International Journal of Entrepreneurship and Management Practices
IJEMP
EISSN : 2600-8750

Volume 7 Issue 25 (June 2024) PP. 497-506
DOI: 10.35631/IJEMP.725037

INTERNATIONAL JOURNAL OF ENTREPRENEURSHIP AND MANAGEMENT PRACTICES (IJEMP)
www.ijemp.com

GLOBAL ACADEMIC EXCELLENCE

EVALUATION OF CLASSIFICATION ALGORITHMS WITH SOLUTION TO CLASS IMBALANCE PROBLEM ON ZAKAT DISTRIBUTION DATASET

Wan Nurshazelin Wan Shahidan[1], Nor Azriani Mohamad Nor[2*], Azlan Abdul Aziz[3], Nurul Fazira Ismail[4], Nor Hayati Shafii[5]

1 College of Computing, Informatics & Mathematics, Universiti Teknologi MARA Cawangan Perlis, Malaysia. Email: shazelin804@uitm.edu.my
2 College of Computing, Informatics & Mathematics, Universiti Teknologi MARA Cawangan Perlis, Malaysia. Email: norazriani@uitm.edu.my
3 College of Computing, Informatics & Mathematics, Universiti Teknologi MARA Cawangan Perlis, Malaysia. Email: azlan172@uitm.edu.my
4 College of Computing, Informatics & Mathematics, Universiti Teknologi MARA Cawangan Perlis, Malaysia. Email: nurulfazira.ismail01@gmail.com
5 College of Computing, Informatics & Mathematics, Universiti Teknologi MARA Cawangan Perlis, Malaysia. Email: norhayatishafii@uitm.edu.my
* Corresponding Author

Article Info:

Article history:
Received date: 18.04.2024
Revised date: 13.05.2024
Accepted date: 15.06.2024
Published date: 30.06.2024

To cite this document:

Shahidan, W. N. W., Nor, N. A. M., Aziz, A. A., Ismail, N. F., & Shafii, N. H. (2024). Evaluation Of Classification Algorithms With Solution To Class Imbalance Problem On Zakat Distribution Dataset. *International Journal of Entrepreneurship and Management Practices, 7* (25), 497-506.

Abstract:

Class imbalance is one of the most serious and significant issues in data mining classification studies. In recent years, there has been a lot of interest in the problem of imbalance in various real-world applications, such as fraud detection, medical diagnosis, and bank marketing datasets. Similarly, the Zakat distribution UiTM Perlis dataset was utilized to assess the imbalance in the dataset classes. The study aims to evaluate the performance of the resampling technique, Synthetic Minority Over-sampling Technique (SMOTE), and to identify the best classifiers for class-imbalanced Zakat datasets by comparing the classifier's performance. A resampling-based approach is proposed in this study to solve the imbalance dataset. This approach aims to enhance the true positive or the detection of the minority class, which in this case is the not eligible class. The evaluation is based on various metrics, which include accuracy, precision, recall, F-measure, and the ROC area. The algorithms considered in the study include Logistic Regression, Decision Tree (C.4.5), and Random Forest. The results indicate that Random Forest consistently performs well across all evaluation metrics after implementing SMOTE. It attains the highest accuracy, precision, recall, F-measure, and ROC area levels. In conclusion, this research highlights the effectiveness of SMOTE in resolving

International Journal of Entrepreneurship and Management Practices
EISSN : 2600-8750
IJEMP
Volume 7 Issue 25 (June 2024) PP. 497-506
DOI: 10.35631/IJEMP.725037

DOI: 10.35631/IJEMP.725037.

class imbalance in the Zakat distributions dataset. The evaluation of various classification algorithms will be conducted using WEKA.

Keywords:

Accuracy, Class Imbalance, Classifiers, Random Forest, SMOTE

Introduction

Class imbalance is one of the greatest and most important problems in data mining classification studies. Imbalanced data is a term used in machine learning to describe a situation in which a dataset's distribution of classes (categories or labels) is not uniform, implying that one of two classes has a much larger number of samples compared to the others (Yıldırım, 2016). Hence, training a predictive model with an imbalanced dataset tends to produce a strong bias toward the majority classes (Folorunso, 2013). This study conducted a comparative classify algorithm analysis of the sampling method. The zakat distribution dataset from Universiti Teknologi MARA (UiTM) Perlis was utilized to evaluate our class imbalance dataset.

The zakat distribution dataset used in this study shows a significant imbalance of samples between eligible and non-eligible students for zakat assistance. Hence, the predictive model tends to be biased and predicts eligibility more frequently rather than non-eligibility. Therefore, the study aims to investigate the performance of the Synthetic Minority Over-sampling Technique (SMOTE), a resampling-based approach in addressing the class imbalance problem and to identify the best classifiers for class-imbalanced zakat datasets by comparing the classifier's performance. The algorithms/classifiers considered in the study include Logistic Regression, Decision Tree (C.4.5), and Random Forest. The performance evaluation is based on various metrics, which include accuracy, precision, recall, F-measure, and the ROC area. WEKA software, a machine–learning tool written in Java, was used to evaluate various classification algorithms (Witten et al., 2011). It allows us to artificially balance class-imbalanced datasets by applying sampling filters in the pre-processing tab (Verma, 2019).

Literature Review

Class imbalance poses a significant challenge in data mining, as noted by (Kuo et al., 2018), especially in domains requiring individualized care for clusters of varying sizes. The prevalence of uneven data distributions necessitates robust pre-processing methods, with many proposed solutions focusing on adjusting sampling procedures to rectify the underrepresentation of minority classes. (Yıldırım, 2016) further emphasizes the inadequacy of conventional data mining techniques for balanced datasets, advocating for data-level interventions such as resampling and algorithm-level strategies employing sophisticated classification methodologies. However, despite these advancements, there remains a lack of comparative evaluations between imbalanced and balanced datasets, leading to disparities in findings and methodological approaches.

In response to these challenges, researchers have conducted empirical analyses to assess the efficacy of machine learning algorithms in handling class imbalances. For instance, (Kuo et al., 2018) propose integrating cluster analysis with granular computing to tackle imbalanced data classification, highlighting the importance of innovative approaches. Conversely, (Verma, 2019) conducted a WEKA-based evaluation of classification algorithms on a bank marketing dataset, exploring resampling techniques like random under-sampling and SMOTE. Data-level

International Journal of Entrepreneurship and Management Practices
EISSN : 2600-8750
IJEMP
Volume 7 Issue 25 (June 2024) PP. 497-506
DOI: 10.35631/IJEMP.725037
techniques such as random under-sampling, random oversampling, and SMOTE have provided better results and enhanced research quality.

Additionally, a study by (Belavagi & Muniyal, 2016) and (Mourabit et al., 2014) compared various classifiers, concluding that Random Forest exhibited superior performance compared with other classifiers. This model can be visually represented and has the ability to handle data of both numerical and categorical nature (Matzavela & Alepis, 2021). (Schoppa et al., 2020) in their study on the performance of Random Forest for large-scale flood discharge simulation, shows that due to its cost-effectiveness in terms of setup and operation, the random forest method could serve as a viable alternative to physical and conceptual hydrological models for conducting large-scale hazard assessments in diverse catchment areas. Despite the effectiveness of Random Forest, (Belavagi & Muniyal, 2016) highlight the importance of considering multiple classifiers for diagnostic purposes, as evidenced by their comparison of algorithms for predicting cardiovascular disease, where Random Forest achieved 100% accuracy. Furthermore, (Rao & Makkithaya, 2017) in their study used WEKA-based classifiers to identify the best classifiers for class-imbalanced health datasets through a cost-based comparison of classifier performance. The results show that the Bayesian classifiers performed well with a high recall and low number of false negatives and were unaffected by the class imbalance.

Meanwhile, the study (Ramchandra & Rajabhushanam, 2022) has integrated four machine learning techniques: Deep Autoencoder (DAN), Deep Belief Network (DBN), Random Forest (RF), and Long Short-Term Memory (LSTM) to develop a traffic prediction model crucial for traffic system administrators. The findings highlight that LSTM achieves the highest accuracy at 95.2%, outperforming the other models, particularly the Decision Tree, making it the most precise approach for predicting traffic in congested areas. (Xu et al., 2019) proposed a Density-based Synthetic Minority Over-sampling Technique (DSMOTE), integration of Density-based Spatial Clustering of Applications with Noise (DBSCAN) with Synthetic Minority Over-sampling Technique (SMOTE) to overcome the imbalanced dataset problem. Their results show that DSMOTE can achieve better results than SMOTE and Borderline-SMOTE in terms of precision, recall, and F-value.

Methodology

The information was gathered from UiTM Perlis students who applied for zakat during the COVID-19 pandemic. A total of 1949 data was retrieved from Unit Zakat, Sedekah dan Wakaf (ZAWAF) available online in Mendeley Data (Aziz, 2023b)(Aziz, 2023c)(Aziz, 2023a)(Aziz, 2023d). The data was recorded from March 2020 to February 2022. The goal dataset of classification is to predict whether the student is eligible or non-eligible for zakat assistance. The number of instances in the minority class is 68, which are not eligible, and the number of instances in the majority class is 1881, which are eligible. The specific variables used in this study are as follows:

International Journal of Entrepreneurship and Management Practices
IJEMP
EISSN : 2600-8750
Volume 7 Issue 25 (June 2024) PP. 497-506
DOI: 10.35631/IJEMP.725037

Table 1: Variable Selection of Zakat

Variable	Description
Gender	Categorical (male, female)
Semester	Categorical (1,2,3,4,5,6,7)
CGPA	Cumulative Grade Point Average (numeric)
Head family's income	Numeric
Mother's income	Numeric
Household size	Number of households (numeric)
Zakat receiver	Have the students ever received zakat assistance before (Categorical:" yes", "no")
Zakat eligibility (Y)	Has the students eligible for zakat (binary: "yes", "no")

An imbalanced dataset can be provided for classification. By imbalance dataset, we mean that one of the two classes has a very small number of samples compared to several samples in the other, C2 << C1. C2 is referred to as the minority class, while C1 is referred to as the majority class. Our interest lies in the minority class. The machine learning algorithm always performs well when given a balanced dataset, but this is not always true. As an example, the dataset for zakat distribution in UiTM Perlis will likely have fewer people who are not eligible than those who are eligible. The prediction of the unbalanced dataset is biased towards the majority class. The approach to solve this problem is by using SMOTE. SMOTE is a resampling approach, particularly an oversampling technique.

Synthetic Minority Over-Sampling Technique (SMOTE)

The problem of overfitting is reduced by using a method of creating synthetic instances of the minority class. The name given to this method is the synthetic minority over-sampling technique (SMOTE). In this case, the training set is altered by adding synthetically generated minority class instances, which leads to a more balanced class distribution. The instances are said to be synthetic because they are new instances of the minority class created from existing instances of the minority class.

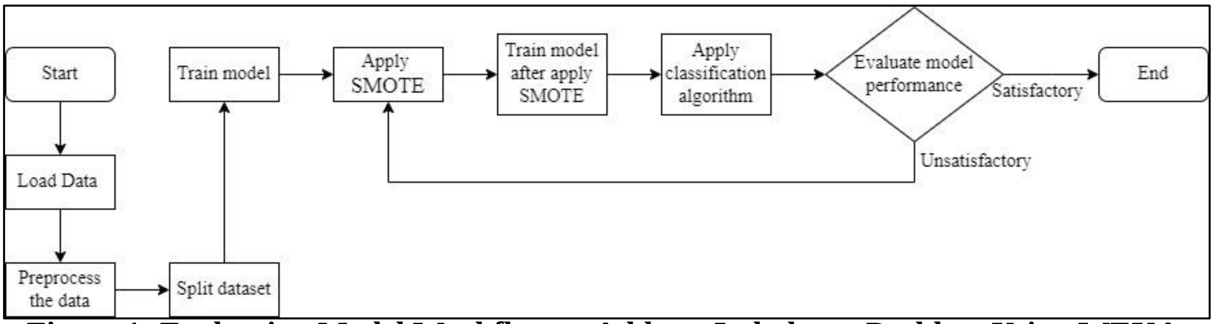

Figure 1: Evaluation Model Workflow to Address Imbalance Problem Using WEKA

WEKA classifiers work well in the class imbalance dataset (Witten et al., 2011). For the Synthetic Minority Over-Sampling Technique, we use WEKA classifiers, which resample a dataset by applying the Synthetic Minority Over-sampling Technique (SMOTE). Figure 1 presents the model workflow to address the imbalance problem using WEKA. The zakat distribution dataset, containing 1949 instances with a significant imbalance (68 minority

International Journal of Entrepreneurship and Management Practices
EISSN : 2600-8750
IJEMP
Volume 7 Issue 25 (June 2024) PP. 497-506
DOI: 10.35631/IJEMP.725037
instances and 1881 majority instances), is imported into the WEKA system. The user may specify the amount of SMOTE and the number of nearest neighbors. The ClassValue in this zakat distribution dataset is set to 0 to detect the non-empty minority class automatically. The dataset is divided into a 70% training set and a 30% testing set to facilitate model training and evaluate performance. The SMOTE filter is applied three times, each time with different parameters. In the first iteration, a percentage of 30 is utilized, and the number of instances in the minority class is increasing from 68 to 88. In the second iteration, SMOTE is once more applied with a percentage of 50, increasing the number of instances of minority classes from 88 to 132. In the third iteration, SMOTE is applied at a rate of 80 percent, resulting in the minority class instances increasing from 132 to 237. Classification algorithms, including Logistic Regression, Decision Trees, and Random Forest, are then applied. The accuracy, precision, recall, F-measure, and ROC metrics for the minority class are recorded and represented visually. The best-performing algorithm is then selected based on comprehensive performance metrics, ensuring improved minority class detection and overall model efficacy.

Results

Original Data Evaluation

Table 2: Original Data Evaluation

	Instances of Majority class (Eligible)	Instances of Minority class (Not Eligible)
Without Handling Class imbalance problem (No Filter Applied)	1881	68
Imbalance Ratio	96.5%	3.5%

Table 2 displays the number of instances for the majority and minority classes in the original dataset and the imbalance ratio. There are a total of 1949 instances in the dataset. The original dataset contains 1881 instances with a 96.5% majority class imbalance ratio. This refers to the class that is more prevalent or has a higher number of occurrences in the dataset, which is the eligibility of students to receive zakat. The original dataset contains 68 instances, representing a 3.5% imbalance ratio of the minority class which represents students who are not eligible to receive zakat assistance. This refers to the class that is less prevalent or has a lower number of occurrences in the dataset. The imbalance ratio measures the imbalance in the class distribution within the dataset. In this case, the imbalance ratio is represented in Table 3., which shows 96.5% for the majority class and 3.5% for the minority class. Approximately 96.5% of the total instances in the dataset are accounted for by the majority class. The statement indicates a significant over-representation of the majority class compared to the minority class. The minority class accounts for only 3.5% of the total instances in the dataset. The statement indicates a notable lack of representation of the minority class. The result highlights the class distribution imbalance, in which the majority class dominates the dataset, while the minority class is severely underrepresented. Algorithms may tend to favor the majority class and perform poorly in correctly predicting instances of the minority class. It is imperative to address this class imbalance to provide a fair and accurate classification outcome. To improve the model's ability to effectively learn from both classes and increase the representation of the minority class, techniques like oversampling the minority class using algorithms like SMOTE (Synthetic Minority Over-sampling Technique) can be used. Another technique is to detect the minority class accurately in the confusion matrices.

International Journal of Entrepreneurship and Management Practices
EISSN : 2600-8750
IJEMP
Volume 7 Issue 25 (June 2024) PP. 497-506
DOI: 10.35631/IJEMP.725037

Confusion Matrix Of Each Algorithm For Original Data

The provided confusion matrices represent the performance of various classification algorithms, including Logistic Regression, Decision Tree, and Random Forest. These algorithms were evaluated on a binary classification problem with two classes: “Not Eligible” (true positive) and “Eligible” (true negative). The confusion matrices below focus on the “Not Eligible” class not being detected well.

Table 3: Confusion Matrix of Logistic Regression, Decision Tree and Random Forest

	Logistic Regression		Decision Tree		Random Forest	
	Observed		Observed		Observed	
Predicted	Not eligibility	Eligibility	Not eligibility	Eligibility	Not eligibility	Eligibility
Not eligibility	0	26	0	26	0	26
Eligibility	1	558	2	557	3	556

The common issue observed in Table 3 is the detection of the “Not Eligible” class being poor. This is indicated by nonzero values in the False Negative cells, which represent cases where Eligible is incorrectly predicted as Not Eligible. This means that some instances, which should have been classified as “Not Eligible,” are being misclassified as “Eligible”. The fact that the True Positive values are relatively low compared to the False Negative values suggests that the models perform poorly in identifying the “Not Eligible” class. Concerning, however, is the misclassification of “Not Eligible” instances. These models also exhibit a small number of accurate positive predictions and a greater number of inaccurate negative predictions. The low sensitivity indicated by the low true positive values suggests that these models do not accurately identify positive cases.

Actual Data Evaluation For Each Algorithm

To acquire a more thorough knowledge of the algorithms’ performance on the dataset, it is required to analyze and evaluate their performance further using other evaluation metrics such as accuracy, precision, recall, F-measure, and ROC area.

Table 4: Evaluation of The Original Dataset

Algorithm	Accuracy	Precision	Recall	F-Measure	Roc Area
Logistic Regression	95.3846	0	0	0	0.695
Decision Tree	95.2137	0	0	0	0.568
Random Forest	95.0427	0	0	0	0.738

According to Table 4., all the algorithms exhibit relatively high accuracies ranging from approximately 95.0% to 95.4%. Regrettably, all algorithms' precision, recall, and F-measure values are 0. The fact that the algorithm did not correctly identify any positive instances in the dataset indicates a lack of true positive detections. Furthermore, if the precision, recall, and F-measure all equal 0, the algorithms did not make any true positive predictions, which means the student did not receive the amount of zakat. This implies that the algorithms have failed to classify positive instances, resulting in false negatives correctly. False negatives occur when the algorithms mistakenly label actual positive instances as negative. The non-zero ROC area

International Journal of Entrepreneurship and Management Practices
IJEMP
EISSN : 2600-8750
Volume 7 Issue 25 (June 2024) PP. 497-506
DOI: 10.35631/IJEMP.725037

values, which are 0.695, 0.568, and 0.738, suggest that the algorithms still exhibited some discrimination between the positive and negative classes. However, they were unable to identify any true positives accurately.

Result After Applied SMOTE Filter

Table 5: SMOTE Technique Evaluation (Eligibility After Applied Smote Technique)

	Instances of Majority class (Eligible)	Instances of Minority class (Not Eligible)
Without Handling Class imbalance problem (No Filter Applied)	1881	237
Imbalance Ratio	89%	11%

Table 5 displays information regarding the distribution after applying the SMOTE technique. The updated minority class consists of 237, while the minority class comprises 68. After applying SMOTE, the instances of the majority class and the instances of the minority class are presented as percentages. This implies that the majority class accounts for 89% of the dataset, whereas the minority class comprises 11%. The purpose of applying SMOTE with Random Over Sampling is to address class imbalance by generating synthetic instances for the minority class. In this case, SMOTE has increased the number of instances in the minority class based on the provided data. This can help improve the representation of the minority class in the dataset and potentially mitigate the bias towards the majority class.

Confusion Matrix Of Each Algorithm (SMOTE)

Table 6: Confusion Matrix of Logistic Regression, Decision Tree, and Random Forest

	Logistic Regression		**Decision Tree**		**Random Forest**	
	Observed		Observed		Observed	
Predicted	Not eligibility	Eligibility	Not eligibility	Eligibility	Not eligibility	Eligibility
Not eligibility	22	49	40	31	48	23
Eligibility	11	553	16	548	8	556

After applying SMOTE (Synthetic Minority Over-sampling Technique) to address the class imbalance in the dataset, the confusion matrices in Table 6 have shown improvement in detecting the “Not eligible” class, as indicated by the increase in true positive (TP) values for this class. In the case of logistic regression, the number of true positives for “Not eligible” rises from 22 to 49, showing that the model can now correctly identify more instances of “Not eligible.” Similarly, the actual positive count for “Not eligible” climbed from 40 to 48 in the decision tree and from 48 to 56 in the random forest. These increased positive values show that after being trained with SMOTE-augmented data, the models have grown more adaptable at recognizing the “Not eligible” class. The effectiveness of SMOTE in enhancing the detection of the “Not eligible” class can be attributed to its capacity to generate synthetic samples by interpolating between existing instances of minority classes. This balances the class distribution, allowing the models to demonstrate enhanced sensitivity or recall for the “Not eligible” class, resulting in higher accurate positive rates. It is essential to acknowledge that

International Journal of Entrepreneurship and Management Practices
EISSN : 2600-8750
IJEMP
Volume 7 Issue 25 (June 2024) PP. 497-506
DOI: 10.35631/IJEMP.725037

utilizing SMOTE enhances the identification of the "Not eligible" category. However, a thorough assessment of model efficacy should encompass additional metrics, including precision. The evaluation of the overall effectiveness of each algorithm after the SMOTE transformation involves the utilization of accuracy, precision, recall, F-measure, and ROC area metrics. However, the observed rise in true positive values suggests that implementing SMOTE has effectively improved the model's capacity to identify instances belonging to the "Not eligible" class.

Algorithm Evaluation Metrics

Table 7: Evaluation for Each Algorithm After Applying the Smote Technique.

Algorithm	Accuracy	Precision	Recall	F-Measure	Roc Area
Logistic Regression	90.5512	0.667	0.31	0.423	0.834
Decision Tree	95.5984	0.714	0.563	0.63	0.855
Random Forest	95.1181	0.857	0.676	0.756	0.951

Table 7 displays the evaluation metrics for the Logistic Regression, Decision Tree, and Random Forest classification algorithms. Random Forest obtained the highest accuracy with 95.1181%, followed by Decision Tree with 92.5984% and Logistic Regression with 90.5512%. This indicates that Random Forest performed the best overall regarding accurate predictions. Random Forest again surpassed the other algorithms, with a precision rating of 0.857. The precision of the Decision Tree was 0.714, whereas the precision of the Logistic Regression was 0.667. Higher accuracy suggests a reduced percentage of false positives, implying that Random Forest performed better in categorizing positive instances. Upon examining the recall values, it is evident that Random Forest attained the highest recall value of 0.676, followed by Decision Tree with a recall value of 0.563, and Logistic Regression with the lowest recall value of 0.310. Recall represents the model's ability to determine positive instances correctly. A higher recall shows that the Random Forest was carried out better to detect the positive class. Next, when considering the F-measure, Random Forest achieved the highest value of 0.756. It was followed by Decision Tree with a value of 0.630, and Logistic Regression had the lowest F-measure of 0.423. The F-measure is a metric that combines precision and recall into a single metric, and a higher F-measure indicates a more balanced relationship between the two. Lastly, Random Forest obtained the maximum ROC area value of 0.951, indicating superior discrimination between positive and negative classes. The ROC area for Decision Tree was 0.855, while the ROC area for Logistic Regression was 0.834. In conclusion, Random Forest consistently demonstrated the highest performance across various evaluation criteria, including accuracy, precision, recall, F-measure, and ROC area. It obtained the highest accuracy, precision, recall, and F-measure values among the three algorithms. Moreover, the highest ROC area indicated that Random Forest exhibited superior discrimination between the positive and negative classes.

Conclusion And Recommendation

This study addresses imbalanced datasets using the zakat distribution dataset and concludes that employing the Synthetic Minority Over-Sampling Technique (SMOTE) significantly improved the models' ability to identify the minority class and effectively resolved the institution's challenges. Various evaluation metrics, including accuracy, precision, recall, F-measure, and ROC area, were utilized to compare the performance of different algorithms when SMOTE was applied. The findings indicate that the Random Forest algorithm is the most

International Journal of Entrepreneurship and Management Practices
EISSN : 2600-8750
IJEMP
Volume 7 Issue 25 (June 2024) PP. 497-506
DOI: 10.35631/IJEMP.725037
suitable for managing class imbalance in this Zakat dataset. It demonstrates superior classification performance and effective handling of imbalanced classes.

However, it's essential to acknowledge that the selection of the optimal algorithm may also depend on the specific characteristics and requirements of the dataset. Further research and experimentation may be necessary to determine the most appropriate algorithm for different contexts. Therefore, while Random Forest proves to be effective in this study, it's essential to consider other factors that may influence algorithm selection.

Several recommendations and suggestions for future research are proposed. Firstly, while SMOTE was applied in this study, it's worth exploring other oversampling techniques, such as Borderline-SMOTE and SMOTE-ENN, to compare their performance and impact on classification results. Further investigation into the optimal algorithm selection process, considering various datasets and scenarios, would contribute to a deeper understanding of class imbalance management in zakat distribution. Continuous exploration and experimentation are vital for refining and improving the effectiveness of imbalance dataset handling techniques in zakat distribution and similar domains to achieve more accurate classification outcomes.

Acknowledgements
The authors would like to express their deepest gratitude to Universiti Teknologi MARA Cawangan Perlis for their support in facilitating the research process.

References
Aziz, A. A. (2023a). Profile Permohonan dan Agihan Zakat April 2021. Mendeley. https://doi.org/10.17632/RMP9TT27NP.1
Aziz, A. A. (2023b). Profile Permohonan dan Agihan Zakat Mac 2020. Mendeley. https://doi.org/10.17632/HCCZ9YZ8XH.1
Aziz, A. A. (2023c). Profile Permohonan dan Agihan Zakat Oktober 2020. Mendeley. https://doi.org/10.17632/5WFT868ZBG.1
Aziz, A. A. (2023d). Profile Permohonan dan Agihan Zakat Oktober 2021. Mendeley. https://doi.org/10.17632/SD3RJB2B6Z.1
Belavagi, M. C., & Muniyal, B. (2016). Performance evaluation of supervised machine learning algorithms for intrusion detection. *Procedia Computer Science*, 89, 117–123. https://doi.org/10.1016/j.procs.2016.06.016
Folorunso, S. O. (2013). Alleviating classification problem of imbalanced dataset. *International Journal of Engineering and Computer Science*, 6(2), 137–144.
Kuo, R. J., Su, P. Y., Zulvia, F. E., & Lin, C. C. (2018). Integrating cluster analysis with granular computing for imbalanced data classification problem: A case study on prostate cancer prognosis. *Computers & Industrial Engineering*, 125(August), 319–332. https://doi.org/10.1016/j.cie.2018.08.031
Matzavela, V., & Alepis, E. (2021). Decision tree learning through a predictive model for student academic performance in intelligent M-learning environments. *Computers & Education: Artificial Intelligence*, 2, 100018.
Mourabit, Y. E. L., Toumanari, A., Bouirden, A., Zougagh, H., & Latif, R. (2014). Intrusion detection system in wireless sensor network based on mobile agent. 2014 *Second World Conference on Complex Systems (WCCS), 248–251*. . https://doi.org/10.1109/ICoCS.2014.7060910
Ramchandra, N. R., & Rajabhushanam, C. (2022). Machine learning algorithms performance evaluation in traffic flow prediction. *Materials Today: Proceedings*, 51, 1046–1050.

International Journal of Entrepreneurship and Management Practices
EISSN : 2600-8750
IJEMP
Volume 7 Issue 25 (June 2024) PP. 497-506
DOI: 10.35631/IJEMP.725037
Rao, R. R., & Makkithaya, K. (2017). Learning from a class imbalanced public health dataset: A cost-based comparison of classifier performance. *International Journal of Electrical and Computer Engineering*, 7(4), 2215–2222. https://doi.org/10.11591/ijece.v7i4.pp2215-2222

Schoppa, L., Disse, M., & Bachmair, S. (2020). Evaluating the performance of random forest for large-scale flood discharge simulation. *Journal of Hydrology*, 590(September), 125349.

Verma, A. (2019). Evaluation of classification algorithms with solutions to class imbalance problem on bank marketing dataset using WEKA. *International Journal of Computer Applications*, 178(25), 14–19.

Witten, I. H., Frank, E., & Hall, M. A. (2011). Data mining: Practical machine learning tools and techniques (3rd ed.). *Morgan Kaufmann*. www.mkp.com

Yıldırım, P. (2016). Pattern classification with imbalanced and multiclass data for the prediction of albendazole adverse event outcomes. *Procedia Computer Science*, 83(Dmdms), 1013–1018. https://doi.org/10.1016/j.procs.2016.04.216