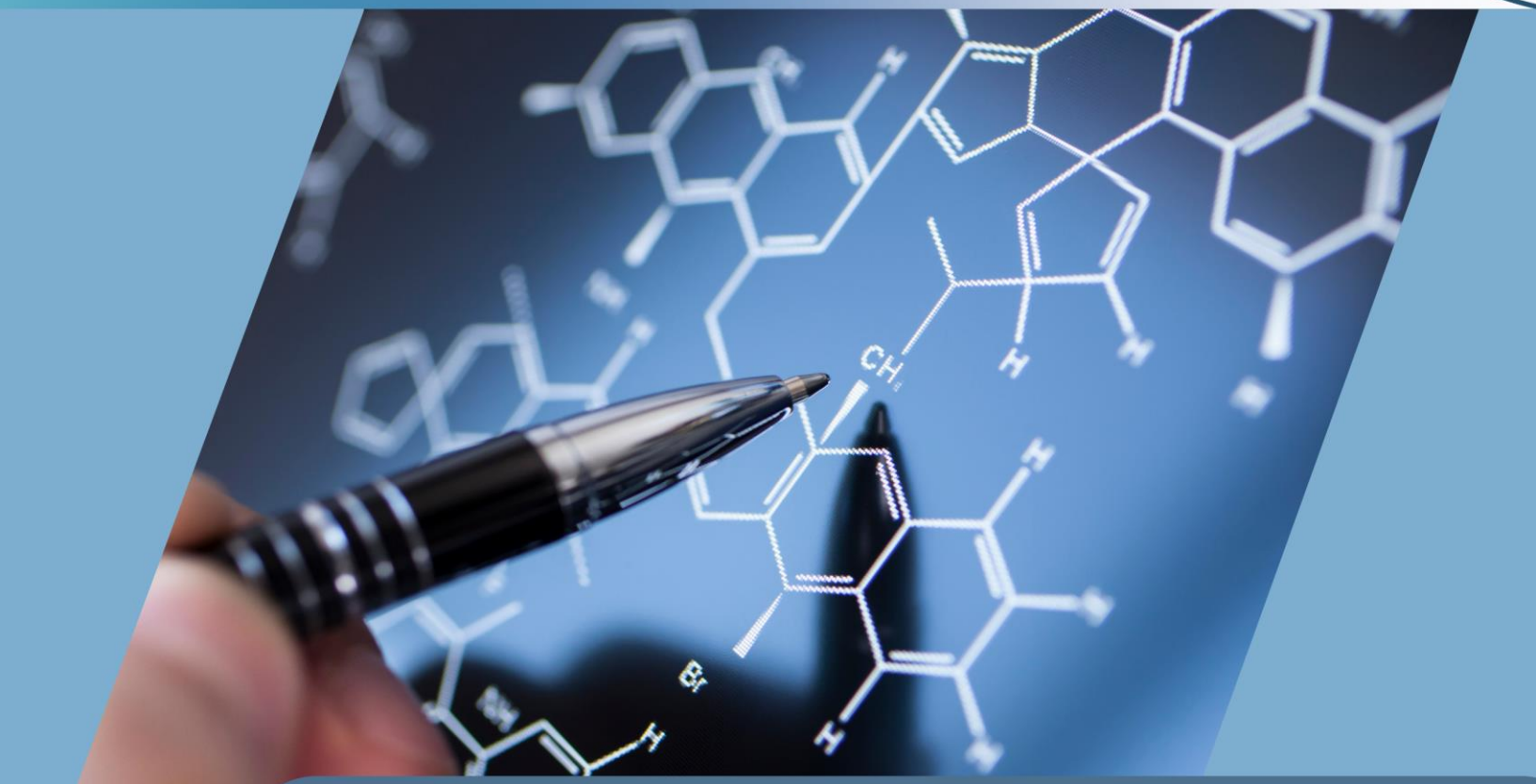
eISSN: 2789-858X

Scientific Journal for the Faculty of Science - Sirte University (SJFSSU)

Bi-annual, Peer- Reviewed, and Open Accessed e-Journal



VOLUME 4 ISSUE 1 APRIL 2024

10.37375/issn.2789-858X

Legal Deposit Number@National Library (Benghazi): 990/2021

jsfsu@su.edu.ly

Scientific Journal for the Faculty of Science-Sirte University Vol. 4, No. 1 (2024) 18-29

Scientific Journal for the Faculty of Science-Sirte University

Journal home page: http://journal.su.edu.ly/index.php/JSFSU/index

DOI: 10.37375/issn.2789-858X

Classifying the 1st Year Academic Performance of Nursing Students at Tobruk University via Data Mining with SQL and WEKA Tool

James Neil B. Mendoza, Dorothy G. Buhat-Mendoza

Nursing College, Tobruk University, Libya.

DOI: https://doi.org/10.37375/sjfssu.v4i1.2581

ARTICLE INFO:

Received: 17 January 2024

Accepted: 14 March 2024

Published: 17 April 2024

Keywords: *Classification, Data Mining, Extraction, Preprocess, SQL, WEKA*

A B S T R A C T

Data mining is a tool that can identify hidden patterns affecting academic success. The objective of this research is to investigate and classify the academic performance of first-year nursing students at Tobruk University. This study concentrates on the preliminary stage of data preprocessing and data mining classification. The methodology to classify academic performance includes data acquisition and preprocessing stage using SQL commands to extract student data from the university database and undergo basic cleaning and transformation. Initial classification and data analysis followed using the preprocessed data, further refined by the WEKA data mining tool algorithms including BayesNet, NaiveBayes, JRIP, and J48. Results of the preliminary data distribution and initial classification show that J48 is the most accurate model creator using regular classification (88.6619) and attribute selector (97.8261). Relative to the other three algorithms, J48 also recorded the highest precision, recall, F1 measure, and the lowest error measurement. The recorded Kapa stat of J48 (0.7779 and 0.9599) also proves the significance of the classification result, interpreted as substantial to near-perfect reliability scores respectively, which BayesNet and JRIP also attained. The results reveal that Finals (final exam result) attribute is the biggest factor in determining the descriptive Rating of a student's grade at the university. The created model can serve as a classifier for future test sets and may provide a foundation for further research and model development. Further modification will help discover what factors contribute to student success and what applicable interventions are needed to improve the academic achievement of students in the nursing program.

1 Introduction

Predicting and understanding student academic performance is crucial for educational institutions to improve learning outcomes and student success (Alyahyan & Düştegör, 2020; Cui et al. 2019). In the field of nursing education, ensuring strong academic foundations in the first year sets the stage for future clinical skills development and professional competence. This research focuses on applying data mining techniques to the academic performance (Nahar et al. 2021) of first-year nursing students at Tobruk University. Utilizing a mixed-methods approach, it combines SQL queries for data extraction and transformation (Kumar & Krishnaiah, 2012; Ordonez et al. 2014) with the WEKA data-mining tool for analysis and classification (Han et al. 2006). This paper presents the initial phases of the research, encompassing the preliminary data preprocessing stage (Garcia et al. 2016) and initial classification attempts (Espinosa et al. 2011).

The transition to nursing education marks a pivotal juncture, where aspiring healthcare professionals

Corresponding author: E-mail: james.mendoza@tu.edu.ly

embark on a challenging yet rewarding journey. While academic success paves the way for future competence and patient care (Keshavarzi, 2022), understanding the factors influencing student performance remains crucial for optimizing educational strategies and fostering excellence (Bressane et al. 2023). This study delves into the intricate landscape of academic performance among first-year nursing students at Tobruk University, employing data mining techniques to uncover hidden patterns and predictive insights (Namoun & Alshanqiti, 2020; Villarica, 2020). As such, large data sets can be processed and valuable information can be extracted from simple data using data mining (Feng & Fan, 2024).

Traditionally, student performance assessment relies on summative measures like grade point average (GPA) (Schwab et al. 2018). While these provide valuable snapshots, they often fail to capture the nuanced interplay of factors contributing to academic success (Zughoul et al. 2018). Data mining, with its ability to identify hidden patterns and relationships within large datasets (Roostaee & Meidanshahi, 2023), offers a powerful lens to delve deeper into this intricate landscape (Schwab, 2018). This study leverages the strengths of data mining to explore the complex interplay of academic, demographic, and other possible variables influencing first-year nursing students' performance (Goundar et al. 2022) at Tobruk University.

The first academic year (AY) is considered an important phase of the laying foundation for future success and this investigation concentrates on it. Two powerful tools: SQL for effective data extraction and processing (Mori et al. 2015), and WEKA for powerful data mining algorithms and classification tasks (Aher et al. 2011; Kabakchieva, 2013) will be utilized. This study does not stop at just identifying factors that influence performance. It was desired to use the information gained through data mining for developing focused interventions and educational measures (Barakeh et al., 2024). Through early identification of students likely to have trouble succeeding, the provision of individualized attention and direction, encouragement of their natural ability, and establishment of a solid base for their nursing endeavors. Furthermore, the outcomes of demographic variables can guide initiatives targeted at fostering equity and diversity in Tobruk University's nursing program.

The primary objective of the study is to investigate the academic performance of first-year nursing students at Tobruk University by applying data mining techniques. This data mining study may shed light on the intricate world of academic success. The study sought to reveal hidden patterns and predictive models in an effort of navigating through the academic maze, to guide interventions towards specific targets as future generations of proficient self-assured nurses advanced. The next section is the related study and literature. The materials and methods section follows, describing the data acquisition and preprocessing stage using SQL, where extracted data will be classified and analyzed by the WEKA mining tool. The results of the process performed will be presented and discussed in the data extraction result and data analysis results section.

2 Related Literature and Study

Finding knowledge from a large set of data is difficult to perform. One tool that stands out to analyze hidden patterns from a huge amount of data is data mining. Since it is impractical for data not to be utilized properly (Hussein et al., 2018), data mining procedures will primarily depend on data quality of the sources, requiring preprocessing to obtain dependable knowledge (Espinosa et al. 2011). Data mining is applied to different industries, but one of the emergent sectors is education (Villarica, 2020), as every academic year, a large amount of data is being generated (Gowri et al., 2017). Data mining with its several algorithms for the extraction of patterns and knowledge will aid in better decision-making (Roostaee, & Meidanshahi, 2023).

Before we can use WEKA for classification, data will undergo extraction and cleansing first. Preprocessing involves cleaning, integrating, and transforming extracted data from sources. Preparing a dataset for analysis requires patience and a lot of time since it involves complex SQL queries, joining of tables, and aggregation of columns (Ordonez et al. 2014). These aggregation functions by SQL include SUM, MIN, MAX, COUNT, and AVG to obtain a summary of data (Kumar & Krishnaiah), besides JOIN and conditional queries, which this current study will implement. Preprocessing follows as the accuracy of data mining classification will improve if missing values are attributed (Panda & Adhikari, 2020). Deletion of row and if possible imputation of missing value must be used to complete a data set. A possibility of skewed results may be present when a large set of complex data extracted has an outlier. Outlier discovery in data

mining means finding a pattern in the data set that may deviate from expected behavior (Dash et al., 2023). Generated data are noisy and dirty which is another preprocessing issue. Data cleansing adheres to better data quality making sure data is ready for the analytic phase (Ridzuan & Zainon, 2019). Performing validation and verification will ensure data quality.

After data cleansing, a selection of data mining techniques follows. One of the most useful data mining techniques is classification, a supervised method responsible for identifying previously hidden class labels (Kawade et al., 2020). In their study, they used WEKA to classify the academic performance of students and used the result to make future decision-making. J48 algorithm gains the highest accuracy relative to other methods used in their experiment. Their study used JRIP, NaiveBayes, and BayesNet together with J48 as their classification tool, which the current study will adopt. Another data mining study used WEKA to classify students who are academically good or poor in the government schools of the Vellore district in Tamil, Nadu (Gowri et al., 2017). The current study will try to classify students based on descriptive ratings of failed, passed, good, very good, and excellent. In the paper of Ahmed & Kaber, they also used classification algorithms J48 and JRIP to find the reasons behind the failure of students. They generated the JRIP rule and J48 pruned tree to analyze the result of their study using the data from 1st-year class results from 2017 to 2022 (Ahmed & Kaber, 2022).

The classifier algorithms that are popular among data miners are BayesNet, NaiveBayes, JRIP, and J48. Conditional probabilities are described graph-wise by the Bayesian Network, also known as BayesNet (Baranyi et al., 2019; Hussain et al, 2018). It uses a direct graph with nodes to represent random attributes and conditional dependencies that symbolize arbitrary variables (Almarabeh, 2017). The Bayesian Network improves speed, accuracy, and ease of computation for large databases. On the other hand, Naive Bayes is a simple classifier used for probabilistic learning and it shows great performance in terms of accuracy when attributes are independent (Almarabeh, 2017; Hussain et al, 2018; Pujianto et al. 2017). Data mining commonly uses JRIP, or Repeated Incremental Pruning to Produce Error Reduction, as a rule-based classification algorithm. It is an enhanced variant of RIPPER, or Reduced Error Pruning, renowned for its effectiveness and capacity to produce clear rules (Ahmed & Kabir, 2022; Walia et al. 2020). Finally, the J48 algorithm is an expansion of the ID3 algorithm created by Ross Quinlan. Frequently referred to as a statistical classifier, J48 is used to generate decision trees that are produced by the C4.5 algorithm (Almarabeh, 2017; Mishra et al. 2014).

3 Materials and Methods

The goal of the preliminary preprocessing stage is to clean, integrate, and transform data from sources. The cleansed dataset will then be applied to initial classification algorithms to detect possible associations and predictive models of academic performance. The university database, TUGS-CON Ver. 2 (Mendoza et al., 2017) is the primary source of data. Section 2.1 describes the data acquisition and preprocessing process to gather and prepare data for analysis. Section 2.2 explains the initial classification and data analysis procedure by the WEKA tool as well as the algorithms and metrics to be used.

3.1 Data Acquisition and Preprocessing:

The research utilizes data from the academic records of first-year nursing students at Tobruk University. SQL queries will be used to extract relevant information from the university database, including student demographics, course grades, class standing, and attendance records. The data extraction procedure is explained below.

Figure 1 shows the data extraction process performed in the study. The procedure is explained below:

1. Connect to the database: The researchers will access TUGS-CON Ver. 2 of the College and browse its database.
2. Identify the tables: Database tables containing the relevant data for the research will be specified. For this study, student information (stud_info), course grades (stud_records), and courses (subject_tb) table were selected.
3. Define the query: An SQL query to extract the desired data from the identified tables will be created. SELECT, JOIN, WHERE, GROUP BY, In, and other clauses to filter, combine, and aggregate data according to research needs will be applied.
4. Extract the data: The query will be executed and the result exported as a dataset in a spreadsheet. Results will then be formatted to CSV in preparation for feeding into WEKA. Before feeding, extracted data that are still unclean will undergo a preprocessing procedure.

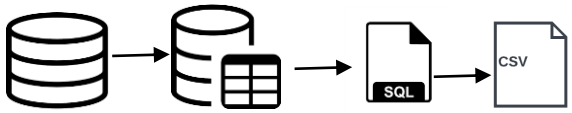


Figure (1) Data Extraction Process

The extracted data undergoes subsequent cleaning and transformation stages. This will be the preprocessing procedures:

1. Imputation and deletion will be used to deal with missing values identified (Panda & Adhikari, 2020).
2. Identification of outliers will be considered for possible influence on the analysis (Dash et al. 2023).
3. Data validation and verification will be performed to rectify data inconsistencies and errors (Ridzuan & Zainon, 2019).
4. The process of feature engineering can be used to develop new features using the existing data (OuahiMariame, 2021).

3.2 Initial Classification and Data Analysis:

Preprocessed data will then be imported into the WEKA software for further analysis and exploration. The descriptive statistics will be computed to determine the distribution of student performance and identify possible influences. Exploratory data analysis techniques will allow us to visualize the relationships between variables and identify patterns.

Selected classification algorithms available in WEKA will be utilized for the initial classifications of experiments. These are BayesNet, NaiveBayes, JRIP, and J48. These algorithms will classify and then try to predict the performance of students based on the features extracted. The created JRIP rule and J48 pruned tree will be presented to identify the main contributor to students' academic performance.

Metrics such as accuracy, precision, recall, and the F1 score will be applied to evaluate every classifier's performance. Kapa stat will be used to gauge the significance of the classification result. Error measurement will include mean absolute error and root mean square error. These metrics will give information on how well each algorithm predicts performance based on the available data.

4 Data Extraction Result

After carefully looking at the tables, we found that the data collected by the system was not as complete as expected. Table stud_info recorded name, control no., gender, and current year level as the only useful information for the study. Table subject_tb has course code, subject, description (if major or minor course), and units (Lecture/theory units, Laboratory Units, and Clinical Units). Table stud_records where students' academic performance was recorded, class standing (Cs, summation of class lecture performance like attendance, quiz, and term exam), Lab/exam (for courses with laboratory), and Finals (final exam). The documented performance however was recorded in summary instead of by category, thus the study can only use Cs as a whole, Lab/exam, and Finals as performance variables. Additional variables include 2nd (reset exam result), and carrier (loading exam result) for students who fail the course after the final exam. A snapshot of the datasheet view of the student's record extracted from the database is shown in Figure 2.

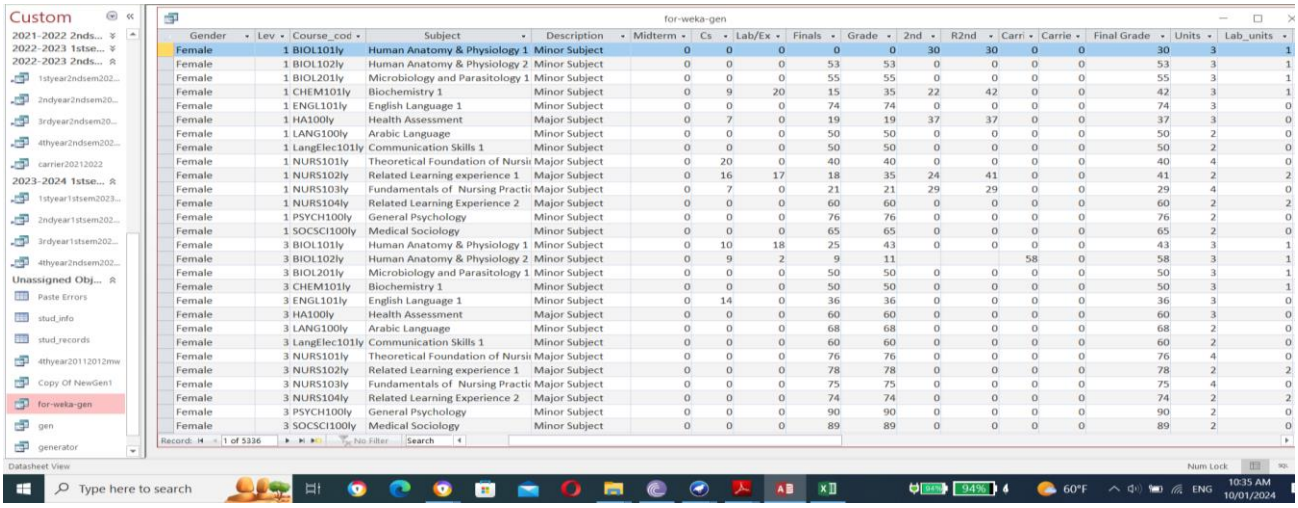


Figure (2) Datasheet view of students' record

The recorded academic performance were noisy and incomplete, a usual suspect for hindering knowledge discovery (Sessa & Syed, 2016). The computation for final grades at the college has different treatments depending on the subject/course description and between theoretical classes and classes with laboratory. Courses described as major (nursing major subject), lecture (theory class), lab (practical or laboratory), or clinical units have a passing rate of 60. All minor (general subject) courses have a passing mark of 50. The grading system and percentage equivalent are shown in Table 1.

Table (1) Grading system and percentage equivalent

Course	Description	CS	Lab Exam	Finals	Clinical	Passing Rate
Lecture (Theory)	Minor	30	0	70	0	50
Lecture (Theory)	Major	30	0	70	0	60
With Lab	Minor	20	20	60	0	50
With Lab	Major	20	40	40	0	60
Clinical	Major	20	0	30	50	60

It is also noteworthy that these were recorded in their summary form instead of raw equivalent. Instead of recording for example Cs=80, Lab=90, and Finals=50 for minor courses and then transmuted, the system shows it recorded instead Cs=16, Lab=18, and Finals=30 with a Grade of 64 for a passing mark. Field Midterm was not used for recording in recent years, instead long quiz that was incorporated with Cs was used. Either task from Cs or midterm was scrapped in the recent school year due to shortened classes and closure from the pandemic and other factors. Figure 3 shows the current Grade computation used in the College of Nursing using SQL's IIf statement.

```
Grade: IIf([lab_units] Or [clinical_units]>0, [Midterm]+ [cs]+[Lab]
+[finals], [midterm]+[cs]+[finals])
```

Figure (3) Grade computation query

Another distinguishing feature of the College's grading system is the recording of a 2nd assessment (reset exam) to replace the result of query Grade. Depending on the subject, lecture courses 2nd assessment results will replace 100% of the student's Final Grade, while courses with lab and clinical units retain their mark. 2nd assessment results will then replace the sum of Finals and Cs. Furthermore, a carrier exam (loading exam) was also given to students who were promoted to the next year's level if the student have only two failing courses after 2nd assessment result. Figure 4 shows the final grade computation with 2nd assessment and carrier (loading exam). There was some school year when even courses with lab were completely replaced by 2nd assessment exam. IIf statements were used to handle carrier, 2nd, and SY in different eras, thus the computation below.

```
Final Grade: IIf([carrier2]<>0,[carrier2], IIf([carrier]<>0,[carrier],
IIf([2nd]=-1,-1,IIf([2nd]=0,[grade], IIf([lab_units]>0 And
[stud_records.SY] >="20162017" Or [stud_records.SY]
<"20132014", [2nd]+[lab],[2nd])))))
```

Figure (4) Final grade computation with 2nd assessment and Carrier Exam (Loading exam)

To capture the core of the 2nd assessment, the researchers created a new query (figure 5) instead of relying on the recorded 2nd (which is only the reset exam result) and computed Final Grade. Computation was recorded in R2nd.

```
R2nd: IIf([2nd]=-1,-1, IIf([2nd]=0,0, IIf([lab_units]>0 And
[stud_records.SY] >="20162017" Or [stud_records.SY]
<"20132014",[2nd]+[lab],[2nd]))
```

Figure (5) Special computation used for courses 2nd Assessment (Reset Exam)

The study will be using the record from the last three (3) school year, 2020-20221, 2021-2022 and 2022-2023. As shown in Figure 6, WHERE clause with an In statement was used to extract the said AY's.

```
WHERE (((stud_records.SY) In
("20202021","20212022","20222023")) AND ((stud_records.sem)
In ("1st","2nd")) AND ((stud_records.code) In ("y1s1","y1s2")))
```

Figure (6) Where and In clause used to extract records from previous three A.Y.

The system produces a final grade status of passing and failing remarks depending on the subject description. For the researcher to create a nominal value better than Status (pass or fail), a Rating query was created to depict a descriptive rating equivalent. Table 2 shows the Final grade equivalent rating and Figure 7 its query.

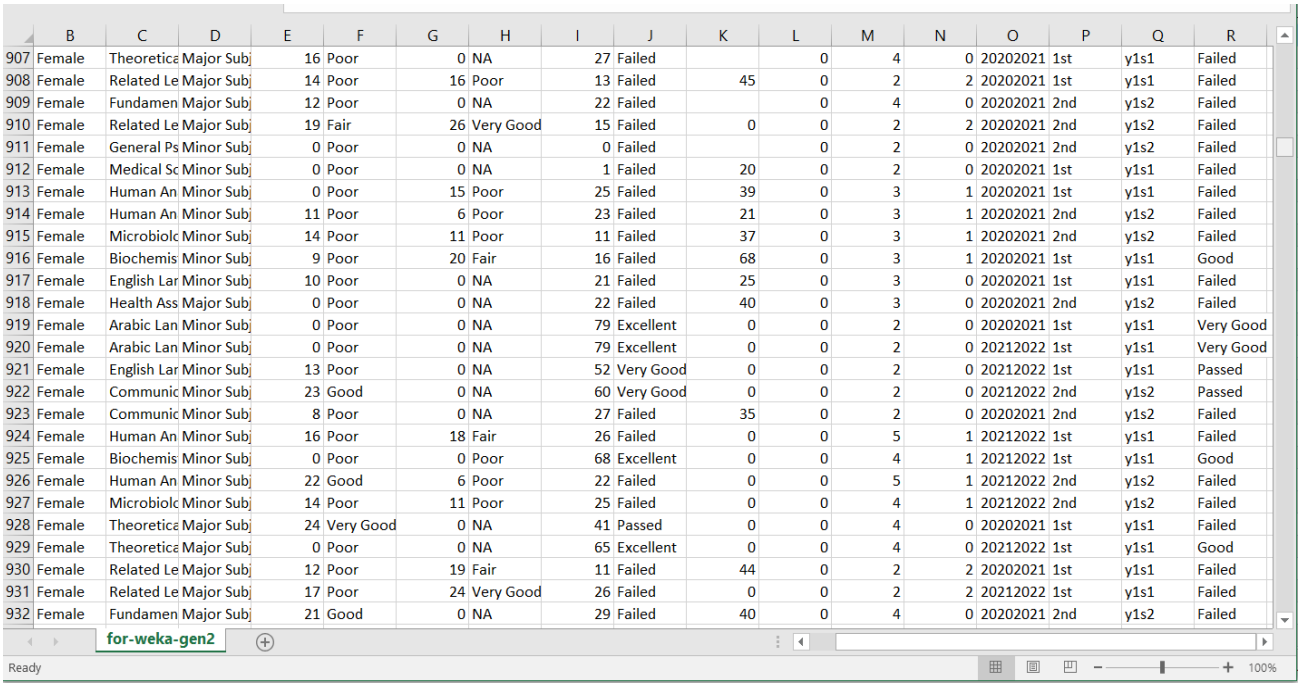
Table (2) Grading system and percentage equivalent

Final Grade	Course Description	Descriptive Rating equivalent
85 to 100	Both	Excellent
75 to <85	Both	Very Good
65 to <75	Both	Good
60 to <65	Major	Passed or Fair
50 to <65	Minor	Passed or Fair
Below 60	Major	Failed or Poor
Below 50	Minor	Failed or Poor

```
Rating: IIf([Final Grade]>=85,"Excellent", IIf([final grade]>=75,
"Very Good", IIf([final grade]>=65,"Good", IIf(([Final grade]>=50
And ([description]="Minor Subject") Or [final grade]>=50 And
[description]="") Or ([final grade]>=60 And [description]="Major
Subject"),"Passed","Failed"))))
```

Figure (7) Final grade rating

Using the SQL code, the data were extracted and exported to a spreadsheet file. The preprocessing stage then follows. Imputation and deletion were used on missing values. Possible outliers were determined. Validation and verification were performed to rectify data inconsistencies and errors. Feature engineering was utilized to reclassify attributes. A total of 5336 rows of records were gathered in a dataset after preprocessing. There are 316 unique students, 280 were female and 36 were male. A total of 14 courses were also retrieved. The file was then formatted to a CSV file in preparation for data analysis. The next section describes the data analysis result including the classification procedure and algorithms used for this experiment. A sample of dataset extracted from the College's database is shown in Figure 8.



	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
907	Female	Theoretica	Major Subj	16	Poor	0	NA	27	Failed		0	4	0	20202021	1st	y1s1	Failed
908	Female	Related Le	Major Subj	14	Poor	16	Poor	13	Failed	45	0	2	2	20202021	1st	y1s1	Failed
909	Female	Fundamen	Major Subj	12	Poor	0	NA	22	Failed		0	4	0	20202021	2nd	y1s2	Failed
910	Female	Related Le	Major Subj	19	Fair	26	Very Good	15	Failed	0	0	2	2	20202021	2nd	y1s2	Failed
911	Female	General Ps	Minor Subj	0	Poor	0	NA	0	Failed		0	2	0	20202021	2nd	y1s2	Failed
912	Female	Medical Sc	Minor Subj	0	Poor	0	NA	1	Failed	20	0	2	0	20202021	1st	y1s1	Failed
913	Female	Human An	Minor Subj	0	Poor	15	Poor	25	Failed	39	0	3	1	20202021	1st	y1s1	Failed
914	Female	Human An	Minor Subj	11	Poor	6	Poor	23	Failed	21	0	3	1	20202021	2nd	y1s2	Failed
915	Female	Microbiolc	Minor Subj	14	Poor	11	Poor	11	Failed	37	0	3	1	20202021	2nd	y1s2	Failed
916	Female	Biochemis	Minor Subj	9	Poor	20	Fair	16	Failed	68	0	3	1	20202021	1st	y1s1	Good
917	Female	English Lar	Minor Subj	10	Poor	0	NA	21	Failed	25	0	3	0	20202021	1st	y1s1	Failed
918	Female	Health Ass	Major Subj	0	Poor	0	NA	22	Failed	40	0	3	0	20202021	2nd	y1s2	Failed
919	Female	Arabic Lan	Minor Subj	0	Poor	0	NA	79	Excellent	0	0	2	0	20202021	1st	y1s1	Very Good
920	Female	Arabic Lan	Minor Subj	0	Poor	0	NA	79	Excellent	0	0	2	0	20212022	1st	y1s1	Very Good
921	Female	English Lar	Minor Subj	13	Poor	0	NA	52	Very Good	0	0	2	0	20212022	1st	y1s1	Passed
922	Female	Communic	Minor Subj	23	Good	0	NA	60	Very Good	0	0	2	0	20212022	2nd	y1s2	Passed
923	Female	Communic	Minor Subj	8	Poor	0	NA	27	Failed	35	0	2	0	20202021	2nd	y1s2	Failed
924	Female	Human An	Minor Subj	16	Poor	18	Fair	26	Failed	0	0	5	1	20212022	1st	y1s1	Failed
925	Female	Biochemis	Minor Subj	0	Poor	0	Poor	68	Excellent	0	0	4	1	20212022	1st	y1s1	Good
926	Female	Human An	Minor Subj	22	Good	6	Poor	22	Failed	0	0	5	1	20212022	2nd	y1s2	Failed
927	Female	Microbiolc	Minor Subj	14	Poor	11	Poor	25	Failed	0	0	4	1	20212022	2nd	y1s2	Failed
928	Female	Theoretica	Major Subj	24	Very Good	0	NA	41	Passed	0	0	4	0	20202021	1st	y1s1	Failed
929	Female	Theoretica	Major Subj	0	Poor	0	NA	65	Excellent	0	0	4	0	20212022	1st	y1s1	Good
930	Female	Related Le	Major Subj	12	Poor	19	Fair	11	Failed	44	0	2	2	20202021	1st	y1s1	Failed
931	Female	Related Le	Major Subj	17	Poor	24	Very Good	26	Failed	0	0	2	2	20212022	1st	y1s1	Failed
932	Female	Fundamen	Major Subj	21	Good	0	NA	29	Failed	40	0	4	0	20202021	2nd	y1s2	Failed

Figure (8) Sample dataset extracted from the College's database

5 Data Analysis Result

The CSV file created in the data extraction process was then loaded to WEKA ver. 3.8.6 for data classification and analysis using its exploration application. Among the different attributes used, continuous data produced a higher accuracy result compared to attributes with nominal data. The result to be presented in this study will be the regular academic performance of students where attributes include course Description, Cs, Lab, and Finals with Rating as nominal classifier. Table 3 show the comparison of the four classifiers wherein J48 got the highest accuracy (88.6619), best in Kappa,

lowest mean absolute error, and root mean square error, while having 2nd to the highest precision, highest recall, and F1 measurement. On the other hand, NaiveBayes has the lowest accuracy (68.9843) relative to the other algorithms used. Accuracy results do not mean that it is the best tool for the model in data mining. However,

when coupled with other metrics, the result clearly shows J48's supremacy among the four tools. In the similar study of Kawade et al., J48 also displayed the highest accuracy when compared to other tools (Kawade et al. 2020). The model created by BayesNet, JRIP, and J48 recorded a Kappa stat between >=61 to <=80 proving their result is of substantial significance.

Figure 9 shows the JRIP rules created by the experiment, showing Finals as the dominant rule among the students' academic performance. This means that the better the performance in Finals, the nearer its descriptive rating will be classified. Simply put, most students with excellent marks in Finals will have a high probability rating of excellent regardless of their descriptive rating in other performance metrics. JRIP rule also shows that Lab/Exam (lab) and course Description (major or minor) may also contribute to the descriptive rating. Similar to the result of Ahmed & Kabir's experiment, JRIP rules show that the better the result in final major or minor classes, the higher the chance of passing rate expected (Ahmed & Kabir, 2022). A total of nine rules were generated by the model in the current set of attributes. The created JRIP rule model also has a higher chance of classifying higher-rating students than lower-rating ones correctly.

One of the most effective methods for data mining and knowledge discovery is the presentation of decision trees (Bhargava et al., 2013). In the generated J48 pruned tree visualizer (figure 10) Finals appeared as the root (top node) adhering to the fact that the better the result in this attribute the nearer it would be to its descriptive result equivalent or Rating. The tree also created several internal nodes of the said attribute that represent test conditions applied. It shows that marks >49 in Finals have a bigger chance of being classified to its equivalent descriptive rating. Lab/exam also appears to influence marks <=49, however to a lesser extent compared to Finals based on the tree. Both JRIP and J48 classify that attribute Finals will most likely determine the Final grade equivalent descriptive Rating. The nominal attribute Rating is the created leaf node of the tree.

Table (3) Comparison of different classifiers using regular academic performance

Algorithm	Accuracy	Kappa Stat	*MAE	**RMSE	Precision	Recall	F1
BayesNet	83.9393	0.6852	0.0819	0.2357	0.844	0.839	0.833
NaiveBayes	68.9843	0.4272	0.137	0.2957	0.716	0.690	0.682
JRIP	87.8748	0.7581	0.0844	0.2054	0.898	0.879	0.873
J48	88.6619	0.7779	0.0751	0.1931	0.897	0.887	0.882

*Mean Abs Error, **Root mean square error.

```
JRIP rules:
===========

(Finals >= 85) => Rating=Excellent (89.0/0.0)
(Finals >= 75) => Rating=Very Good (124.0/0.0)
(Finals >= 65) => Rating=Good (277.0/1.0)
(Lab/Exam >= 14) and (Finals >= 50) => Rating=Good (17.0/4.0)
(Finals >= 50) and (Description = Minor Subject) => Rating=Passed (686.0/0.0)
(Finals >= 60) => Rating=Passed (161.0/0.0)
(Lab/Exam >= 5) and (Finals >= 31) and (Le-R = Fair) => Rating=Passed (47.0/2.0)
(Description = Minor Subject) and (Finals >= 37) and (Lab/Exam >= 13) => Rating=Passed (29.0/0.0)
 => Rating=Failed (3906.0/640.0)

Number of Rules : 9


Time taken to build model: 0.41 seconds
```

Figure (9) JRIP rules created with regular academic performance

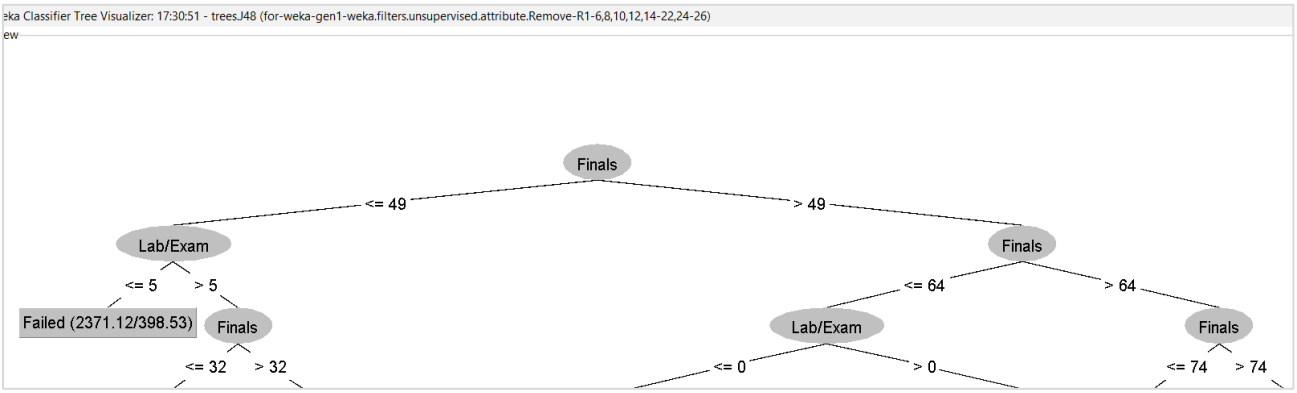


Figure (10) J48 Tree visualizer with regular academic performance (top three levels)

The researchers also used the attribute selector, where WEKA chooses description, Finals, Fe-R (Finals exam descriptive rating), R2nd, and carrier as the best attributes to go along with the Rating. As shown in Table 4, J48 improved and retained the highest accuracy (97.8261), precision, recall, and F1 measure while still having the lowest error stat. The rest of the algorithms vastly improved on their calculation although NaiveBayes remained with the lowest accuracy (79.3853) compared to the rest of the algorithms used. BayesNet, JRIP, and J48 all recorded a KAPA score of >=80, proving the higher reliability of the model, where its significance is interpreted as near perfect. Once again, J48 shows supremacy against the other algorithm using the current set of attributes. The algorithm with a stronger classification and lower error rate was always preferable (Kawade et al., 2020).

Figure 11 shows the JRIP rules created by the experiment, where Finals remained the dominant rule among the students' academic performance. It is the most important attribute in the experiment based on the result. The addition of R2nd also plays a significant part, as it is a result replacer, it mattered considerably in the descriptive Rating result. JRIP rule also shows that Carrier and course Description (major or minor) may contribute to the descriptive rating. The model is created with a total of 14 rules. There are more rules created for this model since there are six attributes selected by the attribute selector compared to the original five attributes selected by the proponents. The higher accuracy and precision of this second JRIP rule model also attributes to its faster creation (0.35 sec) compared to the first set of experiments (0.41 sec).

Likewise in the J48 tree visualizer (figure 12) generated by the WEKA tool, Finals remained as the root of the tree, thus the better the result in this attribute the closer it gets to its descriptive equivalent. R2nd also displays a considerable classification, especially for poorer-performing students. With more attributes compared to the first set of classifications, the J48 pruned tree created more nodes. Metrics like the number of nodes, number of leaves, depth of the tree, and number of attributes used in tree construction define the complexity of a tree (Bhargava et al., 2013).

Table (4) Comparison of different classifiers using an attribute selector

Algorithm	Accuracy	Kappa Stat	*MAE	*RMSE	Precision	Recall	F1
BayesNet	91.9228	0.8489	0.0453	0.1455	0.813	0.919	0.918
NaiveBayes	79.3853	0.5984	0.1069	0.2654	0.799	0.794	0.772
JRIP	97.8073	0.9596	0.0304	0.1024	0.978	0.978	0.978
J48	97.8261	0.9599	0.0171	0.0923	0.979	0.978	0.978

*Mean Abs Error, **Root mean square error

```
JRIP rules:
===========

(Finals >= 85) => Rating=Excellent (89.0/0.0)
(R2nd >= 85) => Rating=Excellent (14.0/0.0)
(Finals >= 75) => Rating=Very Good (124.0/0.0)
(R2nd >= 75) => Rating=Very Good (50.0/0.0)
(Carrier >= 75) => Rating=Very Good (5.0/1.0)
(Finals >= 65) => Rating=Good (277.0/1.0)
(R2nd >= 65) => Rating=Good (85.0/0.0)
(Carrier >= 65) => Rating=Good (6.0/0.0)
(Finals >= 50) and (Description = Minor Subject) => Rating=Passed (703.0/17.0)
(R2nd >= 50) and (Description = Minor Subject) => Rating=Passed (282.0/0.0)
(Finals >= 60) => Rating=Passed (161.0/0.0)
(R2nd >= 60) => Rating=Passed (131.0/0.0)
(Carrier >= 50) => Rating=Passed (45.0/1.0)
 => Rating=Failed (3364.0/97.0)

Number of Rules : 14


Time taken to build model: 0.35 seconds
```

Figure (11) JRIP rules created using an attribute selector

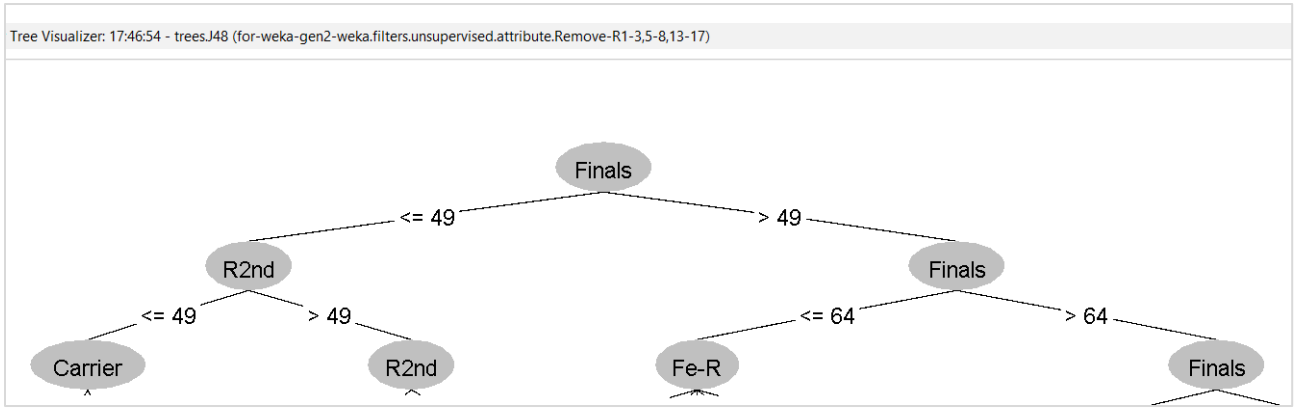


Figure. (12) J48 Tree visualizer using an attribute selector (top three levels)

The preliminary data analysis and initial classification attempts reveal some interesting insights:

- Data distribution: The distribution of different features depends on the data visualization techniques. This implies emerging patterns and relationships to be pursued further.
- Potential influencing factors: The initial analysis shows factors like final exams are predictive of student performance in most courses. There might be other factors contributing to this, but more analysis should be done to validate these relationships and determine other variables affecting them.
- Initial classification results: The analysis of several classification algorithms reveals positive findings. The majority of algorithms obtained accuracy rates over 80% except for NaiveBayes. Nevertheless, these are interim effects, and more fine-tuning as well as cross-validation is needed to evaluate model generalizability and validity.

Due to limited non-academic attributes, the classification algorithm tends to choose academic performance attributes as predictive measures. Despite successfully acquiring and cleaning data in data extraction results, and displaying high accuracy results in data analysis results, further research with additional attributes must be gathered to form a better model. The next section is the study's limitations and recommendations for future work.

6 Limitations and Future Work

This research is currently in its preliminary stages, and some limitations need to be considered:

- Limited data scope: The current analysis is limited to first-year data. Incorporating data from subsequent years may help to paint a broader picture of student performance trajectories.
- Feature selection and engineering: Additional analysis to determine the most pertinent features for prediction and possibly generate new features that can optimize performance is necessary.
- Class imbalance: The distribution of performance grades can be unbalanced and hence there is a need for appropriate treatment to manage this bias in data.

Future research may involve addressing these limitations and further refining the analysis:

- Feature selection and engineering will be performed on the data used for classification.
- Other classification algorithms will be investigated and compared for better performance.
- To verify the generalizability and reliability of developed models, cross-validation methods will be applied.
- The analysis will include data from the subsequent years to track student performance trajectories and long-term academic outcomes.
- Further refinement of classification and predictive models can be achieved by gathering more non-academic performance. Attributes like economic standing, behavior, and other possible factors may affect classification and prediction.

7 Conclusion

Application of the data mining methods promises a great deal in identifying the variables that affect performance scores by students enrolled in nursing studies. As depicted by the algorithms used, Finals attribute is the most important academic performance. However, other attributes may play a significant role in influencing students' Final grades and descriptive ratings. Non-academic factors may also contribute to the result if utilized. The created model may be used as a classification training set for future test sets, although modification and update of attributes are preferable. This study attempts to offer some insights into this area of research by looking at the peculiarities of Tobruk University and tracing patterns that can be used as a starting point for educational interventions, leading students toward success. The preliminary results presented here provide a basis for further analysis and model building, which could lead to a more thorough understanding of academic performance in nursing education at Tobruk University.

Conflict of Interest: The authors declare that there are no conflicts of interest.

References

Aher, S. B., & Lobo, L. M. R. J. (2011, March). Data mining in the educational system using weka. In *International conference on emerging technology trends (ICETT)* (Vol. 3, pp. 20-25).

Ahmed, Md & Kabir, Md. (2022). Analysis of University Students' Performance Using WEKA to Enhance the Education Quality. BAUET Journal. 03. 1-18.

Almarabeh, H. (2017). Analysis of students' performance by using different data mining classifiers. *International Journal of Modern Education and Computer Science*, *9*(8), 9.

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, *17*, 1-21.

Barakeh, A. M., Mezher, M. A., & Alharbi, B. A. (2024). Literature Review for Educational Data Mining Systems—Fahad Bin Sultan University Case Study. In *Artificial Intelligence-Augmented Digital Twins: Transforming Industrial Operations for Innovation and Sustainability* (pp. 435-453). Cham: Springer Nature Switzerland.

Baranyi, M., Gál, K., Molontay, R., & Szabó, M. (2019, November). Modeling students' academic performance using Bayesian networks. In *2019 17th international conference on emerging eLearning technologies and applications (ICETA)* (pp. 42-49). IEEE.

Bhargava, N., Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. *Proceedings of international journal of advanced research in computer science and software engineering*, *3*(6).

Bressane, A., Zwirn, D., Essiptchouk, A., Saraiva, A. C. V., de Campos Carvalho, F. L., Formiga, J. K. S., & Negri, R. G. (2023). Understanding the role of study strategies and learning disabilities on student academic performance to enhance educational approaches: A proposal using artificial intelligence. *Computers and Education: Artificial Intelligence*, 100196.

Cui, Y., Chen, F., Shiri, A., & Fan, Y. (2019). Predictive analytic models of student success in higher education: A review of methodology. *Information and Learning Sciences*, *120*(3/4), 208-227.

Dash, C. S. K., Behera, A. K., Dehuri, S., & Ghosh, A. (2023). An outliers detection and elimination framework in classification task of data mining. *Decision Analytics Journal*, *6*, 100164.

Espinosa, R., Zubcoff, J., & Mazón, J. N. (2011). A set of experiments to consider data quality criteria in classification techniques for data mining. In *Computational Science and Its Applications-ICCSA 2011: International Conference, Santander, Spain, June 20-23, 2011. Proceedings, Part II 11* (pp. 680-694). Springer Berlin Heidelberg.

Feng, G., & Fan, M. (2024). Research on learning behavior patterns from the perspective of educational data mining: Evaluation, prediction and visualization. *Expert Systems with Applications*, *237*, 121555.

García, S., Luengo, J., & Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems*, *98*, 1-29.

Goundar, S., Deb, A., Lal, G., & Naseem, M. (2022). Using online student interactions to predict performance in a first-year computing science course. *Technology, Pedagogy and Education*, *31*(4), 451-469.

Gowri, G. S., Thulasiram, R., & Baburao, M. A. (2017, November). Educational data mining application for estimating students performance in weka environment. In *IOP Conference Series: Materials Science and Engineering* (Vol. 263, No. 3, p. 032002). IOP Publishing.

Han, J., Kamber, M., & Mining, D. (2006). Concepts and techniques. *Morgan kaufmann*, *340*, 94104-3205.

Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, *9*(2), 447-459.

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies*, *13*(1), 61-72.

Kawade, D. R., Oza, K. S., & Naik, P. G. (2020). Student performance classification: A data mining approach. *JIMS8I-International Journal of Information Communication and Computing Technology*, *8*(2), 462-466.

Keshavarzi, M. H., Azandehi, S. K., Koohestani, H. R., Baradaran, H. R., Hayat, A. A., & Ghorbani, A. A. (2022). Exploration the role of a clinical supervisor to improve the professional skills of medical students: a content analysis study. *BMC Medical Education*, *22*(1), 399.

Kumar, V. P., & Krishnaiah, R. V. (2012). Horizontal aggregations in SQL to prepare data sets for data mining analysis. *IOSR Journal of Computer Engineering (IOSRJCE)*, 2278-0661.

Mendoza J., Buhat-Mendoza D., Tan C. (2017). Tobruk University Grading System for College of Nursing Version 2 in Tobruk, Libya. *International Journal for Database Management System (IJDMS)*, Vol. 6, No. 5.

Mishra, T., Kumar, D., & Gupta, S. (2014, February). Mining students' data for prediction performance. In *2014 Fourth International Conference on Advanced Computing & Communication Technologies* (pp. 255-262). IEEE.

Mori, M., Noughi, N., & Cleve, A. (2015). Extracting data manipulation processes from SQL execution traces. In *Information Systems Engineering in Complex Environments: CAiSE Forum 2014, Thessaloniki, Greece, June 16-20, 2014, Selected Extended Papers 26* (pp. 85-101). Springer International Publishing.

Nahar, K., Shova, B. I., Ria, T., Rashid, H. B., & Islam, A. S. (2021). Mining educational data to predict students performance: A comparative study of data mining techniques. *Education and Information Technologies*, *26*(5), 6051-6067.

Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, *11*(1), 237.

OuahiMariame, S. K. (2021). Feature engineering, mining for predicting student success based on interaction with the virtual learning environment using artificial neural network. *Annals of the Romanian Society for Cell Biology*, *25*(6), 12734-12746.

Ordonez, C., Maabout, S., Matusevich, D. S., & Cabrera, W. (2014). Extending ER models to capture database transformations to build data sets for data mining. *Data & Knowledge Engineering*, 89, 38-54.

Panda, B. S., & Adhikari, R. K. (2020, March). A method for classification of missing values using data mining techniques. In *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)* (pp. 1-5). IEEE.

Pujianto, U., Azizah, E. N., & Damayanti, A. S. (2017, October). Naive Bayes using to predict students' academic performance at faculty of literature. In *2017 5th International Conference on Electrical, Electronics and Information Engineering (ICEEIE)* (pp. 163-169). IEEE.

Ridzuan, F., & Zainon, W. M. N. W. (2019). A review on data cleansing methods for big data. *Procedia Computer Science*, *161*, 731-738.

Roostaee, M., & Meidanshahi, R. (2023). Hidden Pattern Discovery on Clinical Data: an Approach based on Data Mining Techniques. *Journal of AI and Data Mining*, *11*(3), 343-355.

Schwab, K., Moseley, B., & Dustin, D. (2018). Grading grades as a measure of student learning. *SCHOLE: A Journal of Leisure Studies and Recreation Education*, *33*(2), 87-95.

Sessa, J., & Syed, D. (2016, December). Techniques to deal with missing data. In *2016 5th international conference on electronic devices, systems and applications (ICEDSA)* (pp. 1-4). IEEE.

Villarica, Mia. (2020). Mining Student Academic Performance on ITE subjects using Descriptive Model Approach. *Res. J. Computer and IT Sci* 4. 1-15.

Walia, N., Kumar, M., Nayar, N., & Mehta, G. (2020, April). Student's academic performance prediction in academic using data mining techniques. In Proceedings of the International Conference on Innovative Computing & Communications (ICICC).

Zughoul, O., Momani, F., Almasri, O. H., Zaidan, A. A., Zaidan, B. B., Alsalem, M. A., & Hashim, M. (2018). Comprehensive insights into the criteria of student performance in various educational domains. *IEEE access*, *6*, 73245-73264.