International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

A Comprehensive Analysis in Predicting Student Success based on School Academic Records using ZeroR and Logistic Regression Algorithms

Zahra Nabila Izdihar*[1], **Simeon Yuda Prasetyo**[2], **Ghinaa Zain Nabiilah**[3]

Submitted: 25/01/2024 **Revised**: 04/03/2024 **Accepted**: 11/03/2024

Abstract: Student success is intricately woven with various factors, including exam scores, class participation, and organizational experiences. This research embarks on a captivating exploration, delving into the multifaceted elements that shape student success based on meticulous scrutiny of school records. This study aims to analyze the determinants of student success based on school records. The examination of academic factors facilitates the prediction of success in both academic and non-academic domains. By analyzing students' academic factors, we can predict student success in academic and non-academic achievements in the future. Because, basically, student success is not only based on academic achievement, but also in non-academic fields. This research compares the effectiveness of the ZeroR and Logistic Regression methods.Leveraging academic records from schools across diverse nations, we navigate through a tapestry of attributes to discern patterns. Logistic Regression proves more adept at discerning nuanced patterns within the data, resulting in more precise predictions. Notably, the Logistic Regression algorithm demonstrated an accuracy of 73.75%, markedly surpassing ZeroR, which achieved an accuracy of 43.96%. This substantial difference underscores the superior performance of Logistic Regression in predictive analysis. In conclusion, the limited predictive capacity of the ZeroR algorithm, ethered to its limitation in predicting a singular class, highlights the necessity of employing advanced models like Logistic Regression for precise student success predictions based on school records.

Keywords: *Prediction, Forecasting, Machine Learning, Logistic Regression, ZeroR Algorithm*

1. Introduction

Student success is a multifaceted outcome influenced by various factors such as exam scores, class engagement, and extracurricular experiences. Previous research has delved into predicting academic performance using Ensemble methods and analyzing educational datasets through X-API to enhance student outcomes. Prior methodologies, including naive Bayes, decision trees, KNN, ANN, and ensemble, have been employed to investigate these aspects.

In this study, the authors aim to conduct a comprehensive analysis of the determinants of student success based on school records. By scrutinizing students' academic factors, the goal is to predict their success not only academically but also in non-academic realms. Acknowledging that student success extends beyond academic achievements, the exploration focuses on understanding how various elements contribute to overall success.

The prediction of success involves classifying student records across different categories. Predictions of non-academic achievements are influenced by how often students check announcements. Regularly perusing notices posted at school provides students with diverse information beyond the school environment. Frequent visits to the announcement board can broaden students' information and perspectives, exposing them to various opportunities outside the school.

Announcements often contain details about competitions, sparking students' interest and encouraging participation. Such engagement, be it in competitions, scientific projects, or other endeavors, contributes to non-academic achievements. For instance, when there's an announcement about a chess competition or a scientific writing contest, students participating and winning would earn them non-academic accolades.

Furthermore, the frequency with which students access educational resources is closely tied to their engagement in scientific work. Opening resources enhances insight and aids in the development of research skills. Therefore, through meticulous analysis of students' academic factors, the aim is to predict not only their academic success but also their achievements in non-academic realms, fostering a holistic understanding of student success.

[1] *Software Engineering Program, Computer Science Department, Bina Nusantara University – 17142, INDONESIA*
ORCID ID : 0009-0003-0394-2095
[2] *Computer Science Department, Bina Nusantara University – 11480, Jakarta, INDONESIA*
ORCID ID : 0000-0002-6077-4003
[3] *Computer Science Department, Bina Nusantara University – 11480, Jakarta, INDONESIA*
ORCID ID : 0000-0001-7638-7449
** Corresponding Author Email: Zahra.izdihar@binus.ac.id*

2. Method

2.1. Logistic Regression Algorithm

Logistic Regression is a statistical technique employed to analyze the relationship between one or more predictors and outcomes that exhibit a dichotomous nature, typically involving binary outcomes like the presence or absence of a specific event. This method has gained widespread popularity as a statistical tool, particularly over the last two decades. Logistic Regression is often regarded as the preferred statistical approach in situations where the prediction of binary outcomes is essential [10]. In essence, it is a well-established technique commonly used for modeling binary outcomes, addressing binary classification problems with two distinct class values.

Logistic Regression offers several advantages [11], enhancing its versatility and applicability across various analysis tasks. Some of its strengths include:

- Flexibility: The algorithm is highly flexible, accommodating various types of input and supporting diverse analysis tasks.
- Prediction using demographics: It can effectively utilize demographic information to make predictions.
- Exploration of contributing factors: Logistic Regression enables the exploration and consideration of factors that contribute to a specific outcome.
- Classification of objects with multiple attributes: It is adept at classifying documents, emails, or other objects characterized by numerous attributes.

However, Logistic Regression also presents certain limitations:

- Lack of support for drillthrough: This limitation arises from the non-necessity for the node structure in a mining model to directly correspond to the underlying data.
- Absence of support for creating data mining dimensions.
- Support for OLAP mining model: Logistic Regression supports the use of the OLAP mining model.
- Non-support for the use of Predictive Model Markup Language (PMML) in creating mining models.

2.2. ZeroR Algorithm

The ZeroR classifier, also known as the 0-R classifier, is a straightforward classification method. While it is relatively simple, its application establishes a performance baseline for certain datasets, providing a reference point for the improvement that more sophisticated classifiers can achieve. Consequently, ZeroR serves as a practical benchmark for evaluating how accurately a class can be predicted without considering additional attributes, serving as a Lower Bound on Performance [7].

ZeroR functions as a target-dependent classifier, disregarding all predictors except for the target attribute [8]. So, ZeroR is an extremely simple and basic classifier, and its predictive accuracy is generally quite limited. The reason for this is that ZeroR makes predictions based solely on the majority class in the dataset, regardless of any features or patterns present in the data. It is essentially a baseline or benchmark model that provides a minimal standard for comparison.

This classifier exclusively predicts the majority class and is constructed based on Frequency Tables. The process involves examining the target attributes and their potential values, creating a frequency table, and selecting the most prevalent values. When applied to a given dataset, ZeroR returns the most frequently occurring values for the target attribute. Notably, ZeroR, as implied by its name, does not involve rules that operate on non-target attributes. To elaborate, ZeroR has the capability to predict the mean (for numeric target attributes) or mode (for nominal target attributes), yet it exclusively predicts a single class [9]. Despite its simplicity, ZeroR is utilized as a predictor and proves valuable in assessing student academic performance.

2.3. Dataset

This article uses a dataset obtained from Kaggle and created by Elaf Abu Amrieh, Thair Hamtini, and Ibrahim Aljarah from the University of Jordan, Amman. The research was conducted by the researchers in 2015 and 2016. Data were collected using a student activity tracker tool called the Experience API (xAPI). xAPI is a component of the Training and Learning Architecture (TLA) that allows monitoring learning progress and learner actions such as reading articles or watching training videos. xAPI helps learning activity providers identify students, activities, and objects that describe the learning experience. The dataset consists of 480 student records and 16 features. Features are grouped into three main categories: (1) demographic features such as gender and nationality, (2) academic background features such as educational stage, grade level, and section, and (3) behavioral features such as raising hands in class, opening resources, answering surveys by parents, and school satisfaction.

The dataset comprises 305 males and 175 females. Students come from various countries with details: 179 students from Kuwait, 172 students from Jordan, 28 students from Palestine, 22 students from Iraq, 17 students from Lebanon, 12 students from Tunisia, 11 students from Saudi Arabia, 9 students from Egypt, 7 students from Syria, 6 students from the United States, Iran, and Libya, 4 students from Morocco, and one student from Venezuela.

The data set was collected over two semesters with details: 245 student records collected during the first semester and 235 student records collected during the second semester. It also includes school attendance features grouped into two categories based on class attendance: 191 students absent (alpha) for more than 7 days and 289 students absent for less than 7 days.

The dataset also includes a new category of features. These features play a role in the educational process. Parental participation features have two sub-features: Parental survey responses and parental satisfaction. There are 270 parents who answered the survey and 210 who did not, and 292 parents are satisfied with the school while 188 are not.

As for the characteristics of this dataset, it is multivariate with a total of 480 instances, 3 class labels, and 16 attributes. Additionally, there are no missing values in this dataset. Therefore, I chose this dataset because the presented data is comprehensive. Table 1 is a table containing metadata, listing the attributes of the dataset used.

Table 1. List of Attributes in the Dataset I

Attribute Name	**Description**	**Data Type**	**Value Span/ Enumeration**
Gender	Gender Determination	Nominal	(F,M)
Nationality	Classification of Nationality	Nominal	(Kuwait, Lebanon, Egypt, SaudiArabia, USA, Jordan, Venezuela, Iran, Tunis, Morocco, Syria, Palestine, Iraq, Lybia)
Place of Birth	Classification of Place of Birth	Nominal	(Kuwait, Lebanon, Egypt, SaudiArabia, USA, Jordan, Venezuela, Iran, Tunis, Morocco, Syria, Palestine, Iraq, Lybia)
StageID	Students' Educational Level	Nominal	(lowerLevel, MiddleSchool, HighSchool)
GradeID	Determining Level Grade	Nominal	(G-02, G-04, G-05, G-06, G-07, G-08, G-09, G-10, G-11, G-12),
SectionID	Determining Students' Classroom	Nominal	(A, B, C)
Semester	Determining the First Semester (F) or Second Semester (S)	Nominal	(F, S)
Relation	Guardian of the student or responsible for	Nominal	(Father, Mum)
Raisehands	Determining classroom participation through raising hands.	Numeric	(Minimum, maximum, mean, Std. Dev)

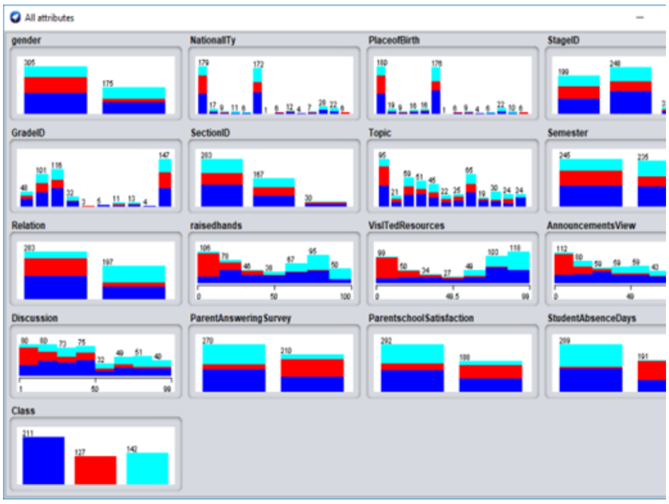


Fig. 1. Data Distribution Graph of all attributes

Table 2 is another example of using tables to explain the characteristics of the dataset used in research.

Table 2. Proportion of The Number of Dataset Class Members

Label Name	Number of Instance	Percentage
Class M	211	43,96%
Class L	127	26,46%
Class H	142	29,58%

Fig. 2. Data Preparation on WEKA

2.4. Experimental Design

In this research, the tool I used was Weka made by the University of Waikato, New Zealand. This experiment using Weka version 3.9.2. The laptop hardware specifications used are 8GB RAM, 8th generation Intel Core i5 processor, 1T laptop hard disk memory and Windows 10 operating system. And the research steps in Information Retrieval research are shown in Figure 3.

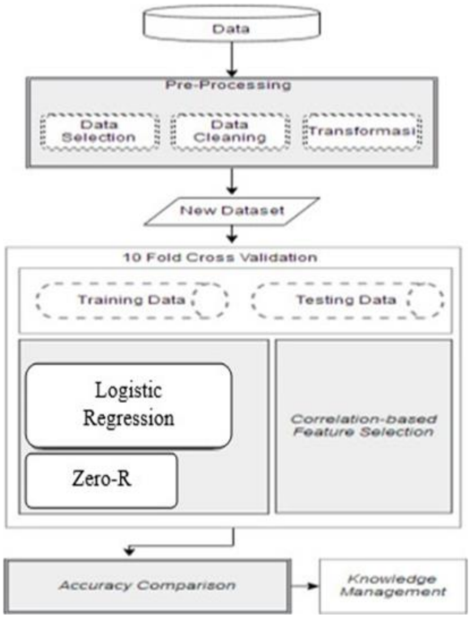


Fig. 3. Research Flow

3. Results and Discussion

3.1. Results

The following are the results of the research which contain tables of algorithm performance measurement results. The performance measures used include accuracy, error rate, precision, recall.

Here's a brief explanation [17]:

- Precision: the number of positive examples correctly classified divided by the number of examples the system labels as positive

- Recall: the number of positive examples correctly classified divided by the number of positive examples in the data

- Fscore: combination of the above.

Notes:

TP = True Positive

TN = True Negative

FP = False Positive

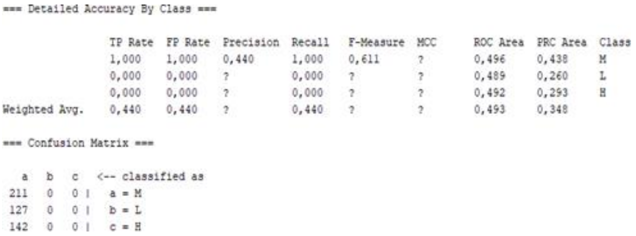
FN = False Negative

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} * 100\%$$

$$Presisi = \frac{TP}{FP+TP} * 100\%$$

$$Recall = \frac{TP}{FN+TP} * 100\%$$

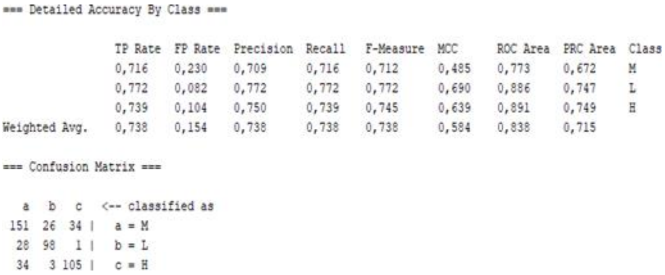
Fig. 4. Formula to calculate Accuracy, Precision and Recall



```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    1,000    0,440      1,000    0,611      ?        0,496     0,438     M
                 0,000    0,000    ?          0,000    ?          ?        0,489     0,260     L
                 0,000    0,000    ?          0,000    ?          ?        0,492     0,293     H
Weighted Avg.    0,440    0,440    ?          0,440    ?          ?        0,493     0,348

=== Confusion Matrix ===

   a   b   c   <-- classified as
 211   0   0 |   a = M
 127   0   0 |   b = L
 142   0   0 |   c = H
```

Fig. 5. Performance Results of Logistic Algorithm Based on Cross

Validation



```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,716    0,230    0,709      0,716    0,712      0,485    0,773     0,672     M
                 0,772    0,082    0,772      0,772    0,772      0,690    0,886     0,747     L
                 0,739    0,104    0,750      0,739    0,745      0,639    0,891     0,749     H
Weighted Avg.    0,738    0,154    0,738      0,738    0,738      0,584    0,838     0,715

=== Confusion Matrix ===

   a   b   c   <-- classified as
 151  26  34 |   a = M
  28  98   1 |   b = L
  34   3 105 |   c = H
```

Fig. 6. Performance Results of ZeroR Based on Cross Validation

Table 3. Classifier with Cross Validation

Performance	ZeroR Algorithm	Logistic Regression
Correctly Classified Instances	211 (43,96%)	354 (73,75%)
Incorrectly Classified Instances	269 (56,04%)	126 (26,25%)
Kappa Statistic	0	0,595
Mean Absolute Error (MAE)	0,43	0,21
Root Mean Squared Error (RMSE)	0,47	0,38
Relative Absolute Error (RAE)	100%	49,27%
Root Relative Squared Error (RRSE)	100%	80,66%

Table 4. Classifier with Percentage Split (60%)

Algorithm Performance	ZeroR Algorithm	Logistic Regression
Correctly Classified	84	123
Instances	(43,75%)	-64%
Incorrectly Classified	108	69
Instances	(56,25%)	(35,9%)
Kappa Statistic	0	0,44
Mean Absolute Error (MAE)	0,43	0,24
Root Mean Squared Error (RMSE)	0,47	0,43
Relative Absolute Error (RAE)	100%	56,12%
Root Relative Squared Error (RRSE)	100%	92,35%

Table 5. Classifier with Training Set

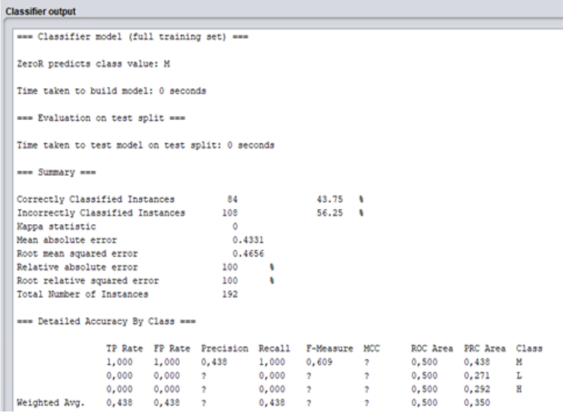
Algorithm Performance	ZeroR Algorithm	Logistic Regression
Correctly Classified Instances	211 (43,96%)	408 (85%)
Incorrectly Classified Instances	269 (56,04%)	72 (15%)
Kappa Statistic	0	0,77
Mean Absolute Error (MAE)	0,43	0,15
Root Mean Squared Error (RMSE)	0,47	0,27
Relative Absolute Error (RAE)	100%	34,76%

Root Relative Squared Error (RRSE)	100%	57,97%

Through experiments comparing the ZeroR and Logistic Regression methods, the results revealed that ZeroR exhibited a lower error value than the Logistic Regression algorithm. Specifically, the accuracy of the Logistic Regression algorithm was determined to be 73.75%, a significantly higher value compared to ZeroR, which achieved an accuracy of 43.96%. This discrepancy underscores the superior performance of Logistic Regression in this predictive analysis.

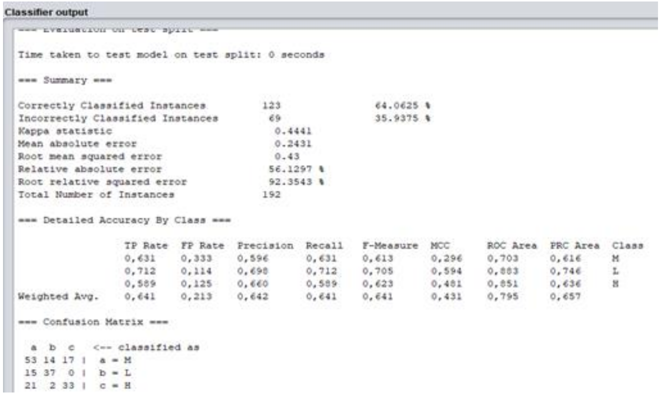
Additionally, the images below is the result of the classification process using Logistic Regression algorithm:

- **Classifier Output from Logistic Regression**



```
=== Classifier model (full training set) ===

ZeroR predicts class value: M

Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances          84               43.75   %
Incorrectly Classified Instances       108               56.25   %
Kappa statistic                          0
Mean absolute error                      0.4331
Root mean squared error                  0.4656
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              192

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    1,000    0,438      1,000    0,609      ?        0,500     0,438     M
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,271     L
                 0,000    0,000    ?          0,000    ?          ?        0,500     0,292     H
Weighted Avg.    0,438    0,438    ?          0,438    ?          ?        0,500     0,350
```

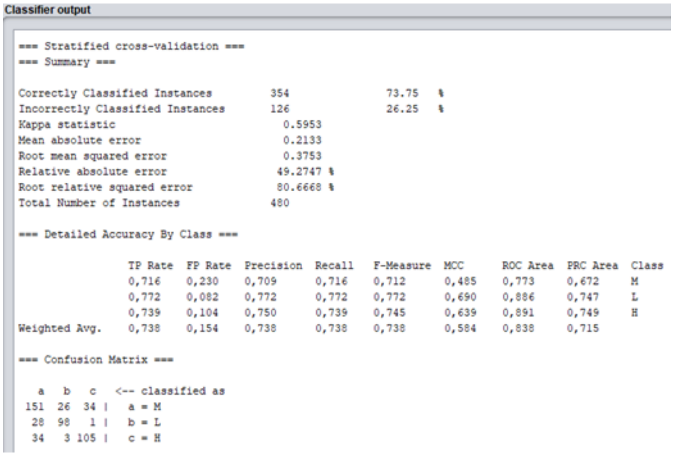
Fig. 7. The Testing Process in Logistic Regression using Cross Validation



```
=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         123               64.0625 %
Incorrectly Classified Instances        69               35.9375 %
Kappa statistic                          0.4441
Mean absolute error                      0.2431
Root mean squared error                  0.43
Relative absolute error                 56.1297 %
Root relative squared error             92.3543 %
Total Number of Instances              192

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,631    0,333    0,596      0,631    0,613      0,296    0,703     0,616     M
                 0,712    0,114    0,698      0,712    0,705      0,594    0,883     0,746     L
                 0,589    0,125    0,660      0,589    0,623      0,481    0,851     0,636     H
Weighted Avg.    0,641    0,213    0,642      0,641    0,641      0,431    0,795     0,657

=== Confusion Matrix ===

  a  b  c   <-- classified as
 53 14 17 |  a = M
 15 37  0 |  b = L
 21  2 33 |  c = H
```

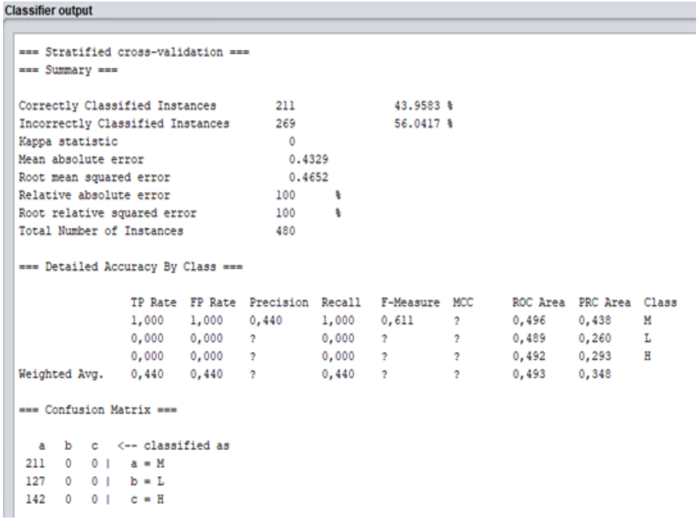
Fig 8. The Testing Process in Logistic Regression using Percentage Split 60%

- **Classifier Output from ZeroR Algorithm**



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

Correctly Classified Instances         354               73.75   %
Incorrectly Classified Instances       126               26.25   %
Kappa statistic                          0.5953
Mean absolute error                      0.2133
Root mean squared error                  0.3753
Relative absolute error                 49.2747 %
Root relative squared error             80.6668 %
Total Number of Instances              480

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,716    0,230    0,709      0,716    0,712      0,485    0,773     0,672     M
                 0,772    0,082    0,772      0,772    0,772      0,690    0,886     0,747     L
                 0,739    0,104    0,750      0,739    0,745      0,639    0,891     0,749     H
Weighted Avg.    0,738    0,154    0,738      0,738    0,738      0,584    0,838     0,715

=== Confusion Matrix ===

   a   b   c   <-- classified as
 151  26  34 |   a = M
  28  98   1 |   b = L
  34   3 105 |   c = H
```

Fig. 9. Testing Process in ZeroR Algorithm using Cross Validation



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

Correctly Classified Instances         211               43.9583 %
Incorrectly Classified Instances       269               56.0417 %
Kappa statistic                          0
Mean absolute error                      0.4329
Root mean squared error                  0.4652
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              480

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    1,000    0,440      1,000    0,611      ?        0,496     0,438     M
                 0,000    0,000    ?          0,000    ?          ?        0,489     0,260     L
                 0,000    0,000    ?          0,000    ?          ?        0,492     0,293     H
Weighted Avg.    0,440    0,440    ?          0,440    ?          ?        0,493     0,348

=== Confusion Matrix ===

   a   b   c   <-- classified as
 211   0   0 |   a = M
 127   0   0 |   b = L
 142   0   0 |   c = H
```

Fig. 10. Testing Process in ZeroR Algorithm using Percentzge Split 60%

3.2. Discussions

In this study, an attempt was made to predict students' success based on the analysis of student records in school. Through experiments comparing the ZeroR and Logistic Regression methods, the results revealed that ZeroR exhibited a lower error value than the Logistic Regression algorithm.

Specifically, the accuracy of the Logistic Regression algorithm was determined to be 73.75%, a significantly higher value compared to ZeroR, which achieved an accuracy of 43.96%. This discrepancy underscores the superior performance of Logistic Regression in this predictive analysis.

Consequently, it can be concluded that the ZeroR algorithm demonstrated poor performance, primarily due to its limitation in predicting only a single class. The findings emphasize the importance of selecting appropriate predictive models, such as Logistic Regression, for more accurate and reliable student success predictions based on school records.

So, the difference in performance between ZeroR and Logistic Regression in this study can be attributed to the inherent limitations of the ZeroR algorithm. ZeroR, being a simplistic baseline model, predicts outcomes based on the majority class in the dataset, lacking the ability to consider multiple classes or intricate patterns within the data. Moreover, ZeroR disregards any additional features, making its predictions solely based on the most prevalent class. In contrast, Logistic Regression excels in handling more complex relationships between multiple features and the predicted outcomes.

By considering a range of attributes, Logistic Regression is better equipped to capture the nuanced patterns present in the data, resulting in more accurate predictions. The experimental results, with Logistic Regression achieving an accuracy of 73.75% compared to ZeroR's 43.96%, underscore the superior performance of Logistic Regression in this predictive modeling task. This discrepancy highlights the significance of utilizing more advanced algorithms, like Logistic Regression, when addressing the intricacies of real-world datasets for improved predictive accuracy.

4. Conclusion

By analyzing students' academic factors, we can predict their success in academic and non-academic achievements in the future. After comparing the ZeroR and Logistic Regression methods, it can be observed that the results of this study indicate that the ZeroR algorithm has poor performance with high error values, as ZeroR only predicts one class. For further development, trials can be carried out using other algorithms for comparison.

5. References and Footnotes

Acknowledgements

This research was supported/partially supported by Bina Nusantara University. We thank Simeon Yuda Prasetyo for assistance with [particular technique, methodology], and Ghinaa Zain Nabiilah, for comments that greatly improved the manuscript.

Author contributions

Zahra Nabila Izdihar: Conceptualization, Writing-Original draft, Methodology, Implementation

Simeon Yuda Prasetyo: Implementation, Investigation, Writing-Reviewing and Editing

Ghinaa Zain Nabiilah: Data curation, Writing-Original draft preparation, Validation.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] M. Ichsan, "Psikologi Pendidikan Dan Ilmu Mengajar," *Edukasi*, Vol. 2, No. 1, Pp. 183–200, 2016.

[2] T.Sukitman, "Internalisasi Pendidikan Nilai Dalam Pembelajaran (Upaya Menciptakan Sumber Daya Manusia Yang Berkarakter)," *J. Pendidik. Sekol. Dasar*, Vol. 2, No. 2, Pp. 85–12, 2016.

[3] T. D. Utama *Et Al.*, "Vol 3 . No 2 . Desember 2014 Issn : 2301 – 7201 Implementasi Algoritma Iterative Dichotomiser 3 Pada," Vol. 3, No. 2, Pp. 74–83, 2014.

[4] M. Sokolova And G. Lapalme, "A Systematic Analysis Of Performance Measures For Classification Tasks," *Inf. Process. Manag.*, Vol. 45, No. 4, Pp. 427– 437, 2009.

[5] F. Gorunescu, *Data Mining: Concepts And Techniques*, Vol. 12. 2011.

[6] J. Han, M. Kamber, And J. Pei, *Data Mining: Concepts And Techniques*. 2012.

[7] M. Koklu, H. Kahramanli, And N. Allahverdi, "Applications Of Rule Based Classification Techniques For Thoracic Surgery," *Jt. Int. Conf. 2015*, No. November, Pp. 1991–1998, 2015.

[8] S. Fitri, "Perbandingan Kinerja Algoritma Klasifikasi Naïve Bayesian , Lazy-Ibk , Zero-R , Dan Decision Tree- J48," *Dasi*, Vol. 15, No. 1, Pp. 33–37, 2014.

[9] 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. Comput. Sci. Technol.*, Vol. 3, No. 11, Pp. 6–11, 2014.

[10] P. Komarek And A. Moore, "Making Logistic Regression A Core Data Mining Tool A Practical Investigation Of Accuracy , Speed , And Simplicity," *Compute*, No. 1, Pp. 1–4, 2005.

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