Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi

Afyon Kocatepe University Journal of Science and Engineering

AKÜ FEMÜBİD **22** (2022) 035101 (570-576) AKU J. Sci. Eng. **22** (2022) 035101 (570-576)

DOI: 10.35414/akufemubid.1011058

Araştırma Makalesi / Research Article

Balancing the Dataset by Generating New Synthetic Data Based on Heinz Mean in Medical Data

Ibrahim Halil GÜMÜŞ[1], Serkan GÜLDAL[2*]

[1]*Adiyaman University, Faculty of Arts and Sciences, Department of Mathematics, Adiyaman, 02040, Turkey*

[2]*Adiyaman University, Faculty of Arts and Sciences, Department of Physics, Adiyaman, 02040, Turkey*

e-mail[1]: igumus@adiyaman.edu.tr ORCID ID: https://orcid.org/0000-0002-3071-1159

Corresponding Author e-mail[2]: sguldal@adiyaman.edu.tr* ORCID ID: https://orcid.org/0000-0002-4247-0786

Submission Date: 18.10.2021 Acceptance Date: 16.05.2022

Abstract

Advances in science and technology have caused data sizes to increase at a great rate. Thus, unbalanced data has arisen. A dataset is unbalanced if the classes are not nearly equally represented. In this case, classifying the data causes performance values to decrease because the classification algorithms are developed on the assumption that the datasets are balanced. As the accuracy of the classification favors the majority class, the minority class is often misclassified. The majority of datasets, especially those used in the medical field, have an unbalanced distribution. To balance this distribution, several studies have been performed recently. These studies are undersampling and oversampling processes. In this study, distance and mean based resampling method is used to produce synthetic samples using minority class. For the resampling process, the closest neighbors for all data points belonging to the minority class were determined by using the Euclidean distance. Based on these neighbors and using the Heinz Mean, the desired number of new synthetic samples were formed between each sample to obtain balance. The Random Forest (RF) and Support Vector Machine (SVM) algorithms are used to classify the raw and balanced datasets, and the results were compared. Additionally, the other well known methods (Random Over Sampling-ROS, Random Under Sampling-RUS, and Synthetic Minority Oversampling TEchnique-SMOTE) are compared with the proposed method. It was shown that the balanced dataset using the proposed resampling method increases classification efficiency as compared to the raw dataset and other methods. Accuracy measurements of RF are 0.751 and 0.799 and, accuracy measurements of SVM are 0.762 and 0.781 for raw data and resampled data respectively. Likewise, there are improvements in the other metrics such as Precision, Recall, and F1 Score.

Keywords
Machine learning; Synthetic data; Balanced data; Heinz mean

Tıbbi Verilerde Heinz Ortalamasına Dayalı Yeni Sentetik Veriler Üreterek Veri Kümesini Dengeleme

Öz

Bilim ve teknolojideki ilerlemeler veri boyutlarının büyük hızda artmasına neden olmuştur. Böylece dengesiz veriler ortaya çıkmıştır. Sınıflar neredeyse eşit olarak temsil edilmiyorsa, bir veri kümesi dengesizdir. Bu durumda sınıflandırma algoritmaları veri setlerinin dengeli olduğu varsayımı ile geliştirildiği için verilerin sınıflandırılması performans değerlerinin düşmesine neden olur. Sınıflandırmanın doğruluğu çoğunluk sınıfını desteklediğinden, azınlık sınıfı genellikle yanlış sınıflandırılır. Özellikle tıp alanında kullanılan veri kümelerinin çoğu dengesiz bir dağılıma sahiptir. Bu dağılımı dengelemek için son zamanlarda çeşitli çalışmalar yapılmıştır. Bu çalışmalar, eksik örnekleme ve aşırı örnekleme süreçleridir. Bu çalışmada, azınlık sınıfı kullanılarak sentetik örnekler üretmek için uzaklık ve ortalama tabanlı yeniden örnekleme yöntemi kullanıldı. Yeniden örnekleme işlemi için, azınlık sınıfına ait tüm veri noktaları için en yakın komşular Öklid uzaklığı kullanılarak belirlendi. Bu komşulara dayalı olarak ve Heinz Ortalaması kullanılarak veri setini dengeye getirmek için her numune arasında istenilen sayıda yeni sentetik numuneler oluşturuldu. Ham ve dengeli veri setlerini sınıflandırmak için Rassal Orman (RF) ve Destek Vektör Makinesi (SVM) algoritmaları kullanıldı ve sonuçlar karşılaştırıldı. Ayrıca, iyi bilinen diğer yöntemler (ROS, RUS ve SMOTE) önerilen yöntemle karşılaştırılmıştır. Önerilen yeniden örnekleme yöntemini kullanan dengeli veri kümesinin, ham veri kümesi ve diğer yöntemlere kıyasla sınıflandırma verimliliğini artırdığı gösterilmiştir. Sırasıyla ham veriler ve yeniden örneklenmiş veriler için RF'nin

Anahtar kelimeler
Makine Öğrenimi; Sentetik veri; Dengesiz veri; Heinz ortalaması

doğruluk ölçümleri 0.751 ve 0.799'dur ve SVM'nin doğruluk ölçümleri 0.762 ve 0.781'dir. Aynı şekilde Kesinlik, Hassasiyet ve F1 Skoru gibi diğer metriklerde de iyileştirmeler vardır.

1. Introduction

Machine learning and deep learning methods have been widely used in the medical diagnosis of diseases (Gopinath *et al*. 2019). However, due to the large volume, multidimensionality, and complexity of medical data, the problem of imbalance between classes arises. In such cases, the direct use of classification algorithms on the raw dataset causes performance degradation. In order to prevent performance losses and increase the predictive power of the classifier algorithms, the classes can be balanced (Mohammed *et al*. 2020). In the medical dataset, some classes are represented by a large number of samples, while others are represented by only a few. As a result of the research, several methods are discussed about this problem. These are considered as data-level resampling (Random Over Sampling, ROS, and Random Under Sampling, RUS), learning algorithm selection according to imbalance situation, and the relationship between class imbalance and cost-sensitive learning (Elreedy and Atiya 2019). The most successful results were obtained from the ROS model, one of the sampling methods generally used at the data level (Fotouhi *et al*. 2019). In addition, hybrid applications have been made to combine data and algorithm levels to achieve more successful results (Krawczyk 2016).

There have been many studies in the literature on balancing unbalanced data in medical and other fields (Chawla *et al*. 2002, Han *et al*. 2005). Various problems arise due to low model accuracies resulting from unbalanced datasets (Chawla *et al*. 2004). In such datasets, the number of majority observations is higher than the number of minority observations. Majority class observations are more effective in classification methods and minority observations are generally ignored. Therefore, some observations belonging to the minority class are misclassified. This case reduces the accuracy of the model. Many resampling approaches have been proposed to improve the model performance of classifiers (Kovács 2019). The Synthetic Minority Oversampling TEchnique-SMOTE algorithm, which produces synthetic observations, is one of the basic samplings that increases the success rate (Chawla, Bowyer, Hall, and Kegelmeyer 2002). SMOTE resamples data by randomly generated synthetic data between minority class data. The SVM-SMOTE algorithm is proposed to generate new minority observations near the boundary between the majority and minority classes (Nguyen *et al*. 2011). In medical and other datasets of different sizes, RUS, which randomly deletes from the majority class, and ROS, which randomly copies from the minority class, can reduce the rates of evaluation metrics according to the dataset. To solve this problem and reduce noisy data in datasets, the hybrid method SMOTE-ENN is proposed (Batista *et al*. 2004). As a result of the performance analysis of these methods proposed in the literature, more successful results were obtained by balancing medical data (Rahman and Davis 2013).

In this study, a synthetic sample generation study is conducted using a Heinz mean-based approach using a dataset with an unbalanced distribution of diabetes patients. The difference between this study and the methods previously used in synthetic sample replication is that Heinz mean generated synthetic samples is a distinctive approach at the mean level. The aim of the study is to balance the raw data with this proposed method and to obtain more successful results. The dataset balanced with the proposed method is classified with Random Forest and Support Vector Machine algorithms and the results are compared. Accuracy, Precision, Recall, and F1 score values are taken into account as performance values.

2. Materials and method

To remedy the imbalanced dataset problem, the medical dataset is selected since it is a common problem to obtain a balanced dataset. By the proposed method, the dataset is balanced and used to feed machine learning methods, namely random forest and support vector machines. Additionally, the findings are compared with the well-known

dataset balance methods. The detailed explanations are given in the following sections.

2.1. Dataset Used

In this study, Pima Indians real diabetes dataset is used from KEEL (Knowledge Extraction based on Evolutionary Learning) opensource software tool site (Int. Ref. 1). Although the study covers the medical data as a case study, the proposed method has the flexibility to be applied to any imbalanced dataset which has positive values. The aim of this dataset is to diagnose whether or not a patient has diabetes based on certain diagnostic measures included in the dataset. In this dataset, a total of 768 patient women were subjected, including 500 of them not having diabetes and 268 of them having diabetes with 8 attributes are available for each sample. These attributes are; Pregnancies, Glucose, Blood pressure (mm Hg), Skin thickness (mm), Insulin (mu U/ml), Body mass index (kg/m^2), Diabetes pedigree function, Age (years) and Outcome (Diabetic = 1 and Non-Diabetic = 0).

The imbalance ratio between the diabetic and non-diabetic classes is 53%. It is clear that the dataset's distribution is unbalanced based on the numbers and this ratio. In this study, the diabetic patient population in the minority class is resampled and approximated to the majority class using the approach suggested in this study, so the dataset is balanced.

2.2. Proposed Method

In this study, a strategy that differs from previous methods has been used to solve the problem of unbalanced data distribution. In this approach, the Euclidean distance metric is used to calculate the distance between the closest neighboring couples in the minority class. By using Heinz mean, among the samples described in the predefined range, synthetic data is generated in an amount equal to the number of balances required. The method steps developed in the study are as follows;

- Firstly, the imbalance ratio is calculated by dividing the number of samples in the majority class by the number of samples in the minority class in the dataset. If the dataset is unbalanced, other steps are applied.
- In order to balance the dataset, a sufficient number of synthetic data is generated from the minority class. The Euclidean distance metric, which calculates the distance between the two samples, was used during this process. If $x = [x_1, x_2, \dots . x_n]^T$ and $y = [y_1, y_2, \dots . y_n]^T$, then the Euclidean distance metric is shown by formula 1.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

- For the samples, a specified zone around every point was determined then the remaining samples within the zone were paired with the selected datum. The range of these areas is determined by brute force until the missing number of data points is obtained.
- We use Heinz mean to generate the synthetic samples. The Heinz mean for vectors with positive components is defined as follows: Let α be a random number in $[0,1]$. For $x = [x_1, x_2, \dots . x_n]^T$, $y = [y_1, y_2, \dots . y_n]^T \in R_+^n$

$$x \,\Delta\, y = \begin{bmatrix} \frac{x_1^{\alpha} y_1^{1-\alpha} + x_1^{\alpha} y_1^{1-\alpha}}{2} \\ \frac{x_2^{\alpha} y_2^{1-\alpha} + x_2^{\alpha} y_2^{1-\alpha}}{2} \\ \vdots \\ \frac{x_n^{\alpha} y_n^{1-\alpha} + x_n^{\alpha} y_n^{1-\alpha}}{2} \end{bmatrix} \quad (2)$$

- For generating synthetic samples, we use the following formula,

$$S_{new} = x \,\Delta\, y \quad (3)$$

- Formula 3 is repeated until all of the pairs are utilized.
- A reasonable amount of synthetic data is generated from the minority class in order to balance the dataset.

2.3. Random Forest Algorithm

There exist several classification techniques such as Decision Trees, K-Nearest Neighbor, and Support Vector Machine. Due to its simplicity, comprehensibility, and high predictive efficiency, decision tree is the most common and effective method among these. Even though decision trees have many advantages over the other classifiers, they have several disadvantages, including inconsistency and many more. However, they are eliminated by constructing a random forest. The random forest is one way to maximize the potential of these decision tree algorithms. The Random Forest algorithm predicts future instances using several classifiers rather than a single classifier to improve the accuracy and correctness of prediction (Breiman 2001). Random Forest takes a random subset of variables to obtain a split at each node of the trees. For classification, the input vector is transferred to each tree in the algorithm, and each tree casts a vote for one of the classes.

The class with the most votes is chosen by the algorithm (Liaw and Wiener 2002). The working system of the Random Forest algorithm is shown in Figure 1.

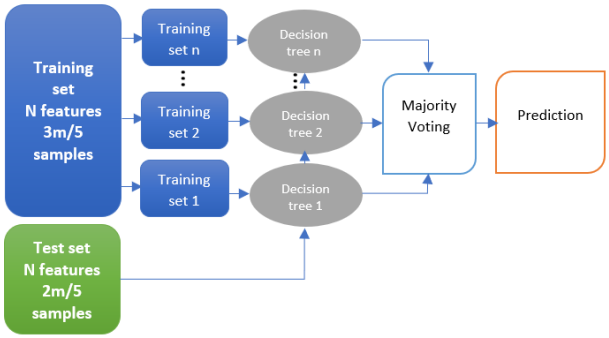

Figure 1. Random Forest algorithm working system. m is the number of total samples and n is the number of the decision tree.

As shown in Figure 1, the Random Forest approach requires two separate data classes. These are a training dataset (in-bag) and a test dataset (out-of-bag, oob). Throughout our analysis, the training dataset makes up 3/5 of the data and the test dataset makes up 2/5. The training dataset is used to train the tree, while the test dataset is used to figure out the generalized error rate (oob error) of the tree. Each tree's training and test datasets are distinct to avoid bias. In this study, the analysis is based on 1500 runs of Random Forest to extract the statistically significant result. If the original dataset has a data group set aside for test, this data group is used to detect the forest's general fault. The average error rates of individual trees and the forest's total error rate are nearly identical.

2.4. Support Vector Machine Algorithm

Vapnik (Vapnik 2013) suggested the Support Vector Machine (SVM) algorithm as a modern approach to solving pattern recognition problems. The SVM algorithm maps the sample points into high-dimensional feature space in order to find the best separating hyperplane by optimizing the margin between two classes. The SVM algorithm is the supervised machine learning algorithm. New objects can be classified using an SVM classifier that has had adequate training and testing results. The SVM algorithm has been successfully used in a variety of applications for different classification problems. Medical diagnostics, text categorization, information extraction, and other applications have used SVM classifiers based on the SVM algorithm (Demidova *et al*. 2017). Equation 4 can be used to represent the separating hyperplane for the objects from the training set.

$$< w, z > + b = 0 \tag{4}$$

where w is a vector-perpendicular to the separating hyperplane and the shortest distance between the origin and the hyperplane is represented by the parameter b. Also $< w, z >$ is the dot product of w and z. We can more confidently classify objects the wider the strip is. The objects that are nearest to the separating hyperplane are precisely on the strip's bounders (Demidova, Klyueva, Sokolova, Stepanov, and Tyart 2017).

2.5. Model Performance Measurements

For a binary class problem, a confusion matrix is used to evaluate the performance of machine learning methods as shown in Table 1. The columns present predicted classes, and the rows present the actual classes. In the confusion matrix, TP, FP, TN, FN represents true positive samples, false positive samples, true negative samples, and false negative samples respectively.

Table 1. Confusion Matrix.

	Predicted class Positive (Diabetic = 1)	Predicted class Negative (Non-Diabetic = 0)
Actual class Positive (Diabetic = 1)	True Positive (TP)	False Negative (FN)
Actual class Negative (Non-Diabetic = 0)	False Positive (FP)	True Negative (TN)

The confusion matrix can be used to create a variety of evaluation metrics. In this study, widely accepted measurements such as Accuracy, Recall, Precision, and F1 score were used.

Accuracy is determined by the ratio of samples correctly classified by a classifier to the number of all samples.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (5)$$

The ratio of correct positive predictions to the total number of positive examples in the dataset is referred to as Recall.

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

Precision is calculated by dividing the number of samples correctly classified as positive by all samples classified as positive.

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

The harmonic mean of Precision and Recall values is F1 score.

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (8)$$

3. Results and Discussion

The experimental results of the resampling strategy based on Heinz mean are presented in this section. The Pima Indians Diabetes dataset is used for the resampling method. Out of 768 patient samples in the dataset, 500 of them fall into the majority (non-Diabetic = 0) category and 268 into the minority (Diabetic = 1) class. Minority class samples were regenerated synthetically using the following resampling methodology in order to minimize the imbalance rate and provide a balanced dataset. 267 synthetic samples were created from minority class samples, yielding a total of 535 samples. Synthetically generated data is an approximation of the original dataset, so the perfect fit of the data topology is not expected. Age and Glucose attributes are used as a basis to illustrate the balanced dataset class distribution, and raw data and resampling data are represented in a two-dimensional plane in Figure 2 and 3, respectively.

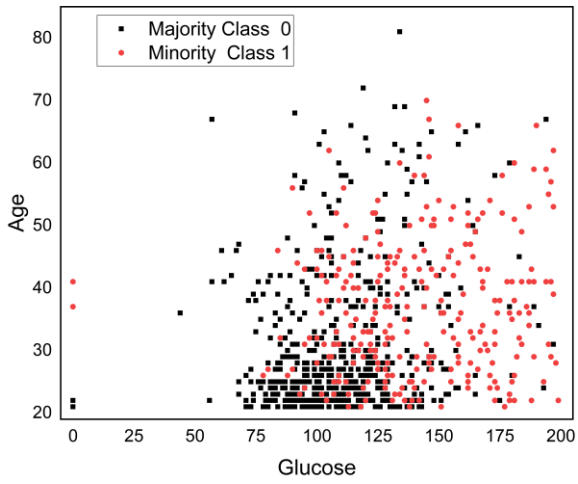

Figure 2. (Color online) The raw dataset is 500 majority and 268 minority classes.

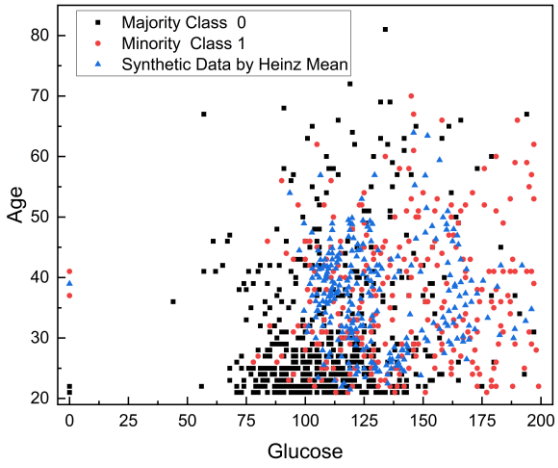

Figure 3. (Color online) Resampled data with the original data. 267 synthetic data is generated by weighted Heinz mean.

In this study, we handled Random Forest (RF) and Support Vector Machine (SVM) as classification algorithms. Firstly, the dataset is balanced by 535 minority class samples and combined with 500 majority classes. The resampled and the raw dataset are classified by RF and SVM. Accuracy (Acc), Precision (P), Recall (R), F1 score values were considered as performance values for the

classification. Classification results of these values for both datasets are shown in Tables 2 and 3 for RF and SVM. The best performance values are shown in bold.

Pima dataset balanced by the proposed method and classified by using RF, an increase in performance values observed, as shown in Table 2. When we measured the overall performance of the proposed model, the Accuracy value increased from 0.751 to 0.799 by the proposed method. Since only the increase of Accuracy performance value is not a criterion in unbalanced data, Precision, Recall, and F1 score values should also be taken as a basis. The minority class (1) outperformed the raw data in all outcomes. Precision increased from 0.685 to 0.800, Recall from 0.539 to 0.816, and F1 score from 0.599 to 0.807 for raw and resampled data. Additionally, the proposed method is compared to another mean based approach, and it is observed slightly better performance metrics (Dal *et al*. 2021). Therefore, the Heinz mean method outperforms the listed methods.

Table 2. Performance Values of Pima Dataset Classification Results with Random Forest Algorithm.

Dataset	Class	Acc	P	R	F1 score
Raw	0	0.751	0.778	**0.865**	0.819
	1		0.685	0.539	0.599
RUS	0	0.737	0.733	0.749	0.738
	1		0.746	0.726	0.733
ROS	0	0.757	**0.891**	0.733	**0.827**
	1		0.505	0.710	0.588
SMOTE	0	0.785	0.789	0.784	0.784
	1		0.784	0.786	0.786
Weigted (Dal, Gümüş, Güldal, and Yavaş 2021)	0 and 1	0.792	0.793	0.792	0.792
Heinz	0	**0.799**	0.800	0.782	0.790
	1		**0.800**	**0.816**	**0.807**

The dataset used in the study was classified with the SVM algorithm and the performance values are shown in Table 3. Considering the performance values of the minority class, there is an increase in the values of the resampled dataset by Heinz mean compared to the raw data. Precision increased from 0.697 to 0.778, Recall from 0.573 to 0.807, and F1 score from 0.625 to 0.792 for raw and resampled data. In addition, the Accuracy value increased from 0.762 to 0.781.

Table 3. Performance Values of Pima Dataset Classification Results with SVM Algorithm.

Dataset	Class	Acc	P	R	F1 score
Raw	0	0.762	0.791	**0.864**	**0.825**
	1		0.697	0.573	0.625
RUS	0	0.738	0.741	0.736	0.737
	1		0.738	0.741	0.738
ROS	0	0.737	**0.895**	0.738	0.808
	1		0.481	0.736	0.579
SMOTE	0	0.764	0.776	0.745	0.759
	1		0.755	0.784	0.768
Heinz	0	**0.781**	0.786	0.754	0.768
	1		**0.778**	**0.807**	**0.792**

According to the classification results shown in Table 2 and Table 3, the minority group of the resampled dataset generally produced more successful results in all metrics. Between all methods, the proposed method provides the best performance. When compared based on classification algorithms, the Random Forest algorithm, which works based on ensemble learning, has been more successful.

4. Conclusions

A synthetic sample generating method with higher performance values has been introduced in this study. The nearest neighbors of the minority group samples were found using the Euclidean distance metric in the proposed method and Heinz Mean was used to generate new synthetic data in the desired number of samples. The balanced dataset is classified using the Random Forest (RF) and Support Vector Machine (SVM). When the raw dataset, resampled by the known methods and resampled by Heinz mean datasets were compared, the Heinz mean resampled dataset outperformed the raw dataset in almost every metric for both classifier algorithms. Also, when compared based on the classifier algorithm, the RF algorithm was more successful than SVM. As a result of the experimental study, it is seen that the data set balanced using the proposed method based on Heinz mean is more successful than the raw dataset and the listed methods.

Acknowledgments

No acknowledgements have been declared by the authors.

Authors' contribution statement

The authors contributed equally to the development, testing, and reporting of the new method.

Ethics committee approval and conflict of interest statement

This study does not require ethics committee permission or any special permission.

5. References

Batista GE, Prati RC, Monard MC, 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, **6** (1), 20-29.

Breiman L, 2001. Random forests. *Machine learning*, **45** (1), 5-32.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP, 2002. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of artificial intelligence research*, **16,** 321-357.

Chawla NV, Japkowicz N, Kotcz A, 2004. Special Issue on Learning from Imbalanced Data Sets. *ACM SIGKDD explorations newsletter*, **6** (1), 1-6.

Dal A, Gümüş İH, Güldal S, Yavaş M, 2021. A New Resampling Approach Based on Weighted Geometric Mean for Unbalanced Data. *Journal of Engineering Science of Adiyaman University*, **8** (15), 343-352. doi:10.54365/adyumbd.940539.

Demidova L, Klyueva I, Sokolova Y, Stepanov N, Tyart N, 2017. Intellectual Approaches to Improvement of the Classification Decisions Quality on the Base of the SVM Classifier. *Procedia Computer Science*, **103,** 222-230.

Elreedy D, Atiya AF, 2019. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. *Information Sciences*, **505**, 32-64.

Fotouhi S, Asadi S, Kattan MW, 2019. A comprehensive data level analysis for cancer diagnosis on imbalanced data. *Journal of biomedical informatics*, **90,** 103089.

Gopinath M, Aarthy S, Manchanda A. 2019. Machine Learning on Medical Dataset. In S. C. Satapathy, V. Bhateja, R. Somanah, X.-S. Yang,R. Senkerik (Eds.), Information Systems Design and Intelligent Applications, Singapore: Springer, 133-143.

Han H, Wang W-Y, Mao B-H. 2005. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Paper presented at the International Conference on Intelligent Computing.

Kovács G, 2019. Smote-Variants: A Python Implementation of 85 Minority Oversampling Techniques. *Neurocomputing*, **366,** 352-354.

Krawczyk B, 2016. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, **5** (4), 221-232.

Liaw A, Wiener M, 2002. Classification and Regression by Random Forest. *R news*, **2** (3), 18-22.

Mohammed AJ, Hassan MM, Kadir DH, 2020. Improving Classification Performance for a Novel Imbalanced Medical Dataset using SMOTE Method. *International Journal*, **9** (3), 3161-3172.

Nguyen HM, Cooper EW, Kamei K, 2011. Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, **3** (1), 4-21.

Rahman MM, Davis DN, 2013. Addressing the Class Imbalance Problem in Medical Datasets. *International Journal of Machine Learning and Computing*, **3** (2), 224-228.

Vapnik V. (2013). The Nature of Statistical Learning Theory (2nd ed.). New York, USA: Springer Science & Business Media. 1-314.

Internet References

1-https://sci2s.ugr.es/keel/dataset.php?cod=21 (05.06.2021)