Predicting school students' academic performance using data mining classification algorithm

Rajinder Singh, Dr. Rajinder Singh Sodhi
Research Scholar, Department of Computer Science and Engineering,
OM Sterling Global University Hisar
Associate Professor & HOD, School of Engineering & Technology,
OM Sterling Global University Hisar
–rajindercse191@osgu.ac.in

Abstract

As educational data continues to grow, the field of educational data mining has gained popularity. This field focuses on extracting hidden patterns from educational data, allowing for a better understanding of students, including their learning styles, and the ability to predict their academic performance. In order to forecast undergraduate students' academic success, this study proposes a model that incorporates data mining techniques. The researchers collected data through questionnaires that included information on demographics, prior GPA, and family history. The data was then analyzed using data mining models, like Decision Tree and Random Forest, and other methods in order to develop the most accurate prediction model. These findings highlight the significant factors that influence students' academic success.

I. Introduction

Data mining, also known as the analysis of significant information from specific data, assists in uncovering concealed patterns and identifying links between parameters in extensive data. Nowadays, numerous researchers utilize data mining to tackle real-world problems in various domains such as marketing, communications, healthcare, medicine, industry, and customer relations. The bioinformatics field heavily exploits data mining and machine learning techniques. Recently, data mining has found extensive application in the realm of education. The academic success of students has become crucial in school education and also in higher learning institutions as it forms a major component of a high-quality educational institution's performance history. Therefore, predicting students' academic achievement has become a critical concern. By accurately predicting their academic performance, early warnings can be provided to students who are at risk. Moreover, analyzing the instructor's performance based on these predictions can also be beneficial. Educational data mining can be employed to explore educational data and uncover hidden patterns in order to utilize machine learning techniques for predicting students' academic achievement.

Data mining employs various methods to examine and process data, including clustering or classification, association rules, and sequence analysis. Each item in a dataset must be classified, thus a classification procedure is utilized.

To ensure reliable prediction of the target class for each case in the dataset, a classification algorithm is applied. In this study, we utilized the Decision Tree algorithm, which is a commonly used prediction method. This approach is favored by many researchers due to its simplicity and the ease with which it can be translated into a set of IF-THEN rules. Numerous previous studies have focused on predicting students' academic performance and learning behavior. Enrollment statistics also include socio-demographic factors such as gender, age, employment position, class, level of education, and disability, as well as study environment factors like course program and course block. Among these factors, ethnicity, course program, and course block are found to be the most crucial in prediction. Thus, we perceive an opportunity to undertake a study that predicts students' academic achievement based on the literature. However, most previous studies have not utilized Random Forest for data classification. Therefore, this study aims to compare different data mining methods for predicting students' academic achievement. Additionally, this research aims to identify the factors that affect pupils' academic achievement.

EDM

Education is a crucial factor in the advancement and development of a nation. It empowers the citizens of a cultured and courteous society. Educational data mining is a burgeoning field that focuses on creating techniques to explore the distinct forms of information derived from an educational database. The process of mining within the educational domain is referred to as educational data mining, and it aims to devise innovative approaches to uncover insights from educational databases (Gallet, 2007) (Erdogan and Timur 2005), in order to examine students' attitudes and behaviors towards their education (Alaa Al-Halis, 2009). The absence of profound and adequate knowledge within the school education system may impede the management from attaining quality objectives, and the methodology of data mining can assist in bridging these knowledge gaps within the school education system.

II. DATA MINING DEFINITION & TECHNIQUES

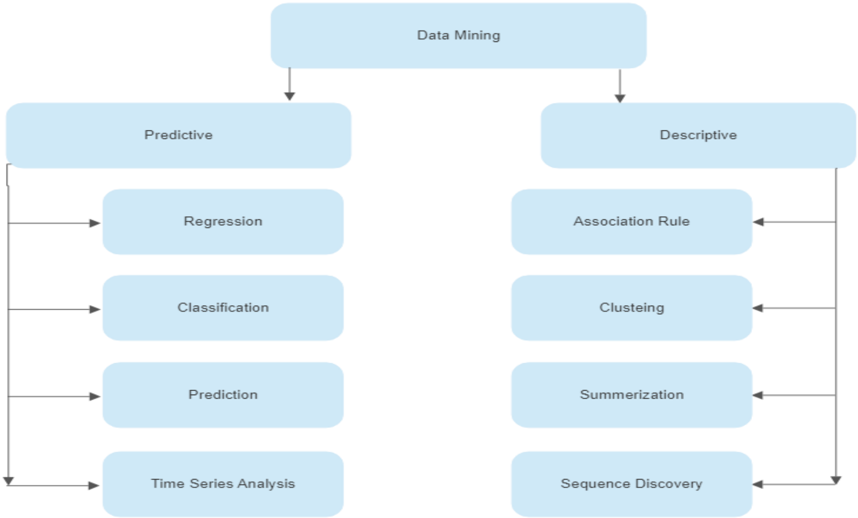

Figure:-1 Data Mining Techniques

Data Classification

Classification is the most commonly utilized technique in data mining, employing a set of pre-classified attributes to generate a model capable of categorizing the majority of information. This approach typically employs classification algorithms based on neural networks or decision trees. Both learning and classification play a role in the data classification process. During learning, a classification algorithm analyzes training data, while data from classification tests are used to evaluate the accuracy of the rules. If the accuracy is deemed acceptable, the rules can be applied to new data tuples. The classifier-training method utilizes these pre-classified attributes to determine the necessary parameters for accurate discrimination. Subsequently, these parameters are incorporated into a classifier model.

Logistic Regression

Logistic regression can be used to model the probability of a specific event or class. It is employed to model a binary dependent variable, providing the likelihood of a single trial. Logistic regression is specifically designed for categorization and allows for an understanding of how independent variables impact a single outcome variable.

Linear Regression

Linear regression is used for regression analysis and is based on supervised learning. It utilizes independent variables to model the predicted value. It is primarily employed to determine the relationship between variables and forecasting.

Decision Trees

Decision trees are the most reliable classification method in data mining. They are represented as flowcharts with a tree-like structure. Each internal node in the tree represents a conditional test, and each branch represents the outcome of the test (true or false). The leaf nodes of the decision tree contain class labels.

Random Forest

The random forest classifier fits multiple decision trees on different sub-samples of the dataset. It utilizes averaging to enhance forecasting accuracy and control overfitting. The sub-sample size is always the same as the input sample size, even though the samples are drawn with replacement.

Naive Bayes

The Naive Bayes method assumes that each feature operates independently of the others and equally influences the outcome. Naive Bayes only requires a small amount of training data to estimate the necessary parameters. Compared to more complex and advanced classifiers, a Naive Bayes classifier is also much faster.

III. RELATED WORK

Due to the potential advantages for educational institutions, data mining in school education is a relatively recent area of research.

According to Rosemarie M. Bautista, Menchita Dumlao, Melvin A. Ballera (2016), Educational Data Mining (EDM) can be utilized to extract patterns that are valuable in analyzing student academic records. The primary aim of this study is to offer recommendations for specialization to engineering students through the use of a data mining algorithm. The attributes that may be significant in making predictions were identified by selecting characteristics based on correlation. The comparative analysis of different algorithms demonstrates that precision was given the highest consideration. The author discovered that a decision tree classification model using WEKA and J48 yielded a precision value of 80.06. The study found that factors such as gender, algebra, calculus, and physics courses greatly influence the prediction of engineering specialization.

In her research, Vandna Dahiya (2018) presents an overview of the various aspects of educational data mining and its objectives. Educational data mining has a profound impact on multiple sectors of the education industry and has the potential for visualizing information, predicting student performance, categorizing students, profiling, planning, and scheduling. The author emphasizes that educational data from diverse sources are constantly changing. Additionally, storing and managing such vast amounts of data becomes challenging. The author also addresses the issue of organizing and comprehending this dynamic educational data. Furthermore, the author outlines several data mining tools such as WEKA, KEEL, R (Revolution), KNIME, and ORANGE.

Surjeet Kumar Yadav, Brijesh Bharadwaj, Saurabh Pal (2012) demonstrate in their paper that a key objective of school and higher education is to provide high-quality education to college students. They propose that one approach to achieve quality in the education system is by using knowledge discovery to predict college student enrollment in courses. This paper presents a data mining project aimed at creating predictive models for student retention management. By analyzing data of new college students, these predictive models can generate accurate lists of students who are likely to require assistance from retention programs. The paper evaluates the effectiveness of machine learning algorithms in generating these predictive models. The results indicate that certain machine learning algorithms are capable of developing robust predictive models based on existing student retention data.

Amirah Mohamed Shahiri, Wahidah Husain, Nur'aini Abdul Rashid (2015) says that in today's era predicting students' performance becomes more challenging due to the huge volume of data in educational databases. In Malaysia, there is no system to analyze and monitor student progress and performance. The two main reasons behind this are as follows. First, the study on existing prediction methods is still not up to the mark and insufficient to identify the most suitable technique and methods for predicting the students' performance. The second is due to the lack of investigations on the various factors affecting students' achievements in particular courses. Therefore, in this paper author proposed a systematic literature review on predicting student performance by using data mining techniques to improve students' achievements. The main objective of this paper is to provide an overview of the various data mining techniques that have been used to analyze and predict students performance. The author said that by using educational data mining techniques we could actually improve students' achievement and success more effectively in an efficient way. It could bring benefits and overall impacts to students, educators, administrators, and academic institutions.

Raza Hasan, Sellappan Palaniappan, Salman Mahmood, Ali Abbas, Kamal Uddin Sarker, and Mian Usman Sattar (2020) Author says that in today's era technology and innovation empower higher educational institutions (HEI) to use various types of learning systems such as video learning system. By analyzing the footprints left behind from these online interactions with students is useful for understanding the effectiveness of video learning systems. This system will help the student to improve their academic performance. In this study, 772 examples of students registered in e-commerce and e-commerce technologies were used. The main aim of this study is to predict student's overall performance semester-end by using video learning analytics and various data mining techniques. Data from various sources like student information systems, mobile applications, and learning management systems were analyzed using eight different classification algorithms. Also, data transformation, preprocessing techniques, genetic search, and principal component analysis were carried out to reduce the features. In end, the CN2 Rule Inducer and multivariate projection can be used to assist faculty in interpreting the rules to gain insights into student interactions. The results of various experiments showed that Random Forest accurately predicted successful students with an accuracy of 88.3% with an equal width and information gain ratio.

Oswaldo Moscoso-Zea, Mayra Vizcaino, Sergio Luján-Mora (2017) The author asserts that Educational Data Mining (EDM) is a developing field that enables the extraction of knowledge from diverse academic contexts through the implementation of data mining techniques on information stored in data repositories of educational institutions. By utilizing various data mining methods and algorithms, institutions can gain a deeper understanding of teaching strategies, student learning patterns, and organizational activities, leading to enhanced decision-making processes. The author conducts experiments employing classification techniques and different decision tree algorithms in EDM to analyze two key performance indicators (KPI): student dropout and graduation rate. Additionally, the author compares these methods and algorithms, suggesting the most accurate ones for specific scenarios.

A.S. Arunachalam, T.Velmurugan (2016) the authors highlight the significant impact of educational data mining (EDM) in the academic field. EDM explores, analyzes, and provides insights into students' behavioral patterns to guide them in making informed career choices. This survey focuses on various techniques used for mining educational data to enhance knowledge. The authors also discuss different tools and techniques employed in EDM. Among these options, they recommend the most effective tools and techniques for real-world applications. Ultimately, the authors conclude that most classification algorithms excel at analyzing and describing current trends in EDM as perceived by students and academicians.

Suhirman, Jasni Mohamad Zain, Haruna Chiroma, and Tutut Herawan (2014) the author emphasizes the importance of continuous efforts by school and higher education management to enhance the quality of educational institutions. Regular evaluations of data collected from multiple sources allow for informed decision-making. Higher authorities should devise plans to utilize data more efficiently, develop tools for data collection and analysis, and provide management information to enhance decision-making processes. The vast amount of collected data can be utilized and analyzed to evaluate quality, diagnose issues, and propose alternative solutions. Data mining methods are well-suited for supporting decision-making processes in educational environments, facilitating the generation and presentation of relevant information and knowledge to improve the quality of education processes.

Atta-Ur-Rahman, Kiran Sultan, Nahier Aldhafferi, Abdullah Alqahtani (2018) Author describes that there are various classroom teaching techniques for effective teaching and learning such as teaching on black/whiteboard, projectors, etc. Some of the students feel comfortable with one of these techniques while others may be comfortable with some other technique. The aim of their study is to discover the trend of the comfort level of students with factors like timetable and teaching techniques. The author designed a questionnaire to acquire students' interest in the teaching-learning process that will show what techniques are mostly liked or preferred by different types or groups of students. Based on the feedback, various machine learning algorithms are applied to extract useful information. They use the WEKA tool to analyze the data and the proposed scheme is then compared with other well-known techniques in the literature.

Saja Taha Ahmed Prof. Dr. Rafah shihab Al-hamdani Dr. Muayad Sadik Croock (2018) Author says that in recent times different methods and algorithms have been adopted in e-learning systems to offer more flexible and effective services for students. Also, they say that the recent smart systems consider the prediction strategies for analyzing and expecting the logical results of different categories in e-learning. In this paper, the researcher goes further with the decision-making process for students, presented as a recommendation for each type of classification method. Moreover, the e-learning systems use various classification and clustering methods for classifying the investigated dataset. In this paper, the author presents a comprehensive study of the newest e-learning decision-making and prediction.

Sedigheh Abbasnasab Sardareh, Mohd Rashid Mohd Saad, Abdul Jalil Othman, RosalamChe Me (2014) Author says that due to the persistent growth and increasing availability of educational data, EDM techniques facilitate data-driven decision making for enhancing teaching-learning. In this analytical study, the author provides an introduction to EDM. Also, the researchers will look at various application areas of data mining in the education domain, and major challenges in mining big educational data. This information enables educators to understand how big data helps students and teachers in improving the teaching and learning processes.

Carla Silva, José (2017) Author says that from last year the adoption of new technology named learning management systems in education has been increased. They also say that various data mining techniques like clustering, prediction, and relationship mining can be applied to huge educational data to study the behavior and performance of students. In this paper, they explore that to build up a new environment and to give new predictions different data mining approaches and techniques can be applied to Educational data.

Cristóbal Romero, Sebastián Ventura (2010) says that Educational Data Mining is an emerging interdisciplinary research area that mainly deals with the development of new methods to explore data collected from various educational institutes. In EDM computational approaches are used to analyze education. The author also surveys the most relevant studies carried out in the EDM field to date. In this paper, author define EDM and describes the various groups of user, different types of educational environments and the data they generate. Then author prepare a list of most typical/common tasks in the educational environment that may be resolved through various data mining techniques.

Ms. Falguni Suthar, Ms. Hiralben Patel, Dr. Bhavesh (2019) Author describes Educational data mining as an emerging field that focuses on analyzing huge educational data to develop models that will help to improve learning experiences and institutional effectiveness. To improve the quality of teaching and learning it provides inherent knowledge about the delivery of education. The mining of huge educational data develops new methods to analyze and discover the knowledge of the educational database and is used for decision-making in the education domain. In this paper, the author presents a study on various components of educational data mining along with its objectives. The objective of the author behind this document is to present a brief general description of EDM methods and techniques.

Saeed Aghabozorgi, Hamidreza Mahroeian, Ashish Dutt, Teh Ying Wah, and Tutut Herawan (2014) According to the author, educational institutions now store massive amounts of data, leading to a continuous growth of data in the field of education. This data is stored in structured formats such as relational databases, as well as unstructured formats like Word or PDF files, images, videos, and geospatial data. As a result, the complexity of education is increasing on a daily basis. The velocity of various data types, along with the processing of streaming data, poses challenges for stakeholders such as educators, instructors, students, research developers, tutors, and others who directly work with educational data.

Mohammad Shiralizadeh Dezfoli, Behzad Soleimani Neysiani, Dr. Naser Nematbakhsh (2019) The primary focus of academic institutions is the performance of their students. These institutions aim to identify the factors that affect student performance and provide approaches to improve their academic levels. In this study, the author examines the factors that influence the identification and prediction of different student educational statuses from two perspectives: academic and algorithmic. The author utilizes a dataset consisting of 26,000 records collected from students at Ashrafi Esfahani University in Sepahanshahr, Isfahan, Iran. The dataset includes 27 different attributes and covers all academic levels, including Bsc., MSc., and Ph.D. The author applies three algorithms - decision tree, Naïve Bayes, and deep learning - using the open-source Rapidminer tool after preprocessing the data. The output of the algorithms is then compared. The results indicate that the decision tree algorithm with the IGR approach has the highest validation performance, achieving an accuracy of about 95%, recall of 68.7%, and precision of 78.2% in predicting student statuses.

IV. PROPOSED WORK

Data Collection & Preparations: The data set utilized in this study was obtained from different educational institutions. Initially, the data consisted of 200 entries. The data set comprises eleven variables for analysis, namely: student's gender, age, father's educational qualification, father's occupation, mother's educational qualification, mother's occupation, family income, percentage obtained in 10th class, stream chosen in 12th class, percentage obtained in 12th class, and final grade. To convert the variables into categorical attributes, we discretized the numerical attributes. For example, let's consider variable X (X = x0, x1, x2…..), which represents the passing percentages of students in the 10th, 12th, and other related factors. All grades were categorized into three groups: Excellent, Good, and Average. The table below illustrates this categorization.

Final Percentage	Final Grade
$X \geq 80\%$	Excellent
$X \geq 60\%$ and $X<80$	Good
$X< 60\%$	Average

TABLE-I: VALUES OF FINAL GRADE

We also descretized other attributes, such as the student's present stream, the passing percentages for the 10th, and 12th. Finally, the following table lists the most specific attributes:

Attribute	Description	Possible Values

Student ID	Student's Unique Identification	{alphabets Characters}
Gender	Gender of Student	{Male, Female, Other}
Age	Students Age	{Below 16, 16 to 18, Above 18}
Father Qualification	Qualification of father	{10^{th}, 12^{th}, graduation, post graduation, not educated}
Father Occupation	Occupation of Father	{Agriculture, Business, Govt. Service, Labour, Private Service}
Mother Qualification	Qualification of Mother	{10^{th}, 12^{th}, graduation, post graduation, not educated}
Mother Occupation	Occupation of Mother	{Govt. Service, House Wife, Private Service}
Family Income	Income of family	{Under 2 lac, 2 to 4 lac, more than 4 lac}
High School Percentage (10^{th} %)	Percentage of marks obtained in 10^{th} class exam.	{ Below 60%, 60% to 70%, 70% to 80%, 80% to 90%, Above 90%}
12^{th} class stream	Stream in 12^{th} class	{Arts, Science, Commerce}
Intermediate Percentage (12^{th} %)	Percentage of marks obtained in 12^{th} class exam.	{ Below 60%, 60% to 70%, 70% to 80%, 80% to 90%, Above 90%}
Final_ Grade	Final Grade obtained after analysis the passing percentage of 10^{th} ,12^{th} and other factors	{ Excellent, Good, Average }

Table II: **THE SYMBOLIC ATTRIBUTE DESCRIPTION**

For the sake of this inquiry, the following definitions of some of the variables' domain values were used:

- **Stream** – Student's Course Stream in which they are enrolled in 12^{th} class. Stream split in three classes: Arts, Commerce, Science.
- **High School Percentage (10^{th} %)** -- Student's passing Percentage (%) in 10^{th} class. 10^{th} % is split into three classes: Below 60%, 60% to 70%, 70% to 80%, 80% to 90%, Above 90%.

- **Intermediate Percentage (12th %)** --Student's passing Percentage (%) in 12th class. For admission in undergraduate courses minimum 50% marks are needed in 12th class. So 12th % is split into two classes: Below 60%, 60% to 70%, 70% to 80%, 80% to 90%, Above 90%.
- **Final Grade** –The value of final grade (X) will be finding after analysis of rule sets of Student's passing percentage (%) in 10th (x_0), 12th (x_1), and other factors. The final grade is divided into three categories: Excellent, Good, Average.

V. SPSS TOOL

SPSS, also known as the Statistical Package for the Social Sciences, is a widely utilized software tool for statistical analysis. It is recognized for its user-friendly interface and extensive range of functionalities. Researchers can effectively manipulate and analyze intricate data in an interactive and intuitive manner using this software. According to the SPSS base user's guide, it offers comprehensive features for generating detailed reports, charts, plots, descriptive statistics, and advanced statistical analyses using data from diverse file formats.

The range of functionalities provided by SPSS encompasses a broad array of statistical techniques, including data manipulation, descriptive statistics, contingency tables, correlation analysis, analysis of variance (ANOVA), multivariate analysis of variance (MANOVA), regression analysis, discriminant analysis, and cluster analysis. In our research, we imported the data in .xls format into IBM's SPSS version 16.0 using the built-in import feature and saved it in the .sav data file format. To gain valuable insights from the student database, we conducted descriptive statistics utilizing frequency tables, cross-tables, and graphs.

SPSS employs two primary file formats: the data file (.sav), which contains the actual data, and the output file (.spv), which presents the analysis results, such as tables and graphs. The software interface is based on a graphical user interface (GUI), providing a user-friendly experience and facilitating smooth navigation. SPSS's Integrated Development Environment (IDE) consists of two main windows.

The Data Editor Window serves as the primary workspace in SPSS, allowing users to view and modify their data. It is the initial window that appears upon opening the software. In this window, users can create new datasets, import existing ones, and directly manipulate the data. The Data Editor Window is organized into rows and columns, with each row representing an observation or case, and each column representing a variable. Users have the flexibility to add, delete, rearrange variables and cases, and apply sorting and filtering based on specific criteria.

Moreover, the Data Editor Window in SPSS offers advanced features for recording, transforming, and aggregating variables. Users can generate new variables based on calculations or logical conditions, merge data from multiple sources, and perform various data manipulation operations. It serves as a robust tool for managing and manipulating data within SPSS, playing a vital role in the software's analytical capabilities.

In this study we use SPSS to analyze the frequency distribution of various factors.

1. **Gender**

		Frequency	Percent
Valid	Female	128	64.0
	Male	72	36.0
	Total	200	100.0

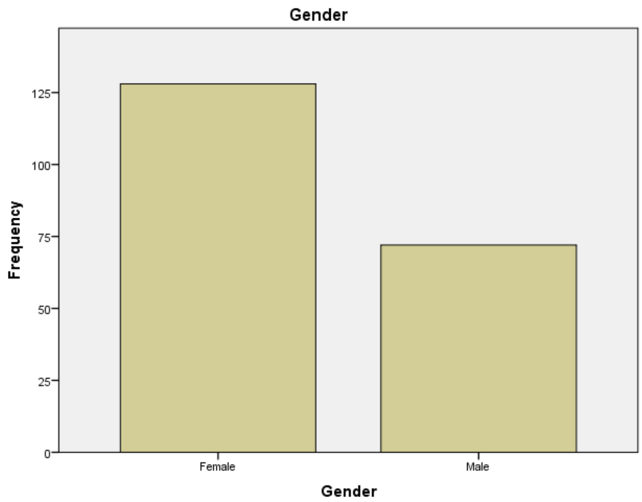

Figure 2

The data presented in the table showcases the distribution of a variable known as "Gender" across two distinct groups: "Female" and "Male". It offers insights into the number of individuals within each group and the corresponding percentage relative to the total sample size. Among the 200 individuals included in the study, 128 individuals (constituting 64% of the total) are categorized as "Female", whereas 72 individuals (making up 36% of the total) are categorized as "Male". The "Total" row denotes the combined frequencies and represents the overall count of individuals in the sample.

2. **Age**

		Frequency	Percent
Valid	16-18	170	85.0
	Above 18	29	14.5
	Below 16	1	.5
	Total	200	100.0

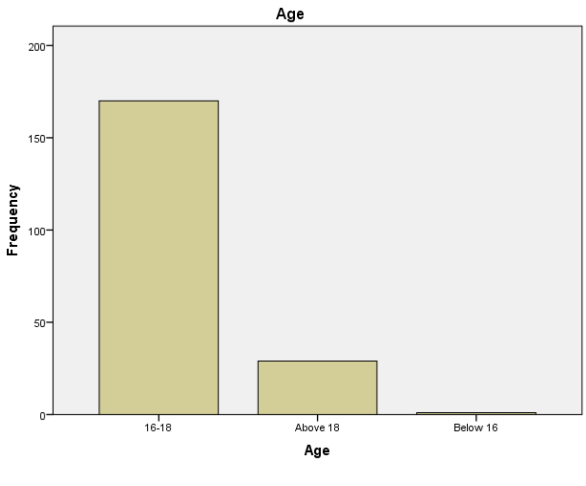

Figure 3

This frequency distribution show that the sample contain maximum 170 records with age 16 to 18 year, whereas only 29 students are there with record above 18 and only one record with age below 16. This data shows that maximum students are in 16 to 18 age block, whereas these are just 1 student which has less than 16 year age.

3. Father Qualification and Occupation CrossTabulation

father_qualification * father_occupation Crosstabulation

		father_occupation					Total
		Agriculture	Business	Govt. Service	Labour	Private Service	
father_qualification	10th	56	6	1	9	1	73
	12th	29	2	4	6	10	51
	Graduation	11	0	11	0	0	22
	Not Educated	30	0	2	5	8	45
	Post-graduation	5	0	4	0	0	9
Total		131	8	22	20	19	200

Figure 4

This information about the education and occupation of student's shows that if education is 10^{th} or 12^{th} the maximum persons are working in agriculture field, whereas Govt. service are contain the maximum graduate people. Person which are in business have qualification 10^{th} or 12^{th} and maximum person which are in not educated category are in agriculture and in labour field.

4. Mother Qualification and Occupation

mother_qualification * mother_occupation Crosstabulation

		mother_occupation			Total
		Govt. Service	House Wife	Private Service	
mother_qualification	10th	1	63	1	65
	12th	1	20	0	21
	Graduation	0	10	0	10
	Not Educated	1	99	1	101
	Post-graduation	3	0	0	3
Total		6	192	2	200

Figure 5

Analysis of this information shows that maximum female belongs to House Wife category. Whether they have qualification like graduate or post graduate they are working as house wife. Only 3 post graduation female are in Govt. Service and only 1 with 10th qualification are in private service.

5. **Family Income**

		Frequency	Percent
Valid	2 to 4 lac	20	10.0
	More than 4 Lac	3	1.5
	Under 2 Lac	177	88.5
	Total	200	100.0

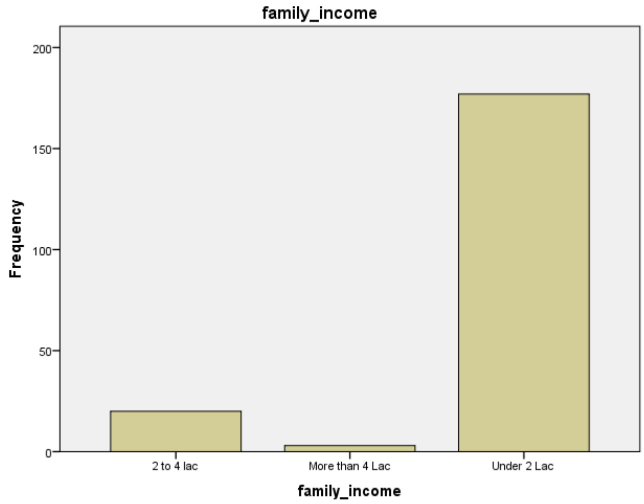

Figure 6

Above frequency Distribution shows that "Under 2 Lac": Out of the total 200 individuals, 177 individuals, or 88.5% of the total, come from families with an income level of under 2 Lac. "2 to 4 lac": There are 20 individuals, accounting for 10.0% of the total, whose families have an income level ranging from 2 to 4 Lac. "More than 4 Lac": Only 3 individuals, representing 1.5% of the total, come from families with an income level exceeding 4 Lac.

6. **10th Class Percentage**

		Frequency	Percent
Valid	60% to 70%	19	9.5
	70% to 80%	59	29.5
	80% to 90%	75	37.5
	Less than 60%	6	3.0
	More than 90%	41	20.5
	Total	200	100.0

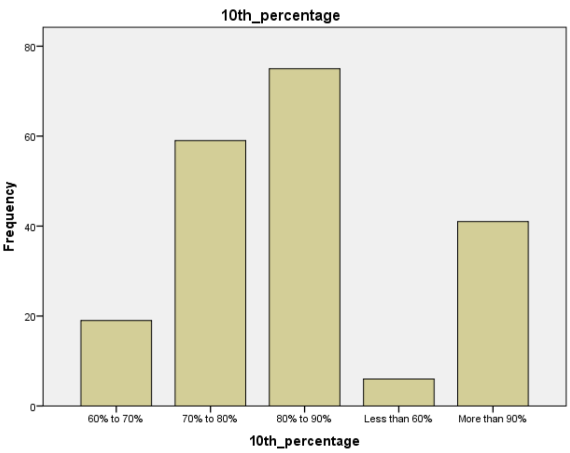

Figure 7

This frequency distribution provides insights into the distribution of individuals based on their percentage scores. It helps in understanding the proportion of individuals falling into different percentage score ranges within the dataset. This shows that maximum student get 80% to 90%, whereas Less than 60% section contain minimum students.

7. 12th class stream and percentage

12th_stream * 12th_percentage Crosstabulation

		12th_percentage					Total
		60% to 70%	70% to 80%	80% to 90%	Above 90%	Below 60%	
12th_stream	Arts	18	39	27	21	24	129
	Commerce	0	7	10	8	6	31
	Science	2	14	8	12	4	40
Total		20	60	45	41	34	200

Figure 8

Analysis of data shows total 129 students are from arts stream, 31 are from commerce stream and 40 are from science stream. Also this data shows that in Arts stream maximum students are in 70% to 80% block. Whereas in Commerce more students present in 80% to 90% and in Science stream max students are in 70% to 80%. This table shows that maximum students of all the streams are in 70% to 80% block where as 60% to 70% contain minimum no. of students.

8. 12th class stream and final grade cross tabulation

12th_stream * Final_Grade Crosstabulation

		Final_Grade			Total
		Average	Excelle nt	Goo d	
12th_stream	Arts	24	48	57	129
	Commerce	6	18	7	31
	Science	4	20	16	40
Total		34	86	80	200

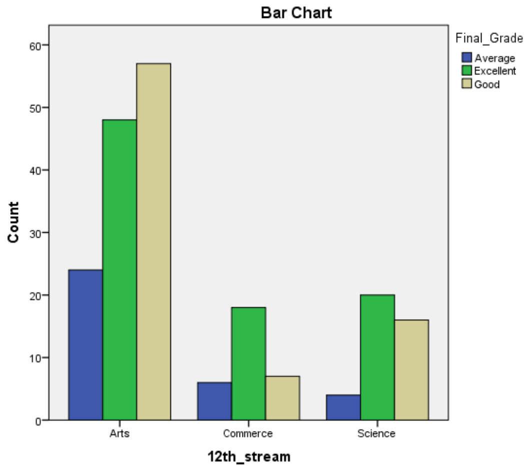

Figure 9

Above data shows that Arts stream have 24 students as average, 48 as excellent and 57 have good final grade. This data clear indicate that students are in all the groups. In commerce stream only 6 students are in average category where 18 students have excellent grade and 7 students have Good grade. Also in science stream 4 students in average, 20 students in excellent and 16 student in Good grade which is maximum for science students.

VI. WEKA TOOL

A variety of visualization tools, algorithms, and graphical user interfaces for quick access to these capabilities are all included in the Weka workbench. It is software that is open source. It can run on practically any current computing platform because it is fully implemented in the Java programming language, making it portable and platform neutral. Data preprocessing, clustering, classification, association, visualization, and feature selection are just a few of the common data mining activities that Weka can perform. The six-button WEKA graphical environment is launched by the WEKA GUI chooser: Simple CLI, Explorer, Experimenter, Knowledge Flow, ARFF-Viewer, & Log.

The Explorer interface features a number of panels that provide access to the workbench's essential elements.

- The Preprocess panel imports data from databases, CSV files, ARFF, etc. and preprocesses it using a filtering method that can change the data's format, such as turning numerical attributes into discrete ones. On the preprocess screen, it is also possible to eliminate instances and attributes in accordance with particular criteria. You may also view the graph for a specific attribute.
- Applying classification and regression techniques (such as the NaiveBays algorithm, ADTree, ID3 Tree, J48 Tree, and ZeroR rules, among others) to the dataset and estimating the model's accuracy are both possible using the Classify panel. Additionally, incorrect predictions, ROC curves, etc., can be seen. In the classifier output area, you can see the classification results.
- To access Weka's clustering methods, such as the simple k-means algorithm, EM, DBScan, and XMeans algorithm, use the Cluster panel. The Omit Attribute button makes it feasible to ignore certain attributes while utilizing the clustering process.
- Access to association rules, such as the Apriori and Predictive Apriori algorithms, is provided by the Associate panel. Once the proper parameter for the association rule has been selected, the result list enables viewing or saving of the result set.

- To find the subset of an attribute that works best for creating predictions, use the Select Attributes panel to search through all possible combinations of attributes in the dataset.
- 2D plots of the present relation are visualized by the Visualize panel.

VII. RESULT AND DISCUSSION

The data set of 200 students used in this study was obtained from the various schools whether they are in rural or urban area. All the schools are in two category one is Co-Educational and other of only for girls. Weka tool is used in this study to analyze the data using various classification method.

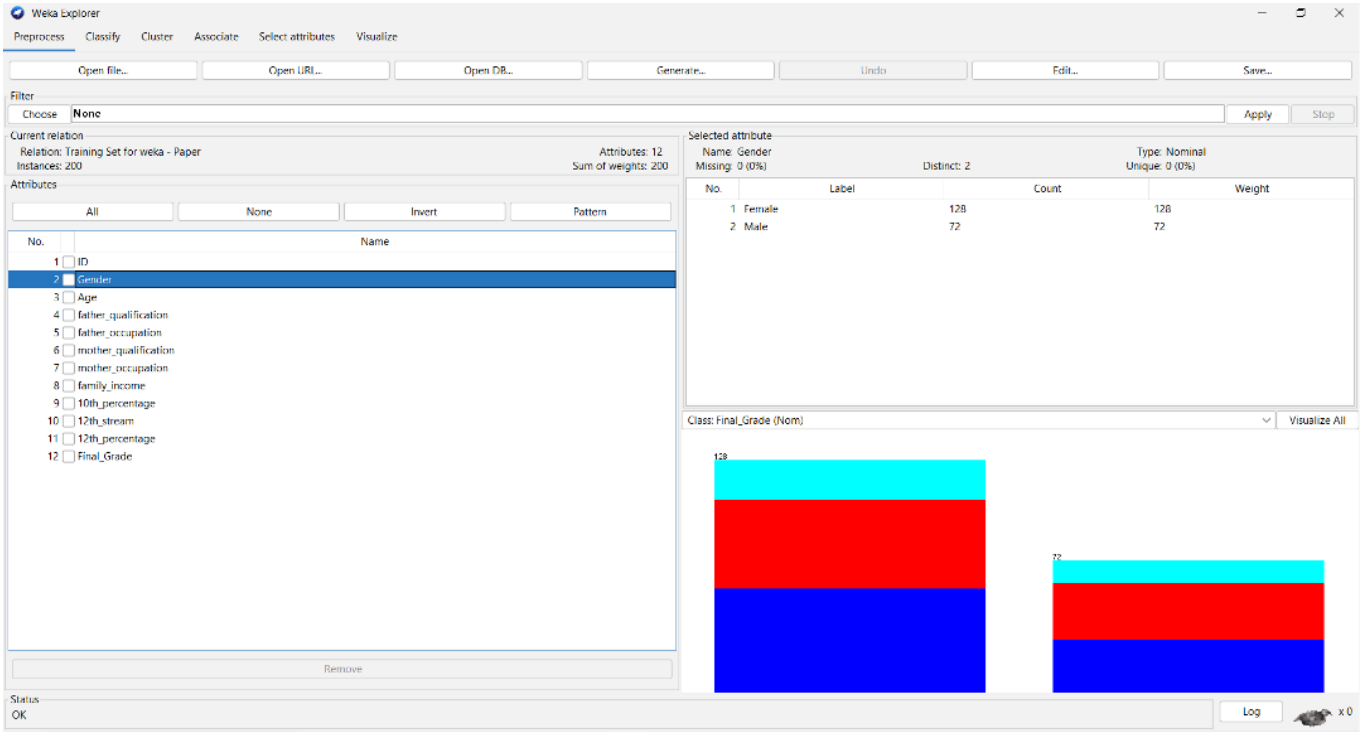

Figure10: Weka 3.9.6 with explorer window open with student's database

This is main window which is used to show all the variables and there information. Left panel in this windows show the variable name where as right panel shows count and weight of selected variable. We can remove unwanted variable from the left panel if that variable is of no use in out study.

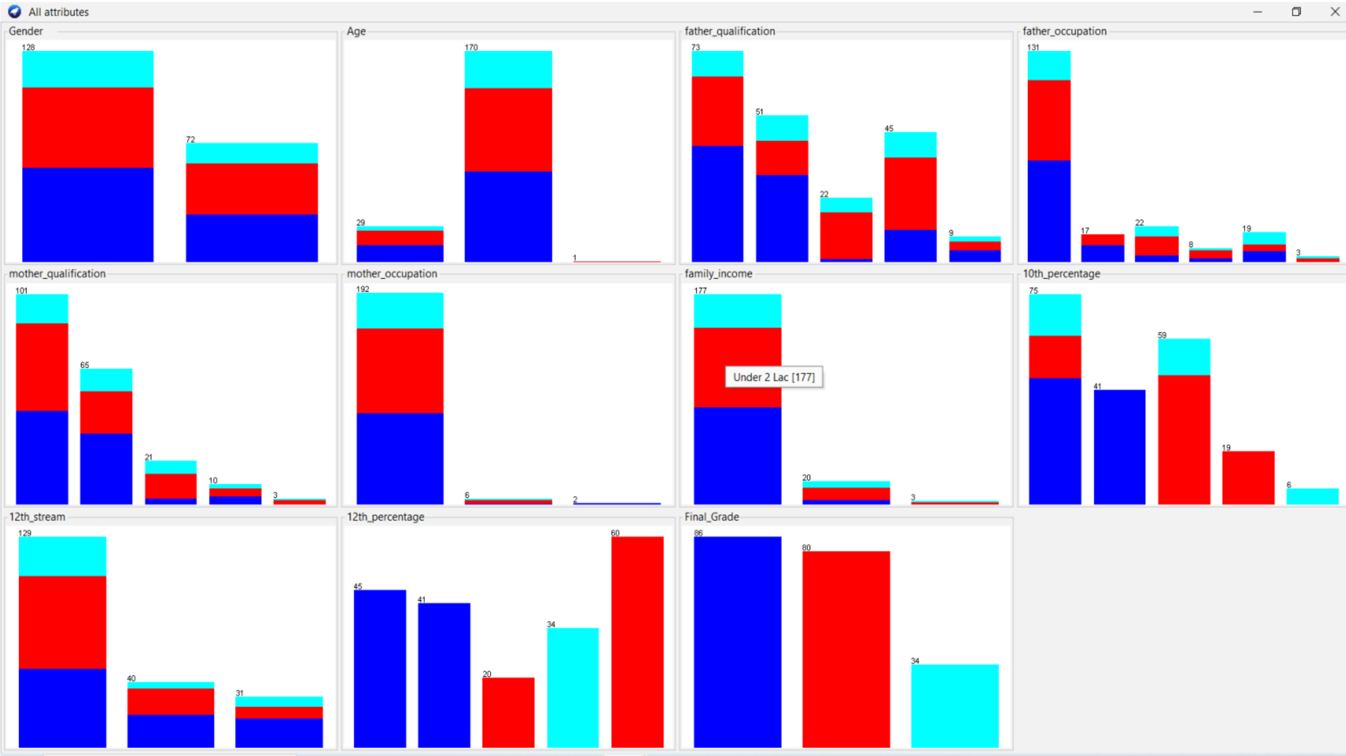

Figure 11: Visualized Result of Student's Dataset from Weka

This window shows the visualization of all variables with count for each variable. We can also check variable count in separate window for each variable.

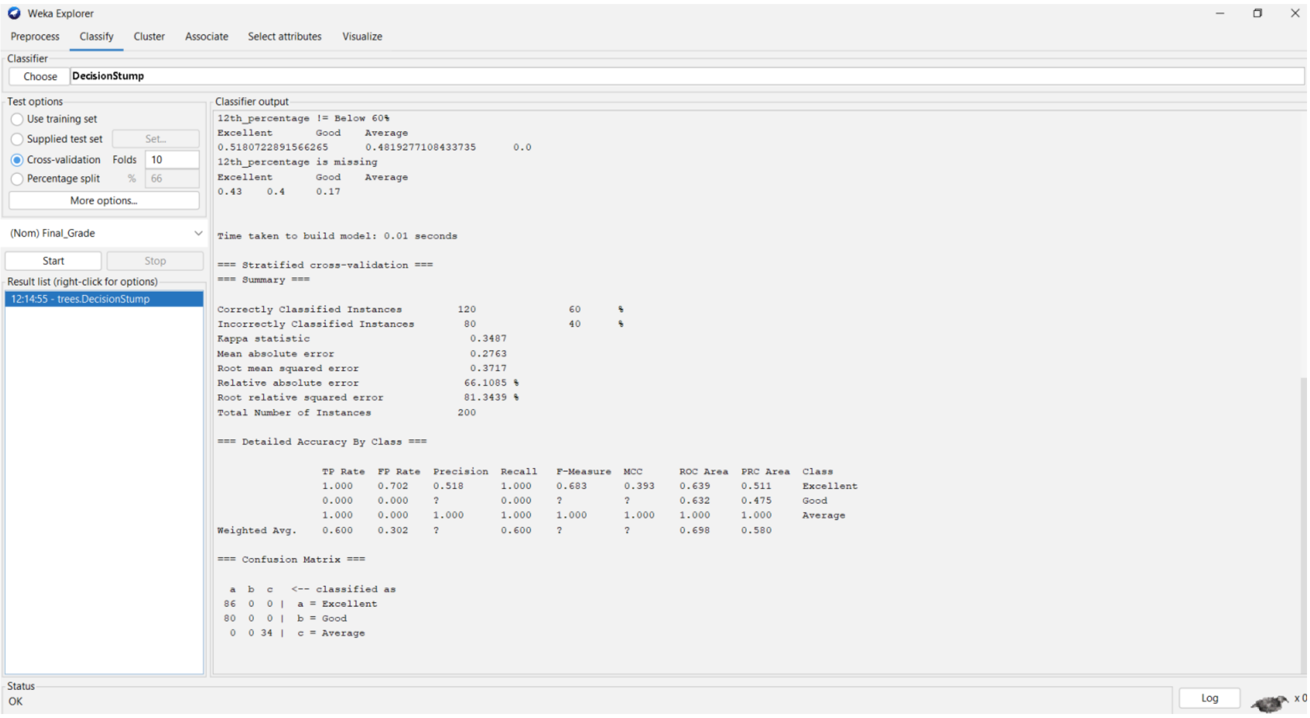

Figure 12: Decision Stump Classifier Result of Student's Dataset from Weka

This window shows the result of classification method used on this data. Confusion matrix and time to build model is also shown here. We can select different classification method using this window and analyze the output. In left panel method which we used are shown whereas in right panel the output is ready for analyze. Also in right panel correctly classified and incorrectly classified records are shown which show the performance of that particular method.

S. No.	**Course**	**No. of Students**	**No. of students Excellent**	**No. of students Good**	**No. of students Average**
1.	BA	129	48	57	24
2.	BSC	40	20	16	4
3.	BCom	31	18	7	6
Total		200	86	80	34
Grade Percentage (%)			43%	40%	17%

TABLE- III: COURSEWISE STUDENT'S FINAL GRADE DETAILS

Classifier Name	**BayesNet**	**Navive Bayes**	**Random Forest**	**IBk**	**Decision Stump**	**J48**	**PART**
Total No. of Instances	200	200	200	200	200	200	200
Correctly Classified Instances	199 (99.5%)	199 (99.5%)	200 (100%)	181 (90.5%)	120 (60%)	200 (100%)	200 (100%)
Incorrectly Classified Instances	1 (0.5%)	1 (0.5%)	0 (0%)	19 (9.5%)	80 (40%)	0 (0%)	0 (0%)
Time Taken to	0.05	0	0.11	0	0.01	0.1	0

build the Model	Second	Second	Second	Second	Second	Second	Second
Confusion Matrix	85 0 1 0 80 0 0 0 34	85 0 1 0 80 0 0 0 34	86 0 0 0 80 0 0 0 34	84 0 2 4 76 0 5 8 21	86 0 0 80 0 0 0 0 34	86 0 0 0 80 0 0 0 34	86 0 0 0 80 0 0 0 34
Kappa Statistic	0.992	0.992	1	0.8449	0.3487	1	1

TABLE IV: RESULT FROM DIFFERENT CLASSIFIER USING WEKA

Classifier Name	**BayesNet**	**Navive Bayes**	**Random Forest**	**IBk**	**Decision Stump**	**J48**	**PART**
Correctly Classified Instances	--	--	Random Forest		--	J48	--
Time Taken to build the Model		Navive Bayes	--	IBk	--	--	PART
Confusion Matrix	--	--	Random Forest			J48	PART
Kappa Statistic	--	--	Random Forest			J48	PART
Total points (4)	0	1	3	1	0	3	3

TABLE V: BEST CLASSIFIERS OF DIFFERENT MEASUREMENTS

The results from the different data mining algorithms, such as BayesNet, Navive Bayes, IBk, Decision Stump, Random forest, J48, and PART, on the data set for the different courses of students are tabulated, and the performance is analysed. A comparison table provides the total number of instances, instances that were correctly and incorrectly classified, the time it required to build a model, the confusion matrix, and the performance statistics such as Kappa statistic.

Based on these parameters, the results are interpreted as follows:

In table III course wise detail of final grade are given. There are 86(43%) Excellent students, 80(40%) Good and 34(17%) Average students in different courses according to calculation. The classifiers Random Forest, IBk, J48 and Decision Table are proven to be very effective and accurate, as shown in Table IV in (i). In this instance, 100% of the examples are correctly classified. In addition, (ii) neither Navive Bayes nor IBk took more than 0 seconds to build the model. (iii) The diagonal elements of the confusion matrix are accurately predicted by the classifiers Random Forest, J48 and PART, however Random forest is shown to be more effective as a learning model when it comes to time complexity. (iv) The classifiers Random Forest, J48, and PART algorithm are recommended in terms of Kappa Statistic. Random forest, J48, Part achieves a total score of 3 out of 4 points, which is in accordance with these classifiers' performance analysis, and produces effective and precise findings for this kind of data set.

VIII. Conclusion & Future Work

The purpose of this research is to improve the standard of school education by using data mining techniques to examine academic data from students. In this work, we used the BayesNet, Naive Bayes, Random Forest, IBk, Decision Table, J48, and PART Classification methods to classify student data. We note that the Random Forest Classifier is the most appropriate algorithm for this kind of student dataset based on trial results. Such a classification model can be used by company executives or the school management's executives to measure or visualize the students' performance in accordance with the extracted knowledge. This study will be valuable for school management, teachers and for parents in the future. With the help of data mining tools, we may generate the information after using other data mining techniques like clustering, prediction, and association rules, etc. on various eligibility requirements of industry recruiting for students. The results and conclusions obtained from the school education data mining research will play a crucial role in determining the next steps and directions for further analysis. These findings will provide valuable insights into the dataset and guide decision-making for enhancing prediction accuracy.

Based on the outcomes of the research, several actions can be considered to improve the predictive models. This may involve exploring different transformations of the dataset, such as feature engineering or scaling, to enhance the data's representation and extract more meaningful patterns. Additionally, incorporating new data sources or variables can contribute to a more comprehensive and informative analysis.

Another aspect that can be addressed is the fine-tuning of the classification algorithms' parameters.

By optimizing the algorithm settings, researchers can enhance the model's performance and increase its prediction accuracy. This may involve adjusting parameters such as learning rates, regularization factors, or decision thresholds, depending on the specific algorithms used.

The research may also uncover insights into the sufficiency and availability of school education data. It can provide recommendations to the school management regarding the data collection process. This may involve identifying gaps or limitations in the current data collection practices and suggesting improvements or additional data sources that can enhance the accuracy and reliability of the models.

Overall, the research findings will serve as a foundation for making informed decisions and implementing strategies to improve the university's data mining efforts. The recommendations provided will help guide future research endeavors and ensure that the data collection process is optimized for accurate predictions and valuable insights.

References

1. Rosemarie M. Bautista, Menchita Dumlao, Melvin A. Ballera "Recommendation System for Engineering Students' Specialization Selection Using Predictive Modeling" Third International Conference on Computer Science, Computer Engineering, and Social Media (CSCESM2016), Thessaloniki, Greece, 2016.
2. Vandna Dahiya "A SURVEY ON EDUCATIONAL DATA MINING" International Journal of Research in Humanities, Arts and Literature (IMPACT: IJRHAL) ISSN (P): 2347-4564; ISSN (E): 2321-8878 Vol. 6, Issue 5, May 2018, 23-30
3. Srinivasu Badugu "Performance Analysis of a Student during a Learning Management System using Classification Algorithms" Test Engineering and Management · Volume 82 Page Number: 7658 - 7665 Publication Issue: January-February 2020
4. Zameer Gulzar, Dr. A. Anny Leema, A.Salman Ayaz "Educational Data Mining Using SCORM Specifications on Learning Management System" International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10 No.82 (2015)

5. Surjeet Kumar Yadav, Brijesh Bharadwaj, Saurabh Pal "Mining Education Data to Predict Student's Retention: A Comparative Study", International Journal of Computer Science and Information Security, Vol. 10, No. 2, 2012
6. Amirah Mohamed Shahiri, Wahidah Husain, Nur'aini Abdul Rashid " The Third Information Systems International Conference A Review on Predicting Student's Performance using Data Mining Techniques" Procedia Computer Science 72 (2015) 414 – 422.
7. Raza Hasan , Sellappan Palaniappan , Salman Mahmood , Ali Abbas , Kamal Uddin Sarker and Mian Usman Sattar "Predicting Student Performance in Higher Educational Institutions Using Video Learning Analytics and Data Mining Techniques" Appl. Sci. 2020, 10, 3894; doi:10.3390/app10113894
8. Oswaldo Moscoso-Zea, Mayra Vizcaino, Sergio Luján-Mora "Evaluation of Methods and Algorithms of Educational Data Mining" Conference: 2017 Research in Engineering Education Symposium At: Bogota, Colombia
9. A.S. Arunachalam, T.Velmurugan " A Survey on Educational Data Mining Techniques" International Journal of Data Mining Techniques and Applications Volume: 05 Issue: 02 December 2016, Page No.167-171 ISSN: 2278-2419
10. Alisa Bilal Zorić " Benefits of Educational Data Mining" Journal of International Business Research and Marketing, vol. 6, issue 1, pp. 12-16, November 2020
11. Suhirman, Jasni Mohamad Zain, Haruna Chiroma, and Tutut Herawan "Data Mining for Education Decision Support:A Review" iJET – Volume 9, Issue 6, 2014
12. Atta-Ur-Rahman, Kiran Sultan, Nahier Aldhafferi, Abdullah Alqahtani "EDUCATIONAL DATA MINING FOR ENHANCED TEACHING AND LEARNING" Journal of Theoretical and Applied Information Technology 31st July 2018. Vol.96. No 14
13. Saja Taha Ahmed Prof. Dr. Rafah shihab Al-hamdani Dr. Muayad Sadik Croock "Studying of Educational Data Mining Techniques" International Journal of Advanced Research in Science, Engineering and Technology Vol. 5, Issue 5 , May 2018
14. Sedigheh Abbasnasab Sardareh, Mohd Rashid Mohd Saad, Abdul Jalil Othman, RosalamChe Me "Enhancing Education Quality Using Educational Data"Scholars Journal of Arts, Humanities and Social Sciences ISSN 2347-5374 (Online) Sch. J. Arts Humanit. Soc. Sci. 2014; 2(3B):440-444
15. Ashish Dutt, Maizatul Akmar Ismail, and Tutut Herawan "A Systematic Review on Educational Data Mining" IEEE Access · January 2017 DOI: 10.1109/ACCESS.2017.2654247
16. Carla Silva, José Fonseca "Educational Data Mining: a literature review" Advances in Intelligent Systems and Computing · September 2017 DOI: 10.1007/978-3-319-46568-5_9
17. Cristóbal Romero, Sebastián Ventura "Educational Data Mining: A Review of the State-of-the-Art" IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews) · December 2010
18. Ms. Falguni Suthar, Ms. Hiralben Patel, Dr. Bhavesh Patel " A Study on Educational Data Mining" International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; Volume 7 Issue II, Feb 2019-

19. Saeed Aghabozorgi, Hamidreza Mahroeian, Ashish Dutt, Teh Ying Wah, and Tutut Herawan "An Approachable Analytical Study on Big Educational Data Mining" International Conference on Computational Science and Its Applications June 2014 DOI: 10.1007/978-3-319-09156-3_50
20. Mohammad Shiralizadeh Dezfoli, Behzad Soleimani Neysiani, Dr. Naser Nematbakhsh "Identifying affecting factors on prediction of students' educational statuses: A case study of educational data mining in Ashrafi Esfahani University of Isfahan of Iran" 10th International Conference on Information and Knowledge Technology (IKT 2019)
21. Mewati Ayub, Hapnes Toba, Steven Yong & Maresha C. Wijanto " Modelling students' activities in programming subjects through educational data mining"
    Global Journal of Engineering Education · November 2017
22. Mussa S. Abubakari *, Fatchul Arifin, Gilbert G. Hungilo "Predicting Students' Academic Performance in Educational Data Mining Based on Deep Learning Using TensorFlow" I. J. Education and Management Engineering, 2020, 6, 27-33 Published Online December 2020 in MECS (http://www.mecs-press.org/) DOI:10.5815/ijeme.2020.06.04
23. Mais Haj Qasem, Raneem Qaddoura, Bassam Hammo "Educational Data Mining (EDM): A Review" Conference: New Trends in Information Technology - (NTIT) 2017
24. Maria P. G. Martins, Vera L. Miguéis, D. S. B. Fonseca "Educational Data Mining: A Literature Review" Conference: 2018 13th Iberian Conference on Information Systems and Technologies (CISTI) June 2018 DOI: 10.23919/CISTI.2018.8399281
25. B.M. Monjurul Alom, Matthew Courtney "Educational Data Mining: A Case Study Perspectives from Primary to University Education in Australia" I.J. Information Technology and Computer Science, 2018, 2, 1-9 Published Online February 2018 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijitcs.2018.02.01
26. Anwar Ali Yahya, Addin Osman, Ahmad Taleb "Swarm Intelligence in Educational Data Mining" Conference: Machine Learning and Data Analytics Symposium – MLDAS.At: Doha, Qatar March 2014
27. ASHISH DUTT, MAIZATUL AKMAR ISMAIL, TUTUT HERAWAN "A Systematic Review on Educational Data Mining" IEEE Access 5:15991-16005 January 2017 DOI: 10.1109/ACCESS.2017.2654247
28. Mewati Ayub, Hapnes Toba, Maresha Caroline Wijanto, Steven Yong "Modelling Online Assessment in Management Subjects through Educational Data Mining" International Conference on Data and Software Engineering (ICoDSE) 2017
29. Smita J. Ghorpade1, Seema S. Patil2, Ratna S. Chaudhari3 "EDUCATIONAL DATA MINING: TOOLS AND TECHNIQUES STUDY" International Journal of Research and Analytical Reviews (IJRAR) Volume 7, Issue 4, November 2020
30. Mudasir Ashraf, Dr. Majid Zaman, Dr. Muheet Ahmed, S. Jahangeer Sidiq "Knowledge Discovery in Academia: A Survey on Related Literature" International Journal of Advanced Research in Computer Science Volume 8, No. 1, Jan-Feb 2017
31. Ebtehal Ibrahim Al-Fairouz1, Mohammed Abdullah Al-Hagery "Students Performance: From Detection of Failures and Anomaly Cases to the Solutions-Based Mining Algorithms" International

Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 10 (2020), pp. 2895-2908

32. K. Touya, Mohamed Fakir "Mining Students' Learning Behavior in Moodle System" Journal of Information Technology Research, 7(4), 12-26, October-December 2014
33. Leena Khanna, Dr. Shailendra Narayan Singh, Dr. Mansaf Alam "Educational Data Mining and its Role in Determining Factors Affecting Students Academic Performance: A Systematic Review" Conference: 2016 1st India International Conference on Information Processing (IICIP) August 2016 DOI: 10.1109/IICIP.2016.7975354
34. Mohammed Alsuwaiket, Christian Dawson, Firat Batmaz "Measuring the Credibility of Student Attendance Data in Higher Education for Data Mining" International Journal of Information and Education Technology, Vol. 8, No. 2, February 2018
35. Dr. Srinivasu Badugu "Performance Analysis of a Student during a Learning Management System using Classification Algorithms" February 2020 Test Engineering and Management 82:7658 – 7665
36. Mudasir Ashraf* and Majid Zaman "Tools and Techniques in Knowledge Discovery in Academia: A Theoretical Discourse" International Journal of Data Mining and Emerging Technologies DOI: 10.5958/2249-3220.2017.00001.5