in develop of a solution" is checked. In the condition that the counter is above the determined limit value, it is considered that the food is finished in the related region. As a result of this, the employed bee turns into a scout bee, and the random solution search process starts again. In case the counter does not exceed the determined limit value, the quality of the v_i source produced between the minimum and maximum limits is calculated, and then the fitness value is assigned. The mathematical expression of the fitness value that presents the output related to the quality of the source is given in equation 6 (Karaboğa, 2020).

$$\text{fitness}_i = \begin{cases} \frac{1}{1+f_i}, & f_i \geq 0 \\ 1+abs(f_i), & f_i < 0 \end{cases} \tag{6}$$

The f_i is the cost value of the vi resource. The amount of food between the x_i and v_i sources represents the fitness value. According to the fitness value, whether the new solution is better than the old solution is investigated. If it offers a more successful solution, the information of the new resource

is memorized. Otherwise, the employed bee goes to the x_i source, but it cannot develop a solution. When the "counter of failure in develop of a solution" reaches the limit value for x_i source, this area is abandoned (Karaboğa, 2020).

Calculating probability values by onlooker bees and choosing food source regions

Information about the resources discovered by the employed bees is transferred to the onlooker bees with dance figures. Then, probabilistic region selection is performed in direct proportion to the quality of the source. In the basic ABC algorithm, probabilistic region selection is done with the roulette wheel. The fitness value determined for each source is proportional to the angles of the slices on the roulette wheel. The mathematical expression of this ratio is given in Equation 7 (Karaboğa, 2020).

$$P_i=\frac{fitness_i}{\sum_{j=1}^{PN} fitness_j} \quad (7)$$

The quality of the source in Equation 7 is expressed by the $fitness_i$ parameter. As the quality of the source increases, the P_i value will increase, and the probability of selecting the related region will increase. After selection by the onlooker bees, the new solution will be evaluated and their suitability compared. As a result of the new solution being better, the new state will be maintained. This will continue until the cycle control limit (counter of failure in develop of a solution) is reached (Karaboğa, 2020).

ABC algorithm based on swarm intelligence is simple, flexible, and it is a controllable algorithm with few control parameters. It aims to find the global optimum values without getting stuck to local optimum values. It is adaptable to real-world problems (Karaboğa, 2020) . In this study, the contribution of the artificial bee colony optimization algorithm to the synthetic data generation process is explained under the title “Adaptatıon of Artıfıcıal Bee Colony Optımızatıon Algorıthm to Synthetıc Data Generatıon Process”.

Adaptation of Artificial Bee Colony Optimization Algorithm to Synthetic Data Generation Process

ABC algorithm is a successful and popular optimization algorithm preferred for solving real-world problems. It is used in numeric, binary, integer, mixed integer and combinatorial optimization problems. At the same time, there are ABC algorithm-based studies for routing, rule mining, team orienteering, timetabling, travelling salesman and vehicle routing in the scope of combinatorial optimization (Kaya et al., 2022). It is inspired by these studies. Then the contribution of the artificial bee colony optimization algorithm to the synthetic data production process, which keeps light on the solution of real-world problems, is evaluated. This contribution is explained in detail below.

In the first step of the improved ABC algorithm to generate synthetic data, the search space is constructed. For this, random locations are generated in the defined number using the parameters whose lower and upper limits are determined. The mathematical expression of this is given in equation 8.

$$X_i=\text{unifrnd}(\text{min},\text{max},\text{size}) \quad (8)$$

The i variable given in Equation 8 is a parameter determined between 1 and FS. FS is the total number of food sources. In the improved ABC algorithm using Equation 1, the food source is created in the total "size" amount and between the parameters whose minimum and maximum values are defined. Then, the whole region is evaluated by equations given in Equation 2 and Equation 3.

The foraging process of all the bees employed in the hive is provided throughout the cycles. At the end of each cycle "counter of failure in develop of a solution" is checked. In case the counter exceeds the determined limit value, the employed bee turns into a scout bee, and the random solution search

process starts again. Otherwise, the quality of the source is evaluated with the fitness function. The fitness value is found using the equation given in equation 9 (Karaboğa, 2022) . For this reason, the cost parameter ($cost_i$) given in equation 9 has been carefully designed to fit the distribution in the original dataset of generated solutions. The pseudo code of the cost function is given below.

Pseudo-Code of the cost function

Algorithm newVariance(A,n):

Input: The A array storing food source in the total "size" amount and between the parameters whose minimum and maximum values are defined

Output: Variance of the array

```
v <- length[A]
v_ort <- mean[A]
mean <- 0
kvd <- 3.25 % The variance value of the sequence in the original data is assigned to the kvd variable. This value is updated depending on the distribution of the array in every instance.
for i <- 1 to v
mean = (x(i) - v_ort) * (x(i) – v_ort)
end
if ((mean/v)> kvd)
        mean=mean-(mean*(1/mean))
end
if ((mean/v)== kvd)
        mean=mean+0
end
if ((mean/v)< kvd)
        mean=mean+(mean*(1/mean))
end
z= mean /v
return z
```

The algorithm named " newVariance " is used as a cost function. In this function, it is aimed that randomly assigned solutions resemble the distribution in the original array. For this, the variance value in the original array is compared with the variance value in the randomly generated array, and two different situations appear. The large variance value representing the distance from the arithmetic mean indicates that the cost is large. Otherwise, the cost will be small. When cost is large, fitness should be small, and when cost is small, fitness should be large. Therefore, in order to adaptively update the synthetic sequences created to resemble the original data, the outputs produced by the cost function should be given as input to the fitness function given in Equation 9.

$$\text{fitness}_i = \begin{cases} \frac{1}{1+cost_i}, cost_i \geq 0 \\ 1 + |cost_i|, cost_i < 0 \end{cases} \quad (9)$$

The variance value obtained adaptively using the function named " newVariance " is assigned to the $cost_i$ variable in equation 9. After this process, the fitness value of each bee is calculated according to its cost. A high fitness value indicates that the solution is high quality. Then, the obtained fitness value is normalized and the bees' probability of making a choice is determined. Onlooker bees select a source they have identified and search for new positions around this source. If a bee reaches its abandonment limit, a new position is found for this bee and the exploration process begins again. The algorithm stores the best solutions obtained so far in each iteration, and these solutions are continuously updated throughout the 100 iterations determined by the assigned 40 bees. The aim of the study is to find the solution set that provides the lowest cost. In order to better express the process of the proposed methodology, an explanation is provided below as items.

1- New Cost Function Proposal

A new cost function is proposed to understand the population, evaluate the population and produce data distribution similar to the original data set.

2- Probabilistic Region Selection

Probabilistic region selection was made using the roulette wheel method used in the basic ABC algorithm.

3- Source Quality and Its Fitness

The quality of the source is expressed by the suitability parameter. As presented in Equation 7, as the quality of the source increases, the P_i value will increase and the probability of selecting of related region will increase.

4- Selection Process of Onlooker Bees

Following the selection process carried out by the onlooker bees, new solutions will be evaluated and their suitability will be compared.

5- Cycle Process

All these processes will continue until the loop control limit is reached.

6- Algorithm Results

The generated data will be used after the algorithm working process is completed. However, since the produced data is not an integer, rounding will be performed to the nearest integer value between the minimum and maximum parameters.

The main loop of the ABC algorithm invokes the cost function for each individual in the population for a certain number of iterations. So the total time complexity is O (Maxit * nPop * cost). The cost function has O (v) time complexity as it processes v number of elements. In this context, the total time complexity of the algorithm is O (Maxit * nPop * v). (Number of Iterations = 100 and Number of Population = 40). In summary the final time complexity is O(N).

Providing Category Definition of The Produced Synthetic Data

Regression analysis is an approach used in the fields of machine learning and computational statistics. It offers a measure of fitness for performance evaluations of developed models (Chicco et al., 2021). The mathematical relationship between the independent and dependent variables in the data sets using this fitness measure is interpreted.

In this study, the R-squared(R^2) coefficient was used to reveal the relevance between the dependent and independent variables in the original datasets. The mathematical expression of the R^2 coefficient is given in Equation 10.

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - Y_i)^2}{\sum_{i=1}^{m}(Y' - Y_i)^2} \quad (10)$$

The parameters X_i and Y_i given in Equation 10 are the estimated i^{th} value and the actual i^{th} value, respectively. The y' value is the average of the actual values. The m is the total number of data in the dataset (Chicco et al., 2021).

The fact that the r^2 coefficient, which evaluates the degree of interpretation of the independent variables and the performance of the closeness of the estimated values to the real values, is close to 1 indicates that the mathematical relationship is strong (Parhi & Patro, 2023).

The R^2 value used in this study is a statistic that measures how well the model explains the relationships between the data. It is also used to determine the effect of independent variables on the dependent variable. For this reason, it is a statistical modeling preferred in determining the categories.

Regression

Supervised learning is a machine learning task. It maps target input to output via sample input-output pairs. Regression problems are included in supervised learning. In regression problems, the result that presents a continuous output is estimated by the attributes in the sample data (Kinaneva et al., 2021).

In this study, five different regression analyzes were performed using Linear Regression, Polynomial Regression, Support Vector Regression, Decision Tree Regression and Random Forest Regression methods from data with binary or multi-category in the original datasets. As a result of the regression analysis, the R2 coefficient was calculated. Among these coefficients, the coefficient value closest to 1 was accepted as the most successful analysis.

The regression model, which was defined as successful as a result of the analysis, was applied to the synthetic data created by the modified ABC algorithm. With this process, the category of synthetic data was estimated. In the final stage, the synthetic data whose category was defined and the original data were written in the same csv document and given as input to the classification algorithms.

Classification

Classification is the process of interrelating outputs that offer potential solutions as a result of analyzing the data set created in the scope of a specific problem. After training the model with the classification process, analyses are made on the data sets, decisions are made and potential predictions are created (Brnabic & Hess, 2021).

Classification problems are included in the supervised learning framework. In classification problems, the algorithm makes a relation between the attributes of the sample data and maps the input variable to a discrete category (Kinaneva et al., 2021).

In this study, classification was provided between the binary or multi-category data in the original datasets with the k-nearest neighbor, logistic, support vector machines, decision tree and random forest-supervised machine learning classifiers. Then, synthetic data was added to the original data and reclassified on the enriched dataset. Thus, the power of making relationships and producing inferences between the data in the original and enriched data sets was evaluated. The flow diagram of the hybrid structure, in which the modified artificial bee colony optimization algorithm, regression and classification for synthetic data production are used together, is given in Figure 1.

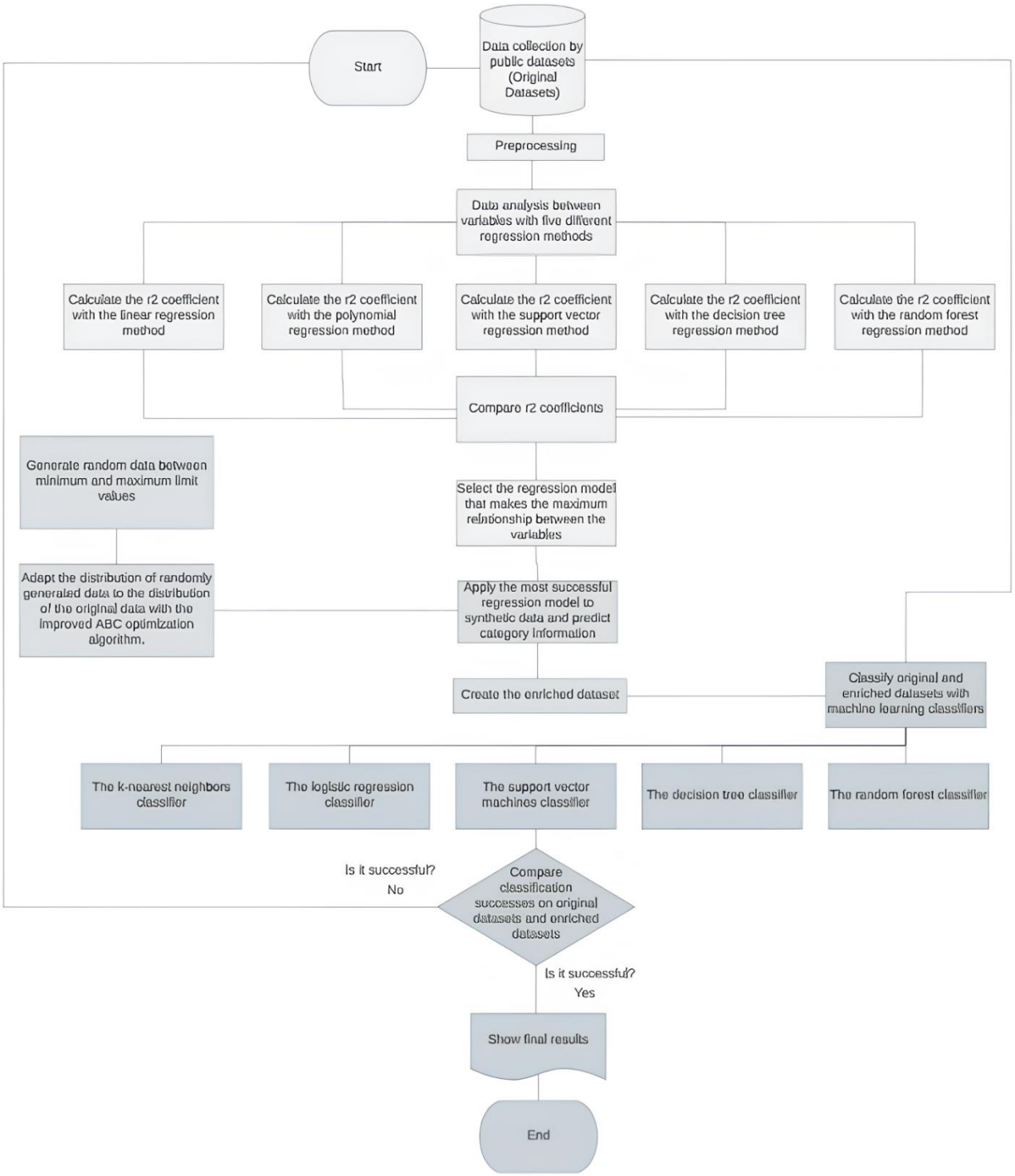


Figure 1. The flow diagram of the proposed hybrid structure

In the first stage of the proposed structure, datasets with little data were supplied from the UCI (UCI, 2024a) public dataset. Then, the data were made suitable for analysis with the preprocessing step. In the second stage, the relationship of the independent variables with the dependent variable in each data set was analyzed with five different regression methods, and the analysis results were evaluated over the r2 coefficients. In the third stage, the regression model with the maximum r2 coefficient was applied to data produced with the improved ABC optimization algorithm. Thus, a category definition was made for synthetic data with a distribution similar to the original data distribution, and an enriched data set was created. In the fourth stage, 5 different machine learning classifiers were used on the enriched dataset with the original dataset. In the fifth stage, the success rates of machine learning classifiers were compared for both original datasets and enriched datasets. As a result of this comparison,

it is expected that the success rate achieved for the enriched dataset was higher than the success rate for the original dataset. Because the amount of data is an important parameter in the training carried out with machine learning algorithms. While discovering patterns among data is easier and possible for a rich dataset, it is difficult and meaningless for a dataset with insufficient data. In addition to this, if the patterns among the generated synthetic data differ from the patterns among the original data, the classification process will fail. For this reason, it is an important criterion to compare the enriched and original datasets to measure the similarity of the synthetic data produced at the end of the classification to the original data. In order to measure the success of the proposed hierarchy, the accuracy rate achieved through the enriched dataset is aimed to be more successful.

RESULTS AND DISCUSSION

Increasing the number of data-based solutions is a crucial goal for the data science world. The synthetic data generation approach is preferred as an alternative preprocessing step in this goal framework. Because synthetic data generation enhances inference power, also it improves educational success depending on the increasing number of data. It creates a solution for cases where confidentiality regulations and ethical limits exist in the original data and strengthens the functionality. However, the distribution of the original data should be preserved among the data arising from real-world problems during the creation of synthetic data. The generated synthetic data should produce patterns similar to the original data. Because data that misrepresents the dataset produces ineffective or incorrect outputs. Therefore, the statistical flow between the data should be analyzed, and reliable synthetic data should be produced by maintaining the distribution of the original data. Then, the efficiency and applicability of the generated data should be evaluated. Human evaluation, statistical difference evaluation, evaluation using a pre-trained machine learning model, training on the synthetic dataset and testing on the real dataset (TSTR), and application-specific evaluation are different evaluation strategies. These strategies should be chosen according to the purpose. At the same time, the use of a combination of these strategies will be useful in evaluating the successful and unsuccessful aspects of the generated synthetic data (Lu et al., 2021).

This study proposes a new synthetic data generation approach using statistical analysis and machine learning approaches. In the first stage of this approach, the distribution of synthetic data produced in the amount of "size" between the minimum and maximum limit values with the ABC optimization algorithm is simulated as the distribution of the original data. Then, the squared(R^2) coefficient is calculated by five different regression analysis methods to find out the relationship between the independent variables with category information of the original data. The R^2 coefficient expresses the relevance between the dependent and independent variables in the original data sets. These relevance values reached in the scope of five different regression methods for each original data set are given in Table 1.

Table 1. Relevance values reaching in the scope of five different regression methods in original datasets.

Squared (R^2)	Linear Regression	Polynomial Regression	Support Vector Regression	Decision Tree Regression	Random Forest Regression
Lenses Dataset	0.598	1.0	0.827	1.0	0.874
COVID Dataset	0.888	1.0	0.808	1.0	0.894
Balloons Dataset 1	0.761	1.0	0.991	1.0	1.0
Balloons Dataset 2	0.761	1.0	0.991	1.0	1.0
Balloons Dataset 3	0.761	1.0	0.991	1.0	0.987
Balloons Dataset 4	0.571	1.0	0.926	1.0	0.796
Caesarian Section Dataset	0.162	0.897	0.373	0.897	0.688
Post-Operative Dataset	0.037	0.807	0.134	0.807	0.665

Table 1 presents the R^2 coefficients reached by the Linear Regression, Polynomial Regression, Support Vector Regression, Decision Tree Regression and Random Forest Regression methods for each dataset. The regression model, which has the maximum R^2 coefficient among these coefficients, indicates that the relationship between the independent variables and the dependent variable is strong. For this reason, the Polynomial Regression model, which has the R^2 coefficient characterizing the maximum relationship for each data set, was applied to the synthetic data created by the modified ABC algorithm. With this implementation, category infomation was determined and an enriched data set was created. In the last step, the classification process was performed on the original and enriched datasets with supervised machine learning classifiers, and the success of the proposed approach was analyzed.

The evaluation criteria used to evaluate the success of the proposed approach have been explained below.

True negative

Examples where the negative class is correctly predicted by the model (Akalın & Yumuşak, 2022).

False positive

Examples where the model predicts a positive sample when it is actually negative (Akalın & Yumuşak, 2022).

False negative

Examples where a sample that is actually positive is predicted as negative by the model (Akalın & Yumuşak, 2022).

True positive

Examples where the positive class is correctly predicted by the model (Akalın & Yumuşak, 2022).

Accuracy rate

It is the ratio of correct predictions for each category to all predictions. Its mathematical expression is given in equation 11 (Akalın & Yumuşak, 2022).

DP+DN/DP+YP+DN+YN (11)

It is aimed that the accuracy rate in the dataset enriched by synthetic data production will be higher than the accuracy rate in the original dataset.

Sensitivity/recall

It is the rate of true positives. Its mathematical expression is given in equation 12 (Akalın & Yumuşak, 2022).

DP/DP+YN (12)

It is aimed that the sensitivity rate in the dataset enriched by synthetic data production will be higher than the sensitivity rate in the original dataset.

Precision

It is the ratio of true positives to total positive predictions. Its mathematical expression is given in equation 13 (Akalın & Yumuşak, 2022).

DP/DP+YP (13)

It is aimed that the Precision ratio in the dataset enriched with synthetic data generation will be higher than the Precision ratio in the original dataset.

F score

It is the harmonic mean of precision and sensitivity values. It enables the evaluation of unbalanced estimates made between categories. Its mathematical expression is given in equation 14 (Akalın & Yumuşak, 2022).

F Score=(2* precision * sensitivity) /(precision + sensitivity) (14)

It is aimed that the F-Score ratio in the dataset enriched by synthetic data generation will be higher than the F-Score ratio in the original dataset. The outputs obtained using k-nearest neighbor, logistic, support vector machines, decision tree and random forest supervised machine learning classifiers for the original dataset and the enriched dataset through these experimental criteria are presented in Table 2-9.

Tablo 2. Lenses Dataset (Original dataset and Enriched dataset)

Original Lenses Dataset	**24 data Acc. Rate**	**Category 1 Prec.**	**Sens.**	**F Sc.**	**Category 2 Prec.**	**Sens.**	**F Sc.**	**Category 3 Prec.**	**Sens.**	**F Sc.**
KNN	0.9	0.5	1	0.666	1	1	1	1	0.857	0.922
SVM	0.9	1	1	1	1	0.5	0.66	0.875	1	0.933
DT	0.9	0.5	1	0.666	1	1	1	1	0.857	0.933
RF	0.7	0.5	1	0.666	0.5	0.5	0.5	0.833	0.714	0.768
Log. Rg.	0.9	1	1	1	1	0.5	0.5	0.875	1	0.933
Enriched Lenses Dataset	**100 data Acc. Rate**	**Category 1 Prec.**	**Sens.**	**F Sc.**	**Category 2 Prec.**	**Sens.**	**F Sc.**	**Category 3 Prec.**	**Sens.**	**F Sc.**
KNN	1	1	1	1	1	1	1	1	1	1
SVM	0.95	1	1	1	1	0.8	0.888	0.925	1	0.961
DT	1	1	1	1	1	1	1	1	1	1
RF	1	1	1	1	1	1	1	1	1	1
Log. Rg.	0.975	1	1	1	0.909	1	0.952	1	0.96	0.979

Tablo 3. COVID-19 Surveillance Dataset (Original dataset and Enriched dataset)

Original COVID Dataset	**13 data Acc. Rate**	**Category 1 Prec.**	**Sens.**	**F Sc.**	**Category 2 Prec.**	**Sens.**	**F Sc.**	**Category 3 Prec.**	**Sens.**	**F Sc.**
KNN	0.5	0.6	1	0.75	0	0	0	0	0	0
SVM	0.5	0.5	1	0.666	0	0	0	0	0	0
DT	0.666	0.75	1	0.857	0.5	0.5	0.5	0	0	0
RF	0.166	0.333	0.333	0.326	0	0	0	0	0	0
Log. Rg.	0.5	0.666	0.666	0.653	0.333	0.5	0.401	0	0	0
Enriched COVID Dataset	**100 data Acc. Rate**	**Category 1 Prec.**	**Sens.**	**F Sc.**	**Category 2 Prec.**	**Sens.**	**F Sc.**	**Category 3 Prec.**	**Sens.**	**F Sc.**
KNN	0.65	0.789	0.714	0.749	0.454	0.454	0.454	0.6	0.75	0.666
SVM	0.8	1	0.761	0.864	0.588	0.909	0.714	0.857	0.75	0.799
DT	0.925	0.952	0.952	0.951	0.9	0.818	0.857	0.888	1	0.936
RF	0.75	0.8	0.761	0.780	0.583	0.636	0.608	0.875	0.875	0.875
Log. Rg.	0.85	0.9	0.857	0.878	0.727	0.727	0.727	0.888	1	0.932

Tablo 4. Balloons Dataset 1 (Original dataset and Enriched dataset)

Original Balloons Dataset 1	**16 data Acc. Rate**	**Category 1 Prec.**	**Sens.**	**F Sc.**	**Category 2 Prec.**	**Sens.**	**F Sc.**
KNN	0.75	0.666	0.666	0.666	0.8	0.8	0.8
SVM	0.5	0.428	1	0.599	1	0.2	0.333
DT	1	1	1	1	1	1	1
RF	0.75	0.6	1	0.75	1	0.6	0.75
Log. Rg.	0.875	0.75	1	0.857	1	0.8	0.888

Continuation of Table 4. Balloons Dataset 1 (Original dataset and Enriched dataset)

Enriched Balloons Dataset 1	100 data Acc. Rate	Category 1 Prec.	Category 1 Sens.	Category 1 F Sc.	Category 2 Prec.	Category 2 Sens.	Category 2 F Sc.
KNN	1	1	1	1	1	1	1
SVM	1	1	1	1	1	1	1
DT	1	1	1	1	1	1	1
RF	1	1	1	1	1	1	1
Log. Rg.	1	1	1	1	1	1	1

Tablo 5. Balloons Dataset 2 (Original dataset and Enriched dataset)

Original Balloons Dataset 2	16 data Acc. Rate	Category 1 Prec.	Category 1 Sens.	Category 1 F Sc.	Category 2 Prec.	Category 2 Sens.	Category 2 F Sc.
KNN	0.875	1	0.75	0.857	0.8	1	0.888
SVM	0.75	0.666	1	0.799	1	0.5	0.666
DT	1	1	1	1	1	1	1
RF	1	1	1	1	1	1	1
Log. Rg.	0.75	0.666	1	0.799	1	0.5	0.666
Enriched Balloons Dataset 2	**100 data Acc. Rate**	**Category 1 Prec.**	**Category 1 Sens.**	**Category 1 F Sc.**	**Category 2 Prec.**	**Category 2 Sens.**	**Category 2 F Sc.**
KNN	1	1	1	1	1	1	1
SVM	1	1	1	1	1	1	1
DT	1	1	1	1	1	1	1
RF	1	1	1	1	1	1	1
Log. Rg.	1	1	1	1	1	1	1

Tablo 6. Balloons Dataset 3 (Original dataset and Enriched dataset)

Original Balloons Dataset 3	16 data Acc. Rate	Category 1 Prec.	Category 1 Sens.	Category 1 F Sc.	Category 2 Prec.	Category 2 Sens.	Category 2 F Sc.
KNN	0.75	0.5	1	0.666	1	0.666	0.799
SVM	0.625	0.4	1	0.571	1	0.5	0.666
DT	0.75	1	1	1	1	1	1
RF	0.875	0.666	1	0.799	1	0.833	0.908
Log. Rg.	1	1	1	1	1	1	1
Enriched Balloons Dataset 3	**100 data Acc. Rate**	**Category 1 Prec.**	**Category 1 Sens.**	**Category 1 F Sc.**	**Category 2 Prec.**	**Category 2 Sens.**	**Category 2 F Sc.**
KNN	1	1	1	1	1	1	1
SVM	1	1	1	1	1	1	1
DT	1	1	1	1	1	1	1
RF	1	1	1	1	1	1	1
Log. Rg.	1	1	1	1	1	1	1

Tablo 7. Balloons Dataset 4 (Original dataset and Enriched dataset)

Original Balloons Dataset 4	16 data Acc. Rate	Category 1 Prec.	Category 1 Sens.	Category 1 F Sc.	Category 2 Prec.	Category 2 Sens.	Category 2 F Sc.
KNN	0.428	0	0	0	0.428	1	0.599
SVM	0.428	0	0	0	0.428	1	0.599
DT	0.857	1	0.75	0.857	0.428	1	0.599
RF	0.571	0.666	0.5	0.571	0.5	0.666	0.571
Log. Rg.	0.428	0	0	0	0.428	1	0.599
Enriched Balloons Dataset 4	**100 data Acc. Rate**	**Category 1 Prec.**	**Category 1 Sens.**	**Category 1 F Sc.**	**Category 2 Prec.**	**Category 2 Sens.**	**Category 2 F Sc.**
KNN	1	1	1	1	1	1	1
SVM	0.85	0.736	0.933	0.823	0.952	0.8	0.869
DT	1	1	1	1	1	1	1
RF	1	1	1	1	1	1	1
Log. Rg.	0.9	1	0.733	0.846	0.862	1	0.925

Tablo 8. Caesarian Section Dataset (Original dataset and Enriched dataset)

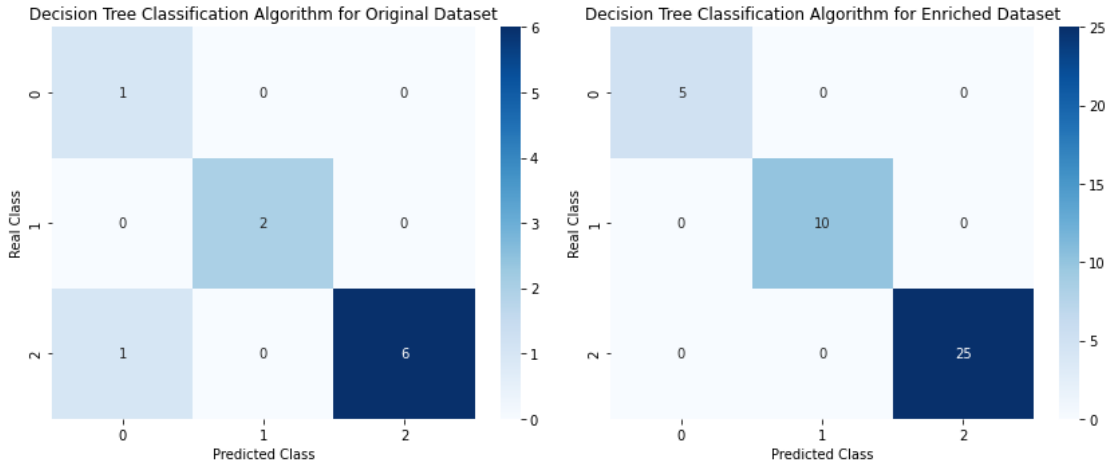
Original Caesarian Section Dataset	**80 data Acc. Rate**	**Prec.**	**Category 1 Sens.**	**F Sc.**	**Prec.**	**Category 2 Sens.**	**F Sc.**
KNN	0.718	0.769	0.625	0.689	0.684	0.812	0.745
SVM	0.562	0.583	0.437	0.499	0.55	0.687	0.611
DT	0.593	0.588	0.625	0.606	0.6	0.562	0.580
RF	0.687	0.687	0.687	0.687	0.687	0.687	0.687
Log. Rg.	0.593	0.615	0.5	0.551	0.578	0.687	0.628
Enriched Caesarian Section Dataset	**250 data Acc. Rate**	**Prec.**	**Category 1 Sens.**	**F Sc.**	**Prec.**	**Category 2 Sens.**	**F Sc.**
KNN	0.76	0.745	0.803	0.772	0.777	0.714	0.744
SVM	0.65	0.605	0.901	0.724	0.791	0.387	0.519
DT	0.85	0.891	0.803	0.845	0.814	0.897	0.854
RF	0.81	0.82	0.803	0.811	0.8	0.816	0.808
Log. Rg.	0.77	0.818	0.764	0.790	0.732	0.836	0.780

Tablo 9. Post-Operative Dataset (Original dataset and Enriched dataset)

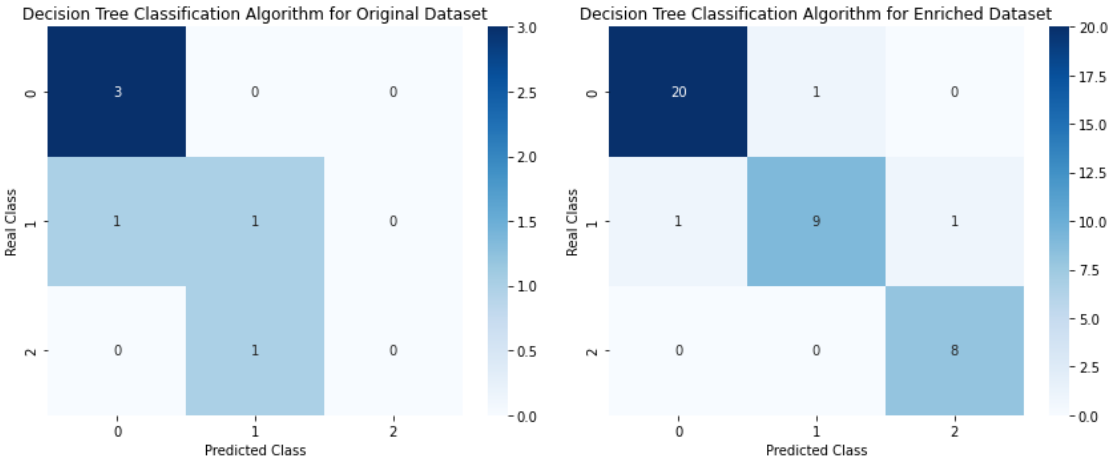
Original Post-Operative Dataset	**86 data Acc. Rate**	**Prec.**	**Category 1 Sens.**	**F Sc.**	**Prec.**	**Category 2 Sens.**	**F Sc.**
KNN	0.571	0.689	0.769	0.727	0	0	0
SVM	0.714	0.735	0.961	0.833	0	0	0
DT	0.571	0.739	0.653	0.693	0.25	0.333	0.285
RF	0.657	0.718	0.884	0.793	0	0	0
Log. Rg.	0.685	0.727	0.923	0.813	0	0	0
Enriched Post-Operative Dataset	**250 data Acc. Rate**	**Prec.**	**Category 1 Sens.**	**F Sc.**	**Prec.**	**Category 2 Sens.**	**F Sc.**
KNN	0.64	0.676	0.745	0.708	0.571	0.487	0.526
SVM	0.76	0.753	0.881	0.812	0.774	0.585	0.666
DT	0.66	0.698	0.745	0.721	0.594	0.536	0.563
RF	0.71	0.714	0.847	0.775	0.7	0.512	0.591
Log. Rg.	0.82	0.788	0.949	0.861	0.896	0.634	0.742

When Table 2-9 is examined, it is seen that the performance criteria reached in the enriched datasets are more successful than the performance criteria achieved in the original datasets. More successful training was carried out with the enriched dataset with patterns similar to the original dataset. Because the increase in the number of data provided more relationships to be learned and more inferences to be made. Experimental results show that patterns between data are detected more successfully with the Decision Tree machine learning classifier among all classifiers. The synthetic data generation approach, which positively affects educational power's success, clearly explains the importance of producing synthetic data. In addition, these criteria summarize the success of the proposed approach. In this study, synthetic data with a similar spread to the original data were produced by the ABC optimization algorithm. Then, the statistical relationship between dependent and independent variables was analyzed. This statistical relationship should have close statistical inferences for the analysis performed on the enriched dataset and the analysis performed on the original dataset. With these inferences, patterns and relationships between data will be discovered with machine learning classifiers, and performance will be evaluated. As a result of classification, the evaluation performance of the enriched datasets should be more successful than the original datasets. Because the sufficient number of data reveals the power of machine learning algorithms. For this reason, the increase in the success rate with the increasing number of data in the presented study indicates that proportional relations are established, and suitable approaches are preferred.

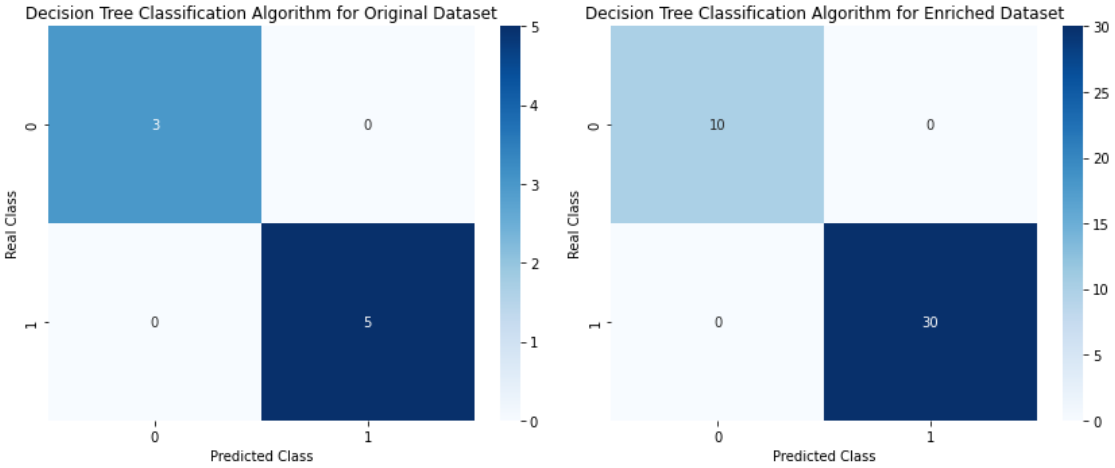
For an deep analysis of the decision tree classification algorithm that showed maximum success within the framework of the performance criteria given in Table 2-9, the confusion matrices obtained in the original and enriched datasets for each dataset are given in Figure 2-10.



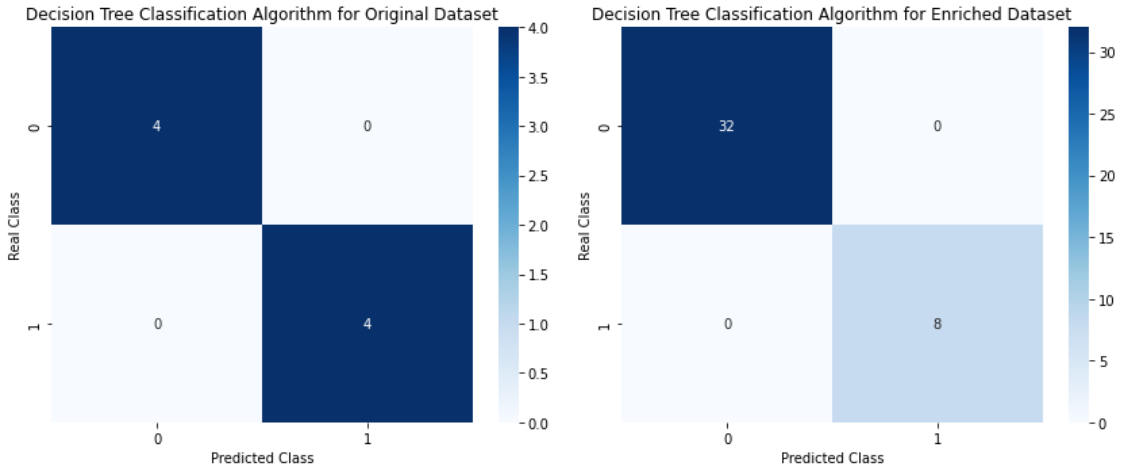
Şekil 3. Confusion matrices obtained for both the original and enriched datasets in the Lenses dataset



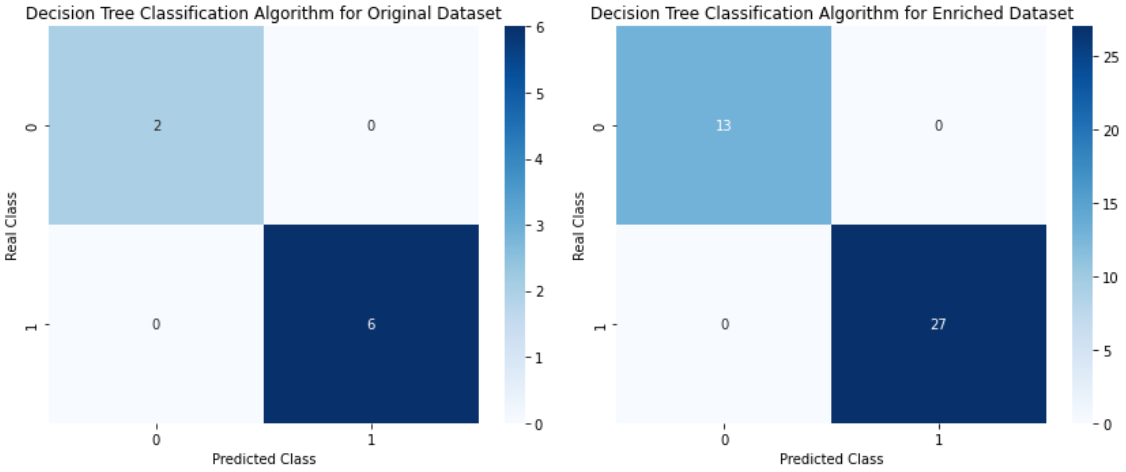
Şekil 4. Confusion matrices obtained for both the original and enriched datasets in the COVID-19 Surveillance dataset



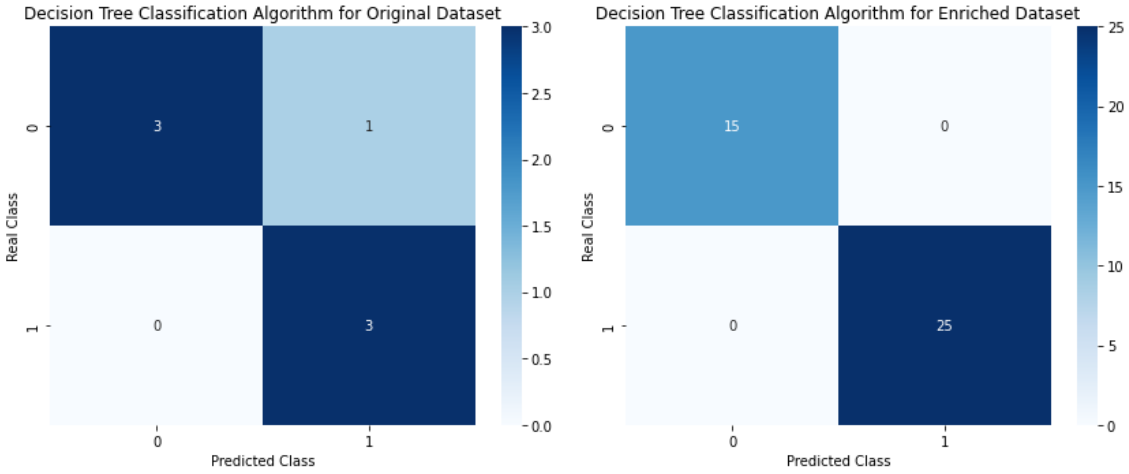
Şekil 5. Confusion matrices obtained for both the original and enriched datasets in the Balloons dataset 1



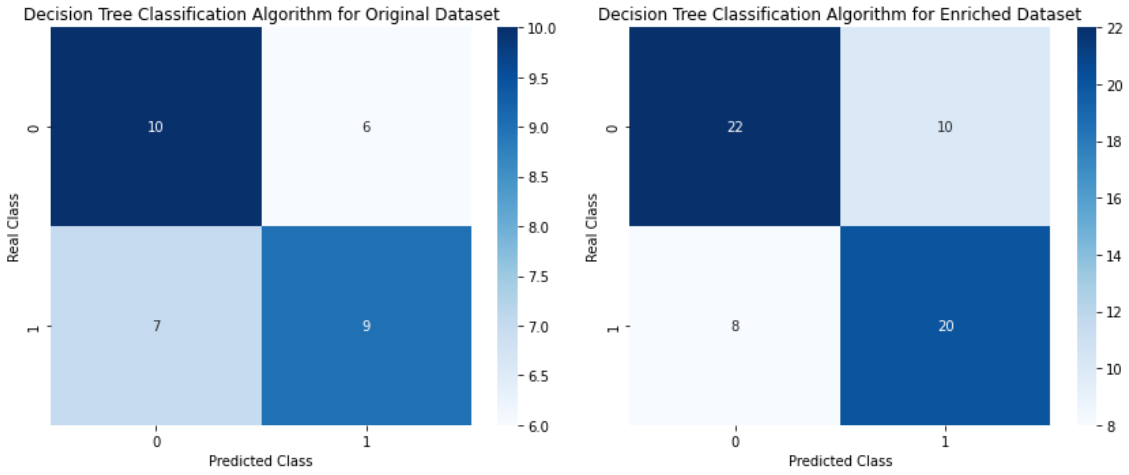
Şekil 6. Confusion matrices obtained for both the original and enriched datasets in the Balloons dataset 2



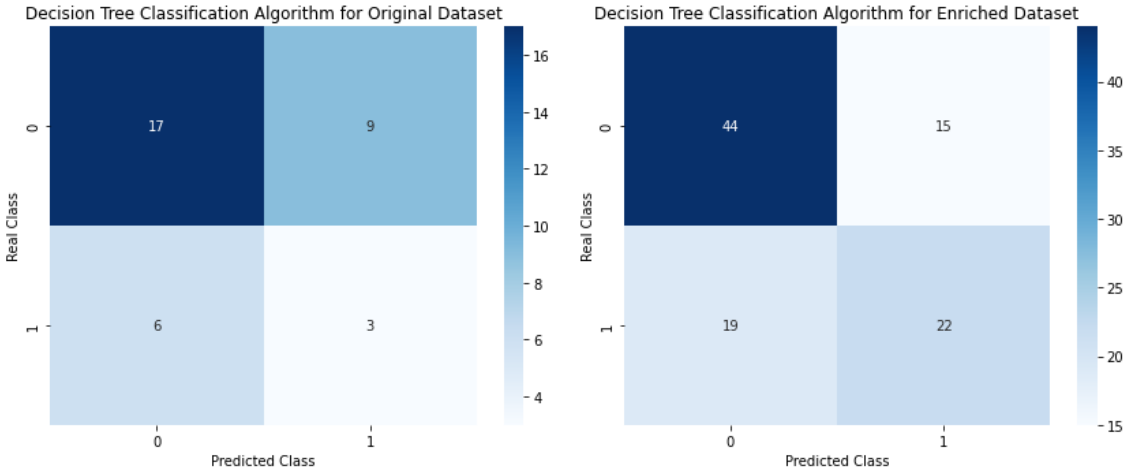
Şekil 7. Confusion matrices obtained for both the original and enriched datasets in the Balloons dataset 3



Şekil 8. Confusion matrices obtained for both the original and enriched datasets in the Balloons dataset 4



Şekil 9. Confusion matrices obtained for both the original and enriched datasets in the Caesarian Section dataset



Şekil 10. Confusion matrices obtained for both the original and enriched datasets in the Post-Operative dataset

When the confusion matrices given in Figure 2-10 are examined, it is seen that the decision tree algorithm that produces maximum performance identifies true positives and true negatives more strongly on the enriched datasets compared to the original dataset. In this context,

-In the enriched Lenses dataset, it is seen that each category is detected correctly compared to the original dataset.

-While there was no correct prediction for category 3 in the original COVID-19 Surveillance dataset, all instances for category 3 were correctly detected in the enriched dataset.

-The Decision Tree classification algorithm produced 100% accurate predictions on all original and enriched datasets within the scope of Balloons Dataset 1, Balloons Dataset 2 and Balloons Dataset 3. On the other hand, while not every category was detected correctly in the original Balloons Dataset 4 dataset, every category was detected correctly in the enriched Balloons Dataset 4 dataset.

-While almost half of the samples classified with the decision tree algorithm in the original Caesarian Section dataset were predicted incorrectly, approximately two-thirds of the predictions for both categories were detected correctly in the enriched Caesarian Section dataset.

- While almost half of the samples classified with the decision tree algorithm in the original Post-Operative Patients dataset were predicted incorrectly, approximately two-thirds of the predictions in the enriched Post-Operative Patients dataset were detected correctly.

This proposed methodology provides a functional process for improving performance and detection accuracy for each dataset. In addition, it is expected that detection performance will increase if a model that is more suitable for the nature of the dataset and more powerful than the decision tree classification algorithm is used. This situation is clearly depicted in Table 9.

Besides the increasing success rate, the proposed approach has limitations that need to be discussed. The first limitation of the proposed approach is that it is optimized with the ABC optimization algorithm after random values are generated. The heuristic ABC algorithm used in the optimization process has a convergence property but does not guarantee the exact solution. For this reason, the algorithm was re-run in case data exceeded the minimum and maximum limits during the production of data similar to the original data distribution. The second limitation of the proposed approach is scaling the data for category detection in synthetic data after the regression model that models the relationship with maximum success is chosen. This will make the being meaningful of inferences produced for category detection. The third limitation of the proposed approach is that the regression result for category definition is not an integer. These results should be rounded to the nearest integer. The fourth limitation of the proposed approach is that an adequate inference cannot be made with insufficient or incomplete data on the target category. Because the proposed method offers a strong performance as a result of successfully learning the relationships and patterns in the original data.

The synthetic data generation approach proposed through this study has presented a novelty to making inferences about the past or producing smart systems about the future. In addition, a solution point has been produced for time, cost and legal problems, and functionality has been improved.

CONCLUSION

The realization of human knowledge, experience and intelligence by machines is the future goal of many sectors. This innovation plan aims to make inferences from existing data and is closely related to data science. Because data science provides to produce successful decisions with the suitability of the selected model to the problem and a sufficient number of data. However, some sectors do not have the working area to produce sufficient data. On the other hand, legal problems for fields that can generate huge datasets may be, and these legal problems prevent the process. Therefore, generating synthetic data instead of identifiable data in the original dataset will tolerate these problems.

In this study, a novel synthetic data generation approach is proposed. The basic steps of this proposed approach are given below.

1- Analyze the distribution of the original data with a small sample space

2- Generate random data between minimum and maximum limit values

3- Create the distribution of randomly generated data similar to the distribution of the original data by means of the modified artificial bee colony optimization algorithm.

4- Calculate the squared(R2) coefficient by means of five different regression analysis methods to find out the relationship between the independent variables in the original data and the category information

5- Choose the regression method that has the maximum squared(R2) coefficient among the coefficients calculated by the Linear Regression, Polynomial Regression, Support Vector Regression, Decision Tree Regression and Random Forest Regression methods

6- Apply the regression model with the maximum R2 coefficient to the synthetic data generated by the improved ABC algorithm, and then write the category of the predicted synthetic data.

7- Classify the original dataset and the enriched dataset containing the original data with the k-nearest neighbour, logistic, support vector machines, decision tree and random forest supervised machine learning classifiers, respectively.

8- Compare the success rates achieved for both datasets and evaluate the success of the proposed approach.

The proposed synthetic data generation approach summarized in eight items produced high-quality, accurate, sufficient datasets with a balanced distribution between categories. Thus, stronger machine-learning analyzes were made using enriched datasets. It is thought that the hierarchy created by this study is a solution point in real-world problems. It is planned to be used as a pre-processing step in the innovation processes of different fields for this approach improving functionality. Moreover with the novel synthetic data generation approach proposed in this study, a new solution has been developed for time, cost, legal problems and data scarcity situations. Functionality has been improved and powerful machine-learning inferences have been made,

In the future, it is planned to integrate the transformer approach, which has a self-attention mechanism, into the proposed methodology to discover stronger relationships between data and make more accurate category determination. Because this mechanism learns long-range dependencies between data points, it also allows each data point to evaluate its relationships with all other data points.

ACKNOWLEDGEMENTS

There are no supported projects.

Conflict of Interest

The authors declare that there is no financial support or relationship that could pose a conflict of interest.

Author's Contributions

The article is single-authored.

REFERENCES

Akalın, F., & Yumuşak, N. (2022). DNA genom dizilimi üzerinde dijital sinyal işleme teknikleri kullanılarak elde edilen ekson ve intron bölgelerinin EfficientNetB7 mimarisi ile sınıflandırılması. Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi, 37(3), 1355–1371. https://doi.org/10.17341/gazimmfd.900987.

Akay, B., Karaboga, D., Gorkemli, B., & Kaya, E. (2021). A survey on the artificial bee colony algorithm variants for binary, integer, and mixed integer programming problems. Applied Soft Computing, 106, 1–35. https://doi.org/10.1016/j.asoc.2021.107351.

Alvarado-Iniesta, A., Garcia-Alcaraz, J. L., Rodriguez-Borbon, M. I., & Maldonado, A. (2013). Optimization of the material flow in a manufacturing plant by use of artificial bee colony algorithm. Expert Systems with Applications, 40, 4785–4790. https://doi.org/10.1016/j.eswa.2013.02.029.

Arab, N., Nemmour, H., & Chibani, Y. (2023). A new synthetic feature generation scheme based on artificial immune systems for robust offline signature verification. Expert Systems with Applications, 213. https://doi.org/10.1016/j.eswa.2022.119306.

Brnabic, A., & Hess, L. M. (2021). Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. BMC Medical Informatics and Decision Making, 21. https://doi.org/10.1186/s12911-021-01403-2.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7, 1–24. https://doi.org/10.7717/PEERJ-CS.623.

Dahmen, J., & Cook, D. (2019). SynSys: A synthetic data generation system for healthcare applications. Sensors, 19(5), 1–11. https://doi.org/10.3390/s19051181.

Dankar, F. K., & Ibrahim, M. (2021). Fake it till you make it: Guidelines for effective synthetic data generation. Applied Sciences, 11, 1–18. https://doi.org/10.3390/app11052158.

Douzas, G., Lechleitner, M., & Bacao, F. (2022). Improving the quality of predictive models in small data GSDOT: A new algorithm for generating synthetic data. PLoS One, 17(4), 1–15. https://doi.org/10.1371/journal.pone.0265626.

El Mrabet, M. A., El Makkaoui, K., & Faize, A. (2021). Supervised machine learning: A survey. In Proceedings of the 4th International Conference on Advanced Communication Technologies and Networking (CommNet 2021). https://doi.org/10.1109/CommNet52204.2021.9641998.

Hashimoto, D. A., Ward, T. M., & Meireles, O. R. (2020). The role of artificial intelligence in surgery. Advances in Surgery, 54, 89–101. https://doi.org/10.1016/j.yasu.2020.05.010.

Karaboğa, D. (2020). Yapay Zeka Optimizasyon Algoritmaları,Nobel Akademik Yayıncılık, 7. Baskı.

Kaya, E., Gorkemli, B., Akay, B., & Karaboga, D. (2022). A review on the studies employing artificial bee colony algorithm to solve combinatorial optimization problems. Engineering Applications of Artificial Intelligence, 115. https://doi.org/10.1016/j.engappai.2022.105311.

Kinaneva, D., Hristov, G., Kyuchukov, P., Georgiev, G., Zahariev, P., & Daskalov, R. (2021). Machine learning algorithms for regression analysis and predictions of numerical data. In 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA).

Li, M., Zhuang, D., & Chang, J. M. (2023). MC-GEN: Multi-level clustering for private synthetic data generation. Knowledge-Based Systems, 264, 1–11. https://doi.org/10.1016/j.knosys.2022.110239.

Li, Z., Zhao, Y., & Fu, J. (2020). SynC: A copula-based framework for generating synthetic data from aggregated sources. In 2020 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 571–578). https://doi.org/10.1109/ICDMW51313.2020.00082.

Lu, Y., Shen, M., Wang, H., & Wei, W. (2021). Machine learning for synthetic data generation: A review. Journal of Big Data, 14(8), 1–18.

Parhi, S. K., & Patro, S. K. (2023). Prediction of compressive strength of geopolymer concrete using a hybrid ensemble of grey wolf optimized machine learning estimators. Journal of Building Engineering, 71. https://doi.org/10.1016/j.jobe.2023.106521.

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The synthetic data vault. In 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (pp. 399–410). https://doi.org/10.1109/DSAA.2016.49.

Ping, H., Stoyanovich, J., & Howe, B. (2017). DataSynthesizer: Privacy-preserving synthetic datasets. In Proceedings of the 29th International Conference on Scientific and Statistical Database Management (SSDBM ’17) (pp. 1–5). https://doi.org/10.1145/3085504.3091117.

UCI. (2024a). (the University of California Irvine Machine Learning Repository). https://archive.ics.uci.edu/.

UCI. (2024b). (the University of California Irvine Machine Learning Repository)- Lenses. https://archive.ics.uci.edu/dataset/58/lenses.

UCI. (2024c). (the University of California Irvine Machine Learning Repository)- COVID-19 Surveillance. https://archive.ics.uci.edu/dataset/567/covid+19+surveillance.

UCI. (2024d). (the University of California Irvine Machine Learning Repository)- Balloons. https://archive.ics.uci.edu/dataset/13/balloons.

UCI. (2024e). (the University of California Irvine Machine Learning Repository)- Caesarian Section. https://archive.ics.uci.edu/dataset/472/caesarian+section+classification+dataset.

UCI. (2024f). (the University of California Irvine Machine Learning Repository)- Post-Operative Patient. https://archive.ics.uci.edu/dataset/82/post+operative+patient.

Wharrie, S., et al. (2022). HAPNEST: An efficient tool for generating large-scale genetics datasets from limited training data. In NeurIPS 2022 Workshop on Synthetic Data Empowering Machine Learning Research (pp. 1–7).