Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 15(2) 2022: 198-208
http://ejournal.undip.ac.id/index.php/media_statistika

COMPARISON OF SMOTE RANDOM FOREST AND SMOTE K-NEAREST NEIGHBORS CLASSIFICATION ANALYSIS ON IMBALANCED DATA

Jus Prasetya[1], Abdurakhman[2]
[1]Master Program in Mathematics, Gadjah Mada University, Indonesia
[2]Department of Mathematics, Gadjah Mada University, Indonesia

e-mail: *jusprasetya777@gmail.com*

DOI: 10.14710/medstat.15.2.198-208

Article Info:
Received: 19 November 2021
Accepted: 2 January 2023
Available Online: 4 April 2023

Keywords:
Machine Learning; Classification; SMOTE; Random Forest; K-Nearest Neighbors.

Abstract: In machine learning study, classification analysis aims to minimize misclassification and also maximize the results of prediction accuracy. The main characteristic of this classification problem is that there is one class that significantly exceeds the number of samples of other classes. SMOTE minority class data is studied and extrapolated so that it can produce new synthetic samples. Random forest is a classification method consisting of a combination of mutually independent classification trees. K-Nearest Neighbors which is a classification method that labels the new sample based on the nearest neighbors of the new sample. SMOTE generates synthesis data in the minority class, namely class 1 (cervical cancer) to 585 observation respondents (samples) so that the total observation respondents are 1208 samples. SMOTE random forest resulted an accuracy of 96.28%, sensitivity 99.17%, specificity 93.44%, precision 93.70%, and AUC 96.30%. SMOTE K-Nearest Neighborss resulted an accuracy of 87.60%, sensitivity 77.50%, specificity 97.54%, precision 96.88%, and AUC 82.27%. SMOTE random forest produces a perfect classification model, SMOTE K-Nearest neighbors classification produces a good classification model, while the random forest and K-Nearest neighbors classification on imbalanced data results a failed classification model.

1. INTRODUCTION

In machine learning study, classification analysis aims to minimize misclassification and also maximize the results of prediction accuracy. The assumption underlying this classification method is that the observed data set has a class that has a balanced number of objects from the available classes. In other words, it assumes that the probabilities of the target class are equal (balanced). But in reality, in case of medical diagnostics, most of the classification data tend to lead to negative class values. Data means imbalanced if at least one of the target variable classes has a much smaller sample size compared to the other classes (Thabtah et al., 2020).

Classification is an important learning for pattern recognition (characteristics in data). Several classification learning algorithms have been developed such as backpropagation on neural network, bayesian network, decision tree, nearest neighbor, and support vector machine (Goyal et al., 2020). Class imbalance is one of the most influential

factors in the predictive performance of classifiers. Imbalanced data is characterized by having more samples in certain classes than others. In this case, the classifier will tend to create biased learning models that have predictions with poorer accuracy in the minority class compared to the majority class (Zheng & Jin, 2020). Longadge et al (2013) define the problem of imbalance class in terms of skewness. That this problem occurs when the data set highly skewed which can lead to high levels of False Negatives (FN).

Based on Yap et al. (2013) if the percentage of the minority class is less than 5%, then it is called a rare event. When the dataset is imbalanced or a rare event occurs, it can be difficult to obtain a good predictive model due to a lack of information to study the rare event. Ouyang et al (2017) investigated the effect of class imbalance on oil spill detection systems. However, with only 10% of spills originating from the ocean floor, there is far less data for images of oil spills than for images without oil spills. According to Brown & Mues (2012) conducted a study of several classifiers based on five real-world credit scoring data sets in addressing the problem of imbalanced credit scoring. Empirical studies create percentage be divided as 25%, 20%, 15%, 10%, 5%, 2.5%, and 1% bad observations to identify whether any classifier is negatively affected in the predictions.

Several approaches have been developed to deal with the problem of balancing training datasets in supervised learning. These approaches have been categorized into two distinct groups: algorithm level and data level (Goyal et al., 2020). Algorithm level approach overcomes the class distribution problem by modifying the learning stages. The most popular methodology is cost-sensitive learning. Data level approach adjusts the class ratios in the input data set to achieve a balanced class distribution, two main methods are oversampling and undersampling (Hoyos-Osorio et al., 2021).

Previous research using SMOTE-Random Forest (SMO-RF) has been carried out by (Goyal et al., 2020), (Lopez et al., 2013) which discusses resampling techniques (SMOTE), cost-sensitive learning, ensemble methods, (Wei et al., 2022) which uses Improved and Random SMOTE (IR-SMOTE). Previous research combining SMOTE and the Boosting Algorithm (SMOTEBoost) was carried out by (Lee & Kim, 2021). Research by (Zhu et al., 2021) discusses the oversampling method combined with the random forest algorithm, bayes algorithm, and K-Nearest neighbor algorithm. Research (Fernandes et al., 2017) used logistic regression and Support Vector Machine (SVM) to observe cervical cancer screening and also (Fernandes et al., 2018) used supervised deep learning to predict cervical cancer diagnosis. Research by (Jatmiko et al., 2019) discusses the performance of CART, Bagging and Random Forest classification methods on object classification. Random Forest method produces the best performance compared to CART and Bagging. Therefore, this research to compare the performance of the SMOTE random forest and K-Nearest neighbor classification techniques on cervical cancer risk factors datasets based on the output values, namely accuracy, sensitivity, specificity, precision, AUC and accuracy with k-fold cross validation and also knowing the predictor variables that influence the classification of cervical cancer risk factors dataset.

2. LITERATURE REVIEW

2.1. Imbalanced Data

The main characteristic of this classification problem is that there is one class that significantly exceeds the number of samples of other classes. Minority classes generally represent the most important concepts to learn and difficult to identify because they are associated with significant exceptional cases. Most standard learning algorithms consider

balanced training data, this can result in a less than optimal classification model, i.e. good prediction results from the majority sample, while the minority sample is often misclassified. Algorithms that obtain good predictive results within the standard classification framework, do not necessarily achieve the best performance for data sets that have imbalanced classes (Lopez et al., 2013).

2.2. SMOTE (Synthetic Minority Oversampling Technique)

Oversampling is another common sampling approach that used to deal with imbalanced class problems. Oversampling methods available are random oversampling, focused oversampling, and synthetic sampling. SMOTE is a technique in which oversampling on a minority class is carried out by producing synthetic samples. This new synthetic minority class sample is the result of interpolation between adjacent minority class samples (Thabtah et al., 2020).

For example, given data with p variable, namely $\boldsymbol{x}^T = [x_1, x_2, x_3, \dots, x_p]$and $\boldsymbol{z}^T = [z_1, z_2, z_3, \dots, z_p]$ then the euclidean distance $d(\boldsymbol{x}, \boldsymbol{z})$generally as following :

$$\boldsymbol{d}(\mathbf{x}, \mathbf{z}) = \sqrt{(\boldsymbol{x_1} - \boldsymbol{z_1})^2 + (\boldsymbol{x_2} - \boldsymbol{z_2})^2 + (\boldsymbol{x_3} - \boldsymbol{z_3})^2 + \cdots + (\boldsymbol{x_p} - \boldsymbol{z_p})^2} \quad (1)$$

Synthetic data generation is done using the following equation:

$$\mathrm{x}_{syn} = \mathrm{x}_i + (\mathrm{x}_{knn} - \mathrm{x}_i)\gamma \quad (2)$$

with x_{syn} is synthetic data, x_i is *i*-th data of the minor class, x_{knn} is data from the minor class has closest distance from x_i, and $\gamma$is a random number between 0 and 1.

2.3. Decision Tree Based on CART Algorithm

CART algorithm or also known as classification and regression tree is a binary decision tree that deals with classification or regression problems. In Rokach & Maimon (2015) it is stated that the Gini Index is a criterion based on impurity that measures the difference between probability distributions of the target attribute value. The Gini index is the most common data splitting rule many used. The Gini index equation is:

$$i(t) = 1 - \sum_{k=1}^{K} p(k|t)^2 \quad (3)$$

where $i(t)$ is the heterogeneity function of the Gini index, $p(k|t)$ is the class proportion k at node t. After looking for the Gini Index, then the Gini Split value will be searched. Gini Split Equation is:

$$Gini_{split}(t) = \frac{N_1}{N} Gini(t_1) + \frac{N_2}{N} Gini(t_2) + \cdots + \frac{N_n}{N} Gini(t_n) \quad (4)$$

2.4. Random Forest

Random forest is a classification method consisting of a combination of mutually independent classification trees. The classification prediction is obtained through a voting process (the highest number) of the classification trees formed. Random forests are an extension of the ensemble method which was first developed by Breiman (2001) and is used to improve classification accuracy.

2.4.1 Out of Bag (OOB) Error Estimation

Out of Bag Error Estimation serves to estimate error testing the model is bagged, without the need for cross-validation or approximation validation set. The results of the OOB error are obtained based on the calculation of the average error of the predictions for each sample of training data x_i by using trees decisions that do not have x_i in the bootstrap sample. Observations which is not included in the bootstrap sample is called out of bag data (Culter et al., 2012).

2.4.2 Variabel Importance

Mean Decrease Gini (MDG) is a measure of the importance (variable importance) of the explanatory variables generated by the random forest method.

$$\text{MDG} = \frac{1}{n}\sum_{t}[d(h,t)I(h,t)] \tag{5}$$

with n is the number of trees formed, $d(h,t)$ is decrease in the Gini index on the x variable when the node is t, and $I(h,t)$ is 1 at node t, and another 0.

2.5. K-Nearest Neighbors

K-Nearest Neighbors is a non-parametric pattern recognition technique that uses the mean of the closest K observations in the training data to provide estimates. K-Nearest Neighbors can be applied to classification and regression problems. Euclidean distance is a commonly used measure (Becker, 2017).

$$d(x,y) = \sqrt{\sum_{i}(x_i - y_i)^2} \tag{6}$$

2.6. k-Fold Cross Validation

k-Fold Cross Validation method partitions the data set D randomly into k folds (sub sets) that are independent of each other: $f_1, f_2, \dots, f_k$ so that each fold contains $\frac{1}{k}$data section. Furthermore, we can build k datasets: $D_1, D_2, \dots, D_k$each containing (k-1) fold for training data, 1 fold for test data. For example, with $k = 5$, the data set D_1 contains four folds: f_2, f_3, f_4, and f_5 as training data and one fold f_1 for test data. And so on for the D_2, D_3, D_4, and D_5 data sets so that each fold has been a test data once (Suyanto, 2019).

James et al (2014) The process of generating k estimates of test error, MSE_1, MSE_2,...,MSE_k. k-fold cross validation is calculated by

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\text{MSE}_i \tag{7}$$

2.7. Classification Evaluation Model

Evaluation of a classification is generally carried out using a test data set, which is not used in the training of the classification, with a certain size. There are a number of measures that can be used to assess or evaluate the classification model, including: accuracy, error rate, recall/sensitivity/true positive rate, specificity/true negative rate, and precision (Han et al., 2011).

1. Accuracy: Part of an instance that is correctly classified

$$Accuracy = \frac{TP + TN}{P + N} \quad (8)$$

2. Error rate: Part of an instance that is misclassified

$$Error\ rate = \frac{FP + FN}{P + N} \quad (9)$$

3. Sensitifity/Recall: the percentage of positive instance that are correctly classified.

$$Sensitifity = \frac{TP}{P} \quad (10)$$

4. Specificity: the percentage of negative instance that are correctly classified

$$Specificity = \frac{TN}{N} \quad (11)$$

5. Precision: the proportion of the outcomes that are relevant

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

Confusion matrix like a table 1.

Table 1. Confusion Matrix to Evaluate the Classification Model

	Prediction : Yes	Prediction : No	Total
Actual :Yes	TP	FN	P
Actual : No	FP	TN	N
Total	P'	N'	P+N

Lopez et al (2013) stated that the size of the Area Under Curve (AUC) is calculated as the area of the ROC curve using the equation :

$$\text{AUC} = \frac{1 + TP_{rate} - FP_{rate}}{2} \quad (13)$$

Gorunescu (2011) explains the basic criteria for concluding the classification results using AUC:

0.90 - 1.00 = Perfect classification;

0.80 - 0.90 = Good classification;

0.70 - 0.80 = Fair classification;

0.60 - 0.70 = Poor classification;

0.50 - 0.60 = Fail

3. MATERIAL AND METHOD

3.1. Source of Data

Data used is cervical cancer risk factors dataset from UCI Machine Learning Repository Data Sets, which is center for machine learning and intelligent systems. Data has been used in the research of Fernandes et al (2017) which discusses transfer learning with partial observability applied to cervical cancer screening, this work focuses on using classification techniques (e.g. logistic regression and support vector machines) and also in Fernandes et al (2018) which discussed supervised deep learning embeddings for the prediction of cervical cancer diagnosis. in this work present a computationally automated

strategy for predicting the outcome of the patient biopsy, given risk patterns from individual medical records.

3.2. Attributes

In this study, the data consisted of 23 independent variables, namely Age, Number of sexual partners, First sexual intercourse, Number of pregnancies, Smokes, Smokes (packs/year), Hormonal Contraceptives, Hormonal Contraceptives (years), IUD, IUD (years), STDs, STDs (number), STDs: condylomatosis, STDs: vaginal condylomatosis, STDs: vulvo-perineal condylomatosis, STDs: syphilis, STDs: pelvic inflammatory disease, STDs: genital herpes, STDs: molluscum contagiosum, STDs: HIV, STDs: HIV, STDs B, STDs: HPV, STDs: Number of diagnoses and 1 target variable (outcome), namely a biopsy consisting of 2 categories, cervical cancer (1) and not cervical cancer (0)

3.3. Analytical Procedures

To achieve this research, the following steps were carried out:

1. Handling missing value data using the `na.omit()` function in R-Studio Software
2. Using Eq. (2) perform a SMOTE technique to overcome imbalanced data
3. Using Eq. (14), (15) and (16) and default ntree which is 500 trees to get the best mtry based on the smallest out of bag error

$$m = \left|\sqrt{p}\right| \tag{14}$$

$$m = 2\left(\left|\sqrt{p}\right|\right) \tag{15}$$

$$m = \frac{1}{2}\left(\left|\sqrt{p}\right|\right) \tag{16}$$

   with p is independent variable
4. After getting the best mtry, the best ntree will be determined by building 100 to 1000 trees with the `do.trace()` function.
5. Perform random forest classification techniques on balanced and imbalanced data with the best mtry and ntree
6. Perform the K-Nearest neighbors classification technique on balanced and imbalanced data by using k = 4
7. Compare the evaluation value of the classification model consisting of the values of accuracy, sensitivity, specificity, precision, AUC and accuracy with 5-fold cross validation
8. Interpretation of results.

4. RESULTS AND DISCUSSION

4.1. SMOTE Technique Resolves Imbalanced Data

Based on the dataset used, namely cervical data cancer risk factors in attribute variables (independent) there are categorical and numeric type variables, so the distance used uses the following equation:

$$ED = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2 + Median^2} \tag{17}$$

the median value was obtained from the median standard deviation of the numerical variables. The SMOTE technique will be repeated in generating synthesis data until the data in the minority class has been balanced. Based on smooth computing using SMOTEfamily

packages in R-Studio Software, the resulting data that has been balanced (balanced data) with the minority class (class worth 1) becomes 585 observation respondents so that the total observation respondents become 1208 as presented in Table 2. Visually imbalanced and balanced data are explained in the Figure 1.

Table 2. Imbalanced dan Balanced Data

		Imbalanced Data	Balanced Data
Biopsy	0	623	623
	1	45	585
Total		668	1208

Figure 1. Imbalanced Data (Left) and Balanced Data (Right)

4.2. SMOTE Random Forest and Random Forest on Imbalanced Data

Based on running program R-Studio for cervical cancer training data risk factor that has been balanced (balanced data) with the best tunning parameters namely mtry 10 and ntree 800 then the resulting random forest with testing data is

Table 3. Confusion Matrix Data Testing SMOTE Random Forest

Biopsy		Predicted Class		Total
		0	1	
Actual Class	0	119	1	120
	1	8	114	122
Total		127	115	242

On Table 3 it is explained that class 0 (not cervical cancer) correctly predicted as class 0 (no cervical cancer) as much as 119 and 1 others are misclassified. Class 1 (cervical cancer) which was correctly predicted as class 1 (cervical cancer) as many as 114 and 8 others misclassified.

Imbalanced data cervical cancer risk factor training data with the best tunning parameters, namely mtry 2 and ntree 100, the resulting random forest with testing data is

Table 4. Confusion Matrix Data Testing Random Forest

Biopsy		Predicted Class		Total
		0	1	
Actual Class	0	125	0	125
	1	9	0	9
Total		134	0	134

On Table 4 it is explained that class 0 (no cervical cancer) is correctly predicted as class 0 (no cervical cancer) as many as 125 and there is no misclassification. Class 1 (cervical cancer) is not correctly predicted as class 1 (cervical cancer) and 9 is predicted as class 0 (not cervical cancer) so there is a misclassification.

4.3. Variable Importance Imbalanced Data and Balanced Data

Mean Decrease Gini is a measure importance of the independent variables generated by the random forest method. Based on the running program of R-Studio using `importance()` and `varImPlot()` function, the mean decrease gini on imbalanced data and balanced data obtained is as Figure 2.

Based on Figure 2 on the MDG Imbalance Data it is explained that the variables age, hormonal contraceptives (years), first sexual intercourse, number of pregnancies, number of sexual partner and smokes packs (year) are the most influential variables in causing cervical cancer, and while the variables STDs: vaginal condylomatosis, STDs: HPV, STDs: molluscum contagiosum, STDs: pelvic inflammatory disease and STDs: genital herpes do not contribute to causing cervical cancer. On the MDG Balanced Data it is explained that the variables number of pregnancies, age, number of sexual partners, first sexual intercourse, hormonal contraceptives (years) and hormonal contraceptives are the most influential variables in causing cervical cancer, while the variables STDs: vaginal condylomatosis, STDs: pelvic inflammatory disease, STDs: hepatitis B, STDs: HPV, and STDs: molluscum contagiosum do not contribute to causing cervical cancer.

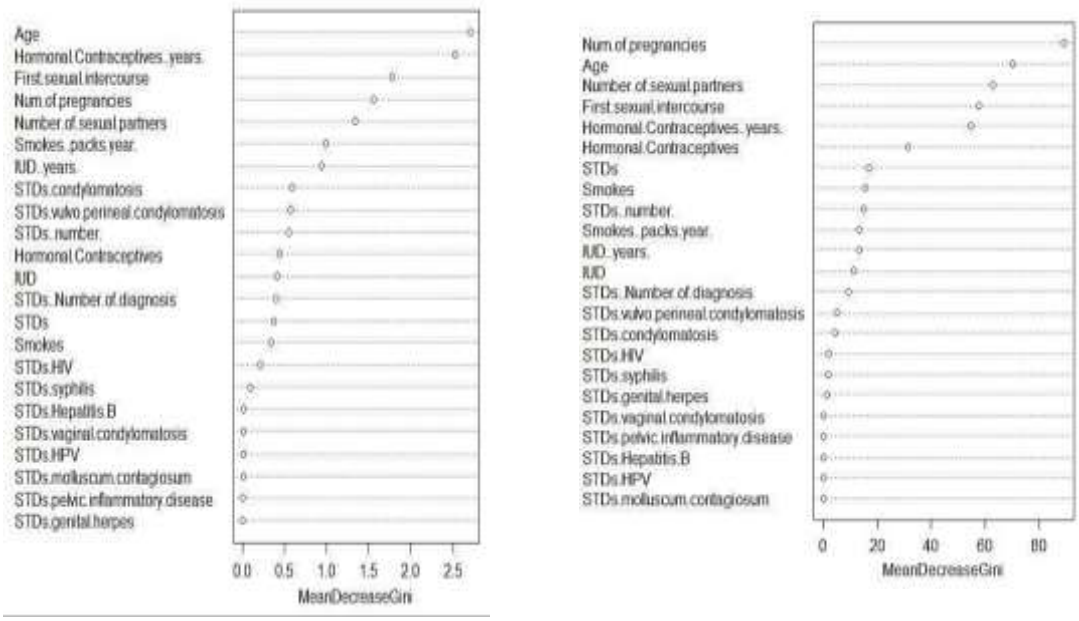

Figure 2. MDG Imbalanced Data (Left) and MDG Balanced Data (Right)

4.4. SMOTE K-Nearest Neighbors and K-Nearest Neighbors on Imbalanced Data

The SMOTE K-Nearest Neighbors model will be formed based on the training data with each sample on the training data measuring the Euclidean distance between samples. After creating a model using the K-Nearest Neighbors SMOTE method, the next step is to evaluate the classification model using the data balanced data testing.

Table 5. Confusion Matrix Data Testing SMOTE K-Nearest Neighbors

Biopsy		Predicted Class		Total
		0	1	
Actual Class	0	93	27	120
	1	3	119	122
Total		96	146	242

On Table 5, it is explained that 93 classes of 0 (not cervical cancer) were correctly predicted as class 0 (no cervical cancer) and the remaining 27 were predicted to be class 1 (cervical cancer). Class 1 (cervical cancer) which was correctly predicted as class 1 (cervical cancer) as many as 119 and 3 predicted as class 0 (not cervical cancer) were misclassified.

In making the K-Nearest Neighbors model on the imbalanced data, the K-Nearest Neighbors model will be formed based on the training data of 534 samples, each of which will measure the euclidean distance. After making a model using the K-Nearest Neighborss method, the next step is to evaluate the classification model using imbalanced data testing.

Table 6. Confusion Matrix Data Testing K-Nearest Neighbors

Biopsy		Predicted Class		Total
		0	1	
Actual Class	0	123	0	123
	1	11	0	11
Total		134	0	134

On Table 6, it is explained that class 0 (no cervical cancer) is correctly predicted as class 0 (no cervical cancer) as many as 123 and there is no misclassification. Class 1 (cervical cancer) is not correctly predicted as class 1 (cervical cancer) and 11 is predicted as class 0 (not cervical cancer) so there is a misclassification.

4.5. Compare Classification Evaluation Model

Based on the results of the classification analysis of the four methods, namely SMOTE random forest, random forest, SMOTE K-Nearest neighbors, and K-Nearest neighbors, the evaluation value of the classification model consists of the values of accuracy, sensitivity, specificity, precision, AUC and accuracy with a 5-fold cross validation will be compared. to conclude whether the classification model made is successful then by using the AUC criteria based on Eq. (13). Table 7 compares the evaluation of the classification model.

Table 7. Compare Classification Evaluation Model

Classification Method	Evaluation Model	Nilai	Conclusion
SMOTE Random Forest	Accuracy	96,28%	Perfect Classification
	Sensitivity	99,17%	
	Specificity	93,44%	
	Precision	93,70%	
	AUC	96,30%	
	Accuracy+5-fold CV	95,65%	
Random Forest	Accuracy	93,28%	Fail Classification
	Sensitivity	100%	
	Specificity	0%	
	Precision	93,28%	
	AUC	50%	
	Accuracy+5-fold CV	93,25%	
SMOTE K-Nearest Neighbors	Accuracy	87,60%	Good Classification
	Sensitivity	77,50%	
	Specificity	97,54%	
	Precision	96,88%	
	AUC	87,27%	

K-Nearest Neighbors	Accuracy+5-fold CV	84,78%	Fail Classification
	Accuracy	91,79%	
	Sensitivity	100%	
	Specificity	0%	
	Precision	91,79%	
	AUC	50%	
	Accuracy+5-fold CV	92,88%	

On Table 7, it can be concluded that the random forest and K-Nearest Neighbors classification technique performed on imbalanced data has an AUC value of 50% so that the conclusion is fails classification. Random forest classification and K-Nearest Neighbors which are carried out on imbalanced data produce a sensitivity value of 0%, which means that the model cannot predict actual data in class 1 (cervical cancer) is precisely predicted as class 1 (cervical cancer) this can cause that patients who should have cervical cancer will be predicted not to have cervical cancer. SMOTE random forest classification technique has an AUC value of 96.30%, which is a perfect classification, while SMOTE K-Nearest Neighbors classification has an AUC value of 87.27% is a good classification.

5. CONCLUSION

Synthetic Minority Oversampling Technique (SMOTE) generates synthesis data in the minority class, namely class 1 (cervical cancer) to 585 observation respondents (samples) so that the total observation respondents are 1208 samples. Based on cervical cancer risk factor dataset used in this study it was concluded that SMOTE random forest classification technique produces a perfect classification model and SMOTE K-Nearest Neighbors classification produces a good classification model, while the random forest and K-Nearest Neighbors classification on imbalanced data results in a failed classification model. If applying predictions using random forest and K-Nearest Neighbors on imbalanced data it will get bad prediction results because patients who should have cervical cancer will be predicted not to have cervical cancer.

REFERENCES

Becker, R. & Thrän, D. (2017). Completion of Wind Turbine Data Sets for Wind Integration Studies Applying Random Forests and K-Nearest Neighbors. *Applied Energy*, *208*(October), 252–262. https://doi.org/10.1016/j.apenergy.2017.10.044

Brown, I. & Mues, C. (2012). An Experimental Comparison of Classification Algorithms for Imbalanced Credit Scoring Data Sets. *Expert Systems with Applications*, *39*(3), 3446–3453. https://doi.org/10.1016/j.eswa.2011.09.033

Fernandes, K., Cardoso, J. S., & Fernandes, J. (2017). Transfer Learning with Partial Observability Applied to Cervical Cancer Screening. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10255 LNCS*, 243–250. https://doi.org/10.1007/978-3-319-58838-4_27

Fernandes, K., Chicco, D., Cardoso, J. S., & Fernandes, J. (2018). Supervised Deep Learning Embeddings for The Prediction of Cervical Cancer Diagnosis. *PeerJ Computer Science*, *2018*(5), 1–20. https://doi.org/10.7717/peerj-cs.154

Gorunescu, F. (2011). *Data Mining : Concepts, Models and Techniques*. Romania: Springer.

Goyal, A., Rathore, L., & Sharma, A. (2021). SMO-RF:A Machine Learning Approach by Random Forest for Predicting Class Imbalancing Followed by SMOTE. *Materials Today: Proceedings*, *xxxx*. https://doi.org/10.1016/j.matpr.2020.12.891

Han, J., Kamber, M., & Pei, J. (2011). *Concepts and Techniques-Chapter 2*.

Hoyos-Osorio, J., Alvarez-Meza, A., Daza-Santacoloma, G., Orozco-Gutierrez, A., & Castellanos-Dominguez, G. (2021). Relevant Information Undersampling to Support Imbalanced Data Classification. *Neurocomputing*, *436*, 136–146. https://doi.org/10.1016/j.neucom.2021.01.033

James, G et al. (2014). *An Introduction to Statistical Learning*. New York: Springer.

Jatmiko, Y. A., Padmadisastra, S., & Chadidjah, A. (2019). Analisis Perbandingan Kinerja CART Konvensional, Bagging dan Random Forest Pada Klasifikasi Objek: Hasil Dari Dua Simulasi. *Media Statistika*, *12*(1), 1-12. https://doi.org/10.14710/medstat.12.1.1-12

Lee, D. & Kim, K. (2021). An Efficient Method to Determine Sample Size in Oversampling Based on Classification Complexity for Imbalanced data. *Expert Systems with Applications*, *184*(May), 115442. https://doi.org/10.1016/j.eswa.2021.115442

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An Insight Into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics. *Information Sciences*, *250*, 113–141. https://doi.org/10.1016/j.ins.2013.07.007

Rokach, L. & Maimon, O. (2015). *Data Mining with Decision Trees*. Edition 2. Singapore : World Scientific Publishing Co.

Team, R. (2020). *RStudio: Integrated Development for R*. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Thabtah, F., Hammoud S., Kamalov, F., & Gonsalves, A. (2020). Data Imbalance in Classification: Experimental Evaluation. *Information Sciences*, *513*, 429–441. https://doi.org/10.1016/j.ins.2019.11.004

Wei, G., Mu, W., Song, Y., & Dou, J. (2022). An Improved and Random Synthetic Minority Oversampling Technique for Imbalanced Data. *Knowledge-Based Systems*, *248*, 108839. https://doi.org/10.1016/j.knosys.2022.108839

Zheng, W. & Jin, M. (2020). The Effects of Class Imbalance and Training Data Size on Classifier Learning: An Empirical Study. *SN Computer Science*, *1*(2), 1–13. https://doi.org/10.1007/s42979-020-0074-0

Zhu, L., Zhou, X., & Zhang, C. (2021). Rapid Identification of High-Quality Marine Shale Gas Reservoirs Based on the Oversampling Method and Random Forest Algorithm. *Artificial Intelligence in Geosciences*, *2*(July), 76–81. https://doi.org/10.1016/j.aiig.2021.12.001