INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Predicting Factors that Affect East Asian Students' Reading Proficiency in PISA

Adeline Hui-Min Low [a], Amy Hui-Lan Lim [a,*], Fang-Fang Chua [a]

[a] *Faculty of Computing and Informatics, Multimedia University, Cyberjaya, 63100, Malaysia*

Corresponding author: [*]amy.lim@mmu.edu.my

***Abstract*— Teachers, schools, and parents contribute to equipping students with essential knowledge and skills during their education years. When students are approaching the end of their education, they are randomly selected to participate in Program for International Student Assessment (PISA) to assess their reading proficiency. Existing work on analyzing PISA achievement results concentrates solely on identifying factors related to Parent or in combination with Student. Limited work has been proposed on how factors related to Teacher and School affect the students' reading proficiency in PISA. This study focuses on identifying the factors related to Teacher and/or School that affect East Asian students' reading proficiency in PISA. The PISA achievement results from East Asian students are chosen as the domain study because they are consistently the top performers in PISA in the past decade. Decision Tree (DT), Naïve Bayes (NB), K-Nearest Neighbors (KNN) and Random Forest (RF) are compared. Hamming score is used as the evaluation metric. The results indicate that RF produces the best predictive models with highest Hamming score of 0.8427. Based on the findings, School-related factors such as the number of school's disciplinary cases, size of the school, the availability of computers with Internet facilities, the quality and educational qualifications of teachers have higher impact on the PISA achievement results. The identified factors can be used as a reference in assessing the current school's teaching, learning environment, and organizing extra activities as part of intervention programs to cultivate reading habits and enhance reading abilities among students.**

***Keywords*— Data mining; PISA; reading domain; teacher.**

Manuscript received 15 Jan. 2023; revised 14 Apr. 2023; accepted 24 Aug. 2023. Date of publication 30 Nov. 2023.

I. INTRODUCTION

Reading is recognizing the words and understanding the semantics of the words in a given text [1]. Books, magazines, and newspapers are examples of reading sources. There are many benefits to being able to read. For students, being able to read can help to improve his/her vocabulary, enhance his/her interaction skills, and enrich his/her knowledge and skills. When a student has a good grasp of knowledge and skills, he/she can generally achieve good academic results in school assessments or evaluations. Many studies have proven that having good reading proficiency contributes to student's achievement in various assessments [2]–[4].

The Program for International Student Assessment (PISA) is an international study that assesses worldwide education systems. PISA assesses how well 15-year-old students can use their Reading, Mathematics, and Science knowledge and skills to tackle real-world problems [5]. Since the year 2000, over 70 countries have participated in PISA. PISA is conducted once every three years, focusing on one domain each time the study is conducted. The reading domain was focused on in the years 2000, 2009, and 2018. With the advent of technology, reading sources have expanded to digital versions such as e-books, e-magazines, and websites. Therefore, the PISA 2018 reading framework has been revised to incorporate additional assessments to assess the new form of reading known as digital reading literacy. This includes assessing students' ability to type the relevant keywords to search for digital reading sources. PISA provides a few questionnaires to solicit feedback from students, teachers, school principals, and parents. Students are given a few questionnaires to obtain feedback on their learning styles, Well-being, Financial Literacy, and ICT familiarity, with the latter three questionnaires being optional. Besides, students also answer the questionnaires related to three domains: assessment questions related to Reading, assessment questions related to Science, and assessment questions related to Mathematics. There are two sets of questionnaires for teachers: the General Teacher questionnaire and the Test Language Teacher questionnaire. Teachers are given the

flexibility to decide as to whether they want to attempt the questionnaires. Next, school principals are required to answer the school questionnaire, which covers school's management and the learning environment. Finally, a set of Parent questionnaires is distributed to parents of the students who are participating in the PISA study. This questionnaire will mainly focus on gathering feedback from parents regarding the extent of their involvement in their children's studies.

The compilation of responses from PISA studies from the year 2000 to 2018 can be downloaded from the Organization for Economic Co-operation and Development (OECD) website. Based on the outcome of the past cycles of PISA studies that have been released over the years, generally East Asian students were top achievers in PISA assessments [5]. From the compilation of past cycles of PISA studies, numerous research studies have been done to forecast academic achievement using data mining techniques. However, majority of the research studies have concentrated on determining key factors that affect academic achievement across three domains from aspects such as such as learning time for a subject [6], [7], parents' education [6], and the usage of ICT [8], [9].

A student's academic achievement is not solely based on their hard work and parental support. Teachers also play important roles in molding the students because they spend most of their time in schools during their schooling years [10]. The presence of teachers and parents has formed a triangle of support for students during the learning process in their schooling years. Despite many research studies on a compilation of responses from PISA studies over the years, there are less research studies that predict students' reading proficiency using the compilation of PISA responses from Teacher questionnaires [11], [12], School questionnaires [13]–[15] or combination of both Teacher and School questionnaires [16]. Specifically, to the best of our knowledge, no specific research study analyzes the aspects of Teachers and/or School and how they impacted East Asian students' reading proficiency in PISA. The research from these aspects is crucial since the outcome of this research can aid teachers and school administrators in developing better teaching and learning strategies and set a more conducive learning environment for students. It is believed that better teaching and learning strategies can increase students' engagement and understanding of a particular topic taught in school, while a conducive learning environment ensures that students have the necessary facilities to facilitate the learning process.

This research study aims to achieve the following three objectives. First is to identify factors associated with Teachers and/or Schools that contribute to the academic achievement of East Asian students in the Reading domain. The second is to determine the most appropriate predictive model for predicting East Asian students' proficiency level in Reading domain. Finally, it is to analyze how the factors associated with Teachers and/or School will impact students' reading proficiency. The compilation of responses based on PISA study in 2018 is used in this research study because they are the latest data released publicly by OECD. Specifically, the responses from a total of 6 out of 8 East Asian countries as listed by World Population Review, namely China, Hong Kong, Japan, Macau, South Korea and Taiwan are used in this research study.

This paper is organized as follows. The literature review and research methodology are discussed in Section II. The results of the proposed methodologies are shown in Section III. The conclusion of this research is presented in Section IV, and references are provided in the last part of this paper.

II. Materials and Method

A. Literature Reviews

Many studies use the compilation of responses based on PISA studies over the years to identify factors that affect students' academic achievement. Since the PISA study is participated by 79 countries and the size of the responses is huge, commonly researchers will focus on one or combination of the following data dimensions for analysis such as geographical locations, domains, and responses from different type of participants in the PISA studies.

Using data responses from PISA in the year 2018, the researchers [17] compare a series of machine learning algorithms to determine the relevant algorithm that can accurately predict the reading achievement of Macau students. The machine learning algorithms that are selected for analysis include multiple tree-based ensemble machine learning such as Random Forest (RF), Gradient and Extreme Gradient Boosting, Extra Tree, and TreeBag. The researchers chose RF as their main statistical method since it has outperformed other machine learning algorithms with R2 of 0.43 and Root Mean Square Error (RMSE) of 66.17. Hierarchical Linear Modeling (HLM) is proposed to ascertain each factor's impact as discovered through RF. They have concluded that the most crucial category is personal, consisting of 13 factors that represent student characteristics.

The authors also conclude that there is a need to include more factors in the analysis by considering other questionnaires, not just the student questionnaire. Using micro-level data from PISA study in the year 2018 for China, the researcher [14] sets out to investigate the causes of the gap in academic performance between schools in urban and rural locations in China. As schools in urban and rural locations may differ significantly, the study employed Shapley flow, a graph-based approach, to analyze causal relations between a set of factors and average PISA scores that have been obtained concerning the best and worst schools in urban and rural areas. The researcher has also used XGBoost and Linear Regression to determine the causal structure. The Shapley values are hypothesized to show how certain factors affect how well the schools perform academically. The researcher has commented that intermediate learning outcomes and student characteristics affect the academic performance of schools in urban areas. For schools in rural areas, the characteristics of schools affect the school's academic performance. The researcher has highlighted that although the researcher is using the latest data from the PISA study, the data is limited to responses from Beijing, Shanghai, Jiangsu and Zhejiang, which do not contain responses from participants in other provinces in China.

In a research study, the researchers [18] have sought to identify the factors that affect high and low performers in Global Competence test that is administered to students in

Hong Kong, China. The publicly available responses from PISA study in the year 2018 is used in the study. Support Vector Machine (SVM) is proposed in this study and compared with SVM-based recursive feature elimination Cross Validation (SVM-RFE-CV). Generally, the researchers have reported that SVM is a good classifier with performance metrics comprising of accuracy (ACC), F-score and Area under curve (AUC) of more than 0.80. SVM-RFE-CV has identified 30 optimal factors. This study has discovered that students' global competency is affected by their perspective-taking capacity, adaptability, awareness of intercultural communication and respect for people from different cultures. The researchers [13] have conducted research to find the factors that genuinely affect Singaporean students' reading proficiency using the responses from PISA study in the year 2015. SVM is proposed, and accuracy is used to measure the effectiveness of SVM. SVM-based recursive feature elimination (SVM-RFE) is used to identify and rank the factors. The outcome reveals that the SVM model produces an ACC of 0.78 and the most important factor is the learning time that students spent on test language (LMINS).

The researchers [9] use the Activity Region Finder (ARF) algorithm, to uncover factors contributing to high achievement in the Reading domain for students from Turkey and China. Two sets of data from the PISA study are created by separating students' responses from Turkey and China. Feature selection using RF is applied on both sets of data with the total factors reduced to 20. RF produces an ACC in the form of percentage which is 75% when applied to Turkish dataset and in the Chinese dataset, RF produces an accuracy in the form of percentage which is 77%. The ability to comprehend text is the main factor determining the high achievement in the Reading domain for students from Turkey and China.

There is also research work conducted using the entire PISA dataset. The researchers [19] compare the efficiency of RF and XGBoost in predicting the self-efficacy of students from 74 different nations. The researchers have reported that XGBoost is a slightly better predictive algorithm and students' non-cognitive factors are the most important factors. XGBoost is reported to have RMSE of 9.776, R2 of 0.458 and Mean Absolute Error (MAE) of 7.271, which are lower than RF when these algorithms are applied to test data. Other researchers [20] propose an educational data mining approach consisting of a combination of clustering and classification techniques to detect and analyze factors related to country, school, and student that might affect students' academic performance. The researchers group the schools according to their average performance levels in Science, Mathematics and Reading domains using k-means clustering. Three groups have been established: low performance, high performance, and medium performance. Socioeconomic country indicators such as Gross Domestic Product (GDP), GDP adjusted by Purchasing Power Parity (PPP), and GDP per capita are added to existing data for further analysis. C4.5 algorithm is used to build the decision tree, and a confusion matrix is reported. The results reveal that socioeconomic factors are important in determining students' academic performance. The researchers [21] have used RF to determine significant attributes that contribute to the reading proficiency of Filipino students in PISA. RF has identified 53 attributes of which 26 of them are selected based on the mapping with Bronfenbrenner's bioecological framework. These 26 attributes will serve as input into the HLM model. The result shows that 26 attributes do contribute to students' reading proficiency.

There are other studies on predicting the risk of dropouts using log datasets [22]–[26]. In work by [22], weighted attributes are introduced prior to SVM Classifier, resulting in better performance than non-weighted attributes. In work by [23], DT, RF, SVM and Deep Neural Network (DNN) are compared with RF, outperforming the others. FWTS-CNN [24] combines a weighted features approach with a time series and convolutional neural network (CNN). It outperforms CNN when it is applied to KDD Cup 2015 dataset. DeepFM [25] is DNN and factorization machine hybrid, achieving 99% in validation data. In work by [26] multiple linear regression (MLR), multilayer perceptron (MLP) and classification and regression tree (CART) are compared with MLP and CART performing better than MLR. Other than predicting the risk of dropout, there are studies on using machine learning to predict graduation [27], [28][29] has created a mobile application with deep learning to enhance English and Arabic vocabulary among children. The mobile application has recorded more than 90% accuracy for image classification.

In summary, many studies have been conducted in order to identify the factors that influence students' performance in the PISA assessments. However, only a few research studies focus on examining how the factors related to Teacher and School can influence students' performance. To our knowledge, no research analyzes the impact of factors related to Teacher and/or School to East Asian students' reading proficiency levels. Hence, in this paper, we would like to investigate how factors related Teacher and/or School are associated with East Asian students' academic achievement in Reading domain using the compilation of responses from PISA study in the year 2018.

B. Research Methodology

Here, we describe the proposed methodology with its graphical illustration as shown in Fig.1. Each step will be explained as follows.

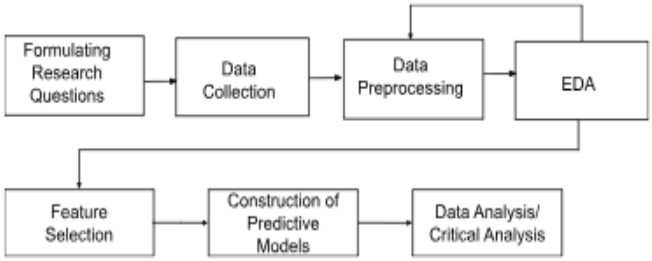

Fig. 1 Proposed Methodology

1) Formulating Research Questions: Based on the objectives as described in the second last paragraph of Section I, the following research questions have been formulated.

- What is the supervised learning technique that is reliable for predicting East Asian students' reading proficiency in PISA?
- Among the factors associated with Teacher and/or School, which factors are more important in

determining East Asian students' reading proficiency in PISA?

- How are the important factors associated with Teacher and/or School have an impact to the East Asian students' reading proficiency in PISA?

2) Data Collection: As mentioned in the second last paragraph in Section I, the compilation of responses based on PISA study in the year 2018 is used in this research study because they are the latest data that is released publicly by OECD. The responses from the student questionnaire, School questionnaire and Teacher questionnaires are chosen to be part of the dataset that will be used in this study.

3) Data Pre-processing: The data pre-processing is applied for the first time on the original datasets that are downloaded from the OECD website so that they can be suitably used in exploratory data analysis to gain preliminary understanding on the characteristics of the datasets. Based on the preliminary insights, the data pre-processing is later re-applied for the second time to merge and prepare several sets of datasets as input for more complex exploratory data analysis and machine learning algorithms. The following three paragraphs describe the data pre-processing steps that are conducted on the first round.

Students' responses from East Asian countries are extracted from the original data of students' responses in the PISA study. The extracted data consists of 41871 rows. In PISA 2018 data, each domain has its own PISA band definition where levels 2 and below are classified as low performers while level 5 and 6 are classified as top performers. OECD does not explicitly define the remaining levels 3 and 4, but students who obtain these levels are assumed to be medium performers. Since machine learning algorithms can only accept numeric as nominal values, the value '0' represents low performers, while the value '1' represents medium performers and the value '2' represents top performers. On the whole, a total of 4 attributes are selected. The attribute CNT is used to filter rows of data from East Asian countries only. Furthermore, the attributes CNT and CNTSCHID allow us to perform data merging. Since our aim is to predict reading proficiency, the attribute Proficiency Level Read_Mean is considered. Table I shows the attributes that represent the columns after pre-processing the student questionnaire dataset.

TABLE I
LIST OF ATTRIBUTES IN THE STUDENT QUESTIONNAIRE DATASET

Attribute(s)	Description
CNT	Country
CNTSCHID	The unique ID that represents a school in a country
CNTSTUID	The unique ID that represents a student from a country
Proficiency Level Read_Mean	Performance level in Read

There are two questionnaires: the General Teacher and the Test Language Teacher. The General Teacher questionnaire is to be answered by teachers who teach Science and mathematics, while the Test Language Teacher questionnaire is to be answered by language teachers. Responses from both questionnaires are selected to form the responses from teachers. Although the responses from Test Language Teacher Questionnaire seem to be more relevant due to the association of reading with language, the PISA reading framework in 2018 has been revised to focus not only on assessing the traditional forms of Reading but also on assessing students' digital literacy skill in searching for digital reading sources. This refers to technical skills. Hence, the responses from both questionnaires are selected. The responses from teachers who work in East Asian countries are selected. Although there are seven East Asian countries, only teachers from Chinese Taipei, Macao, Hong Kong, and Korea have answered the Teacher questionnaire. Hence, only teachers' responses from four East Asian countries are selected. Firstly, the rows that capture teachers' IDs without any responses will be deleted. Such a situation happens because the teacher questionnaire is optional for teachers, and teachers can opt not to answer the questionnaire. A total of 26 relevant attributes related to Teacher are selected. Additional cleaning activity is done where rows that have more than 17 missing values are dropped. The numerical columns that have missing values will be set to 0. Eventually, the data is reduced to 14,105 rows. Table II shows the list of attributes that represent columns in the Teacher questionnaire dataset and the rationale for selecting these attributes to form the Teacher questionnaire dataset.

TABLE II
LIST OF ATTRIBUTES IN THE TEACHER QUESTIONNAIRE DATASET

Attribute(s)	Description / Rationale
CNT	Country
CNTSCHID	The unique ID that represents a school in a country
CNTTCHID	The unique ID that represents a teacher from a country
TEACHERID	To analyze whether the subject taught (General, Reading) could affect students' performance in specific domains.
STTMG1, STTMG2, STTMG3	To understand whether the subject taught overlaps with the initial education and affect student proficiency level in the subject
NTEACH1, NTEACH2, NTEACH3	To analyze whether the subjects that the teacher taught could improve students' assessment.
EXCHT, COLT	To understand whether the exchange and co-ordination of teaching practices could help students improve their results.
SATJOB, SATTEACH	To determine how teachers' attitudes towards their current job environment help in students' assessment.
SEFFCM	To determine how teacher, control the classes environment
SEFFREL	To determine how teacher, maintain the positive relations with students help students to achieve good results.
SEFFINS	To determine whether teachers provide clear instructions to students.
TCOTLCOMP	To determine whether teachers use computers in teaching
TCSTIMREAD, TCSTRATREAD	To determine whether the strategy used by language teachers could bring impact to students' proficiency in Reading
TCICTUSE	To determine whether technology could help students in their proficiency.

Attribute(s)	Description / Rationale
TCDISCLIMA	To determine how language teachers, manage the students' disciplinary.
TCDIRINS	To determine how teachers, provide instruction
FEEDBACK, FEEDBINSTR	To determine whether the practice of giving feedback to students could bring impact to students' proficiency.
ADAPTINSTR	To determine whether students could follow teachers' instructions.

Since our focus is on East Asian countries, the responses from school principals in East Asian countries are selected. A total of 16 relevant attributes related to School are selected. Rows with at least 13 missing values are dropped. Furthermore, there are few rows with missing values for attributes SCHLTYPE and CLSIZE which are also deleted. The numerical columns that have missing values will be set to 0. Eventually, the data is reduced to 1069 rows. Table III shows the list of attributes that represent columns in the school questionnaire dataset and the rationale of selecting these attributes.

TABLE III
LIST OF ATTRIBUTES IN THE SCHOOL QUESTIONNAIRE DATASET

Attribute(s)	Description / Rationale
CNT	Country
CNTSCHID	The unique ID that represents a school in a country
SCHLTYPE	To identify whether school type (private or public) could affect student performance.
SCHSIZE	To determine whether the size of the school have an impact on students' performance.
CLSIZE	To determine whether the class size have an impact on students' performance.
RATCMP1, RATCMP2	To determine whether the quantity of computers with internet access may influence students' performance.
PROATACE	To determine whether the proportion of fully certified teachers could result in excellent student performance.
PROAT5AB	To determine whether the proportion of teachers with an ISCED 5A bachelor qualification could result in excellent student performance.
PROAT5AM	To determine whether the proportion of teachers with an ISCED 5A master qualification could result in excellent student performance.
PROAT6	To determine whether the proportion of teachers with an ISCED level 6 qualification could result in excellent student performance.
CREACTIV	To understand how creative extra-curricular activities at school could assist students in achieving excellent results.
STAFFSHORT, EDUSHORT	To understand how staff and education materials shortage could affect students' academic performance.
STUBEHA, TEACHBEHA	To investigate how school climate and teacher's behavior could affect overall students' performance.

This paragraph describes the data pre-processing that is conducted on the second round. Since we need to identify whether the attributes related to Teacher and/or School are significantly influencing the students' proficiency level, a total of three datasets are created from this step as follows.

- A dataset having Teacher-related attributes/factors with student proficiency level. This dataset is formed by merging the processed Student questionnaire dataset with the processed Teacher questionnaire dataset.
- A dataset having School-related attributes/factors with student proficiency level. This dataset is formed by merging the processed Student questionnaire dataset with processed School questionnaire dataset.
- A dataset having Teacher-related and School-related attributes/factors with student proficiency level. This dataset is formed by merging the processed Student questionnaire dataset with the processed School and Teacher questionnaire datasets.

The key attributes to be used when merging the datasets are CNTSCHID and CNT. Further data pre-processing will also be conducted where key attributes and attributes that uniquely identifies the student and teacher are removed. Table IV shows the dimensions of the three datasets after pre-processing.

TABLE IV
DIMENSION OF THE THREE DATASETS AFTER PRE-PROCESSING

Shape / Data	Teacher-related Factors	School-related Factors	Teacher-related and School-related Factors
No. of rows	739975	39814	699004
No. of columns	26	17	40

4) Exploratory Data Analysis: Similar to data pre-processing step, the Exploratory Data Analysis (EDA) is conducted twice. The first round is to obtain a preliminary understanding of the characteristics of the datasets which serve as guide to further pre-process the datasets. The second round will provide more complex EDA after the data merging process. The results from both rounds of EDA are presented in Section III.

5) Feature Selection: In this step, the most relevant features from the Teacher and School variables are chosen using a feature selection method. This study employs recursive feature elimination cross-validation (RFE-CV) to determine the most important features (attributes) affecting students' reading proficiency level. RFE-CV chooses the best features by eliminating features of low importance using recursive feature elimination and then picking the best subset based on the model's cross-validation score. RFE-CV is chosen as the feature selection algorithm over recursive feature elimination (RFE) because RFE requires a user to specify the total number of features to be retained. RFE-CV will show the features that have high importance by fitting the model multiple times and at each step, removing the weakest features according to the importance of the features. The outcome of this step helps to determine the important attributes that should be selected to form the input dataset to our predictive algorithms.

6) Construction of Predictive Models: Different supervised data mining algorithms have been selected to compare the accuracy of predicting students' reading proficiency level. The algorithms such as Random Forest (RF), Decision Tree (DT), Naïve Bayes (NB) and K-Nearest

Neighbors (KNN) are used to predict and compare the accuracy of students' reading proficiency level. These algorithms are chosen because they are used by most of the research to predict students' academic performance based on Student-related factors in the PISA dataset. Hence, in this study, we would like to find the most suitable algorithm that can produce high accuracy in predicting reading proficiency level using Teacher-related and/or School-related factors. Since this study is predicting multiclass labels, classification evaluation matrices such as Accuracy, Precision, Recall, F1-score are not suitable be used to evaluate and compare the performance of each predictive model that is generated. In this research, Hamming score [30] is used to evaluate and compare the performance of each supervised machine learning algorithms. Hamming score is one of the metrics used to evaluate the performance of any classification algorithm by calculating the percentage of its correct predictions. Hamming score and accuracy are interchangeable for binary and multiclass cases, but the Accuracy score is calculated using the number of True Positives, True Negatives, False Positives and False Negatives whereas the Hamming score is calculated using the number of correct predictions. Hamming score differs from Hamming loss. To calculate Hamming score, a multi-class problem confusion matrix must be built first. Then the total number of correctly predicted classes will be divided by the total number of samples used to build the predictive model. A Hamming score of more than 0.7 is regarded as a good score. The study's outcome will be further discussed as the outcome contributes as answer to the third research question.

III. Results and Discussion

A. *Results from EDA*

Here we describe the outcome from the first and second round of EDA. For each graphical representation, a short illustration is provided. Fig. 2 shows the total number of students that have participated in the PISA assessments. Beijing-Shanghai-Jiangsu-Zhejiang (B-S-J-Z) have the highest number of students participating in the assessment, while Macao has the lowest. This is because China has the largest population compared to other East Asian countries. Besides, China provides free education to students for both primary and secondary schools. As a result, many parents will send their children to school. Hence, there will be more students participating in PISA assessments.

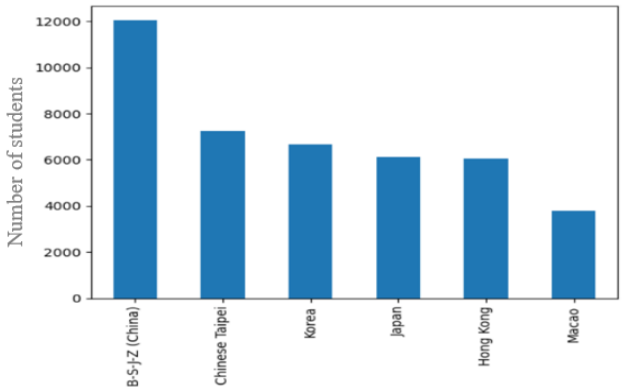

Fig. 2 Total number of East Asian students who participated in PISA 2018

Fig.3 shows the proficiency level of reading for each country. B-S-J-Z has the highest number of students who obtained medium and high proficiency levels in Reading. It can be seen that most of the students only achieve a medium proficiency level in Reading. It is possible that the difficulty of reading instruction is why most students gain reading competency at the medium level. To comprehend the language and grammar of a document, Reading requires effort. For students to grasp a language, they must put in more time and practice.

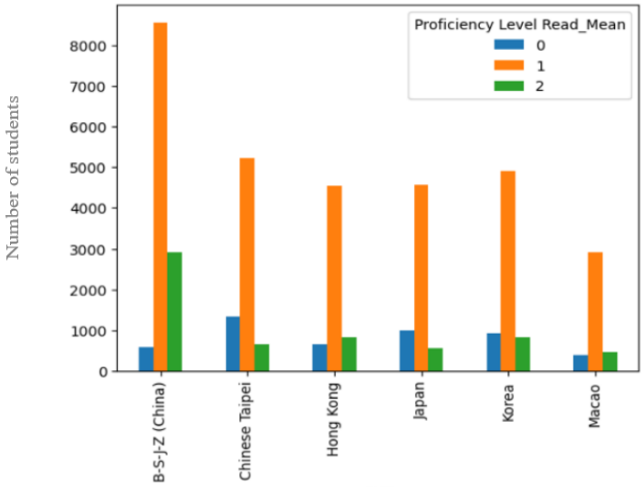

Fig. 3 Reading proficiency level of each country

Fig. 4 shows the total number of teachers in each subject. It shows that most teachers teach general subjects, including mathematics and Science, and that only about 5223 teachers teach Reading or language-specific subjects. The reason as to why the amount of language teachers are lesser than general teachers maybe because teaching language is much more complicated than general subjects. Teachers who teach languages need to understand the grammatical structure of the language. As a result, most teachers prefer to teach subjects unrelated to languages because they are easier to teach, and students are more engaged in learning general subjects.

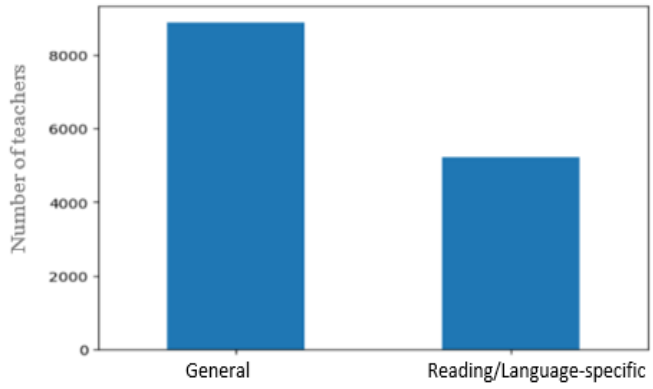

Fig. 4 Total number of teachers in each category of subjects

The correlation between factors related to language teachers is shown in the heat map in Fig.5. There is high positive correlation between FEEDBINSTR and ADAPTINSTR. Teachers who have high adaptivity of instructions (ADAPTINSTR) will tend to provide feedback to students' work (FEEDBINSTR). TCSTIMREAD also shows

a high correlation between FEEDBINSTR and ADAPTINSTR. Teachers who always try to let students engage in reading (TCSTIMREAD) will likely provide and receive feedback and always change the structure of the lessons if a student has difficulty in a lesson.

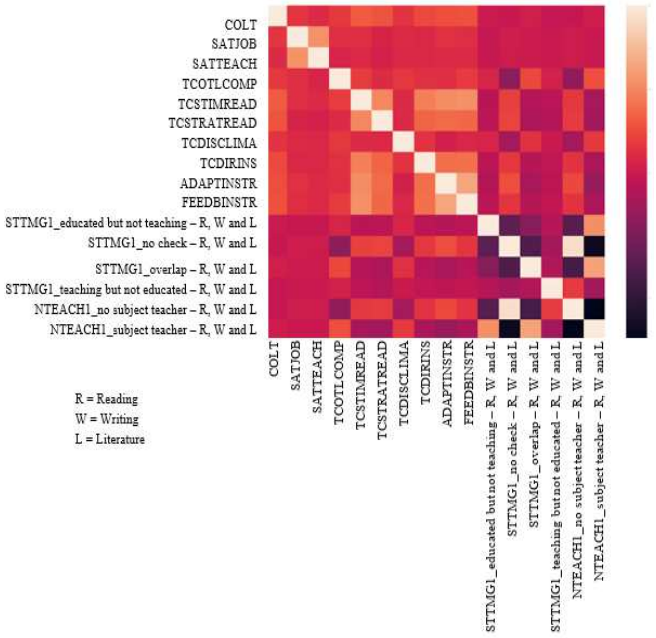

Fig. 5 Correlation heat map of factors in language teacher's questionnaire

Fig. 6 shows the correlation between factors related to general teacher variables. The self-efficacy variables have a high correlation with each other. Teachers who practice self-efficacy in classroom management (SEFFCM) will most likely have self-efficacy in instruction setting (SEFFINS). Besides, teachers who have self-efficacy in maintaining positive relations with students (SEFFCM) are able to manage the classes and students are willing to listen to instructions.

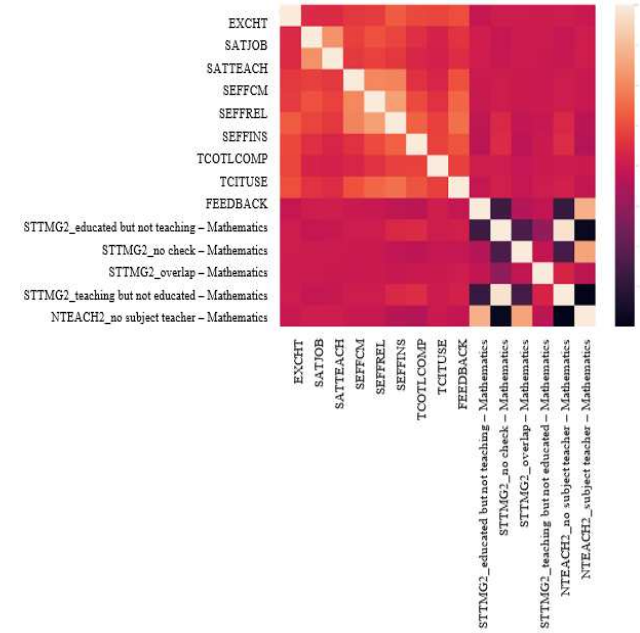

Fig. 6 Correlation heat map of factors in general teacher's questionnaire

As reflected in Fig.7, China has the highest number of public schools and Hong Kong has the highest number of private, government-dependent schools. This diagram shows that most of the schools that participated in this assessment are public schools.

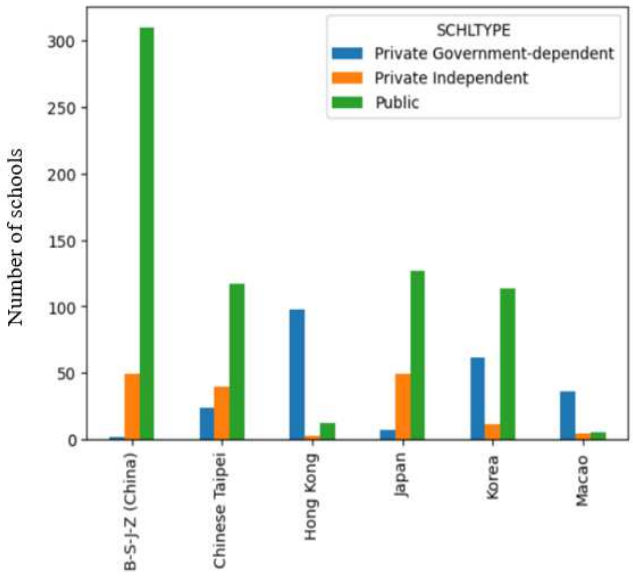

Fig. 7 Total number of schools that participated in PISA assessment

In Fig. 8, the majority of schools have 36 to 40 students in each class. There may not be enough classes offered at the institution or there may be too many students enrolled.

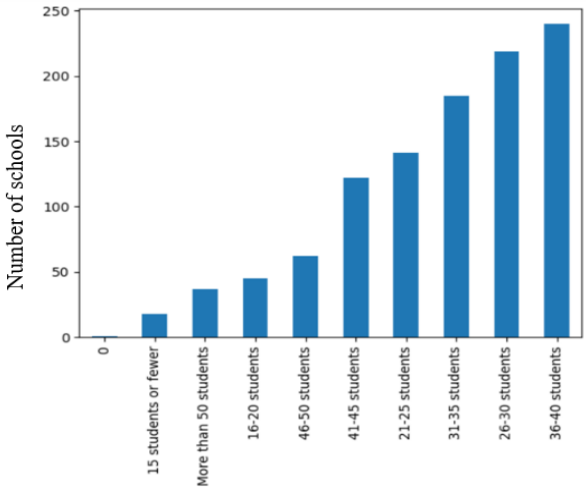

Fig. 8 Total number of schools arranged by class size

Fig. 9 shows a high correlation between EXCHT and students' reading proficiency level. Teachers who always exchange teaching materials with colleagues will likely to be able to improve their teaching methods. Teachers will try different methods to make teaching interesting and engage students more in class. Besides, teachers who are satisfied with their job environment (SATJOB) and teaching profession (SATTEACH) will also be able to improve students' language proficiency.

	Proficiency Level Read_Mean
Proficiency Level Read_Mean	1.000000
EXCHT	0.051496
TCDISCLIMA	0.045446
SATJOB	0.041584
TCSTRATREAD	0.034020
SATTEACH	0.024476
NTEACH1_subject teacher - Reading, writing and literature	0.024353
TEACHERID_Reading/Language-specific	0.021336
STTMG1_educated but not teaching - Reading, writing and literature	0.017058
STTMG1_overlap - Reading, writing and literature	0.011764
TCSTIMREAD	0.007891
TCOTLCOMP	0.006010
TCDIRINS	-0.003492
FEEDBINSTR	-0.005942
ADAPTINSTR	-0.007666
STTMG1_teaching but not educated - Reading, writing and literature	-0.013045
STTMG1_no check - Reading, writing and literature	-0.019389
TEACHERID_General	-0.021336
NTEACH1_no subject teacher - Reading, writing and literature	-0.024353

Fig. 9 Correlation values for Teacher-related attributes towards students' reading proficiency

From Fig.10, the highest correlation value is recorded between students reading proficiency level and creative extra-curricular activities (CREATIV) factor, which is 0.16. The second highest variable is PROAT5AB. Teachers who have ISCED 5A bachelor degrees are most likely able to provide better teaching methods and can teach students better. Next, the lowest correlation value is -0.126322, which is student behavior (STUBEHA). It shows that schools that have less disciplinary problems can lead to produce students that have higher proficiency level in Reading.

	Proficiency Level Read_Mean
STUBEHA	-0.126322
EDUSHORT	-0.095418
RATCMP1	-0.091663
STAFFSHORT	-0.083754
SCHLTYPE_Private Government-dependent	-0.076299
CLSIZE_21-25 students	-0.073965
CLSIZE_16-20 students	-0.066998
TEACHBEHA	-0.054060
CLSIZE_26-30 students	-0.051721
CLSIZE_15 students or fewer	-0.015363
CLSIZE_31-35 students	-0.006977
PROAT6	-0.003770
SCHLTYPE_Private Independent	-0.000752
PROAT5AM	0.009777
CLSIZE_More than 50 students	0.023133
RATCMP2	0.034753
CLSIZE_46-50 students	0.048463
PROATCE	0.059298
CLSIZE_36-40 students	0.065153
CLSIZE_41-45 students	0.068253
SCHLTYPE_Public	0.069277
SCHSIZE	0.085123
PROAT5AB	0.109237
CREACTIV	0.160004
Proficiency Level Read_Mean	1.000000

Fig. 10 Correlation values for School-related attributes towards students' reading proficiency

From Fig. 11, the highest correlation value is CREACTIV, which is 0.13. Schools that provide extra-curriculum can help to increase students' interest in Reading. The lowest correlation value is -0.187398, which is STUBEHA. A safe and positive learning environment can be built if the school has minimal or absence in disciplinary cases. This will be able to improve students' attention and reduce anxiety.

	Proficiency Level Read_Mean
Proficiency Level Read_Mean	1.000000
CREACTIV	0.135928
CLSIZE_36-40 students	0.075972
SCHLTYPE_Private Independent	0.063558
TCDISCLIMA	0.045066
SATJOB	0.040660
TCSTRATREAD	0.033018
SCHSIZE	0.031053
CLSIZE_31-35 students	0.030559
RATCMP2	0.026861
CLSIZE_26-30 students	-0.035024
PROAT5AM	-0.035918
RATCMP1	-0.035929
CLSIZE_16-20 students	-0.035940
EDUSHORT	-0.037541
PROATCE	-0.047133
SCHLTYPE_Public	-0.064655
STAFFSHORT	-0.104596
TEACHBEHA	-0.128329
STUBEHA	-0.187398

Fig. 11 Correlation values for Teacher and School-related attributes towards students' reading proficiency

B. Output from Construction of Predictive Models

The predictive models are built using the three datasets that have been obtained from the second round of data pre-processing. Table V, VI and VII show the Hamming scores of each predictive model that is generated using the three datasets.

TABLE V
HAMMING SCORE FOR EACH PREDICTIVE MODEL THAT IS GENERATED USING TEACHER-RELATED ATTRIBUTES

	Reading			
Algorithm	RF	DT	NB	KNN
Hamming Score	0.8110	0.8093	0.5716	0.7791

TABLE VI
HAMMING SCORE FOR EACH PREDICTIVE MODEL THAT IS GENERATED USING SCHOOL-RELATED ATTRIBUTES

	Reading			
Algorithm	RF	DT	NB	KNN
Hamming Score	0.8427	0.8420	0.6533	0.8155

TABLE VII
HAMMING SCORE FOR EACH PREDICTIVE MODEL THAT IS GENERATED USING TEACHER AND SCHOOL-RELATED ATTRIBUTES

	Reading			
Algorithm	RF	DT	NB	KNN
Hamming Score	0.8201	0.8178	0.6583	0.7843

The tables above show the outcomes of each model's Hamming scores. School-related attributes have the greatest influence on students' reading proficiency levels. Since the input based on School-related attributes generally produces high Hamming scores for the above four predictive algorithms, we will only show the list of important attributes related to School. Additional experiments are conducted to determine the Hamming scores of the predictive algorithms by providing these predictive algorithms with the input data without employing feature selection. The results reveal a minor difference ranging from 0.001 to 0.0001 with slightly better Hamming scores been reported when we use the input data that has undergone feature selection.

C. *Output from Feature Selection*

Fig. 12 shows the list of important School-related attributes based on important scores. The higher the importance score, the more important the attribute is. From the total of 24 School-related attributes, 22 School-related attributes are selected after the feature selection step.

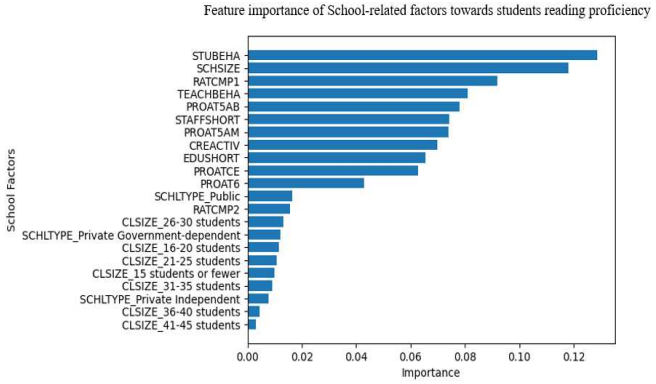

Fig. 12 The list of features sorts according to importance scores

D. *Critical Analysis*

Based on Table V, VI and VII, it is clear that RF performs well in predicting students' reading levels. The Hamming scores for the predictive models are higher than DT, NB, and KNN. From Fig. 12, it is possible to establish that eleven features or factors have significant values greater than 0.04. Significant scores of greater than 0.1 are found in student behavior (STUBEHA) and school size (SCHSIZE). Student behavior has the greatest influence on students' reading proficiency level. Students' misbehavior such as truancy, tendency to skip lessons, lacking respect for teachers, threatening or bullying other students, and lacking attention in class can affect other students' academic achievement due to other students' feeling of insecurity and disharmony in school. A school with a higher disciplinary incidence will create a bad classroom environment. In an environment such as high absenteeism, students who constantly miss a class struggle to learn and may find it difficult to concentrate or catch up with language-related subjects. In reality, the key to mastering reading skills is to concentrate in lessons have constant reading practice on reading materials.

School size affects students' academic achievement because a bigger size school will have better infrastructure and facilities. With better support facilities, teachers can deploy various teaching and learning strategies to make the lessons more entertaining, such as using Kahoot and other online games. A highly populated school can promote better interactivity for students where many discussion activities can be organized. Discussion activities can enhance communication skills as well as motivate the students to read, recognize, and use the words effectively in their conversations. The proportion of computers available in schools (RATCMP1) also influences students' reading competency. Having computers diversifies reading resources and not only confines reading resources to traditional printed materials. Students can use computers to search for e-books to supplement their reading materials and study a wide range of texts, articles, and online books that are relevant to their interests and reading levels. Furthermore, teachers can use computers to support teaching such as organizing quizzes using Kahoot and other online games. In short, schools with adequate computer resources can optimize the potential benefits of technology in education.

Following that, teacher behaviors (TEACHBEHA) may influence students' reading proficiency levels. If teachers are regularly absent, students will miss important lessons. Aside from teachers' behaviors, teachers' qualifications (PROAT5AB, PROAT5AM, PROATCE) may also influence students' achievement. Teachers that are fully certified and possess higher educational qualifications are adept at developing assessments that accurately assess student learning outcomes. They may identify areas of difficulty that students will likely face and provide important comments to help students better understand their weaknesses.

Other significant features or factors that influence students' reading proficiency levels are the number of extracurricular activities available at schools (CREATIV), shortage of staff (STAFFSHORT) and educational materials (EDUSHORT). Book Club and story-telling competitions are examples of extracurricular activities that can be conducted to cultivate reading habits among students. Shortage of manpower and educational material can affect the daily operations of a school and student's learning.

IV. Conclusion

This study focuses on predicting factors that affect East Asian students reading proficiency levels using PISA data. Using Teacher and School datasets, supervised machine learning approaches based on RF, DT, NB, and KNN are used to predict East Asian students' reading proficiency levels. To find the most relevant attributes from Teacher and School dataset, RFE-CV is used as feature selection. Based on the result of this study, School-related factors have the greatest influence on students' achievement. This can be seen from Table V, VI and VII where the results of Hamming scores for School-related factors are greater than the results of Hamming scores for Teacher-related and the combination of Teacher-related and School-related factors.

Furthermore, RF outperforms DT, NB, and KNN regardless of the type of datasets being applied. The school management needs to provide a conducive teaching and learning environment because it provides a sense of security for students to pursue learning easily. With better staff quality, infrastructure and availability of essential learning materials, school can be a great learning ground for students when more engaging lessons can be prepared, and various exciting learning activities can be organized for students.

Nevertheless, not all teachers in East Asian countries provide responses to the teacher's questionnaires because these questionnaires are optional for teachers. The lack of responses from certain East Asian countries can potentially introduce bias into our analysis. In future, other supervised data mining approaches can be utilized to improve the performance of the predictive models.

REFERENCES

[1] R. C. Anderson, "Becoming a nation of readers: The report of the Commission on Reading.," 1985.

[2] A. Talwar *et al.*, "Early Academic Success in College: Examining the Contributions of Reading Literacy Skills, Metacognitive Reading Strategies, and Reading Motivation," *Journal of College Reading and Learning*, vol. 53, no. 1, pp. 58–87, Jan. 2023, doi: 10.1080/10790195.2022.2137069.

[3] K. Nyarko, N. Kugbey, C. C. Kofi, Y. A. Cole, and K. I. Adentwi, "En4glish Reading Proficiency and Academic Performance Among Lower Primary School Children in Ghana," *Sage Open*, vol. 8, no. 3, p. 215824401879701, Apr. 2018, doi: 10.1177/2158244018797019.

[4] L. Stoffelsma and W. Spooren, "The Relationship Between English Reading Proficiency and Academic Achievement of First-Year Science and Mathematics Students in a Multilingual Context," *Int J Sci Math Educ*, vol. 17, no. 5, pp. 905–922, Jun. 2019, doi: 10.1007/s10763-018-9905-z.

[5] Oecd, "PISA 2018 results: Combined executive summaries," *J Chem Inf Model.*, vol. 53, no. 9, pp. 1689–1699, 2019.

[6] N. Aksu, G. Aksu, and S. Saracaloglu, "Prediction of the Factors Affecting PISA Mathematics Literacy of Students from Different Countries by Using Data Mining Methods," *International Electronic Journal of Elementary Education*, vol. 14, no. 5, pp. 613–629, 2022.

[7] A. Bozak and E. C. Aybek, "Comparison of Artificial Neural Networks and Logistic Regression Analysis in PISA Science Literacy Success Prediction.," *International Journal of Contemporary Educational Research*, vol. 7, no. 2, pp. 99–111, 2020.

[8] O. Lezhnina and G. Kismihók, "Combining statistical and machine learning methods to explore German students' attitudes towards ICT in PISA," *International Journal of Research & Method in Education*, vol. 45, no. 2, pp. 180–199, Mar. 2022, doi: 10.1080/1743727X.2021.1963226.

[9] S. Kılıç Depren and Ö. Depren, "Cross-Cultural Comparisons of the Factors Influencing the High Reading Achievement in Turkey and China: Evidence from PISA 2018," *The Asia-Pacific Education Researcher*, vol. 31, no. 4, pp. 427–437, Aug. 2022, doi: 10.1007/s40299-021-00584-8.

[10] C. Nunes, T. Oliveira, M. Castelli, and F. Cruz-Jesus, "Determinants of academic achievement: How parents and teachers influence high school students' performance," *Heliyon*, vol. 9, no. 2, p. e13335, Feb. 2023, doi: 10.1016/j.heliyon.2023.e13335.

[11] S. Li, X. Liu, Y. Yang, and J. Tripp, "Effects of Teacher Professional Development and Science Classroom Learning Environment on Students' Science Achievement," *Res Sci Educ*, vol. 52, no. 4, pp. 1031–1053, Aug. 2022, doi: 10.1007/s11165-020-09979-x.

[12] J. G. Mora-Ruano, M. Schurig, and E. Wittmann, "Instructional Leadership as a Vehicle for Teacher Collaboration and Student Achievement. What the German PISA 2015 Sample Tells Us," *Front Educ (Lausanne)*, vol. 6, p. 582773, Feb. 2021, doi: 10.3389/feduc.2021.582773.

[13] X. Dong and J. Hu, "An Exploration of Impact Factors Influencing Students' Reading Literacy in Singapore with Machine Learning Approaches," *Int J Engl Linguist*, vol. 9, no. 5, p. 52, Aug. 2019, doi: 10.5539/ijel.v9n5p52.

[14] H. Lee, "What drives the performance of Chinese urban and rural secondary schools: A machine learning approach using PISA 2018," *Cities*, vol. 123, p. 103609, Apr. 2022, doi: 10.1016/j.cities.2022.103609.

[15] H. Lee and J.-W. Lee, "Why East Asian students perform better in mathematics than their peers: An investigation using a machine learning approach," 2021.

[16] C. Ding, "Examining the context of better science literacy outcomes among U.S. schools using visual analytics: A machine learning approach," *International Journal of Educational Research Open*, vol. 3, p. 100191, 2022, doi: 10.1016/j.ijedro.2022.100191.

[17] Y. Wang, R. King, J. Haw, and S. on Leung, "What explains Macau students' achievement? An integrative perspective using a machine learning approach (*¿Cuál es la explicación del rendimiento de los estudiantes macaenses? Una perspectiva integradora mediante la adopción del enfoque del aprendizaje automático*)," *Journal for the Study of Education and Development*, vol. 46, no. 1, pp. 71–108, Jan. 2023, doi: 10.1080/02103702.2022.2149120.

[18] T. Luo and Y. Peng, "The analysis of influencing factors on the value dimension of Asian students' global competence - based on PISA 2018," in *2021 16th International Conference on Computer Science & Education (ICCSE)*, IEEE, Aug. 2021, pp. 1130–1134. doi: 10.1109/ICCSE51940.2021.9569461.

[19] B. Tan and M. Cutumisu, "Employing Tree-based Algorithms to Predict Students' Self-Efficacy in PISA 2018," in *Proceedings of the 15th International Conference on Educational Data Mining*, 2022, p. 634.

[20] A. Gamazo and F. Martínez-Abad, "An Exploration of Factors Linked to Academic Performance in PISA 2018 Through Data Mining Techniques," *Front Psychol*, vol. 11, p. 575167, Nov. 2020, doi: 10.3389/fpsyg.2020.575167.

[21] J. Y. Haw and R. B. King, "Understanding Filipino students' achievement in PISA: The roles of personal characteristics, proximal processes, and social contexts," *Social Psychology of Education*, vol. 26, no. 4, pp. 1089–1126, Aug. 2023, doi: 10.1007/s11218-023-09773-3.

[22] Z. Yujiao, L. W. Ang, S. Shaomin, and S. Palaniappan, "Dropout Prediction Model for College Students in MOOCs Based on Weighted Multi-feature and SVM," *Journal of Informatics and Web Engineering*, vol. 2, no. 2, pp. 29–42, 2023, doi: 10.33093/jiwe.2023.2.2.3.

[23] H. S. Park and S. J. Yoo, "Early Dropout Prediction in Online Learning of University using Machine Learning," *JOIV : International Journal on Informatics Visualization*, vol. 5, no. 4, p. 347, Dec. 2021, doi: 10.30630/joiv.5.4.732.

[24] Y. Zheng, Z. Gao, Y. Wang, and Q. Fu, "MOOC Dropout Prediction Using FWTS-CNN Model Based on Fused Feature Weighting and Time Series," *IEEE Access*, vol. 8, pp. 225324–225335, 2020, doi: 10.1109/ACCESS.2020.3045157.

[25] N. M. Alruwais, "Deep FM-Based Predictive Model for Student Dropout in Online Classes," *IEEE Access*, vol. 11, pp. 96954–96970, 2023, doi: 10.1109/ACCESS.2023.3312150.

[26] Y. Tong and Z. Zhan, "An evaluation model based on procedural behaviors for predicting MOOC learning performance: students' online learning behavior analytics and algorithms construction," *Interactive Technology and Smart Education*, vol. 20, no. 3, pp. 291–312, Sep. 2023, doi: 10.1108/ITSE-10-2022-0133.

[27] D. Fahrudy and S. 'Uyun, "Classification of Student Graduation using Naïve Bayes by Comparing between Random Oversampling and Feature Selections of Information Gain and Forward Selection," *JOIV : International Journal on Informatics Visualization*, vol. 6, no. 4, p. 798, Dec. 2022, doi: 10.30630/joiv.6.4.982.

[28] R. Mehdi and M. Nachouki, "A neuro-fuzzy model for predicting and analyzing student graduation performance in computing programs," *Educ Inf Technol (Dordr)*, vol. 28, no. 3, pp. 2455–2484, Mar. 2023, doi: 10.1007/s10639-022-11205-2.

[29] H. Mohd Nasir, N. M. A. Brahin, F. E. Mohd Sani @ Ariffin, M. S. Mispan, and N. H. Abd Wahab, "AI Educational Mobile App using Deep Learning Approach," *JOIV : International Journal on Informatics Visualization*, vol. 7, no. 3, p. 952, Sep. 2023, doi: 10.30630/joiv.7.3.1247.

[30] Hasty.ai, "Hamming score." Accessed: Jun. 01, 2023. [Online]. Available: https://hasty.ai/docs/mp-wiki/metrics/hamming-score