CTU Journal of Innovation and Sustainable Development

ISSN 2588-1418 | e-ISSN 2815-6412

DOI:10.22144/ctujoisd.2024.321

Students' performance prediction employing Decision Tree

Abtahi Ahmed[1], Farzana Akter Nipa[1], Wasi Uddin Bhuyian[1], Khaled Md Mushfique[1], Kamrul Islam Shahin[2], Huu-Hoa Nguyen[3*], and Dewan Md. Farid[1]

[1]Department of Computer Science and Engineering, United International University United City, Bangladesh

[2]Software Engineering, The Maersk Mc-Kinney Moller Institute University of Southern Denmark, Denmark

[3]College of Information and Communication Technology, Can Tho University, Viet Nam

**Corresponding author (nhhoa@cit.ctu.edu.vn)*

Article info.

Received 28 Jun 2024
Revised 30 Aug 2024
Accepted 30 Sep 2024

Keywords

Classification, educational data mining, machine learning, students' performance prediction

ABSTRACT

An optimized educational community is a must in this modern era. The intersection of educational activities and the transformative potentials of Educational Data Mining (EDM) should be traversed, highlighting the reasoning behind the importance of EDM. Prior prediction of how a student stands academically, can facilitate them towards a much safer approach with their life decisions. This study uses the vast power and analytical domain of EDM, combining it with machine learning models, upholding an accurate prediction of students' academic performance. The study consists of a dataset containing academic, demographic and social data of undergraduate students. The paper aims to analyze comprehensively the features that act behind academic performance. Lastly, it compares the impact of non-academic data separately on a student's performance and with academic data as well. Traditional machine learning algorithms perform quite well in general, with SVM giving a best accuracy of around 95% with academic data, while training and testing the model without academic data still gives a good performance of 93%. The hierarchical tree from Decision Tree visualizes the key features, which include past results, family members' qualification levels and their jobs, hobbies of the student, commute time, and more.

1. INTRODUCTION

Education is the most potent weapon at the hands of a human. A well-educated person has the capability to achieve the best possible outcomes in life through sheer knowledge. In this modern era, people can stumble upon many educational programs and make a certain career within that accordingly. However, quite often, people make wrong decisions in choosing the path even after studying a specified educational program, leading them to a dead end and causing them loss of money and time, which is a major concern. In this day and age where there are a variety of professional degrees to choose from, it can be quite overwhelming to select one that is best suited for an individual, due to societal or peer pressure, or the lack of basic understanding of the degree. A study conducted by the Bangladesh Government (2022) shows that most of the students who enroll for tertiary education (undergraduate degree) tend to drop out quite soon, and a low percentage of them actually complete their degrees. Such occurrences waste a country's resources, and the students are misled into carrying out activities which may harm their growth and productivity. In such cases, Educational Data Mining (EDM) comes in handy.

EDM is a field of study that uses machine learning algorithms to analyze educational data and provides outcomes accordingly. It analyzes the student's learning behavior using traditional mining algorithms (Romero & Ventura, 2010; Mohamad & Tasir, 2013). Educational Data Mining methods explore the multiple layers of consequential grouping in educational data (Ocumpaugh et al., 2014). It is possible to determine a student's learning patterns and build a new individualized learning experience, as well as provide future performance predictions using educational data accordingly. Predicting students' performances have shown promising effectiveness in identifying which students are at-risk of dropping out and what could be the dropout rate (Marbouti et al., 2016; Kumar et al., 2017; Alamri & Alharbi, 2021). There are also some other factors such as demographic, social and psychological attributes that have a great impact on how a student is going to perform. A study shows that it is possible to identify the students who are prone to dropout through their demographic data (Kotsiantis et al., 2003). Academic achievements are seen as the foundation of predicting a person's future occurrences, therefore making it critical for a person to have a strong academic background. However, if a person chooses a study path that they are not comfortable with, it might be too late by the time he or she discovers it. Therefore, it is essential to address such issues by investigating the factors associated with students' success and finding ways for early intervention to help poorly performing students (Jayaprakash et al., 2020). Such a study is just a small step in creating a personalized digital education system for the near future. Moreover, carrying out EDM studies in Bangladesh will also reduce gaps in the education system of the country and identify which features are more important in the region compared to other regions.

Initially, the study focused solely on predicting academic performance in the Object-Oriented Programming (OOP) course. To broaden the scope and improve generalizability, the dataset has been expanded to include participants from the Structured Programming Language (SPL) course. By incorporating data from both OOP and SPL, this study now provides insights into student performance across multiple technical courses, making the predictive model applicable to a wider range of academic contexts.

In this paper, we have investigated the issues faced by students throughout their academic time and have quantified the data they provided to conduct performance analysis of ML algorithms. We have made a comprehensive analysis of the type of data that impacts students' performance through surveys and in-person interviews. We collected raw data through assessments from students enrolled in the Object-Oriented Programming course at United International University. Merging the academic data with additionally collected social and demographic data helped to create a dataset that is attentive to detail. The literature review helped us identify the features from existing published datasets, and it also helped us figure out which algorithms work well in the EDM domain. Creating a classification model from these data helped us make grade predictions and reduce the hassle of students who are facing uncertainty with their study approach as well as conduct an investigation to identify the features that are key to the students' success.

Further parts are sequenced in the way section (2) presents the materials and methods. Section (3) discusses the results and discussion. Lastly, Section (4) provides a conclusion for the study and suggests future research directions.

2. MATERIALS AND METHOD

This section brings forth the tasks engaged while working on the proposed system, the methodological framework, as well as the rational thought for the chosen approach. The major steps are outlined below:

1. The creation of the survey and the reasoning behind the selection of the features of the dataset have been explained.
2. The data collection strategy and the preprocessing to remove inconsistency and keep the quality well maintained of the data.
3. We have studied and executed multiple machine learning algorithms to find the one with the best accuracy rate.
4. Lastly, we have identified the impact of demographic and social data separately on the students' performance.

2.1. Data preparation and preprocessing

The survey section can be concluded as the crucial strait of this study. We grouped the data into 3 types, which are prior academic data, demographic data, and social data. Demographic data consists of background information such as age, gender, parents' qualifications, commute time to campus, etc. Social data grasps how the student interacts with the people around and the surroundings. It contains data such as relationship status, perceived support received from family and friends, group study hours, club or extracurricular activities, etc. Finally academic data consists of current and past academic results of computer-related courses as well as English. In the survey, we have added the attributes that have more impact on how the students can perform. Relationships among these attributes illustrate how they mutually influence each other. These attributes have been collected from previous work in the EDM field (Cortez & Silva, 2008; Amrieh & Hamtini, 2016; Dataset-2, 2018; Dataset-3, 2019). Some of the impactful attributes which have been used in prior work that have more influence on a students' progress include age, living with family or not, mother's qualification, parental relationship status, economic status, weekly study hours, reading frequency, relationship status, prior grades and hobbies. The aim is not only to input and analyze academic data, but more importantly to identify the demographic and social data associated with the performance of students. To start off, we conducted in-person interviews of teachers and students to obtain their feedback on what makes a student perform well. Along with this, the datasets which we reviewed above provided us with the necessary features that we can include in our dataset. The survey contained carefully crafted questions that allowed us to extract extensive data about each individual. Since the data collected is confidential, we are unable to publish them at this moment. However, we are working with the university authority to collect and organize more data so that we can make it publicly available.

Initially, the dataset consisted of 250 participants from a single Object-Oriented Programming (OOP) course. To address the limitation of sample size, the study was expanded to include data from an additional 300 participants enrolled in the Structured Programming Language (SPL) course. This brings the total number of participants to 550, allowing for a more comprehensive analysis of student performance across multiple courses. The inclusion of participants from both OOP and SPL provides a more diverse dataset, enhancing the generalizability of the findings.

The demographic of the audience for the survey, conducted through Google Forms, consisted of Bangladeshi university undergraduate students studying in their 1st year who have completed their Object-Oriented Programming (OOP) and Structured Programming Language (SPL) course. We have a collection of around 550 participants' data, with the median age being 21. The collection of the qualitative data from these participants and their convergence with preprocessing procedures allowed us to create a reliable dataset. There are 59 columns in our dataset, comprising demographic (29), social (7) and academic (22) features and the last column for the class (grade obtained in OOP & SPL course). The dataset was divided into a training set (70%) and test set (30%). To preprocess the data, we used the OrdinalEncoder from scikit-learn library to encode the categorical (discrete) features into integers, and the LabelEncoder from scikit-learn library to encode the target class values into n integers (in this case 11 values for the 11 grades provided for a course by the university) for transformation to better fit the ML models.

Table3 shows some of the questions in our survey, labeled as (D), (S) and (A) representing Demographic, Social and Academic features, respectively. In addition to this we have also collected secondary and higher secondary results of some courses which were English, Physics, Math and ICT along with grades of Introduction to Computer Science (ICS), Discrete Mathematics, Structured Programming Language and English from the first year of tertiary education.

Similarly, we conducted the same survey for the prediction of SPL course grades as well. For that, the main difference in the questions was that we did not include the grades of the SPL course itself.

2.2. learning algorithms

Classification and clustering are two major domains in machine learning. Since our dataset comprises textual or numerical values, and we want to build a multi-class supervised classification model, the classical ML algorithms would work just fine. These offer good accuracy while using minimal computational resources. Many algorithms have been developed over the past in the vast history of classification algorithms, and many have stood tall after being tried and tested. Some of them include Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Naive Bayes (NB). The top most used algorithms in EDM for performance prediction and feature analysis are Decision Tree, Naive Bayes, Support Vector Machine, and K-Nearest Neighbor (Jordan &

Mitchell, 2015). A brief description of the 5 chosen algorithms is given below.

Support Vector Machine (SVM) is a supervised machine learning algorithm for classification which works for both linear and non-linear data. The original training data is transformed into a higher dimension using nonlinear mapping. Then it finds the linear optimal separating hyperplane through support vectors in the new dimension. It separates the hyperplane from the following equation:

$$W.X + b = 0$$

Decision Tree (DT) has a hierarchical tree structure with vertices (root, internal and leaves) and edges. It splits an internal vertex by splitting criteria and repeats the process until all or most rows are classified accordingly. Gini Impurity is a method to split attributes. It is the probability of incorrectly classifying a sample, found by using the formula below, where pi represents the probability. The attribute with the lowest Gini impurity is chosen as the splitting attribute.

$$GiniImpurity = 1 - \sum_{i} \square \ (pi)^2$$

Logistic Regression (LR) is a statistical model that estimates the probability of an event occurring based on independent variables. It first calculates the beta parameters, or coefficients based on the maximum likelihood estimation (MLE) of all the independent variables, shown in the equation below, and after applying mathematical operations (logging and summing), a predictive probability is generated for each input.

$$ln(\pi/(1-\pi)) = \beta 0 + \beta 1 * X1 + ... \beta k * Kk$$

Naive Bayes (NB) is a probabilistic classifier based on Bayes' Theorem. It uses conditional probability, which represents the probability of an event occurring given some other event has occurred. A simple calculation of the probability of an event Y occurring given that an event X has occurred can be calculated from the following equation:

$$P\ (Y\ |X) = P\ (X|Y\ )P\ (Y\ )/P\ (X)$$

K-Nearest Neighbors (KNN) is a non-parametric algorithm that uses proximity to predict classes, an idea that similar instances will be close to each other based on their features. The distance between the features of an instance and other data points is calculated using any of the distance metrics, such as Euclidean distance, Manhattan distance, etc. with the value of k determining how many neighbors will be checked.

We want to train our dataset in the supervised learning algorithms mentioned above to identify the best one, which we would fine-tune to get the best results.

3. RESULTS AND DISCUSSION

This section describes the background study and discusses the results and its inferences.

3.1. Background study

In terms of educational data mining (EDM), researchers have designed multiple frameworks and models to predict students' academic performance. Regarding this field, authors Nosseir and Fathy developed a framework applying a neural network for GPA prediction, additionally a mobile application for topic-based testing, and a fuzzy model for performance estimation. Their dataset involved collecting academic data from university students, preprocessing it, and applying the Levenberg-Marquardt technique. The mobile application focused on testing students in Relational Database Management Systems, while the fuzzy model employed fuzzy theory for performance percentage estimation (Nosseir & Fathy, 2020). On the other hand, Pathan et al. (2014) proposed a model for enhancing students' programming skills, using a decision tree-based mining model to predict C programming skills gave us some insights. The model achieved 87% accuracy and identified factors such as prior programming experience, class attendance, participation, and online resource usage linked to student success. Limitations of the study had included a small sample size and a single-context setting. Another decision tree approach from Nahar et al. (2021) emphasized EDM's role in predicting student performance, using decision trees, naive Bayes, and support vector machines. Their decision tree classifier achieved 82% accuracy, highlighting the importance of previous course performance, attendance, CGPA, and extracurricular activities. Similarly in the paper (Sokkhey et al., 2020) the authors explored the use of EDM to anticipate mathematics performance, categorizing EDM models into statistical analysis, machine learning, and deep learning and identified previous math performance, attendance, gender, parents' education, and extracurricular activities as key predictors which later on achieved an 84% accuracy with the Random Forest algorithm.

Diving deeper into more machine learning models, Alturki and Alturki (2021) employed six supervised data mining algorithms to predict graduation results, with naive Bayes performing the best overall. The study discussed classifier performances and applied various feature selection techniques, emphasizing the need for predictive accuracy (Alturki & Alturki, 2021). Another similar approach was taken by Mustafa Yağcı, who utilized EDM to forecast academic outcomes, auditing machine learning algorithms such as Random Forest, Support Vector Machines, Logistic Regression, Naïve Bayes, and k-Nearest Neighbors. With a dataset of 10,000 students, the Random Forest achieved the highest with 85% accuracy rate, and Yağcı concluded that EDM is valuable for predicting performance and identifying at-risk students (Yağcı, 2022). Tomasevic et al. (2020) also provided a comprehensive analysis and comparison of supervised machine learning methods for evaluating students' exam performance. On the Open University Learning Analytics Dataset (OULAD), different combinations of student-related variables, such as past performance, engagement, and demographics, were used as input for training and testing the algorithms. When only historical performance and engagement data were used as inputs, the ANN technique beat models that included demographic data in both classification and regression, obtaining the highest performance rate. Using this methodology, it was feasible to analyze a number of supervised learning algorithms in-depth quantitatively and determine which features worked best for predicting students' exam success.

There were hybrid approaches as well which were proposed by the authors (Feng et al., 2022) a method combining EDM and convolutional neural networks to analyze and predict students' performances, preventing dropout risks. They introduced an improved K-Means Cluster algorithm and achieved high accuracy in clustering effects and cross-validation techniques. In a systematic literature review, Roslan and Chen identified research trends in EDM studies, emphasizing the impact of student's past records and demographics. They found that classification was the most used data mining approach, with the Decision Tree Classifier being the most popular algorithm. Demographic factors such as age, family socioeconomic level, gender, and ethnicity were linked to student performance, along with additional attributes like student psychology, activities, e-learning platforms, and instructor/course attributes (Roslan & Chen, 2022). Also, Shafiq et al. (2022) focused on EDM and Predictive Analytics for student retention, identifying popular models like Decision Trees and Random Forests. Both of the papers emphasized the importance of selected dataset classes and features for effective prediction. The review highlighted the need for more generalized studies, larger sample sizes, and correlation exploration among different studies.

Yang and Li (2018) introduced analyzing tools so that student performance can be analyzed, identify what are the affecting variables, how students can be more advanced, and whether students have the capacity to perform better. They incorporated both performance and non-performance related features and used the Student Attribute Matrix (SAM) for categorizing the students. Moreover, the paper included a tool which used Back Propagation Neural Network (BP-NN) for estimating the performance of the students. BP-NN was also used to identify the attributes that are affecting mostly students' performance (Yang & Li b, 2018). Overall, we can see that quite a few machine learning approaches have been used, including the less frequently used unsupervised learning and neural networks. The pie chart in figure 1 shows the top different machine learning algorithms identified in the literature review that have been proven to work well in the field of EDM.

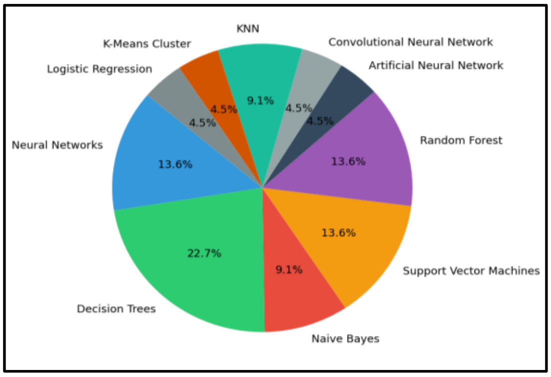

Figure 1. Identification of algorithms used

3.2. Performance analysis

In this segment, we have provided an in-depth analysis of the outcomes and performances of all models. Table- 1 shows us the overall accuracy of the models that we have employed. Among them, SVM has given the highest accuracy in terms of predicting future performance, which aligns with our study of the literature.

While SVM, DT and LR performed quite well, we can clearly see that NB and KNN had poor results. Such a fall in accuracy can be explained using the fundamental concepts of these algorithms. While SVM, DT and LR all get down to optimizing their equations until all instances are classified accurately in the training phase, NB uses the idea that features are independent of each other, where the probability of each feature is separately calculated. This makes the model perform poorly in datasets where some features are closely related to each other, for example, number of group study hours per week and commute time to the university campus. The same can be said for KNN, where it doesn't give more priority to some features than others. KNN uses the concept of majority voting, so if the majority of the surrounding data points (in this case, 9.09% of the data points, since there are 11 classes in our dataset) vote for one of the classes, KNN labels the sample to that class. This value of 9.09% is too small, and therefore, can easily lead to misclassification. In our model, the number of neighbors for KNN was 5, but even after increasing the number of neighbors, we found worse results. So, we can say that attention to each feature along with the relationships between features would provide a good output for datasets with lots of features, as is the case with our dataset, which is accomplished by SVM, DT and LR algorithms.

To address the issue of focus, we have expanded our study to include the prediction of academic performance in the SPL course alongside the OOP course. By incorporating two distinct courses, we have broadened the applicability of our model beyond a single subject, ensuring a more comprehensive evaluation of student performance. The addition of SPL not only increases the sample size but also introduces variability in course content, teaching methods, and assessment styles. This allows us to assess whether the factors influencing student success in one course are similar or different in another, providing more refinement into the predictors of academic performance.

The initial results, showing a high accuracy of 95%, raised concerns about potential overfitting. To address this, we increased the dataset size for OOP to 300 and 250 for SPL, increasing the variability in the dataset. This additional data helps mitigate overfitting, as the model is now trained on a more diverse set of student performance metrics.

3.3. Impact of academic information

Results and discussion We wanted to figure out the influence of academic features on a student's performance, mainly the lack of it. Therefore, we wanted to train the model without including any academic features such as grades in past courses and past board exams, and keeping only the demographic and social data, as well as some behavioral data (such as hours spent studying weekly, type of education resource preferred, etc.). Hence, after separating the academic features from our dataset, the SVM-trained model gave us the best results, though a lower accuracy (93%) than before. Table 2 shows the evaluation metrics of the model trained without academic data. This confirms the obvious idea that past academic performance does influence future performance, but the minute difference between the model's performance with and without academic data shows other factors (demographic, social and behavioral) have a significant impact on the performance. Decision Tree has a hierarchical tree structure with vertices (root, internal and leaves) and edges. It splits an internal vertex by splitting criteria and repeats the process until all or most rows are classified accordingly. Using a decision tree allows us to visualize the root and various internal vertices and identify the features which play a part in determining the splitting criteria. Figures 2 & 3 show simple decision trees for our model with a depth of 3, with academic data and without academic data. With academic data in the dataset, we can see study hours per week, father's qualification level and results of past courses such as Structured Programming Language, Introduction to Computer Science, Discrete Math and English are considered to be the splitting points. For the model without academic data, crucial factors include whether the student has any sibling or cousin with higher qualifications than the student, whether the student has any corporate job holder, engineer, scientist or teacher as close relatives, whether the student likes reading, writing, watching movies and cooking, and how long it takes for the student to commute to and from the university campus.

Similarly, the decision trees for the SPL model have been created, both with and without academic data, each limited to a depth of 3. In the scenario where academic data is included, key splitting criteria emerge, such as weekly study hours, parent's qualification level, and the student’s performance in relevant courses like Introduction to Computer Science, Discrete Math and English. These academic factors strongly influence the model’s decision-making process, highlighting their impact on predicting performance in SPL.

With more depth in the tree, we would see more features as splitting points in the hierarchy.

3.4. Tables and figures

Table 1. Performance metrics of all algorithms with academic data

Algorithms	Accuracy OOP (%)	Accuracy SPL (%)
SVM	95%	93.4%
DT	93.3%	91%
LR	93.3%	90%
KB	68.3%	54.4%
KNN	56.7%	40.2%

Table 1 presents the performance metrics while considering the academic data, whereas Table 2 presents the performance metrics of the students without considering their previous academic data.

Table 2. Performance metrics without academic data

Algorithms	Accuracy OOP (%)	Accuracy SPL (%)
SVM	93%	87.5%
DT	91.7%	82.2%
LR	91.7%	83%
KB	65%	49.5%
KNN	55%	41.3%

Figures 2 and 3 present the decision trees for both academic data and non-academic data cases.

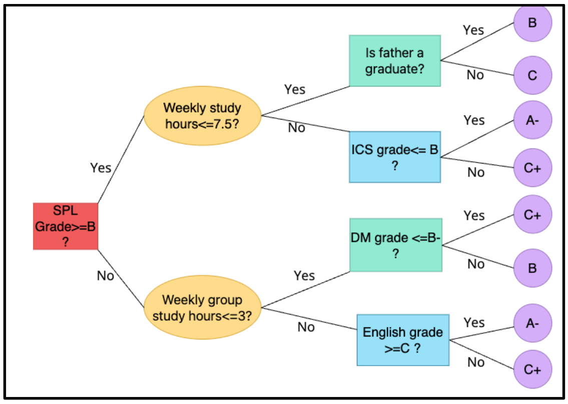

Figure 2. Decision Tree with academics

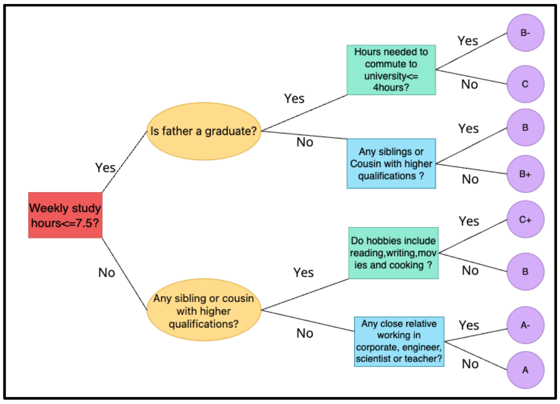
Weekly study hours<=7.5?
Yes
No
Is father a graduate?
Yes
No
Any sibling or cousin with higher qualifications?
Yes
No
Hours needed to commute to university<= 4hours?
Yes
No
B-
C
Any siblings or Cousin with higher qualifications ?
Yes
No
B
B+
Do hobbies include reading,writing,movies and cooking ?
Yes
No
C+
B
Any close relative working in corporate, engineer, scientist or teacher?
Yes
No
A-
A

Figure 3. Decision Tree without academics

Table 3. Sample survey questions

Questions	Options
Age	Input Numerical Value
Gender	Male or Female
Estimated total number of hours needed to commute to and from university everyday	Input Number
Are you living with your family?	Yes or No
Number of people in your family	Input Number
Mother’s highest academic qualification	Not graduate/ graduate/ post graduate
Father’s highest academic qualification	Not graduate/ graduate/ post graduate
Parents relationship status	Still married/ Separated
Do you have any siblings or closer relatives who have higher qualifications than you and are closely involved with your studies?	Yes or No
How much would you say your family is involved with your studies?	On a scale of 5
Tick all the jobs your family or close relatives have.	Corporate, Engineer, Healthcare Specialist, Entrepreneur, Artist, Teacher, Businessman, Scientist, Civil service, Technician, Politician, NGO Employee, Lawyer, Banker
Family’s socio-economic status	Middle class/ Upper class/ Lower class
Do you have any job outside your studies?	Yes or No
Identify some of your hobbies.	Reading, Gaming, Cooking, Sports, Writing or watching movies, Travelling and film making, Arts

Questions	Options
	and Crafts, Dance and Drama, Music, Gardening, Bike touring, Learning new things, Chess, Programming, Exploring new things, learning new technical staffs, Gaining knowledge, Collecting productive information, Teaching, Tinkering with things related to Linux, Robotics/ Photography
Are you currently involved in any romantic relationship?	Yes or No
How many hours a week do you study with your classmates/ friends in group sessions?	Input Numbers
How supportive are your friends/classmates?	On a scale of 5
How supportive are your family members?	On a scale of 5
How healthy and supportive is the education environment in your university?	On a scale of 5
How healthy and supportive is the education environment in your place/ residence?	On a scale of 5
Are you involved in any extracurricular activities that are not related to computer science and engineering?	Yes or No
Identify resources you use to study.	Documents provided by class teacher, textbooks, video tutorials on different platform like YouTube, Other learning platforms on the internet, AI platform
Tick the Platforms you use regularly.	Coursera, W3Schools, GeeksForGeeks, Stack Overflow, Quora , JavaTPoint, Programiz
How often do you read or use resources such as textbooks, internet resources, etc?	On a scale of 5 (5 means 5 days a week)
How often do you access the E-learning management system?	On a scale of 5 (5 being most frequent at once everyday)
Do you participate in any competitive programming?	Yes or No
Are you currently under any scholarship/ waiver scheme?	Yes or No

4. CONCLUSION

Using the data mining field for educational purposes is a powerful way to enhance educational outcomes. From gaining insights into students' learning patterns to developing personalized learning experiences, EDM can genuinely bring changes in the educational system and make the students learn how and when to approach a specific field of study. With the help of EDM, there will be a reduction in the rate of students experiencing depression due to their academic results, as well as a decrease in the number of students making incorrect decisions regarding their career path. From our research, we found that traditional ML algorithms like SVM and DT work just fine, though there are some exceptions (NB and KNN) given the diversity and size of the feature set. The study only considered one particular course for the dataset, which does not consist of a lot of students. So, the dataset is quite small to provide the best accuracy of the model. Other crucial factors which were not considered for our research due to the lack of expertise are the mental and psychological aspects. It is not certain that the students are at their full potential when they are sitting for the assessments, due to physical or mental health troubles. So, the accuracy is also not consistent for all samples. Therefore, it is suggested that bigger research be carried out to include these aspects as well in the dataset to gain a better understanding of the students. Furthermore, teaching methods and teachers ' data should also be considered as features. Complex algorithms can be used to include all the courses of an undergraduate program, which can be used to predict the grade of any course at any time of the student's undergraduate time. This has the potential to generate a degree plan, helping the student figure out which courses to take in which semester based on the performances of past courses. Such models would also assist students in determining the major that they may declare, perhaps using unsupervised learning algorithms such as K-Means Clustering.

REFERENCES

Alamri, R., & Alharbi, B. (2021). Explainable student performance prediction models: A systematic review. *IEEE Access*, *9*, 33132-33143. https://doi.org/10.1109/ACCESS.2021.3061368

Alturki, S., & Alturki, N. (2021). Using educational data mining to predict students' academic performance for applying early interventions. *Journal of Information Technology Education: JITE. Innovations in Practice: IIP*, *20*, 121-137.

Amrieh, E. A., & Hamtini, T. (2016). *Students' academic performance dataset*. https://www.kaggle.com/datasets/aljarah/xAPI-Edu-Data

Cortez, P., & Silva, A. M. G. (2008). *Student Performance*. https://archive.ics.uci.edu/dataset/320/student+performanc

Bangladesh Government. (2022). *Education Statistics of Bangladesh Bureau of Educational Information and Statistics*. https://banbeis.portal.gov.bd/40

Feng, G., Fan, M., & Chen, Y. (2022). Analysis and prediction of students' academic performance based on educational data mining. *IEEE Access*, *10*, 19558-19571. https://doi.org/10.1109/ACCESS.2022.3151652.

Jayaprakash, S., Krishnan, S., & Jaiganesh, V. (2020). Predicting students' academic performance using an improved random forest classifier. In *2020 international conference on emerging smart computing and informatics (ESCI)* (pp. 238-243).

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260. https://doi.org/10.1126/science.aaa841

Kaggle. (2019). *Higher education students performance evaluation*. https://www.kaggle.com/datasets/csafrit2/higher-education-students-performance-evaluation

Kotsiantis, S. B., Pierrakeas, C., & Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. In *Knowledge-Based Intelligent Information and Engineering Systems: 7th International Conference, KES 2003, Oxford, UK, September 2003. Proceedings, Part II 7* (Vol. 2774, pp. 267-274).

Kumar, M., Singh, A. J., & Handa, D. (2017). Literature survey on educational dropout prediction. *International Journal of Education and Management Engineering*, *7*(2), 8-19. https://doi.org/10.5815/ijeme.2017.02.02

Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers and Education*, *103*, 1-15. https://doi.org/10.1016/j.compedu.2016.09.005

Mohamad, S. K., & Tasir, Z. (2013). Educational data mining: A review. *Procedia - Social and Behavioral Sciences*, *97*, 320-324. https://doi.org/10.1016/j.sbspro.2013.10.240

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*(6), 6051-6067.

Nosseir, A., & Fathy, Y. (2020). A mobile application for early prediction of student performance using fuzzy logic and artificial neural networks. *International Journal of Interactive Mobile Technologies*, *14*(2), 4-18. https://doi.org/10.3991/ijim.v14i02.10940

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology*, *45*(3), 487-501. https://doi.org/10.1111/bjet.12156

Pathan, A. A., Hasan, M., Ahmed, M. F., & Farid, D. M. (2014). Educational data mining: A mining model for developing students' programming skills. In *The 8th International Conference on Software, Knowledge, Information Management and Applications (SKIMA 2014)* (pp. 1-5). IEEE. https://doi.org/10.1109/SKIMA.2014.7083552

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews)*, *40*(6), 601-618. https://doi.org/10.1109/TSMCC.2010.2053532

Roslan, M. B., & Chen, C. (2022). Educational data mining for student performance prediction: A systematic literature review (2015-2021). *International Journal of Emerging Technologies in Learning (iJET)*, *17*(5), 147-179.

Shafiq, D. A., Marjani, M., Habeeb, R. A. A., & Asirvatham, D. (2022). Student retention using educational data mining and predictive analytics: A systematic literature review. *IEEE Access*, *10*, 72480 - 72503. https://doi.org/10.1109/ACCESS.2022.3188767

Sokkhey, P., Navy, S., Tong, L., & Okazaki, T. (2020). Multi-models of educational data mining for predicting student performance in mathematics: A case study on high schools in Cambodia. *IEIE Transactions on Smart Processing and Computing*, *9*(3), 217-229.

Tomasevic, N., Gvozdenovic, N., & Vranes, S. (2020). An overview and comparison of supervised data mining techniques for student exam performance prediction. *Computers and Education*, *143*, 103676. https://doi.org/10.1016/j.compedu.2019.103676

Western OC2 Lab. (2018). *Student-Performance-and-Engagement-Prediction-eLearning-datasets.* https://github.com/Western-OC2-Lab/Student-Performance-and-Engagement-Prediction-eLearning-datasets

Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1), 11. https://doi.org/10.1186/s40561-022-00192-z

Yang, F., & Li, F. W.(2018, August). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. *Computers and Education*, *123*, 97-108.

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for educational data mining models: A case study in affect detection. *British Journal of Educational Technology*, *45*(3), 487-501.