International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 4, Issue 11, pp 43-49, November-2020

https://doi.org/10.47001/IRJIET/2020.411006

Using Formative and Summative Assessments in Data Mining to Predict Students' Final Grades

[1]Iliyasu Yahaya Adam, [2]Hassan Bello, [3]Abdullahi Abba Abdullahi, [4]Musa Dan-Azumi, [5]Nura Abdullahi

[1,3,4]Kano State Institute for Information Technology, Kura, Kano State, Nigeria
[2]Kano State Polytechnic, Kano State, Nigeria
[5]Ahmadu Bello University, Zaria, Nigeria

***Abstract* - The study intends to devise an earlier and more accurate analytical means of predicting the class of degree a student will graduate with. This will help in decision making and the design of academic programs and curriculum development; also it will help in student guidance and counseling. Because the earlier we can predict the class of degree a student is likely to graduate with, the better it will be for the student to keep it up or improve on his performance and this will consequently assist in mitigating and minimizing the student dropout, attrition, or dismissal from school after wasting reasonable amount of time without acquiring the certificate. Academic data of some Communication & Information Technology students; such as year of admission, year of completion, individual grades obtained from the courses he/she offered at a 1 year diploma program, 1 year advance diploma, and the class of degree he obtains from a 1year top-up degree program was introduced to a Classification Data Mining algorithms to extract a pattern and a model for students' final grade prediction. The study's result shows that timely completion of the first two programs, a high score in computer architecture course, programming, network, and discrete mathematics courses are determining factors that can be used to predict students' final grades at graduation.**

Keywords: Data Mining, Summative Assessments, Students, Final Grades, formative.

I. INTRODUCTION

The study is all about a 3 year degree program. Students undergo a 3 step continuing education program that leads to the award of a bachelor's degree in computer science or Information Technology. These 3 steps are a 1 year intentional diploma, a 1 year international Advance diploma, and a 1 year top-up degree program respectively. It is aimed at the study to analyze using data mining algorithms to see if students' academic records in the first 2 levels can be used to predict the student's final performance and success at graduation. This will help to minimize student dismissal from school instead of successfully graduation.

II. LITERATURE REVIEW

Data mining is an emerging technology that helps organizations to efficiently learn with historical data and use the learned knowledge to predict future behavior for concerned areas, the adoption of Data mining technology as a futuristic strategic management tool can be used to enhance the growth of current educational system (Agarwal, Gandey, & Tiwari, 2012).

2.1 Data Mining

Data mining is defined as an analysis methodology that is used to identify some hidden patterns from a large data set (Durairaj & Vijitha, 2014).We can use it to discover an unknown and useful knowledge, but whether the knowledge we discover is useful and of interest to us is very subjective and depends upon the applications and the user (Zaïane & Osmar, 1999). Kovačić (2010) reported Elder and Miner (2009, p 17) to be defining data mining as the " the use of machine learning algorithm to find faint of the relationship between data elements in the large, noisy and messy dataset which can lead to actions to increase benefit in some form (diagnosis, profit detection, detection, etc.)".Data mining is defined as an analysis methodology that is used to identify some hidden patterns from a large data set (Durairaj & Vijitha, 2014). Data mining also refers to extracting or "mining" knowledge from large amounts of data. Data mining algorithm s are used to work on large volumes of data to discover hidden patterns and relationships that help make a decision (Baradwaj & Pal, 2011). With data mining, we can discover unknown and useful knowledge, but whether the knowledge we discover is useful, useful, and of interest to us is very subjective and depends upon the applications and the user (Zaïane & Osmar, 1999). Kovačić (2010) reported Elder and Miner (2009, p 17) to be defining it as the "the use of machine learning algorithm to find faint of the relationship between data elements in the large, noisy and messy dataset which can lead to actions to increase benefit in some form (diagnosis, profit detection, detection, etc.)". It is a stage in knowledge discovery in database (KDD) as can be seen from figure 1.0 below.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 4, Issue 11, pp 43-49, November-2020

https://doi.org/10.47001/IRJIET/2020.411006

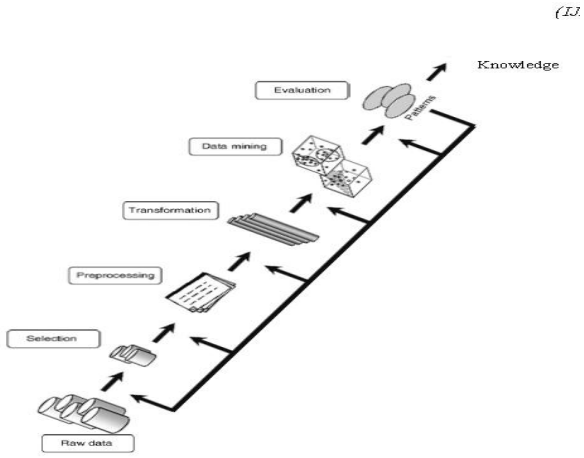

Figure 1: Data mining and other stages in KDD

2.2 Data Mining Concepts and Algorithms

The field of data mining is an interdisciplinary field of study that results from a fusion of many different areas such as machine learning, statistics, Pattern Reorganization, Databases Artificial Intelligence, and Computation capabilities, etc. To have a better discussion several techniques and algorithms like Clustering, Classification Artificial Intelligence, Regression, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, and Nearest Neighbour methods used to discover knowledge from warehouses and databases. Baradwaj and Pal (2011) explain some frequently used data mining techniques as follows.

Classification

This data mining algorithm employs a set of pre-classified examples to develop a model that can be used to classify a larger population of records. It is achieved through learning the training data and classification of the training data. When learning a given training data, classification algorithms are used to analyze the data set; while during classification, test data are used to estimate the accuracy of the classification rules, so that they can be applied to new data columns if the accuracy is acceptable.

Clustering

This is used to identify similar classes of objects; it can be used further to identify densely and parse regions in object space as well as to discover overall distribution pattern and correlation among data attributes.

Decision Trees

This technique generates rules that classify datasets. It used to be a tree-shaped structure representing sets of decisions.

Educational Data mining (EDM)

The history of data mining in other sciences such as physics, biology, and geography can be traced to the earlier 1970s. Contrarily, analytics in learning sciences is relatively late and emerging, being that conferences on EDM commenced in 2008 followed by the publication of the journal of educational data mining in 2009. Though some workshops held since 2000 and 2004 (Baker, 2014). The recent increase in the emergence of data analytics can be connected to the increase in the use of mobile technologies in most of the educational settings around the globe. As learners interact with these digital devices, some data about the interaction can be grabbed easily and this data is applied to subsequent analysis. EDM has been defined as ‘‘an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings and using those methods to better understand students, and the settings which they learn in’’ (IEDMS, 2009). Akinola, Akinkunmi, & Alo (2012) Contended that EDM techniques can be applied on the educational dataset to extract the hidden knowledge for predictions concerning the enrolment of students into a particular course, alienation of traditional classroom teaching model, detection of improper values in the result sheets of the students, exam malpractices and predicting student performance.

2.3 Related Study

Duraraj and Vijitha (2014) used a clustering algorithm called K-means clustering on student real-time data such as details of different subject marks in semester wise that determines the relationship the student's learning behavior and their academic performance and they find it to be a very accurate means of predicting student performance. Baradwaj and Pal (2011) investigate the accuracy of technique to predict student performance using a data set like Attendance, Class test, Seminar, and assignment marks collected from the student's management system, to predict the performance at the end of the semester. The ID3 decision tree technology is used on student data such as the student sex, grades obtain from their senior secondary school certificate examination (SSCE), entry examination scores, and the grades obtained by the students during graduation. The research shows that there is a significant relationship between the scores students obtained in some specific subjects from their senior secondary school certificate examinations and the classes of degree they graduate with (Ogunde & Adibaje, 2014). A study conducted by(Yehuala, 2015) on the different variations of the student records using decision tree and Bayes as classification techniques shows that data mining techniques can be applied to by higher education institutions and universities to determine the success and failure rate so that managing the

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 4, Issue 11, pp 43-49, November-2020
https://doi.org/10.47001/IRJIET/2020.411006

student's enrolment can assist the students earlier before they reach the risk of failure also to ensure effective resource utilization, cost minimization, helping and guiding administrative officers to be successful in management and decision making Akinola, Akinkunmi, & Alo, (2012) researched at the University of Ibadan on student dataset like ordinary level results, Mathematics and Physics scores obtained in year one, marks obtained from programming course by year two in the department of computer science. These datasets were suggested to a data mining system using an artificial neural network algorithm called the Multi-layer perception feed-forward back propagation technique. It shows that prior knowledge in physics and mathematics is central to student prosperity in computer programming and that those students at risk could be detected earlier be given the necessary treatment before it is too late. A group (Minaei-Bidgoli, Minaei-Bidgoli, Minaei-Bidgoli, & Punch, 2003) Used and compared six different classifiers on the LON-CAP a Michigan state University web-based application database. These multiple and commonly used in solving practical problem classifiers include Quadratic Bayesian Classifier, 1-Nearest Neighbour, K-nearest Neighbour Parzen- Window, Multilayer, Perceptron, and Decision Tree. Finally, the result shows that data mining efforts can use in predicting student learning outcomes.

III. METHODOLOGY

3.1 CRISP-DM Methodology (Cross Industry Standard Process for Data Mining)

The study chooses to employ some basic steps of the CRISP-DM 1.0 model as the research approach being freely available, non-proprietary, and application neutral standard for data mining projects. Steps include business understanding, data understanding, data preparation, modeling evaluation, and deployment (Chapman (NCR), et al., 2000). Figure 3.1 shows the stages involved in Crisp-DM and the possible feedbacks that may exist between the stages.

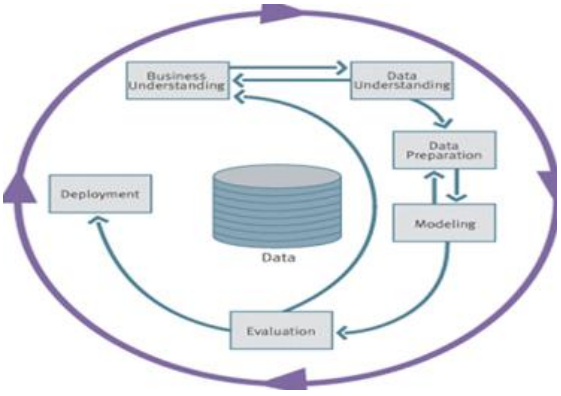

Figure 2: Stages involved in CRISP-DM 1.0 (source: http://www.crisp-dma.org/CRISP-0800.pdf)

Every phase in these processes have some other steps under it, Figure 2 shows how the above data mining tasks are elaborated in more details.

Business Understanding

The business is all about a 3 year degree program. Students undergo a 3step continuing education program that leads to the award of a bachelor's degree in computer science or Information Technology. These 3 steps are a 1 year intentional diploma, a 1 year international Advance diploma, and a 1 year top-up degree program respectively. It is aimed at the study to analyze and see if students' academic records in the first 2 levels can be used to predict the student's final performance and success at graduation.

Data Understanding

The study covers the course grades obtained by Kano Informatics students from their international diploma in computing or international diploma in information technology, advanced diploma in computing, or advanced diploma in the network and computer security from the sub-franchise partnership of Informatics academy Singapore through Jigawa state institute of information technology. The grades and the classes of degree the students graduated with from the IUEA Uganda are fed into data mining algorithms and used both in the training and test data to build a model for predicting the class of degree a student is going to graduate with.

Data Collection

Student details were captured to feed a Microsoft Access database table. The table holds the student details. The values entered into the table two using two different Ms-Access forms. This section describes the attribute used in the training dataset.

The Dependent Variable (Class Variable)

The variable used to classify the dataset as in the research is called *class* and it represents the class of degree a student graduates within his final year top-up degree. This target variable also called the Label variable is the one to be predicted by the model.

The Independent Variables (Predictor Variables)

These are variables whose values are used in the study to predict the target variables are classified into an *identifier*, *socio-demographic*, *formative assessment*, and *summative assessment* variables and potentially useful in predicting the Class variable.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 4, Issue 11, pp 43-49, November-2020

https://doi.org/10.47001/IRJIET/2020.411006

The Identifier Variables

This is the serial number in the list given to every student in the 197 batch1 students who have completed their Top-Up degree.

Formative Assessment Variables

In normal circumstances, formative assessments are an internal form of assessments in which there is room for agreement and understanding between the teachers and the students, such as marks obtained from class quizzes, assignments, continuous assessment tests, and attendance. However, throughout this study it means any academic historical data used to show the students' area of study, duration diploma, advance diploma, whether he has been able to finish his diploma before he enrolled into an advance diploma, or he finishes his diploma later than his advance diploma.

Summative Assessment Variables

These sets of variables represent the grades scored by the students from the individual courses in their diploma and advance diploma external exams. (External Assessments)

Data Preparation, Selection, and Integration

Additional data sources were to get the students' scores from individual courses and the student final grades at graduation obtained from 1year top-up degree at the International University of East Africa (IUEA) UGANDA.

Data Cleaning

About130 records out of 197 records. Nevertheless, these 130 records are used as training instances. Only 34 attributes out of 38 existing attributes are used.

Data Construction

Here new variables such as "SEQUENCE", "DDURY" (duration of the diploma), "ADURY" (duration of adv. Diploma), and "STATUS" were derived from other attributes such as (admission date, date of diploma completion, date of adv. diploma completion) that were used to derive new attributes later they were discarded and replaced with the new attributes derived from them. Some excel built-in functions (such as IF FUNCTION and DAYS360 FUNCTION) are used to calculate the values of the derived attributes.

Data Transformation (Formatting)

The MS-Access database was exported to an excel file for further processing to be transformed into a format that is more appropriate for a data mining tool (WEKA Tool). Mostly the data mining tools used the flat-file format to pre-process a given dataset hence it is very important to convert training and testing dataset to the specified format known as an *attribute relation file format* (ARFF).

Modeling Technique

Classification is chosen to be the data mining technique for this research because the aim of the study is the prediction of students' final grades. The J48 decision tree classification algorithm is used in the research as the data mining technique because it is very suitable for generating pruned and unpruned trees.

Model Test Design

The percentage split is selected as a testing procedure in the study because of the limited number of training instances. 66% of the data records are designed to be used as the training data while the remaining records used as the test data using the percentage split option of the

Model Building

Initially, the default parameters of the configuration panel of the WEKA tool were used to create the model and a less accurate model was generated and this called for reconfiguring the modeling parameters. The result obtained initially is not uniform; hence the configuration panel is adjusted so the minimum number of instances per leaf becomes 10 instead of the default value (2). With this little change, a more uniform and more accurate model is generated. The generated tree is visualized as follows as in figure 3 below. The tree size is 6 and has 9 leaves.

Figure 3: Last decision tree visualized (generated model)

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 4, Issue 11, pp 43-49, November-2020

https://doi.org/10.47001/IRJIET/2020.411006

Model Assessment

The learning and model validation shows to be 48.7% 0r approximately 50% accurate and reliable. The number of correctly classified instances is 19 and the number of the incorrectly classified instances is 20 as can be seen in figure 4 below.

Classifier output

```
Correctly Classified Instances          19               48.7179 %
Incorrectly Classified Instances        20               51.2821 %
Kappa statistic                          0.0714
Mean absolute error                      0.1617
Root mean squared error                  0.2973
Relative absolute error                 90.8814 %
Root relative squared error             99.5874 %
Total Number of Instances               39
Ignored Class Unknown Instances                 5

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.944     0.905      0.472     0.944      0.63       0.535    SECOND CLASS UPPER
                 0.182     0.036      0.667     0.182      0.286      0.715    FIRST CLASS
                 0         0          0         0          0          0.588    SECOND CLASS LOWER
                 0         0          0         0          0          0.429    PASS
                 0         0          0         0          0          0.686    1 RETAKE
                 0         0          0         0          0          0.5      2 RETAKE
                 0         0          0         0          0          0.686    3 RETAKE
                 0         0          0         0          0          ?        SUSPENDED
Weighted Avg.    0.487     0.428      0.406     0.487      0.371      0.592

=== Confusion Matrix ===

  a  b  c  d  e  f  g  h   <-- classified as
 17  1  0  0  0  0  0  0 |  a = SECOND CLASS UPPER
  9  2  0  0  0  0  0  0 |  b = FIRST CLASS
  4  0  0  0  0  0  0  0 |  c = SECOND CLASS LOWER
  2  0  0  0  0  0  0  0 |  d = PASS
  1  0  0  0  0  0  0  0 |  e = 1 RETAKE
  2  0  0  0  0  0  0  0 |  f = 2 RETAKE
  1  0  0  0  0  0  0  0 |  g = 3 RETAKE
  0  0  0  0  0  0  0  0 |  h = SUSPENDED
```

Figure 4: Model accuracy and the confusion matrix

3.2 Result Evaluation

The result of the model can be evaluated as follows:

a. The aim of the research as mentioned in to examine and determine the extent to which the available dataset can be used in data mining to predict student final grade. *It is evident from the research that the available data can be used to predict a student's final grade at graduation with approximately 50% precision.*

b. One of the research questions asked: *Can we find classes of students? Meaning do groups of students exist who have similar performance and similar learning outcomes based on measurable criteria?*
It is evident from the model that, there are existing measurable criteria that can be used to classify the student's final performance at graduation. These include the time it takes a student to finish the two-year bridging program (diploma and advance diploma) and the grade he scores from some brain appealing courses such as computer architecture, discrete mathematics, introduction to java, program design (algorithms), etc.

c. Another research question asked: *Can we use the model obtained to further classify and predict student performance in the future?*
Been that 34% of the dataset has been used as test data to evaluate the model and still an accuracy of approximately 50% has been obtained. This means that future students' final grades can be predicted by the model with about 50% accuracy.

IV. RESULT AND DISCUSSION

From the model generated the extracted pattern shows that the most important variables in the model prediction are the STATUS, CADC and DDURY Normally a student has to finish his diploma or advance diploma program each in eight months that is two complete semesters, both two programs can be completed in 18 months. But some students may not be able to do so even though all courses are repeated every semester therefore it takes them more than 18 months to finish their diploma and advance diploma. The attribute (STATUS) is used in the study to show whether a student has been able to

ISSN (online): 2581-3048
Volume 4, Issue 11, pp 43-49, November-2020
https://doi.org/10.47001/IRJIET/2020.411006

finish his two courses in time or not and this attribute became the root node (starting point) of the extracted tree. This show is timely completion. Computer architecture and data communication (CADC) is a course taken at an advanced diploma. The course covers some aspects of computer architecture, operating systems, networks, and some calculations. (DDURY) is another attribute that tells how long a student takes to finish his diploma alone and it shows an important duration of a diploma over an advanced diploma. It is also obvious from the model that those students who scored fewer marks in these courses and do not usually complete tasks given to them are vulnerable to failure, dropout, dismissed, or graduating with lower classes of degree. Some attributes also appeared as predictors but with lesser precision. However, the maximum accuracy was obtained after adjusting the minimum number of objects from the configuration. The changes made the subsequent attributes to disappear from the model and the maximum accuracy obtained. Summarily the followings are the findings of the research.

1. The industrious and most time-conscious students are more likely to graduate with higher classes of degree.
2. Mastery of programming, operating system, and networks are good predicates of graduating with a higher class of degree.
3. The success rate in this 3 step degree program is high compared to the failure rate. Hence this form of the lifelong learning process is another good alternative to the conventional and traditional University programs.
4. Tertiary institutions can be used to minimize the burden and on universities due to the large number of prospective students looking for a degree qualification.

V. CONCLUSION AND RECOMMENDATIONS

The work has to some extent achieved its aims and objectives in coming up with a model that can be used in predicting students' final performance at graduation so that some corrective and preventive measures are taken earlier. Fortunately, some attributes in the student dataset were found to be appealing to serve the purpose. This shows that the use of data mining techniques on academic data can be counted as one of the critical success factors in educational management and planning, curriculum development, and forms of decision making in an academic setting.

5.1 Recommendations

Based on the findings of the work, the study comes with some recommendations as follows.

1. parents/guardians/students should find information from the relevant authorities in the school Management on what factors determine the student success at the end of studies so that they can cope with challenges that interfere with early and timely completion of their programs and the school authority should prepare and readily provide this information when needed
2. Government parastatals such as the national Universities council, National business, and technical examination board, Universities, Polytechnics, and other institutions should collaborate and design diploma programs that suit the corresponding levels of the degree programs to exploit the available human and material resources in these institutions and to reduce the workload and backlog of student waiting to get admission into universities. This in turn will boost up the national economy and reduce national insecurity and a crime rate that is affecting our youth and the nation at large.

5.2 Suggestions for Further Studies

Future research should deeply incorporate socio-economic and socio-demographic variables such economic background of students' family, educational background of students parent, the geographical area a student is brought up whether rural or urban, the hobbies and games a student engage in, whether the student is an on-campus student or off-campus, class attendance, quizzes, presentations, group assignments, home works, tests, lab practices and other means of internal assessments.

REFERENCES

[1] Agarwal, S., Gandey, G. N., & Tiwari, M. D. (2012, April). Data Mining in Education: Data Classification and Decision Tree Approach. *International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 2, No. 2, April 2012*, 140-144.

[2] Akinola, O. S., Akinkunmi, B. O., & Alo, T. (2012). A Data Mining Model for Predicting Computer Programming Proficiency of Computer Science Undergraduate Students. *African Journal of Computing & ICT, 5* (1), 43-52.

[3] Baker, R. S. (2014). *Educational Data Mining and Learning Analytics George Siemens.* Athabasca: Athabasca University.

[4] Baradwaj, K., & Pal, S. (2011). Mining Educational Data to Analyze Students‟ Performance. *International Journal of Advanced Computer Science and Applications, 2* (6).

ISSN (online): 2581-3048
Volume 4, Issue 11, pp 43-49, November-2020
https://doi.org/10.47001/IRJIET/2020.411006

[5] Behrouz, M.-B. (2004). *Data Mining for a Web-Based Educational System.* Michigan: Michigan State University.

[6] Chapman (NCR), P., Clinton (SPSS), J., Kerber (NCR), R., Khabaza (SPSS), T., Reinartz (DaimlerChrysler), T., Shearer (SPSS), C., et al. (2000). *CRISP-DM 1.0 Step-by-step data mining guide.* Amsterdam: CRISP-DM Consortium.

[7] Durairaj, M., & Vijitha, C. (2014). Educational Data mining for prediction of student performance using clustering algorithms. *International Journal Computer science and information technologies, 5(4)*, 5987-5991.

[8] IEDMS. (2009). *International Educational Data Mining Society.*

[9] Kovačić, Z. J. (2010). Early Prediction of Student Success: *Informing Science & IT Education Conference (InSITE)* (pp. 648-665). New Zealand: Open Polytechnic, Wellington, New Zealand.

[10] Minaei-Bidgoli, B., Minaei-Bidgoli, D. A., Minaei-Bidgoli, G., & Punch, W. F. (2003). Predicting Student Performance: An Application of Data Mining Methods with Educational Web-Based Systems LON-CAPA. *33rd ASEE/IEEE Frontiers in Education Conference* (pp. 1-6). Michigan State: Boulder, CO.

[11] Nourein, A. A. (2010). *A Heuristic Approach To Predict Malaysian Students Academic Performance.* Malaysia: Faculty of Computer Science And Information Technology University Of Malaya Kuala Lumpur.

[12] Ogunde, A., & Adibaje, D. (2014). A Data Mining System for Predicting University Students' Graduation Grades. *Journal of Computer Science and Information Technology*, *2*, 21-46.

[13] Yehuala, M. A. (2015). Application Of Data Mining Techniques For Student Success And Failure Prediction (The Case Of Debre_Markos University). *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, *4* (04).

[14] Zaïane, & Osmar, R. (1999). *Principles of Knowledge Discovery in Databases.* Alberta: University of Alberta.

Citation of this Article:

Iliyasu Yahaya Adam, Hassan Bello, Abdullahi Abba Abdullahi, Musa Dan-Azumi, Nura Abdullahi, "Using Formative and Summative Assessments in Data Mining to Predict Students' Final Grades" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 4, Issue 11, pp 43-49, November 2020. DOI of article https://doi.org/10.47001/IRJIET/2020.411006

****

