 doi:10.20944/preprints202207.0088.v1

A Machine Learning Prediction of Academic Performance of Secondary School Students Using Radial Basis Function Neural Network

[1,2]Olusola Aanu OLABANJO and [3]Ashiribo Senapon WUSU

[1]*Department of Mathematics, Morgan State University, MD, USA*

[2]*Department of Computer Science, Lagos State University, Lagos Nigeria*

[3]*Department of Mathematics, Lagos State University, Lagos Nigeria*

Abstract

Introduction: Academic success is primary goal of every student. It is described as the extent to which a student has successfully achieved his or her short and long-term educational goals. Several factors have been established to predict academic performance of students. Machine learning techniques have been employed in predicting students' performance, but it has not been prevalent in developing countries like Nigeria and most studies did not consider class teachers' end-of-the-year rating.

Aim*:* The aim of this work is to develop a Radial Basis Function Neural Network (RBFNN) for prediction of secondary school students' performance.

Materials and Methods*:* We obtained data from school repository containing students' raw score and classteachers' rating from year one to year six. The data was labelled into pass or fail given the actual outcome of their examinations. Subjects were categorized into Mathematics, English and major, depending on the student's specialization. Class-teachers' ratings were also included in the dataset. The preprocessed dataset was used to train the RBFNN model. The impact of Principal Component Analysis (PCA) was also measured.

Results: We set up four experiments in order to achieve our aim. The best result gave the sensitivity of 93.49%, specificity of 75%, accuracy of 86.59% and an AUC score of 94%. Other experiments gave a relatively low performance.

Conclusion: This study helps students to get a projection of academic success even before sitting for the examination. This will also help parents and counsellors in knowing the direction of their counseling to each student. Teachers and parents should pay attention to class teacher ratings of the students as this is discovered to affect the prediction accuracy of their examination success.

Keyword: Academic performance; Machine Learning; RBFNN

INTRODUCTION

The primary goal of a student in any school is to achieve success. Academic performance is described as the extent to which a student has successfully achieved his or her short and long-term educational goals[1, 2]. Completion of academic benchmarks such as secondary school certificates, bachelor's degrees or numerical figures of Cumulative Grade Point Average (CGPA) altogether represent academic performance or achievement. Academic performance is a common-ground measurement through examinations and continuous assessments either via procedural knowledge such as skills or declarative knowledge such as facts[3].

Studies have shown that several factors have a significant impact on students' academic progress in secondary and postsecondary institutions. Individual differences such as academic intelligence and personality, non-cognitive characteristics such as attitudes, behaviors, and personal tactics, motivation, self-control, and students' participation in extracurricular activities are some of the established factors which affect students' performance at various levels in schools[4, 5].

Secondary school is a place of learning which produces the main manpower of a nation. National development, research, professional training, leadership and empowerment, policy formulations as well as technological advancements are evident manifestations of the existence of strong secondary learning environment [6]. In Nigeria, the rate of University dropouts is very high and this is blamed on several contributory factors one of which is the quality of secondary education [7]. The system of admission into the University is fair and, more importantly, robust enough to select only those who are qualified [8].

Educational data mining is a current trend and interdisciplinary study topic concerned with the development of ways to explore data originating in a learning environment. Educational data mining (EDL) has been used for analysis and visualization of educational data, providing feedback or reports to instructors, providing efficient recommendations for students, predicting student performance, performing student modeling, detecting undesirable student behavior, efficiently grouping students, hidden network analysis, planning and scheduling, and the development of robust courseware, among other applications. [9].

Although evaluating data from educational settings is not new, recent technological advancements, increasing computing power, and the capacity to log fine-grained data have led to a greater focus on the development of tools for analyzing vast amounts of data. The four cardinal goals of EDM are: predicting future learning behavior in students, developing or enhancing domain models, investigating the effects of educational support and advancing scientific understanding of learning and learners[10]. Studies also reveal that the goal of EDM is to turn raw data into relevant information about the learning process using data mining for education in order to make better judgments[9]. EDM, however, faces a long-term criticism vis-à-vis generalizability of outcomes, individual privacy, plagiarism and acceptance[9-12].

SECONDARY EDUCATION SYSTEM IN NIGERIA

Secondary education is an important stage in Nigeria's educational system. It serves as a bridge between primary and University education. It is the type of education that students receive after they have completed their primary school and/or before they begin their University education. It is designed for students in ages 11 through 17. Secondary education provides the foundation for the discovery and classification of specific categories of professions, as well as the training ground for future professionals. Secondary education in Nigeria was only five years long before independence to 1982. Those who earned the needed requirements after five years were eligible for the two-year Higher School Certificate, which qualified them for university studies.

The secondary education as currently practiced in Nigeria is stratified into Junior and Senior categories each expected to run for three years. The Junior category teaches students in all subjects; at the end of which the student chooses a department among or is allocated to one of Science, Commercial or Arts departments. The Science department focuses on teaching Physics, Chemistry, Agricultural Science and Biology as core subjects. The Commercial department immerses their students in Commerce, Financial Accounting and Economics. Arts department has Government,

 doi:10.20944/preprints202207.0088.v1

History and Literature-in-English. All three departments have English, Mathematics and one Nigerian Language as compulsory across board.

At the third year in Senior Secondary category, all students are expected to write West African Examinations Council (WAEC) examinations which requires that a student passes English and Mathematics and three other core subjects. This is what qualifies them for entrance examination into tertiary institutions.

RELATED WORKS

Prediction of student academic performance is an important aspect of research in many academic disciplines including education, social sciences, mathematics and computer science. Many works have been done on the use of computational and statistical methods to predict students' performance at various school levels. For instance, decision trees, random forests, neural networks and support vector machine have been used to predict some Portuguese students' performance using such factors as student's previous grade, demographic social and school-related features, attendance frequency and alcohol consumption [13]. Working status, position in family birth, scholarship opportunity and parent's income status were presented as direct indicators of students' performance using the decision tree technique [14]. Research has also shown that social behavior of students can, to a reasonable extent, predict their performance [12, 15] .

Studies also showed that social network computing can be used to identify student's most important attributes from available data which predict their academic performance. Naïve Bayes, neural network and decision tree used on students' dataset proved very efficient with the first technique emerging as the major predictor (with 86% accuracy) in predetermining student performance [16]. The authors in [17] successfully predicted success or failure of students in tertiary institutions using classification techniques and regression. It was found out that SVM and decision trees were used in classification and Random Forest for regression. By quantitative research, studies have been carried out to affirm that the student dropout probability can be predicted even if the model operated on single data [18].

Literature consulted in this work shows that much work has not been sufficiently done in Nigeria's context and in predicting students' performance at secondary level. Also, most of the works available do not consider the behavioral, cognitive and psychomotor ratings of the teachers concerning the students as predicators of students' academic performance. This study considers the performance of secondary school students from entry to final year including their average cognitive and psychomotor rating for each year to predict if they will pass their unified examination (with minimum of six credits including Mathematics and English) and if they will pass their tertiary entrance examination (Joint Admissions and Matriculation Board - JAMB) with a minimum score of 180 out of the maximum attainable score of 400. Included in the features are the classteachers' rating of the students cognitive and psychomotor skills.

MATERIALS AND METHODS

Study Population

This study investigated the results database of 1927 students in a secondary school located in Okokomaiko axis of Lagos State. Results contained subjects (courses) undertaken by the students

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 July 2022 doi:10.20944/preprints202207.0088.v1

from entry to graduation years. The six year results consists of raw scores of students in subjects such as Mathematics, English, local language subject (French, Yoruba, Igbo or Hausa) and other subjects depending on the students' major. This comprises Science, Commercial and Arts Majors. There was no direct contact with human participants, and there is no institutional board or equivalent board that gives approval of data driven study in the school. In view of this, there was no ethical approval for the study. Additionally, we did not obtain informed consent from participants because the study was conducted on the data available.

Data Description

Since the data came in its very raw form, there was need for preprocessing. The data was first anonymized and then converted to a tidier format using the standard grading system for secondary school as approved by the central education body in Nigeria. The grading system as encoded is given as follows: 0-39% = 0, 40% -44% = 1, 45%-49% = 2, 50%-59% =3, 60% -69% =4, 70% - 100% = 5. The final features of our preprocessed dataset are presented in Table 1.

Table 1: Features of Datasets Used in this Study

Features Category	Number of Features	Description
MATH	6	Average score in Mathematics for each year from year 1 to 6
ENG	6	Average score in English for each year from year 1 to 6
MAJOR	6	Average score in 3 major subject for each year from year 1 to 6
PSYCH	6	Average psychomotor rating for each year from year 1 to 6
COG	6	Average cognitive rating for each year from year 1 to 6
TARGET	1	An indicator (0/1) to indicate if the student has six credit passes or more including Mathematics and English in the final examination (West African Examination Council - WAEC).
	TOTAL: 31	

Model Building

In this study, a supervised deep learning model (known as the Radial Basis Function Neural Network) which predicts a student's final grade in secondary education is proposed. This model uses the students' psychomotor ratings, average score in Mathematics, English and other major subjects to predict if s/he would pass WAEC examination. The framework of the proposed model is given in Figure 1.

 doi:10.20944/preprints202207.0088.v1

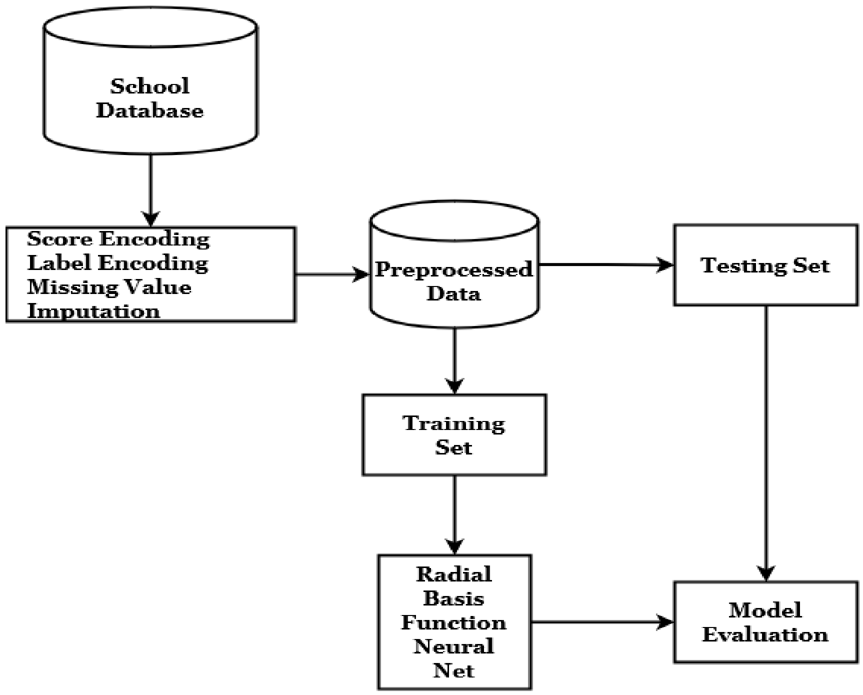

Figure 1: Framework of the Proposed Model

The major activities in this study include data collection, data preprocessing, data splitting and cross validation, model training and model evaluation.

a. ***Data Collection:*** Students' raw scores were extracted from the school's database. The relational database maintained by the school contains the raw scores of students in various subjects from year 1 to 6. In addition, the teachers' ratings for each student were collected.
b. ***Data Preprocessing:*** The data collected went through some stages of preprocessing. The identity of the students was anonymized, and the average raw scores were encoded using Nigeria's grading system. The average rating of the classteachers were also collected for each year.
c. ***Data Splitting:*** The dataset was divided into training and testing and k-fold cross validation was done on the training set to reduce bias and avoid overfitting[19, 20].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 July 2022 doi:10.20944/preprints202207.0088.v1

Figure 2: 10-fold Cross Validation of the Proposed Model

In this study, we used the k-fold cross-validation with k=10 (see Figure 2). K-fold cross-validation improves machine learning model by checking iteratively whether the test performance varies based on which samples you used to train / test. By running through the train/test comparison several times we can get a better estimate of model performance, and sanity check that the model is not performing wildly differently after being trained on different segments of our dataset.

d. **The Radial Basis Function Network (RBFN):** This a special type of feedforward neural network in which radial basis functions are used as activation functions. RBFN can be used for classification, regression, and time-series prediction and have an input layer, a hidden layer, and an output layer[21-23].

   An important property of the RBF is that multi-dimensional space nonlinearity, such as the relation of subjects and psychomotor ratings, can be taken to be a linear combination of the nonlinear RBF. A common feature of the RBFNN is its fast training as compared with the backpropagation networks [24]. Figure 3 illustrates the architecture of an RBFNN. Three layers are involved in RBFNN architecture, and each layer is made up of several nodes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 July 2022 doi:10.20944/preprints202207.0088.v1

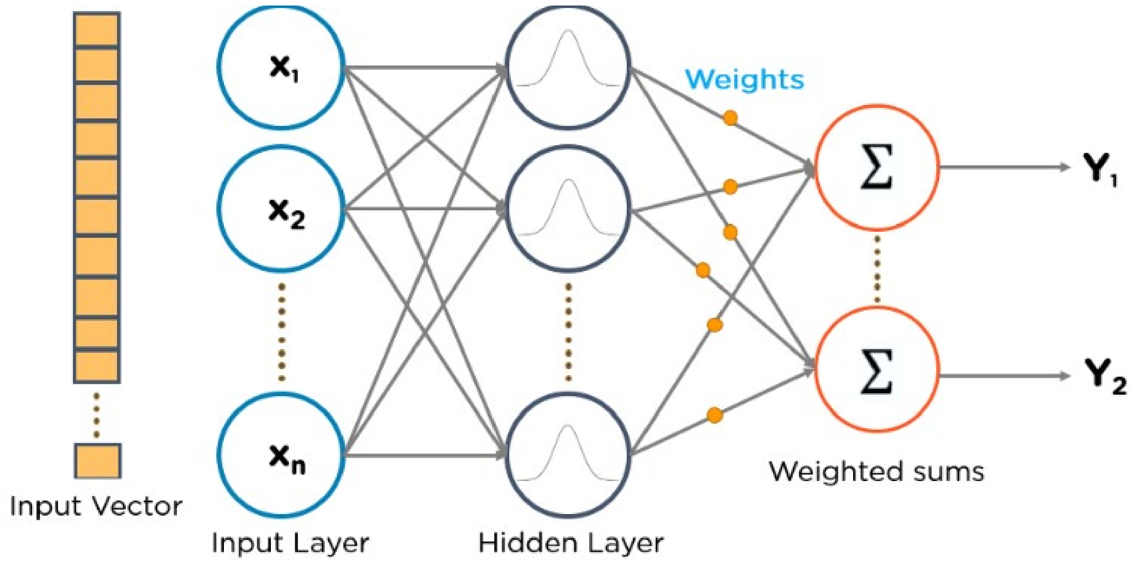

Figure 3: The Structure of RBFNs (adopted from https://www.simplilearn.com/tutorials/deep-learning-tutorial/deep-learning-algorithm)

The input layer (x_1, x_2 ... x_{30} nodes) introduces the outside information into the network. The only hidden layer, which has j nodes, processes the input information with a nonlinear transformation. The transformation associated with each node of hidden layer implemented herein is a Gaussian function defined as

$$Z_j(x) = e^{\left(-\frac{d_j^2}{2\sigma_j^2}\right)} \quad \ldots\ldots(1)$$

$$d_j = \left\|x - \mu_j\right\| \quad \ldots\ldots(2)$$

where $Z_j(x)$ is the output of the j^{th} hidden node and bias node, $Z_0(x)$ is fixed at l; x is the input with n dimension; σ_j is the width of the receptive field of the j^{th} hidden node; μ_j is the center of the j^{th} hidden node; and d_j is the Euclidean distance between x and μ_j. The third layer, the output layer with Y nodes, fully interconnects to each hidden node. The output of the network is a linear combination of the nonlinear radial basis functions.

RBFNs perform classification by measuring the input's similarity to examples from the training set. They have an input vector that feeds the input layer and a layer of RBF neurons. The function finds the weighted sum of the inputs (i.e. features 1 to 30), and the output layer has one node per category or class of data (Y_1 for pass and Y_2 for fail). The neurons in the hidden layer contain the Gaussian transfer functions, which have outputs that are inversely proportional to the distance from the neuron's center. The network's output is a linear combination of the input's radial-basis functions and the neuron's parameters.[25, 26].

e. **Model Evaluation:** The proposed model was tested with the testing test and its performance was measured using sensitivity, specificity, accuracy and Area Under Curve (AUC). These metrics are calculated from the confusion matrix generated by the model testing. The confusion matrix from two-classed dataset consists of four values: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). The calculations are given in equations 3, 4 and 5.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (4)$$

$$Specificity = \frac{TN}{TN+FP} \quad (5)$$

f. Experimental Setup: In this study, four experiments were set up in order to give meaning to the dataset. These are given in Table 2.

Experiment #	*Description*	*Technique*
A	Full Features with Psychomotor Ratings	RBFNN
B	Full Features without Psychomotor Ratings	RBFNN
C	Feature Extraction with Psychomotor Ratings	RBFNN after Principal Component Analysis (PCA)
D	Feature Extraction without Psychomotor Ratings	RBFNN after Principal Component Analysis (PCA)

Feature extraction is a good way of reducing bias and avoid both underfitting and overfitting in a machine learning model. It is usually a preferred feature engineering technique because it reduces redundant features into the important ones only without losing any relevant information[27-29]. The choice of feature extraction technique is based on the known strengths of PCA[30].

RESULTS

After collection, the dimension of the dataset was 1927 x 31. Mean imputation was done for missing ratings. The preprocessed dataset has 1723 x 31 dimension. The percentage of those who passed to those who failed was 68% and 32% respectively. 61% were male and 59% were females. Table 2 shows the descriptive statistics of the data used to train the model. Our PCA-experiment reduced the dataset to 1723 x 17 dimension. The performance of the RBFNN in the four experimental setups are given in Table 3.

Table 3: Performance of RBFNN in the Experiments

Exp. #	*TP*	*TN*	*FP*	*FN*	*Accuracy*	*Sensitivity*	*Specificity*	*ROC(AUC)*
A	182	82	51	28	76.97	86.67	61.65	0.83
B	167	87	66	23	74.05	87.89	56.86	0.87
C	201	96	32	14	86.59	93.49	75.00	0.94
D	194	90	39	20	82.80	90.65	69.77	0.89

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 July 2022 doi:10.20944/preprints202207.0088.v1

The results of the various experiments show significant differences in the performance of the proposed model. Figure 4 is the pictorial representation of the results as shown in Table 3.

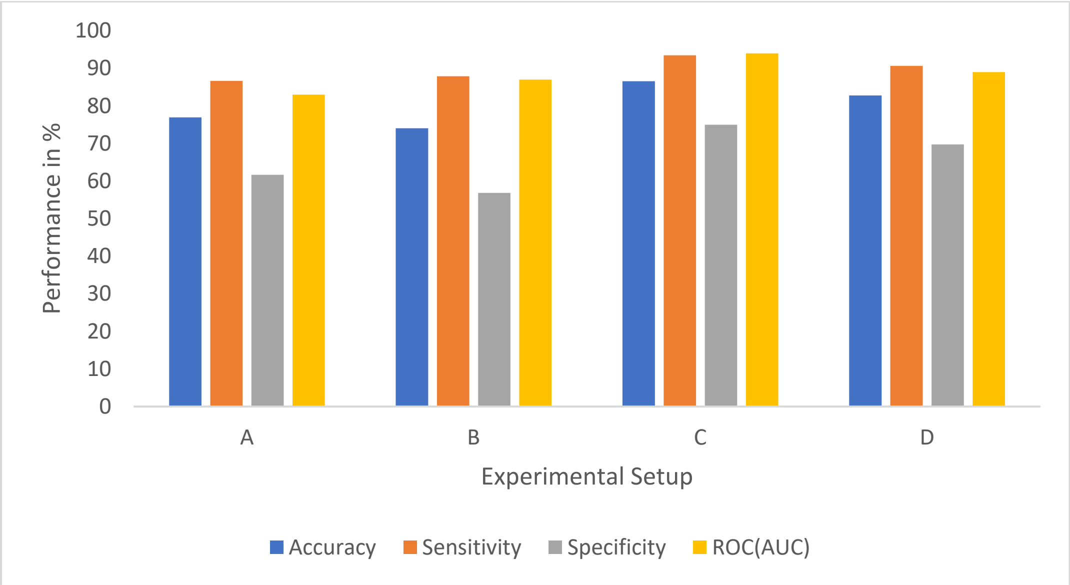

Figure 3: Model Performance Under Experiments A, B, C, D

The AUC as presented provides an aggregate measure of the performance of the proposed model across all possible classification thresholds. One way of interpreting AUC is that is the probability measure that the model ranks a random positive feature vector more highly than negative.

DISCUSSION

The aim of this study is to investigate the performance of the proposed RBFNN model in predicting if students at a secondary school will pass their final board examination using their six-year performance as predictors. One step further is the inclusion of the students' six years psychomotor ratings to measure the effectiveness of their class teachers' observations over time.

In this work, we discovered that student's performance can be predicted using a machine learning classifier. Student's performance while in school can predict the success of their board examinations. Also, we discovered that while, in this part of the world, parents only pay attention to their academic grades of their wards, it is important to also pay attention to their end-of-year psychomotor ratings[31-33]. Experiments B and D supports this argument.

It was also noticed that feature extraction enhanced the performance of our proposed model. The PCA-reduced feature set gave a relatively better accuracy, sensitivity, specificity and computation time. This is in conformity with the reviewed literature[34-36].

One impact of this study is the fact that students can get a projection of academic success even before sitting for the examination. This will also help parents and counsellors in knowing the direction of their counseling to each student. Another strength is the fact that teachers and parents should pay attention to class teacher ratings of the students as this is known to add to the prediction accuracy of their examination success. A good test of if a student will pass the examination is

 doi:10.20944/preprints202207.0088.v1

reflected in the sensitivity analysis. A result as high as 93% shows that this model is reliable in predicting success than predicting failure.

Some limitations have been identified in this work which we hope to consider in the future. The first is the fact that our model fed on 1723 observations. It is known that a model is more reliable with more data than with less. In the future, we hope to harvest more data for more reliability. Also, we hope, in the future, to finetune the model and data to improve the performance of the model vis-à-vis sensitivity, specificity and AUC score.

CONCLUSION

Radial Basis Function Neural Network has been proposed in this study for predicting students' success in their final board examinations. The impact of class teachers' end-of-year rating has also been investigated. We also reported the impact of feature extraction on the performance of our model. The strengths and weaknesses in the study have also been discussed. Recommendations have been presented in the study on how parents and teachers can know potentially weak students in order to better prepare them for their final examinations. This study will aid better performance if the proposed machine learning system is put to use.

 doi:10.20944/preprints202207.0088.v1

References

1. Lavin, D.E., *The prediction of academic performance.* 1965.
2. Camacho-Morles, J., et al., *Activity achievement emotions and academic performance: A meta-analysis.* Educational Psychology Review, 2021. **33**(3): p. 1051-1095.
3. Iglesias-Pradas, S., et al., *Emergency remote teaching and students' academic performance in higher education during the COVID-19 pandemic: A case study.* Computers in Human Behavior, 2021. **119**: p. 106713.
4. Echeverría-Ramírez, J.A. and C. Mazzitelli, *A study of the perception of the institutional factors that influence the academic performance of students of the Distance State University of Costa Rica.* Revista Electrónica Educare, 2021. **25**(2): p. 326-344.
5. Kumar, S., M. Agarwal, and N. Agarwal, *Defining And Measuring Academic Performance of Hei Students-A Critical Review.* Turkish Journal of Computer and Mathematics Education (TURCOMAT), 2021. **12**(6): p. 3091-3105.
6. Srikatanyoo, N. and J. Gnoth, *Country image and international tertiary education.* Journal of Brand Management, 2002. **10**(2): p. 139-146.
7. Bature, I.J., *Productive pedagogies for reforming secondary school mathematics classroom practice in Nigeria*. 2014, Curtin University.
8. Aribisala, B.S., et al., *Comparative analysis of university matriculation examination and post university matriculation examination admission models in Lagos State University.* Ethiopian Journal of Science and Technology, 2018. **11**(2): p. 129-143.
9. Romero, C. and S. Ventura, *Educational data mining: a review of the state of the art.* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2010. **40**(6): p. 601-618.
10. Baker, R.S. and K. Yacef, *The state of educational data mining in 2009: A review and future visions.* JEDM| Journal of Educational Data Mining, 2009. **1**(1): p. 3-17.
11. Ahmad, F., N.H. Ismail, and A.A. Aziz, *The prediction of students' academic performance using classification data mining techniques.* Applied Mathematical Sciences, 2015. **9**(129): p. 6415-6426.
12. Baradwaj, B.K. and S. Pal, *Mining educational data to analyze students' performance.* arXiv preprint arXiv:1201.3417, 2012.
13. Cortez, P. and A.M.G. Silva, *Using data mining to predict secondary school student performance.* 2008.
14. Quadri, M.M. and N. Kalyankar, *Drop out feature of student data for academic performance using decision tree techniques.* Global Journal of Computer Science and Technology, 2010.
15. Bayer, J., et al., *Predicting Drop-Out from Social Behaviour of Students.* International Educational Data Mining Society, 2012.
16. Mueen, A., B. Zafar, and U. Manzoor, *Modeling and predicting students' academic performance using data mining techniques.* International Journal of Modern Education and Computer Science, 2016. **8**(11): p. 36.
17. Strecht, P., et al., *A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance.* International Educational Data Mining Society, 2015.
18. Aulck, L., et al., *Predicting student dropout in higher education.* arXiv preprint arXiv:1606.06364, 2016.
19. Refaeilzadeh, P., L. Tang, and H. Liu, *Cross-validation.* Encyclopedia of database systems, 2009. **5**: p. 532-538.
20. Stone, M., *Cross-validation: A review.* Statistics: A Journal of Theoretical and Applied Statistics, 1978. **9**(1): p. 127-139.

 doi:10.20944/preprints202207.0088.v1

21. Mai-Duy, N. and T. Tran-Cong, *Approximation of function and its derivatives using radial basis function networks.* Applied Mathematical Modelling, 2003. **27**(3): p. 197-220.
22. Park, B., C.J. Messer, and T. Urbanik, *Short-term freeway traffic volume forecasting using radial basis function neural network.* Transportation Research Record, 1998. **1651**(1): p. 39-47.
23. Mustafa, M., et al., *Prediction of pore-water pressure using radial basis function neural network.* Engineering Geology, 2012. **135**: p. 40-47.
24. Hassoun, M.H., *Fundamentals of artificial neural networks*. 1995: MIT press.
25. Dua, M., et al., *Biometric iris recognition using radial basis function neural network.* Soft Computing, 2019. **23**(22): p. 11801-11815.
26. Lin, G.-F. and L.-H. Chen, *A non-linear rainfall-runoff model using radial basis function network.* Journal of Hydrology, 2004. **289**(1-4): p. 1-8.
27. Guyon, I., et al., *Feature extraction: foundations and applications*. Vol. 207. 2008: Springer.
28. Nevatia, R. and K.R. Babu, *Linear feature extraction and description.* Computer Graphics and Image Processing, 1980. **13**(3): p. 257-269.
29. Sachar, S. and A. Kumar, *Survey of feature extraction and classification techniques to identify plant through leaves.* Expert Systems with Applications, 2021. **167**: p. 114181.
30. Uddin, M.P., M.A. Mamun, and M.A. Hossain, *PCA-based feature reduction for hyperspectral remote sensing image classification.* IETE Technical Review, 2021. **38**(4): p. 377-396.
31. Rourke, B.P. and J.D. Strang, *Neuropsychological significance of variations in patterns of academic performance: Motor, psychomotor, and tactile-perceptual abilities.* Journal of pediatric psychology, 1978. **3**(2): p. 62-66.
32. Ramnarain, U. and S. Ramaila, *The relationship between chemistry self-efficacy of South African first year university students and their academic performance.* Chemistry Education Research and Practice, 2018. **19**(1): p. 60-67.
33. Shephard, R.J., *Curricular physical activity and academic performance.* Pediatric exercise science, 1997. **9**(2): p. 113-126.
34. Kuncheva, L.I. and W.J. Faithfull, *PCA feature extraction for change detection in multidimensional unlabeled data.* IEEE transactions on neural networks and learning systems, 2013. **25**(1): p. 69-80.
35. Park, M.S., J.H. Na, and J.Y. Choi. *PCA-based feature extraction using class information*. in *2005 IEEE International Conference on Systems, Man and Cybernetics*. 2005. IEEE.
36. Jade, A., et al., *Feature extraction and denoising using kernel PCA.* Chemical Engineering Science, 2003. **58**(19): p. 4441-4448.