International Journal of Advances in Scientific Research and Engineering (ijasre)

E-ISSN : 2454-8006

DOI: 10.31695/IJASRE.2022.8.10.4

Volume 8, Issue 10
October - 2022

# Clustering, Classification, and Association Rule Mining for Educational Datasets

**EBELOGU Christopher U[1], AGU Edward O[2]**

Research Scholar[1], Lecturer[2]

[1]Department of Computer Science, University of Abuja, Abuja-Nigeria.

[2]Department of Computer Science, Federal University Wukari, Taraba State, Nigeria.

---

## ABSTRACT

*Understanding students/pupils within the context of school is the focus of the expanding research topic known as educational data mining. Before the adoption of data mining tools in student analysis, it was challenging to identify students who were at risk of failure. However, using educational data mining tools has made it straightforward to examine student performance on previous exams. It helps teachers understand the kind of children they are working with. This knowledge can assist teachers to tailor their lecture notes to each student's needs and help difficult students focus more throughout the class. In this study, 1428 newly admitted student records from a tertiary institution in Nigeria were examined. Decision trees, association rules, and K-means clustering techniques were used to analyze the data. The results showed that arts courses were primarily responsible for low scores, social science courses for average scores, and the high JAMB scores came from science courses. Additionally, students from the geopolitical zones of the South-East, South-South, and North-East performed better than those from the geopolitical zones of the North-West, North Central, and South-West. The students with the highest JAMB marks were those who offered Physics and had an English score of at least 70. Most males offered Physics, Chemistry, Mathematics, and Biology while most females offered Economics, Government, CRS, and Literature in English. We highly recommend that Students who did not perform well should be given more attention and extra lessons/classes should be held for them to improve their grades in the various departments they were admitted.*

***Key Words:*** Artificial Neural Networks, Clustering, Classification, Decision tree, K-Nearest Neighbor, Machine Learning, Naive Bayes, Support Vector Machine.

---

## 1. INTRODUCTION

Data mining is a technology used to find the significant relationship among variables in a large database. It is used to extract knowledge from databases to find necessary information according to the requirement which can be used in the future [1]. The information obtained from mining data is new, correct, and potentially useful. It has been known for its powerful role in uncovering hidden information from large volumes of data and because of its advantages; it has now been applied to Educational Databases.

Educational data mining can be defined as an emerging discipline, concerned with developing methods for exploring the unique types of data that come from the educational setting; and using those methods to better understand students, and the settings in which they learn (www.educationaldatamining.org).

It can also be defined as a branch of data mining and 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 in which they learn [2]. Data mining techniques are used to extract hidden knowledge from a large database and this knowledge can be used to predict student behavior.

Some factors that can be used in analyzing student data include performance in the last examination, family income, background, communication skills, technical skills, and leadership quality [3]. Some techniques of data mining that can be used in educational data mining include clustering, rule mining, neural network, and classification.

### 1.1 Aim and Objectives

This research aims to evaluate the students admitted into the University of Abuja to understand their relationships with one another by applying data mining algorithms to a collection of student data.

**The objectives of this study are;**

i. To use the features and characteristics of the admitted students to determine the clustering, classification, and association rule approaches.

ii. To evaluate and contrast the results of the three different machine learning techniques utilized to analyze the student data.

iii. To proffer relevant and appropriate recommendations to the institution and the educators.

## 2. LITERATURE REVIEW

Educational data mining is a dissertation field concerned with the application of data mining, machine learning, and statistics to information generated from an educational setting to discover new insights about how people learn in the context of such settings. It refers to techniques and dissertations designed for automatically extracting meaning from large repositories of data generated by or related to people's learning activities in educational settings [4]. Educational data mining examines data that are generated by the educational institution such as prediction of the student performance, grouping the student according to their performance, and recommendations to the student [5].

Educational data mining cannot be explained without first understanding what educational data is. Educational data is the information that educators, schools, districts, and state agencies get from students, such as personal information (e.g., a student's age, gender, etc) enrollment information (e.g., name of the school, course duration, etc), academic information (e.g., the courses a student takes, students grades, etc) and many types of data gathered and used by educators and educational institutions (e.g., information related to medical and health issues, learning disabilities, etc.) (edglossary.org, 2015)

The use of technologies and software applications has allowed schools, districts, and state agencies to be able to gather a list of detailed data on student data easily. Advances in educational software, computing technologies, internet access, and innovations such as cloud-based data storage and "big data" analytics have fueled a dramatic increase in the collection and use of student-level data in recent years (edglossary.org, 2015)

The data obtained from students provides a large resource of educational data that can be explored to understand how students learn [6]. Understanding and analyzing the factors for poor performance is a complex process hidden in past and present information congregated from academic performance and students' behavior. Technology such as data mining will be required to analyze and predict the performance of students scientifically [7].

The information obtained from educational data mining helps students, teachers, school administrators, and educational policymakers to make informed decisions about how to manage students and educational resources.

The overall goal of educational data mining is to understand how students learn and identify those aspects that can improve learning and educational aspects [1].
Other goals of educational data mining include:
i. *Predicting students' future learning behavior* – With the use of student modeling, this goal can be achieved by creating student models that incorporate the learner's characteristics, including detailed information such as their knowledge, behaviors, and motivation to learn. The user experience of the learner and their overall satisfaction with learning are also measured.

ii. *Discovering or improving domain models* – Through the various methods and applications of EDM, the discovery of new and improvements to existing models is possible. Examples include illustrating the educational content to engage learners and determining optimal instructional sequences to support the student's learning style.

iii. Studying the effects of educational support that can be achieved through learning systems.

iv. Advancing scientific knowledge about learning and learners by building and incorporating student models, the field of EDM dissertation, and the technology and software used.

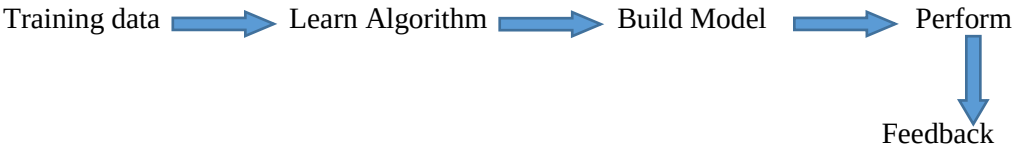


DOI: 10.31695/IJASRE.2022.8.10.4

### 2.1 Machine Learning and Data Mining

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves. Data mining on the other hand refers to extracting knowledge from a large amount of data. Data mining is the process to discover various types of patterns that are inherited in the data and which are accurate, new, and useful.

Machine learning is a way to discover a new algorithm from the experience. Machine learning involves the study of algorithms that can extract information automatically. Machine learning uses data mining techniques and another learning algorithm to build models of what is happening behind some data so that it can predict future outcomes.




Data mining and Machine learning Comparison Table is shown in Table 2.1

**Table 2.1** is the comparison table between data mining and machine learning.

| Basis for comparison | Data mining | Machine learning |
|---|---|---|
| Meaning | Extracting knowledge from a large amount of data | Introduce new algorithms from data as well as experience |
| History | Introduce in 1930, initially referred to as knowledge discovery in databases | introduce in near 1950, the first program was Samuel's checker-playing program |
| Responsibility | Data mining is used to get the rules from the existing data. | Machine learning teaches the computer to learn and understand the given rules. |
| Origin | Traditional databases with unstructured data | Existing data as well as algorithms. |
| Implementation | We can develop our models where we can use data mining techniques for | We can use machine learning algorithms in decision trees, neural networks, and, some other areas of artificial intelligence. |
| Nature | Involves human interference more manually. | Automated, once the design is self-implemented, no human effort |
| Application | used in cluster analysis | Used in web search, spam filter, credit scoring, fraud detection, computer design |



DOI: 10.31695/IJASRE.2022.8.10.4

| | | |
|---|---|---|
| Abstraction | Data mining abstract from the data warehouse | Machine learning reads machine |
| Techniques involve | Data mining is more of research using methods like machine learning | Self-learned and trained system to do intelligent tasks. |
| Scope | Applied in the limited area | Can be used in a vast area. |

### 2.2 Educational Data Mining Techniques

Several data mining techniques can be used for educational data. Knowing the best technique to use to produce the best result can be complex. These techniques include classification, regression, and clustering. The most popular task to predict a student's performance is classification. There are several algorithms under classification task tasks have been applied to predict students' performance. Among the algorithms used are Decision trees, Artificial Neural Networks, Naive Bayes, K-Nearest Neighbor, and, Support Vector machines [8]. We need to know the different categories which EDM techniques are classified under.
There are three (3) main types of machine learning namely;
1. Supervised learning
2. Unsupervised Learning
3. Reinforcement Learning

## 3. SYSTEM ANALYSIS AND DESIGN

### 3.1 Analysis of the Proposed System

To perform this research, we first need to understand the domain from which we are going to collect the data from. The data to be used will be from the newly admitted students. These Students are all expected to take a mandatory examination called JAMB (Joint Admission and Matriculation Board). This Examination consists of four subjects, English is mandatory, to be taken with three other subjects of the student's choice. The result from this examination is used to admit students into the university and different departments. The JAMB cut-off mark for each department varies. For example, Medical Students will need to have a JAMBscore of 200 or higher while Art Students will require just about 180 to be admitted. This data will be useful in understanding the newly admitted students. The data will include first-year students' JAMB data. This data will include the Age, Gender, JAMB score, State of Origin, Department, and course of the Students.
In the building of this system, the concept of educational data mining was well incorporated as discussed in the previous session. The collected data will be organized in a Microsoft Excel sheet.

There are usually four educational data mining steps to be followed, they are;

1. Data collection:
2. Data preprocessing
3. Apply data mining algorithm
4. Interpret and evaluate results

### 3.2. Attributes Description

Here, the attributes for this dissertation are described along with their possible values as in Table 3.1

**Table 3.1 Attributes description and possible values**

| Attribute | Description | Possible Values |
|---|---|---|
| Sex | Student's gender | {Male =1, Female=2} |
| Id | Matric Number | {Numbers} |
| State Of Origin | Student's State of Origin | {Abia =1, FCT = 37} |
| Age | Student's Age | {Numbers} |
| TotalScore | Student's Total JAMB Score | {Numbers} |
| dept | Student's department | {Business Administration =1, Civil Engineering = 3} |



DOI: 10.31695/IJASRE.2022.8.10.4

| | | |
|---|---|---|
| Duration | Duration of the course of study | {Numbers} |
| Course | Student's course | { Business Administration =1, Civil Engineering = 3} |
| DeptSn | Department serial number of Student | {Numbers} |

### 3.3 Data Preprocessing and Preparation

During this phase, data would have to be pre-processed for the mining techniques. Some irrelevant attributes would be eliminated, e.g. student name, LGA, first choice, faculty, and duration. Each student will have the following attributes: Id, Age, Sex, StateOfOrigin, Dept, and TotalScore.

The classification was chosen because the objective of classification techniques in educational data mining is to identify what are the important factors that contribute to categorizing students' final grades. Decision trees are the most popular classification technique in data mining. They represent the group of classification rules in a tree form, and they have several advantages over other techniques.

- The simplicity of its presentation makes them easy to understand
- They can work for different types of attributes, nominal or numerical
- They can classify new examples fast.

This would be done using R.

**Types of Data;**

There are two types of data used in this research and they are;

1. Categorical data
2. Numerical data

**Categorical data:** A categorical variable is a variable that can take on one of a limited and usually fixed number of possible values, assigning each individual or other unit of observation to a particular group or nominal category based on some qualitative property. Categorical data is data that is collected in groups or topics; the number of events in each group is counted numerically. Qualitative or categorical data have no logical order, and can't be translated into a numerical value. Eye color is an example; because 'brown' is not higher or lower than 'blue'. For this research, the categorical data are Sex, Dept, and State of origin.

**Numerical data:** Numerical data is data that is measurable, such as time, height, weight, amount, and so on. You can help yourself identify numerical data by seeing if you can average or order the data in either ascending or descending order. The numerical data to be used in this research are Age, Sex, StateOfOrigin, TotalScore, and Dept.

**Table 3.2: The data before applying the coding scheme**

| Serial No | StateOf Origin | Sex | Age | Total Score | Department | TotalScore Description |
|---|---|---|---|---|---|---|
| 1 | IMO | M | 18 | 276 | Business Administration | High |
| 2 | ABIA | M | 24 | 270 | Business Administration | High |
| 3 | ANAMBRA | M | 21 | 256 | Business Administration | High |
| 4 | CROSS-RIVER | M | 24 | 255 | Business Administration | High |
| 5 | KOGI | F | 23 | 254 | Business Administration | High |
| 6 | AKWA-IBOM | F | 17 | 254 | Business Administration | High |
| 7 | KWARA | F | 21 | 252 | Business Administration | High |
| 8 | EDO | M | 18 | 250 | Business | High |



DOI: 10.31695/IJASRE.2022.8.10.4

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | Administration | |
| 9 | ENUGU | M | 26 | 248 | Business Administration | High |
| 10 | BENUE | F | 23 | 248 | Business Administration | High |
| 11 | KOGI | M | 20 | 248 | Business Administration | High |
| 12 | KOGI | F | 17 | 247 | Business Administration | High |
| 13 | DELTA | F | 20 | 247 | Business Administration | High |
| 14 | ENUGU | F | 21 | 246 | Business Administration | High |

**Table 3.3: The data after applying the coding scheme**

| Serial No | StateOfOrigin | Sex | Age | TotalScore | Dept | TotalScore Description |
|---|---|---|---|---|---|---|
| 1 | 16 | 1 | 18 | 270 | 1 | High |
| 2 | 1 | 1 | 24 | 276 | 1 | High |
| 3 | 4 | 1 | 21 | 256 | 1 | High |
| 4 | 9 | 1 | 24 | 255 | 1 | High |
| 5 | 22 | 2 | 13 | 254 | 1 | High |
| 6 | 3 | 2 | 17 | 254 | 1 | High |
| 7 | 23 | 2 | 21 | 252 | 1 | High |
| 8 | 23 | 1 | 18 | 250 | 1 | High |
| 9 | 14 | 1 | 26 | 248 | 1 | High |
| 10 | 7 | 2 | 23 | 248 | 1 | High |
| 11 | 22 | 1 | 20 | 248 | 1 | High |
| 12 | 22 | 2 | 17 | 247 | 1 | High |
| 13 | 10 | 2 | 20 | 247 | 1 | High |
| 14 | 14 | 2 | 21 | 246 | 1 | High |

### 3.4 Implementation Tools

The tools used for this study include; R and Python programming languages and Excel Spreadsheet.

### 3.5 Methodology

The two methodologies to be used are Clustering and Classification.

#### 3.5.1 Clustering

This is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense. It is used to identify groups of similar objects in multivariate data sets. Clustering is a method of unsupervised learning and a common technique for statistical data analysis used in many fields. For this study, I will be using the K-means clustering algorithm. K-means clustering is a popular unsupervised machine learning algorithm aimed at partitioning several observations into k clusters in which objects in each cluster are similar to each other. It does this to discover the underlying patterns. It looks for a fixed number of clusters in a dataset.

##### 3.5.1.1 Algorithmic Steps for K-Means Clustering

Let $X = \{x_1,x_2,x_3,\ldots\ldots,x_n\}$ be the set of data points and $V = \{v_1,v_2,\ldots\ldots,v_c\}$ be the set of centers.



DOI: 10.31695/IJASRE.2022.8.10.4

1) Randomly select *'k'* cluster centers.

2) Calculate the distance between each data point and the cluster center.

3) Assign the data point to the cluster center whose distance from the cluster center is the minimum of all the cluster centers.

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

5) Recalculate the distance between each data point and newly obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat step 3).

### 3.5.2 Classification

Classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observations.

#### 3.5.2.1 Decision Tree

It builds a top-down tree-like model from a given dataset's attributes. The decision tree is a predictive modeling technique used for predicting or categorizing given data objects based on the previously generated model using a training dataset with the same features (attributes). The structure of the generated tree includes a root node, internal nodes, and leaf (terminal) nodes. The root node is the first node in the decision tree which have no incoming edges and one or more outgoing edges; an internal node is a middle node in the decision tree which have one incoming edge and one or more outgoing edges; the leaf node is the last in the decision tree structure which represents the final suggested (predicted) class (label) of a data object. For this study, TotalScore description is the target variable, and Age, Sex, StateOfOrigin, and dept are the predictors.

#### 3.5.2.2 Decision Tree Algorithm

Step 1: Select the feature (predictor variable) that best classifies the data set into the desired classes and assign that feature to the root node.

Step 2: Traverse down from the root node, whilst making relevant decisions at each internal node such that each internal node best classifies the data.

Step 3: Route back to step 1 and repeat until you assign a class to the input data. [9].

#### 3.5.2.3 Classification and Regression Tree (Cart):

It is a dynamic learning algorithm that can produce a regression tree as well as a classification tree depending on the dependent variable which is the factor variable. [10].

**a. Information Gain:**

$$\text{Information Gain} = I(p, n)$$

$$I(p, n) = \frac{-p}{p+n} \log_2 \left(\frac{p}{p+n}\right) - \frac{n}{p+n} \log_2 \left(\frac{n}{p+n}\right)$$



DOI: 10.31695/IJASRE.2022.8.10.4

Here what are P and n? So to find p and n we check our class attribute or outcome which is binary (0, 1). So for p, we take the true value 1 (in case of binary) and for no, we take the false value 0 (binary value). We go deeper into the mathematical part just here introduction.

**b. Entropy:**

Entropy is used to create a tree. We find our entropy from an attribute or class.

Entropy $E(A) = \sum_{i=1}^{v} {}^{p}{}_{i} + n_i / {}^{p+n}(I(p,n))$

**c. Gain:**

Gain = information entropy
Gain = l (p,n)-E(A)
Gain is used to finding one-by-one attributes of our training set.

## 4. DESIGN AND IMPLEMENTATION

The main purpose of implementation is usually to verify that the algorithm is correctly implemented and functioning as planned. It is used to identify errors in the algorithm and to improve accuracy and reliability.

When testing it is important to know how to choose a test case that is, a set of inputs, execution conditions, and expected results. A good test case shouldn't be too simple or complex.

### 4.1. Decision Tree

For this dissertation, a decision tree was used to classify data into High, Average, and Low based on their JAMBscore. This JAMBscore comprises their score in English and any other 3 subjects of their choice.

**Table 4.1 Classification of JAMB Scores**

| TotalScore(JAMBscore) | Description |
|---|---|
| 180-199 | Low |
| 200-239 | Average |
| 240-Above | High |

When writing the code for the decision tree, the data was split into two; the training set and the testing set. The training set will be used to create a decision tree and the result from the decision tree will be used on the test data to test the accuracy of the classification. The data was divided into 2, 80% for training and 20% for testing. The total number of students for this study is 1428; there are 1142 in training and 286. There are 37 states of origin including 28 departments and the age difference is from 15 to 45. The training test was classified, and the result of the classification is shown in Figure 4.1.



DOI: 10.31695/IJASRE.2022.8.10.4

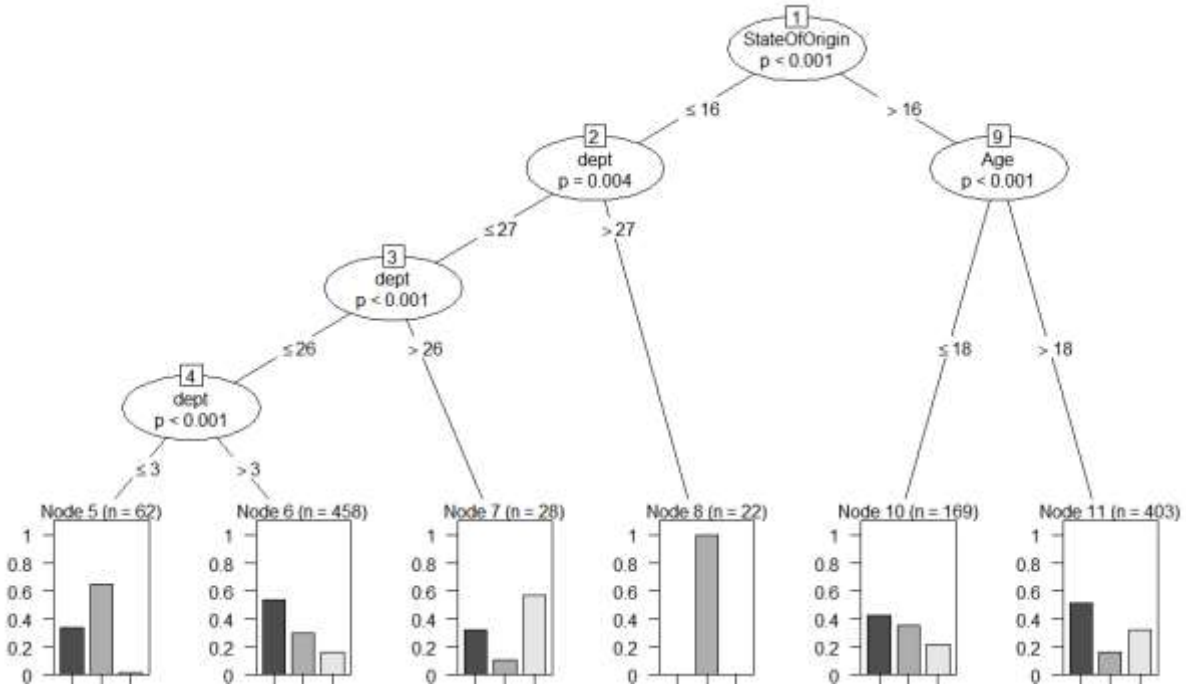



**Figure 4.1 Classifications by the State Of Origin, Age, and Department**

### 4.1.1. JAMBSCORE, State Of Origin, Department and Age Analysis

This decision tree in figure 4.1 represents the State of Origin, Age, and department of the students.

From the tree above, it is easy to compute that population from Jigawa, Kaduna, Kano, Katsina, Kebbi, Kogi, Kwara, Lagos, Nasarawa, Niger, Ogun, Ondo, Osun, Oyo, Plateau, Rivers, Sokoto, Taraba, Yobe, Zamfara and FCT.

i. Above 18 years (50% scored average, 15% scored high and 30% scored low)

ii. Below 18 years (40% scored average, 30% scored high and 20% scored low)

Students from Abia, Adamawa, Akwa-Ibom, Anambra, Bauchi, Bayelsa, Benue, Borno, Cross river, Delta, Ebonyi, Edo, Ekiti, Enugu, Gombe, Imo.

i. In Medical Sciences Department, 100% scored High, 0% scored average, and 0% scored low in JAMB.

ii. In the Agriculture department, 50% scored high, 30% scored average and 10% Scored low

iii. In Counseling and Educational Psychology, Educational Management, Science and Environmental Education, Arts and social science education, Banking and finance, Biological science, Computer science, Economics, Electrical, and Electronic Engineering, Electrical Engineering, Geography, and environmental protection, History, Linguistics, and African language, Mathematics, Mechanical Engineering, Philosophy and religion, Physics, Political Science, Public Administration, Sociology, Statistics, Theatre Arts and Veterinary medicine departments ( 50% scored average, 30% scored high and 10% scored low)

iv. In Business Administration, Chemical Engineering, and Civil Engineering departments (60% score high, 30% scored average, and 15% scored low)

From this analysis, the medical science department and agriculture department have the highest JAMB scores followed by Business administration, chemical engineering, and civil engineering.

Also, Students from South-East, South-South, and North-East Geo-political zones performed better than those from North-West, North-Central, and South-West.



DOI: 10.31695/IJASRE.2022.8.10.4

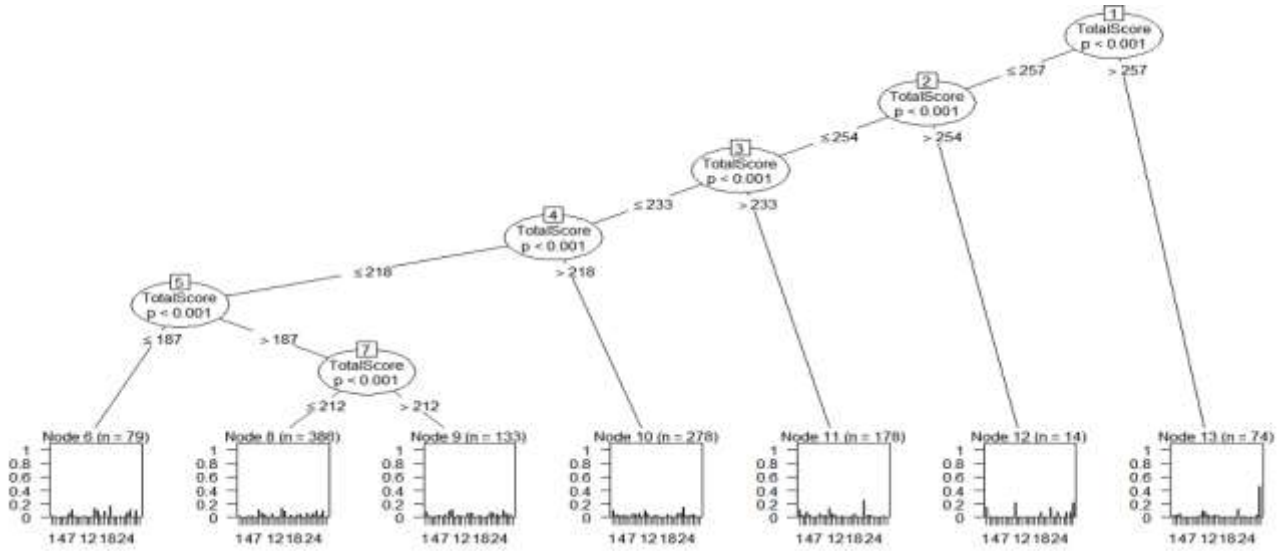



**Figure 4.2: Classification o by Department and JAMBscore**

### 4.1.2. JAMBSCORE and Department Analysis

The students were also classified into different departments and the tree in figure 4.2 showed that;

i. 74 students scored above 257 in JAMB (50% from Medical science, 15% from Political science, 10% from Computer Science, and 8% from economics departments)

ii. 14 students scored between 255 and 256 in JAMB (20% from Medical science, 22% from computer science, and 18% from business administration departments)

iii. 178 students scored between 234 and 254 in JAMB (30% from Political Science, 10% from business administration, and 12% from Computer science departments)

iv. 278 students scored between 233 -219 in JAMB (10% from business administration, 12% from economics, 18% from public administration departments)

v. 133 students scored between 213- 218 in JAMB (10% from biological science, 9% from theatre arts, 8% from banking and finance departments)

vi. 386 students scored between 212 – 188 in JAMB (10% from arts and social science, 16% from geography and environmental protection, and 10% from Agriculture departments)

vii. 19 students scored 187 and below in JAMB (18% from philosophy and religion, 16% from geography and environmental protection, 10% from Arts and Social Science education, and 10% from Theatre arts departments).



DOI: 10.31695/IJASRE.2022.8.10.4

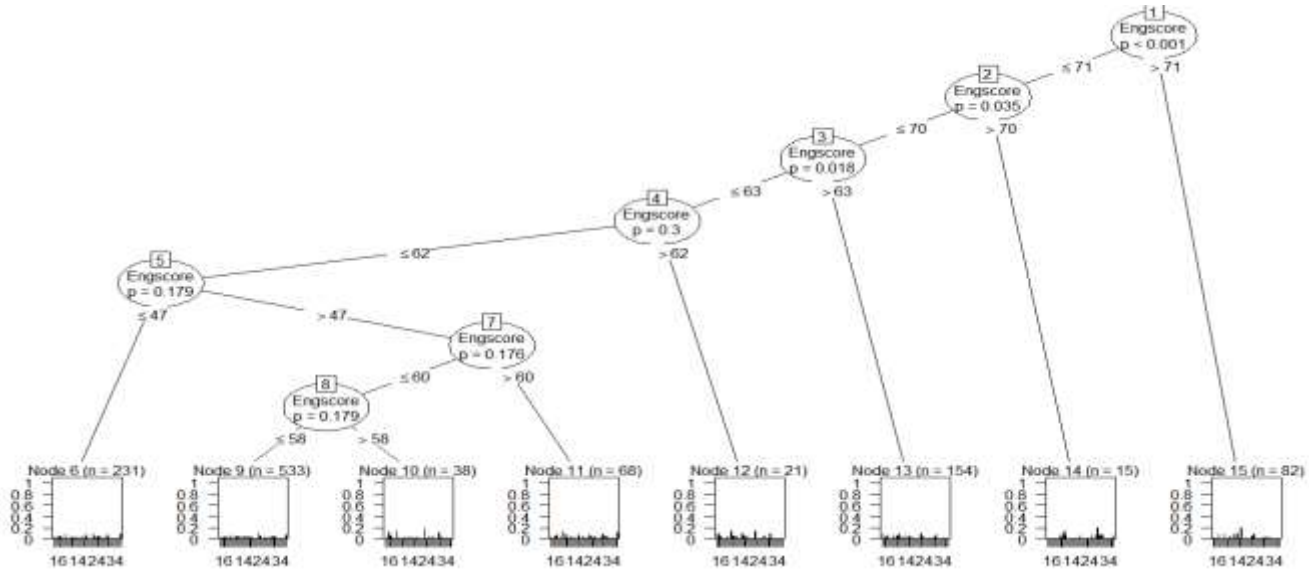



**Figure 4.3: Classification by English Score and the State Of Origin**

### 4.1.3. English Score and State Of Origin Analysis

In this analysis students who had high scores in English were mostly the from Imo, Enugu, Anambra, Ondo, and Delta States respectively while those with the lowest scores in English are from Zamfara and Abuja States.

### 4.1.4. Subjects and State of Origin Analysis

Students who sat for Music, Arab and Hausa subjects were from Kwara and Katsina states. Students who sat for Geography and CRS were mostly from Benue State. Students who offered government and economics were mostly from Kogi and Kwara states respectively.

### 4.1.5. Subjects and JAMBSCORE Analysis

Students who offered physics had the highest JAMB scores while students who offered History, Yoruba, Art, Igbo, Music, Hausa, and Arabic had the lowest JAMB scores. Students who scored above 70 in English had the highest JAMB scores.

### 4.1.6. Result of Analysis

From this analysis, out of the 1142 students trained, 266 had high JAMB scores, 797 scored average and 19 students scored low. The low scores came from Arts courses, the average scores came mostly from social science courses and the high JAMB scores came from science courses. Also, Students from South-South and north-east geo-political zones performed better than those from north-west, north central, and south-west geo-political zones. Students who offered Physics and those who scored higher than 70 in English had the highest JAMB scores.

## 4.2. K- Means Clustering

For k – means clustering, the records were clustered as shown in Figure 4.3 and the number of clusters is given. We graph the relationship between the number of clusters and Within Cluster Sum of Squares (WCSS) then we select the number of clusters where the change in WCSS begins to level off (elbow method). For this data, 9 clusters are given, 3 and 6 are the points where the clusters begin to level off but 6 will be used for the value of k.



DOI: 10.31695/IJASRE.2022.8.10.4

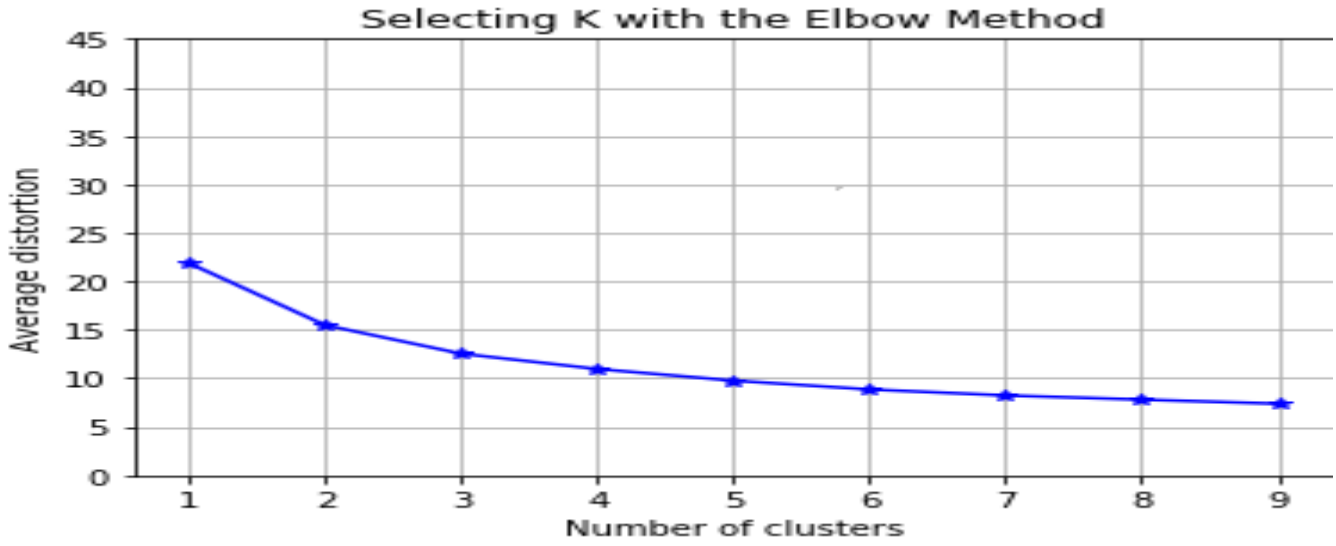



**Figure 4.4: Determining the number of clusters**

### 4.2.1. JAMBSCORE and State Of Origin Analysis

As shown in figure 4.4, the data has been clustered into 6 clusters, from the image it's easy to depict that JAMBscore from 180 – 198 is more from Niger and Ogun states, JAMBscores from 200 -210 are more from Borno, cross river and delta. JAMBscore of 210-218 are from Ogun and Ondo States, JAMBscores from 220-239 are from Cross-River state, JAMBscore of 240 – 269 is from Imo state and JAMB scores from 270 and above are from Imo state. It also shows that there are more students from Benue, Kogi, Imo, Anambra, and Abuja states. There are fewer students from Bayelsa, Jigawa, Zamfara, Rivers, Sokoto, Katsina, Yobe, and Kebbi.

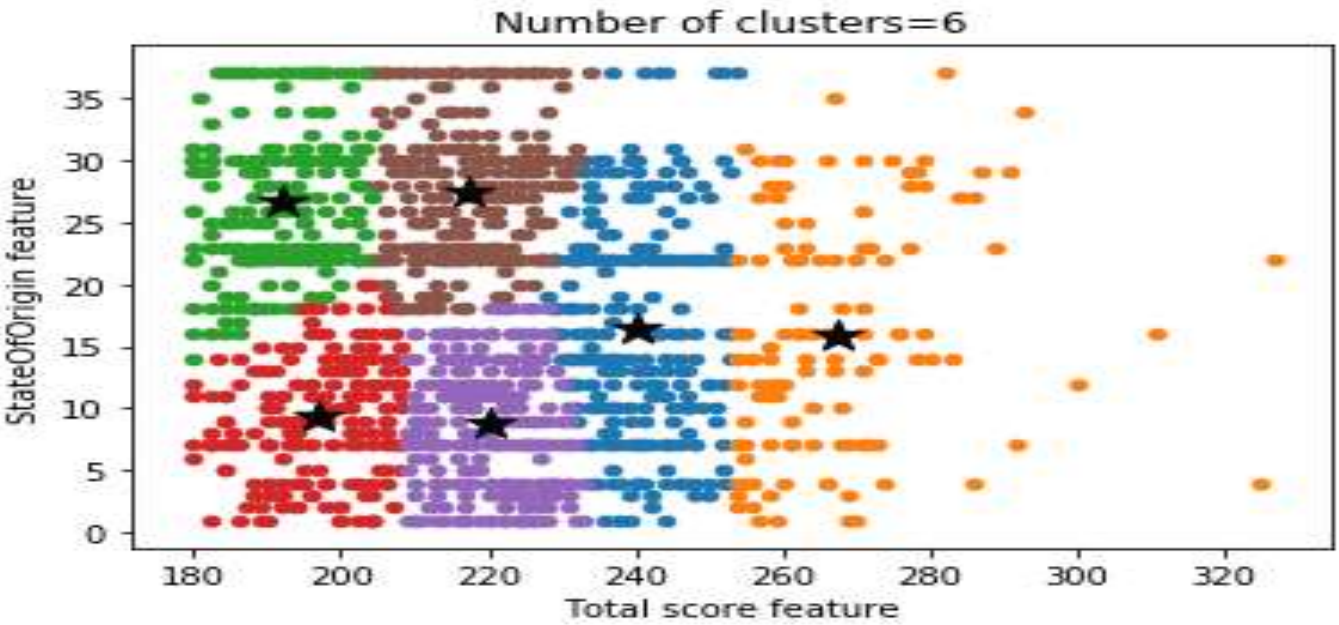



**Figure 4.5: Clustering by JAMB score and state of origin**

### 4.2.2. JAMBSCORE and Age Analysis

As shown in figure 4.5, students with JAMBscores from 180 to 196 are mostly 21 years old, students with JAMBscores from 197 – 210 are mostly 21 years too, students with JAMBscores from 211-230 are mostly 20 years old, students with JAMBscores from 231 and 250 are mostly 20 years old, students with JAMBscores from 251 -265 are mostly 19-year-olds and students with JAMBscores from 266 and above are mostly 19 and 18-year-olds



DOI: 10.31695/IJASRE.2022.8.10.4

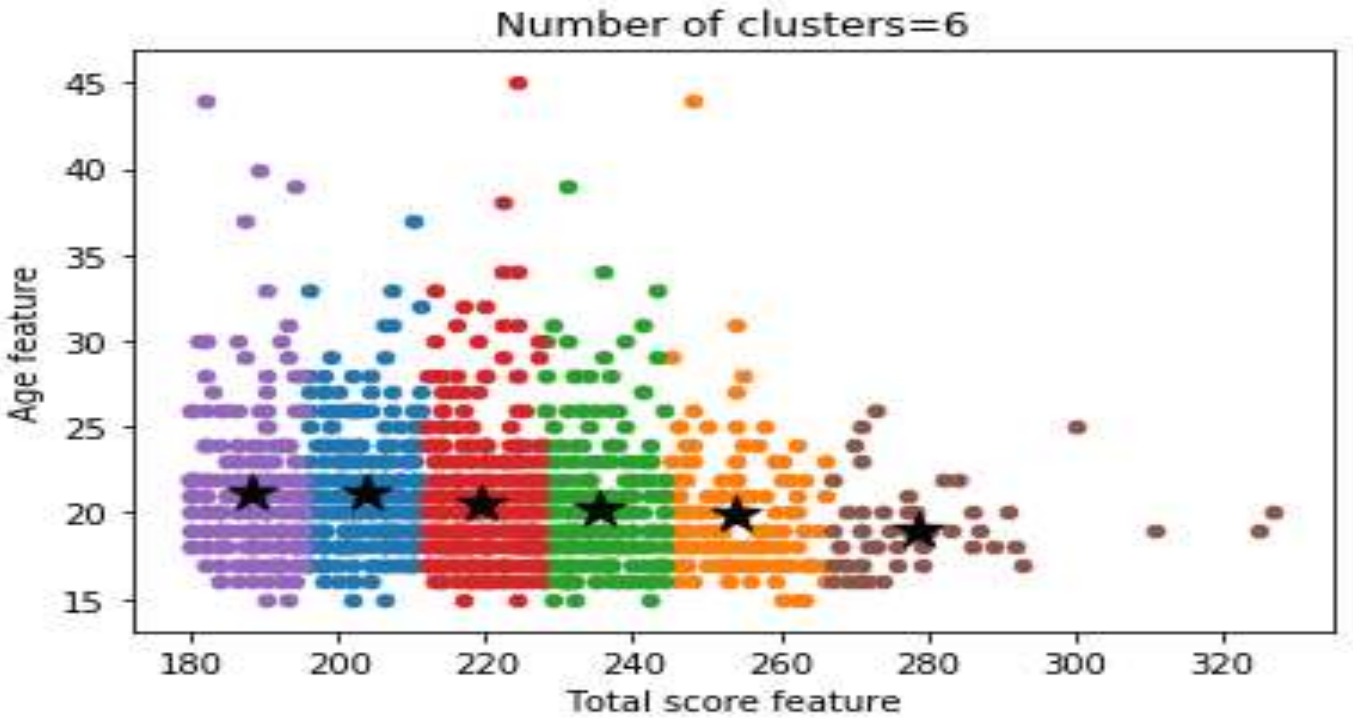



**Figure 4.6: Clustering by JAMBscore and Age**

### 4.2.3. JAMBSCORE and Department Analysis

As shown in figure 4.6, we can say that Political Science, Public Administration, Sociology, Statistics, Theatre Arts, and Veterinary medicine departments have more people that scored between "220 – 240" in their JAMB. The Medical Science department has the highest JAMBScores. Computer Science, Political Science, Science, and environmental education, and business administration have more students who scored between "241-260" in JAMB. Statistics, theatre arts, Agriculture, History, Linguistics, and African language and Arts and Social Science education departments have more students with JAMBscores between "180-195". Medical Science, Political science, Electrical and Electronics Engineering, Computer Science, and Civil Engineering departments have students with the highest JAMBscores. Theatre arts, Physics, Linguistics and African language, History, Science, and environmental education have students with the least JAMBscores.

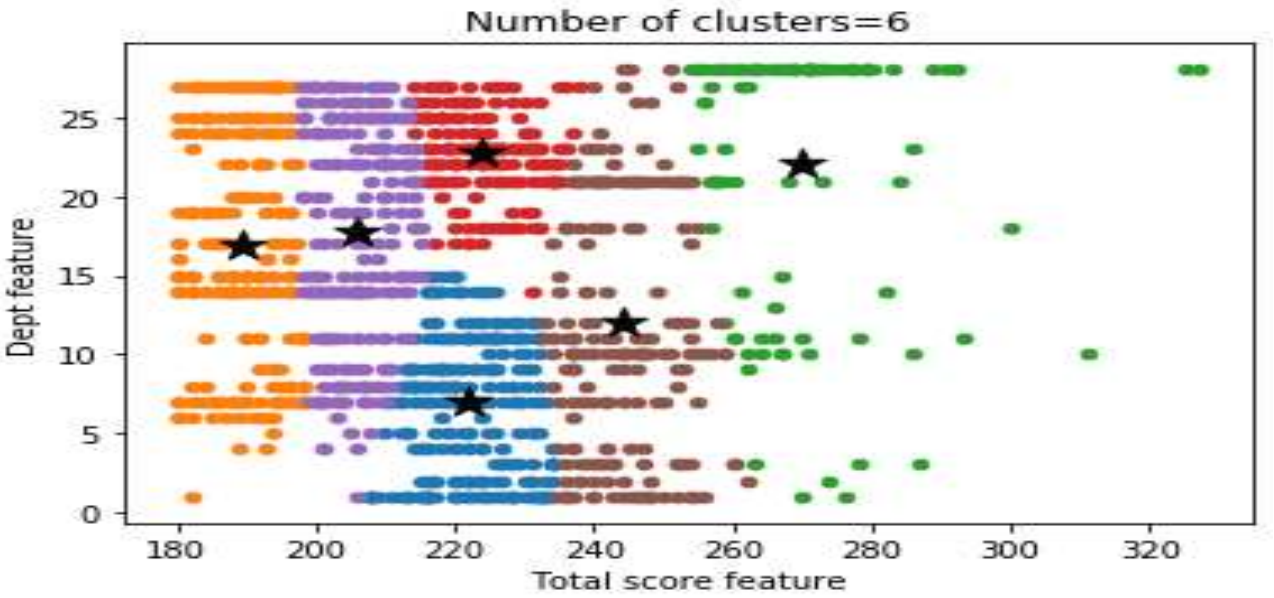



**Figure 4.7: Clustering by JAMB score and Department**



DOI: 10.31695/IJASRE.2022.8.10.4

### 4.3. Association Rules

The data was coded and each subject was put in a different column. Two packages were installed, arules and arulesViz. These are the two packages needed to find the rules in the dataset. The Apriori algorithm was used and 52 rules were generated. These rules were plotted in a graph and from the graph, it was deduced more females offered government, CRS, and literature while most males offered physics, chemistry, and mathematics.

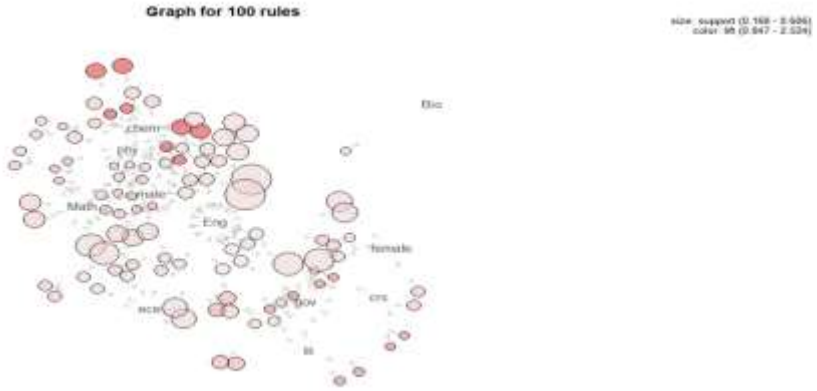



**Figure 4.8: Association rule graph**

In this dissertation, we were able to analyze the JAMB results of 1428 students to understand their performance using machine learning techniques.

### 5.1. Summary

This study sought to analyze the results of students. It has been observed that the algorithm has distinguished similar class objects, and as a result, multiple objects were noticed being classified to their nearest similar class. This dissertation has helped to analyze and categorize different students based on their departments, states of origin, and JAMBscore. The result of this dissertation will help educators give more attention to the students especially those from departments that score the least in JAMB.

### 5.2. Conclusion

This study shows the potential of data mining in higher education. Using techniques such as decision tree, k- means clustering, and association rule to analyze the students and their performances. The result of this dissertation showed that the low scores came from Arts courses, the average scores came mostly from social science courses and the high JAMB scores came from science courses. Also, Students from South-East, South-South, and North-East geo-political zones performed better than those from North-West, North-Central, and South-West geo-political zones. Students who offered Physics, and those who scored higher than 70 in English had the highest JAMB scores. Most males offered Physics, Chemistry, Mathematics, and Biology while most females offered Economics, Government, CRS, and Literature in English.

### 5.3. Recommendation

1. More research should be carried out on the dataset and more algorithms should be designed to use the dataset and predict with higher accuracy.
2. Anyone who wants to try to design more algorithms should consider existing algorithms used on the data and its result.
3. The articles referenced in this documentation can give a good understanding of this topic for improving on it.
4. The public and private educational institutions in Nigeria should pay more attention to efficient prediction analysis in education. This could be achieved by initiating and funding a dissertation on the prediction analysis in education, as it will aid educational administration and planning.



DOI: 10.31695/IJASRE.2022.8.10.4

### 5.4. Further Work

More studies should be done on Educational data mining. Universities and various tertiary institutions should adopt it to help them understand the students they admit and help educators plan their lecture notes. This will improve the quality of education in the country and the quality of graduates too.

## REFERENCES


1. Manjarres, A. V., Sandoval, L. G. M. and Suárez, M. J. S. (2018) Data mining techniques applied in educational environments: A literature review. Digital Education Review - Number 33, June 2018.
2. Prabha S.L, M Shanavas A.R.M, (2015) Application of Educational Data mining techniques in E-learning- A Case Study) / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (5) , 2015, 4440-4443.
3. Natarajan B, Kasozi J A, Chepete P and Arundhathi T (2016) impact of family background and Study skills on the Academic Performance of Higher Education Students: The case of Botho University, International Journal of Multidisciplinary and Current Dissertation.
4. Tetsuya, J. (2019) Why is Educational Data Mining Important in the dissertation, Towards data science
5. Kumar A.D., Selvam R.P and Kumar K.S. (2018) Review on Prediction Algorithms in Educational Data Mining, International Journal of Pure and Applied Mathematics Volume 118 No. 8 2018, 531-537.
6. Bhagoriya, N. and Pande P. (2017). Educational Data Mining In the Field Of Higher Education-A Survey, International Journal of Engineering Sciences & Dissertation Technology ISSN: 2277-9655.
7. Thakar, P., Mehta, A., and Manisha (2015) Performance Analysis and Prediction in Educational Data Mining: A Dissertation Travelogue, International Journal of Computer Applications (0975 – 8887) Volume 110 – No. 15, January 2015.
8. Shahiria, A. M, Husaina,W., and Nur'aini, A.R (2015). The Third Information Systems International Conference A Review on Predicting Student's Performance using Data Mining Techniques, School of Computer Sciences Universiti Sains Malaysia 11800 USM, Penang, Malaysia.
9. Lateef Z (2019), A Complete Guide On Decision Tree Algorithm, edureka.com.
10. Leonel J, (2019) Decision Trees, Medium.com




DOI: 10.31695/IJASRE.2022.8.10.4