Gazi Üniversitesi
Fen Bilimleri Dergisi
PART C: TASARIM VE TEKNOLOJİ

Gazi University
Journal of Science
PART C: DESIGN AND TECHNOLOGY

GU J Sci, Part C, 12(3): 746-757 (2024)

Feature Selection in the Diabetes Dataset with the Marine Predator Algorithm and Classification using Machine Learning Methods

Fuat TÜRK[1*], Nuri Alper METİN [2], Murat LÜY [3]

[1]*Kırıkkale University, Faculty of Engineering and Natural Sciences, Computer Engineering, Kırıkkale, Turkey*

[2]*Kırıkkale University, Kırıkkale Vocational School, Electronic Communication Program, Kırıkkale, Turkey*

[3]*Kırıkkale University, Faculty of Engineering and Natural Sciences, Electrical and Electronics Engineering, Kırıkkale, Turkey*

Article Info

Research article
Received: 25/11/2023
Revision: 25/02/2024
Accepted: 11/03/2024

Keywords

Machine learning
Marine predator optimization algorithm
Classification diabetes
Feature selection

Makale Bilgisi

Araştırma makalesi
Başvuru: 25/11/2023
Düzeltme: 25/02/2024
Kabul: 11/03/2024

Anahtar Kelimeler

Makine öğrenmesi
Deniz yırtıcısı algoritması
Şeker hastalığı sınıflandırması,
Özellik seçimi

Graphical/Tabular Abstract (Grafik Özet)

The diabetic dataset from Kaggle is preprocessed and then submitted to the feature extraction module, where key features are identified and categorized after optimization. Results are compared to those without optimization based on metrics like F1 score, Recall, and accuracy. / Kaggle'dan elde edilen diyabet veri seti, ön işleme sonrası özellik çıkarım modülüne sunulur; burada optimizasyondan sonra anahtar özellikler belirlenir ve kategorize edilir. Sonuçlar, F1 skoru, Recall ve doğruluk gibi metrikler açısından optimizasyon uygulanmadan elde edilenlerle karşılaştırılır.

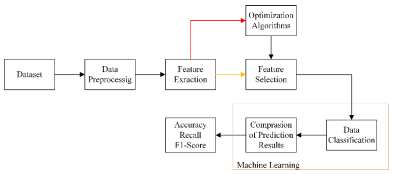

Figure A: *System block diagram /* ***Şekil A:*** *Sistem blok diyagramı*

Highlights (Önemli noktalar)

- *The study uses Kaggle's diabetes dataset for feature selection block analysis.*
- *The feature selection block identifies prominent features.*
- *The selected features are categorized using the categorization module.*
- *Performance metrics are compared with the dataset without the marine predator optimization algorithm (MPOA).*
- *The LR classification approach achieves an accuracy of 77.63% without feature selection.*
- *When MPOA is used for feature selection, accuracy increases to 79.39%.*

Aim (Amaç): *The study aims to improve diabetes classification accuracy by using the Marine Predator Optimization Algorithm (MPOA) for feature selection on a Kaggle dataset, enhancing performance metrics like accuracy, F1 score, Recall, and Precision. / Çalışma, Kaggle veri setinde özellik seçimi için Deniz Yırtıcısı Optimizasyon Algortimasını kullanarak diyabet sınıflandırma doğruluğunu artırmayı ve doğruluk, F1 skoru, Recall ve Precision gibi performans metriklerini iyileştirmeyi amaçlamaktadır.*

Originality (Özgünlük): *Diabetes, a complex metabolic disorder, showed an increase in accuracy from 77.63% to 79.39% in a study using the Marine Predator Optimization Algorithm (MPOA) for feature selection on a dataset obtained from Kaggle. / Diyabet, karmaşık bir metabolik bozukluk olup, yapılan bir çalışmada Kaggle'dan elde edilen veri seti ile Deniz Yırtıcısı Optimizasyon Algoritması kullanılarak yapılan özellik seçiminde doğruluk oranı %77,63'ten %79,39'a yükselmiştir.*

Results (Bulgular): *The study results shows that the Marine Predator Optimization Algorithm (MPOA) boosts Logistic Regression accuracy from 77.63% to 79.39% and improves overall performance metrics. / Çalışma, Deniz Yırtıcısı Optimizasyon Algortimasının Logistic Regression doğruluğunu %77,63'ten %79,39'a artırdığını ve genel performans metriklerini iyileştirdiğini göstermektedir.*

Conclusion (Sonuç): *This study compares diabetes classification performance metrics using machine learning algorithms with and without feature selection, showing that the Marine Predator Algorithm increases accuracy from 77.63% to 79.39% and suggests future research into alternative optimization strategies. / Bu çalışma, makine öğrenimi algoritmalarıyla özellik seçimli ve özelliksiz diyabet sınıflandırma performans metriklerini karşılaştırarak, Marine Predator Algoritması'nın doğruluğu %77,63'ten %79,39'a artırdığını ve alternatif optimizasyon stratejileri üzerine gelecekteki araştırmalar için potansiyel sunduğunu göstermektedir.*

*Corresponding author, e-mail: fturk@kku.edu.tr
DOI: 10.29109/gujsc.1396051

GU J Sci, Part C, 12(3): 746-757 (2024)

Gazi Üniversitesi
Fen Bilimleri Dergisi
PART C: TASARIM VE TEKNOLOJİ

Gazi University
Journal of Science
PART C: DESIGN AND TECHNOLOGY

http://dergipark.gov.tr/gujsc

Feature Selection in the Diabetes Dataset with the Marine Predator Algorithm and Classification using Machine Learning Methods

Fuat TÜRK [1*], Nuri Alper METİN [2], Murat LÜY [3]

[1]*Kırıkkale University, Faculty of Engineering and Natural Sciences, Computer Engineering, Kırıkkale, Turkey*

[2]*Kırıkkale University, Kırıkkale Vocational School, Electronic Communication Program, Kırıkkale, Turkey*

[3]*Kırıkkale University, Faculty of Engineering and Natural Sciences, Electrical and Electronics Engineering, Kırıkkale, Turkey*

Article Info

Research article
Received: 25/11/2023
Revision: 25/02/2024
Accepted: 11/03/2024

Keywords

Machine learning
Marine predator optimization algorithm
Classification diabetes
Feature selection

Abstract

Diabetes is now classified as one of the leading causes of death. Diabetes is a chronic and complex metabolic disorder characterized by carbohydrate, fat, and protein metabolism disturbances. Type 1 diabetes is categorized along with other different types of diabetes, including Type 2 diabetes as well as gestational diabetes. Both acute and chronic complications occur in individuals with diabetes due to decreased insulin secretion and disruptions in carbohydrate, fat, and protein metabolism. In this study, after completing the data preparation step, the diabetes dataset from Kaggle is sent to the feature selection block for analysis. Once the optimization process is complete, the feature selection block will determine the most prominent features. The selected features discussed earlier are categorized using the categorization module. The findings are compared to the performance metrics results with the dataset without the marine predator optimization algorithm (MPOA) technique applied, especially regarding metrics such as F1 score, Recall, Accuracy, and Precision. The results show that the LR classification approach achieves an accuracy of 77.63% without feature selection. On the other hand, when MPOA is used for feature selection, the accuracy increases to 79.39%.

Diyabet Veri Setinde Deniz Yırtıcısı Algoritması ile Özellik Seçimi ve Makine Öğrenimi Yöntemleri Kullanılarak Sınıflandırma

Makale Bilgisi

Araştırma makalesi
Başvuru: 25/11/2023
Düzeltme: 25/02/2024
Kabul: 11/03/2024

Anahtar Kelimeler

Makine öğrenmesi
Deniz yırtıcısı algoritması
Şeker hastalığı sınıflandırması,
Özellik seçimi

Öz

Diyabet günümüzde önde gelen ölüm nedenlerinden biri olarak sınıflandırılır. Diyabet hastalığı karbonhidrat, yağ ve protein metabolizmasındaki bozulmalarla tanımlanan kronik ve karmaşık bir metabolik bozukluktur. Tip 1 diyabet, Tip 2 diyabetin yanı sıra gestasyonel diyabet de dahil olmak üzere diğer farklı diyabet türleriyle birlikte kategorize edilir. Diyabetli bireylerde azalan insülin salgısı ve karbonhidrat, yağ ve protein metabolizmasındaki aksaklıklar nedeniyle hem akut hem de kronik komplikasyonlar ortaya çıkmaktadır. Bu çalışmada, Veri hazırlama adımının tamamlanmasının ardından, Kaggle'dan alınan diyabet veri seti analiz için özellik çıkarma bloğuna gönderilir. Optimizasyon süreci tamamlandıktan sonra, özellik seçimi bloğu hangi özelliklerin en çok öne çıktığını belirleyecektir. Daha önce tartışılan seçilen özellikler, kategorizasyon modülü kullanılarak çeşitli kategorilere ayrılır. Bulgular, özellikle F1 puanı, Geri Çağırma, Doğruluk ve Kesinlik gibi ölçütler açısından, deniz yırtıcısı optimizasyon algoritması (MPOA) tekniği uygulanmamış veri setiyle performans metrikleri sonuçlarıyla karşılaştırılır. Bulgular, LR sınıflandırma yaklaşımının özellik seçimi olmadan %77,63'lük bir doğruluk oranına ulaştığını göstermektedir. Öte yandan özellik seçimi için MPOA kullanıldığında, doğruluk oranı %79,39'a yükselmektedir.

1. INTRODUCTION (GİRİŞ)

Diabetes is one of the major health problems seen worldwide and increasing daily. The financial difficulty caused by the increasing disease during the healing process is becoming an undeniable reality. The main reason for the increase in the disease is factors such as changes in the social structure, heavy work standards, rising obesity, and unhealthy lifestyles. These reasons have caused diabetes to become common in the 21st century.

*Corresponding author, e-mail: fturk@kku.edu.tr
DOI: 10.29109/gujsc.1396051

Diabetes is divided into two: type-1 and type-2. Type-1 diabetes is a chronic disease that causes insulin deficiency due to damage to the beta cells in the pancreas. Studies have proven a high rate of beta cell damage in type 1 diabetes patients. It has been determined that type-1 diabetic patients will likely suffer from chronic diseases such as celiac, hepatitis, and vitiligo. Type 1 diabetes has become one of the most common diseases in childhood. It is formed by the destruction of beta cells produced in the pancreas. According to many reports worldwide, it has been revealed that the genetic scheme of type 1 diabetes is genetically transmitted. Some research reports have revealed that diabetes can occur even if there is genetic transmission in the family. Notably, most diabetic diseases that are not genetically transmitted are in patients at a young age. Type 2 diabetes has become one of the most important chronic diseases increasing in our country and worldwide. Improving the quality of life of type 2 diabetes patients and regular treatment is crucial. According to data received in 2021, 537 million adults worldwide are living with diabetes, predicted to reach 643 million in 2045. Type 2 diabetes occurs due to a sedentary lifestyle and irregular nutrition, allowing the disease to progress continuously. Type 2 diabetes is a metabolic disease in which insulin resistance occurs due to the disorder caused by beta cells in insulin secretion, which is related to high glucose levels. This study performs feature selection with the Marine Predator Optimization Algorithm (MPOA) diabetes dataset obtained from Kaggle. This received feature selection set is classified with Logistic Regression (LR), Random Forest (RF), k-nearest neighbors (k-NN), Gradient Boosting (GB), XgBoost, Support Vector Machine (SVM), and Decision Tress (DT) classification methods. The data set with feature selection was applied, and the data without feature selection is used. The data set classification results are compared regarding performance metrics [1–4].

Machine learning is a field focused on creating algorithms that can recognize patterns and forecast future occurrences through the analysis of extensive data sets. Classification is a crucial subfield in machine learning that entails assigning samples from a collection to preset classes. Feature selection is a crucial stage in classification issues to construct a precise model and get optimal performance. Feature selection involves determining the most useful characteristics in the dataset and removing redundant or low-impact features. This simplifies the model, decreases training time, and minimizes the danger of overfitting. Machine learning models frequently deal with several characteristics. Therefore, feature selection methods can help the model choose the most suitable features. These algorithms may utilize statistical approaches, information acquisition, and feature significance ratings. Effective feature selection can enhance the model's performance and promote the creation of more generalizable models by preventing overfitting. Effective feature selection is crucial in this context for classification issues to enhance model accuracy and get more dependable and generalizable outcomes.

The main diabetes dataset in this study is obtained from the aggregated data set. The secondary data is categorized without feature selection. Performance metrics data are collected following the categorization procedure. The sea predator optimization technique, a unique optimization approach, is utilized for feature selection on the diabetes dataset. Performance metrics are produced as a consequence of this approach. These performance metrics are compared with each other and with the results from the literature review in the last stage. Utilizing MPOA for feature selection results in a significant enhancement in performance measures. Traditional optimization algorithms for feature selection yield inferior performance metrics compared to new optimization algorithms.

This research comprises five chapters, each serving a distinct purpose. The first chapter provides an introduction, the second chapter presents a comprehensive literature review, the third chapter outlines the materials and methods employed, the fourth chapter presents the results and discussion, and the fifth chapter concludes the study.

2. LITERATURE REVIEW (LİTERATÜR TARAMASI)

Sisodia et al., diabetes is one of the worst blood sugar-raising diseases. Diabetes left untreated has numerous implications. After the problematic identification process, patients meet physicians at diagnostic facilities. Increasing machine learning solves this problem. This study seeks to develop a diabetes prediction model. This experiment diagnoses diabetes early using the Decision Tree, SVM, and Naive Bayes. Experiments employ the UCI machine learning repository Pima Indians Diabetes Database (PIDD). All three methods are evaluated using Precision, Accuracy, F-Measure, and Recall. Accuracy depends on right and wrong categorization. With 76.30% accuracy, Naive Bayes is the most accurate algorithm. Methodically using ROC curves confirms these conclusions [5].

Kaur et al., Effective diabetes prediction utilizing patient medical data is the subject of this effort. Diabetes currently affects all ages and communities. Diabetes increases cardiac, kidney, nerve, blood vessel, and blindness risk. Mining diabetes data effectively is vital. It utilizes the Pima Indians Diabetes Data Set, which comprises diabetics and non-diabetics. The enhanced J48 classifier boosts data mining accuracy. WEKA was used as MATLAB API to create J-48 classifiers. Trials showed a significant improvement over the J-48 algorithm [6].

Febrian et al., diabetes may cause blindness, kidney failure, heart attacks, and death. For 2019, the International Diabetes Federation estimated 463 million diabetics. If projections are correct, 578 million will be added by 2030 and 700 million by 2045. In 2020, the Ministry of Health named Indonesia one of the 10 countries with the highest diabetes incidence in 2019. Diagnosing diabetes takes skill. Many people examined have significant conditions due to delayed diagnosis. Severe diabetes prevention needs detection technology. It helps doctors identify ailments quickly and accurately nowadays. By constructing an AI model to identify diabetes, we can apply machine learning to prevent death. We may compare k-NN and Naive Bayes to see which predicts diabetes better. Finally, the study examined two k-Nearest Neighbor algorithms and Naive Bayes to predict diabetes using numerous health indicators in the dataset using supervised machine learning. Our tests and Confusion Matrix assessments suggest Naive Bayes trumps k-NN [7].

Liu et al., Early diabetes complications like Diabetic Retinopathy (DR) are a primary cause of blindness. Frequent fundus imaging screenings may prevent DR in clinical diagnosis and treatment. Most DR screening studies use fundus images with a limited imaging range, field of vision, and lesion information, which leads to poor automated DR grading. We develop 101 ultra-wide-field (UWF) DR fundus images and propose Deep Learning (DL) automated classification system based on a unique preprocessing method to improve DR grading accuracy. Expanded UWF fundus images give more lesion information and a wide imaging range. UWF picture classification improves with data denoising and contrast and brightness augmentation. We use multiple DL classification models to evaluate our dataset and preprocessing. Experiments show the backbone model alone classifies well. The simplest ResNet50 model has ACA 0.66, Macro F1 0.6559, and Kappa 0.58. The best Swin-S model has ACA 0.72, Macro F1 0.7018, and Kappa 0.65. Clinicians benefit from UWF images' improved DR grading accuracy and efficiency [8].

Mercaldo et. al., Medical research has provided evidence indicating a notable rise in the prevalence of diabetic pathology during the past few decades, with no apparent indication of this trend abating. This study presents a proposed approach for classifying individuals who have diabetes, to aid and expedite the diagnostic process. The method involves utilizing a collection of characteristics that have been selected based on the criteria outlined by the World Health Organization. By employing cutting-edge machine learning techniques, we have successfully assessed real-world data and achieved an accuracy score of 0.770 and a recall score of 0.775 by utilizing the HoeffdingTree algorithm [9].

Wu et. al., The worldwide occurrence and frequency of Diabetes Mellitus (DM) have attained epidemic levels. According to projections, the number of individuals impacted by DM is expected to surpass 360 million by 2030. All of these patients are susceptible to the development of DR. The categorization, classification, and staging of DR are important in determining appropriate therapeutic interventions. By using effective management strategies, it is possible to prevent over 90% of instances resulting in vision impairment. The primary objective of this present study is to conduct a comprehensive examination of the categorization of diabetic retinopathy (DR), with particular attention given to the International Clinical Disease Severity Scale for DR. The newly proposed categorization system is characterized by its user-friendly nature, ease of memorization, and foundation in empirical scientific research. Specialized exams such as optical coherence tomography or fluorescein angiography are not necessary. The determination is made by a clinical assessment and the use of the Early Treatment of Diabetic Retinopathy Study 4:2:1 guideline [10].

Nahzat et al., The utilization of artificial intelligence in healthcare systems has seen significant advancements recently. Machine learning (ML) is extensively employed in the field of medical diagnostics for a diverse range of applications. Machine learning methodologies are employed in the prediction and detection of a diverse range of potentially fatal medical conditions, encompassing cancer, diabetes, cardiovascular disease, thyroid disorders, and other similar ailments. Chronic diabetes is a prevalent global ailment, and expediting and streamlining the diagnostic procedures will significantly impact the subsequent treatment protocols. The primary objective of this

study is to employ diverse machine learning approaches for diabetes prediction, followed by analyzing the output generated by these techniques. This research aims to identify the most effective classifier with the highest level of accuracy. This study investigates diabetes prediction by considering many variables associated with the condition. The Pima Indian Diabetes Dataset was utilized in this study to investigate the effectiveness of several Machine Learning classification approaches, including k-NN, RF, SVM, Artificial Neural Network (ANN), and DT, in predicting diabetes. The models employed in this investigation exhibit varying levels of accuracy. This study presents a predictive algorithm that demonstrates accurate forecasting capabilities for diabetes. According to the findings of this study, the random forest algorithm has superior accuracy in predicting diabetes compared to other machine learning approaches [11].

Feature selection aims to decrease the number of characteristics in a dataset or identify the significant ones. This approach efficiently selects the characteristics to be used as input for a machine learning model. Datasets often contain multiple factors, but not all improve the model's performance, and some may even cause overfitting issues. Feature selection is used to improve the model's performance, reduce training time, and minimize the impact of irrelevant or duplicated features. Efficient feature selection is essential in machine learning applications because it improves the model's performance, reduces training time, and enhances interpretability. It is important to note that feature selection should be objective and avoid subjective evaluations. The literature review used traditional optimization algorithms for feature selection. The study employed the innovative marine predator optimization system. Comparison of the categorization results with existing algorithms shows superior outcomes. The diabetes dataset is obtained from Kaggle. The dataset definition is explained in Sub-section 3.1.

3. MATERIALS AND METHODS (MATERYAL VE METOD)

3.1. Dataset Description (Veri Seti Tanımı)

Diabetes is a metabolic illness with a multifaceted origin. Several variables influence the incidence of the disease. Genetic susceptibility can increase the likelihood of developing diabetes, especially in individuals with a family history of the disease. Type 1 diabetes occurs when the immune system attacks the insulin-producing cells in the pancreas. Type 2 diabetes is often associated with obesity, advanced age, a sedentary lifestyle, and genetic predisposition. Obesity increases the risk of type 2 diabetes by causing insulin resistance in the body's fat tissue. Insulin resistance is when the body cannot efficiently use its insulin. Aging can also decrease the pancreas's ability to produce insulin, which can lead to the development of diabetes. Gestational diabetes may develop due to hormonal fluctuations during pregnancy. Environmental factors, including viral infections, may influence type 1 diabetes. A combination of factors affects the development of diabetes. Understanding the beginning of diabetes requires considering genetic, environmental, and behavioral factors that differ among individuals. Personalized techniques should be employed in addressing these intricate challenges in treatment and preventative efforts. The dataset originates from the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset aims to use diagnostic measurements to predict the presence or absence of diabetes in patients.

Several limitations were imposed on selecting these examples from a more extensive database. Specifically, the patient population in this study comprises exclusively of adult females at least 21 years of age and Pima Indian descent [12].

- The topic of pregnancies is being discussed. Frequency of pregnancies
- The plasma glucose concentration at the 2-hour mark in an oral glucose tolerance test is called glucose.
- The variable of interest in this study is diastolic blood pressure, measured in millimeters of mercury (mm Hg).
- Skin thickness is measuring the thickness of the triceps skin fold, expressed in millimeters.
- The measurement of insulin in the serum after 2 hours is denoted as 2-hour serum insulin (mu U/ml).
- The body mass index (BMI) is a metric used to assess an individual's body weight relative to height. It is calculated by dividing the weight in kilograms by the square of the height in meters.
- The variable "DiabetesPedigreeFunction" refers to the diabetes pedigree function.
- Age is typically defined as the number of years a person has lived.

- The outcome is represented by a class variable that can take 0 or 1 values.

In Figure 1, the graphs of the values of each feature of the data set are shown separately.

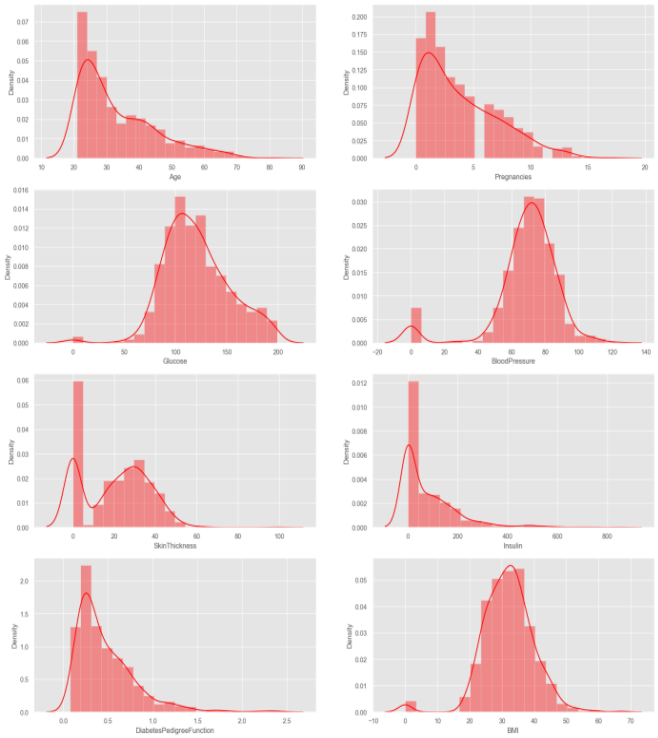

Figure 1. Diabets data analysis (Diyabet veri seti analizi)

3.2. System Description (Sistem Tanımı)

The diabetic dataset obtained from Kaggle is subsequently submitted to the feature extraction module following the data preprocessing phase. Following the optimization procedure, the feature selection block identifies the prominent features. The aforementioned determined features are categorized within the classification module. The results are contrasted with those produced without the optimization procedure concerning machine learning metrics such as F1 score, Recall, and accuracy. Figure 2 shows the block diagram of the system description.

.

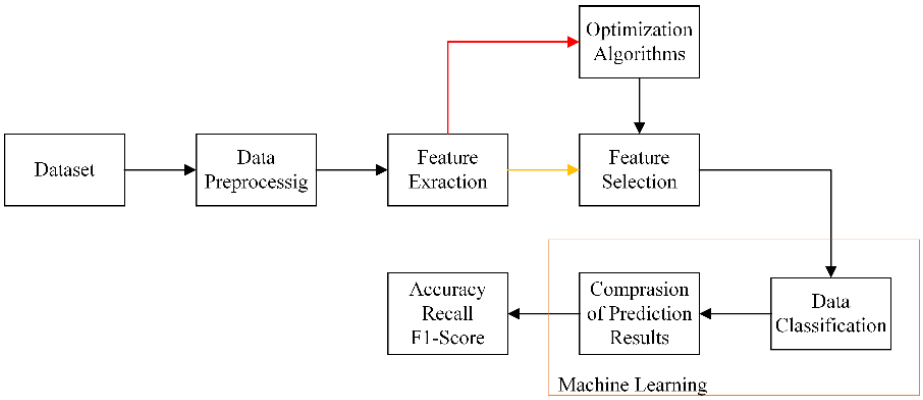

Figure 2. System description block diagram (Sistem tanımı blok diyagramı)

3.3. Marine Predator Optimization Algorithm (Deniz Yırtıcısı Optimizasyon Algorimasi)

The MPOA was formulated by Faramarzi, drawing inspiration from the interplay between marine predators and their prey in a social context. The MPOA algorithm is a heuristic optimization technique developed utilizing the encounter rate between marine predators and their prey. The initial solution for the MPOA algorithm commences by employing a stochastic approach to distribute the search space randomly. The algorithm's phase transitions are determined by the speed ratio between the prey and the hunter, per the principles of MPOA. Marine predators employ a three-phase

approach while pursuing and capturing their prey. The primary characteristic that sets the initial phase of the algorithm apart is its notable efficiency level. The concepts of oneness and a low ratio gain prominence in various stages. In phase 1 of optimization, prey travels in Brownian motion. Given the homogeneous distribution of prey in initial iterations and the considerable distance between predator and prey, Brownian move can aid prey in exploring their neighborhood independently, leading to effective domain exploration. Then, the prey in a new position is assessed for fitness and replaced if it is better. The saving technique is the prey's Recall of bountiful food regions in their fitting settings. Prey is a predator if it forages better. The top predator gets replaced with a better-fitted prey based on its fitness value. While prey is still searching for food, predators should start feeding. The second optimization step begins here. During phase 2 of the algorithm, both the prey and the predator move at the same pace. This phase also contains the second two-thirds of the algorithm. This place employs a variety of movement approaches. At this stage, the predator utilizes the Brownian motion, while the prey utilizes the Levy motion. During this process stage, the prey is multiplied by a vector of random integers derived from Levy's motion normal distribution. The algorithm needs good exploitation capabilities in the final optimization step. In this phase, the predator switches from Brownian to Lévy strategy to seek a neighborhood more efficiently. Using the adaptive convergence factor in this phase helps predators focus on a specific neighborhood for exploitation, reducing wasted effort from lengthy step sizes in Lévy strategy for unproductive regions [13–15].

3.4. Machine Learning Classification Algorithms
(Makine Öğrenmesi Sınıflandırma Algoritmaları)

The k-NN method is considered one of the fundamental example-based learning algorithms. In the context of example-based learning algorithms, acquiring knowledge or skills is undertaken. The procedure is executed using the information included inside the training set. The classification of a newly encountered example is determined based on its resemblance to the examples already in the training set. The k-NN method is popular for classification problems. It is favored in many classification tasks owing to the ease with which it can be interpreted and the small amount of time it takes to compute. Within the context of the k-NN algorithm, the choice of the k parameter is of the most significance. Figure 3 shows the diagram of the k-NN algorithm. [16–18].

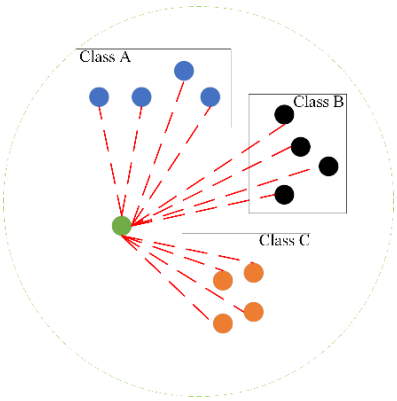

Figure 3. k-NN classification algorithm block diagram (k-NN sınıflandırma algoritması blok diyaramı)

SVM is a machine learning algorithm founded on convex optimization and functions according to the maximization of structural risk reduction. The approach in issue belongs to a class of learning algorithms known as distribution-free learning algorithms since it does not need prior knowledge of the joint distribution function of the data. The SVM aims to locate the best possible separation hyperplane to differentiate the classes. Put another way. The goal is to achieve the most significant possible separation between the support vectors associated with the various classes [19, 20]. Figure 4 shows the diagram of the SVM algorithm.

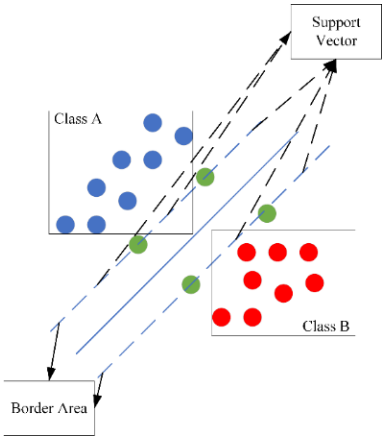

Figure 4. SVM classification algorithm block diagram (SVM sınıflandırma algoritması blok diyaramı)

In 1999, Friedman proposed the concept of Gradient Boosting as a method for ensemble learning in the fields of regression and classification. The process of constructing a robust ensemble classifier is iterative. This approach integrates less powerful models sequentially and opportunistically to get more robust estimators. As the model progresses, further trees are formed by considering the prediction errors made by previous trees [21].

GB is defined as the algorithm of evolving weak individuals into strong individuals through GB. The working principle of this algorithm is that the newly calculated decision trees are run to minimize the errors of the previous decision trees. In this algorithm, the primary decision tree is created by random guessing. The next decision tree is compared to the primary decision tree. These operations are calculated in specified iterations. It is aimed to minimize the error value [22, 23].

The DT classification method is a collection of rules used to analyze, identify, categorize, and predict statistically significant groups or communities. The DT classification approach involves iteratively partitioning the dataset to optimize the discrimination of dependent variables. The DT is a graphical representation that depicts various options and their corresponding outcomes as a tree structure. In the graph, the nodes symbolize events or choices, while the edges reflect decision rules or conditions. Every tree is composed of nodes and branches. In this classification framework, nodes symbolize traits within a particular group, while branches symbolize the potential values that each node can assume [24, 25].

The XGBoost algorithm can be considered as a refined version of the Gradient Boosting technique. The primary factor driving the extensive adoption of XGBoost is its notable advantages compared to its predecessors. The XGBoost algorithm utilizes the maximum depth parameter during the tree construction process. The pruning process is executed if the generated tree has an excessive downward trajectory. The prevention of overlearning is observed. The Gradient Boosting algorithm employs a first-order function to compute the loss function, whereas XGBoost utilizes second-order functions for these calculations. The parallel working feature enables the attainment of results within a reduced timeframe compared to alternative methods [26, 27].

A method known as RF, which is an ensemble learning classification and regression approach, is used to categorize the collected data and then arrange it into classes. Several decision trees are formed during training, which are then used for class prediction once refined. During the calculation process, the classes of each tree are considered, and the class that gained the most significant number of positive votes is regarded as the procedure's outcome [28, 29].

The LR-based classification process involves assigning an arbitrary set of inputs to a function, which then generates the output by categorizing the input data. The classification function produces binary outputs of 0 or 1 to enhance computational efficiency, representing the two distinct classes. Based on the identified requirements and the analysis above, it can be concluded that the range of the function argument mentioned above spans from positive infinity to negative infinity. The range of the dependent variable is limited to the values of 0 or 1. A multitude of functions exist that fulfill the requirements above. The 0-1 step function is often considered the most straightforward option. Nevertheless, the step function lacks differentiability at the step point, rendering it unsuitable for mathematical manipulation [30].

3.5. Performance Metrics (Performans Metrikleri)

The use of classification accuracy as a sole statistic may be undermined when there is an imbalance in the distribution of pictures across various classes in the dataset. This imbalance might result in misleading findings. A study of the confusion matrix obtains the determination of performance metrics for each class within the dataset. The abovementioned criteria include accuracy, Recall, precision, and F1 score [31, 32]. Accuracy is a performance measure that indicates the proportion of right predictions out of all predictions made by a classification model. Accuracy is frequently employed to assess the effectiveness of a model. Accuracy is a vital metric frequently employed to assess a model's performance. However, this might be deceptive in datasets that are not balanced or when the number of samples varies significantly between classes. Accuracy may not evaluate false positives and negatives fairly, especially when unusual classes are involved. Thus, other performance measures like precision, Recall, and F1 scores are utilized since accuracy might be deceptive [33, 34]. The mathematical equation for accuracy is given in Equation 1.

$$Acc = \frac{TP}{Total\ Instances} \quad (1)$$

Recall is a statistic that evaluates a classification model's ability to anticipate all positive cases correctly. Precision specifically aims to reduce false negatives and avoid missing real positives. Recall and accuracy are crucial performance indicators that complement each other. The Recall is critical to detecting a condition to reduce the number of false negative samples. It is frequently essential to strike a balance between precision and sensitivity when evaluating performance indicators. The F1 score combines these two criteria for a more comprehensive performance evaluation [35, 36]. The mathematical equation for Recall is given in Equation 2.

$$Recall = \frac{TP}{Total\ Actual\ Yes} \quad (2)$$

Precision is a statistic that calculates the proportion of cases correctly predicted as positive by the classification model out of all the actual positive instances. Precision specifically aims to manage false positives, which are occurrences predicted as positive by the model but are really negative. Precision is the proportion of true positive samples among all samples anticipated as positive. High precision suggests a high likelihood that the model's positive predictions are correct. Yet, accuracy alone may not suffice as a performance metric as it might be deceptive without taking into account the model's Recall. Various measures, including accuracy, sensitivity, and F1 score, are collectively employed to assess the effectiveness of classification algorithms. Equation 3 is given precision [33, 36].

$$Precision = \frac{TP}{Total\ Predicted\ Yes} \quad (3)$$

The F1 score is a metric that assesses the effectiveness of a classification model by combining precision with Recall. It is particularly useful for datasets with imbalanced classes and equally addresses false positives and negatives. Equation 4 is given F1 score [34, 36].

$$F1\ score = \frac{2 * Prec * Recall}{Prec + Recall} \quad (4)$$

Mean Squared Error (MSE) is a commonly used metric in statistics and machine learning to evaluate performance. It assesses the discrepancy between the model's predictions and the actual values, particularly in regression analysis. MSE is calculated as the average of the squared errors of the model. Equation 5 gives MSE [37, 38].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (5)$$

In Equation 5, n is the number of data points, y_i is the true value and $\hat{y}_i$ is the model's predicted value. MSE is a method that squares the errors to equalize negative and positive errors. However, this method may accentuate significant faults more, as the impact of big errors is amplified due to the errors being squared. A low MSE suggests accurate predictions by the model, whereas a large MSE implies poor performance. MSE is commonly used as a loss function in model optimization to improve forecasts by minimizing it [38, 39].

4. RESULTS AND DISCUSSION (SONUÇLAR VE TARTIŞMA)

This study performs feature selection with the marine predator diabetes dataset obtained from Kaggle. This feature selection set was classified using LR, RF, k-NN, GB, XgBoost, SVM, and DT classification methods. The feature selection dataset was applied and compared with the non-feature selection dataset regarding accuracy.

Table 1 is given the performance metrics table for the diabetes dataset obtained from Kaggle. The dataset was subjected to machine learning classification algorithms without the use of feature selection. The machine learning classification technique that had the best accuracy rate was LR, achieving a precision of 77.63%. Moreover, RF and GB's machine learning classification methods demonstrated promising results.

Table 1. Performance metrics without feature selection (Özellik seçilimsiz performans metrikleri)

Model	Precision (%)	Recall (%)	F1 Score (%)	Accuracy (%)
LR	79	88	83	77.63
RF	80	86	83	77.19
k-NN	80	83	81	75.88
GB	79	86	82	75.88
XgBoost	80	84	82	75.88
SVM	81	81	81	75.44
DT	79	79	79	73.25

The diabetes dataset obtained from Kaggle is subjected to feature selection using the marine predator optimization technique. The resulting performance metrics values for several machine learning classification algorithms are presented in Table 2. The machine learning classification approach that exhibited the highest accuracy rate was LR, with 79.39%. Furthermore, the RF and GB machine learning classification methods revealed favorable outcomes, as given in Table 2.

Table 2. Performance metric with feature selection (Özellik seçilimiyle performans metrikleri)

Model	Precision (%)	Recall (%)	F1 Score (%)	Accuracy (%)
LR	82	88	84	79.39
RF	79	84	81	77.63
k-NN	80	84	82	75.88
GB	82	84	83	77.63
XgBoost	80	84	82	75.88
SVM	80	86	83	76.75
DT	78	79	78	72.73

ROC curves are employed for several objectives, including assessing the test's discriminating ability, evaluating result quality, and comparing diagnostic performance across two or more conditions. Additionally, the ROC curve aids in the comprehension of graphical data. In this study, Figure 5 shows the Receiver Operating Characteristic (ROC) analysis of the data acquired without optimization, whereas Figure 6 shows cases of the analysis of the results achieved with optimization. In each case, all findings indicated levels deemed acceptable for categorization purposes. Upon closer examination, it was seen that the k-NN, LR, and XGBoost models exhibited higher levels of success in Figure 5. Conversely, in Figure 6, the LR and GBDT methods yielded superior outcomes.

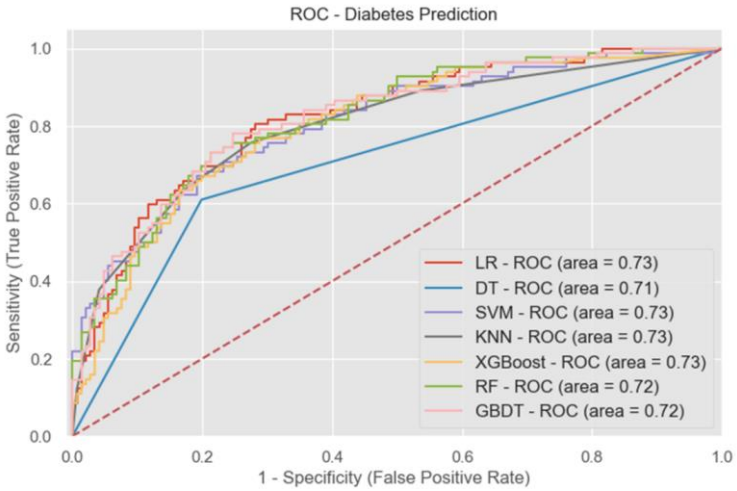

Figure 5. RoC curve without feature selection (Özellik seçimlimsiz RoC eğirisi)

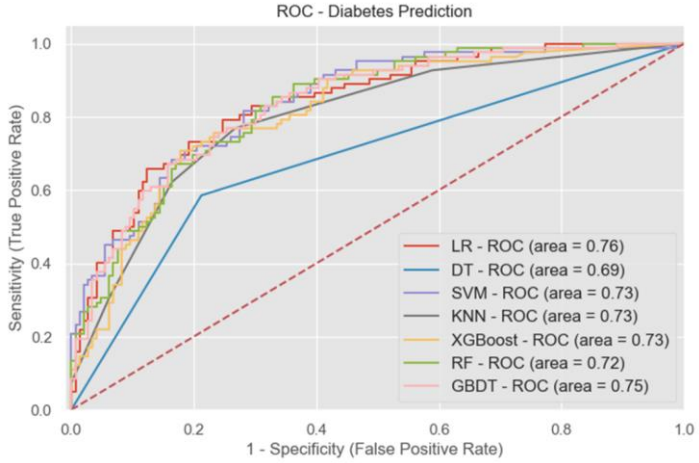

Figure 6. RoC curve with feature selection (Özellik seçilimli RoC eğirisi)

Upon comparing the collected results with existing literature reviews, it becomes evident that the findings have yielded encouraging outcomes. It is anticipated that modifying the classification approach and feature selection algorithm would improve the accuracy rate. The comparison between the current study and the existing literature is given in Table 3. an accuracy of 77.63%/75.44% was achieved with the LR and SVM model without using the optimization method. As a result of the classification of the features determined by MPOA, an accuracy value of 79.39% / 76.75% was achieved with LR and SVM. What is important here is the classification success of the improvement made with MPOA rather than the highest accuracy value

Table 2. Classification of diabetes set and comparison of accuracies with literature review

Article	Classification Algorithm	Feature Selection Optimization Algorithm	Accuracy
Sisodia et al.[5]	Naive Bayes	-	76.30%
Kaur et al.[6]	Improvement J48	-	99.87%
Liu et al. [8]	Swin-S	-	72%
Nahzat et al. [11]	RF	-	88%
In this study	LR/SVM	MPOA	79.39%/ 76.75%
In this study	LR/SVM	-	77.63%/ 75.44%

4. CONCLUSIONS (SONUÇLAR)

This study undertakes a comparative examination of performance metrics for the diabetes dataset by employing machine learning classification algorithms, with and without character selection. The main finding is that incorporating optimization techniques into machine learning frameworks might enhance performance. Previous research has primarily concentrated on utilizing machine learning methods for categorizing diabetes datasets, and this study exhibits similar conditions to those encountered in past investigations. However, recent empirical findings have supported the notion that implementing efficient optimization techniques leads to notable enhancements in performance measures by carefully selecting pertinent features. Moreover, alongside implementing optimization strategies, significant enhancements were achieved in workload management and cost reduction. As a result, the focus was primarily on analyzing the dataset's attributes, with less emphasis on the parameters' importance. The results suggest that the LR classification method attains a 77.63% accuracy rate without feature selection. Nevertheless, when using the marine predator algorithm for property selection, the accuracy rate notably increased, reaching 79.39%. Future research is expected to explore other optimization and classification strategies, offering alternate approaches with more effective algorithms. Furthermore, it is possible to ascertain the primary components of diabetes within the dataset.

DECLARATION OF ETHICAL STANDARDS (ETİK STANDARTLARIN BEYANI)

The author of this article declares that the materials and methods they use in their work do not require ethical committee approval and/or legal-specific permission.

Bu makalenin yazarı çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal, özel bir izin gerektirmediğini beyan ederler

AUTHORS' CONTRIBUTIONS (YAZARLARIN KATKILARI)

Fuat TÜRK: He analyzed the results and contributed to the writing process.

Deney sonuçlarını analiz etmiş ve makalenin yazımına katkıda bulunmuştur.

Nuri Alper METİN: He analyzed the results and contributed to the writing process.

Deney sonuçlarını analiz etmiş ve makalenin yazımına katkıda bulunmuştur.

Murat LÜY: He conducted the experiments, analyzed the results, and contributed to the writing process.

Deneyleri yapmış, sonuçlarını analiz etmiş ve makalenin yazımına katkıda bulunmuştur.

CONFLICT OF INTEREST (ÇIKAR ÇATIŞMASI)

There is no conflict of interest in this study.

Bu çalışmada herhangi bir çıkar çatışması yoktur.

REFERENCES (KAYNAKLAR)

[1] İ. Kabalı and S. Özan, "Communication with Chronic Patients and Patient Relatives in the Example of Diabetes Disease," Tıp Eğitimi Dünyası, vol. 19, no. 57, pp. 109–119, 2020, doi: 10.25282/ted.576901.

[2] B. Aydoğan, A. Aydın, M. B. İnci, and H. Ekerbiçer, "TİP 2 Diyabet Hastalarinin HastaliklariyİlgiliBilgi, Tutum Düzeyleri İlişkili Faktörleri Değerlendirilmesi," Sak. Med. J., 2020, doi: 10.31832/smj.743455.

[3] T. Gülsün and S. Şahhn, "Diyabet ve Diyabete Bağlı Fizyolojik ve Farmakokinetik Değişiklikler," Hacettepe Univ. J. Fac. Pharm., vol. 37, no. 2, pp. 105–123, 2017.

[4] A. Abac, "Tip 1 Diyabet türkçe," no. 8, pp. 1–10, 2007.

[5] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," Procedia Comput. Sci., vol. 132, no. Iccids, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.

[6] G. Kaur and A. Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes," Int. J. Comput. Appl., vol. 98, no. 22, pp. 13–17, 2014, doi: 10.5120/17314-7433.

[7] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, "Diabetes prediction using supervised machine learning," Procedia Comput. Sci., vol. 216, no. 2022, pp. 21–30, 2022, doi: 10.1016/j.procs.2022.12.107.

[8] H. Liu, L. Teng, L. Fan, Y. Sun, and H. Li, "A new ultra-wide-field fundus dataset to diabetic retinopathy grading using hybrid preprocessing methods," Comput. Biol. Med., vol. 157, no. 2699, p. 106750, 2023, doi: 10.1016/j.compbiomed.2023.106750.

[9] F. Mercaldo, V. Nardone, and A. Santone, "Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques," Procedia Comput. Sci., vol. 112, pp. 2519–2528, 2017, doi: 10.1016/j.procs.2017.08.193.

[10] L. Wu, "Classification of diabetic retinopathy and diabetic macular edema," World J. Diabetes, vol. 4, no. 6, p. 290, 2013, doi: 10.4239/wjd.v4.i6.290.

[11] S. NAHZAT and M. YAĞANOĞLU, "Makine Öğrenimi Sınıflandırma Algoritmalarını Kullanarak Diyabet Tahmini," Eur. J. Sci. Technol., no. 24, pp. 53–59, 2021, doi: 10.31590/ejosat.899716.

[12] Kaggle, Available: https://www.kaggle.com/datasets/mathchi/diabetes-data-set

[13] A. Faramarzi, M. Heidarinejad, S. Mirjalili, and A. H. Gandomi, "Marine Predators Algorithm: A nature-inspired metaheuristic," Expert Syst. Appl., vol. 152, p. 113377, 2020, doi: 10.1016/j.eswa.2020.113377.

[14] Z. Garip, M. Çimen, and A. Boz, "Otomatik Gerilim Regülatör Sistemi için Deniz Yırtıcıları Algoritmasının Performans Analizi," Acta Infologica, vol. 0, no. 0, pp. 0–0, 2022, doi: 10.26650/acin.1026494.

[15] S. Mugemanyi et al., "Marine predators algorithm: A comprehensive review," Mach. Learn. with Appl., vol. 12, no. June, p. 100471, 2023, doi: 10.1016/j.mlwa.2023.100471.

[16] O. ULUDAĞ and A. GÜRSOY, "Financial Risk Estimation with KNN Classification Algorithm on Determined Financial Ratios," Eur. J. Sci. Technol., no. 29, pp. 26–29, 2021, doi: 10.31590/ejosat.1001663.

[17] M. Lüy, N. A. Metin "Classification of heart disease dataset with k-NN optimized by pso and gwo algorithms," 2023, doi: 10.51271/JCEEES-0009.

[18] E. Akkur, "Investigatıon of the effect of feature selection and hyperparameter optimizatıon method on machine learning," no. July, 2023.

[19] A. G. Kakisim, Z. Turgut, and T. Atmaca, "XAI Empowered Dual Band Wi-Fi Based Indoor Localization via Ensemble Learning," 2023 14th Int. Conf. Netw. Futur., pp. 150–158, 2023, doi: 10.1109/NoF58724.2023.10302788.

[20] E. Akkur, F. Turk, and O. Erogul, "Breast cancer diagnosis using feature selection approaches and bayesian optimization," Comput. Syst. Sci. Eng., vol. 45, no. 2, pp. 1017–1031, 2023, doi: 10.32604/csse.2023.033003.

[21] K. Çoşkun and G. Çetin, "a Comparative Evaluation of the Boosting Algorithms for Network Attack Classification," Int. J. 3D Print. Technol. Digit. Ind., vol. 6, no. 1, pp. 102–112, 2022, doi: 10.46519/ij3dptdi.1030539.

[22] V. A. Dev and M. R. Eden, "Formation lithology classification using scalable gradient boosted decision trees," Comput. Chem. Eng., vol. 128, pp. 392–404, 2019, doi: 10.1016/j.compchemeng.2019.06.001.

[23] P. Li, C. J. C. Burges, and Q. Wu, "McRank: Learning to rank using multiple classification and gradient boosting," Adv. Neural Inf. Process. Syst. 20 - Proc. 2007 Conf., no. 1, 2008.

[24] D. Altaş and V. Gürpınar, "Karar ağaçları ve yapay sinir ağlarının sınıflandırma performanslarının karışılaştırılması: avrupa

birliği örneği," Trak. Üniversitesi Sos. Bilim. Derg., vol. 14, no. 1, pp. 1–22, 2012.

[25] A. Çalış, S. Kayapınar, and T. Çetinyokuş, "Veri madenciliğinde karar ağacialgoritmalari ile bilgisayar ve internet güvenliği üzerine bir uygulama," Endüstri Mühendisliği, vol. 25, no. 3,pp.2–19, 2014, Available: http://dergipark.org.tr/endustrimuhendisligi/issue/46771/586362

[26] M. Trafi, D. Sald, M. Shap, A. C. Kelle, M. Queuing, and T. Transport, "Araştırma Makalesi / Research Article," vol. 3, no. 1, pp. 50–62, 2022.

[27] M. Tokmak, "XGBoost Algoritması ile ikili parçacık sürü optimizasyonu öznitelik seçme tabanlı jar kötü amaçlı yazılımlarının tespiti jar malware detection with xgboost algorithm based on binary particle swarm optimization feature selection," vol. 10, no. 1, pp. 140–152, 2023.

[28] C. D. Kumral, A. Topal, M. Ersoy, R. Çolak, and T. Yiğit, "Performing Performance Analysis by Implementing Random Forest Algorithm on FPGA," El-Cezeri J. Sci. Eng., vol. 9, no. 4, pp. 1315–1327, 2022, doi: 10.31202/ecjse.1134799.

[29] Ö. Akar and O. Güngör, "Rastgele orman algoritması kullanılarak çok bantlı görüntülerin sınıflandırılması," J. Geod. Geoinf., vol. 1, no. 2, pp. 139–146, 2012, doi: 10.9733/jgg.241212.1t.

[30] X. Zou, Y. Hu, Z. Tian, and K. Shen, "Logistic Regression Model Optimization and Case Analysis," Proc. IEEE 7th Int. Conf. Comput. Sci. Netw. Technol. ICCSNT 2019, pp. 135–139, 2019, doi: 10.1109/ICCSNT47585.2019.8962457.

[31] E. Sivari and S. Sürücü, "Prediction of heart attack risk using linear discriminant analysis methods," J. Comput. Electr. Electron. Eng. Sci., vol. 1, no. 1, pp. 5–9, 2023, doi: 10.51271/jceees-0002.

[32] Ö. Vupa Çilengiroğlu and A. Yavuz, "Lojistik regresyon ve cart yöntemlerinin tahmin edici performanslarının yaşam memnuniyeti verileri için karşılaştırılması," Eur. J. Sci. Technol., no. 18, pp. 719–727, 2020, doi: 10.31590/ejosat.691215.

[33] A. Göde and A. Kalkan, "Performance comparison machine learning algorithms in diabetes disease prediction," Eur. Mech. Sci., vol. 7, no. 3, pp. 178–183, 2023, doi: 10.26701/ems.1335503.

[34] M. İ. Gürsoy and A. Alkan, "Investigation Of Diabetes Data with Permutation Feature Importance Based Deep Learning Methods," Karadeniz Fen Bilim. Derg., vol. 12, no. 2, pp. 916–930, 2022, doi: 10.31466/kfbd.1174591.

[35] Z. PAMUK and C. KAYA, "Classification of Type 2 Diabetes Using Machine Learning Techniques," Eur. J. Sci. Technol., no. 28, pp. 1265–1268, 2021, doi: 10.31590/ejosat.1014878.

[36] Ö. N. ERGÜN and H. O.İLHAN, "Early Stage Diabetes Prediction Using Machine Learning Methods," Eur. J. Sci. Technol., no. 29, pp. 52–57, 2021, doi: 10.31590/ejosat.1015816.

[37] Y. GÜLTEPE, "Makine Öğrenmesi Algoritmaları ile Hava Kirliliği Tahmini Üzerine Karşılaştırmalı Bir Değerlendirme," Eur. J. Sci. Technol., no. 16, pp. 8–15, 2019, doi: 10.31590/ejosat.530347.

[38] F. M. sakran Alamery, "Cryptocurrency analysis using machine learning and deep learning approaches," J. Comput. Electr. Electron. Eng. Sci., vol. 1, no. 2, pp. 29–33, 2023, doi: 10.51271/jceees-0007.

[39] U. Tanyeri, T. Dindar, Y. Kökver, and N. F. Koçak, "Machine learning methods on quantized vectors," J. Comput. Electr. Electron. Eng. Sci., vol. 1, no. 2, pp. 46–49, 2023, doi: 10.51271/jceees-0010.