P-ISSN: 2811-2512
E-ISSN: 2811-2520

THE NEXUS
A Publication of Al-Hikmah University, Ilorin, Nigeria
(Science Edition)

Available online at https://www.alhikmahuniversity.edu.ng

COMPARING THE EFFICIENCY OF LOGISTIC REGRESSION CLASSIFIER AND K-NEAREST NEIGHBOURS CLASSIFIERS FOR PREDICTING STUDENTS' PERFORMANCE IN COMPUTER SCIENCE PROGRAMME

*[1]Ibrahim, S.A., [2]AbdulRauf, U.T., [2]Mustapha, I.O. and [1]Oni, A.A.

[1]Department of Physical Sciences, Al-Hikmah University, Ilorin, Nigeria
[2]Department of Computer Science, Al-Hikmah University, Ilorin, Nigeria

ARTICLE INFO

Article history:
Received: May 3, 2023
Revised: December 20, 2023
Accepted: January 22, 2024
Published online: May 30, 2024

Citation:
Ibrahim, S.A., AbdulRauf, U.T., Mustapha, I.O. and Oni, A.A. (2024). Comparing the Efficiency of Logistic Regression Classifier and K-Nearest Neighbours Classifiers for Predicting Students' Performance in Computer Science Programme. The Nexus (Science Edition). 3(1): 64-67.

***Corresponding Author:**
Ibrahim, S. A.
Al-Hikmah University, Ilorin, Nigeria
*e-mail: adeshinas2010@alhikmah.edu.ng

ABSTRACT

Accurate prediction of Master's program eligibility from Computer Science Bachelor's performance is vital. Despite K-nearest neighbours' (KNN) common use in predictions, there's a gap in comparing it with the Logistic Regression Classifier (LRC). This study aimed to address this gap by identifying the most suitable classifier between LRC and KNN for accurately predicting students' performance in the computer science programme. In order to evaluate the performance metrics of two classification algorithms, LRC and KNN were modelled through 10-fold cross-validation in WEKA, with a comprehensive evaluation of performance metrics for each classifier. The study used secondary data from Al-Hikmah University, Ilorin, Nigeria (2009-2015) on computer science students' academic performances. It included 7 attributes and 478 instances for each, comprising three categorical and four numeric features. Class labels Y_i(YES, NO) reflected meeting minimum admission requirements, with grade scales for class labels including 1.0-1.49(pass), 1.50-2.3(Third class honor) 2.40-3.49 (Second class honor lower division), 3.5-4.49 (Second class honor upper division) and 4.5-5.0 (First class honor). LRC showcased superior performance over KNN, when tuning parameter k = 1with Euclidean distance used as distance metrics, across multiple metrics, including accuracy (94.7699% vs. 89.9582%), precision (96.1% vs. 92.7%), recall (96.9% vs. 93.8%), F-measure (96.5% vs. 93.3%), ROC Area (97.5% vs. 85.6%), and error rate (5.2301% vs. 10.0418%). Notably, KNN exhibited faster processing time (0.01 sec vs. 0.07 sec) when compared to LRC. The optimal KNN configuration for the model was observed when k = 3. The study recommends utilizing LRC as the preferred predictive model for students' performance in a computer science programme.
Keywords: *LRC, KNN Classifiers, Cross Validation, and Performance Metrics.*

1.0 INTRODUCTION

In recent years, there has been an increasing interest in using machine learning algorithms for educational data mining and predictive analytics in the field of computer science education. Predictive models can aid in identifying students at risk, implementing appropriate interventions, and improving educational outcomes. In the admission process for a postgraduate programme, such as Computer Science, predicting students who meet the minimum University admission requirement for a second degree (M.Sc.) through their first degree (B.Sc.) obtained in the Computer Science programme is a critical task. K-nearest neighbours (KNN) have been widely used for predicting student outcomes in various educational settings. However, there is a lack of research that directly compares the performance of KNN with a Logistic Regression Classifier (LRC) in predicting students' performance in a computer science programme, based on empirical evidence (Amra & Maghari, 2017; Asril & Isa, 2020; Sathe & Adamuthe, 2021; Wiyono *et al.,* 2020), among others.

Therefore, this study aims to address this gap by comparing and identifying the most suitable classifier between LRC and KNN for accurately predicting students' performance in a computer science programme, which can help improve the efficiency of the admission process. The study also contributes to the existing literature on educational data mining and predictive analytics in the field of computer science education. The rest of this study is organized as follows: Section 2, related work. Section 3, describes the materials and methods used. Section 4: deals with results and discussion and Section 5, deals with the conclusion.

2.0 Related Works

Researchers have employed various approaches to investigate classification algorithms and performance metrics for assessing students' academic performance. Amra and Maghari (2017) proposed a predictive model using KNN

and Naive Bayes to enhance student performance in secondary schools in the Gaza Strip. The study demonstrated that Naive Bayes outperformed KNN with a 93.6% accuracy. Similarly, Devasia *et al.* (2016) explored data mining techniques, favoring Naive Bayesian mining for its accuracy over methods like regression and decision trees. Wiyono et al. (2020) compared KNN, SVM, and Decision Tree, finding SVM to be the most accurate (95%). Shamsi and Lakshmi (2016) used data mining to predict students' grades and dropouts, selecting techniques based on accuracy and suitability. Vyas and Gulwani (2017) proposed a decision tree system for predicting student performance, aiding faculty in identifying and assisting struggling students.

Shingari *et al.* (2017) discussed using data mining for predicting higher education students' performance, emphasizing the utility of educational data mining. Yaacob *et al.* (2019) used supervised data mining with Naïve Bayes, identifying factors like A+ scores in specific subjects as significant for predicting excellent students. Sathe and Adamuthe (2021) compared various classifiers, finding Random Forest and C5.0 to perform better. Tripathi et al. (2019) focused on the naive Bayes classification model, comparing its accuracy and execution time. Wiyono and Abidin (2019) found SVM to have the best accuracy in predicting student performance.

Asril and Isa (2020) used K-Nearest Neighbor to predict study periods based on final grades, demonstrating high accuracy. Deepika *et al.* (2019) proposed a hybrid Feature Selection method for predicting Student Academic Performance, achieving improved accuracy compared to existing models. Zulfiker *et al.* (2020) used machine learning to predict student grades with multiple classifiers, achieving 81.73% accuracy. Akuma and Abakpa (2021) predicted students' performance based on CGPA, achieving 87.84% accuracy. Yakubu and Abubakar (2022) predicted academic performance using machine learning and early detection indicators, revealing insightful predictors.

While previous studies explored either Logistic Regression Classifier (LRC) or K-nearest neighbors classifier (KNN) for predicting student performance, this study aims to fill the gap by directly comparing the performance of these two algorithms in the context of a computer science programme.

3.0 Materials and Methods

3.1 Materials

3.1.1 Data/Experimental setup

Data used for this study was obtained through secondary sources from the Department of Physical Sciences Al-Hikmah University Ilorin-Nigeria, from 2009 to 2015 on students' academic performance in a computer science programme, with 7 attributes and 478 instances.

Thus, the following are the attributes included in the data collected: Age, state of origin, gender, cumulative grade point average (cgpa), total credits passed (tcp), mode of entry and class of degree for the various graduate students. Hence the class of degree is known as a study variable called class label, coded as follows: Y_i *(YES. NO), if* Y_i $(2^2, 2^1, 1)$ = *YES* or *if* Y_i *(pass,* 3^{rd}*)* = *NO,* Where Y_i $(2^2, 2^1, 1)$ represents the students who met the minimum University admission requirement for a second degree (M.Sc.) through their first degree (B.Sc.) obtained in Computer Science programme in the Department of Physical Sciences, and Y_1 (Pass, 3^{rd}) represent the students who did not meet the minimum University admission requirement for a second degree (M.Sc.) through their first degree (B.Sc.) obtained in Computer Science programme in the Department of Physical Sciences.

The grading scale for class label includes 1.0-1.49 (Pass), 1.50-2.3 (Third class honor), 2.40-3.49 (Second class honor lower division), 3.5-4.49 (Second class honor upper division) and 4.5-5.0 (First class honor).

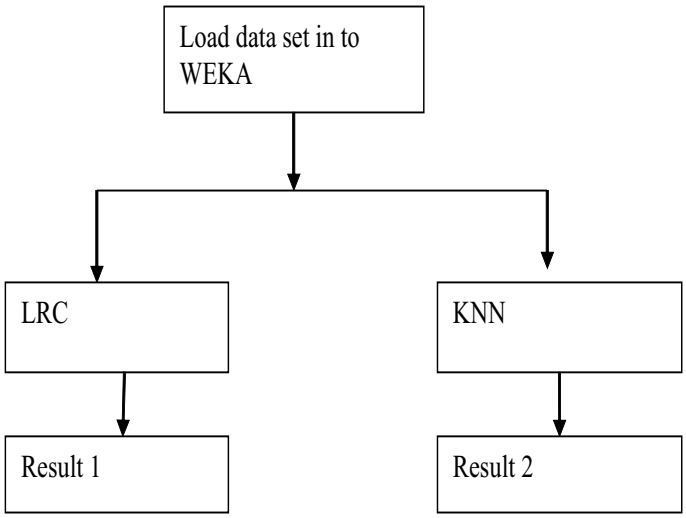

Figure 1: Technique workflow

3.2 Methods

3.2.1 The Description of Technique Workflow

The technique employee in the implementation of this study is described as follows:

3.2.2 Mathematical Description of Algorithms

Brief descriptions of the mathematical development of classification algorithms used are provided in what follow:

3.2.2.1 Logistic Regression Classifier

Consider a collection of k-categorical or continuous features (or predictor variables) to be denoted by vector $X^1 = x_1, x_2, \ldots x_k$). let the conditional probability that the label class is present be denoted by $P(Y-1|x_1, x_2, \ldots x_k)$, then the logit or log odds of having $Y-1$ is modeled as a linear function of features (or predictor variables) as:

$$In\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

(Ibrahim & Fadil, 2020; Ibrahim *et al.,* 2021) (1)

$$\frac{p}{1-p} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k} \quad (2)$$

$$p = \frac{p}{1+e^{-z}} \quad (3)$$

Where $Z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ (4)

and β_0 is the constant or intercept and $\beta_1, \beta_2, \ldots, \beta_k$ are the regression coefficients $x_1, x_2, \ldots x_k$ respectively.

Thus, the decision boundary for two-class logistic regression lies where the prediction probability is 0.50. i.e.

$$p\,(Y = 1 \mid x_1, x_2, \ldots x_k) = \frac{1}{1 + e^{\ 0 + \ 1 x_1 + \ 2 x_2 + \cdots + \ k x_k}} = 0.50 \quad (5)$$

This occurs when

$$- \ \ _0 - \ \ _1 x_1 - \ \ _2 x_2 - \cdots - \ \ _k x_k = 0 \quad (6)$$

Because this is a linear equality in the attribute values, the boundary is a plane, or hyperplane, in an instance space. It is easy to visualize sets of points that cannot be separated by a single hyperplane, and these cannot be discriminated correctly by logistic regression.

3.2.2.2 KNN classifier

KNN classifier attempts to predict the class (i.e. categories) to which the class label belongs by computing the local probability. In KNN, increasing in *K* value will tend to smooth out decision boundaries, avoiding overfitting at the cost of some resolution. There is no single value of *K* that will work for every single dataset. For classification models, especially if there are only two classes, an odd number is usually chosen for *K*.

Similarly, distance metrics measure how 'close' two points are to each other, which is measured in different ways. The most commonly used distance metric is Euclidean. Another metric is the so-called Manhattan. More generally, these are both forms of what is called Minkowski, whose formula is:

$$d\,(X, Y) = \left(\ \ _{i-1}^{p} \mid x_i \quad y_i \right|^p)^{\frac{1}{p}} \text{ (Amra \& Maghari, 2017; Asril \& Isa, 2020)} \quad (7)$$

when p = 1, this formula is the same as the Manhattan distance, and when p = 2, Euclidean distance is defined.

3.3 Performance Metrics

3.3.1 Confusion matrix

Table 1: The Descriptions of the Confusion Matrix

		Observed	
		YES	*NO*
Predicted model	*YES*	*TP*	*FP*
	NO	*FN*	*FN*

where, True positives ***(TP)*** are when the class label is correctly predicted to be positive ***(YES)***, and it is observed to be positive ***(YES)***. False positives ***(FP)*** are when the class label is predicted to be positive ***(YES)***, and it is observed to be negative ***(NO)***. False Negatives ***(FN)*** are when the class label is predicted to be negative ***(NO)***, and it is observed to be positive ***(YES)***. True Negatives ***(TN)*** are when the class label is correctly predicted to be negative ***(NO)***, and it is observed to be negative ***(NO)***.

3.3.1.1 Classifier's Performance Metrics

The performance metrics of classifiers include the following: Accuracy, Error rate, Precision, Recall, F-measure and Time taken to build the model. The terms are defined as follows:

$$\text{Accuracy } (AC) = \frac{(TP + TN)}{(TP+TN+FP+FN} \quad (8)$$

$$\text{Error rate } (ER) = 1 \quad ACC \quad (9)$$

$$\text{Precision } (P) = \frac{TP}{(TP+FP)} \quad (10)$$

$$\text{Recall } (R) = \frac{TP}{(TP+FN)} \quad (11)$$

$$\text{F-measure (F)} = \frac{2 * Recall * Precision}{Recall + Precision} \quad (12)$$

4.0 Results and Discussion

In this study, students' performance was modeled by LRC and KNN through 10-fold cross-validation in WEKA version 3.8.6 and each classifier was evaluated in what follows. The results obtained were reported in a confusion matrix, in Table 2 and performance metric, in Table 3 respectively.

Table 2: Confusion matrix for LRC and KNN

LRC	Observed			TOTAL
Predicted model		*YES*	*NO*	
	YES	343	11	354
	NO	14	110	124
	TOTAL	357	121	478
KNN	Observed			TOTAL
Predicted model		*YES*	*NO*	
	YES	332	22	354
	NO	26	98	124
	TOTAL	358	120	478

Table 2: Shows how many predictions were correct and incorrect per class label based on LRC and KNN algorithms. Hence, Table 2 suggested that LRC prediction was more accurate than KNN.

Table 3: Metric performance of LRC and KNN

Classifier	Accuracy	Error rate	Precision	Recall	F-measure	ROC Area	Time taken
LRC	94.7699%	5.2301%	96.1%	96.9%	96.5%	97.5%	0.07sec
KNN	89.9582%	10.0418%	92.7%	93.8%	93.3%	85.6%	0.01sec

According to Table 3, The comparison of LRC and KNN was done, and the results showed that LRC was better than KNN classifier in terms of higher accuracy (94.7699 % against 89.9582%), error rate (5.2301% against 10.0418%), Precision (96.1% against 92.7%), Recall(96.9% against 93.8%), F-measure(96.5% against 93.3%), ROC Area (97.5% against 85.6%), except time taken where KNN (0.01sec against 0.07sec) had strength.

5.0 Conclusion

In this study, LRC and KNN classifiers were implemented successfully on WEKA using students' academic performance in computer science. This study suggests that LRC is a better classifier than KNN when tuning parameter k=1 with Euclidean distance used as distance metrics for the given dataset used, considering the higher performances in most of the evaluated metrics, except processing time. In future studies, LRC as a classifier may be compared with other classifiers, using the same data sets or different data sets from various domains, such as healthcare, finance, or social sciences, to assess the generalizability and robustness of LRC.

References

Akuma, S., & Abakpa, H. (2021). Predicting Undergraduate Level Students' Performance Using Regression. *Nigerian Annals of Pure and Applied Sciences, 4*(1), 109-117.

Amra, I. A. A., & Maghari, A. Y. (2017). *Students performance prediction using KNN and Naïve Bayesian.* Paper presented at the 2017 8th International Conference onInformation Technology (ICIT). Pp. 909-919.

Asril, T., & Isa, S. M. (2020). Prediction of students study period using K-Nearest Neighbor algorithm. *International Journal, 8(6), 2585-2593*

Deepika, K., Sathyanarayana, N., & Sathyanarayana, N. (2019). Relief-F and budget tree random forest based feature selection for student academic performance prediction. *International Journal of Intelligent Engineering and Systems, 12*(1), 30-39.

Devasia, T., Vinushree, T., & Hegde, V. (2016). *Prediction of students performance using Educational Data Mining.* Paper presented at the 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE). Pp. 91-95

Ibrahim, S. A., & Fadil, N. A. (2020). A Binary Logistic Regression Analysis of Students' Performance in Computer Science Programme. *Transaction of the Nigerian Association of Mathematical Physics, 12*(1), 69-72

Ibrahim, S. A., Isiaka, K. S., & Mustapha, I. O. (2021). A Comparison of Classification Algorithms on Students' Performance in Computer Science Programme using WEKA. *The Transaction of the Nigerian Association of Mathematical Physics 15,* 43-46.

Sathe, M. T., & Adamuthe, A. C. (2021). Comparative Study of Supervised Algorithms for Prediction of Students' Performance. *International Journal of Modern Education & Computer Science, 13*(1), 1-21.

Shamsi, M. S., & Lakshmi, J. (2016). A Comparative Analysis of classification data mining techniques: Deriving key factors useful for predicting students performance. *arXiv preprint arXiv:1606.05735.*

Shingari, I., Kumar, D., & Khetan, M. (2017). A review of applications of data mining techniques for prediction of students' performance in higher education. *Journal of Statistics and Management Systems, 20*(4), 713-722.

Tripathi, A., Yadav, S., & Rajan, R. (2019). *Naive Bayes classification model for the student performance prediction.* Paper presented at the 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 1, 1548-1553.

Vyas, M. S., & Gulwani, R. (2017). *Predicting student's performance using cart approach in data science.* Paper presented at the 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), 1, 58-61.

Witten, I. H., & Frank, E. (2005). DATA MINING Pratical Machine Learning Tools and Techniques. *Elsevier Inc.*(Second Edition).

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). Data Mining Practical Machine Learning Tools and Techniques *Elsevier Inc.*(Fourth Edition).

Wiyono, S., & Abidin, T. (2019). Comparative study of machine learning KNN, SVM, and decision tree algorithm to predict student's performance. *International Journal of Research-Granthaalayah, 7*(1), 190-196.

Wiyono, S., Wibowo, D. S., Hidayatullah, M. F., & Dairoh, D. (2020). Comparative study of KNN, SVM and decision tree algorithm for student's performance prediction. (IJCSAM) *International Journal of Computing Science and Applied Mathematics, 6*(2), 50-53.

Yaacob, W. F. W., Nasir, S. A. M., Yaacob, W. F. W., & Sobri, N. M. (2019). Supervised data mining approach for predicting student performance. *Indones. J. Electr. Eng. Comput. Sci, 16*(3), 1584-1592.

Yakubu, M. N., & Abubakar, A. M. (2022). Applying machine learning approach to predict students' performance in higher educational institutions. *Kybernetes, 51*(2), 916-934.

Zulfiker, M. S., Kabir, N., Biswas, A. A., Chakraborty, P., & Rahman, M. M. (2020). Predicting students' performance of the private universities of Bangladesh using machine learning approaches. *International Journal of Advanced Computer Science and Applications, 11*(3), 672-679

‘

‘

‘

‘

‘

‘

‘