Analyzing Student Academic Performance with a Decision Tree Predictive Workflow Model Using Classification Technique

Hasanien K Kuba[1], Oluwaseun A. Adelaja[2], Hussein Alkattan[3,4], Ali Subhi Alhumaima[5], Suhaf Ali Hussen[6] and Maad M. Mijwil[7]

[1]*University of Information Technology and Communications (UoITC), 10001 Baghdad, Iraq*
[2]*Department of Information Communication and Technology, Lagos State University, 102101 Lagos, Nigeria*
[3]*Department of System Programming, South Ural State University, 454080 Chelyabinsk, Russia*
[4]*Directorate of Environment in Najaf, Ministry of Environment, 54001 Najaf, Iraq*
[5]*Electronic Computer Centre, University of Diyala, 32001 Baqubah, Iraq*
[6]*College of Administration and Economics, Al-Iraqia University, 10001 Baghdad, Iraq*
[7]*Department of Computer Techniques Engineering, College of Engineering Technologies, Al-Iraqia Science University, 10001 Baghdad, Iraq*
hasanien.k.a@uoitc.edu.iq, oluwaseun.adelaja@lasu.edu.ng, alkattan.hussein92@gmail.com, alhumaimaali@uodiyala.edu.iq, Suhafali9@gmail.com, maad.m.mijwil@aliraqia.edu.iq

Keywords: KNIME, Decision Tree Workflow, WEKA, Prediction, Students' Academic Performance, Classification.

Abstract: In the area of Educational Data Mining and Learning Analytics, the prediction of students' academic performance is one of the most paramount aspects as it has improved both teaching approach of lecturers and the learning skills adopted by the students. This study aimed to design a decision tree predictive workflow model which performed classification approach on the students' dataset by utilizing the KNIME software. The students' dataset was splits into 80% training dataset while 20% equally for both test and validation dataset. The students were examined with MOODLE platform; confusion matrix generated with WEKA to obtain some evaluation measures (TP rate/Recall, FP rate, Precision, F-Measure, MCC, PRC area and ROC area) and some statistical representations (Bar Chart and Pie Chart) was plotted. We obtained a value of 74% overall accuracy of the model based on the students examined across the four departments as predicted in the WEKA and a value of 26% for the overall error rate.

1 INTRODUCTION

Most educational institutions such as the Colleges of Education, Polytechnics and Universities in Nigeria today have placed a higher priority in monitoring the performance of their students in examinations conducted every academic session (both the first and second semester). The academic performance of students is an important aspect in the educational sector, thus higher institutions should play a vital role by identifying and providing support to the low and average performing students at an early stage to mitigate failure rate, late graduation and dropouts [1]. The prediction techniques conducted by researchers on the students' academic performance during their duration of study in the higher institution is one of the most suitable approaches in achieving this [2]. Classification is a systematic data mining process for learning or building models which determine classes of given objects based upon the attributes of these objects, where the semantic of the classes is known beforehand [3]. The application of a learning algorithm to training dataset to learn a model and applying this model to assign labels to unlabeled instances are the major steps involved in classification techniques [4]. In recent times, most researchers especially the academicians have focus primarily on establishing learning processes that allow certain analytical tools understand students, their learning patterns, extracting vital information and knowledge from large educational database resulting to the utilization of this information to predict the academic performance of students [5]. Educational Data Mining and Learning Analytics (EDM/LA) use software such as RapidMiner, WEKA, Orange and KNIME for the purpose of effective analysis on student's academic performance [6]. The KNIME data analytics software

is used in this paper to model a workflow with interconnected nodes to produce a decision tree classifier which predicted the student's data sets in the csv file format which were initially loaded and read by one of the connecting nodes. The procedures taken in achieving the workflow simulation involves the following.

Loading a csv file in a file reader containing the input data set of students in the Faculty of Management Sciences consisting of Business Administration, Marketing, Public Administration and Project Management Technology department;

1) Creating a Column Filter for both input data set to filter the selected column to achieve the both the actual and predicted values in the outcomes;
2) A partitioning node which splits the input data into two partitions (test and train data set);
3) Decision tree learner which is the node that induces a classification decision tree;
4) A decision tree predictor which is the node that use the existing decision tree from the learner to predict the class value for new patterns;
5) Finally, the decision tree to image which is node that renders a decision tree view on an image.

The WEKA data analytics platform was also used to obtain the values for the TP rate, FP rate, Precision, Recall, F-Measure, MCC, ROC Area, PRC Area for the class considered which were both the student's department and their gender. Confusion Matrix for the class (department and gender) was also derived using the WEKA platform.

Decision Tree is a simple model for supervised classification techniques [7]. It basically implements classification of single discrete target feature [8]. Decision Tree is a structured classifier that consists of nodes which forms a rooted tree (this implies a directed tree with a node known as a root) which has no incoming edges [9]. All other nodes of a decision tree have exactly one incoming edge is known as the leaves/terminal or decision nodes while a node with outgoing edges are known as the internal or test node [10]. For decision trees outcome, each internal node splits the instance space into two or more sub spaces based on certain discrete function of the input attribute values [11].

2 RELATED WORKS

There are other researchers who have focused on the predicting student's performance by applying several analytical techniques. This section will mainly discuss the techniques adopted by those researchers.

Hamsa Hashmia et al. [12] conducted research focusing on the educational data mining aspect by developing a students' academic performance prediction model utilizing two selected classification methods which were decision tree and fuzzy genetic algorithms for both bachelor and masters' students in the Computer Science and Electronics and Communication department. The authors selected parameters such as the internal marks of bachelor students which combined their attendance marks and average marks obtained for their exams and assignment; while they used the admission score of the master's student which included their entrance and degree marks. This prediction model helped the lecturers to easily identify, classify their performances and take necessary actions to effectively improve their learning.

Kolo David K et al. [13] applied the Chi-Square Automatic Interaction Detection (CHAID) in generating the decision tree structure to predict the students' academic performance with the aid of IBM SPSS. The authors were able to discover that factor such as the students' financial status, their motivation towards learning and gender had an effect on their performance. The outcome of the prediction in their research helped the lecturers discover that 66.8% of the students passed while the remaining 33.2% failed. The work also revealed that much larger percentage of students was likely to have good grades and there is also a higher likely of male students passing than female students.

Gotardo M.A [14] utilized the J48 algorithm data mining technique to create the decision tree model which predicted students' performance in the (Data Structures and Algorithms) subject. The researcher used both the K-fold cross validation and Receiving Operating Characteristics (ROC) Curve for the model accuracy. The decision tree model further revealed in the research conducted that for students to pass the subject, they should have a grade higher than 66.12% in their Midterms and a grade higher than 72.30% in their finals. This prediction benefited both the students and professors, however enabling the professors to implement proactive measures in helping the students learn and ultimately improving their academic performance.

Bharadwa B.K and Pal. S [15] utilized the decision tree method to analyze and classify the data of 50 students and 8 attributes from the VBS Purvanchal University. The outcome was to provide crucial assistance to the lecturers to improve the results of their students.

Mehta.S.H et al [16] conducted research with WEKA machine learning platform by using the J48 classification algorithm to generate a decision tree which predicted students' academic performance, evaluated the teaching skills adopted by the lecturers and enhance the potentials of the students across other specialization. The researchers attained a classification accuracy of about 78.2% which was computed by the J48 algorithm.

Khin. K.L and San.S.L [17] using the ID3 (Iteractive Dichotomies 3) decision tree algorithm to predict and classify the IT undergraduate student's final grade based on their performance. The authors utilized information such as the students' attendance, class quizzes, presentation and marks obtained in their assignments to perform this prediction at the end the semester.

3 METHODS AND DATASET DESCRIPTION

The decision tree classification approach is used to predict the students' dataset by creating a workflow model in KNIME and also the same set of data were loaded in WEKA to obtain some evaluation parameters/measures such as (ROC Area, PRC Area, MCC, F-Measure, Recall, Precision, FP rate, TP rate/ Sensitivity and confusion matrix) for the classes (departments and gender) considered. The students' ages, gender, grades obtained in their examination, and departments were the attributes we utilized in achieving the analysis. The flowchart shown in Figure 1 describes steps implemented in the connected nodes to produce the workflow model in KNIME.

We utilized in this paper the information of students across four departments (Business Administration, Public Administration, Project Management Technology and Marketing) who had taken the examination known as Basic Statistics and Its Applications (STA 115). The student's grades were collected from the Learning Content Management System (LCMS) known as MOODLE in which they were examined as a Computer Based Test. Appendix A (Fig. A1 and Fig. A3) section shows the platform in which the students were being examined. MOODLE is an acronym for Modular Object-Oriented Dynamic Learning Environment [18]. MOODLE is one of the most user friendly and flexible global free open-source courseware product which is specifically designed to help educators provide graded assignments, examinations, forums, chats, share documents to enhance quality learning [19]. The course (STA 115) was graded over 70% marks and with a total number of 29 consisting of (15 female and 14 male) students who took the examinations across the four departments. The workflow model has ten individual connecting nodes with dataflow lines which executes various operations as shown in Figure 2.

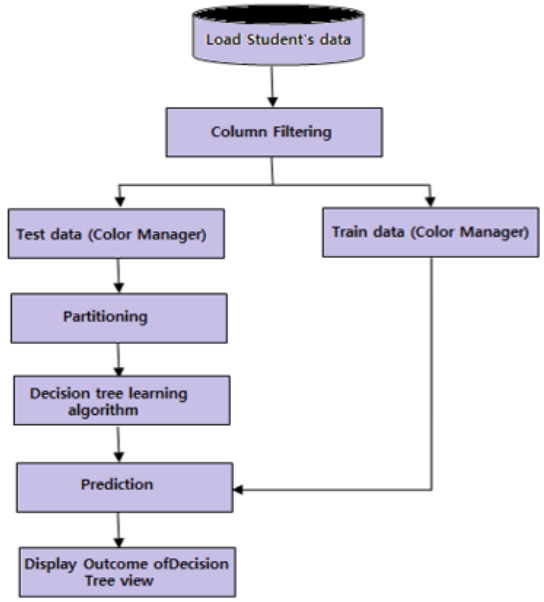


Figure 1: Flowchart of the workflow in KNIME.

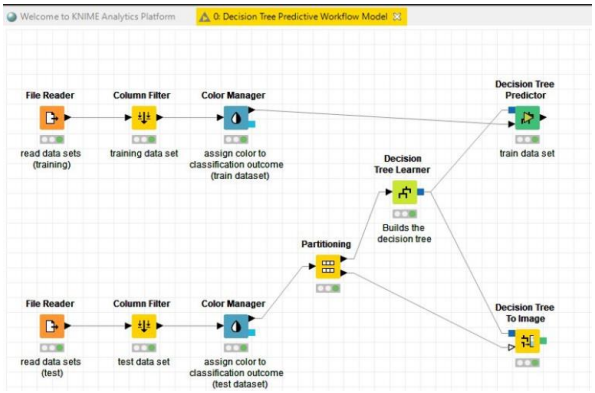


Figure 2: Workflow Model generated in KNIME platform.

The student dataset was loaded into the file reader nodes with a proportion of 80% of training data; the remaining 20% equally split for both test data and validation data to achieve an effective accuracy in prediction. The value of the accuracy of model after the prediction with WEKA revealed 0.74 which is equivalent to 74% and the overall error rate value as calculated as follows (1 – accuracy of model = 0.26)

which implies 26%. The Decision Tree to image node produced a pictorial view of the decision tree classifier generated from learner shown in Appendix A (Fig. A4). The confusion matrix for the classes (departments and the gender) obtained in the WEKA platform as shown in Appendix A (Fig. A5 and Fig. A6) below helps to understand calculation of the evaluation parameters/measures. The gender class produced a 2 by 2 matrix structure (Binary Class with two distinct data categories) with two identifiers a and b representing female and male students respectively, while the department class produced a 4 by 4 matrix structure with four identifiers a,b,c and d representing Business Administration, Public Administration, Marketing and Project Management respectively. For ease in calculating the evaluation formula assigned to the department class with four distinct categories, the 4 by 4 matrix will be converted into a 2 by 2 matrix structure. Confusion matrix is used for evaluating the performance of a machine learning classification model [20]. The confusion matrix schema is shown in Figure 3.

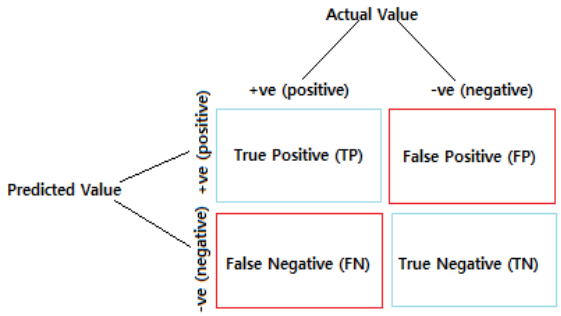


Figure 3: A Schema for confusion matrix.

4 RESULTS AND DISCUSSION

This section describes the calculation of evaluation metrics based on the confusion matrices presented in Appendix A (Fig. A5 and Fig. A6). By comparing the confusion matrix schema in Figure 3 with the gender-based confusion matrix shown in Appendix A (Fig. A6), the values of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) were identified for both male and female classes, as illustrated in Figure 4. These values form the basis for evaluating the performance of the classification model.

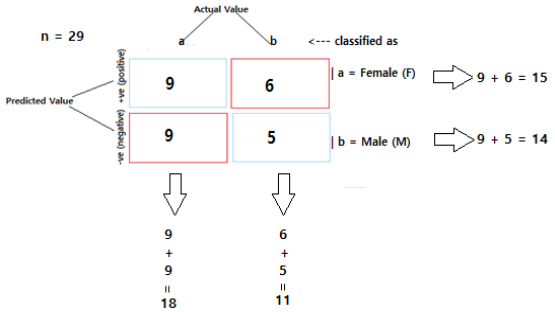


Figure 4: Confusion matrix for gender class with binary unique identifier: a) female: b) male.

We can obtain the TP rate (also known as Recall or Sensitivity, FP rate and other evaluation measures with the following (1)-(5):

$$\text{TP}\frac{\text{Rate}}{\text{Recall}} = \frac{\text{TP (True Positive)}}{\text{actual yes}}, \tag{1}$$

$$\text{FP rate} = \frac{\text{FP (False Positive)}}{\text{actual no}}, \tag{2}$$

$$Precision = \frac{TP\ (True\ Positive)}{predicted\ yes}; \tag{3}$$

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall}, \tag{4}$$

$$\begin{gathered} N = (TP \times TN) - (FP \times FN). \\ D_1 = (TP + FP)(TP + FN). \\ D_2 = (TN + FP)(TN + FN). \\ MCC = \frac{N}{\sqrt{D_1 \times D_2}}. \end{gathered} \tag{5}$$

4.1 Evaluation Metrics Calculation

The results presented in Appendix F (see Fig. A7 and A8) were used to compute the evaluation metrics for both female and male classes. These results indicate moderate classification performance with relatively low correlation values.

To calculate the TP Rate/Recall for Female (F) and Male (M):

$$TP\ Rate/Recall\ (Female) = \frac{9}{15} = 0.600,$$

$$TP\ Rate/Recall\ (Male) = \frac{5}{14} = 0.3571 \approx 0.357.$$

To calculate the FP Rate for Female (F) and Male (M):

$$\text{FP Rate (Female)} = \frac{9}{14} = 0.6428 \approx 0.643,$$

$$\text{FP Rate (Male)} = \frac{6}{15} = 0.400.$$

To calculate the Precision for Female (F) and Male (M):

$$\text{Precision (Female)} = \frac{9}{18} = 0.500,$$

$$\text{Precision (Male)} = \frac{5}{11} = 0.4545 \approx 0.455.$$

To calculate the F-Measure for Female (F) and Male (M):

$$F - Measure\ (Female) = \frac{2 * 0.5 * 0.6}{0.5 + 0.6} = \frac{0.6}{1.1} = 0.5454 \approx 0.545,$$

$$F - Measure\ (Male) = \frac{2 * 0.455 * 0.357}{0.455 + 0.357} = \frac{0.32487}{0.812} = 0.400.$$

To calculate the Matthews Correlation Coefficient (MCC) for Female (F) and Male (M):

$$MCC\ (Female) = \frac{[(9 * 5) - (6 * 9)]}{\sqrt{[(9+6) * (9+9) * (9+5) * (5+6)]}} = \frac{-9}{\sqrt{41580}} = \frac{-9}{203.91} = -0.044$$

$$MCC\ (Male) = \frac{[(5 * 9) - (9 * 6)]}{\sqrt{[(9+6) * (9+9) * (9+5) * (5+6)]}} = \frac{-9}{\sqrt{41580}} = \frac{-9}{203.91} = -0.044$$

4.2 Data Visualization and Analysis

Bar charts illustrating the distribution of students based on their grades and gender are presented in Figures 5 and 6, respectively.

Figure 7 presents the proportion of students across different departments.

Business Administration: $\frac{\text{Sum of students}}{\text{Total number of students examined}} * 360° = \frac{10}{29} * 360° = 124.1°$

Marketing: $\frac{\text{Sum of students}}{\text{Total number of students examined}} * 360° = \frac{10}{29} * 360° = 124.1°$

Project Mgt. Technology: $\frac{\text{Sum of students}}{\text{Total number of students examined}} * 360° = \frac{5}{29} * 360° = 62.1°$

Public Administration: $\frac{\text{Sum of students}}{\text{Total number of students examined}} * 360° = \frac{4}{29} * 360° = 49.7°$

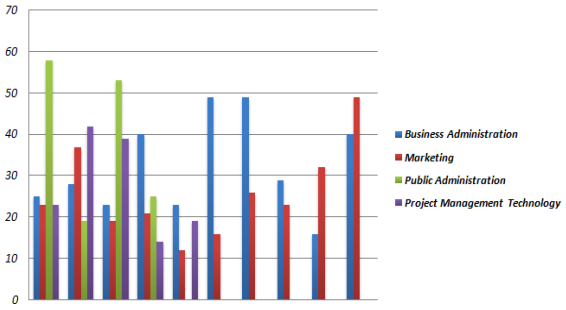


Figure 5: Distribution of students based on their grades.

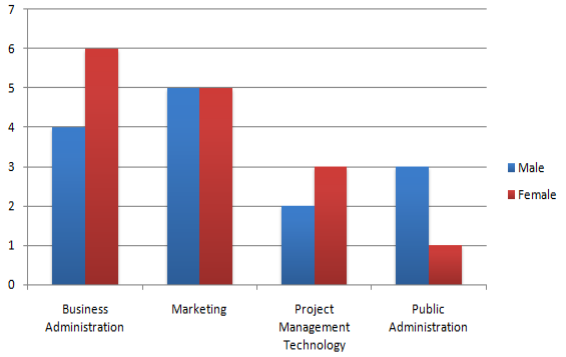


Figure 6: Distribution of students by gender.

Figure 7: Proportion of students across departments.

5 CONCLUSIONS

The dataset (students' grade sheet) utilized in the work revealed that only 4 students (consisting of 3 male and 1 female) out of a total of 10 from Business Administration department obtained above the average of 35 marks out of the total grade of 70 marks which the STA 115 examination was graded; also 2 students (consisting of 1 male and 1 female) out of a total of 10 from Marketing department obtained above the average of 35 marks; while 2 students (consisting of 1 male and 1 female) students out of the total of 5 from Project Management Technology obtained above the average of 35 marks; and only 2 (consisting of 2 female) students out of 4 from Public Administration obtained above the average of 35 marks. The remaining 19 students across the four departments (consisting 9 male and 10 female) obtained grades below the average of 35 marks. The decision tree classification approach adopted in this work to predict the performance of the students across the four selected departments based on the examination conducted will further help academicians make proactive decisions to monitor the grades of those who had low performance. This will also provide a friendly teaching verses learning environment between the students and the lecturers, thereby improving the results of the students in subsequent academic's sessions.

REFERENCES

[1] D. Alboaneen, M. Almelihi, R. Alsubaie, R. Alghamdi, L. Alshehri, and R. Alharthi, "Development of a Web-Based Prediction System for Students' Academic Performance," Data, vol. 7, p. 21, 2022.

[2] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," Computers in Human Behavior, vol. 73, pp. 247-256, 2017.

[3] W. Ibrahim, S. Abdullaev, H. Alkattan, O. A. Adelaja, and A. A. Subhi, "Development of a Model Using Data Mining Technique to Test, Predict and Obtain Knowledge from the Academics Results of Information Technology Students," Data, vol. 7, p. 67, 2022.

[4] S. Sumanthi and S. N. Sivanandam, Introduction to Data Mining and its Applications, Springer, Studies in Computational Intelligence, vol. 29.

[5] A. Daud, M. D. Lytras, N. R. Aljohani, F. Abbas, R. A. Abbasi, and J. S. Alowibdi, "Predicting Student Performance Using Advanced Learning Analytics," in Proc. 26th International World Wide Web Conference Companion (WWW 2017), Perth, Australia, Apr. 2017, pp. 415-421.

[6] R. Mikut and M. Reischl, "Data mining tools," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 1, pp. 431-443.

[7] G. Alice, "Decision Trees," University of Toronto Computer Science Department, 2021, [Online]. Available: https://www.cs.toronto.edu/~axgao/cs486686_f21/lecture_notes/Lecture_07_on_Decision_Trees.pdf.

[8] M. Vranić, D. Pintar, and Z. Skočir, "The use of data mining in education environment," in Proc. 9th International Conference on IEEE, Winchester, UK, Jul. 2007, pp. 243-250.

[9] K. Alsabti, S. Ranka, and V. Singh, "CLOUDS: A Decision Tree Classifier for Large Datasets," in Proc. Knowledge Discovery and Data Mining Conference (KDD-98), Aug. 1998.

[10] L. Rokach and O. Maimon, "Decision Trees," Department of Industrial Engineering, Tel-Aviv University, Data Mining and Knowledge Discovery Handbook, [Online]. Available: https://www.researchgate.net/publication/225237661_Decision_Trees.pdf.

[11] "Aktif Elektrotenik: Data Mining Decision Trees," [Online]. Available: https://aktif.net/en/data-mining-decision-trees/.

[12] H. Hamsa, S. Indiradevi, and J. J. Kizhakkethotteem, "Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm," Procedia Technology, pp. 327-332, 2016.

[13] D. K. Kolo, A. A. Solomon, and K. A. John, "A Decision Tree Approach for Predicting Students Academic Performance," [Online]. Available: http://www.mecs-press.net/ijeme.

[14] M. A. Gotardo, "Using Decision Tree Algorithm to Predict Student Performance," Indian Journal of Science and Technology, Feb. 2019.

[15] B. K. Bharadwaj and S. Pal, "Mining Educational Data to Analyze Students Performance," International Journal of Advanced Computer Science and Applications, pp. 63-69, Jan. 2012.

[16] S. H. Mehta and A. Ashish, "Predicting Students' Performance using J48 Decision Tree," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 4, pp. 122-129, 2019.

[17] K. L. Khin and S. N. San, "Using ID3 Decision Tree Algorithm to Student Grade Analysis and Prediction," pp. 1392-1395, Aug. 2019.

[18] "What is Moodle?," TechTerms, [Online]. Available: https://techterms.com/definition/moodle.

[19] B. William and M. Dougiamas, Moodle for Teachers, Trainers and Administrators of Remote-Learner.net, Moodle.org, 2005.

[20] J. Manuel, "Confusion Matrix," [Online]. Available: https://www.scribd.com/presentation/447230331/CONFUSION-MATRIX-pptx.

[21] R. Kundu, "Confusion Matrix: How to use it & Interpret Results," V7labs Machine Learning Blog, Sep. 2022, [Online]. Available: https://www.v7labs.com/blog/confusion-matrix-guide.

APPENDIX A

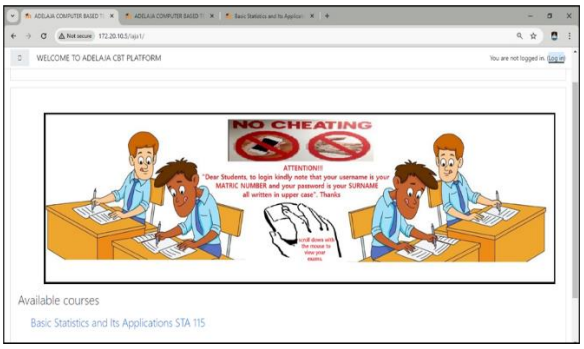


Figure A1: Moodle platform utilized for students' examination (Welcome Page).

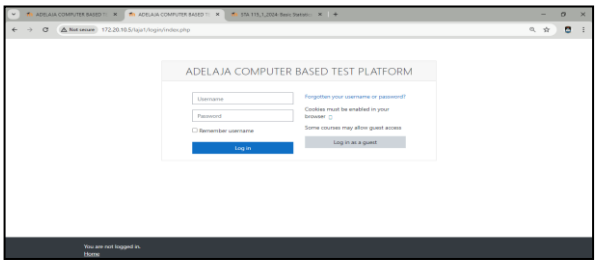


Figure A2: Moodle platform utilized for students' examination (Admin Login Page).

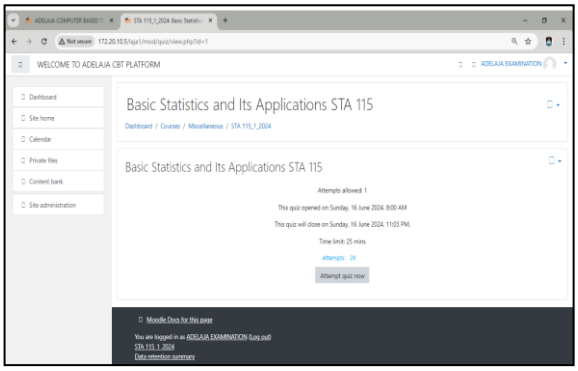


Figure A3: Admin page to monitor students who attempted the STA 115 examinations.

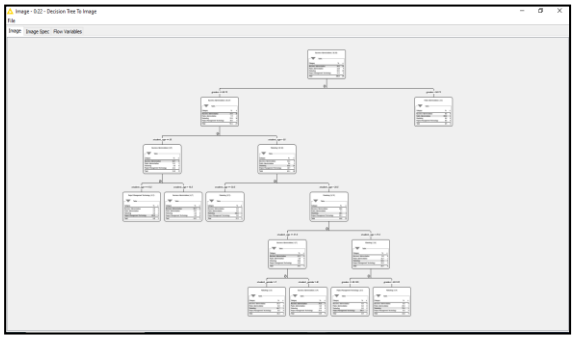
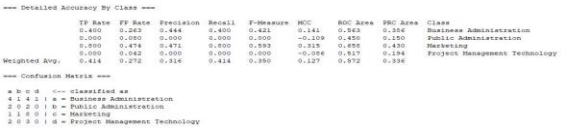


Figure A4: Decision tree dendrogram in KNIME.



```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.400    0.263      0.444   0.400      0.421    0.141     0.563     0.356    Business Administration
                 0.000    0.050      0.000   0.000      0.000   -0.109     0.450     0.150    Public Administration
                 0.500    0.474      0.471   0.500      0.593    0.315     0.655     0.430    Marketing
                 0.000    0.042      0.000   0.000      0.000   -0.056     0.517     0.194    Project Management Technology
Weighted Avg.    0.414    0.272      0.316   0.414      0.350    0.127     0.572     0.336

=== Confusion Matrix ===

 a b c d   <-- classified as
 4 1 4 1 | a = Business Administration
 2 0 2 0 | b = Public Administration
 1 1 8 0 | c = Marketing
 2 0 3 0 | d = Project Management Technology
```

Figure A5: Confusion matrix generated for the departments in WEKA.

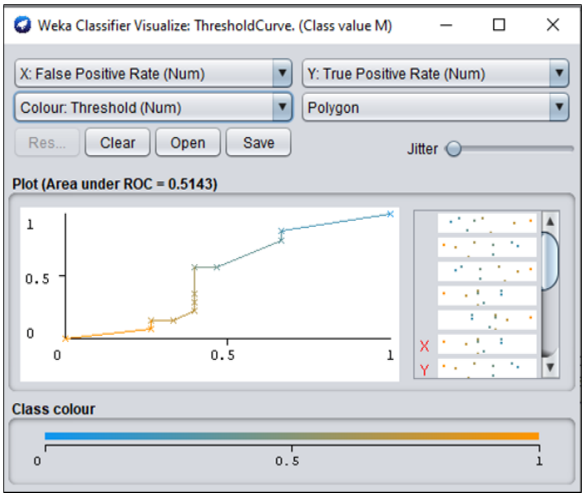


Figure A6: Confusion matrix generated for the genders in WEKA.

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.600    0.643      0.500   0.600      0.545   -0.044     0.514     0.592    F
                 0.357    0.400      0.455   0.357      0.400   -0.044     0.514     0.466    M
Weighted Avg.    0.483    0.526      0.478   0.483      0.475   -0.044     0.514     0.531

=== Confusion Matrix ===

 a b   <-- classified as
 9 6 | a = F
 9 5 | b = M
```

Figure A7: ROC for male in WEKA.

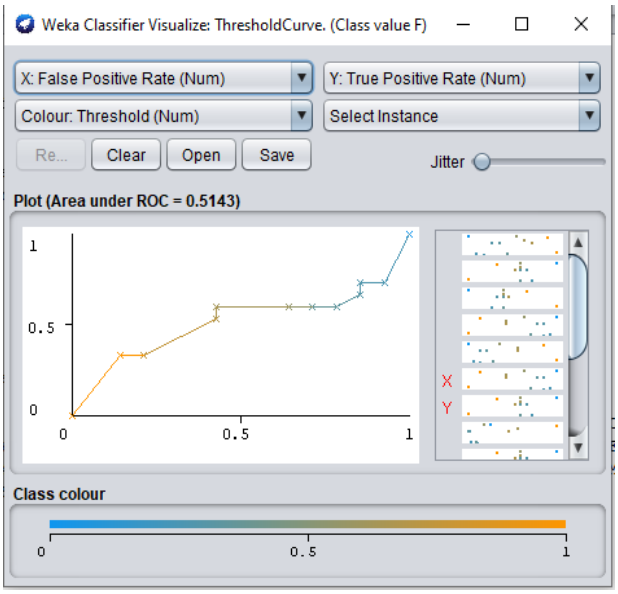


Figure A8: ROC for female in WEKA.