IJSGS VOL. 8(4), DECEMBER, 2022 https://doi.org/ 10.57233/ijsgs.v8i4.365 URL: https://fugus-ijsgs.com.ng

IJSGS

ISSN: 2488-9229
FEDERAL UNIVERSITY
GUSAU-NIGERIA

INTERNATIONAL JOURNAL OF SCIENCE FOR GLOBAL SUSTAINABILITY

Comparative Between Three Machine Learning Algorithms to Predict and Improve Students' Academic Performance

[1]Bashiru Aliyu Sani, [2]Samaila Baoku I.G, [3]Bashir Jamilu Ahmed and[4]Samaila Musa

[1] Department of Computer Science, Abdu Gusau Polytechnic Talata Mafara Zamfara tate Nigeria,

[2]Department of Mathematical Sciences, Faculty of Physical science, Federal University Dutsin-ma Katsina, State, Nigeria,.

[3]Department of Cyber Security, Faculty of Computer Science and Artificial Intelligence. Federal University Dutsin-ma Katsina State, Nigeria,

[4]Department of Computer Science, Federal University Gusau, Zamfara State, Nigeria

***Corresponding Author's Email**: baliyusani@gmail.com: **Phone NO.:** 2348062477796

Received on: October, 2022 **Revised and Accepted on:** November, 2022 **Published on:** December, 2022

ABSTRACT

The greatest aim of every educational setup is giving the best educational experience and knowledge to the students. Discovering the students who need extra support and guidance so as to carry out the necessary actions to enhance their performance plays an important role in achieving that aim. In this research work, three machine learning algorithms have been used to build a classifier that can predict the performance of the students in higher institutions considering three Tertiary institutions which are: Federal University Dutsinma, Katsina State, Abdu Gusau Polytechnic Talata Mafara and College of Education Maru, Zamfara State. The machine learning algorithms includ: Support Vector Machine, Linear Regression and Stochastic Gradient descent algorithms. The models have been compared using the Mean Absolute Error, Mean Square Error and Root Mean Square Error classification accuracy. The dataset used to build the models is collected based on a survey given to the students and the students' grade book. The support vector machine model achieved the best performance that is equal to 99.1%.

Keywords: support vector machine, Linear Regression and Stochastic algorithm

1.0 INTRODUCTION

1.1 Background of the study

With the wide usage of computers and internet, there are recently a huge increase in public available data that need to be analyzed. Be it online sales information, website traffic, or user habits, data is generated every day. Such a large amount of data presents both a problem and an opportunity. The problem is that it is difficult for humans to analyze such large data. The opportunity is that this type of data is ideal for computers to process, because it is stored digitally in a well-formatted way, and computers can process data much faster than humans.

The concept of machine learning is something born out of this environment. Computers can analyze digital data to find patterns and laws in ways that is too complex for a human to do. The basic idea of machine learning is that a computer can automatically learn from experience (Hussain *et al.*,2019). Although machine learning applications vary, its general function is similar throughout its applications. The computer analyzes a large amount of data, and finds patterns and rules hidden in the data. These patterns and rules are mathematical in nature, and they can be easily defined and processed by a computer. The computer can then use those rules to meaningfully characterize new data. The creation of rules from data is an automatic process, and it is something that continuously improves with newly presented data. Depending on the type of input data, machine learning algorithms can be divided into supervised and unsupervised learning. In supervised learning, input data comes with a known class structure (Ahmad and Shahzadi, 2018). This input data is known as training data. The algorithm is usually tasked with creating a model that can predict one of the properties by using other properties. After a model is created, it is used to process data that has the same class structure as input data. In unsupervised learning, input data does not have a known class structure, and the task of the algorithm is to reveal a structure in the data (Hoffait and Schyns, 2017). To determine

which algorithm provides the best results, Mean Absolute Error, Mean Square Error and Root Mean Square Error were used and the implementation was done on Python.

2.0 Review of Related Works

Waheed *et al.* (2020) designed a model with artificial neural networks on students' records related to their navigation through the LMS. The results showed that demographics and student clickstream activities had a significant impact on student performance. Students who navigated through courses performed higher. Students' participation in the learning environment had nothing to do with their performance. However, he concluded that the deep learning model could be an important tool in the early prediction of student performance.

Xu *et al.* (2019) determined the relationship between the internet usage behaviors of university students and their academic performance and he predicted students' performance with machine learning methods. The model he proposed predicted students' academic performance at a high level of accuracy. The results suggested that Internet connection frequency features were positively correlated with academic performance, whereas Internet traffic volume features were negatively correlated with academic performance. In addition, he concluded that internet usage features had an important role on students' academic performance.

Bernacki *et al.* (2020) tried to find out whether the log records in the learning management system alone would be sufficient to predict achievement. He concluded that the behaviour-based prediction model successfully predicted 75% of those who would need to repeat a course. He also stated that, with this model, students who might be unsuccessful in the subsequent semesters could be identified and supported.

Cruz-Jesus *et al.* (2020) predicted student academic performance with 16 demographics such as age, gender, class attendance, internet access, computer possession, and the number of courses taken. Random forest, logistic regression, k-nearest neighbours and support vector machines, which are among the machine learning methods, were able to predict students' performance with accuracy ranging from 50 to 81%.

Fernandes *et al.* (2019) developed a model with the demographic characteristics of the students and the achievement grades obtained from the in-term activities. In that study, students' academic achievement was predicted with classification models based on Gradient Boosting Machine (GBM). The results showed that the best qualities for estimating achievement scores were the previous year's achievement scores and unattendance. The authors found that demographic characteristics such as neighbourhood, school and age information were also potential indicators of success or failure. In addition, he argued that this model could guide the development of new policies to prevent failure. Similarly, by using the student data requested during registration and environmental factors.

Hussain *et al.* (2019) examine a study to predict student performance based on internal assessment with the use of Artificial Neural Network (ANN) and other algorithms. However, the highest classification accuracy gotten in this study was 95.34% produced by ANN. The evaluation matrices used were Precision, Recall, F-Score, Accuracy, and Kappa Statistics.

3.0 METHODS

This study used Machine Learning techniques to predict students who may be at risk based on available student data. The predictions needed to be early in the semester so appropriate measures can be applied to help at risk students succeed in the course. The prediction task was treated as a binary classification task to determine whether students are in good standing or at risk. The target attribute that was employed to classify students in good standing and at-risk was based on students' CGP in a particular course, where students with a grade worse than "C" was considered at-risk. The Python platform was used to perform this study. Python is a software environment for mathematical and statistical computations and graphics that provides a variety of implementations for machine learning algorithms and numerous data mining tasks. In this study, the Classification and Regression Training package was extensively utilized for the execution of all steps in the research process. The package provides a unified interface that streamlines the

process for training and evaluating predictive models since it contains tools for the preparation, training, evaluation, and visualization of ML models and datasets. figure 1 depicts an overview of the study approach. In the following sections, a description of each step of the research process is presented. To address the common issues of above review literatures such as class imbalance, data hi-dimensionality and classification errors, this study has proposed a model which have following phases**.**

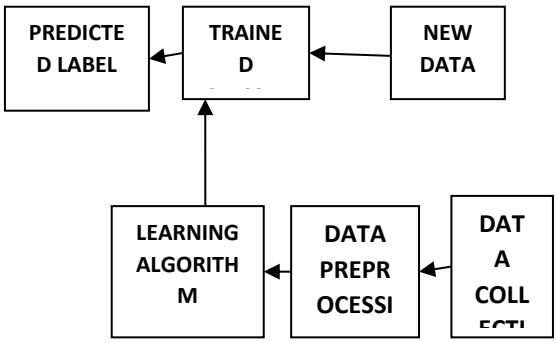

Figure 1: show the main steps of proposed methodology

4.0 RESULT AND DISCUSSION

For our experiments, three classifiers which are Support Vector Machine, Linear Regression and Stochastic Gradient descent classifiers have been evaluated using 10 folds cross validation technique. This technique divides the data set into 10 subsets of equal size; 80% of the subsets are used for training, while 20% is left out and used for testing. The final result is estimated as the average error rate on test. Figure .1 shows the percentage for training and testing.

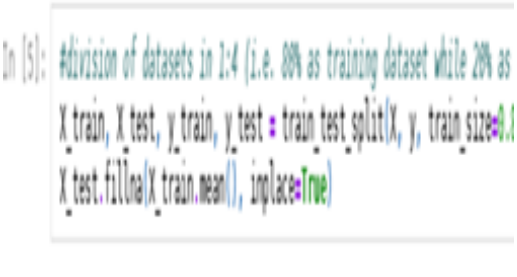

Figure 2: percentage of training and testing

After the implementation and the improvement on better algorithm, the following result was gotten. For the evaluation measure we used Mean Absolute Error, Mean Square Error and Root Mean Square Error.

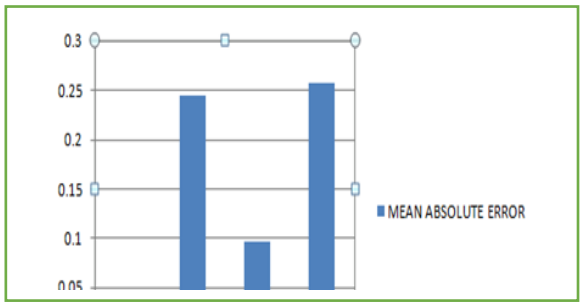

Figure 4: Mean absolute Error for the classifiers

It was observed the Mean absolute error **for** Linear Regression is 0.24, for Support Vector Machine its 0.009 and SGD Regression shows 2.26, therefore, it means SVM has the lowest number of mean absolute error which indicates that the accuracy will be higher that the two other classifiers.

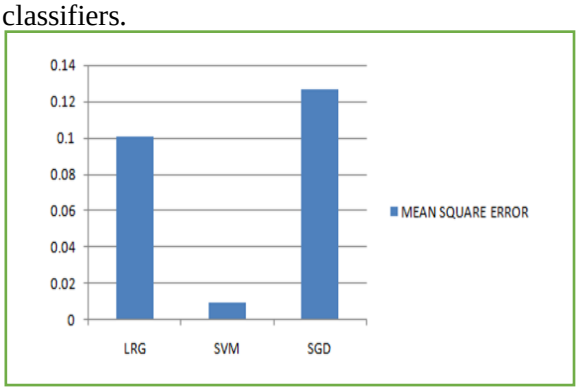

Figure 3: Mean Square Error on the classifiers

From Figure 3, Linear Regression shows 0.100, SVM shows 0.001 while SGD regression shows 0.126, by indication SVM shows the lower error in mean square error, which means its own accurate prediction will be higher and be the best.

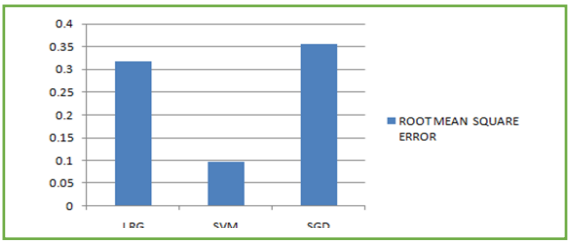

Figure 5: Root Mean Square Error for the classifiers

Fig .4 indicates that Linear Regression has 0.32, SVM has 0.09 and SGD has 0.36, this implies that SGD has the highest root mean square error, followed by Linear Regression, while SVM show the least, that means that the prediction is better.

5.0 RESULT ANALYSIS

In the experiment, three classification algorithms Namely: (Linear Regression, Support Vector Machine and Stochastic Gradient Descent) were simply executed all together on dataset. In the light of results shown in graphical representation of Figure 5, it has been observed that the highest accuracy achieved by Support Vector Machine which was not enough but was improved and 99.1% was realized as the accuracy as compared to previous studies and lowest accuracy has achieved by Stochastic Gradient Descent Regression which is 88.1%.

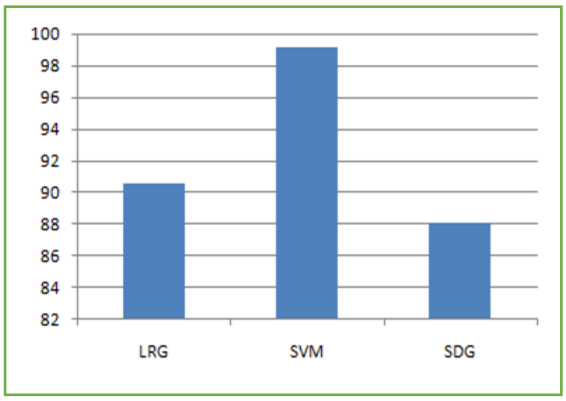

Figure 6: Classifier Accuracy comparative

6.0 CONCLUSION

To predict student's performance in advance is a very important issue. We concluded after deep studies that various datasets of student provides different results with different attributes. This is the reason that the results are vary with different evaluation measures like Mean Square Error, Root Mean square error etc. It was concluded after these studies that every algorithm result are varied according to the dataset and variable attribute used for prediction. However, if we use the Linear Regression, Support Vector Machine and Stochastic Gradient Descent, according to our requirements these algorithms provide extra ordinary accurate results for future prediction and help in the betterment of education system. However, Support Vector Machine algorithm performed better which gave 99.1%.

REFERENCES

Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). predicting academic performance of students from vle big data using deep learning models. computers in human behavior, *104*(October 2019), 106189. https://doi.org/10.1016/j.chb.2019.106189

Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. Computers in Human Behavior, 98(January), 166–173. https://doi.org/10.1016/j. chb.2019.04.015

Bernacki, M. L., Chavez, M. M., & Uesbeck, P. M. (2020). Predicting achievement and providing support before STEM majors begin to fail. Computers & Education, 158(August), 103999. https://doi.org/10.1016/j.compedu.2020.103999.

Cruz-Jesus, F., Castelli, M., Oliveira, T., Mendes, R., Nunes, C., Sa-Velho, M., & Rosa-Louro, A. (2020). Using artifcial intelligence methods to assess academic achievement in public high schools of a European Union country. Heliyon. https://doi.org/10.1016/j.heliyon.2020.e04081

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research, 94(February 2018), 335–343. https://doi.org/10.1016/j.jbusres.2018.02.012.

Hofait, A., & Schyns, M. (2017). Early detection of university students with potential difculties. Decision Support Systems, 101(2017), 1–11. https://doi.org/10.1016/j.dss.2017.05.003.

Hussain, S., Muhsin, Z. F., Salal, Y. K., Theodorou, P., Kurtoğlu, F., & Hazarika, G. C. (2019). Prediction Model on Student Performance based on Internal Assessment using Deep Learning. International Journal of Emerging Technologies in Learning, 14(8), 4-22.

Ahmad, Z., & Shahzadi, E. (2018). Prediction of students' academic performance using artificial neural network. Bulletin of Education and Research, 40(3), 157–164