The Behaviour of the Ensemble Learning Model in Analysing Educational Data on COVID-19

Manargul Mukasheva, Ainur Mukhiyadin*, Ulzhan Makhazhanova, and Sandugash Serikbayeva

***Abstract*—This study delves into the emerging opportunities and challenges arising from the integration of education and artificial intelligence in the unique backdrop of the COVID-19 pandemic. Its primary objective is to develop an optimized ensemble model that sheds light on the surge in learning engagement among secondary school students during Emergency Distance Learning (EDL) amid the pandemic. To achieve this, we explored three distinct methodologies: the k-Nearest Neighbor method (KNN), Random Forest (RF), and Gradient Boosting (XGB). Our approach involved constructing an ensemble model that synthesized the strengths and weaknesses of these individual models based on their training outcomes. In contrast to prevailing beliefs that Emergency Distance Learning (EDL) negatively impacts education, our study's findings underscore a positive upswing in students' learning activity during EDL. Furthermore, our ensemble model effectively identifies the underlying reasons behind this increased engagement, achieving an impressive overall accuracy rate of 87% in processing the survey responses. Our research encompassed a comprehensive sample, targeting 35,950 secondary school students from 16 regions and cities of significant importance within Kazakhstan. This diverse sample included students from urban, rural, and small schools, providing a well-rounded perspective on territorial affiliation. Data collection was conducted through an online survey using a methodologically verified structured questionnaire.**

***Index Terms*—COVID data, KNN, Random Forest, XGBoost, emergency distance learning, learning activity**

I. Introduction

The COVID-19 pandemic has disrupted education systems worldwide, forcing schools to switch to emergency distance learning to ensure the continuity of schoolchildren's education. The integration of EDL has posed many problems for teachers, schoolchildren and parents [1]. One of the main problems of EDL integration is the "digital divide", which means unequal access to Internet technologies and infrastructure [2]. Many schoolchildren from disadvantaged families need access to the necessary technologies and the Internet, which prevents them from participating in distance learning and puts them at a disadvantage compared to their peers [3].

Another area for improvement is the difficulty of providing quality education through EDL. Distance learning can be less fun and interactive than face-to-face classes, and it can be difficult for educators to teach and support schoolchildren in a distance learning setting effectively. Moreover, some schoolchildren may find adapting to new teaching methods challenging and may need additional support and resources [4].

"COVID data" refers to data related to the COVID-19 pandemic. This data includes information on the number of confirmed cases, deaths, recoveries, hospitalisations, testing rates, and other statistics related to the spread and impact of COVID-19 in a particular region or country. Various sources provide COVID data, including government agencies, medical organisations, and independent initiatives. Researchers, policymakers, and the public often use this data to monitor the pandemic and make informed public health and safety decisions.

With the emergence of new versions of COVID-19 and ongoing vaccination, the collection and analysis of COVID data have become even more critical in understanding the current state of the pandemic and predicting its future trajectory. There are many reliable sources of data on COVID-19 on the Internet. These include websites of national and local health authorities such as the US Centers for Disease Control and Prevention (CDC), the World Health Organization (WHO), and the European Center for Disease Prevention and Control (ECDC). Other sources of data on COVID-19 include independent initiatives such as the COVID tracking project, which provides daily updates on COVID-19 testing and hospital admissions in the United States [4, 5].

The COVID-19 pandemic has also significantly affected the education system worldwide, including in Kazakhstan. Schools and universities have had to quickly adapt to new teaching methods, such as online and distance learning, to avoid interrupting schoolchildren learning and minimize the risk of virus transmission. In Kazakhstan, the Ministry of Education and Science has taken steps to ensure the continuity of teaching during the pandemic. For example, the Daryn online learning platform was launched to provide schoolchildren and teachers access to online learning resources. The government has also provided Internet access to schoolchildren in remote areas who do not have access to online learning.

However, the pandemic has also highlighted pre-existing inequalities in the education system, as disadvantaged schoolchildren have struggled to access distance learning resources. Moreover, some schoolchildren found adapting to the new learning environment challenging, affecting their learning outcomes. The situation is similar in other countries, where many schoolchildren worldwide experience learning difficulties and underachievement. In response, many countries have implemented strategies such as expanded curriculums, tutoring and additional support for disadvantaged schoolchildren to mitigate the impact of the pandemic on education. Indeed, learning activity is essential

Manuscript received May 22, 2023; revised July 3, 2023; accepted September 12, 2023.

Manargul Mukasheva is with National Academy of Education named after Y. Altynsarin, Astana, Kazakhstan.

Ainur Mukhiyadin, Ulzhan Makhazhanova, and Sandugash Serikbayeva are with Faculty of Information Technology, L.N. Gumilyov, Eurasian National University, Astana, Kazakhstan.

*Correspondence: amukhiyadin@gmail.com (A.M.)

doi: 10.18178/ijiet.2023.13.12.2000

for both schoolchildren and educational institutions.

II. Literature Review

Distance learning is a teaching method; the main characteristic that distinguishes it from other teaching methods is teaching and instructing schoolchildren without the teacher's physical presence in the classroom [6]. In distance learning, schoolchildren often feel isolated, so communication between the schoolchildren and teacher, as well as with other schoolchildren, is an essential parameter for the success of a distance learning program [7, 8].

Schoolchildren engagement is the most important indicator of educational progress in any country. Thus, the teaching activity of schoolchildren depends on gender, age, teaching staff and schoolchildren's learning. Predicting schoolchildren's achievement has generated much interest in education. In other words, schoolchildren's achievement refers to the extent to which schoolchildren achieve both immediate and long-term learning goals [9].

Analysing schoolchildren learning is an integral part of the teaching and learning process and can help identify areas where schoolchildren can improve their learning experience. Using a combination of qualitative and quantitative indicators, better understand how schoolchildren interact with the learning environment and make data-driven decisions to improve the learning experience. Predicting schoolchildren's performance on time allows teachers to identify those with low learning activity and intervene in time to apply appropriate measures.

Using machine learning models to predict and analyse schoolchildren's learning activities at EDC during the pandemic can help teachers identify areas for improvement and enable their schoolchildren to learn more effectively. However, there is a lack of systematic and comprehensive research in this area, so further research is needed to fully appreciate the potential of machine learning models in this context.

Machine learning and big data analytics can provide valuable information [10] on the growth of schoolchildren's engagement in the learning process during e-learning (EDL). Here are some examples of how these tools can be used to show the positive dynamics of this growth:

- Ensemble learning model. An ensemble learning model can combine multiple machine learning algorithms to accurately and reliably predict schoolchildren's learning activities. This model identifies patterns and trends in schoolchildren's performance that cannot be detected with a single algorithm.
- Time series analysis. Time series analysis can be used to track changes in schoolchildren's learning activities over time. By analysing data over multiple points, identify trends in schoolchildren activity that may take time to appear when looking at a single data snapshot.
- Natural language processing. Natural language processing can be used to analyse schoolchildren's feedback and comments about online learning. By analysing these comments, one can understand the factors influencing schoolchildren's engagement and participation in online education.
- Big data analytics. Big data analytics can be used to analyse large datasets and identify patterns and trends in schoolchildren's learning activities. This type of analysis can be used to identify relationships between different schoolchildren activities and areas where improvements can be made to improve learning.

In general, using high-precision ensemble data learning models and various big data analysis techniques, it is possible to show positive growth dynamics of schoolchildren's learning activity during EDL. These tools can provide valuable insights into schoolchildren's behavior and engagement and can also be used to improve the quality of online learning.

The main purpose of this work is to show the positive dynamics of the growth of educational activity of schoolchildren in the period of pre-school education. Using various big data analysis techniques, we used a high-fidelity ensemble data learning model.

The research hypothesis suggests that an ensemble learning model is better than using a single model to analyse COVID data in education. This problem is because combining multiple models can produce more accurate and reliable forecasts by exploiting the strengths of each model and minimizing their weaknesses.

The initial data are the materials of the answers of an online survey conducted by the National Academy of Education named after I. Altynsarin under the Ministry of Education and Science of the Republic of Kazakhstan with the participation of 35,950 schoolchildren of secondary schools. The online survey aimed to identify topical issues of distance learning by studying the opinions and positions of schoolchildren in general education schools, who are key participants in the educational process. The online survey was conducted from 29 April to 6 May 2020, after the completion of 2.5 months of the academic year, which was conducted in the conditions of EDL.

The novelty of this study lies in its application of machine learning models, particularly the high-precision ensemble learning model, to analyze and predict schoolchildren's learning activities during the period of distance learning caused by the COVID-19 pandemic. By using an ensemble learning approach that combines multiple machine learning algorithms, the authors aimed to enhance the accuracy and reliability of predictions, taking advantage of the individual strengths of each algorithm and mitigating their weaknesses. This novel approach provides valuable insights into schoolchildren's behavior and engagement in online education, enabling teachers and educators to identify areas for improvement and optimize the quality of distance learning.

The contributions of the authors in this work are multifaceted. Firstly, they conducted an online survey involving a substantial number of schoolchildren from secondary schools, totaling 35,950 participants. The survey aimed to capture the opinions and positions of schoolchildren, who are essential stakeholders in the educational process, during the period of EDL. This comprehensive data collection process provides a solid foundation for the subsequent analysis and prediction.

In AI-assisted COVID-19 research, exciting work has been presented to understand the genetic factors associated with severe disease, predict hospitalisation and mortality, and

detect disease based on medical data. Asteris *et al*.' presented a new heuristic algorithm for modelling and assessing the risks of the COVID-19 pandemic [11]. This algorithm can help predict and make decisions in the early stages of a pandemic. Another work by Asteris *et al*. proposes a predictive model based on artificial intelligence methods and only five laboratory parameters for early prediction of COVID-19 outcomes. Artificial intelligence can help determine the severity of the disease in patients and take appropriate treatment measures [12].

Mahanty *et al*.' study used transfer learning and fuzzy ensemble models to detect COVID-19 on chest X-rays. Artificial intelligence makes it possible to quickly and accurately detect a disease based on medical images [13].

These studies demonstrate the importance of artificial intelligence in combating the COVID-19 pandemic, detecting and predicting the disease, and developing prevention and treatment strategies. This opens up new vistas for better responses to pandemics and better public health. In recent years, a new field of using machine learning methods for educational purposes, also known as data mining for education, has been gaining momentum. Researchers study data from computing educational institutions to uncover meaningful patterns. S.B. Kocyantis *et al*.' highlights this area and presents a case study of predicting schoolchildren grades using key demographics and written assignments as training data. In addition, a prototype teacher support software tool was developed [14].

Kaddoura highlighted the role of machine learning in survey management during the COVID-19 pandemic. Technological advances have enabled online learning and security exams. The systematic review evaluates 135 studies on the relevance of machine learning throughout the examination cycle, from preparation to assessment. The review covers aspects of authentication, scheduling, proctoring, fraud detection, and protecting at-risk schoolchildren and adaptive learning. Problems and solutions for integrating machine learning into the examination system are discussed [15].

Iman Rahimi briefly analysed COVID-19 trends in Australia, Italy and the UK. Mathematical models such as susceptible, infected, and recovered SIR and susceptible, exposed, infected, quarantined, and recovered SEIQR are offered for epidemiological forecasts, and optimisation algorithms improve model performance. The Prophet algorithm is suitable for data with growing pandemic trends. The study highlights the need for different algorithms for different cases [16].

Nguyen compares decision tree and Bayesian network algorithms for predicting academic performance in two institutions with different schoolchildren populations. Both algorithms achieve the same level of accuracy in predicting schoolchildren's performance in different grade categories. The decision tree consistently outperforms the Bayesian network. Case studies provide valuable information about accurately predicting schoolchildren's achievement and comparing data mining algorithms [17].

Differences in how they work are normal as the schoolchildren dataset is different. The same algorithms can show various performances for different datasets [18–20]. Moreover, each algorithm has some uncertainties depending on the type of data it is applied to, making it difficult to determine a universally acceptable algorithm. Therefore, it is recommended to use ensemble learning models that combine the predictions of different algorithms to surpass the generalizability and reliability of a single learning algorithm and make the predictions more accurate [21].

In general, the application of machine learning in education and epidemiology shows promising results. However, the algorithm's performance depends on the data type, and ensemble learning models are recommended to improve accuracy. In addition, developing new ensemble models for predicting educational activities in distance learning demonstrates progress in this area.

This study proposes a new ensemble model for predicting the level of educational activity of schoolchildren during distance learning. The model was trained on schoolchildren's responses using machine learning algorithms such as decision trees, extreme gradient improvement, and K-nearest neighbour. The proposed model showed a high accuracy of 0.87, indicating its significance compared to the baseline study in the reduced feature set literature. This study represents an essential step in developing methods for predicting the level of educational activity of schoolchildren in distance learning.

III. Materials and Methods

A. Data Collection

Factors that affected the level of educational activity of schoolchildren during the pandemic were selected for study. Empirical methods of observation, survey of respondents and analysis of results were used. The survey was conducted in the form of an online questionnaire of 32 questions. The survey data were considered in 5 contexts.

- Context block;
- Content block;
- Ergonomic block;
- Technical support;
- Psycho-emotional block.

Since the characteristics of survey participants are heterogeneous, with schoolchildren of different ages, social classes and family support, many factors may influence the acceptance of a new type of education, emergency distance learning. A suitable home learning environment with good technological resources, digital skills, and a place of study are important factors that positively influence attitudes towards EDL and ensure better adaptation and progress in distance learning.

35,950 schoolchildren of general education schools in 16 regions and cities of republican significance of the Republic of Kazakhstan took part in the survey. Of these, 17,170 schoolchildren are in urban schools, 18,780 rural schools and Ungraded Schools (UGS) (Table I). UGS are specific to Kazakhstan as they are a particular category of schools often found in remote and less populated areas with low population density.

It is important to emphasize that our sample was carefully selected to reflect a variety of characteristics and educational contexts to best reflect the variety of contexts in which students may face pressing needs for distance learning. Including students from different age groups, social classes

and levels of family support allowed us to analyze a variety of factors, including their influence on attitudes towards EDL and success in distance learning. Moreover, the diversity of educational environments in different regions of Kazakhstan and the inclusion of specific types of schools, such as small schools in remote areas, highlight our research as significant for the development of individual educational strategies and technological support that contribute to improving the accessibility and quality of education for all schoolchildren in different parts republics.

This diversity of educational environments can affect the accessibility and quality of education in different regions. The lack of educational facilities in remote areas can create challenges for schoolchildren and require the development of specific educational strategies and support.

TABLE I: PROFILE OF SURVEY PARTICIPANTS

		Frequency (n)	Percentage (%)
School status	urban	17,170	47.76
	rural	18,080	50.29
	UGS	700	1.95
Schoolchildren status	elementary grades	7962	22.14
	middle classes	20,927	58.21
	senior classes	7061	19.64
Language of instruction	Kazakh	20,622	58.3
	Russian	14,750	41.7

The questionnaire comprises 32 closed questions grouped into five blocks, each highlighting a particular aspect. Table II presents the structure of the questionnaire.

TABLE II: THEMATIC ANALYSIS OF THE STRUCTURE OF THE QUESTIONNAIRE

Subject	Subtopic	Response types
Context block	school status	urban, rural, MKSH
	schoolchildren status	elementary grades, middle grades, high grades
	Language of instruction	Kazakh, Russian
	Region of residence	Astana, Almaty, Shymkent, etc.
	Device Availability	No tablet, laptop, or work area
Content block	Distance learning format	Use of video communication and TV lessons by the teacher. Independent work of schoolchildren on materials, on assignments using various resources
	Successes, skills, and actions in DO	Subject grades, computer skills to prepare for lessons, homework form
	Advantages of DO	Exciting lessons, interactivity, independence, responsibility, individuality, spending less time
	Learning Activity	Causes and factors
Ergonomic block	Convenience TO	Time spent at the computer, completing tasks
	Physical activity	Performing a warm-up, stress on physical health
Technical support	Internet presence	Weaknesses of the Internet signal
	Use of educational resources	Gadgets, textbooks, TV lessons, online communication with the teacher
	Quality of digital content	Using various modules and activities to engage schoolchildren
	Teacher's Presence	Provide timely feedback, virtual work hours
Psycho-emotional block	parental concerns	Help with lessons, connection problems, mastering new material
	Norms and traditions	Lack of social activity at school, quarrels with family members
	adaptation	Difficulty in adapting to do

In this article, we analysed experimental data containing 35,950 records. To process a large amount of information, we applied the methods of descriptive statistics and contingency tables. The Statistical Package for the Social Sciences (SPSS) package was used for statistical data analysis, making conducting a detailed dataset study possible. Using descriptive statistics, we identified the main characteristics of the data, including measures of central tendency, data scatter, skewness, and kurtosis. We created a contingency table to analyse the relationships between different variables, enabling us to examine the data relationships thoroughly.

For a more visualisation and interpretation of the results, we presented the data in tables. This approach gave a more complete understanding of the data structure and patterns. In addition, the results of the survey, in particular the experimental database used to train the developed models, are presented in Excel format on GitHub (https://github.com/Ainur-Mukhiyadin/DB_Schoolchildren_35950.git).

B. Research Methods

Ensemble learning in the analysis of COVID-19 educational data using the KNN, Random Forest and XGBoost algorithms can improve the accuracy and stability of forecasting. Ensemble learning combines the results of different models to obtain a more reliable and robust final prediction.

The process of using ensemble learning with KNN, Random Forest and XGBoost algorithms can be as follows (Fig. 1). It is collecting COVID-19 Educational Data: First, to collect data related to the education and incidence of COVID-19. It may include data on schoolchildren's progress, teaching methods, distance learning conditions, and other indicators that may affect schoolchildren's success in a pandemic. Data pre-processing: To ensure a high-quality analysis, researchers need to pre-process the data, which includes removing outliers, filling in missing values, normalising or standardising data, and applying other processing techniques. This includes removing outliers, filling in missing values, normalising or standardising data, and other processing techniques. Applying KNN, Random Forest, and XGBoost algorithms are then used to train educational data. Each of these algorithms has its unique characteristics and ability to process data. Ensemble formation: After training each of the KNN, Random Forest, and XGBoost algorithms, an ensemble is created that combines the predictions from all three models. Accept final prediction: The ensemble makes the final decision by combining the predictions from each model. This process may be based on voting or other strategies for combining results.

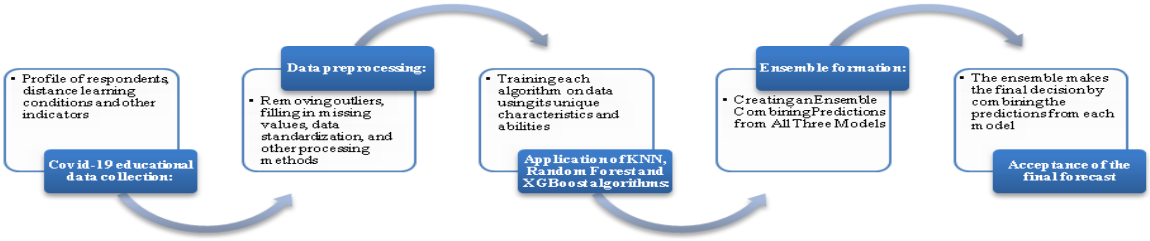

Fig. 1. Application process diagram.

Thus, ensemble learning with KNN, Random Forest and XGBoost algorithms allows us to combine their strengths and make more accurate and reliable predictions on COVID-19 educational data. This approach may be beneficial in the face of uncertainty and volatility associated with a pandemic. The selection of KNN, Random Forest, and XGBoost algorithms for ensemble learning in the analysis of COVID-19 educational data can be well-founded for several reasons:

1) Variety of algorithms: Each has characteristics and abilities in working with data. KNN (k-Nearest Neighbors) is based on the principle of object proximity and performs well in classification and regression tasks. Random Forest (Random Forest) is based on multiple decision trees and can accurately process large amounts of data. XGBoost (Extreme Gradient Boosting) is a robust gradient-boosting algorithm that can process complex data and provide highly accurate predictions.
2) Complementary Algorithms: Each of these algorithms has its strengths and limitations. By combining them into an ensemble, balance their advantages and disadvantages. For example, KNN can better process data with local dependencies, while Random Forest and XGBoost can find more complex and global patterns.
3) Ensemble Stability: Ensembling models can improve the stability and reliability of forecasts. If one of the models gives incorrect predictions due to a data feature or overfitting, the other models can smooth out this error and provide a more accurate result.
4) Minimizing overfitting: Ensemble methods can reduce overfitting, especially when dealing with limited COVID-19 data.

Availability and popularity: KNN, Random Forest and XGBoost are widely used algorithms in machine learning. In some cases, they may already be pre-implemented in Python libraries, making them easier to use and manipulate data.

By combining these algorithms into an ensemble, researchers and analysts can make more accurate and robust predictions when analysing COVID-19 educational data, helping to make more informed decisions during the pandemic and support schoolchildren learning.

The choice of KNN, Random Forest, and XGBoost algorithms in ensemble learning combines their unique characteristics and strengths, resulting in more accurate and robust predictions when analysing educational data in the context of the COVID-19 pandemic. The KNN algorithm assumes that similar data points exist nearby. The model does not train on previous training data but waits until a prediction for a new instance is requested before continuing. There is no predefined form of the mapping function in KNN. The choice of the K value is critical because it plays a crucial role in the classification and avoids overfitting the data.

In addition, we used the KNN algorithm to calculate the distance between the new data and the feature values in the training data. Then we chose K ($K \geq 1$) nearest neighbors for classification or regression. If $K = 1$, the new data will be assigned to the neighboring class. As a result of long-term experiments, it was found that the model works best at $K = 11$.

KNN is based on the principle of object proximity, which allows the algorithm to perform well on classification and regression tasks in the context of COVID-19, where it may be essential to determine whether data belongs to a particular class or category. There can be various categories in COVID-19 educational data, such as schoolchildren achievement (e.g. high, medium, low), making KNN a suitable candidate for classification.

The random forest classifier is one of the ensemble learning methods and supervised learning algorithms that are used for both classification and regression [22, 23]. This classifier generates a group of decision trees based on a randomly selected part of the training sample. For each sample, it generates a decision tree and predicts the results based on it. The more trees there are, the higher the accuracy. Then, to obtain the final result, a combination and an average of all predictor votes obtained from various decision trees are generated [24].

Random Forest was used as a classic ensemble learning algorithm for comparison with XGBoost. Our work used the initial bootstrap method to select a specific sample from the initial training sample randomly. In total, n_tree = 100 samples were selected to create training sets. Each split of each decision tree model was based on obtaining information to select the best feature. Thus, each tree was split until all node training examples belonged to the same class. The final result of the classification was determined by the votes of several classifiers in the tree. There can be many features and dependencies in COVID-19 educational data that Random Forest can handle efficiently, resulting in more reliable and generalisable predictions.

The XGBoost algorithm enhances Gradient Boosted Decision Tree (GBDT) and can be used for classification and regression problems. It should be noted that XGBoost is also one of the boosting tree algorithms that combine many weak classifiers to create a robust classifier [25]. In the context of COVID-19, where there is rapid change and uncertainty, XGBoost can be very useful, as it can work with complex data and find patterns quickly, which will help predict educational outcomes.

IV. Results

For a more in-depth analysis of the survey results and to identify the relationship between the selected questions "Degree of learning activity" and "Reasons for changes in activity", we conducted an additional study. First, a

cross-tabulation (contingency table) was performed between these two questions (Table III). This allowed us to determine the number of schoolchildren related to each combination of the learning activity level and the reasons for the change in activity. For example, how many schoolchildren with high activity will indicate increased parental control.

TABLE III: COMPARISON TABLE OF QUESTIONS "DEGREE OF LEARNING ACTIVITY" AND "REASONS FOR CHANGING LEARNING ACTIVITY"

		Causes of changing learning activity				Total
		strengthening the control of teachers	increased parental control	no fear of the public	Learning just got more fun with a variety of resources	
Degree_of_ learning_activity	high activity	6891	1570	2505	4443	15,409
	average activity	5701	2360	3329	4288	15,678
	low activity	1428	1280	982	1173	4863
Total		14,020	5210	6816	9904	35,950

The links between these two questions were then analysed. Statistical methods, such as the chi-square (χ^2) test (Table IV), were applied to determine if there is a statistically significant relationship between the degree of learning activity and the reasons for the change in activity. Table IV presents the results of all chi-square tests indicate a statistically significant relationship between the selected questions. This confirms our assumption that there is a relationship between the degree of learning activity and the causes of changes in schoolchildren activity. The results obtained will be an essential basis for further study and analysis of the data. They will also allow a more accurate understanding of the relationships and factors influencing the learning activity of schoolchildren.

TABLE IV: CHI-SQUARE TESTS THE STATISTICAL RELATIONSHIP BETWEEN THE QUESTIONS "DEGREE OF LEARNING ACTIVITY" AND "REASONS FOR CHANGING LEARNING ACTIVITY"

	Meaning	St. St.	Asymptotic significance (2-tailed)	Significance of Monte Carlo (2-sided)			Significance of Monte Carlo (1-sided)		
				Significance	99% confidence interval		Significance	99% confidence interval	
					Bottom line	Upper bound		Bottom line	Upper bound
Pearson's chi-square	1079.097[a]	6	0.000	0.000[b]	0.000	0.000			
Likelihood ratios	1017.577	6	0.000	0.000[b]	0.000	0.000			
Fisher's exact test	1017.572			0.000[b]	0.000	0.000			
Line-to-line connection	44.978[c]	1	0.000	0.000[b]	0.000	0.000	0.000[b]	0.000	0.000
Number of valid observations	35,950								

a. A cell count of 0 (0.0%) is assumed to be less than 5. The minimum expected number is 704.76.
b. Based on selecting 10000 tables with a seed value of 2000000.
c. The standardised statistic is 6,707.

The analysis results made it possible to identify whether there is a significant relationship between the level of learning activity and the causes of changes in schoolchildren activity. For example, schoolchildren with high activity are more likely to mention the fun of learning with various resources. In contrast, schoolchildren with low activity tend to indicate increased control from teachers or parents.

These results are essential for further understanding schoolchildren's motivation and factors influencing their learning activity. Understanding the relationship between these two issues can help educational institutions and educators develop more effective learning strategies, considering the needs and interests of schoolchildren at different levels of activity. It can also encourage the development of personalised learning approaches that will help increase schoolchildren's motivation and engagement, leading to improved learning outcomes.

As can be seen from Table V, data on the degree of educational activity of schoolchildren were analysed, divided into three groups: "high activity", "medium activity", and "low activity", and summary statistics were provided for each of them. The total number of observations in the study is:

- "high activity": 15,409 observations;
- "medium activity": 15,678 observations;
- "low activity": 4863 observations.

Each group contains only valid data, which indicates that there are no missing values in this variable. Thus, the study has a sufficiently large sample for each group, which allows statistically significant data analysis and reliable results. The absence of missing values improves the quality of the research since all observations can be fully used to calculate statistical indicators.

These results allow us to be confident in the reliability and validity of the conclusions drawn from this study. The article can mention that the sample has a sufficient amount of data for each group, making the analysis results more representative and interpretable. Also, the absence of missing values contributes to the accuracy and validity of the study, analysis, and conclusions.

TABLE V: SUMMARY REPORT ON OBSERVATIONS

		Observations					
		Valid		Missed		Total	
		N	Interest	N	Interest	N	Interest
Degree of learning activity	high activity	15409	100.0%	0	0.0%	15409	100.0%
	average activity	15678	100.0%	0	0.0%	15678	100.0%
	low activity	4863	100.0%	0	0.0%	4863	100.0%

Table VI includes an analysis of the degree of educational activity of schoolchildren divided into three groups: "high

activity", "medium activity", and "low activity". For each group, descriptive statistics were calculated better to understand the characteristics of the data in each category. In the "high activity" group, the average degree of learning activity is 1.59 (95% confidence interval: 1.58–1.60). The median value is also 2.00, indicating significant outliers in the data as median and mean are significantly different. The skewness is close to 0, indicating a slight deviation from the normal distribution. Also, a negative kurtosis value (−1.119) indicates flatter peaks in the data. In the «moderately active» group, the average value of the degree of learning activity is 1.53 (95% confidence interval: 1.52–1.54). The median value is also 2.00, again indicating outliers. The skewness is slightly positive (0.251), which may also indicate the presence of several outliers. A negative kurtosis value (−1.169) indicates flatter peaks in the data than expected for a normal distribution. In the "low activity" group, the average value of the degree of learning activity is 1.42 (95% confidence interval: 1.41–1.43). This group also has a median value of 1.00, which again indicates the presence of outliers. The asymmetry is positive (0.649), which may indicate the presence of several outliers. A negative kurtosis value (-0.882) also indicates flatter peaks in the data.

TABLE VI: DESCRIPTIVE STATISTICS

				Statistics	standard error
Degree of learning activity	high activity	Average		1.59	0.004
		95% Confidence interval for the mean	Bottom line	1.58	
			Upper bound	1.60	
		The sample mean truncated by 5%		1.58	
		Median		2.00	
		Dispersion		0.284	
		RMS deviation _		0.533	
		Minimum		1	
		Maximum		3	
		Range		2	
		Interquartile range		1	
		Asymmetry		0.048	0.020
		Excess		−1.119	0.039
	average activity	Average		1.53	0.004
		95% Confidence interval for the mean	Bottom line	1.52	
			Upper bound	1.54	
		The sample mean truncated by 5%		1.51	
		Median		2.00	
		Dispersion		0.287	
		RMS deviation _		0.536	
		Minimum		1	
		Maximum		3	
		Range		2	
		Interquartile range		1	
		Asymmetry		0.251	0.020
		Excess		−1.169	0.039
	low activity	Average		1.42	0.008
		95% Confidence interval for the mean	Bottom line	1.41	
			Upper bound	1.43	
		The sample mean truncated by 5%		1.39	
		Median		1.00	
		Dispersion		0.275	
		RMS deviation _		0.524	
		Minimum		1	
		Maximum		3	
		Range		2	
		Interquartile range		1	
		Asymmetry		0.649	0.035
		Excess		−0.882	0.070

Table VII presents the results of normal distribution tests for data on the degree of educational activity of schoolchildren divided into different groups of school types. The table shows the values of statistics and the significance level for two normality tests: Kolmogorov-Smirnov and Shapiro-Wilk.

TABLE VII: CRITERIA FOR NORMAL DISTRIBUTION

		Kolmogorov-Smirnov [a]			Shapiro-Wilk criterion		
		Statistics	St. St.	Significance	Statistics	St. St.	Significance
Degree of learning activity	high activity	0.349	15409	0.000			
	average activity	0.328	15678	0.000			
	low activity	0.384	4863	0.000	0.662	4863	0.000

a. Liljefors Significance Correction

The results of both criteria (Kolmogorov-Smirnov and Shapiro-Wilk) show that the data did not pass the normality test for all three groups of the degree of learning activity. Significance levels are less than 0.05 (usual significance level), which means that the data in each group have a statistically significant deviation from the normal distribution.

It is important to note that failure to meet the normality criteria does not necessarily make using machine learning algorithms impossible. In some cases, especially with a sufficiently large amount of data, algorithms can resist normality violations. However, the presence of deviations from normality should be considered when interpreting the results, and it is possible to apply appropriate data transformation methods if necessary.

The overall analysis indicates the presence of outliers in the data of all three groups, which may affect the interpretation of the results. Data analysis should consider the possible impact of outliers on the accuracy and reliability of the results. Also, a negative kurtosis across all groups may indicate that the data is more flat than expected for a normal distribution.

In general, the results of descriptive statistics provide helpful information about the characteristics of the data in different groups of schoolchildren with varying levels of learning activity and school types. These results can be used to develop and analyse machine learning models and implement additional data pre-processing steps to improve the analysis results.

We train and test machine learning models using a process that involves loading data from an Excel file and pre-processing it, which includes renaming columns and encoding categorical variables. Then, we split the data into features (X), and a target variable (Y), where the features contain all columns except the 'Activity' column, and the target variable includes only this column.

Several machine learning models are applied, including Random Forest (RF), K-Nearest Neighbors (KNN), and XGBoost (XGB). Each model is trained on the training dataset and evaluated on the test dataset to measure its performance and ensure its implementation. The ensembling method, namely the VotingClassifier, is used to increase the generalising ability of models. This method combines predictions from various models (KNN, RF, XGB) and decisions based on voting, improving predictions' overall accuracy and stability. Thus, the described method ensures the correct separation of data into training and test sets and also applies various models and ensembles to achieve the best result when training and testing machine models.

We test the model resulting from the algorithm's training on a dataset containing previously unknown examples to determine whether a learning algorithm is appropriate for the domain and data used. When testing a model, the data source remains the same, but the selections must be new. For this reason, the preferred method is to split the dataset into train and test parts. In cases of insufficient samples, the success of data generalisation may decrease due to incomplete training of the learning algorithm. So the split training/testing method is more suitable for large datasets.

We split the data into two sets: training and testing sets. On the training set, we trained the model, and on the test set, we checked the model's results, evaluating how well this or that model works. Since our dataset is quite large, we split it 90:10. That is, 90% of the data set falls into the training sample and 10% into the test sample (Fig. 2).

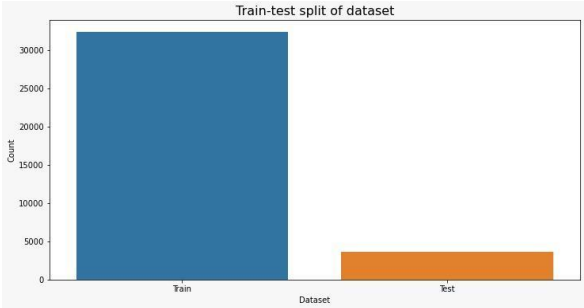

Fig. 2. Train-test split of the dataset.

We used the following hyperparameters to improve the results of the different models. With these parameters, the models perform better:

RF: n_estimators=[5,20,50,100,200,500], max_features=['auto', 'sqrt'] , bootstrap=[True, False]

KNN: n_neighbors = [5,7,9,11,13,15], weights = ['uniform', 'distance'], metric = ['minkowski', 'euclidean', 'manhattan']

XGB: Eta = [0.1, 0.3, 0.5, 0.8, 1], max_depth = [1, 3, 6, 9]

We have selected the most optimal ones among them for different types of models:

RF: n_estimators=100, max_features=auto, bootstrap=True

KNN: n_neighbors=11, metric=euclidean, weights=distance

XGB: max_depth = 3, eta = 0.8

Table VIII presents the test statistics of the learning algorithms used in the study. Accordingly, the most successful models in TN results that show correct predictions of learning activity growth in the test dataset were Ensemble Model (EM) and XGB. In contrast, the most unsuccessful models were RF and KNN.

TABLE VIII: THE TEST STATISTICS OF THE LEARNING ALGORITHMS USED IN THE STUDY

Models	TN	FP	FN	TP
Random Forest (RF)	3098	445	36	16
K-nearest neighbours (KNN)	3090	432	44	29
XGBoost (XGB)	3131	458	3	3
Ensemble Model (EM)	3126	451	8	10

True positives (TP), True negatives (TN), False positives (FP), and False negatives (FN)

Table IX shows the calculated performance in the range [0, 1] based on the prediction results obtained by the models on the test dataset. The most successful models in sensitivity values (TPR) showing the correct measures of schoolchildren learning activity growth in the dataset were XGB and EM.

TABLE IX: PERFORMANCE OF INDIVIDUAL ALGORITHMS AND ENSEMBLE MODELS

Models	Specificity or TNR	Sensitivity or TPR	Precision
RF	0.874262	0.35000	0.030369
KNN	0.877342	0.39726	0.062907
XGB	0.872388	0.50000	0.006508
EM	0.873846	0.50000	0.021692

At the end of the search process, we selected a candidate model formed by the RF, KNN, and XGB sub-models as the final ensemble model. We identified the "hard" voting type and model weights as the best parameters for optimisation in the final ensemble model. Fig. 3 displays the error matrix of

the selected final ensemble model.

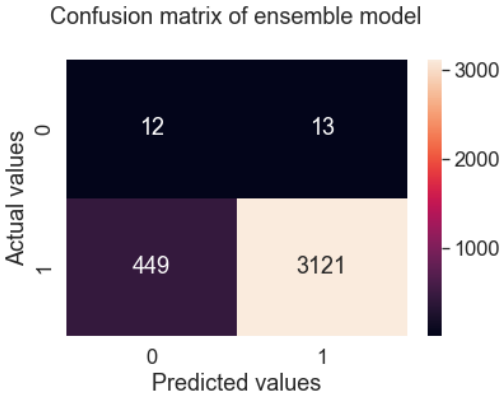

Fig. 3. Confusion matrix of the final ensemble model.

V. Discussion

Fig. 4 shows the cross-validated specificity scores of the individual algorithms and the ensemble model. We visually compared the specificity scores for the RF, KNN, and XGB models and the optimised ensemble model using a box plot with a tenfold CV (Cross-Validation). The bars are medians, defining the average scores obtained at each cross-validation fold.

As can be seen, the median score obtained by the final ensemble model in the far right corner of the graph shows the mean value, which is optimal.

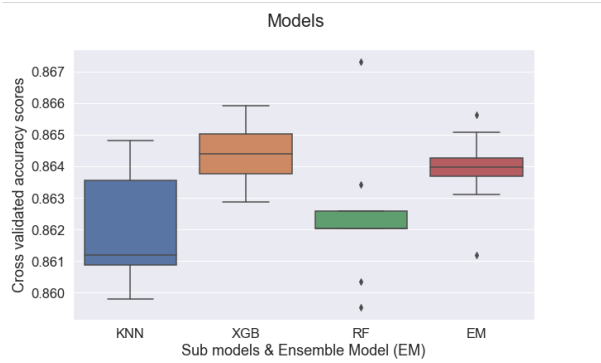

Fig. 4. Comparison of individual models and the ensemble model.

The results show that an ensemble model consisting of combinations of Random Forest (RF) and k-Nearest Neighbour (KNN) classification algorithms and the gradient bootstrap method (XGBoost) is best suited for processing the questionnaire data. The specificity of this model, namely the TPR, is 87%. In previous studies, the best algorithms used to predict schoolchildren performance were Random Forest [26, 27], fuzzy logic [28–30], K-means clustering [31], and naive Bayes analysis. [32], decision tree [33], support vector machine [34], artificial neural network [35], and KNN.

However, this study identified an ensemble model that gives the best prediction result when combining different classification algorithms instead of identifying an individual algorithm that best predicts an increase in schoolchildren's learning activity. For schoolchildren with moderate or low levels of involvement in distance learning, increased adult supervision and the ability not to speak publicly in front of the class were the main factors. For these two categories of schoolchildren, the teacher's checking of each task could have been of better importance. Moreover, such a factor as an increase in interest in learning, since one can study not only from textbooks, is of little importance for the group of schoolchildren who have shown the most significant activity in distance learning [36, 37].

The ensemble learning model provides high accuracy in many areas. Its use as a new approach is rapidly spreading in education. This study concluded that the ensemble learning model approach, which consists of combinations of k-nearest neighbour (KNN), random forest (RF), and XGBoost (XGB) classification algorithms, is best suited for predicting schoolchildren learning activity.

The above work of the authors sheds light on a new area of application of machine learning methods for educational purposes and the role of machine learning in exam management during the COVID-19 pandemic. The research discussed in the review highlights the potential of machine learning in predicting schoolchildren's achievement, managing exams, and solving real-world problems. In addition, the review touches on the importance of ensemble learning models and the need for different algorithms for specific cases.

A notable contribution to this area is the proposal of a new ensemble model for predicting the level of educational activity of schoolchildren in conditions of distance learning. We trained this model using machine learning algorithms, including decision tree, extreme gradient improvement, and K-nearest neighbor. The results of this study show that the proposed ensemble model achieves a high accuracy of 0.87. This level of precision is significant compared to a baseline study with a reduced feature set. Therefore, the developed ensemble model is promising for accurately predicting the level of educational activity in distance learning.

This study protects schoolchildren's engagement and activity during distance learning. Ensemble learning methods allow to combine several algorithms' predictions, increasing the generalisation and reliability of the model. Using diverse algorithms, the ensemble model can effectively capture the various patterns and nuances in educational data.

The implications of this research are valuable for both teachers and schoolchildren. By accurately predicting learning activity levels, tutors and educators can gain insight into individual schoolchildren's engagement and performance. This information can help with early intervention and targeted support for at-risk schoolchildren, improving learning outcomes. In addition, the developed ensemble model can contribute to the optimisation of distance learning processes, providing a more individual and practical educational experience [38].

However, despite the ensemble model's promising results, some issues still need to be addressed. One of the potential limitations lies in the diversity of the schoolchildren population and the specifics of distance learning. Different schools and regions may have different educational environments and infrastructures, affecting the model's generalizability. Future research should focus on testing the model's performance in various educational and demographic contexts to ensure its robustness.

Moreover, relying on schoolchildren's responses can lead to bias or inaccuracies, as self-reported data may only sometimes reflect the actual level of engagement. Combining

multiple data sources and using advanced data cleaning and pre-processing techniques can help mitigate these issues.

The proposed ensemble model for predicting the learning activity of schoolchildren in distance learning demonstrates the potential of combining heterogeneous machine learning algorithms to achieve high accuracy. This study is of practical importance for expanding the educational experience of schoolchildren in distance learning. However, further research and testing on a larger scale are needed to ensure the effectiveness and applicability of the model in different educational contexts.

This work's novelty lies in ensemble learning's use to analyse COVID-19 educational data using various algorithms such as KNN, Random Forest and XGBoost. A feature of the study is the inclusion of questionnaire results as the primary data for analysis, which allows for taking into account various issues related to education and the impact of the pandemic on the educational process and schoolchildren's well-being. Considering data in the context of the COVID-19 pandemic provides relevant information for educational decision-making. Thus, the application of questionnaire-based ensemble learning offers more profound and accurate insights better to understand the pandemic's impact on education and learning.

One of the main limitations of this study is the presence of incomplete data in the original database of online survey responses. In online surveys, the data may need to be completed and contain answers that do not correspond to the direction of the research since schoolchildren in schools may misinterpret questions or provide biased solutions, which may affect the accuracy and validity of the results. Although there are various methods for validating data, eliminating the possibility of errors, inaccuracies, or distortions in respondents' responses in social research has been. It remains an issue that requires even more research in this direction [39].

VI. Conclusions

Other significant limitations in our study were the small amount of data (responses from 35,950 respondents) and their heterogeneity in terms of content. Since insufficient data can limit the possibility of developing more complex models, considering various factors that affect the efficiency and accuracy of forecasts and contextual content heterogeneity makes it difficult to generalise the results and limit the applicability of the developed model in other studies. The problems of forecasting based on the analysis of heterogeneous data can initiate the development of more flexible and valuable machine learning models for studying the results of sociological research.

Using an ensemble learning model to analyse data on COVID in education could be a valuable approach to improve the accuracy and reliability of predictions. Combining several models' strengths and minimising their shortcomings, the ensemble learning model can cover a broader range of factors that affect outcomes, including gender differences, family relationships, social factors, and geographic location, among others.

In practice, we can use an ensemble learning model as an example by combining the predictions of several machine learning models, such as decision trees, random forests, and neural networks, each trained on various aspects of the data. Combining these forecasts enables us to obtain more accurate and reliable predictions by considering a more comprehensive range of factors.

Using ensemble learning models to analyse educational data during the COVID-19 pandemic can provide valuable insights that can aid decision-making, policy development, and public health interventions. However, it is essential to note that the effectiveness of an ensemble learning model depends on the quality and relevance of the data, the choice of appropriate models, and the careful tuning of the ensemble parameters. Close attention to these factors is critical to the success of any ensemble learning model used to analyse COVID data in education.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Author Contributions: Conceptualization, M.M. and A.M.; methodology, U.M., A.M. and S.S.; validation, U.M., M.M., A.M., S.S; analysis, U.M., A.M.; investigation, M.M., A.M., U.M., S.S.; resources, U.M., S.S.; writing – original draft preparation, , M.M., A.M., U.M., S.S.; writing – review and editing, A.M., U.M.; visualization, A.M; su-pervision, U.M. and S.S.; project administration, M.M. All authors have read and agreed to the published version of the manuscript.

References

[1] A. Mukhiyadin, U. Makhazhanova, S. Serikbayeva, A. Kassekeyeva, G. Muratova, S. Karauylbayev *et al*., "Application of information technologies and methods for processing big data to the management of the educational process during the pandemic," *Journal of Theoretical and Applied Information Technology*, vol.101, no. 2, pp. 458–470, 2023.

[2] M. Yessenova, G. Abdikerimova *et al*., "The effectiveness of methods and algorithms for detecting and isolating factors that negatively affect the growth of crops," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 2, pp. 1669–1679, 2023.

[3] S. S. Jadimath and J. Sheetlani, "Effectiveness of ICT Tools in Education Sector with Special Reference to Artificial Intelligence," *Neuroquantology*, 2022, no. 20(10), pp. 5300–5306.

[4] M. Mukasheva, O. Chorosova, Z. Zhilbayev, and Y. Payevskaya, "Integrated approach to the development and implementation of distance courses for school computer science teachers," in *Proc. 2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT)*, Tashkent, Uzbekistan, 2020, pp. 1–6.

[5] J. S. Schwind, D. J. Wolking *et al*., "Evaluation of local media surveillance for improved disease recognition and monitoring in global hotspot regions," *PLoS ONE*, vol. 9, no. 10, e110236, 2014. https://doi.org/10.1371/journal.pone.0110236

[6] Meta. (2020). Infection prevention and control and surveillance for coronavirus disease (COVID-19) in prisons in EU/EEA countries and the UK. [Online]. Available: http://www.ristretti.it/commenti/2020/luglio/pdf3/ecdc_covid.pdf

[7] Meta. (2023). *The Situation with the Coronavirus Officially*. [Online]. Available: https://www.coronavirus2020.kz

[8] Meta. (2023). *Mellega al pincho...@virus2020*. [Online]. Available: https://t.me/virus2020

[9] Meta. (2023). The Astana International Financial Center (AIFC). [Online]. Available: https://aifc.kz

[10] Meta. (2023). Open Data gov. [Online]. Available: https://data.egov.kz

[11] P. G. Asteris, M. G. Douvika, C. A. Karamani, A. D. Skentou, K. Chlichlia *et al*. "A novel heuristic algorithm for the modeling and risk assessment of the COVID-19 pandemic phenomenon," *CMES-Computer Modeling in Engineering & Sciences*, 2020, vol. 125, no. 2, pp. 815–828.

[12] P. G. Asteris, S. Kokoris *et al*., "Early prediction of COVID-19 outcome using artificial intelligence techniques and only five laboratory indices," *Clinical Immunology (Orlando, Fla.)*, vol. 246, 109218, 2023. https://doi.org/10.1016/j.clim.2022.109218

[13] C. Mahanty, R. Kumar, P. G. Asteris, and A. H Gandomi, "COVID-19 patient detection based on fusion of transfer learning and fuzzy ensemble models using CXR images," *Appl. Sci.* 2021, *11*, 11423. https://doi.org/10.3390/app112311423

[14] S. Kotsiantis, K. Patriarcheas, and M. Xenos, "A combinational incremental ensemble of classifiers for predicting schoolchildren's performance in distance education," *Knowledge-Based Systems*, 2010, no. 23(6), pp. 529–535.

[15] S. Kaddoura, D. E. Popescu, and J. D. Hemanth, "A systematic review on machine learning models for online learning and examination systems," *Peer J. Computer Science*, vol. 8, e986, 2022. https://doi.org/10.7717/peerj-cs.986

[16] I. Rahimi, A. H. Gandomi, P. G. Asteris, and F. Chen, "Analysis and prediction of COVID-19 using SIR, SEIQR, and machine learning models: Australia, Italy, and UK cases," *Information*, 2021, vol. 12, p. 109. https://doi.org/10.3390/info12030109

[17] N. Thai-Nghe, A. Busche, and L. Schmidt-Thieme, "Improving academic performance prediction by dealing with class imbalance," in *Proc. 2009 Ninth International Conference on Intelligent Systems Design and Applications*, Pisa, Italy, 2009, pp. 878–883.

[18] M. N Injadat, A. Moubayed, A. B. Nassif, and A. Shami. "Systematic ensemble model selection approach for educational data mining." *Knowledge-Based Systems*, 2020, vol. 200, no. 105992.

[19] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: A classification and combining techniques review," *Artificial Intelligence Review*, vol. 26, no. 3, pp. 159–190, 2006.

[20] T. V. Batura, A. M. Bakiyeva, and M. V. Charintseva, "A method for automatic text summarisation based on rhetorical analysis and topic modelling," *International Journal of Computing*, vol.19, no. 1, pp. 118–127, 2020.

[21] S. Kotsiantis, K. Patriarcheas, and M. Xenos, "A combinational incremental ensemble of classifiers for predicting schoolchildren's performance in distance education," *Knowledge-Based Systems*, vol. 23, no. 6, pp. 529–535, 2010.

[22] A. Idris, M. Rizwan, and A. Khan, "Churn prediction in telecom using Random Forest and PSO-based data balancing in combination with various feature selection strategies," *Computers & Electrical Engineering*, vol. 38, no. 6, pp. 1808–1819, 2012.

[23] A. Idris, A. Khan, and Y.S. Lee, "Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification," *Applied Intelligence*, vol. 39, no. 3, pp. 659–672, 2013.

[24] J. Pamina, R. Beschi, S. SathyaBama, S. Soundarya, M. S. Sruthi, S. Kiruthika, V. J. Aiswaryadevi, and G. Priyanka, "An Effective classifier for predicting churn in telecommunication," *Jour of Adv Research in Dynamical & Control Systems*, vol. 11, no. 01, 2019.

[25] L. Wang, X. Wang, A. Chen, X. Jin, and H. Che "PrEDLction of type 2 diabetes risk and its effect evaluation based on the XGBoost model," *Healthcare (Basel)*, vol. 8, no. 247, 2020.

[26] C. C. Gray and D. Perkins, "Utilising early engagement and machine learning to predict schoolchildren outcomes," *Computers & Education*, no. 131, pp. 22–32, 2019.

[27] M. Kumar and A. J. Singh, "Evaluation of data mining techniques for predicting schoolchildren's performance," *International Journal of Modern Education and Computer Science*, vol. 8, no. 4, pp. 25–31, 2017.

[28] O. Yildiz, A. Bal, and S. Gulsecen, "Improved fuzzy modelling to predict the academic performance of distance education schoolchildren," *International Review of Research in Open and Distance Learning*, vol. 14, no. 5, pp.144–165. 2013.

[29] U. T. Makkhazhanova, F. A. Murzin, A. A. Mukhanova, and E. P. Abramov "Fuzzy logic of Zadeh and decision-making in the loan field," *Journal of theoretical and applied Information Technology*, vol. 98, no. 06, pp. 1076–1086, 2020.

[30] U. Makhazhanova, S. Kerimkhulle, A. Mukhanova, A. Bayegizova, Z. Aitkozha, A. Mukhiyadin, B. Tassuov, A. Saliyeva, R. Taberkhan, and G. Azieva, "The evaluation of CrEDLtworthiness of trade and enterprises of service using the method based on fuzzy logic," *Applied Sciences*, vol.12, no. 22, 11515, 2022.

[31] S. Sisovic, M. Matetic, and M. B. Bakaric, "Clustering of imbalanced Moodle data for early alert of schoolchildren failure," in *Proc. 2016 IEEE 14th international symposium on applied machine intelligence and informatics (SAMI)*, pp. 165–170, 2016.

[32] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Predicting schoolchildren's performance in distance learning using machine learning techniques," *Applied Artificial Intelligence*, vol.18, no. 5, pp. 411–426, 2004.

[33] K. Bunkar, U.K. Singh, B. Pandya, and R. Bunkar, "Data: PrEDLction for performance mining improvement of graduate schoolchildren using classification," *IFIP International Conference on Wireless and Optical Communications Networks, WOCN*, 2012.

[34] A. Tekin, "Early prediction of schoolchildren' grade point averages at graduation: A data mining approach," *Eurasian Journal of Educational Research*, vol. 14, no. 54, pp. 207–226, 2014.

[35] Ş. Aydogdu, "Predicting schoolchildren final performance using artificial neural networks in online learning environments," *Education and Information Technologies*, vol. 25, no. 3, pp. 1913–1927. 2019.

[36] S. Sandugash, T. Jamalbek, S. Madina, Y. Akbota, and A. Ainur, "Building a standard model of an information system for working with documents on scientific and educational activities" *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 9, pp. 445–455, 2021.

[37] S. S. Kurmanbekovna, S. Z. Bakirbaevna, B. A. Gabitovich, S. M. Aralbaevna, and Y. A. Sembekovna, "Development of technology to support large information storage and organization of reduced user access to this information," *International Journal of Advanced Computer Science and Applications*, vol.12, no. 7, pp. 493–503, 2021.

[38] S. Sudman and N. M. Bradburn, "Asking questions—A practical guide to survey instrument design," *San Francisco: Jossey-Bass Publishers*, 1989.

[39] J. Tussupov, K. Kozhabai, A. Bayegizova, L. Kassenova, Z. Manbetova, N. Glazyrina, M. Bersugir, and M. Yeginbayev, "Applying machine learning to improve a texture type image," *Eastern-European Journal of Enterprise Technologies*, vol. 2, pp. 13–18, 2023.

Copyright © 2023 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).