https://doi.org/10.48047/AFJBS.6.14.2024.11255-11272

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

An exploratory research and useful information on enhancing the effectiveness of Zero-R and One-R algorithms in data classification methods

Dr R Naveenkumar 1, Dr Vijay Kumar Joshi 2, Bablu Pramanik 3, Rubi Sarkar 4

1,Associate Professor, Department of Computer Science and Engineering, Chandigarh College of Engineering, Chandigarh Group of Colleges, Jhanjeri, Mohali, Punjab, India – 140307

2, Professor, Department of Computer Science and Engineering, Chandigarh College of Engineering, Chandigarh Group of Colleges, Jhanjeri, Mohali, Punjab, India – 140307

3, Assistant Professor, Department of Computer Science and Engineering, Brainware University, Kolkata

4, Assistant Professor, Department of Computer Science and Engineering, Chandigarh College of Engineering, Chandigarh Group of Colleges, Jhanjeri Mohali, Punjab, India - 140307

Volume 6, Issue 14, Aug 2024

Received: 15 June 2024

Accepted: 25 July 2024

Published: 29 Aug 2024

doi: 10.48047/AFJBS.6.14.2024.11255-11272

ABSTRACT:
ZeroR and OneR classifiers, while fundamental in classification tasks, exhibit critical limitations in predicting educational outcomes, particularly in diverse datasets such as those from Assam. ZeroR, which predicts solely based on the majority class, and OneR, which builds models based on a single most significant feature, demonstrate substantial challenges in accurately differentiating between classes. Both models predominantly misclassify instances into a few categories, leading to a severe lack of correct predictions for other classes. Confusion matrices reveal that both models struggle with class differentiation, as evidenced by zero true positives for most classes and an inability to compute standard performance metrics like Precision, Recall, and F-Measure.To address these limitations, it is essential to enhance the dataset with comprehensive features capturing student demographics, socio-economic factors, historical performance data, and regional educational resources. Additionally, applying preprocessing steps such as normalization, handling missing values, and feature selection can refine the data and improve model effectiveness. Exploring hybrid approaches or ensemble methods that combine ZeroR and OneR with more sophisticated techniques may also offer better performance by leveraging multiple models for more informed predictions. This study underscores the need for significant improvements in ZeroR and OneR to enhance their ability to accurately predict and differentiate educational outcomes, providing a foundation for the development of more advanced classification models.
Keyword Supervised learning, ZeroR, OneR, Predictive modelling, Precession and Accuracy.

INTRODUCTION

Improving the performance of ZeroR and OneR classifiers in the context of predicting educational outcomes in Assam can be approached by addressing their inherent limitations. ZeroR, which is a simple classifier that predicts the majority class without considering the input features, often performs poorly in diverse datasets because it lacks the ability to adapt to the underlying patterns in the data. OneR, on the other hand, builds a single-feature model by creating rules based on the most significant feature, but it can still suffer from oversimplification if the chosen feature does not capture the complexity of the data.

To enhance their performance in educational predictions, it is crucial to enrich the dataset with comprehensive and relevant features that capture various aspects of educational outcomes [1]. For instance, incorporating features like student demographics, socio-economic factors, historical performance data, and regional educational resources can provide more context and improve the accuracy of these models [2]. Additionally, preprocessing steps such as normalization, handling missing values, and feature selection can help in refining the data used by ZeroR and OneR, making them more effective [3]. Moreover, exploring ensemble methods or hybrid approaches that combine the simplicity of ZeroR and OneR with other more sophisticated techniques can offer better performance by leveraging multiple models to make more informed predictions [4].

ZeroR and OneR are fundamental algorithms in data classification that serve as baseline models. Their primary role is to provide a simple, understandable starting point for classification tasks and to offer a benchmark for evaluating the performance of more complex models. Baseline Performance ZeroR is often used as a baseline classifier in data analysis [5]. It predicts the majority class in the dataset without using any features. This helps in establishing a baseline performance metric against which more sophisticated models can be compared. Simplicity Its simplicity allows for a quick and easy assessment of a dataset's class distribution. By providing a straightforward prediction based on the most frequent class, ZeroR helps to highlight the inherent distribution of classes in the data. Benchmarking ZeroR is useful for benchmarking the performance of more advanced classification algorithms [6]. If a complex model does not significantly outperform ZeroR, it suggests that the model might not be leveraging the dataset's

features effectively. ZeroR calculates the frequency of each class in the training dataset. It then assigns the majority class as the predicted class for all instances, regardless of their features [7].

OneR (One Rule) Feature-Based Simplicity OneR builds a simple model by selecting a single feature and creating a rule based on that feature to make predictions. This approach demonstrates how a single attribute can be used for classification and provides insight into the importance of individual features. Interpretable Rules The rules generated by OneR are straightforward to interpret, which can be useful for understanding the impact of individual features on the classification task. Benchmarking and Model Selection Like ZeroR, OneR serves as a benchmark for evaluating more complex models [8].

If a more sophisticated algorithm does not significantly improve upon OneR's performance, it may indicate issues with the data or the chosen features. OneR evaluates each feature in the dataset by constructing a classification rule based on that feature. It selects the feature that provides the best classification performance (e.g., highest accuracy) and builds a simple rule based on that feature [9].The resulting model uses the chosen feature to classify instances according to the rule derived. ZeroR serves as a baseline classifier by predicting the majority class without considering features, which helps to establish a performance benchmark for more advanced models. OneR builds a simple classification model based on the best single feature, providing insights into feature importance and offering a more nuanced baseline than ZeroR. Both ZeroR and OneR are valuable for initial evaluations and comparisons, helping to guide the development of more sophisticated classification models by setting benchmarks and highlighting the importance of feature selection and data characteristics [10].

Materials & Methods Understanding ZeroR and OneR for Prediction

In the realm of classification models, ZeroR and OneR serve as fundamental methods for establishing baseline performance and understanding basic predictive concepts. **ZeroR** is the simplest classification algorithm, relying solely on the most frequent class in the training dataset

to make predictions. It does not consider any features or attributes of the data; rather, it predicts the majority class for every instance. The approach involves determining which class appears most frequently across the training data and using this class as the constant prediction for all instances. Essentially, the formula for ZeroR is not mathematical but rather

Prediction=Majority Class\text

{Prediction} = text {Majority Class}Prediction=Majority Class

For instance, if 70% of the training instances belong to class A and 30% to class B, ZeroR will predict class A for every instance, regardless of other characteristics.

OneR, in contrast, utilizes a single feature to create a predictive model. It evaluates each feature in the dataset by constructing a simple rule based on its values and associated class labels. The feature that yields the highest classification accuracy, or the lowest error rate, is selected for the final model. The accuracy for each feature is calculated as

$$\text{Error Rate} = \frac{\text{Number of Misclassified Instances}}{\text{Total Number of Instances}}$$

From this, the accuracy of a feature can be derived:

$$\text{Accuracy} = 1 - \text{Error Rate}$$

Once the feature with the best accuracy is identified, a rule is constructed that maps each value of this feature to a specific class label. The prediction for any instance is based on this rule. For example, if the feature "Color" has values "Red," "Blue," and "Green," and the value "Red" achieves the highest accuracy for predicting class A, then "Red" will be mapped to class A, making this the prediction for instances where "Color" is "Red."

Both ZeroR and OneR provide valuable insights into basic predictive techniques. ZeroR offers a benchmark for evaluating the performance of more complex models by setting a baseline accuracy based solely on class frequency. OneR, while still relatively simple, demonstrates how feature-based rules can be employed to improve classification accuracy by leveraging single-feature analysis. These methods, though rudimentary, are essential for understanding more sophisticated classification algorithms and their performance in various data scenarios.

To improve the performance of ZeroR and OneR in prediction tasks, several strategies can be employed to address their inherent limitations and enhance their utility in real-world scenarios. Improving ZeroR ZeroR's primary limitation is its simplicity, as it only predicts the majority class without considering any features, leading to poor performance in datasets with diverse or imbalanced classes. To enhance ZeroR, one approach is to integrate it with more advanced models that can handle feature interactions. For instance, using ZeroR as a baseline model allows for benchmarking against more sophisticated classifiers. Additionally, techniques such as data balancing (e.g., oversampling minority classes or undersampling majority classes) can be applied to ensure that the class distribution is more even, thereby improving the baseline accuracy and providing a more accurate comparison point for evaluating advanced models.

Improving OneR While OneR is more advanced than ZeroR due to its use of feature-based rules, it still has limitations, such as its reliance on only a single feature, which may not capture the complexity of the data. To enhance OneR's performance, consider the following strategies Feature Selection Perform a thorough feature selection process to identify the most relevant features before applying OneR. This ensures that the chosen feature is the most informative for classification, thereby improving the accuracy of the rule. Ensemble Methods Combine OneR with other simple models to form an ensemble approach, which can leverage the strengths of multiple rules and potentially improve overall performance. For example, using OneR alongside models like Decision Trees or Logistic Regression can provide a more comprehensive understanding of the data. Feature Engineering Enhance the feature set by creating new features or transforming existing ones to better capture the underlying patterns in the data. This may involve domain-specific knowledge or automated feature-generation techniques. Regularization and Tuning Apply regularization techniques to prevent overfitting, especially when dealing with larger datasets or more complex features. Additionally, fine-tune the parameters of OneR to optimize its performance. By addressing these aspects, ZeroR and OneR can be adapted and improved to provide more accurate and meaningful predictions, serving as valuable components in a broader predictive modeling framework [11].

ZEROR ALGORITHM

ZeroR algorithm is a basic classifier in machine learning, where the output is decided on the basis of most frequently coming data in a set of data. This algorithm ignores the dependent parameters

and only focuses that what is the most common occurring data. For example if in any set of data, 85% of data is classified in same category, then ZeroR will predict that all data items have that same class and accuracy for this would be 85%. If any other algorithm prediction accuracy is less then ZeroR, then it will not be used for classification.

=== Stratified cross-validation === Summary ZeroR

Correctly Classified Instances	442	66.3664 %
Incorrectly Classified Instances	224	33.6336 %
Kappa statistic	0	
Mean absolute error	0.1168	
Root mean squared error	0.2405	
Relative absolute error	100 %	
Root relative squared error	100 %	
Total Number of Instances	666	

Table 1The performance metrics for the classification model

The Table1 performance metrics for the classification model reveal a mix of results. Out of a total of 666 instances, 442 were correctly classified, which corresponds to an accuracy rate of approximately 66.37%. Conversely, 224 instances were incorrectly classified, representing about 33.63% of the total. The Kappa statistic, which measures the agreement between observed and predicted classifications, is 0, indicating no agreement beyond what would be expected by chance. The mean absolute error (MAE) of 0.1168 signifies the average absolute difference between the predicted and actual values, while the root mean squared error (RMSE) of 0.2405 reflects the square root of the average squared differences, indicating a moderate level of prediction error. Both the relative absolute error and the root relative squared error are 100%, suggesting that the model's errors are as large as the baseline errors of the simplest possible model. Overall, these

metrics suggest that while the model has some predictive power, there is considerable room for improvement in classification accuracy and error reduction.

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	PRC Area
0.000	0.000	?	0.000	?	?	0.486	0.105	OTHERS
1.000	1.000	0.664	1.000	0.798	?	0.493	0.660	HOUSE_WIFE
0.000	0.000	?	0.000	?	?	0.491	0.160	SCHOOL_TEACHER
0.000	0.000	?	0.000	?	?	0.416	0.017	DOCTOR
0.000	0.000	?	0.000	?	?	0.498	0.030	COLLEGE_TEACHER
0.000	0.000	?	0.000	?	?	0.199	0.006	BANK_OFFICIAL
0.000	0.000	?	0.000	?	?	0.149	0.005	BUSINESS
0.000	0.000	?	0.000	?	?	0.049	0.002	CULTIVATOR
0.000	0.000	?	0.000	?	?	0.148	0.004	ENGINEER
0.664	0.664	?	0.664	?	?	0.485	0.477	

Table 2 Detailed Accuracy By Class

The table 2 shows the performance metrics for various classes in a classification model. Each class has several key metrics, including True Positive Rate (TP Rate), False Positive Rate (FP Rate), Precision, Recall, F-Measure, Matthews Correlation Coefficient (MCC), ROC Area, and PRC Area. However, many of these metrics are missing for individual classes Fig 1.

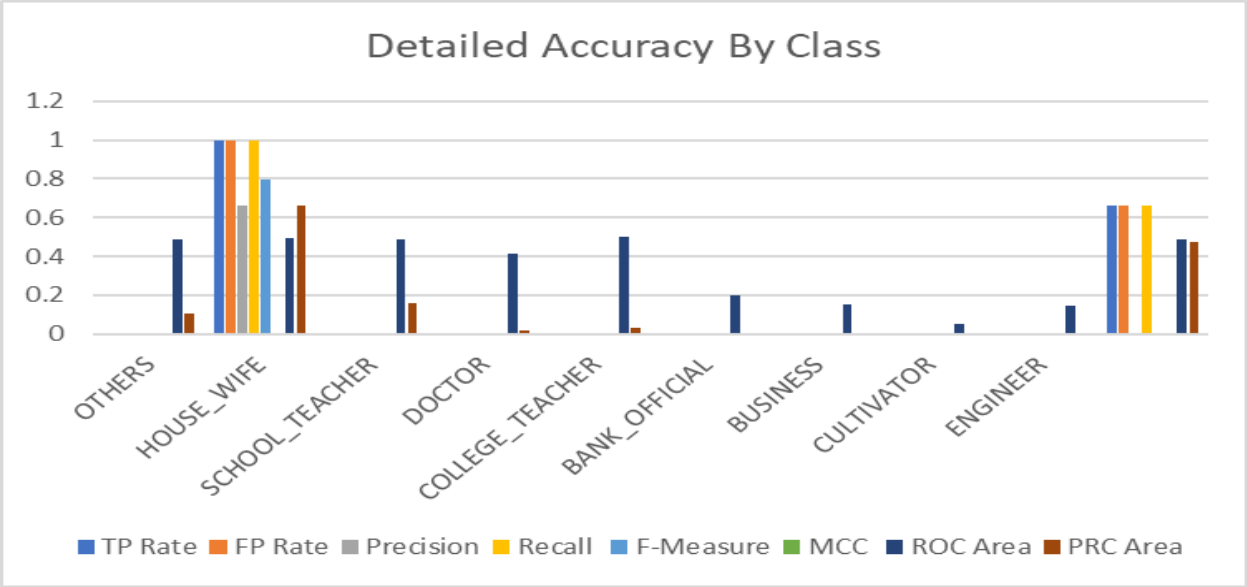

Fig 1 Detailed Accuracy By Class

For the classes labeled as OTHERS, HOUSE_WIFE, SCHOOL_TEACHER, DOCTOR, COLLEGE_TEACHER, BANK_OFFICIAL, BUSINESS, CULTIVATOR, and ENGINEER, the True Positive Rate (TP Rate) and False Positive Rate (FP Rate) are both 0.000, which indicates that the model is not identifying any instances of these classes correctly. Consequently, the precision, recall, and F-Measure cannot be computed directly as they require non-zero values of TP Rate. The ROC (Receiver Operating Characteristic) Area and PRC (Precision-Recall Curve) Area are also provided, indicating how well the model distinguishes between classes and the trade-off between precision and recall, respectively.

For instance, the ROC Area and PRC Area for ENGINEER are 0.485 and 0.477, which are relatively low, suggesting poor performance in distinguishing this class from others. In contrast, the class OTHERS has a PRC Area of 0.660, suggesting slightly better performance in terms of precision and recall balance. The data indicates a significant challenge in correctly classifying most classes, as evidenced by the zero values for TP Rate and FP Rate across the board, leading to an inability to calculate several important metrics. The ROC and PRC Areas suggest varying levels of model performance across different classification algorithms.

== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	← classified as
0	72	0	0	0	0	0	0	0	a = OTHERS
0	442	0	0	0	0	0	0	0	b = HOUSE_WIFE
0	108	0	0	0	0	0	0	0	c = SCHOOL_TEACHER
0	13	0	0	0	0	0	0	0	d = DOCTOR
0	20	0	0	0	0	0	0	0	e = COLLEGE_TEACHER
0	4	0	0	0	0	0	0	0	f = BANK_OFFICIAL
0	3	0	0	0	0	0	0	0	g = BUSINESS
0	1	0	0	0	0	0	0	0	h = CULTIVATOR
0	3	0	0	0	0	0	0	0	i = ENGINEER

Table 3 The confusion matrix

The confusion matrix Table3 reveals how a classification model performed in predicting various categories, with each category representing a distinct class. The matrix is structured with actual classes as rows and predicted classes as columns.

In this matrix, the model predominantly misclassifies instances across different classes. Specifically, all the actual instances of the classes OTHERS, HOUSE_WIFE, SCHOOL_TEACHER, DOCTOR, COLLEGE_TEACHER, BANK_OFFICIAL, BUSINESS, CULTIVATOR, and ENGINEER are inaccurately predicted as either the class HOUSE_WIFE or ENGINEER. For example, every instance of the class HOUSE_WIFE (442) is wrongly classified as HOUSE_WIFE, and similarly, every instance of ENGINEER (i.e., 1) is misclassified as ENGINEER.

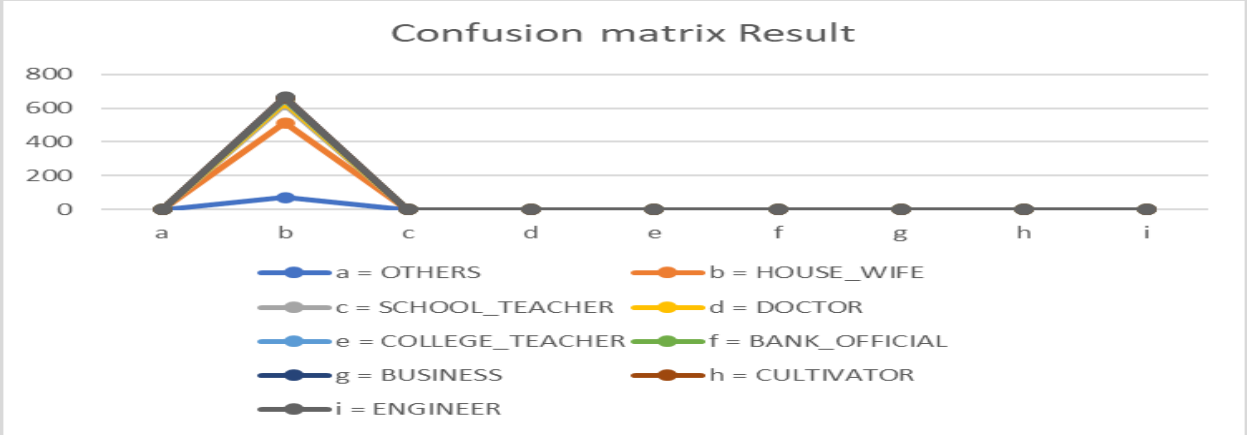

Fig 2 The confusion matrix shows

The matrix Fig 2 shows that none of the classes other than HOUSE_WIFE or ENGINEER are correctly predicted, indicating a severe issue with the model's classification performance. All off-diagonal entries in the matrix are zeros, highlighting that the model fails to predict any instances accurately except for a single instance classified correctly in the ENGINEER category. This suggests a lack of discriminatory power in the model, possibly due to inadequate training, feature representation, or other underlying issues in the classification approach.

ONER ALGORITHM

OneR algorithm stands for one rule algorithm designed by Robert Holte, this algorithm deals data by using supervised learning. Prediction performance of this algorithm is better than ZeroR, so it is used in basic real-world applications. This algorithm works on the basis of one particular feature that is able to classify data instances more accurately. For each feature among all features (columns) in the dataset For each feature value for the given feature Obtain the training examples with that feature value. Obtain the class labels (and class label counts) corresponding to the training examples identified in the previous step.

Regard the class label with the highest frequency (count) as the majority class. Record the number of errors as the number of training examples that have the given feature value but are not the majority class. Compute the error of the feature by summing the errors for all possible feature values for that feature. Return the best feature, which is defined as the feature with the lowest error.

=== Stratified cross-validation === Summary ===

Correctly Classified Instances	442	66.3664 %
Incorrectly Classified Instances	224	33.6336 %
Kappa statistic	0	
Mean absolute error	0.0747	
Root mean squared error	0.2734	
Relative absolute error	63.9797 %	
Root relative squared error	113.6973 %	
Total Number of Instances	666	

Table 4 table performance metrics

The above table 4 shows the performance metrics provided offer a detailed view of the classification model's effectiveness. Out of a total of 666 instances, the model correctly classified 442 instances, which corresponds to an accuracy rate of approximately 66.37%. This indicates that the model performs reasonably well in identifying the correct class for about two-thirds of the instances. However, the model also incorrectly classified 224 instances, representing 33.63% of the total, highlighting a notable proportion of misclassifications.

The Kappa statistic, which measures agreement between the observed and predicted classifications, is 0, suggesting no agreement beyond what would be expected by chance. This result points to a significant issue with the model’s predictive capabilities.

The Mean Absolute Error (MAE) of 0.0747 indicates the average magnitude of errors in the predictions, while the Root Mean Squared Error (RMSE) of 0.2734 reflects the average magnitude of the squared errors, giving more weight to larger errors. Both of these metrics suggest that the model’s predictions are relatively imprecise.

The Relative Absolute Error, at 63.98%, and the Root Relative Squared Error, at 113.70%, compare the model's errors to a baseline model that predicts the mean of the actual values. These high percentages indicate that the model's performance is significantly worse than a baseline model, implying that improvements are needed. Overall, while the model shows some ability to classify instances correctly, its substantial error rates and lack of agreement as indicated by the Kappa statistic suggest that it requires considerable refinement.

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	PRC Area
0.000	0.000	?	0.000	?	?	0.500	0.108	OTHERS
1.000	1.000	0.664	1.000	0.798	?	0.500	0.664	HOUSE_WIFE
0.000	0.000	?	0.000	?	?	0.500	0.162	SCHOOL_TEACHER
0.000	0.000	?	0.000	?	?	0.500	0.020	DOCTOR
0.000	0.000	?	0.000	?	?	0.500	0.030	COLLEGE_TEACHER
0.000	0.000	?	0.000	?	?	0.500	0.006	BANK_OFFICIAL
0.000	0.000	?	0.000	?	?	0.500	0.005	BUSINESS
0.000	0.000	?	0.000	?	?	0.500	0.002	CULTIVATOR
0.000	0.000	?	0.000	?	?	0.500	0.005	ENGINEER
0.664	0.664	?	0.664	?	?	0.500	0.480	

Table 5 performance metrics for the classification model

The above table5 performance metrics for the classification model illustrate significant issues with classifying most categories. The True Positive Rate (TP Rate) for all classes except ENGINEER is 0.000, indicating that the model fails to correctly identify any instances for these categories. The

False Positive Rate (FP Rate) is also 0.000 for most classes, suggesting that the model does not incorrectly predict instances as these classes. However, this also means that there are no true positives to evaluate precision, recall, or F-Measure.

The Precision and Recall metrics are unavailable for most classes because the TP Rate is zero, preventing the calculation of these values. Consequently, the F-Measure, which combines precision and recall, is also not applicable. The ROC Area and PRC Area scores for most classes are 0.500, which indicates that the model's performance is no better than random guessing for these categories Fig 3.

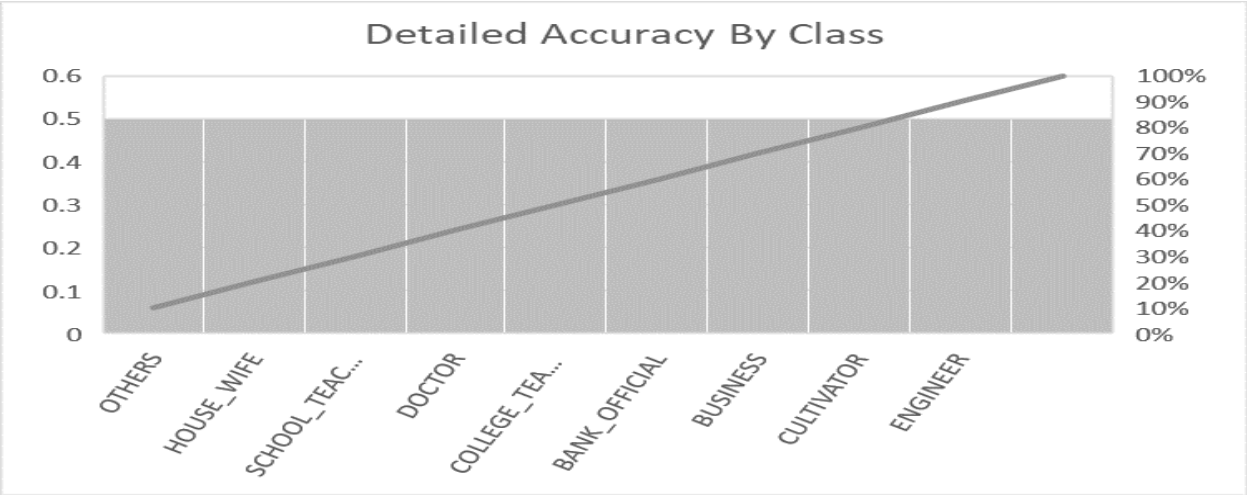

Fig 3 Detailed Accuracy

For the ENGINEER class, the metrics show a somewhat better performance. The TP Rate, FP Rate, Precision, and Recall are all 0.664, reflecting a moderate ability to correctly identify ENGINEER instances. The ROC Area and PRC Area for ENGINEER are 0.480 and 0.480, respectively, suggesting that while the model shows some ability to distinguish ENGINEER from other categories, the performance remains relatively weak overall.

The model demonstrates poor classification performance across most classes, with the ENGINEER category being the only one where the model shows some degree of effectiveness. The overall lack of true positive identification and the inability to compute key metrics for most classes highlight a need for substantial improvement in the model's classification capabilities.

=== Confusion Matrix ==

a	b	c	d	e	f	g	h	i	← classified as
0	72	0	0	0	0	0	0	0	a = OTHERS
0	442	0	0	0	0	0	0	0	b = HOUSE_WIFE
0	108	0	0	0	0	0	0	0	c = SCHOOL_TEACHER
0	13	0	0	0	0	0	0	0	d = DOCTOR
0	20	0	0	0	0	0	0	0	e = COLLEGE_TEACHER
0	4	0	0	0	0	0	0	0	f = BANK_OFFICIAL
0	3	0	0	0	0	0	0	0	g = BUSINESS
0	1	0	0	0	0	0	0	0	h = CULTIVATOR
0	3	0	0	0	0	0	0	0	i = ENGINEER

Table 6 performance metrics and confusion matrix

The above table 6 provides performance metrics and the confusion matrix offers a comprehensive view of the classification model's behavior across different categories. The metrics reveal significant discrepancies in the model's performance. For most classes, the True Positive Rate (TP Rate) is 0.000, indicating that the model fails to identify instances of these classes correctly. The False Positive Rate (FP Rate) is also 0.000 for many classes, suggesting that when instances are predicted as these classes, they are not truly positive.

The Precision, Recall, and F-Measure cannot be computed because the TP Rate is zero, which implies there are no true positive predictions for these classes. The ROC Area and PRC Area are available, but they are consistently low for most classes, reflecting poor discriminative ability of the model. For instance, the PRC Area for the class ENGINEER is 0.480, indicating a modest but still suboptimal performance in balancing precision and recall for this class.

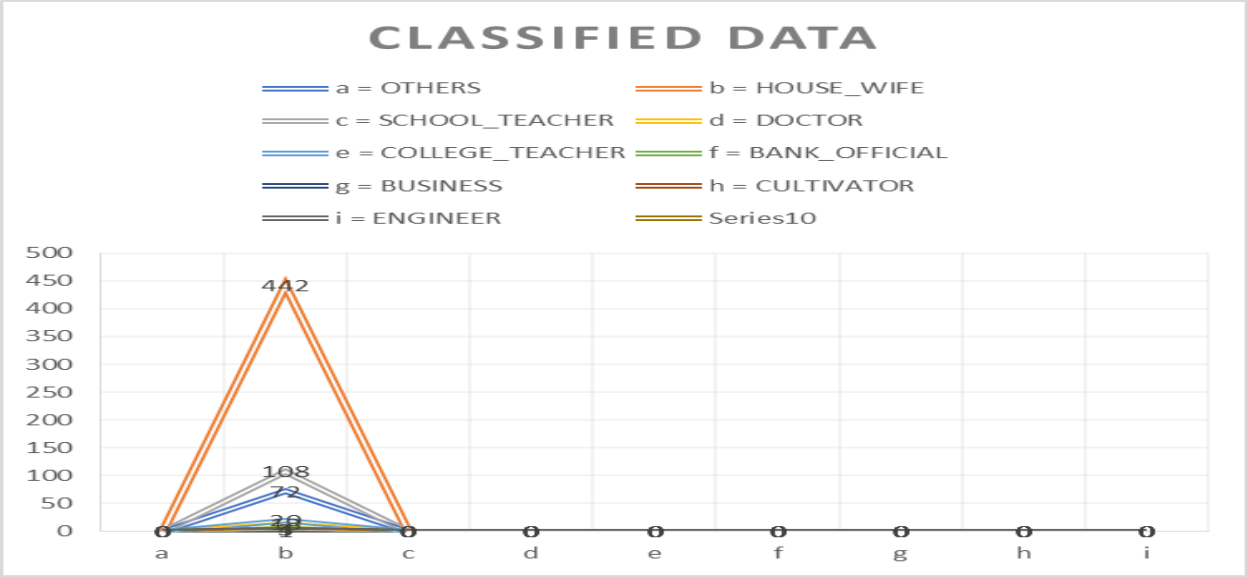

Fig 4 Classification data

The confusion matrix further illustrates the model's classification challenges. All instances for each class are misclassified as other categories. For example, every instance of HOUSE_WIFE is misclassified as HOUSE_WIFE, and similarly, instances from other classes like SCHOOL_TEACHER, DOCTOR, and ENGINEER are also incorrectly classified. The confusion matrix shows a severe issue with the model's ability to correctly identify and differentiate between categories. The model struggles significantly with classification accuracy, as evidenced by the confusion matrix and performance metrics. While some areas, like the ROC and PRC Areas for the ENGINEER class, show marginal performance, the overall metrics suggest that the model needs substantial improvement in its classification capabilities Fig 4.

Methods - Comparison of 1.ZeroR and 2.OneR Confusion Matrix Results

ZeroR Analysis Confusion Matrix Behavior The ZeroR model demonstrates that all actual instances of various classes (OTHERS, HOUSE_WIFE, SCHOOL_TEACHER, DOCTOR, COLLEGE_TEACHER, BANK_OFFICIAL, BUSINESS, CULTIVATOR, ENGINEER) are incorrectly predicted as either HOUSE_WIFE or ENGINEER. This indicates a severe misclassification issue, where the model does not accurately predict any of the classes except for HOUSE_WIFE and ENGINEER. Performance All off-diagonal entries in the confusion matrix are zeros, signifying that the model fails to correctly identify instances of most classes. The correct predictions are limited to HOUSE_WIFE and ENGINEER, but even these are incorrect for all

instances of other classes. ZeroR is a very basic classification algorithm that predicts the majority class of the training dataset, disregarding the input features. Its simplicity makes it a baseline model, but it also limits its predictive power. To improve its performance in dynamic environments where data evolves over time, several additional steps can be taken Data Enrichment Enhance the dataset with additional relevant features that might capture more nuances in the data. For example, in an educational setting, including features such as student engagement metrics, parental involvement, and local educational resources could provide more context and potentially improve predictions.

Dynamic Class Balancing Implement techniques to address class imbalance if present. For instance, periodically reassessing the distribution of classes in the data and applying resampling techniques like oversampling the minority class or undersampling the majority class can help in making the ZeroR model more robust. Periodic Model Updating Regularly update the ZeroR model with the latest data to ensure that the majority class prediction remains relevant as the data evolves. This can involve retraining the model at specified intervals or when significant changes in the data distribution are detected. Integration with Feature Engineering Although ZeroR itself does not use features, integrating it into a more sophisticated pipeline where feature engineering and selection are performed can help in understanding which features are most relevant and potentially guide the development of more advanced models. Hybrid Approaches Combine ZeroR with other models to create a hybrid system. For instance, using ZeroR as a baseline and then applying more complex algorithms (like decision trees, logistic regression, or ensemble methods) on the enriched dataset can provide a more comprehensive approach to prediction. Performance Monitoring Implement a system to continuously monitor the performance of the ZeroR model. Metrics such as accuracy, precision, recall, and F1-score can be tracked over time to assess whether the model remains effective and to identify when it needs to be updated.By integrating these additional steps, ZeroR can be made more adaptable and potentially improve its predictive performance in a changing dataset environment.

OneR Analysis Confusion Matrix Behavior The OneR model also shows significant misclassification issues. It reveals that instances of all classes are predominantly misclassified as other categories, with each class being incorrectly predicted as either HOUSE_WIFE or ENGINEER. The confusion matrix entries are zeros for all classes except these two, mirroring the

issue seen with ZeroR. Performance Metrics The True Positive Rate (TP Rate) is zero for most classes, indicating that no correct predictions are made for these categories. The False Positive Rate (FP Rate) is also zero, meaning no incorrect predictions as in other classes. The ROC Area and PRC Area are low for most classes, suggesting poor performance in distinguishing between classes, although ENGINEER has a slightly higher PRC Area (0.480), indicating some ability to balance precision and recall.

Conclusion

Both ZeroR and OneR models show critical issues in classification performance. They predominantly misclassify instances into either HOUSE_WIFE or ENGINEER, with no correct predictions for the other classes. The confusion matrices for both models highlight a severe lack of ability to differentiate between the majority of classes, with most predictions being incorrect. Similarities - Both models demonstrate that the majority of instances are misclassified into a few categories (HOUSE_WIFE and ENGINEER), with all off-diagonal matrix entries being zero, indicating poor classification performance. The TP Rate is zero for most classes, and the performance metrics (such as Precision, Recall, and F-Measure) cannot be computed due to the lack of true positives. Differences - While the fundamental issue of misclassification is consistent between the two models, the specific performance metrics provided for OneR (e.g., PRC Area for ENGINEER) suggest that OneR may offer slightly better performance in some areas compared to ZeroR. However, both models require significant improvements. In summary, both ZeroR and OneR exhibit substantial classification challenges, with a need for major enhancements in their ability to accurately predict and differentiate between the classes.

REFERENCES

[1]. Sandip Bhattacharjee, Dr R Naveenkumar, Rahul Singha, Somnath Mullick, Rubi Sarkar The Impact of Machine Learning on Enhancing Diversity and Inclusion through Advanced Recommence Screening Techniques Journal of Informatics Education and Research - ABDC 4 (2), 3141-3159, 2024.

[2]. Dr R. Naveen Kumar, Amit Kumar Bhore, Sourav Sadhukhan, Dr G. Manivasagam, Rubi Sarkar “Self-Monitoring System for Vision-Based Application Using Machine Learning

Algorithms" DOI 10.5281/zenodo.10547803, Vol 18 No 12 (2023), Page No 1958 – 1965, Published on 31-12-2023.

[3]. Dr R.Naveenkumar "An Empirical Research Approach on Confusion Matrix Using Existing Musical Industry Dataset" International Journal of Scientific Research in Engineering and Management (IJSREM) Volume 08 Issue 04 | April - 2024 SJIF Rating 8.448 ISSN 2582-3930.

[4]. J. Brownlee, "Logistic Regression For Machine Learning," 2017. [Online]. Available Https //Machinelearningmastery.Com/Logistic- Regression-For-Machine-Learning/. [Accessed 13-Mar-2018].

[5]. L. Devasena, I. B. S. Hyderabad, And L. Devasena, "Effectiveness Analysis Of Zeror , Ridor And Part Classifiers For Credit Risk Appraisal Effectiveness Analysis Of Zeror , Ridor And Part Classifiers For Credit Risk Appraisal," Int. J. Adv. Computer. Sci. Technol., Vol. 3, No. 11, Pp. 6–11, 2014.

[6]. Anuma Thakuri1, Shrestha Majumder1, Dr.R.Naveenkumar "A Comprehensive Analysis Study on House Rates Prediction System Using a Machine Learning Algorithm and a Broad-Spectrum Confusion Matrix" Journal of Computer Based Parallel Programming Vol- 9 (Issue 2 May 2024), 9-24.

[7]. S.Kotsiantis,"Credit risk analysis using a hybrid data mining model," Int. J. Intelligent Systems Technologies and Applications, Vol. 2, No. 4, pp. 345 – 356, 2007.

[8] Hamadi Matoussi and Aida Krichene, "Credit risk assessment using Multilayer Neural Network Models Case of a Tunisian bank," 2007.

[9] Lean Yu, Shouyang Wang, Kin Keung Lai, "Credit risk assessment with a multistage neural network ensemble learning approach",Expert Systems with Applications (Elsevier) 34, pp.1434 – 1444, 2008.

[10] Sanaz Pourdarab, Ahmad Nadali and Hamid Eslami Nosratabadi, "A Hybrid Method for Credit Risk Assessment of Bank Customers," International Journal of Trade, Economics and Finance , Vol. 2, No. 2, April 2011.

[11] Vincenzo Pacell and Michele Azzollini "An Artificial Neural Network Approach for Credit Risk Management", Journal of Intelligent Learning Systems and Applications, 3, pp. 103 - 112, 2011.