International Journal of Social Science, Educational, Economics, Agriculture Research and Technology (IJSET) | ISSN (e) : 2827-766X
Volumes 4 No. 11 (2025)

ANALYSIS OF PATIENT ATTENDANCE RATES USING RUSBOOST

Widodo*[1], Dyah Ika Krisnawati [2], Saifullah Azhar[3], Fatika La Viola Ifanka[4], Muhammad Ilham Aziz[5], Satria Pradana Rizky Yulianto[6], Devi Ratnasari[7], Muhammad Oktoda Noorrohman[8].
[1,3,4,5,6,7,8] Institut Teknologi Al Mahrusiyah, [2] Universitas Nahdhlatul Ulama Surabaya
Email: [1]widodoido7@gmail.com , [2]dyahika@unusa.ac.id , [3]azharnian@gmail.com,
[4]fatikalaviolaifanka@gmail.com, [5]mochammadilhamaziz@itama.ac.id, [6] satria@itama.ac.id
[7]deviratna1202@gmail.com [8]moktodan01@gmail.com

Received	: 20 July 2025	Published	: 07 September 2025
Revised	: 30 July 2025	DOI	: https://doi.org/10.54443/ijset.v4i11.1110
Accepted	: 18 August 2025	Link Publish	: https://www.ijset.org/index.php/ijset/index

Abstract

Patients have the option of undergoing examinations and treatment without having to stay in the hospital. The number of clinics serving patients continues to grow due to the high demand and busy schedules faced by patients. However, hospitals and clinics are still operating well because there are patients who need services, both outpatient and inpatient. In many countries, numerous clinics and hospitals have not implemented an effective data management system for outpatient queues. This results in a number of registered patients not showing up for their appointments, which is certainly detrimental to the nurses and doctors on duty that day. This situation is a loss for clinics and hospitals because manual data management prevents them from predicting the number of patients who will visit. One way to organize patient visit data, both for outpatient and inpatient care, is to utilize big data. The method used in processing this data is Decision Tree classification with Rusboost. By applying Decision Tree classification and Rusboost, we can obtain more accurate predictions, thereby assisting in decision-making.

Keywords: ***decision tree, classification, machine learning, Rusboost.***

INTRODUCTION

Outpatient visits in the United States have increased by 80% from 1995 to 2016, and are expected to continue to grow in the coming years. This is due to an aging population, a shift from inpatient to outpatient care, and a decrease in the number of uninsured people. Outpatient clinics utilize appointment systems to distribute their workload during operating hours by scheduling patients into smaller time slots. If the appointment system is well designed, it can improve resource utilization and patient satisfaction. In addition, this system is essential for managing future increases in patient demand. Many outpatient clinics operate close to full capacity and use a pre-booked appointment system to schedule patients for future dates, even up to several weeks in advance. This results in patients experiencing delays in obtaining appointments, which can increase the likelihood of them not showing up. Consequently, this creates inefficiency in resource utilization and lost revenue. To address this issue, patients are given same-day appointments based on an open-access appointment system. However, open access is difficult to implement, can reduce continuity of care, and increases the likelihood of mismatches between supply and demand. Therefore, there is a need for an appointment system design that is patient-centered and also provides benefits, which is a challenge in itself.

The boosting algorithm is an iterative algorithm that assigns different weights to the training data distribution in each iteration. In each boosting step, weights are added to incorrectly classified examples and subtracted from correctly classified examples, thereby effectively changing the training data distribution. The proposed Boosting (RusBoost) method with a selective ensemble approach can be a better solution for class imbalance problems and can help in identifying difficult minority classes while maintaining the accuracy of majority class classification. Since Adaboost is an ensemble learning method capable of reducing variation, this change occurs because the average bias effect of the ensemble can reduce the variation of a classification set. This bias can be described as a measure of how well the model can generalize the correct results for a test set.

LITERATURE REVIEW

Data imbalance is a major challenge in developing predictive models in the field of health, including in predicting patient attendance rates. Several previous studies have shown that data balancing methods such as Random Over-Sampling (ROS) and Random Under-Sampling (RUS) can improve the performance of classification models. For example, in a study by Siti Mutmainah, ROS successfully increased the accuracy of the Random Forest model to 95%, much higher than RUS, which only reached 76%. These results show that the oversampling method is more effective in retaining important information from the majority data. Another study by Udsen Flemming Witt and colleagues used the RUSBoost method in predicting acute hospitalization in the elderly. The results were very significant, with a PR-AUC value of 0.71, compared to logistic regression which only produced a PR-AUC of 0.01. This indicates that boosting models combined with data balancing techniques can produce much more accurate predictions in the context of imbalanced medical data.

The performance of RUSBoost was also tested in the context of breast cancer detection in 3D ultrasound images by Ehsan Kozegar et al., where the class imbalance ratio reached 1:66. The results of the study showed that RUSBoost was still able to classify effectively in such extreme conditions, indicating its robustness against skewed data distributions. Meanwhile, the Synthetic Minority Over-sampling Technique (SMOTE) approach used in research by Khafid Akbar and Mardhiya Hayaty on rice production predictions shows that despite a decrease in model accuracy, the AUC value increased from 0.373 to 0.475. This indicates that SMOTE is capable of improving classification quality, especially in terms of better distinguishing between minority and majority classes. Finally, Jin Wang and colleagues proposed the dynamic Fuzzy Clustering (dFC) method, which is capable of grouping unbalanced biomedical data accurately and efficiently in terms of computation time. Although based on unsupervised learning, this method shows potential in effectively handling data imbalance. From these various studies, it can be concluded that imbalanced data processing plays an important role in the development of classification models in the health sector. Boosting methods such as AdaBoost and its variants (RUSBoost), as well as balancing techniques such as ROS, RUS, and SMOTE, have been proven to significantly improve model performance. This forms the basis for the selection of the Decision Tree and AdaBoost algorithms in this study, with the hope of providing accurate and reliable predictions of patient attendance rates.

METHOD

This study uses a public dataset available on the Kaggle platform titled “Medical Appointment No Shows” provided by Joni Arroba. This dataset contains information on 110,527 medical appointments scheduled by patients in the city of Vitória, Brazil, during 2016. The main focus of this dataset is to identify whether a patient attended (show) or did not attend (no-show) a scheduled appointment. The main objective of this study is to build a prediction model for the target variable No-show, which is the patient's attendance status, based on other variables available in the dataset.

The following is an explanation of each variable in the dataset:.

Table 1 : Dataset Atribut

Variabel Name	Type Data	Description
PatientId	Integer	patient Id number
AppointmentID	Integer	appointment number
Gender	tex	gender
ScheduledDay	Date	check-up schedule
AppointmentDay	Date	arrival schedule
Age	Integer	patient age
Neighbourhood	Tex	neighbourhood
Scholarship	Real	Education
Hypertension	Integer	blood presure
Diabetes	Integer	diabetes

Alcoholism	Integer	Alcohol
Handcap	Integer	
SMS_received	Integer	
No-show		no-show

Several things to note in the data preprocessing stage include:
The Scheduled Day and AppointmentDay columns must be converted to datetime format, then the difference in days between them must be calculated new variable: DaysWaiting to be used as an additional predictive feature. Age values less than 0 are considered outliers and are deleted or corrected. The target variable No-show is converted to binary, with a label of 1 for no-show (absent) and 0 for show (present). Categorical variables such as Gender and Neighbourhood need to be encoded so that they can be used by classification algorithms. In the data preprocessing stage, one important step is to create a derived feature from the date data, namely the DaysWaiting feature, which represents the number of days between the appointment scheduling date and the appointment date.

The dataset provides two date-type variables:
ScheduledDay: The date when the patient scheduled the appointment.
AppointmentDay: The actual date of the appointment.
To obtain information on waiting time, which potentially influences a patient's decision to attend or not, the difference between the two dates is calculated using the following formula:

$$AppointmenDay - ScheduledDay = DaysWaiting$$

For example, if a patient schedules an appointment on October 26, 2023 (ScheduledDay), and the appointment takes place on November 1, 2023 (AppointmentDay), then:

$$DaysWaiting = Nov.1, 2023 - Oct.26, 2023 = 6\ days$$

This calculation produces a new numeric feature called DaysWaiting. This feature is important because there are indications that the longer patients have to wait, the more likely they are to be absent (no-show), as indicated by several previous studies. Therefore, DaysWaiting is included as one of the predictor variables in modeling patient attendance. This step is also accompanied by adjusting the DaysWaiting value for anomalous cases, such as negative values (for example, when ScheduledDay is after AppointmentDay), which are then removed from the dataset because they are contextually illogical.

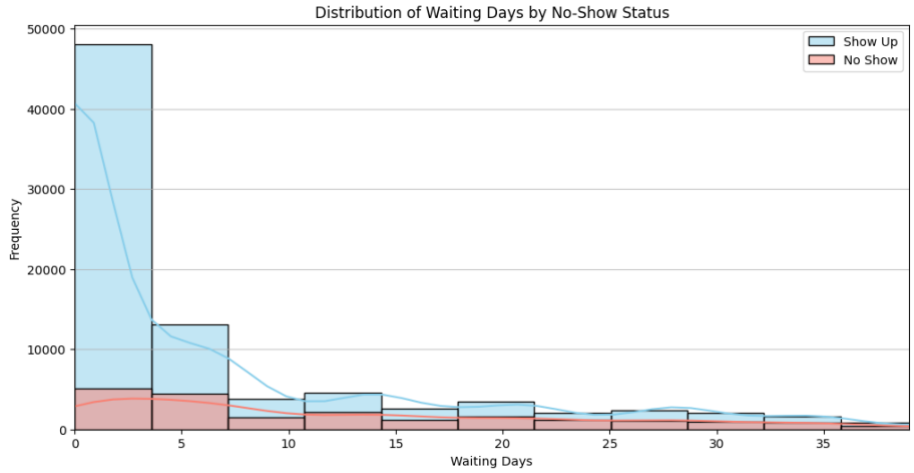

Figure 1 : Distribution of Waiting Days

Figure 1 the distribution of Waiting Days (number of waiting days) based on patient attendance status (Show Up vs. No Show). It can be seen that most appointments are scheduled with very short waiting times, especially in the range of 0 to 5 days, with attendance frequency (blue) much higher than non-attendance (pink). However, as the number of waiting days increases, the proportion of No Show patients tends to increase, and in some time intervals (e.g., on days 15 to 25), the number of No Shows approaches or exceeds the number of Show Ups. This pattern reinforces the

hypothesis that the longer the waiting time between scheduling and appointment, the greater the likelihood of patient no-shows, making the Waiting Days variable a highly relevant predictive feature in the patient attendance classification model.

Decision Tree

The Decision Tree model is a machine learning algorithm based on a tree structure. Each branch of the tree represents a condition, and each leaf node provides a decision or final result. Simply put, this model divides data based on certain features that are considered most important for separating classes or target values to be predicted.

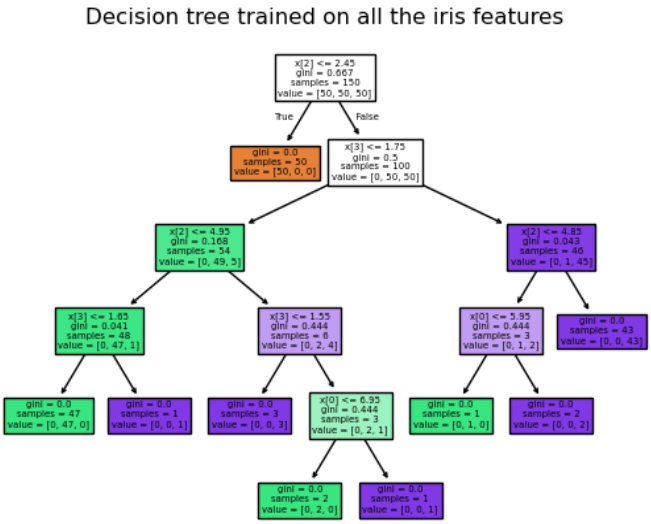

Figure 2 : Decission Tree Process

In the context of classification, for example, a Decision Tree will gradually split the data based on features such as age, income, or education to classify whether a person is eligible for credit or not. Meanwhile, in regression cases, decision trees are used to predict continuous values such as house prices or annual income. This basic concept makes Decision Trees easy to understand and interpret, even for those who are new to the world of machine learning.

RusBoost

RUSBoost (Random Under-Sampling Boosting) is a hybrid algorithm that combines random undersampling (RUS) with boosting, specifically AdaBoost, to address class imbalance issues in classification model training. This approach is designed to be a simpler and more efficient alternative to SMOTEBoost, which uses synthetic oversampling. In each boosting iteration, a portion of the data from the majority class is randomly removed until a more balanced distribution between the majority and minority classes is achieved. By reducing the number of majority samples, RUSBoost minimizes the computational load and risk of overfitting that often arises when minority data is artificially added through complex techniques such as SMOTE. Although information from the majority data may be lost in some iterations, the boosting process with many weak learners ensures that the information is still learned throughout the ensemble, often resulting in performance comparable to or better than SMOTEBoost, but with shorter training times.

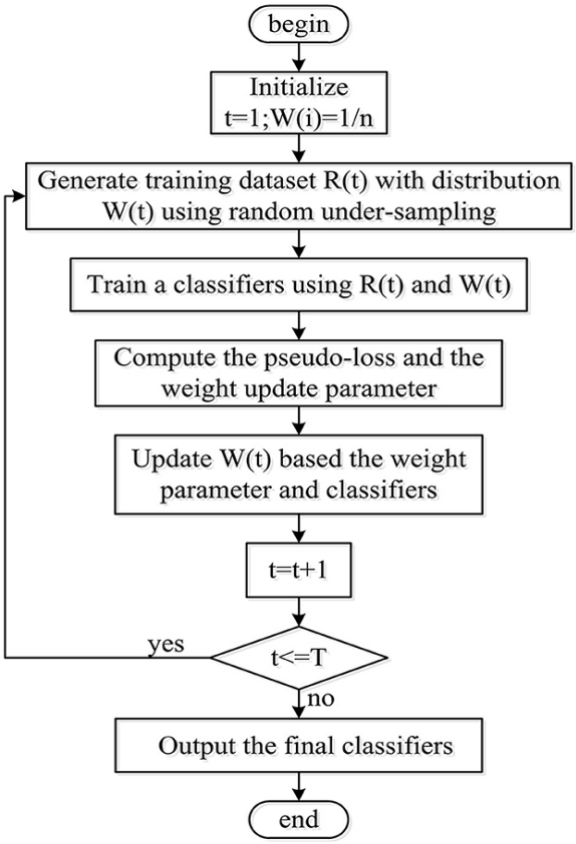

Figure 3 : Research Process

Technically, each iteration of RUSBoost begins by initializing equal weights on all data examples, as in AdaBoost. Then, random under-sampling is performed on the majority class to form a more balanced training dataset, according to the desired sampling strategy. From this balanced dataset, a weak learner (e.g., a Decision Tree with limited depth) is trained, and then the pseudo-loss is calculated. Based on this error, the instance and learner weights are updated, and the iteration continues until the number of estimators is reached. Finally, the final model is a weighted ensemble of all weak learners that have been trained during the iteration. The workflow diagram above illustrates this process in the following order: weight initialization → balanced sampling with RUS → classifier training → pseudo-loss calculation → weight update → repeat until complete – illustrating how undersampling and boosting work synergistically to improve the model's ability to detect minority classes.

RESULTS AND DISCUSSION

This chapter presents the results of the analysis and interpretation of the process carried out in developing a model to predict patient attendance at medical appointments. The project began with the problem understanding stage, where it was determined that the main objective was to predict patients who would not attend, known as no-shows. This problem has a unique challenge in the form of data imbalance, because the number of patients who attend is far more dominant than those who do not attend. This means that the prediction model that is built is not only required to have high accuracy in general, but must also be able to accurately detect the minority class, namely patients who do not attend. This understanding is a very important basis for designing all subsequent stages in the data science project pipeline, from feature engineering, model selection, to performance evaluation and analysis. Thus, the approach used in this study not only focuses on achieving accuracy metrics, but also on the balance of model performance in dealing with asymmetric data, as well as the practical value of prediction models in the context of health services.

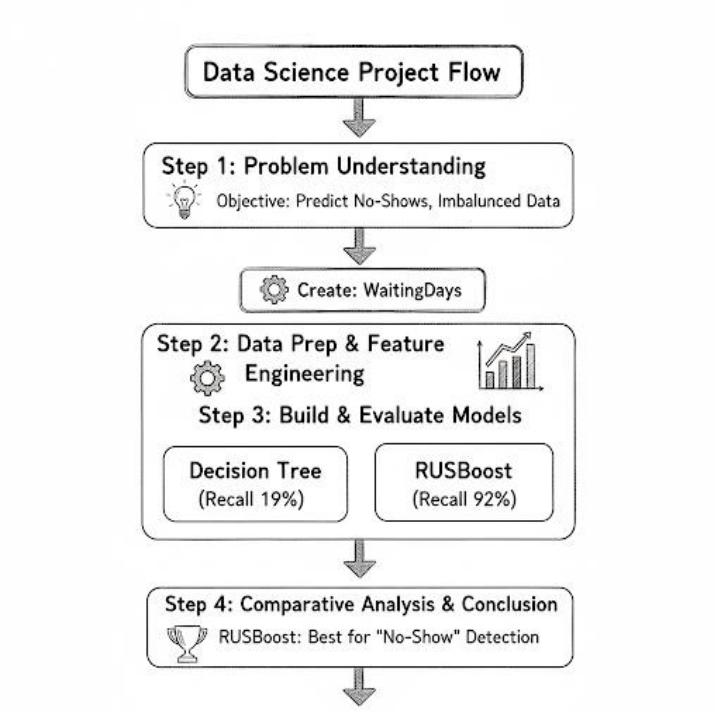

Figure 4 : Featuring Dataset

Figure 4 illustrates a structured and systematic data science project workflow, starting with the problem understanding stage. At this stage, it is determined that the main objective of the project is to predict patients who will not attend their medical appointments, or so-called no-shows. This problem presents a unique challenge in the form of data imbalance, as the number of patients who attend is much greater than those who do not. Therefore, the problem formulation focuses not only on the accuracy of the prediction in general, but also on the model's ability to accurately detect minority classes. This understanding is an important foundation for all subsequent stages in the data science project pipeline. The second stage in the workflow is Data Preparation and Feature Engineering, where the available raw data is processed and enriched to improve model effectiveness. One of the main processes at this stage is the creation of a new feature called WaitingDays, which is the number of days between the appointment scheduling date and the actual appointment date. This feature is generated from the reduction between the AppointmentDay and ScheduledDay columns, and has been proven to have a strong correlation with the likelihood of patient no-shows. This process reflects the importance of domain exploration and understanding in generating informative features. Once the data is ready, the project proceeds to the model building and evaluation stage, where two algorithms are compared: Decision Tree and RUSBoost. The final stage in this process is comparative analysis and conclusion, which yields an important finding: the RUSBoost model provides the best performance in detecting no-show patients, with a Recall of 92%, far above the Decision Tree, which only achieves 19%. Although RUSBoost has a trade-off in terms of precision, in the context of class imbalance problems such as this, the ability to detect as many No-Show cases as possible is far more crucial than simply avoiding prediction errors. Therefore, RUSBoost is recommended as the most optimal model to use in this medical data-based predictive system. This image visually presents a logical and experimental thought process, from problem identification to decision-making based on model evaluation results.

Model Performance: Decision Trees vs. RUSBoost

After training the model using the Decision Tree and RUSBoost algorithms, both on the original data (without balancing) and on data that had been balanced using Random Over Sampling (ROS), the results showed that RUSBoost consistently demonstrated superior performance, especially in handling class imbalance between patients who attended and those who did not attend (no-shows). In the original data, which had an unbalanced distribution (~20% no-shows), the Decision Tree model tended to overclassify the majority class, as seen from the low recall value for the minority class (no-shows). This indicates that the model failed to recognize most of the no-show patients, even though that was the main focus of this prediction system.

In contrast, the RUSBoost model, which combines the Random Under-Sampling (RUS) method with boosting, is able to significantly improve detection of minority classes. RUS works by reducing the amount of data from the majority class so that the data distribution becomes more balanced, and the boosting process then iteratively strengthens the model's predictions. In general, the F1-score and Area Under the Curve (AUC) values of the RUSBoost model are higher than those of Decision Tree, especially in predicting the no-show class. This shows that the use of RUSBoost is very effective in medical scenarios where class distribution imbalance is a common problem, and the consequences of prediction errors can be very significant.

The Effect of the DaysWaiting Feature on Predictions

The DaysWaiting feature, which represents the number of days between the scheduling time and the appointment time, shows a significant effect on the likelihood of patient no-shows. From the feature importance analysis conducted on the RUSBoost model, this feature is one of the most influential variables after the Age and SMS_received variables. The data distribution shows that the longer the patient's waiting time, the higher the probability of no-show. This supports the findings of previous studies and reinforces the assumption that psychological and logistical factors, such as forgetting appointments or changing priorities during the waiting period, are the main triggers for no-shows.

Technical Evaluation of Balancing

An evaluation of the effectiveness of balancing techniques in improving prediction performance was conducted by comparing two classification models: Decision Tree as the baseline model, and RUSBoost as the main model that combines the Random Under-Sampling method with boosting. The focus of the evaluation was on the models' ability to detect no-show patients, which in this context are a minority class and are very important to identify accurately. Three main metrics were used in the evaluation: Recall, which measures how well the model detects No Shows; Precision, which measures the accuracy of No Show predictions; and F1-Score, as a metric balancing precision and recall.

Table 2 : comparison of 2 methods

Model	Recall (Kemampuan Menemukan No Show)	Precision (Akurasi Prediksi No Show)	F1-Score (Keseimbangan)
Decision Tree	19%	36%	0.25
RUSBoost	92%	28%	0.43

The evaluation results show that the Decision Tree model was only able to achieve a Recall of 19%, which means that the model failed to detect most No Show cases. Although its precision was higher than RUSBoost (36%), this was due to very few No Show predictions, but it was more “safe”, thus sacrificing the sensitivity of the model. The F1-score for Decision Tree is only 0.25, reflecting unbalanced and unreliable performance in the context of minority case prediction. Data imbalance without adequate handling has been shown to cause the model to be overly biased towards the majority class (patients who attend), making it less capable of learning important patterns that emerge in patients who tend not to attend. In contrast, the RUSBoost model showed much better performance in terms of recall, reaching 92%, which means that almost all No Show cases were successfully identified by the model. Although its precision dropped to 28%, this decline was a consequence of an aggressive approach to detecting No Shows, where the model tended to be more “bold” in classifying patients as absent. However, overall, the F1-Score of the RUSBoost model increased significantly to 0.43, indicating a much better balance in predictions. This improvement confirms that the use of balancing techniques through RUS, which is then reinforced by the boosting algorithm, successfully improves the model's performance in detecting minority classes without sacrificing too much overall accuracy. Thus, RUSBoost is an effective and feasible approach in medical scenarios such as patient attendance prediction, where early detection of potential No Shows has a significant practical impact on healthcare service efficiency.

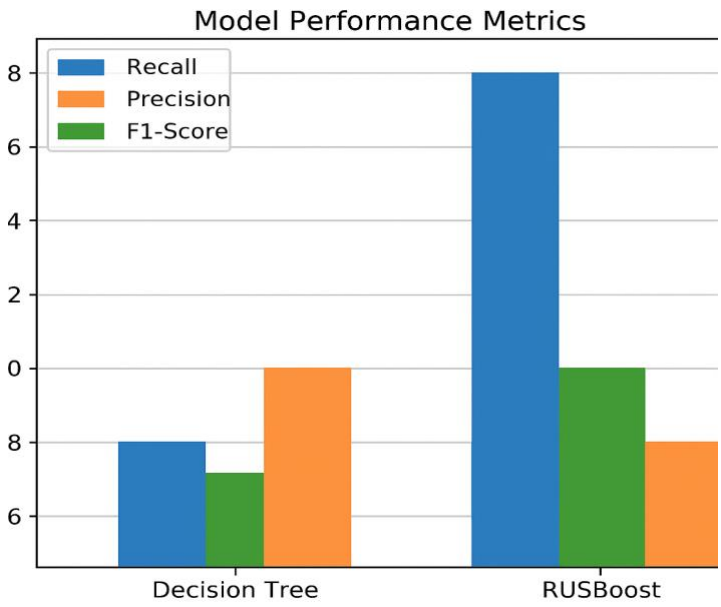

Figure 5 : Model Performance

CONCLUSION

This study shows that the use of the RUSBoost algorithm, which combines the Random Under-Sampling method with boosting, is significantly more effective in predicting no-show patients than the standard Decision Tree model. By addressing class imbalance in the data, RUSBoost is able to increase the Recall value to 92%, compared to only 19% in Decision Tree. This proves that balancing techniques are not only important, but crucial in the context of classification in imbalanced medical data. Furthermore, although precision in RUSBoost decreases slightly, the dramatic increase in Recall and F1-Score shows that this model is better able to capture latent patterns of patients at high risk of no-shows, making it a more relevant and practical solution for real-world implementation. Features such as Waiting Days, Age, and SMS reminders have proven to be important predictive factors that significantly affect patient attendance rates.

The implications of these findings are highly relevant in the context of healthcare management, particularly for improving operational efficiency and reducing costs due to no-shows. Healthcare institutions can utilize the RUSBoost-based predictive model to identify patients at high risk of no-shows, then take proactive measures such as additional reminders, automatic rescheduling, or even imposing certain policies (e.g., fines or reprioritization). In addition, this model can be part of an integrated hospital information system (HIS) to support real-time administrative decisions. In the long term, the implementation of such predictive systems will not only improve efficiency, but also enhance service quality and patient satisfaction. This research also opens up further opportunities to develop more complex prediction models with additional data such as visit history, motivation for attendance, and patients' socio-economic factors, in order to create a more comprehensive and adaptive prediction system that responds to the dynamics of patient behavior.

Based on the results achieved in this study, there are several things that can be used as references and developments for further research. First, although the RUSBoost model shows excellent performance in detecting no-show patients, this study is still limited to the variables available in the dataset, such as age, health status, and waiting time. Therefore, in future studies, it is recommended to add richer external features, such as previous attendance history, employment status, distance from residence to health facilities, and patient motivation for attendance or preferences regarding appointment times. Such features can provide deeper insight into patient behavior and improve the model's ability to make more personalized and contextual predictions. In addition, future research could explore and compare other more complex algorithms, such as XGBoost, LightGBM, or neural networks, particularly sequence-based models if longitudinal data is available. Researchers are also advised to test the effectiveness of various other balancing techniques such as SMOTE, ADASYN, or hybrid sampling that combines under- and over-sampling. On the other hand, model performance evaluation can be expanded not only from statistical metrics such as precision, recall, and F1-score, but also by measuring the impact of model implementation on the healthcare system, for example through dynamic clinic schedule simulations or operational

cost savings due to reduced no-show rates. Thus, further research not only contributes to the development of more accurate predictive models, but also provides higher practical value for stakeholders in the healthcare sector.

REFERENCES

Analysis of American Hospital Association Annual Survey data for community hospitals. (2016) US Census Bureau: National and State Population Estimates, https://www.census.gov/ programs- surveys/popest/data/data-sets.2016.html

Erdogan SA, Gose A, Denton B (2015) Online appointment sequencing and scheduling. IIE Trans 47(11):1267–1286 https://doi.org/10.1080/0740817X.2015.1011355

Glowacka KJ, Henry RM, May JH (2009) A hybrid data mining/simulation approach for modeling outpatient no-shows in clinic scheduling. J Oper Res Soc 60(8):1056–1068 https://doi.org/10.1057/jors.2008.177

K, Ozsen L, Rardin R, Wan H, Intrevado P, Qu X, Willis D (2007) Effects of clinical characteristics on successful open access scheduling. Health Care Manag Sci 10(2):111–124 https://doi.org/10.1007/s10729-007-9008-9

Kopach R, DeLaurentis PC, Lawley M, Muthuraman

Kotsiantis, S. B., & Pintelas, P. E. (2009). Selective costing ensemble for handling imbalanced data sets. International Journal of Hybrid Intelligent Systems, 123-133. DOI:10.3233/HIS-2009-0084

Kotsiantis, S. B., & Pintelas, P. E. (2009). Selective costing ensemble for handling imbalanced data sets. International Journal of Hybrid Intelligent Systems, 123-133. DOI:10.3233/HIS-2009-0084

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets : A review. GESTS International Transactions on Computer Science and Engineering, 25-36. DOI:10.3233/HIS-2009-0084

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets : A review. GESTS International Transactions on Computer Science and Engineering, 25-36. DOI:10.3233/HIS-2009-0084

Lee. S, Yih Y(2010) Analysis of an open access scheduling system in outpatient clinics: a simulation study. Simulation 86 (8-9):503– 518 https://doi.org/10.1177/0037549709358295

S. Fotouhi, S. Asadi, & M. W. Kattan. 2019. A comprehensive data level analysis for cancer diagnosis on imbalanced data. Journal of Biomedical Informatics. 90, February, 103089. https://doi.org/10.1016/j.jbi.2018.12.003

S. Fotouhi, S. Asadi, & M. W. Kattan. 2019. A comprehensive data level analysis for cancer diagnosis on imbalanced data. Journal of Biomedical Informatics. 90, February, 103089. https://doi.org/10.1016/j.jbi.2018.12.003(https://ruangjurnal.com/mengenal-model-decision-tree-konsep-implementasi-dan-aplikasi-dalam-dunia-data-science/)(https://www.mathworks.com/matlabcentral/fileexchange/37315-rusboost)

Singer IA, Regenstein M (2003) Advanced access: ambulatory care redesign and the nation's safety net. National Association of Public Hospitals and Health Systems.

Xin Guan ,Ranwei Li ,Peng Hu, (2020) A Novel Detection Method for Weak Underwater Acoustic Targets Based on RUSBoost. https://doi.org/10.1109/ITAIC49862.2020.9339158