STUDENT PERFORMANCE PREDICTION USING RANDOMDECISION (RD) CLASSIFICATION ALGORITHM

Jayasree R[1]
Research Scholar, Department of Computer Science
Sri Krishna Arts and Science College, Coimbatore, Tamilnadu, India.

Sheela Selvakumari N.A[2]
Associate Professor, Department of Computer Science,
Sri Krishna Arts and Science College, Coimbatore, Tamilnadu, India.

Abstract — Any educational institution that wants to enhance its students' learning experience and academic success must be able to predict their academic performance. Due to the various factors that influence students' learning processes and academic success, even though the problem of predicting their performance has received several studies educational institutions continue to face difficulties with it. The most important phase in developing a prediction model for student performance is feature selection and machine learning classification, which can increase prediction accuracy and make it easier to identify variables that have a big influence on student performance. This paper presents an improved Wrapper Feature with Rank Selection (IWFRS) with new RandomDecision (RD) Classification model that is proposed for the prediction of student performance dataset. The results have demonstrated that the unique strategy performs better than existing Classification of Decision Tree and Random Forest classification algorithms. As the Experimental results show that proposed RandomDecision (RD) classification with (IWFRS) feature selection methods clearly outperform than existing methods.

I. INTRODUCTION

In recent years, there has been a lot of interest in using data mining techniques in the educational setting. Data mining (DM) is the process of finding data. The application of conventional DM techniques to address issues pertaining to education is known as educational data mining (EDM) [1-2]. EDM refers to the use of DM techniques to educational data, including information about the student, their academic history, their exam scores, their involvement in class, and the frequency of their questioning. EDM is currently a helpful tool for identifying hidden patterns in educational data, predicting academic success, and improving the learning and teaching environment.

Data mining methodologies have multiplied greatly. Predictive (like classification) and descriptive (like association analysis) are the two primary categories according to their goals [3]. In contrast to predictive approaches, which use the data to make predictions, descriptive techniques extract characteristics from the data. The goal of the paper is to identify intriguing characteristics that affect student success. Descriptive techniques may uncover intriguing trends in student characteristics, but they are unable to classify individuals according to a certain result. The predictive approaches, on the other hand, classify pupils into groups based on behaviors they share, but their major goal is to create models that are more accurate at predicting future outcomes. Through the use of subgroup discovery [4] techniques, the research proposes to extract the student traits that have the greatest influence.

One of the most crucial concerns in the realm of EDM is performance prediction for students. In order to intervene and provide guidance in advance, it is important to anticipate student performance in order to recognize the possibility of academic failure in kids as early in the learning process as possible. Additionally, it can serve as a basis for individualized learning recommendations and help decision-making by educational administrators by examining the variables influencing student performance [5, 6]. Machine learning classification problems in the realm of EDM include predicting student success. Using labeled historical academic data from students, researchers build a prediction model based on a classification algorithm for supervised learning. According on student demographics, prior academic achievement, and other characteristics, the trained prediction model produces the class of students' performance.

The goal of feature selection is to choose a subset of characteristics from the input that may accurately characterize the input data while minimizing noise or unimportant features [7]. Many feature selection techniques have been presented over the course of the last few decades; these techniques can be categorized into 3 groups: filtering techniques, wrapper techniques, and embedded techniques [7]. The most widely used filtering techniques are ReliefF [8-10] techniques.

Data mining techniques like classification, clustering, and association rules, to mention a few, can all be used to analyze educational data. These methods will aid in obtaining hidden information and useful knowledge. One of the supervised learning strategies is classification, which develops a model to categories a data point into a specified class label. By using the facts at hand, classification attempts to forecast future results. Therefore, classification is one of the strategies better suited for educational analysis since educational institutions are attempting to forecast the future output of their enrolled students based on their existing prior and current student's data.

By collecting student data from secondary education at two Portuguese institutions, this paper estimates student accomplishments at the end of the semester. The purpose of this paper is to forecast student's final grades so that teachers can protect vulnerable kids. To improve the prediction model's accuracy rate, preprocessing of the data is used. The ideal feature subset was identified using an improved wrapper method for feature selection. Next, a new hybrid RandomDecision (RD) classification model that extends the Random forest and Decision tree classification models is introduced. Initializing a variety of tree variables and dimensions is the first step in this categorization. Initializing the classification process with a variety of tree factors and feature dimensions are performed. These classification inputs include the entire pre-processed feature set as the training feature and the final feature selection as the test feature.

II. REVIEW OF LITERATURE

R. Kaviyarasi and T. Balasubramanian, 2018 [11] addressed how an increase in student enrollment has led to the construction of educational facilities at all levels. Teachers today have a wide range of duties. Teachers have a duty to assist students in selecting a career path that fits with their skills and interests. To enhance the quality of educational processes, the data mining area extracts educational data from massive amounts of data. The modern educational system must help students improve their ability to make decisions, solve problems, and develop social skills. Finding hidden patterns and knowledge in educational institutions is one of the uses of data mining called "Educational data mining". Fast learners, average learners, and slow learners are the three significant student groupings that have been found. Students likely suffer with a variety of issues. The writers concentrated on identifying the key elements that have the most potential to influence college students' performance.

In order to identify the variables influencing student success, Helal, et al., 2018 [12] obtained course evaluation along with student demographic and academic data recorded at registration. Their findings have shown both the general efficacy of the subgroup discovery approach in identifying the components as well as the advantages and disadvantages of various widely used subgroup discovery algorithms used in this research. According to the results of the experiments, students who come from poor socioeconomic backgrounds or who were accepted based on specific entry requirements are more likely to fail.

Using supervised machine learning and multiple linear regression, Ouafae El Aissaoui et al. (2020) [13] suggested a methodology to develop a student performance prediction model (MLR). Three main parts make up their methodology: the first step analyses and preprocesses the students' characteristics or variables using a variety of statistical analytic techniques; the second step involves choosing the most crucial variables using various techniques. In the third phase, various MLR models will be built based on the chosen variables, and their performance will be compared using the k-fold cross-validation method.

Data mining techniques will be used in Croatian higher education institutions' educational data in 2020, according to Snjezana Kri zanic [14]. The data used for the study are event logs that were taken from a real setting for an e-learning course. The research used two data mining methods: decision trees and cluster analysis. Using decision trees was the preferred method for developing a representation of decision-making that allowed distinguishing classes of items with the intention of undertaking a more extensive research of how children learn.

Mudasir Ashraf and Yass Salal discussed ensemble approaches in 2021 [15], which combine different learning classifiers based on heuristic machine learning techniques to create prediction paradigms. These learning ensemble methods are frequently more accurate than individual classifiers. In order to predict student performance, researchers have discovered a common learning classifier called bagging among a variety of ensemble techniques.

According to a study by Sarah A. Alwarthan et.al., 2022 [16], understanding the elements that have the greatest influence on a student's performance level will be beneficial to both students and policymakers and will provide in-depth insights into the issue. As a result of their superior performance in numerous earlier investigations, the RF and ensemble models were shown to be the most accurate models. There is a need to address this issue because in earlier studies could not agree on whether or not the entrance requirements have a strong association with students' accomplishment. Additionally, it has been noted that there aren't many studies that use student data from arts and humanities majors to predict academic performance.

In order to improve students' accomplishments, Jayasree R and Sheela Selvakumari N.A, 2022 [17] provided an extensive literature review on predicting student overall success through the application of data mining methodologies. Their study's major objective is to provide an overview of the data mining techniques that have been employed to forecast students' final grades. With the aid of academic record mining techniques, they hope to significantly improve students' academic achievement and success.

III. PROPOSED METHODOLOGY

The proposed research methodology performs a predicting students' overall performance using data preprocessing, improved wrapper feature selection method and new RandomDecision (RD) Classification process is derived. A data preprocessing work is a crucial data mining method that involves transforming the original properties of the data into a logical structure. In order to assess the quality of feature subsets and produce high-quality results, the Improved Wrapper Feature Selection (IWFS) approach finds a relevant subset of features or attributes from a pre-processed dataset using machine learning technique. The predicting students' overall performance process flow diagram is described in figure 1.

A. Data Preprocessing

Data preprocessing is a vital data mining method that involves transforming the original properties of the data into a logical structure. The preprocess method aims to lessen the amount of noise, irrelevant data, and NaN (Not a Number) in the dataset. Preprocessing the input data set for knowledge discovery target techniques typically takes up the majority of the effort allocated to the entire task in data mining.

In this dataset, 33 attributes (Table 1) are used in each record to characterize student achievement in secondary education of two Portuguese schools. The attributes include student grades, demographic, social, and educational factors. The data was collected through school reports and surveys. Two datasets are presented in relation to the performance in the two distinct subjects of Portuguese language (por) and mathematics (mat). First period grade, second period grade, and final grade are the three categories into which the features in the data set are divided. Because duplicate data and minor features frequently cause the classification algorithm to become confused and produce wrong or ineffective results, data preparation is necessary. Furthermore, using all features will lengthen processing time. Preprocessing also helps to put the data into a uniform format and remove duplicated and incomplete data.

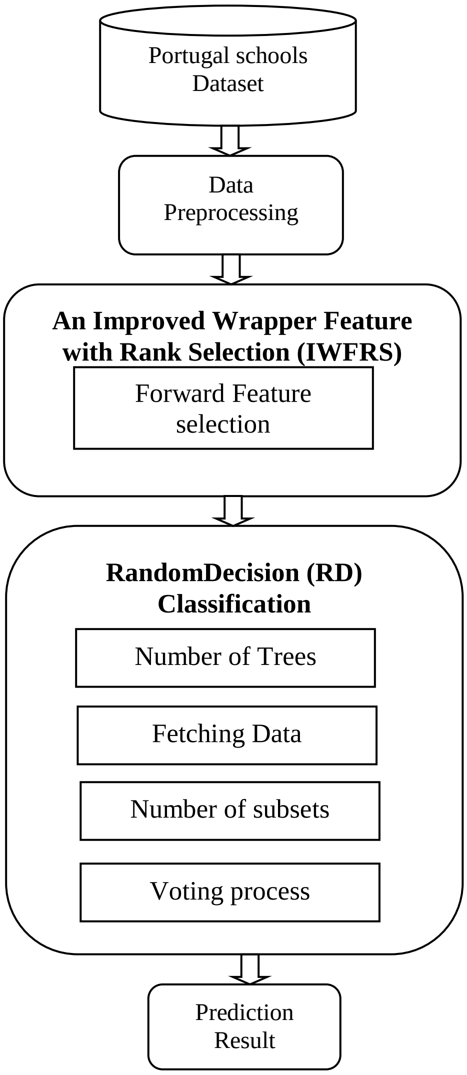

Fig.1: Proposed Workflow Diagram

Datasets on performance in the discipline of mathematics are provided. The goal attributes G3 and the attributes G2 and G1 are highly correlated. This is due to the fact that G3 is the final year grade whereas G1 and G2 are the first and second period grades. Finding that the data in the dataset was clean during the data pre-processing prevented the need for data preprocessing techniques.

Table 1: Lists of Features Available in Student Performance Dataset

S.no	Feature name	Feature Type
1	F_1- School	Binary
2	F_2- Sex	Binary
3	F3 – age	Numeric
4	F4 – address	Binary
5	F5 – family size	Binary
6	F6 – parent’s status	Binary
7	F7 – Mothers Education	Nominal
8	F8 – Fathers Education	Nominal
9	F9 – Mother’s job	Nominal
10	F10 – Father’s job	Nominal
11	F11- Reason	Nominal
12	F12 –Students Guardian	Nominal
13	F13-Student travel time	Numeric
14	F14-Student study time	Numeric
15	F15 – Failures	Numeric
16	F16 – Schoolup	Binary
17	F17 – famsup	Binary
18	F18 – Paid	Binary
19	F19 – Activities	Binary
20	F20 –Nursery	Binary
21	F21 – Higher	Binary
22	F22 – Internet	Binary
23	F23 – Romantic	Binary
24	F24 – famrel	Numeric
25	F25 – Freetime	Numeric
26	F26-goout	Numeric
27	F27 – Dalc	Numeric
28	F28 – Walc	Numeric
29	F29 – Health	Numeric
30	F30 -Absences	Numeric
31	F31 – G1	Numeric
32	F32 – G1	Numeric
33	F33 – G3	Numeric(target)

As in many nations, the final grade in the raw dataset ranges from 0 to 20, with 0 representing the lowest score and 20 the highest. Since the students' final grades are given as integers and the anticipated class is given as categorical values, the data had to be translated into categories in accordance with a grading policy. The grade

F, which is the lowest grade and translates to "fail," is represented by the numbers 0 through 9. The remaining ranges (10-11, 12-13, 14-15, and 16-20) are equivalent to the class labels D (sufficient), C (acceptable), B (good), and A (excellent/very good), F (Fail) respectively.

B. An improved Wrapper Feature with Rank Selection (IWFRS)

Machine learning activities need the use of a hybrid feature selection technique called Improved Wrapper Feature with Rank Selection (IWFRS), which can considerably increase performance by removing redundant and irrelevant features while also expediting the learning task. This approach starts with a null model, then starts fitting the model with each individual feature one at a time, and selects the feature with the lowest p-value. It builds a subset of features from ranking features (FR) using the wrapper method of forward selection search. By attempting combinations of the previously chosen feature with all other surviving features, fit a model with two features at this point. The final Feature Selection (FS) result is obtained by selecting the feature with the lowest p-value once more, selecting the subset with the highest classifier accuracy as the feature set (Featset).

The IWFRS primarily consists of three components, including the input dataset, classification algorithms for assessing feature subsets, and ranking selection method. The performance of the various assessment criteria is fairly similar, and the prediction based on the correlation Pearson coefficient is more established.

$$C_r(feat_k, class) = \frac{Covariance(feat_k, class)}{\sqrt{Variance(feat_k, class)}} \quad eq(1)$$

where C_r(*$feat_k$*, *class*) denotes correlation among feature *$feat_k$* and class, Covariance (*$feat_k$*, *class*) denotes covariance of features and class, and Variance (*$feat_k$*, *class*) indicates variance of features and class.

The gain of every attribute g_k can be achieved by assessing every feature *$feat_k$* in the dataset using Pearson correlation coefficient, and All features can be ranked descendingly based on their gains to acquire *FR*. If there are *k* features in dataset, then FR= $\{feat_1{:}g_1, feat_2{:}\ g_2, \ldots, feat_k{:}\ g_k\}$, while $g_k \geq g_m$ and $1 \leq k < m \leq k$.

Algorithm 1: An Improved Wrapper Feature with Rank Selection (IWFRS)

```
Input: Student performance dataset (Dt).
Output: Selecting Best Feature (BF)
Process
Initialize:  Ac = 0; // Assessment criteria
  C = 0 // Classifier for evaluating feature subset
  Candidateset⟵NULL
  BF⟵0.
REPEAT
    For every feature F in Dt
      FG ← Gain (Dt, Ac) using eqn. (1)
      FL ← SortByDescending (FG)
        FS  ← By focusing on one feature at a time fit every simple regression model possible. Choose the component with the highest FG-value.
```

```
ENDFOR
FOR EACH (m in FS)
  gain ← C(Dt, m)
  IF gain > BF THEN
    Candidateset ← gain
    BF ← gain
  ENDIF
ENDFOR
```

As stated in Algorithm 1, IWFRS gains all features in Dt using the supplied Assessment criterion AC assembles all features in descending order based on their gains to generate FL. To create feature subsets, the wrapper Forward selection is employed; if FL = {$feat_1$, $feat_2$,…, $feat_n$}, then *FS* = {{$feat_1$}, {$feat_2$},…, {$feat_1$, $feat_2$,…, $feat_n$}. The performance of each subset in *FS* is assessed using the designated classifier. In order to arrive at the final feature selection of Best feature, IWFRS lastly searches in the candidate set of features using the given forward search method (BF).

C. RandomDecision (RD) Classification

An extension of the Random Forest and Decision Tree Classification models, the new RandomDecision (RD) classification method is presented in this paper. Both supervised learning and a geometrical classification method are represented by the classifier. Assumes an underlying probabilistic model, and by calculating the likelihoods of the possible outcomes, it enables us to incorporate multinomials in the multiclass case model in a meaningful manner. It can address diagnostic and predictive issues.

This RD classification procedure began with a decision tree that was calculated and acted as a representation of the full dataset. After that, when finding the nodes in a tree, entropy is calculated. It simulates choices based on effectiveness, outcomes, and resource costs. The community of decision trees trained using the bagging approach; one of the ensemble methods, then creates the established RD. To increase prediction accuracy and stability, it constructs several decision trees and combines them.

A number of FOLD (First order logical decision) methods are used to classify RD classification cases in order to evaluate the prediction class. With the use of their IWFRS features, the random selection feature classification achieves the class (A, B, C, D, and F) score.

$$Range_{feat} = \sum_{k=1}^{cl} \sum_{j=1}^{IWFRS_f} search\big(Vote_{feat} == prediction_{cl}\big) \ \ eq(2)$$

Where *feat* is features; *cl* is number of class (*A*,*B*,*C*,*D and F*,); In the training dataset, search is a distinct feature.. After that feature ranging process, RandomDecision process is achieve the classification accuracy is using 3.

$$Classification_{acc} = \sum_{p=1}^{len(search)} Choice\big(searchIndex_p\big) \ \ eq(3)$$

IV. Results and Discussion

The results have been estimated using the proposed IWFRS with RD classification technique. On a PC running Windows 10 with Python 3.8 simulations, the findings were implemented using an Intel I5-6500U series processor running at 3.21 GHz and 8GB of main memory.

To assess how well the suggested Classification works in comparison to existing Decision Tree [18] and Random Forest [18] classification, The student performance dataset could be used in an experiment to evaluate the proposed RandomDecision (RD) classification with IWFRS feature selection approach, which has the flexibility to be customized to forecast the accuracy of data sets to satisfy a variety of test criteria. The classification accuracy rate for the datasets from the Portuguese and math lesson is shown in Tables 2 and 3.

Table 2: Rates of classification accuracy for the dataset from the Portuguese lesson dataset

Methods	Decision Tree	Random Forest	Proposed RandomDecison (RD)
Five-level grading	67.80	73.50	**92.28**
Binary Grading	91.37	93.07	**98.63**

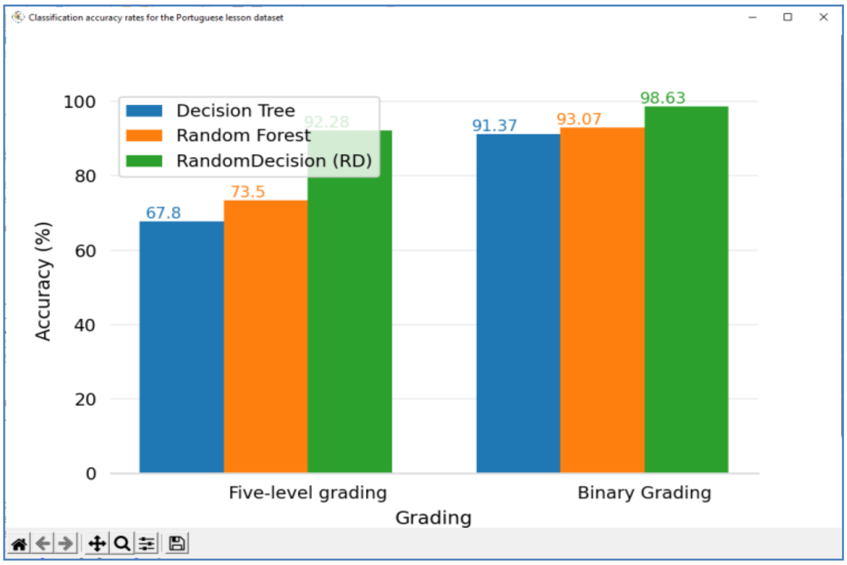

Fig.2: Rates of classification accuracy for the dataset from the Portuguese lesson.

Table 3: Rates of classification accuracy for the dataset from the mathematics lesson dataset

Methods	Decision Tree	Random Forest	Proposed RandomDecison (RD)
Five-level grading	73.42	71.14	**86.55**
Binary Grading	89.11	91.39	**95.18**

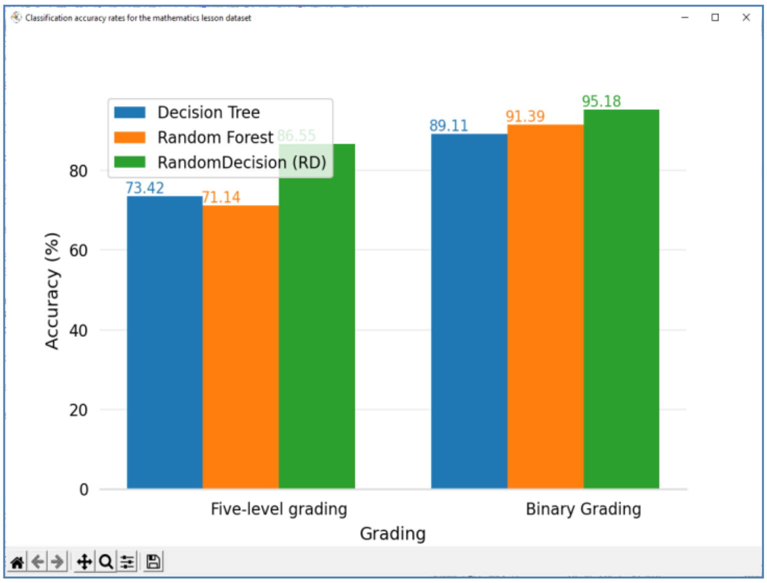

Fig.3: Rates of classification accuracy for the dataset from the mathematics lesson dataset.

According to Figure 2 and 3, the RandomDecision (RD) classification method produced the best results for the five-level grading version for this dataset, with an accuracy rate of Portuguese lesson dataset is 92.28% and mathematics lesson dataset is 86.55%. However, the binary grading version of this dataset improved the accuracy rate. The accuracy rate was raised to Portuguese lesson dataset is 98.63% and mathematics lesson dataset is 95.18% in the dataset where the final grade is classified in binary form (pass or fail).

The ratio of the amount of accurate prediction samples to the total samples is used to determine the classifier's accuracy and assess the performance of the feature subset. To employ tenfold cross validation [19] when employing the classifier for verification accuracy is calculated as the mean of the accuracy of ten rounds. To evaluate the performance of proposed IWFRS feature selection correlation accuracy with existing RnkHEU method [20] and Propsoed IWFRS method described in table 4.

$$Featacc = \frac{\text{amount of accurate prediction samples}}{\text{total samples}} \times 100\% \quad eq\ (4)$$

Table 4: Result and performance Feature Correlation Accuracy rate of Portuguese lesson and mathematics lesson dataset

Datasets	RnkHEU	Proposed IWFRS
Portuguese lesson	59.75	**63.20**
mathematics lesson	68.88	**72.08**

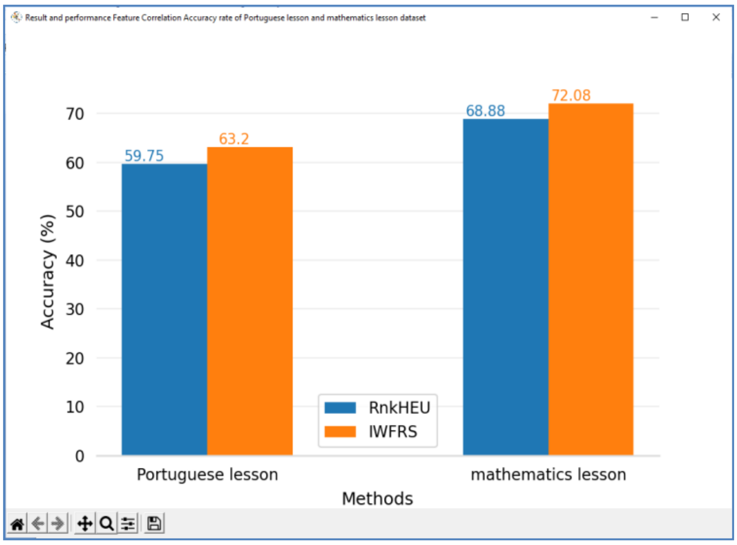

Fig.4: Feature Correlation Accuracy rate of Portuguese lesson and mathematics lesson datasets.

V. CONCLUSION

The final grades of students are predicted in the current study using a new RandomDecision (RD) Classification model that is proposed for the prediction of student performance dataset. To enhance the classification performance, a feature selection technique called Improved Wrapper Feature with Rank Selection (IWFRS) was developed. The percentage of accuracy in the classification method increased as a result of preprocessing operations on the dataset, which included grouping the final grade field into five and two categories. Finally predicts the RandomDecision (RD) classification for predicting accuracy. Overall, the binary class strategy improved accuracy rates for both the Portuguese dataset and mathematics dataset.

VI. REFERENCES

[1] Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3–17.

[2] Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining : Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research, 94(February 2018), 335–343.

[3] Han, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc, Burlington (2005)

[4] Klösgen, W.: Explora: a multipattern and multistrategy discovery assistant. In: Advances in Knowledge Discovery and DataMining, pp. 249–271 (1996)

[5] N. A. Yassein, R. G. M. Helali, and S. B. Mohomad, "Predicting student academic performance in KSA using data mining techniques," Journal of Information Technology & Software Engineering, vol. 07, no. 05, 2017.

[6] S. Helal, J. Li, L. Liu, E. Ebrahimie, S. Dawson, D. Murray and Q. Long, "Predicting academic performance by considering student heterogeneity", Knowledge-Based Systems, vol. 161, pp. 134-146, 2018.

[7] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," Computers & Electrical Engineering, vol. 40, no. 1, pp. 16–28, 2014.

[8] N. Spolar, E. A. Cherman, M. C. Monard, and H. D. Lee, "ReliefF for multi-label feature selection," in Proceedings of the 2013 Brazilian Conference on Intelligent Systems, IEEE, Fortaleza, Brazil, October 2013.

[9] B. K. Yousafzai, M. Hayat, and S. Afzal, “Application of machine learning and data mining in predicting the performance of intermediate and secondary education level student,” Education and Information Technologies, vol. 25, no. 3, 2020.

[10] L. E. Raileanu and K. Stoffel, “(eoretical comparison between the gini index and information gain criteria,” Annals of Mathematics and Artificial Intelligence, vol. 41, no. 1, pp. 77–93, 2002.

[11] Kaviyarasi, R., & Balasubramanian, T. (2018). Exploring the high potential factors that affects students’ academic performance. International Journal of Education and Management Engineering, 8(6), 15.

[12] Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., Murray, D.J. (2018). Identifying key factors of student academic performance by subgroup discovery. International Journal of Data Science and Analytics, 7(3), 227–245.

[13] Ouafae El Aissaoui, Yasser E, Alami El Madani, Lahcen Oughdir, Ahmed Dakkak , and Youssouf El Allioui, "A Multiple Linear Regression-Based Approach to Predict Student Performance", Advanced Intelligent systems for Sustainable Development (AI2SD’2019) (pp.9-23), 2020.

[14] Snjezana Kri zanic, "Educational data mining using cluster analysis and decision tree technique", International Journal of Engineering Business Management, vol 12, 1-9, 2020.

[15] Mudasir Ashraf and Yass Salal, "Educational Data Mining Using Base (Individual) and Ensemble Learning Approaches to Predict the Performance of Students", Data Science, Transactions on Computer Systems and Networks, March 2022.

[16] Sarah A. Alwarthan, Nida Aslam and Irfan Ullah Khan, "Predicting Student Academic Performance at Higher Education Using Data Mining: A Systematic Review", Hindawi, Applied Computational Intelligence and Soft Computing, 2022.

[17] Jayasree R and Sheela Selvakumari N.A, "Recent Exploration on Student Performance Analysis using Educational Data Mining Methods", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-11 Issue-9, August 2022

[18] Ferda Ünal, "Data Mining for Student Performance Prediction in Education", In book: Data Mining - Methods, Applications and Systems , 2020

[19] K. Battula, “Research OF machine learning algorithms using K-fold cross validation,” International Journal of Engineering and Advanced Technology, vol. 8, no. 6S, pp. 215–218, 2021.

[20] Wen Xiao, Ping Ji, and Juan Hu, "RnkHEU: A Hybrid Feature Selection Method for Predicting Students’ Performance", Hindawi, Scientific Programming, Volume 2021.