I.J. Modern Education and Computer Science, **2024, 6, 94-107**
Published Online on December 8, 2024 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2024.06.07

Predicting Education Level of the Farmers' Children of a Developing Country during COVID 19 Using Machine Learning

Md. Mehedi Rahman Rana
Department of CSE, Army University of Science and Technology (BAUST), Khulna
Email: mehedi.rahman@baust.edu.bd
ORCID iD: https://orcid.org/0009-0004-1017-0932

Md. Nasim Adnan*
Department of CSE, Jashore University of Science and Technology, Jahore-7408, Bangladesh
Email: nasim.adnan@just.edu.bd
ORCID iD: https://orcid.org/ 0000-0001-9210-2896
*Corresponding Author

Md. Moradul Siddique
Department of CSE, University of Information Technology and Sciences, Dhaka-1212, Bangladesh
Email: moradul_siddique@uits.edu.bd
ORCID iD: https://orcid.org/0000-0003-3264-5383

Md. Tahadur Rahman
Department of CSE, Northern University of Business & Technology, Khulna-9100, Bangladesh
Email: trtanmoy1@gmail.com
ORCID iD: https://orcid.org/0009-0003-7319-8871

Ferdib-Al-Islam
Department of CSE, Northern University of Business & Technology, Khulna-9100, Bangladesh
Email: ferdib.bsmrstu@gmail.com
ORCID iD: https://orcid.org/0000-0002-0758-2790

Received: 05 February, 2024; Revised: 15 April, 2024; Accepted: 16 May, 2024; Published: 08 December, 2024

Abstract: Education is one of the necessities of an individual's life, as it enhances the self-morality and nobility that leads one towards the challenging pathways of the competitive world. In the agricultural based country, education is scarce among the children of the farmers as they suffer from poverty. After affecting with COVID-19, study dropout rate of farmers' children is increased. We collected raw data from rural areas of different countries, and pre-processed this data before applying the machine learning algorithm to improve the performance. We used advanced machine learning models to predict whether farmer's children will run or drop out of their education. Based on the outcomes it was viewed that, machine learning strategies substantiate to be suitable in this area. This research proposes preventive steps for dropping out of the farmers' children. It also shows that, the Random Forest being the highest reliable model for foreseeing dropout rate and education level.

Index Terms: Dropout level of farmer's children, education level, education dropout, development of farmer's children education, predicting farmer's children education level

1. Introduction

Education is one of the basic and promising fields in the world. Such a kind of fact leads a country towards positive prosperity and well recognition. The efficiency of education in the developing country is downward. Education for all and ensuring quality education continues to get persistent attention and commitment of governments across the

globe. Some developing countries has made remarkable progress in escalating access to education over the previous decade. In case, the common issue for the instruction framework in developing country is inadequate quality. Overpopulation and under-qualified teachers have prompted high dropout rates at the essential and optional levels. Indeed, while essential enlistment might be at 98%, it drops considerably to 22% at the higher auxiliary level. This is particularly apparent in-country networks where different social and financial elements can forestall safe admittance to schooling – things like cataclysmic events, kid work, family neediness, inability, sex misuse, youngster marriage, and helpless sterilization [1].

The education sector of the developing country is facing immense challenges due to Covid-19. Because of the pandemic, there could be a critical ascent in the number of dropouts in the instructive establishment. If the school is closed, the students will be involved in income-earning instead of education. Poverty rates will affect young people, especially students, as dropout rates may increase. Experts fear that if the Covid-19 shutdown reopens after 2022, more than 45% of the developing country's secondary students will not return to school. They say if the epidemic continues for a long time, the dropout rate is likely to be even higher. Losing a job for many parents, especially in the informal sector, is the main reason for dropping out. Different factors like social and financial foundation, populace, and family foundation influence a youngster's tutoring [2]. The public authority's reaction plan of COVID 19 incorporates following and bringing youngsters back to school to forestall dropouts, which should be completely carried out.

On the other hand, Agriculture is the largest business sector in almost every developing country. The presentation of this zone mainly affects major macroeconomic goals such as working age, poverty reduction, human resource improvement, food safety, and other monetary and communal capabilities [3]. As the farmers ensure all human food resources, likewise the government should ensure their future. The maximum data of our paper has been collected from different regions of Bangladesh. In this country, like farmer needs, the people need to be educated to lead the country forward. It is known that Bangladesh is a developing country, and the poor financial condition stopped their children's education midway. Only a specific level of education is barely enough to ensure both qualitative and quantitative education with accuracy. Suitable steps should be occupied so that the children of farmers do not drop out of education. Furthermore, this research will help predict at which stage the farmer's children are dropping out of education.

In this research, we plan to design a framework for understudies utilizing current instructive information mining and progressed machine learning strategies that give guidance ahead of time by anticipating the stage at which the farmer's children are exiting the review. In India and Bangladesh, the children of farmers are dropping out of study every year due to many problems. If the problem is solved by finding out the students who are likely to drop out of the study, the dropout rate will be reduced.

In this work, a dataset consuming 506 samples and 22 features with five dissimilar classes labelled 'Above H.S.C.', 'Above S.S.C. but Below H.S.C.', 'Above Class 5 but Below S.S.C.', 'Below Class 5', and 'Illiterate' has been created. The attributes of the dataset are contained to become goodly pertinent for anticipating the objective. We have collected raw data from the farmers in rural areas of India and Bangladesh. In any case, inconsistent dissemination of information tests among classes is seen, demonstrating a lopsided circumstance. Utilizing Machine Learning (ML), tremendous amounts of information can be rethought, and explicit plans uncovered would not be promptly recognizable to us. ML approaches have been utilized to dissect instructive information on holding and graduation rates [4]. Data mining is evoking concealed information from immense crude datasets. At last, getting information disclosure is applied to ML strategies for prediction and equivalent decisions [5]. Present-day ML techniques convey material expectations of understudy assessment appraisals like imprints, achievement, etc. [6]. Five notable model-based methodologies, specifically Support Vector Machines (SVM), Decision-Tree, Gradient Boosting Classifier, Ada-Boost, and Random Forest, have been applied for administered characterization undertakings. It has been noticed that the presentation of Random Forest is better than all.

As a result, this research lies in its potential to inform targeted interventions and policies aimed at supporting the education of farmers' children in developing countries during and beyond the COVID-19 pandemic. Moreover, by identifying factors contributing to educational disparities, policymakers can design holistic strategies to address systemic issues, such as poverty, access to healthcare, and digital divide, thereby fostering equitable and inclusive education systems. By accurately predicting education levels, stakeholders can identify at-risk individuals and implement tailored interventions to ensure continued learning and educational attainment. Additionally, this research contributes to our understanding of the complex socio-economic factors influencing educational outcomes in vulnerable populations, shedding light on disparities exacerbated by the pandemic. Ultimately, this research has the potential to empower farmers' children and their communities, enabling them to break the cycle of poverty and contribute to sustainable development in countries.

This thesis aims to investigate the education levels of farmers' children in developing countries, particularly in the context of the unprecedented challenges posed by the COVID-19 pandemic. By leveraging advanced machine learning techniques, this research seeks to predict educational outcomes and elucidate the socio-economic factors that influence educational attainment among vulnerable populations. The study's significance lies in its timely response to the heightened educational challenges exacerbated by the pandemic, offering insights into the long-term effects of COVID-19 on educational equity and access. Through its innovative approach, this research strives to inform evidence-based interventions and policy initiatives aimed at mitigating the adverse impacts of the pandemic on educational attainment and promoting equitable access to education for all.

The remainder of this proposition is composed as follows. Section II discuss the literature review. Section III expounds on the utilized thesis methodology. Section VI designates the result and discussion. Finally, section V indicates the conclusion and future work.

2. Review of Literature

Throughout the long term, numerous analysts have effectively done many works in regards to investigating the purposes for foreseeing understudy dropouts in instruction. A portion of these proposal works incorporate just examining the properties which have immediate or roundabout consequences for understudies' exhibition and some additionally incorporate foreseeing other understudies' outcomes dependent on the concentrated on ascribes utilizing distinctive learning calculations. The learning algorithm such as heuristics, ML, artificial neural networks, Random Forest, and Bayesian classifiers are part of this is (AI).

Naïve-Bayesian data-mining procedure was useful to foresee the understudy execution dependent on nineteen (19) ascribes like sexual orientation, food propensity, the mode of educating, family status, family pay, understudies' grade, and child on [7]. Classification's algorithms, for example, the Decision-Tree calculation (J48), Naïve Bayesian, Bayesian-classifiers, k-Nearest-Neighbor (KNN) calculation, and two guideline student's calculations (JRip and OneR) were utilized to assess the understudy's exhibition. Supervised learning depends on gaining from a bunch of marked models in the preparation set so it able to recognize unlabeled models in the test-set by the most noteworthy conceivable precision. The worldview of this learning is productive and it generally discovers answers for linear and non-linear issues like grouping, plant control, anticipating, forecast, mechanical technology thus numerous others [8].

Yukselturk et. al. [9] analyzed the forecast of dropouts utilizing Data-mining methods in an internet-based schooling programme. Data was gathered by web-based polls which comprised factors like sexual orientation, age, and instructive level, self-viability, past internet-based insight, web-based learning preparation, and so forth. To arrange information, four Data-mining methods were utilized dependent on k Nearest Neighbor, Decision-Tree, Naïve-Bayes, and Neural-Network. Also, the Genetic-Algorithm was utilized toward track down the critical determinant issue(s) among the variables referenced previously.

Edward Wakelam et al. [10] predicted understudy execution in little accomplices with negligible accessible traits utilizing ML and Data mining strategy. To lead their examination, they utilized an understudy associate of 23 and recognize the understudies who are in danger. A relative investigation was performed on three unique calculations named k Nearest Neighbor's, Decision Tree (DT), and Random Forest (RF). Toward assess & look at their regression methods, they utilized mistake/exactness, Correlation Coefficient (CC), and, Mean Squared Error (MSE).

Khasanah et. al. [11] led an investigation to discover that high-impact credits might be chosen cautiously to foresee understudy execution. Element choice might be utilized before grouping for such a task. The student information was from at Indonesia University Islam Indonesia of Department of IE (Industrial Engineering). They utilized Bayesian-Network & Decision-Tree calculations for grouping & forecasting understudy execution. The Variables choice techniques presented that understudies' participation and GPA (Grade-Point-Average) in the main semester beat the rundown of elements. At the point once the precision was thought of, the Bayesian-Network outflanked the Decision-Tree characterization for their situation. Ankita A Nichat et. al. [12] constructed characterization method utilizing Decision-Tree and ANN procedures. They utilized a few ascribes to get to the strength and shortcomings of the understudies to work on their exhibition of the understudies.

Nikolovski et al. [13] proposed in his paper black propagation neural network system is measured by the creator to check the likelihood of the education drop-out of University pupils once they applied in some scholastic programme. As indicated by the examination of the dataset, accommodating characteristics are courses programming application and mathematical.

Miguel Gil, Norma Reyes et al [14] said that in his paper, the creator measured a Black-Propagation ANN framework, which is utilized to quantify the likelihood of the instruction dropout of college understudies once they joined up with some scholarly program. They additionally tracked down that the consequence of the examination might be conflicting because of understudy information filled in the review. The sexual existences of the understudy are additionally being considered as one of the significant components for the expectation.

P. Sunil Kumar, D. Jena et al [15] said that in his paper, the creators have deliberated 7 (seven) unique features & discovered the connection between them that power to drop out from study. The investigation refers, they discovered that the understudies who are not inspired studying often drop out of school when contrasted with showing poverty and teaching environment. They likewise attempted AV investigation, Correlation, and conviction examination and found that neediness, just as a showing climate, additionally makes understudies unengaged in the review.

El-Halees [16] proposed a contextual analysis that utilized instructive information mining to examine understudies' learning conduct. The aim of his audit is to display in what way valuable Data-mining be able to be utilized in higher study to work on understudies' exhibitions. They applied information mining strategies to find significant data from enormous data sets, for example, affiliation rules and order rules utilizing a DT, bunching, and exception examination.

A mixed-methods study investigating the impact of the COVID-19 pandemic on children's education, particularly in disadvantaged and rural areas of Indonesia. The research involves 900 parents, 943 children, 15 teachers, and education officials across 594 villages in 9 provinces. Key findings include a significant number of children

discontinuing learning, limited access to online education, challenges with offline learning methods, negative effects on children's mental health, parental difficulties in supporting learning alongside livelihood activities, instances of domestic violence, financial constraints hindering distance learning provision, and an increased risk of school dropouts [17].

A study conducted to assess the impact of the COVID-19 pandemic on learning outcomes in Mexico, specifically in reading and numeracy skills. The researchers compared data from household surveys conducted in 2019 and 2021, involving 3161 children aged 10 to 15 years. They found a significant learning loss in both reading and mathematics, ranging from 0.34 to 0.45 standard deviations (SD) in reading and 0.62 to 0.82 SD in mathematics. Additionally, there was an increase in learning poverty, with percentages ranging from 15.4% to 25.7% in reading and 28.8% to 29.8% in numeracy [18]. F.J. Hevia et. al. [19] introduces the application of artificial intelligence (AI) in predicting student performance during the COVID-19 pandemic in the Guelmim Oued Noun region of Morocco. Whereas, the use of AI algorithms to develop a recommendation system aimed at enhancing student outcomes in the area.

The study highlights the significance of data mining in education during the pandemic, providing insights for evaluating student performance. Especially during the COVID-19 pandemic, and explores factors influencing e-learner satisfaction. It aims to evaluate these factors and understand how data mining techniques can predict student performance. The study employs various data mining algorithms on a dataset of 1000 instances with 35 attributes related to student performance in e-learning. Results from algorithms like Decision Tree, Random Forest, Naive Bayes, and others are compared to assess the impact of e-learning features on student performance [20].

F. Faisal et al. [21] investigates the use of machine learning algorithms to predict school closures amid the COVID-19 pandemic. By analyzing UNESCO data with ten supervised ML algorithms, it identifies Decision Tree as the most accurate predictor with 90.75% accuracy, followed by Artificial Neural Network at 80.37%. The findings suggest ML algorithms hold promise in forecasting school closure scenarios due to COVID-19, offering valuable insights for education system decision-making.

M.M Nishat [22] presents the performances of these algorithms for detecting postgraduate students' psychological impact during this pandemic. An online survey dataset is employed from Mendeley data repository consisting of the responses of Malaysian students based on the questions of General Anxiety Disorder (GAD-7) and concerns about research progress, academic delay, and daily life. Among the classifiers, artificial neural network showcased the highest accuracy of 95.45% whereas Logistic Regression, Linear Kernel Support Vector Machine, and Gaussian Process exhibited accuracies of more than 90%."

In conclusion, these papers collectively contribute to understanding the multifaceted impacts of the COVID-19 pandemic on education globally. They employ various methodologies, including statistical analysis, machine learning, and mixed-methods approaches, to assess learning loss, predict academic performance, detect psychological impacts, and identify challenges faced by different populations. Each paper underscores the importance of targeted interventions and innovative solutions to address the educational disparities exacerbated by the pandemic.

3. Methodology

3.1 Dataset Collection and Description

Every variable of this dataset is characterized by probing the provisions just as element significance is represented in different papers [23, 24]. This way, twenty features' highlights fill in as likely markers for forecast, and one yield include has been indicated in such manner displayed in Table 1. Input variables are set so that they adequately affect the yield variable. The elucidating type of other thought about input variables; Farmer's Age, how many family members, Number of children's, number of spouses, farmer qualifications, farmer wife's qualification, how many earning person's present in the family, total family income in a month, amount of own land, do you have any loan, etc. exposed in Table 1. The target features hold five unalike classes: 'Illiterate', 'Below Class 5', 'Above Class 5 but Below S.S.C.', 'Above S.S.C. but Below H.S.C.', 'Above H.S.C.'.

Samples collected from rural farmers are divided into five categories based on the level of education of their children. The dataset contained 549 instances with 21 attributes after data pre-processing. We have collected raw data from the farmers in six rural areas Bangladesh and India.

Table 1. Dataset Description

Attributes	Description	Values
Age	Farmer's age	(35 to 80) Age
FM	How many family members	(3 to 12) person per family
N.F.	Number of children's	(1 to 7) child each family
N.S.	Number of spouses	1 and 2
F.Q.	farmer qualifications	Illiterate, BC5=Below Class 5, AC5BS=Above class 5 but Below S.S.C., ASBH=Above S.S.C. but Below H.S.C., AH=Above H.S.C. or Study continue and Job-holder

F.W.Q.	farmer wife's qualification	Illiterate, BC5=Below Class 5, AC5BS=Above class 5 but Below S.S.C., ASBH=Above S.S.C. but Below H.S.C., AH=Above H.S.C. or Study continue and Job-holder
N.E.P.	How many earning person's present in the family	(1 to 4) earning person
F.M.I.	Total family income in month (in B.D.T.)	If F.M.I. <=8000 then Low, If FMI>8000 but FMI<13000 them Medium If F.M.I. >=14000 them High
H.E.P.	Have any higher educated person in the family	(Yes, No)
O.H.	Do you have your own house	(Yes, No)
H.T.	Type of house	(Mud-Built, Brick-Built)
E.I.S.	Have any extra income source	(Yes, No)
AOL	Amount of own land (in Bigha)	The values are numeric in Bigha
A.L.C.	How much land does the farmer cultivate	The values are numeric in Bigha
SC	Do you work as a sharecropper in another's land	(Yes, No)
LOAN	Do you have any loan	(Yes, No)
F.D.A.	Is the farmer drug addicted	(Yes, No)
C.D.A.	Are the children's drug addicted	(Yes, No)
S.F.M.	There any severely sick family member	(Yes, No)
TDBHC	The time duration between home and school (in Km)	If TDBHC<2 km then Short If 3<TDBHC<4 then Average If TDBHC > 4 then Long

3.2 Chi-Square Test for Features

A Chi-square test is a statistical technique. Two normal Chi-square tests include checking if one classification's audited frequencies match anticipated frequencies. The Chi-square test is the same as the χ^2 test. Chi-square helps explore such differences in categorical factors, particularly nominal ones. In this dataset, we applied the chi-square test, and the results are given in Table 2.

Table 2. Chi-square test score in the dataset

Sl. No.	Top Features	Score
1	HEP	232.839693
2	HT	76.651898
3	SC	39.275916
4	AOL	38.629991
5	FQ	37.950048
6	ALC	27.507991
7	LOAN	24.823032
8	FWQ	21.923386
9	EIS	19.553995
10	CDA	13.368949
11	FDA	12.374883
12	FMI	8.717475
13	SFM	2.648445
14	NS	1.923617
15	Age	1.661077
16	TDBHC	1.328238
17	NEP	0.696036
18	NC	0.547295
19	FM	0.237990
20	OH	0.078808

A piled bar diagram is introduced in Figure 1, displaying the data index relation between farmers' education level and their children's education level. This figure shows the literacy rate of farmers is very low; where the farmer is educated, his children are more likely to be educated. This data set shows that where the farmer is uneducated, most children are also prone to be uneducated.

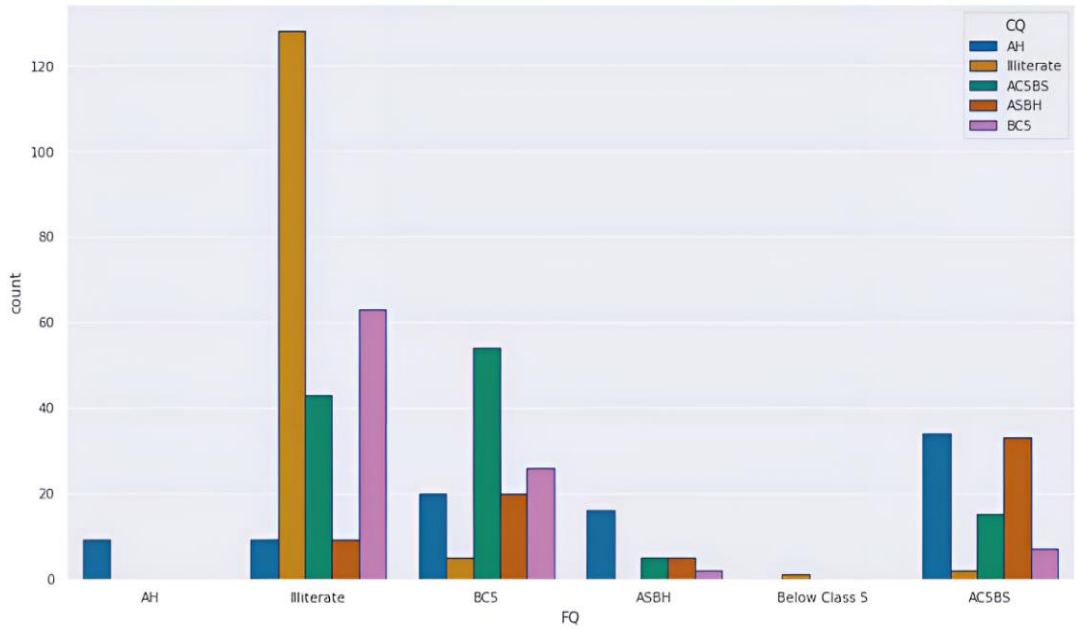

Fig. 1. Relation between C.Q. and F.Q.

In this data set where the farmer is illiterate, the farmer's children are 120 people illiterate, 10 people A.H., 44 people AC5BS, 12 people ASBH, and 64 people BC5. A bar diagram is introduced in Figure 2, displaying the data index relation between farmer wives' education level and their children's education level.

This data set shows that the farmer's wife and her children are also educated. In this data set where the farmer's wife is illiterate, the farmer's children are 130 people illiterate, 8 people A.H., 43 people AC5BS, 17 people ASBH, and 52 people BC5.

3.3 Data Preparation

The dataset contained 506 instances with 23 attributes. As the feature name of the farmer does not convey any importance, we eliminated it from the rundown of the characteristics. At last, 21 features were chosen after data cleaning.

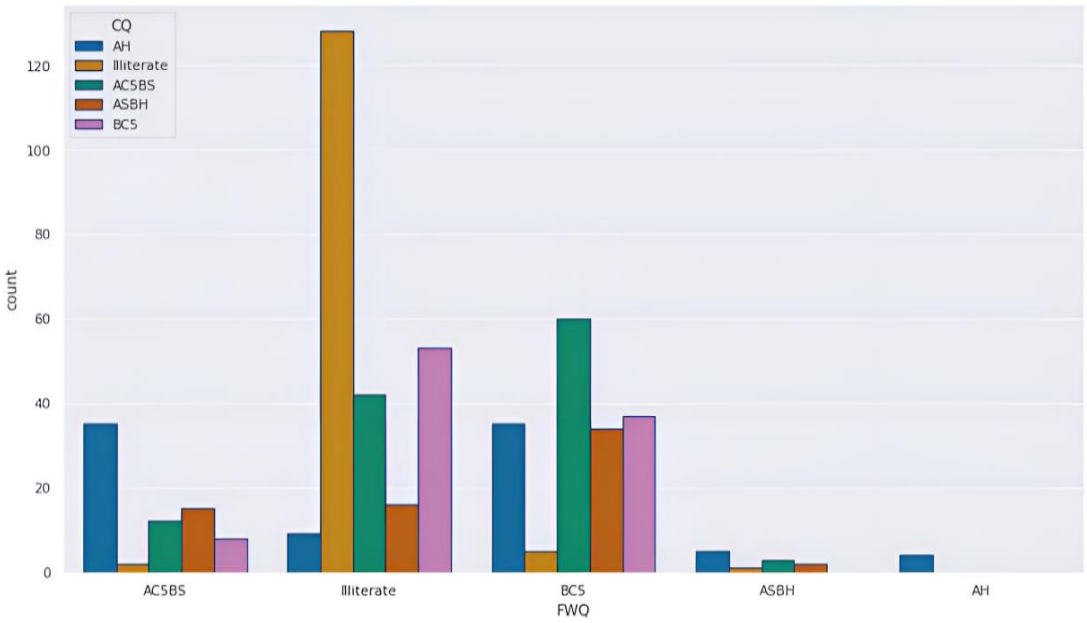

Fig. 2. Relation between C.Q. and F.W.Q.

Figure 3 shows the chosen credits with their potential qualities. A few pre-processing steps like managing missing qualities, variable encoding, Data-Normalization, and Data-Standardization are useful to my dataset as my ML algorithm improves managing our information [25]. The progression refers to the scale of my dataset and converts the data range [0, 1].

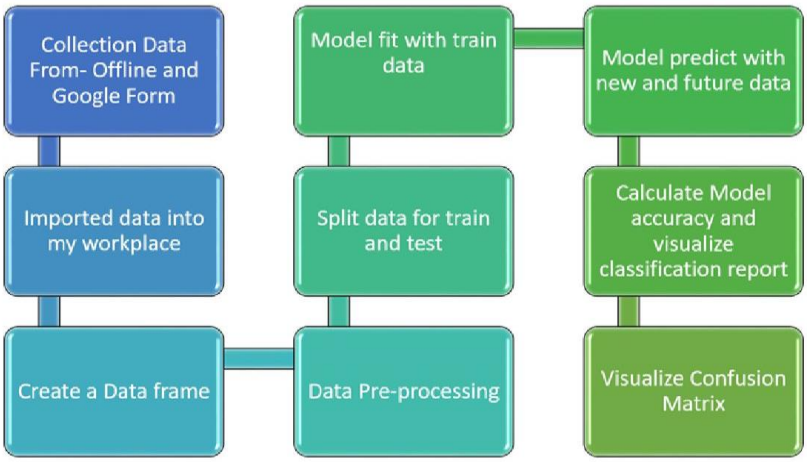

Fig. 3. Proposed methodology of the study.

A. One-hot Encoding

One-Hot encoding is a strategy by which absolute factors are changed into a structure given to machine learning algorithms to make a superior showing in the forecast. The most ordinary approach to converting categorical features to a suitable format for input to an ML model is one-hot encoding. Continuing with predicting a person's salary, the person's type of employment may be a vital factor to consider. For instance, a lawyer tends to make more money than a student. Assuming that we want to differentiate between four types of employment - student, teacher, doctor, and banker - it is possible to represent this information using one-hot encoding, as shown in Figure 4 [26].

One-Hot encoding is the most extreme broadly utilized coding structure. It compares each level of the categorical feature to a stable direction level. One-hot encoding changes over a solitary variable with n perceptions and separate qualities to d binary factors with n perceptions each. Every perception shows the dichotomous binary feature's presence (1) or nonattendance (0).

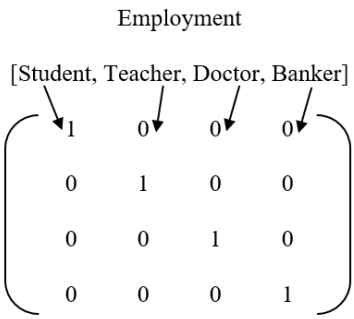

Fig. 4. Graphical representation of one-hot encoding

B. Label Encoding

In ML, we ordinarily contract with datasets that hold numerous labels in at least one than one segment. These components can be as numbers or words. In ML, we ordinarily manage datasets that hold numerous tags in at least one segment. These tags can be as numbers or words. The preparation information is frequently labelled in words to make the data fathomable or an intelligible structure. Label Encoding alludes to interpreting the names into a numeric structure to change them into machine-meaningful structures. ML algorithm would then be able to pick how those labels should be worked in a superior manner. It is a significant pre-processing venture for the organized dataset in regulated learning. Assuming that we want to differentiate between three types of Height - Tall, Medium, Short - it is possible to represent this information using one-hot encoding as shown in Figure 5 where 0 is the name for tall, 1 is the name for medium, and 2 is for short.

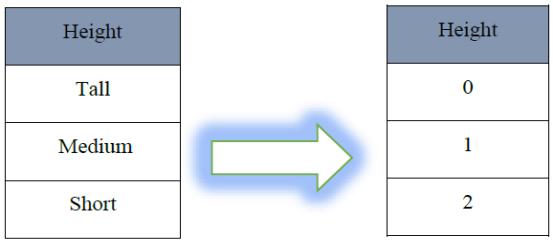

Fig. 5. Graphical representation of Label Encoding

C. Normalization

Normalization is a scaling method where esteems are moved and rescaled with the goal that they wind up going somewhere in the range of 0 and 1. It is otherwise called Min-Max scaling. This technique rescales the provisions or the yields in any range into another range. Generally, the elements are scaled between 0-1 or (- 1)-1. The uniformity utilized in the strategy is as underneath where X_{min} shows least worth, X_{max} greatest worth, X_i input worth, and X' normalization data:

$$X' = \frac{X_i - X_{min}}{X_{max} - X_{min}} \quad (1)$$

D. Cross-Validation

Cross-Validation (CV) is a procedure used to test a model's capacity to foresee inconspicuous data, data not used to prepare the model. CV is helpful if we have restricted data when our test set is not adequately enormous. There is a wide range of ways of performing out a CV. As a rule, CV parts the training data into k folds. The model trains on k-1 fold in every cycle and is approved utilizing the last folds. We can utilize many cycles of CV to decrease variability. We utilize the average blunder over every one of the cycles to assess the model. It is constantly liked to utilize a model with better CV execution. Likewise, we can utilize CV to tune model parameters. In Figure 6 shows the procedure of CV.

Steps of K-fold Cross-Validation

- The training data is split into k equivalent folds.
- The model fit on the k-1 parts and the fitted model on the kth part, calculate test error.
- Repetition k times, utilizing every data subset as the test set once. (typically, k= 5~20)

E. Performance analysis

This paper intends to anticipate farmer children's education level based on their collected data from farmers. The suggested model applied distinctive ML algorithms to assess which model played out the best for performing farmer children's education level. We measured the exhibition utilizing different measurements, including classification accuracy, Recall (Sensitivity), F-Measure, and precision to guarantee the prescient methods were appropriate to create precise outcomes. Specifically, coming up next is the hypothetical model utilized as a premise to build our multi-class forecast model:

Gradient boosting

The fundamental aim behind this algorithm is to construct models successively, and these resulting models attempt to decrease the errors of the past model. However, the entrancing goal behind Gradient Boosting is that instead of fitting an indicator on the data at every cycle, it fits another predictor to the remaining mistakes made by the past predictor. This is made by building another new model on the mistakes or residuals of the past model. The target is to limit this loss function by adding powerless learners utilizing gradient descent.

AdaBoost

AdaBoost consolidates numerous classifiers to rise the accuracy of classifiers. AdaBoost is also known as the iterative ensemble technique. AdaBoost classifier assembles a potent classifier by consolidating multiple weakly performing classifiers with the goal that you will get a high potent classifier [27]. The essential idea driving AdaBoost is to set weights of classifiers and train the data specimen in every emphasis to such an extent that it guarantees the exact forecasts of uncommon perceptions. Any ML method can be utilized as a base classifier if it accepts weights over the train set. Two conditions should meet AdaBoost's:

- The classifier ought to be prepared interactively on different weighted training examples.
- Every cycle attempts to give a great fit to these models by decreasing training errors.

How to work AdaBoost algorithm is given in Figure 6.

Decision Tree

Decision-Tree is generally utilized in a few multi-class classifications that can deal with lost qualities with high dimensional information. D.T. a generally utilized in a few multi-class classifications that can deal with missing qualities by high dimensional information [28].

Support Vector Machine

Support-Vector-Machine (SVM) depends on the idea of choice planes that conditions choice limits that manage order issues effectively [29]. It takes an arranged dataset then forecasts which of 2 possible groups incorporates the data, creation the Support-Vector-Machine, a non-probabilistic binary direct classifier.

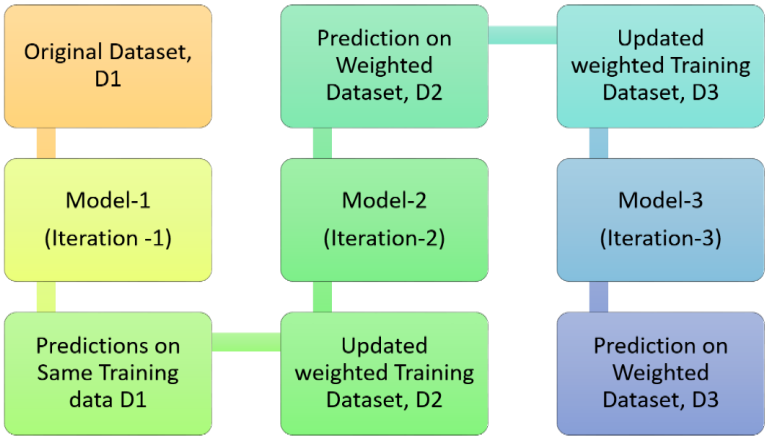

Fig. 6. AdaBoost Algorithms Working process

Random Forest

Random Forest is the classifier that relies on group discovery that utilizes various choice trees on a different subset to track down the top provisions for extraordinary exactness & forestalls the issue of over-fitting. Random Forest is generally vigorous to exceptions & clamor that works adequately in classification [30].

A confusion matrix assists with imagining the grouping execution of each prescient model. Table 3 presents the disarray network utilized for farmer posterity instruction-level expectation where P, Q, X, Y, and Z denote the classes for farmer children education name (E.L.) name as being 'A.H.', 'AC5BS', 'ASBH', 'BC5' and 'Illiterate'. The class name addresses in a structure an articulation [25]:

$$EL \in \{P, Q, X, Y, Z\} \tag{2}$$

Table 3. Confusion matrix for farmer children education label prediction classification

				Predicted		
		P	Q	X	Y	Z
A	P	PP	PQ	PX	PY	PZ
C	Q	QP	QQ	QX	QY	QZ
T	X	XP	XQ	XX	XY	XZ
U	Y	YP	YQ	YX	YY	YZ
L	Z	ZP	ZQ	ZX	ZY	ZZ

The exhibition measurements of the confusion matrix are still up in the air utilizing accuracy, recall, precision, and F-Measure in the accompanying condition [25].

$$Accuracy\ (A) = \frac{(PP+QQ+XX+YY+ZZ)}{\sum N} \tag{3}$$

Here, the number of data is N.

$$\text{Precesion (P)} = \frac{1}{5}\left(\frac{\text{PP}}{\text{PP+ QP+ XP +YP+ ZP}} + \frac{\text{QQ}}{\text{PQ+QQ+XQ+YQ+ZQ}} + \frac{XX}{PX+QX+XX+YX+ZX} + \frac{Y.Y.}{PY+QY+XY+YY+ZY} + \frac{Z.Z.}{PZ+QZ+XZ+YZ+ZZ}\right) \quad (4)$$

$$Recall\ (R) = \frac{1}{5}\left(\frac{PP}{PP+\ PQ+\ PX+PY+\ PZ} + \frac{QQ}{QP+QQ+QX+QY+QZ} + \frac{XX}{XP+XQ+XX+XY+XZ} + \frac{Y.Y.}{YP+YQ+YX+YY+YZ} + \frac{Z.Z.}{ZP+ZQ+ZX+ZY+ZZ}\right) \quad (5)$$

$$F - Measure = 2\frac{P*R}{p+R} \quad (6)$$

Here the F-measure refers to the weighted-harmonic mean of recall and precision.

The choice of ML models for predicting the education levels of farmers' children in developing countries during the COVID-19 pandemic was informed by several factors, including model performance in preliminary tests, computational efficiency, interpretability, and robustness to hyperparameters. Preliminary tests were conducted to evaluate the performance of various machine learning models on a subset of the dataset. Models assessed included Random Forest, Gradient Boosting Classifier, AdaBoost, Decision Tree, and Support Vector Machine (SVM). Performance metrics such as accuracy, precision, recall, and F1-score were computed to assess each model's predictive capabilities. The results of these preliminary tests revealed that ensemble learning methods, particularly Random Forest and Gradient Boosting Classifier, consistently outperformed other models in terms of accuracy and F1-score. These models demonstrated superior predictive performance and robustness to variations in the dataset, making them suitable candidates for further evaluation.

Additionally, the computational efficiency of each model was considered, with Random Forest and Gradient Boosting Classifier exhibiting relatively faster training times compared to other models such as SVM. This scalability is crucial for handling large datasets and optimizing model performance in real-world applications. Furthermore, the interpretability of the models was assessed, with Decision Tree and Random Forest offering intuitive insights into feature importance and decision-making processes. This interpretability is essential for stakeholders, such as policymakers and educators, to understand the factors influencing educational outcomes among farmers' children and inform targeted interventions accordingly.

4. Results and Discussion

To execute this proposition, five regulated ML algorithms named Random Forest, Support-Vector-Machine (SVM), Decision-Trees, Gaussian-Naïve-Bayes, and K-Nearest Neighbours to foresee have been used for settling grouping assignments. For carrying out Random Forest, we checked for an alternate quantity of D.T. where the quantity is greater than (500) to accomplish elite and yield the outcome after the greatest democratic methods. I have more than 500 samples in our dataset, so Random Forest algorithms performance is well.

Our fundamental target is to look at the prescient model dependent on the exactness execution in this segment. Here, five chosen algorithms were utilized to prepare the farmer children dataset, and their expectation precision was assessed. The trained and tested ratio is 80:20 of the models. Using the function name train_test_split is denoted as "train_test_split (x, y, test_size = 0.2, random_state = int)". At this point, test_size is 0.2, which implies 80% information was used for preparing and 20% for testing.

4.1 Performance of the Models

The exhibitions of classification-based Machine learning models are determined, and the accuracy percentage is determined. We calculate the performance utilizing different measurements with classification accuracy, recall (Sensitivity), precision, and F-Measure to guarantee that the prescient methods were appropriate for delivering precise outcomes. Table 4 refers to the expected execution performance of the various classifiers on the farmers' information.

It very well may be realized from Table 4 that the outcomes showed Gradient Boosting Classifier & R.F. accomplish the best forecast execution with an accuracy-score separately 0.85 and 0.86 while followed by AdaBoost with 0.75. In the interim, Decision Tree and SVM got an accuracy, respectively 0.84 and 0.61. The most reduced model is accomplished by SVM with 0.61. Notwithstanding, on the grounds that the classes in my dataset are imbalanced, the expected outcomes were regularly prompted by misclassification choices of the minority class that was made while preparing the dataset. For generalizability purposes, different tests in managing the issues were directed to diminish the proportion of each class which is depicted in the following subsection.

Table 4. Comparison performance of the models.

Algorithms	Accuracy	Precision	Recall	F1-score
Support Vector Machines	0.61	0.61	0.58	0.59
Decision Trees	0.84	0.86	0.84	0.85
Gradient Boosting Classifier	0.85	0.87	0.85	0.85
Random Forest	0.86	0.87	0.85	0.86
AdaBoost	0.75	0.73	0.72	0.72

From this data set, Random Forest performed best to predict the level of education of farmer's children. The F1-Measure is 0.85 for the micro-average, in which computes worldwide metrics by count the total T.P., F.N., and F.P. Nonetheless, this can be misdirecting for imbalanced informational sets, which is the situation here. A more practical measure is the macro average, which registers measurements for each name, and tracks down their un-weighted mean, which for this situation is 0.85. This measurement doesn't consider name lop-sidedness and shows that the model would not perform accurately in its present arrangement and it is probably going to make classification blunders for under-addressed instances. In any case, the outcomes are empowering with a weighted average of 0.85.

It is known that, Random Forest (RF) is an ensemble learning method that combines the predictions of multiple decision trees. Each tree in the forest is trained independently on a random subset of the data and features, and the final prediction is made by averaging the predictions of all trees (for regression) or taking a majority vote (for classification). This ensemble approach helps reduce overfitting and improve generalization performance. Random Forest provides a measure of feature importance, which indicates the relative importance of each feature in predicting the target variable. This information can help identify the most relevant features for predicting educational outcomes among farmers' children, providing valuable insights for policymakers and educators. Moreover, RF is less sensitive to noise and outliers in the data, resulting in more robust predictions and capable of capturing complex non-linear relationships between input features and the target variable. Finally, Random Forest is relatively robust to hyperparameter tuning, making it easier to train and optimize compared to other models like Decision Tree (DT) or Gradient Boosting Machines (GBM). This makes Random Forest a practical choice for predictive modeling tasks with large and complex datasets.

4.2 Cross-Validation Check

Cross-validation refers, you make a fixed number of folds (or parts) of the data, executing the analysis on each part, and then the overall error estimate is averaged. Using 5-folds cross validation techniques each iteration gives a new accuracy and then we have got new 5 accuracy. In the term, Cross validation check results shows in Table 5.

Table 5. Using 10-folds cross validation

	5-Fold Accuracy					
	1st	**2nd**	**3rd**	**4th**	**5th**	**Mean**
DT	0.80	0.84	0.84	0.92	0.84	0.84
RF	0.80	0.86	0.84	0.91	0.84	0.85
SVM	0.62	0.66	0.53	0.58	0.58	0.60
GBC	0.81	0.85	0.89	0.93	0.85	0.86
AdaBoost	0.72	0.74	0.80	0.77	0.70	0.74

After using 10-folds cross validation, Decision tree gives the best accuracy is 0.92 and lowest accuracy is 0.80 and the mean accuracy is 0.84. Random Forest gives the best accuracy is 0.91 and lowest accuracy is 0.80 and the mean accuracy is 0.85. And Support Vector Machine gives the best accuracy is 0.66 and lowest accuracy is 0.53 and the mean accuracy is 0.60. Gradient Boosting Classifier gives the best accuracy is 0.93 and lowest accuracy is 0.81 and the mean accuracy is 0.86. AdaBoost gives the best accuracy is 0.80 and lowest accuracy is 0.70 and the mean accuracy is 0.74.

The R.O.C. curve for multi-class classification models can be determined in Figure 7. The AUC-ROC curve is just for binary classification issues. Be that as it may, we can extend it to multi-class classification issues by utilizing the one vs all technique. We have calculated five R.O.C. curve and above R.O.C. curve is the model of Random Forest because of this model gives the best accuracy in our dataset.

The predictive performance of various machine learning models was rigorously evaluated to determine their efficacy in predicting the education levels of farmers' children in developing countries during the COVID-19 pandemic. The models assessed included Random Forest, Gradient Boosting Classifier, AdaBoost, Decision Tree, and Support Vector Machine (SVM). Performance metrics such as accuracy, precision, recall (sensitivity), F1-score, and area under the receiver operating characteristic curve (AUC-ROC) were computed to assess the models' predictive capabilities. Additionally, statistical significance tests, such as cross-validation and hypothesis testing, were conducted to validate

the observed differences in model performance. Furthermore, the F1-scores for Gradient Boosting Classifier and Random Forest were also notably higher compared to other models, indicating their superior balance between precision and recall. This suggests that these models achieved better overall predictive performance in identifying both positive and negative instances of education levels among farmers' children. The robustness of these models, as validated through statistical significance tests, highlights their potential utility in informing targeted interventions and policy initiatives aimed at supporting vulnerable populations in accessing quality education.

In the future, we will collect more data and compare the dropout rate of children of farmers among the agricultural based developing country of the world.

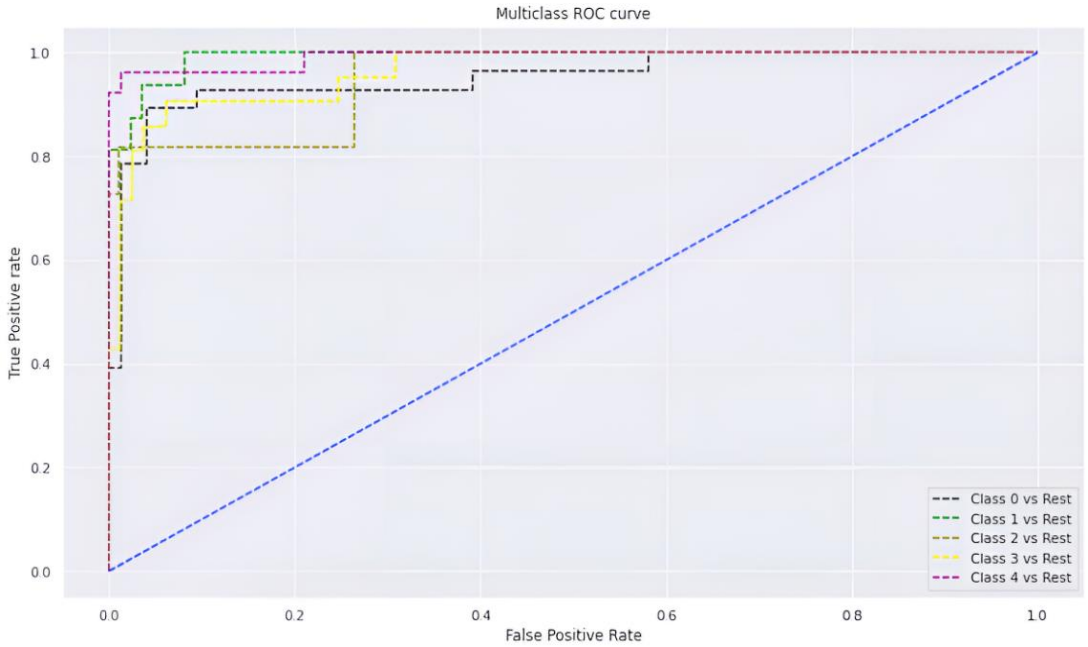

Fig. 7. Multi-class classification R.O.C. curve of Random Forest Model

5. Conclusion

The point of this research is to make farmers mindful and support farmer posterity as they can make convenient moves, such as wanting to decrease the dropout rate by early giving the farmer children education level test outcome utilizing prescient techniques for ML. In this research we propose a multi-class forecast model with five foretelling models to predict farmers' children education level based on farmer family information. A similar examination has been achieved amid the methods through footage assessed the accommodation utilizing the greatest helpful measurements (Accuracy, Precision, Sensitivity, F1-measure), and based on this exhibition best re-inspecting procedure is picked for this dataset. Along with the results search guide, a class method called Random Forest has been selected as a recommended method of excellence. We believe that the proposed method will be successful to predict the fall of the farmer's children from education. A straightforward outline of significant provisions is additionally given in this research. One significant challenge in implementing ML based models in real-world scenarios is the availability and quality of data. Rural areas in developing countries may lack comprehensive and reliable datasets necessary for training robust models. Particularly, there has several concerning issues such as privacy, fairness, and algorithmic bias. As a component of future-task, supplementary arising instructive Data-mining and ML algorithms can be remembered for the general investigation and correlation. Incorporating real-time data analysis into the study can offer insights into the dynamic nature of educational challenges. Educational organizations are able to examine which categories of farmers' children may need more attention using this research as its base. Similar to Random Forest, Forest CERN [31], Forest PA [32] and BDF [33] will also be used in comparison spectrum. Access to all experimental code and dataset can be found at https://github.com/TR-Tanmoy/Predicting-education-level-of-farmer-s-offspring-using-machine-learning-technique.

References

[1] Teach the World Foundation, 2021, [online]. Available at: https://www.teachtheworldfoundation.com/whom-we-impact/bangladesh
[2] R. Kominski, "Estimating the national high school dropout rate", Demography, vol. 27, no. 2, pp.303-311, 1990.
[3] "C.I.A. – The World Factbook". Central Intelligence Agency, 2019 [online]. Available at: https://www.cia.gov/the-world-factbook/

[4] T.A. Cardona, "Predicting student retention using support vector machines", Procedia Manufacturing, vol. 39, pp.1827-1833, 2019.
[5] B.W. O'Bannon and K.M. Thomas, "Mobile phones in the classroom: Preservice teachers answer the call", Computers & Education, vol. 85, pp.110-122, 2015.
[6] M. Hussain, W. Zhu, W. Zhang, S.M.R. Abidi and S. Ali, "Using machine learning to predict student difficulties from learning session data", Artificial Intelligence Review, 52, pp.381-407, 2019.
[7] T. Devasia, T.P. Vinushree and V. Hegde, March. "Prediction of students performance using Educational Data Mining". In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE), Ernakulam, India, March 16-18, 2016, pp. 91-95.
[8] R. Sathya and A. Abraham, "Comparison of supervised and unsupervised learning algorithms for pattern classification", International Journal of Advanced Research in Artificial Intelligence, vol .2 no. 2, pp.34-38, 2013.
[9] E. Yukselturk, S. Ozekes and Y.K. Türel, "Predicting dropout student: an application of data mining methods in an online education program", European Journal of Open, Distance and e-learning, vol. 17, no. 1, pp.118-133, 2014.
[10] E. Wakelam, A. Jefferies, N. Davey and Y. Sun, "The potential for student performance prediction in small cohorts with minimal available attributes", British Journal of Educational Technology, vol. 51, no. 2, pp.347-370, 2020.
[11] A.U. Khasanah, "A comparative study to predict student's performance using educational data mining techniques", In IOP Conference Series: Materials Science and Engineering, vol. 215, no. 1, p. 012036, 2017.
[12] N. Ankita and, R. Anjali, "Analysis of Student Performance Using Data Mining Technique", International Journal of Innovative Research in Computer and Communication Engineering, vol. 5, no. 1, 2017.
[13] V. Nikolovski, I. Mishkovski, R. Stojanov, I. Chorbev and G. Madjarov, Educational data mining: Case study for predicting student dropout in higher education, 2015.
[14] M. Gil, N. Reyes, M. Juárez, E. Espitia, J. Mosqueda and M. Soria, "Predicting early students with high risk to drop out of university using a neural network-based approach", ICCGI, p.302, 2013.
[15] P.S. Kumar, A.K. Panda and D. Jena, "Mining the factors affecting the high school dropouts in rural areas", International Journal of Advanced Computer Engineering and Communication Technology (IJACECT), ISSN (Print), pp.2278-5140, 2013.
[16] A. El-Halees, Mining students data to analyze e-Learning behavior: A Case Study, 2009.
[17] M. Indrawati,C. Prihadi and A. Siantoro, "The COVID-19 Pandemic impact on children's education in disadvantaged and rural area across Indonesia". International Journal of Education (IJE), vol 8, no. 4, pp.19-33, 2020.
[18] A. Tarik, H. Aissa, and F. Yousef, "Artificial intelligence and machine learning to predict student performance during the COVID-19". Procedia Computer Science, 184, pp.835-840, 2021.
[19] F.J. Hevia, S. Vergara-Lope, A. Velásquez-Durán and D. Calderón, "Estimation of the fundamental learning loss and learning poverty related to COVID-19 pandemic in Mexico", International Journal of Educational Development, 88, p.102515, 2022.
[20] I.L.H Alsammak, A.H. Mohammed and I.S. Nasir, "E-learning and COVID-19: predicting student academic performance using data mining algorithms", Webology, vol. 19, no. 1, pp.3419-3432, 2022.
[21] F. Faisal, M.M. Nishat, M.A. Mahbub, M.M.I. Shawon and M.M.U.H. Alvi, "Covid-19 and its impact on school closures: a predictive analysis using machine learning algorithms" "In 2021 International Conference on Science & Contemporary Technologies (ICSCT), pp. 1-6, IEEE, 2021.
[22] M.M. Nishat, M.A. Mahbub, A. Ahmed and M.A. Al Mamun, "Performance assessment of machine learning classifiers in detecting psychological impact of postgraduate students due to COVID-19", In 2021 6th IEEE international conference on recent advances and innovations in engineering (ICRAIE), vol. 6, pp. 1-6, IEEE, 2021.
[23] S. Hussain, N.A. Dahan, F.M. Ba-Alwib and N. Ribata, "Educational data mining and analysis of students' academic performance using WEKA", Indonesian Journal of Electrical Engineering and Computer Science, vol. 9, no. 2, pp.447-459, 2018.
[24] M.A.A. Walid, S.M. Ahmed and, S.S. Sadique, "A Comparative Analysis of Machine Learning Models for Prediction of Passing Bachelor Admission Test in Life-Science Faculty of a Public University in Bangladesh", In 2020 IEEE Electric Power and Energy Conference (EPEC), Edmonton, AB, Canada, November 09-10, 2020, pp. 1-6.
[25] S.D.A. Bujang, A. Selamat, R. Ibrahim, O. Krejcar, E. Herrera-Viedma, H. Fujita and N.A.M. Ghani, "Multiclass prediction model for student grade prediction using machine learning", IEEE Access, vol. 9, pp.95608-95621, 2021.
[26] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance", Applied Soft Computing, vol. 97, p.105524, 2020.
[27] C. Seger, An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing, 2018.
[28] B. Predić, G. Dimić, D. Rančić, P. Štrbac, N. Maček and P. Spalević, "Improving final grade prediction accuracy in blended learning environment using voting ensembles", Computer Applications in Engineering Education, vol. 26, no. 6, pp.2294-2306, 2018.
[29] A.K. Srivastava, D. Singh, A.S. Pandey and T. Maini, "A novel feature selection and short-term price forecasting based on a decision tree (J48) model", Energies, vol. 12, no. 19, p.3665, 2019.
[30] X. Zhang, R. Xue, B. Liu, W. Lu and Y. Zhang, "Grade prediction of student academic performance with multiple classification models", In 2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Huangshan, China, July 28-30, 2018, pp. 1086-1090.
[31] M.N. Adnan and M.Z. Islam, "Forest CERN: a new decision forest building technique," In Advances in Knowledge Discovery and Data Mining: 20th Pacific-Asia Conference, PAKDD, Auckland, New Zealand, pp. 304-315, 2016.
[32] M.N. Adnan and M.Z. Islam, "Forest PA: Constructing a decision forest by penalizing attributes used in previous trees", Expert Systems with Applications, vol. 89, pp.389-403, 2017.
[33] M.N. Adnan, R.H. Ip, M. Bewong and M.Z. Islam, "BDF: a new decision forest algorithm", Information Sciences, vol. 569, pp.687-705, 2021.

Authors' Profiles

Engg. Md. Mehedi Rahman Rana has been working as an Assistant Professor in the Department of Computer Science and Engineering at Bangladesh Army University of Science and Technology Khulna. He is a member of The Institution Of Engineers, Bangladesh and his Membership No is M - 43865. He received M.Sc. in Computer Science and Engineering (CSE) degree with thesis and B.Sc. in CSE from Khulna University at Computer Science and Engineering Discipline in 2021 and 2016 respectively. Previously he worked as a lecturer in Northern University of Business and Technology Khulna. He worked on a variety of topics including object detection and classification, image recognition and segmentation, spectral image reconstruction, document image analysis and image quality analysis. His research activities are focused in the area of computer vision, image processing, machine learning, optimization and data mining.

Dr. Md. Nasim Adnan is an Assistant Professor in the Department of Computer Science and Engineering, Jashore University of Science and Technology. Nasim received his B.Sc. in Computer Science and Engineering degree from Khulna University, Bangladesh, M.Sc. in Computer Science and Engineering degree from Bangladesh University of Engineering and Technology (BUET), Bangladesh and Ph.D. from the Charles Sturt University, Australia in 2002, 2010 and 2017 respectively. Previously he worked as an Assistant Professor at the Department of Computer Science and Engineering in the University of Liberal Arts Bangladesh (ULAB). He also served as a Systems Analyst in Bangladesh Bank (the Central Bank of Bangladesh). His research interest includes Data Mining, Software Engineering and E-Commerce. He authored and co-authored several research papers on Data Mining, Software Engineering and E-Commerce.

Md. Moradul Siddique is a lecturer in the Department of Computer Science and Engineering, University of Information Technology and Sciences (UITS). Along with, Moradul studying M.Sc. in Computer Science and Engineering from Jashore University of Science and Technology (JUST). He received the Bachelor degree in the Department of Computer Science and Engineering, Jashore University of Science and Technology (JUST). Moradul received a Diploma in Engineering (Automobile Technology) degree from Dhaka Polytechnic Institute. His research interests lie in the fields of Machine Learning, Deep Learning, Natural Language Processing, Data Mining and Bioinformatics. He is also a multi-disciplinary practitioner with experience in IoT products, 3D Designs and printing using a 3D Printer.

Md. Tahadur Rahman received B.Sc. in Computer Science and Engineering (CSE) degree with thesis in 2022 from Northern University of Science and Technology Khulna. His research activities are focused in the area of Machine Learning, Internet of Things and Data Mining.

Ferdib-Al-Islam received B.Sc. in CSE from Bangabandhu Sheikh Mujibur Rahman Science & Technology University in 2018. Now he has been working as a lecturer in Northern University of Business and Technology Khulna. Previously he worked as a Jr. Software Engineer (Internet of Things, R&D) in W3 Engineers Ltd. His research activities are focused in the area of Machine Learning, Deep Learning, Internet of Things, Health Informatics, Data Science, and Computer Vision.

How to cite this paper: Md. Mehedi Rahman Rana, Md. Nasim Adnan, Md. Moradul Siddique, Md. Tahadur Rahman, Ferdib-Al-Islam, "Predicting Education Level of the Farmers' Children of a Developing Country during COVID 19 Using Machine Learning", International Journal of Modern Education and Computer Science(IJMECS), Vol.16, No.6, pp. 94-107, 2024. DOI:10.5815/ijmecs.2024.06.07