Educational Data Mining (EDM) for Predicting Students' Performance Using K-Nearest Neighbour Technique"

Mustapha I. O. & Mustapha M. L.
Department of Computer Sciences
Faculty of Natural Sciences,
Al-Hikmah University
Ilorin, Nigeria.
Email: salnet2002@ alhikmah.edu.ng

Adewole K.S.
Department of Computer Science
Faculty of Communication and Information Sciences
University of Ilorin
Ilorin Nigeria.
Email: adewole.ks@unilorin.edu.ng

ABSTRACT

Education tends to be the bed rock for the development of any country and the need to plan for future is an essential ingredient for the achievement of many educational objectives. Therefore, based on the current data, this research work is targeted towards prediction of the number of graduating student for different classes of degree that is to be awarded in Alhikmah University so as to aid academic planner of the university in decision making processes, improve students" academic performance, trim down failure rate and to monitor student behavior. Therefore, Educational data mining (EDM) and students' performance prediction was carried out using the technique of K-Nearest Neighbour which is an instance based learning for classification based on a similarity measure. Dataset of computer science students of Al-Hikmah University was used. Consequently, K-NN base model performed better than other model used in this study with 75% accuracy.

Keywords: Education, Data Mining, Prediction, Students' Performance & K-Nearest Neighbour Technique Models.

iSTEAMS Conference Proceedings Paper Citation Format
Mustapha, I.O., Mustapha, M.I. & Adewole, K.S. (2018): Educational Data Mining (EDM) for Predicting Students' Performance Using K-Nearest Neighbour Technique" Proceedings of the 14th iSTEAMS International Multidisciplinary Conference, AlHikmah University, Ilorin, Nigeria, Vol. 14, No 2 Pp 229-238

1. INTRODUCTION

The development of most country is traceable through the standard of education possed by the inhabitant, this is to say that education tends to determine the economic, social, political, spiritual and moral standard of a nation. Cognitive domain is a vital aspect of education that can be measure through academic performance, but such measure is only viable if it actually represent the exact academic and mental ability of the student or candidate. Most business organization today, rely on certificate with the assumption that it is representing the academic ability of the individual . The use of certificate for the representation of the mental ability of individual may be deceiving atimes because there are some factors that tends to affect the measure of academic performance. Examination and some other tools which are used to measure the academic performance of the university students can be affected by various factors and consequently if the tool to be used to measure is affected then meaningless or vague decision as per the cognitive ability of student may arise, and as such, certificate will not be able to represent individual mental ability. Kabakchieva (2012) analyzed about twenty (20) factors that affect student performance in university degree examinations, including gender, birth year, birth place, living place and country, type of previous education, profile and place of previous education, total score from previous education, university admittance exam and achieved score, total university score at the end of the first year, etc.

Jia and Mareboyana (2013) identified total credit hours (TCH) taken as one of the main factors that affect students retention ability. Acharya and Sinha (2014) also analysed a number of factors that affect student performance which include gender, caste, religion, family size, board, state of origin, family income, board marks, study hours per day, attendance, mid semester score, student's medium of study, type of secondary school attended and private tuition. All these factors has correlation with students performance in the university, hence the focus of this study is to implement predictive system for student academic performance. The implementation will be carried out through the use of data mining technique via K-NN classifier to predict the main attributes that may affect the student"s performance in Al-Hikmah University. This is paramount to elicit which attribute may indirectly disallow students certificate to represent their mental ability.

2. DATA MINING AND EDUCATION RESEARCH:

Data mining refers to the set of computational methods that extract important patterns from original data. It is factual that high percentage of real life phenomena from which huge data are collected, generates data following a particular trend although some might be schiotasic in nature. Therefore, knowledge can be derive from most dataset through the use of some techniques refers to as data mining. Knowledge Discovery from Databases (KDD) is a cyclic and interactive processes that as to do with automatic, exploratory data analysis and modeling of large data sources. Katare and Dubey (2017) described data mining as the foundation of the KDD process, relating the linking of algorithms that search the data, build up the model and determine previously unknown patterns. KDD process is concerned with manipulation of immense data, scaling algorithms for better presentation, appropriate analysis of retrieved information, and human interaction with the overall process.

Educational Data Mining is a rising authority, that is concerned with development of methods or pattern from unique types of data that is available from educational settings, and the use of those methods for better understanding of students, and achievement of educational goals. In other word, educational data mining as to do with the development of methods or pattern through computational approaches on data that emerge from learning processes and educational settings for better achievement of educational goals.

EDM technique can be used to elicit hiding pattern and the relationship between various factors of educational settings together with human and environmental factors. Within EDM a section for analysis of students cognitive outcome based on classification approach study and the usage of statistical algorithm to position students' score data according to their height is the unique domain of this study. Researchers effort in the area of application of data mining to education cannot be overemphasis. As an instance, standard large database were constantly subjected to appropriate data mining technique so as to expose hiding information and thus providing information for future improvement. Different kind of data mining technique have been applied to educational data mining problem area, such technique include rule learner, decision tree, neural network, and k-nearest neighbor (Kabakchieva, 2012), logistic regression, naive bayes and SVM (Jia & Mareboyana, 2013). Asif, Merceron, and Pathan, (2014) used pre-university grades and first year grades of students for students' performance prediction. Abdullahi, Malibari, and Alkhozae, (2014). used multiagent technique to predict students' performance.

The research noted and identify the fact that Prediction of students' performance is more beneficial for identifying the low academic performance students and that student retention as some correlation with academic performance. Muralidharan, Pravien and Balaji, (2017) investigate the issue of limited data size in relation to students result prediction.

3. METHODOLOGY & KNN CONCEPT

3.1 Dataset Description:

The independent variables that were recognized as predictive attributes includes UTME score, five WAEC grades (English, Mathematics, First Science Subject, Second Science Subject and Third Science subject). These data were sourced from the academic section of the school management for the last two current years, and were used for prediction of final degree of class for the student (Dependent Variable). The dependent variable can be any of the target class, which is also expected to be any of the final degree class (First Class, Second Class, and Third class).

Sourced data of students' pre-university results that was used consist of forty-eight (48) students' records in a computer science department of the university. The pre-processed dataset contains 6 fields representing six predictive features and one target class. The pre-processed dataset contains 48 records. Table (I) shows a sample of the dataset having five columns. The second, third, fourth, fifth, sixth and seventh column in table represents the normalized values of UTME score, English-WAEC grade, Mathematics-WAEC grade, First Science Subject-WAEC grade, Second Science Subject-WAEC grade and Third Science subject-WAEC grade. The last column represents the target predicted class. The value 1 represents First Class degree, 2 for Second Class degrees and 3 for Third Class degree respectively.

The data fields of the last five attributes are converted to numeric format using the representation; 5 for A, 4 for B, 3 for C, 2 for D, 1 for E and 0 for F respectively. The sample of normalized dataset is shown in Table I.

Table I: Sample of Raw dataset

S/N	UTME Score	English-WAEC	Maths-WAEC	First Science subject-WAEC	Second Science subject-WAEC	Third Science subject-WAEC	Target Class
CS001	185	3	3	3	3	3	2ND
CS003	170	3	3	3	4	5	3RD
CS005	190	3	4	3	4	2	2ND
CS006	180	3	2	3	3	3	3RD
CS007	184	5	3	3	4	5	2ND
CS008	175	3	4	4	3	4	3RD
CS011	190	3	3	3	3	4	2ND
CS012	187	3	4	3	4	3	2ND
CS014	193	3	3	4	4	4	1ST
CS016	186	3	3	3	3	3	2ND

3.2 Data Preparation and Feature Selection:

Data preparation is carried out by eliminating outliers in the dataset and filling up missing values. Most of the data sourced were either in quantitative or qualitative format, therefore discretization of the dataset is performed to achieve a useable format for the machine learning. The normalization technique used in the pre-processing is Max-min normalization technique because of limited range in the data and less variability between minimum and maximum values of predictive features. The Max-min normalization technique is expressed in equation 3.1

The following attributes were selected through feature selection algorithm so as to improve the prediction performance and provides cost-effective and faster predictors.

i. UTME score
ii. English-WAEC grade
iii. Mathematics-WAEC grade
iv. First Science Subject-WAEC grade
v. Second Science Subject-WAEC grade; and
vi. Third Science subject-WAEC grade.

The pre-processed data is divided into training data (70%) and testing data (30%). Training of the classifiers is done by feeding it with training data set. The methodology is subdivided into four stages as follows: Data pre-processing, Analysis of the algorithms, Training and Testing of classifiers respectively. Figure II gives the students performance predictive model architecture. K-Nearest Neighbor technique was then used to develop the predictive classifiers for the group of students' data under study.

$$NewValue = \frac{(f_{value} - f_{\min})}{(f_{\max} - f_{\min})} \quad (3.1)$$

Where
f_{value}, is the feature value to be normalized
$f_{min}$is the minimum feature value and
$f_{max}$is the maximum feature value respectively.

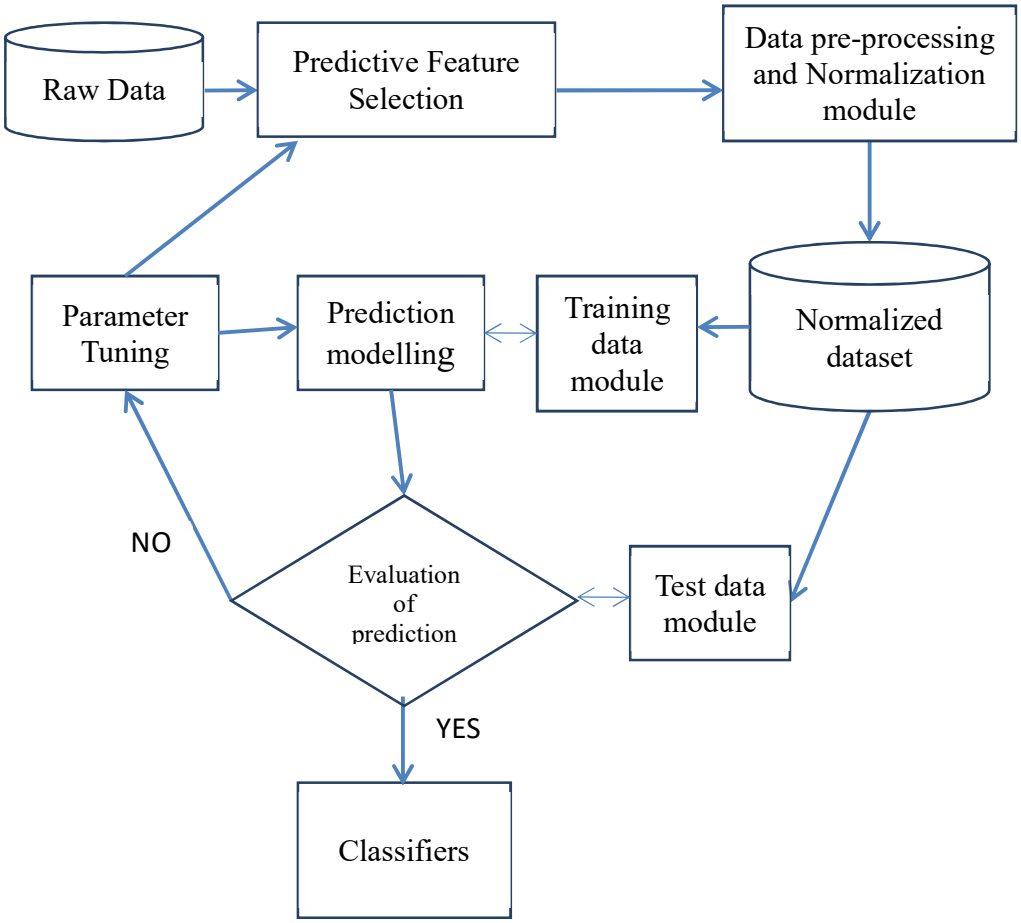

Fig. 1: Student Performance Predictive Model Architecture

3.3 Concept and Analysis of K-Nearest Neighbour Classifier:

KNN algorithm is the supervised machine learning algorithm used in this study to classify given instances into conflict or non-conflict prone states given a number of predictive features (n = 4). It is insensitive to outliers and makes no assumptions about data. It works well with numeric and nominal values. The K-Nearest Neighbour is an instance based learning which carries out its classification based on a similarity measure, like Euclidean, Mahanttan or Minkowski distance functions. The first two distance measures work well with continuous variables while the third suits categorical variables. Manhattan distance computes the absolute differences between coordinates of pair of objects while Minkowski is a generalized metric distance. Mahanttan and Minkowski distance functions are shown in equations (3.2) and (3.3) respectively.

$$D_{xv} = \left| X_{im} - X_{jm} \right| \qquad (3.2)$$

$$D_{xv} = \left(\sum_{m=1}^{n} \left| X_{im} - X_{jm} \right|^{\frac{1}{p}} \right)^{p} \qquad (3.3)$$

When p equals 2, then equation (3.3) becomes Euclidean distance. Euclidean distance computes the root of square difference between co-ordinates of pair of objects. The Euclidean distance (D_{xv}) between two input vectors (Xi, Xj) is given by:

$$D_{xv} = \sqrt{\sum_{m=1}^{n} \left(X_{im} - X_{jm} \right)^2} \qquad m = 1,2,\ldots,n \qquad (3.4)$$

For a given dataset, where

$X = \{x_1, x_2, x_3, \ldots\ldots\ldots, x_n\}$ (3.5)

$V = \{v_1, v_2, \ldots\ldots\ldots, v_c\}$ (3.6)

Where
X is the set of data points, V is set of centers of clusters and n is the number of features.

The Euclidean distance measure is used in this study for the K-NN classifier. For every data point in the dataset, the Euclidean distance between an input data point and current point is calculated.

These distances are sorted in increasing order and K items with lowest distances to the input data point are selected. The majority class among these items is found and the classifier returns the majority class as the classification for the input point. The parameter K is tuned for K = 1, 3, 5,..,O (where O is an odd number) to achieve optimal performance.

3.4 Calculation of the target class:

With respect to Table 3.2 , the calculation of the target class is as follows

Table 3.3: Normalized students' performance dataset with an unknown label (in Bold)

UTME Score	English-WAEC	Maths-WAEC	First Science subject-WAEC	Second Science subject-WAEC	Third Science subject-WAEC	Target Class
0.5862	0.5	0.5	0.5	0.5	0.5	2ND
0.0689	0.5	0.5	0.5	0.75	1	3RD
0.7586	0.5	0.75	0.5	0.75	0.25	2ND
0.4137	0.5	0.25	0.5	0.5	0.5	3RD
0.5517	1	0.5	0.5	0.75	1	2ND
0.2413	0.5	0.75	0.75	0.5	0.75	3RD
0.7586	0.5	0.5	0.5	0.5	0.75	2ND
0.6551	0.5	0.75	0.5	0.75	0.5	2ND
0.8620	0.5	0.5	0.75	0.75	0.75	1ST
0.6206	0.5	0.5	0.5	0.5	0.5	2ND
1	**0.5**	**0.5**	**0.5**	**0.75**	**0.5**	**?**

Using equation 3.4, the Euclidean distance calculation between the unknown data point and the preceding (10) rows are computed and tabulated in Table 3.4. For instance, the Euclidean distance between the first three vectors and the unknown vector with data points is computed as;

$$D_{xv} = \sqrt{(0.5862-1)^2 + (0.5-0.5)^2 + (0.5-0.5)^2 + (0.5-0..5)^2 + (0.5-0.75)^2 + (0.5-0.5)^2}$$

= 0.483450858

$$D_{xv} = \sqrt{(0.0689-1)^2 + (0.5-0.5)^2 + (0.5-0.5)^2 + (0.5-0..5)^2 + (0.75-0.75)^2 + (1-0.5)^2}$$

= 1.056799512

$$D_{xv} = \sqrt{(0.7586-1)^2 + (0.5-0.5)^2 + (0.75-0.5)^2 + (0.5-0..5)^2 + (0.75-0.75)^2 + (0.25-0.5)^2}$$

= 0.428093414

Table 3.4: Euclidean distance between vectors and unknown data point

Vectors	Euclidean distance to unknown data point
Row vector 1	0.483450858
Row vector 2	1.056799512
Row vector 3	0.428093414
Row vector 4	0.684571783
Row vector 5	0.837228313
Row vector 6	0.90857325
Row vector 7	0.428093414
Row vector 8	0.425917908
Row vector 9	0.379506219
Row vector 10	0.454286625

After computing the Euclidean distances to the unknown target class label, the K-Nearest vectors is found by sorting the distances in decreasing order. Following the distance calculation, the distances are sorted from least to the highest (shown in Table 3.5). Next, the first k or lowest k distances are used to vote on the unknown target class of input vector. The input k should always be a positive integer. Take K=3. Then, the three closest vectors are vectors 9, 8 and 3 respectively. The K-NN algorithm takes the majority vote from these three vectors to determine the class of the unknown input vector. Because all three nearest vectors (row vectors 9, 8 and 3) have class labels 1st, 2nd and 3rd respectively then the unknown input vector will be 2nd (Second Class Hons). This is because the outcome is based on majority vote from the three nearest vectors.

Table 3.5: Euclidean distance between vectors and unknown data point sorted in ascending order

Vectors	Euclidean distance to unknown data point
Row vector 9	0.379506219
Row vector 8	0.425917908
Row vector 3	0.428093414
Row vector 7	0.428093414
Row vector 10	0.454286625
Row vector 1	0.483450858
Row vector 4	0.684571783
Row vector 5	0.837228313
Row vector 6	0.90857325
Row vector 2	1.056799512

Note that row vector 3 and 7 had same Euclidean distance value, thus row vector takes priority in order of hierarchy based on position of data point in dataset. It is important to note that there are various factors that impact the performance of a k-NN based model classifier, such as settings of the classifier and the nature of dataset. Different algorithms perform differently on different datasets.

4. EVALUATION RESULTS FOR THE CLASSIFIERS

The performance of the kNN model classifier is measured with the error rate. In classification, the error rate is the number of misclassified pieces of data divided by the total number of data points tested. An error rate of 0 means a perfect model classifier is achieved, and an error rate of 1.0 means it is always wrong. Therefore, the testing of the classification is done using training and testing error metrics. To evaluate these classifiers, 70% each of the datasets is used for training while 30% is set aside for validating and testing. Performance evaluation is carried out based on true positive, true negative, false positive, false negative rates and total accuracy of classification. True positives are instances classified as positives which are actually correct classification or prediction. False positives are instances classified as otherwise which are incorrect classification or prediction.

$$TPR = \frac{TP}{P} \tag{3.7}$$

$$FPR = \frac{FP}{N} \tag{3.8}$$

where
TPR, FPR, means true positive rate and false positive rates.

Prediction accuracy was also used as evaluation criteria due to its widespread relevance in most related literature. The prediction accuracy as shown in equation (3.8) for n number of test cases is given as;

$$PA = \frac{1}{n}\sum_{i=1}^{n} TP_i \qquad (4.1)$$

where
n = total number of test cases (that is one third of 48 = 15), TP_i = number of true positives

The test results of the K-NN model on the dataset consisting of all six features set is shown in Table 4.1.

Table 4.1: Test Results of kNN model classifier

Metrics	II,UR,EF,RD
TP rate	73.33%
FP rate	26.66%
PA	73.33%

*TP = True Positive, FP = False Positive, PA = Prediction Accuracy

It could be seen from Table 4.2 that kNN classifier rather showed better true positive rate performance (TP rate) than false positive rate (FP) rates. This means that the classifier is adequate for predicting final students' performance in a degree program accurately.

Thus the prediction accuracy of the k-NN model classifier for students' performance prediction is 73.33%. The performance evaluation of the model with reviewed works is shown in Table 4.2.

Table 4.2: K-NN classifier prediction accuracies

Student performance prediction models	Kabakchieva (2012) model	Asif et al (2014) model	Agrawaland Mavani (2015)	Muralidharanet al (2017)	Proposed K-NN based model
Prediction Accuracy (PA)	70.49%	73.08%	70%	57.14%	73.33%

* PA = Prediction Accuracy

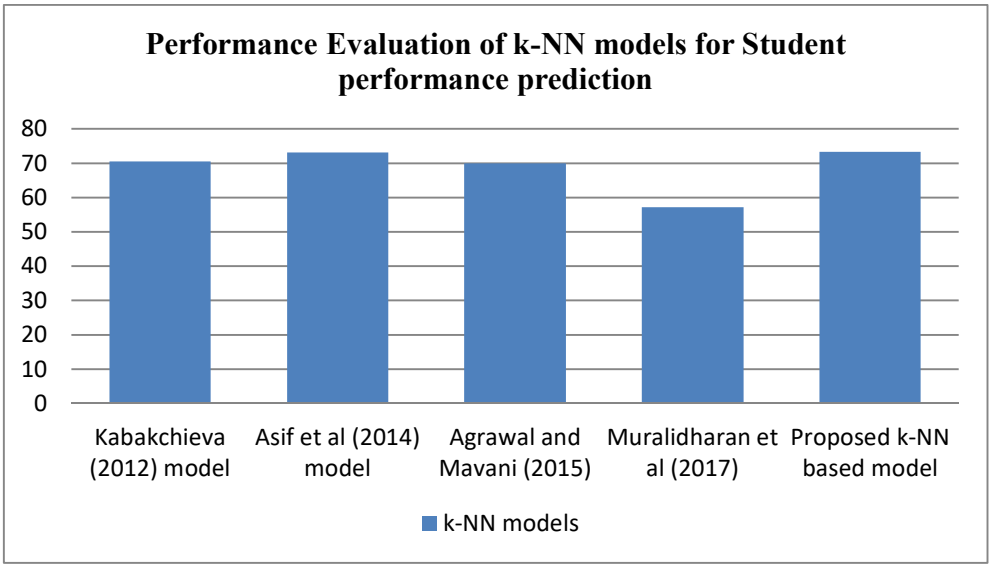

Fig. 4.1: Performance evaluation of K-NN based Students' Performance Prediction model classifiers with other related models

5. Conclusion

The evaluation of the proposed K-NN based students' performance prediction model as shown in figure 4.1 with other models in related works depict better performance. The k-NN model developed in Kabakchieva (2012) had a prediction accuracy of 70.49% while in Asif et al (2014), the accuracy recorded was 73.08%. Agarwal and Mavani (2015) models showed 70% accuracy while that of Muralidharanet had an average accuracy of 57.14% due to the limited dataset used. Conclusively, it implies by the results of this study, that the choice for an optimal student performance prediction model is greatly dependent on the feature(s) used, together with the quality and dimensionality of dataset involved. An optimal current student performance prediction model (using kNN technique) was also established which classified computer science students of Alhikma University into first, second and third class degrees.

6. FUTURE

The promising areas of future work for this study include: Modeling of students performance prediction using a more comprehensive and large datasets and Investigating the viability of hybrid machine learning techniques in actualizing prediction of student final degree performance.

REFERENCE

1. Abdullah, A. L., Malibari, A., &Alkhozae, M. (2014). Students' Performance Prediction System using Multi Agent Data Mining Technique. International Journal of Data Mining & Knowledge Management Process, 4(5), 1.
2. Acharya, A., & Sinha, D. (2014). Early prediction of students performance using machine learning techniques. International Journal of Computer Applications, 107(1).
3. Agrawal, H., &Mavani, H. (2015). In Student Performance Prediction using Machine Learning. International Journal of Engineering Research and Technology.
4. Asif, R., Merceron, A., &Pathan, M. K. (2014). Predicting student academic performance at degree level: a case study. International Journal of Intelligent Systems and Applications, 7(1), 49.
5. Jia, J. W., &Mareboyana, M. (2013). Machine Learning Algorithms and Predictive Models for Undergraduate Student Retention. In Proceedings of the World Congress on Engineering and Computer Science (Vol. 1).
6. Kabakchieva, D. (2012). Student performance prediction by using data mining classification algorithms. International Journal of Computer Science and Management Research, 1(4), 686-690.
7. Katare, A., and Dubey, S., (2017). A Study of various Techniques for Predicting student Performance under Educational Data Mining, International Journal of Electrical Electronics and Computer Engineering, 6(1): 24-28, ISSN: 2277-2626 (Online)
8. Muralidharan, V., Pravien, M., and Balaji, J. (2017). Result Prediction Using K-Nearest Neighbor Algorithm for Student Performance Improvement, International Research Journal of Electronics & Computer Engineering (IRJECE), 3(1): 7-10