Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 44 No. 11
November 2023

An Analysis On School Dropouts During Covid-19 Using Machine Learning Algorithms

Jenifer Jothi Mary A[1] Shantha Mary Joshitta R[2] Jasmine B[3] Princitta R[4]

[1]St. Joseph's College (Autonomous), Tiruchirappalli, Affiliated to Bharathidasan University, Tiruchirappalli, Tamilnadu,.

[2,3] Jayaraj Annapackiam College for Women (Autonomous), Periyakulam, Affiliated to Mother Teresa Women's University, Kodaikanal, Tamilnadu, India.

[4]Solaimalai College of Engineering, Maduari, Affiliated to Anna University, Tamilnadu, India.

Abstract—Education is the dream of every Indian citizen and they are even ready to sell whatever they have for their livelihood for good education. That is why the national education percentage, which was significantly higher when the country gained independence in 1947. But the bitter truth is that education has not escaped the clutches of the Corona monster that brought the world under its control in 2019. Many rural Indian students were deprived of the opportunity to study and were forced to go for a lucrative job to survive and maintain their loved ones. One of the main reasons for this is not only the indefinite closure of educational institutions due to the corona epidemic but also the paralysis of employment, loss of income, rising prices of basic needs and total paralysis of unorganized industries. This research analysis the reason for school dropout during Covid-19 using machine learning algorithms. The current qualitative case study aims to better comprehend the nature of school dropouts in India, Tamilnadu state. The K-means algorithm is used for clustering the dataset and Random Forest algorithm is used for classifying the reason for dropouts.

Keywords— School Dropouts, Covid-19, Machine Learning Algorithms, K-means Algorithm, Random Forest Algorithm, Education

INTRODUCTION

India always considers education as a great asset. From 'Gurugulam' in ancient India to primary schools in rural India shows its priority towards education. In 1951, only 18.33 persons were literate in India whereas, now this count is rocket flying with 74.04 persons in 2011 [1]. It has great strides in refining access to quality education, growing elementary school enrollment, and dropping the number of out-of-school children. No one could have imagined the impact of Covid-19 on the Indian education system. The government has decided to temporarily close the schools to reduce the widespread Corona virus and many schools started their teaching through various new initiatives such as google classroom, Kalvi channel, radio talk, and so on [2]. On the other hand, lots of students have struggled to obtain the gadgets required for online classes. The bitter truth is that many Indian children come to school for their midday meals. As this was stopped by the closure of the schools, many kids become child labor to support their families [3,4]. This paper analyzed the reasons of school dropouts in Tamilnadu state of India during covid-19 using machine learning algorithms. The reason for dropping out is analyzed using the K-means algorithm. Canopy Algorithm is used to select K Centroids and the Random Forest algorithm is applied for the classification of the data collected.

RELATED RESEARCH WORK

The research in [5] analyzed the school dropout of Scheduled Tribal students of Wayanad District, Kerala. There is a high level of school dropouts in the above-said district and found an increase in later 2011-12. The researchers carried out a qualitative analysis to provoke the reasons for the increase in the Paniya tribes' students' dropping out of the Wayanad district. The authors proposed a solid constructivist pedagogy and class-oriented education approach in the tribal areas.

The paper noted in [6] has presented the problems of the existing K-modes algorithm based on rough set theory. Weighted overlap

distance-based K-modes clustering algorithm was proposed by the researchers. The proposed algorithm was applied to obtain a new unsupervised intrusion detection model. Performance of the intrusion detection model has been verified on the KDD Cup 99 dataset and experimental results showcased efficient results.

The authors in [7] evaluated the impact of Higher Education Institutions and faculty support for the students, available academic resources, and social concerns of students during the pandemic. an online national survey was conducted among 11,114 HEIs' students across the Sultanate of Oman. The research model was verified using Regression and factor analysis and the results were presented by the HEIs. But, the availability of academic resources was not of the required level for the students because of their social concerns

In the paper presented in [8], the canopy clustering algorithm was used for pre-processing by Machine Learning with Apache Mahout. The results disclosed that the Canopy pre-processing step has reduced the time of managing the healthcare insurance dataset, and it also sped up the execution time of the k-means by selecting initial centroids for the dataset. Through Hadoop multi-node, the working of the parallel k-means was offered regarding the time needed for enactment and the number of nodes.

The authors of the research in [9] have used a new distance metric to manipulate the similarity between the categorical data points. Dynamic attribute weight and frequency probability were used as new distance metrics to differentiate the data points. A different technique was used in finding out the number of clusters on the data density distribution. Seeds were selected in view of the density distribution of the proposed method to ensure the initial seeds and they reduced the iterations required for the convergent solution.

In research [10], the researchers used four different K-value selection algorithms namely Elbow Method, Silhouette Coefficient, Gap Statistic, and Canopy. They have presented the pseudo-codes and verified them using the Iris data set. As the computational overhead was very large, it was found that that algorithm could not be used for large-scale data sets. The data set was divided by the Canopy algorithm into several overlapping subsets by a pre-determined distance threshold and repeated deletion and aggregation through distance comparisons were used until the original data set was empty.

The research in [11] provided a structured and synoptic overview of the k-means algorithm to overcome shortcomings. The experimental analysis along with a detailed evaluation among different k-means clustering algorithms differentiated the proposed research work from other existing survey papers. Furthermore, it outlined a clear understanding of the k-means algorithm along with its diverse research directions.

COVID-19 AND DROPOUTS IN SCHOOLS

India has been greatly affected by the biggest catastrophe of recent centuries, the COVID-19 Pandemic. As of 20.03.2022, India has registered 4,30,07,841 total cases across India and 5,16,479 deaths are confirmed all over India.[12] In this national crisis, around 15% of students of Delhi government schools have been dropped out of their schools [13]. According to a Right to Education Forum policy, almost ten million girls in India are out of school in India due to the COVID-19 pandemic [14] Many have sent their kids for domestic work to shoulder the economic pressures caused by COVID-19 which affects available study time and access to remote learning opportunities [15]. It hits the Indian Education System miserably as students of the country to be at palisade for an organized education, discipline, peer management, personality development, and particularly, life skills.

MOTIVATION

Covid-19 creates many risks and challenges for the Indian Education System. On one hand, many migrant workers returned home due to the financial hardships caused by this disaster, on the other hand, many kids lost their parents. Moreover, patriarchal social norms such as dowry, child marriage, and restrictions on girls' mobility are also considered major concerns in this pandemic situation. These situations disrupted the education of millions of students in India and produced a mounting demand for a study on school dropouts in this period.

According to the Unified District Information System for Education Plus (UDISE+) 2019-20 report, the dropouts rate in India is more than 17 % at the secondary school level while at upper-primary (VIII – X) and primary level, it is 1.8 % and 1.5 % during the Pandemic period. Most of the dropouts are due to financial crises in the rural family and many girls at the high school level were got married [16]. To find out dropout ratio, the researchers use machine learning algorithms for analyzing the data after collecting it from the districts of southern Tamilnadu, India.

THE PROPOSED METHODOLOGY

The methodology used in this study is depicted in the following Figure 1.

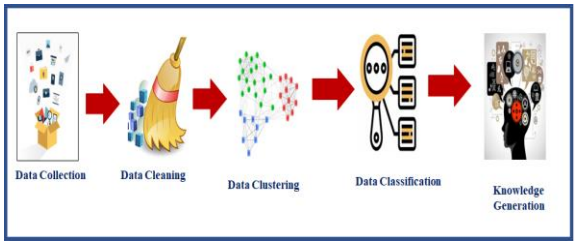

Figure 1. The Proposed Methodology

The procedure of the proposed methodology is as follows:

Step 1. Data collection for the study
Step 2. Collected data is cleaned using a data cleaning mechanism
Step 3. Data clustering using K-means algorithm. K value is calculated using Canopy algorithm [17]
Step 4. Clustered data is classified using Random Forest classification algorithm
Step 5. Knowledge generation from the result

A. Data Collection

The researchers used the survey method to collect the data from the people of Theni district, Tamilnadu, India. Theni is a southern east district of Tamilnadu state in southern India with a total population of 1,245,899 according to the Sensex of 2011 [18]. It has an average literacy rate of 77.26 % with a total literate of 870,080. This district has 479,403 male literates whereas, its female literates are 390,677 [19].

A questionnaire is used to collect data on the school dropouts during Covid-19 pandemic. The items and reasons on the questionnaire have gone through a review committee and a field test with respondents and data users to ensure their relevance [20]. Data are collected for particular units of the target population of more than 5000 from many villages of the district, therefore, sampling is done. The authors perform the data capture activities and follow-up of non-respondents. Contact with respondents is maintained for subsequent follow-up. The data are stored as Microsoft excel files for further analysis.

B. Data Cleaning

After storing the collected data, the processes of cleaning the data is performed. The data cleaning mechanisms such as finding noisy data, missing value imputation, removing duplicates, converting data types were performed and corrected data is saved in the datasheet [21]. After performing data cleaning processes, a total of 4878 data are taken for further analysis. Out of this, 30 % of data are selected as test data and the remaining were taken as train data. Further, clustering is carried out on the train data of the selected dataset. The result of the data cleaning process is depicted in Figure 2.

```
>summary(out_For, explicit = TRUE)
Filter edgeBoostFilter applied to dataset jeni
Call:
edgeBoostFilter(formula = Species ~ ., data = jeni)
Parameters:
m: 15
percent: 0.03
threshold: 0
Results:
Number of removed instances: 8 (3.333333 %)
Number of repaired instances: 0 (0 %)
Additional information:
Highest edge value kept: 0.0665357382344
Explicit indexes for removed instances:
58 78 84 107 120 130 134 139
```

Figure 2. The result after data cleaning

C. Data Clustering

After cleaning the data, the researchers cluster the cleaned data based on some attributes. The famous and most effective machine learning algorithm, K-means, is used to cluster the data. The value of K is decided using another machine learning algorithm called as canopy Algorithm. The K-means algorithm is a

simple iterative clustering algorithm. Using the distance as the metric and given the K classes in the data set, calculate the distance mean, giving the initial centroid, with each class described by the centroid. For a given data set DS containing n multidimensional data points and the category K to be divided, the Euclidean distance is selected as the similarity index and the clustering targets minimize the sum of the squares of the various types; that is, it minimizes $Distance = \sum_{k=1}^{k}\sum_{i=1}^{n}\|(DS_i - C_k)\|^2$ (1)

where k represents K cluster centers, C_k represents the k^{th} center, and DS_i represents the i^{th} point in the data set. The solution to the centroid C_k is as follows:

$$\frac{\partial}{\partial_{C_k}} = \frac{\partial}{\partial_{C_k}}\sum_{k=1}^{k}\sum_{i=1}^{n}(DS_i - C_k)^2 = \sum_{k=1}^{k}\sum_{i=1}^{n}\frac{\partial}{\partial_{C_k}}(DS_i - C_k)^2 \quad (2)$$

$$= \sum_{i=1}^{n} 2(DS_i - C_k)$$

Let us assume that the Eq.(2) is zero, then $C_k = \sum_{i=1}^{n} DS_i$.

The main idea of the algorithm is to randomly extract K sample points from the training dataset as the center of the cluster: Each cluster is grouped with sample points by the nearest main point. Later the main points of the sample points are again grouped as the center point of the cluster. These steps are repeated until the center point is kept unchanging. The choice of the K value is determined by the Canopy machine learning algorithm of [22]. The pseudo-code for the canopy algorithm is as follows.

Algorithm Canopy
Input:dropout=datasets.load_dropout(),X=dropuout.data[:,2:]
Output: k
1. def G1, G2 where G1>G2; delete_X=[]; Canopy_X=[];
2. for P X do
3. d=;
:G2 then
5. Delete_X = [d];
6. Else Canopy_X = [d];
7. Until X=;
8. End;

The K value is identified after applying Canopy Algorithm as shown in Figure 3. After deciding the K value, the K-means clustering algorithm, is implemented to carry out clustering [23]. After applying the K-means clustering algorithm in the given dataset, the No. of Clusters Within groups sum of squares are calculated. After performing the hierarchical clustering, Cluster Dendrogram is developed as shown in Figure 4. After completing the clustering of data, classification is carried out using Random Forest classification algorithm [24].

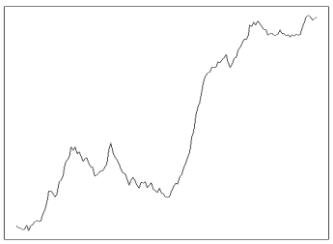

Figure 3. Identification of K value after

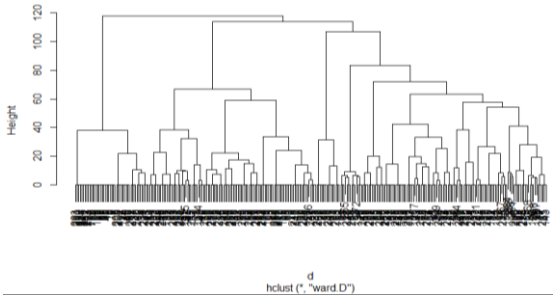

Figure 4. Cluster Dendrogram after applying Canopy Algorithm hierarchical clustering

D. Data Classification

Data classification is carried out using Random Forest Classification algorithm [25]. The first step of Random Forest algorithm is finding error rate. Finding error rates for the individual classes will help us to detect the imbalance in outputs of the given dataset. It is found out that the rate of error is very less in the collected data set and the result is depicted in Figure. 5. The final output of a forest of 500 trees on this data is only 0.54%. It is clearly understood that the proposed dataset has a low overall test set error. Second, we find the number of nodes to each tree. It is found that the predicted number of nodes to tree in the dataset is close to 80 as given in Figure 6. It is very easy to understand that majority of the trees in the

dataset has an average number which is close to 80 nodes.

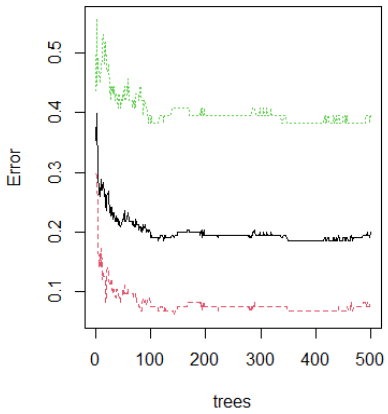

Figure 5. Finding Error Rates

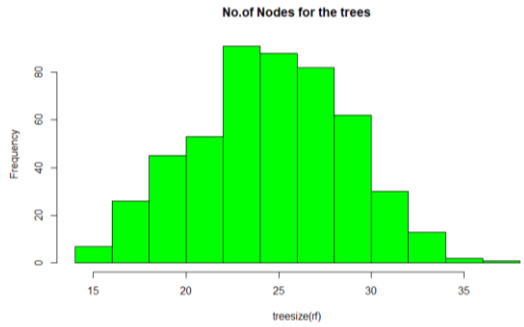

Figure 6 Average number of nodes per Tree

Next, the mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves. The higher the value of mean decrease accuracy or mean. Decrease Gini score, the higher the importance of the variable in the model. Here, we can find out the increase the mean value, so it's more accurate. The final classification result is given Figure 7.

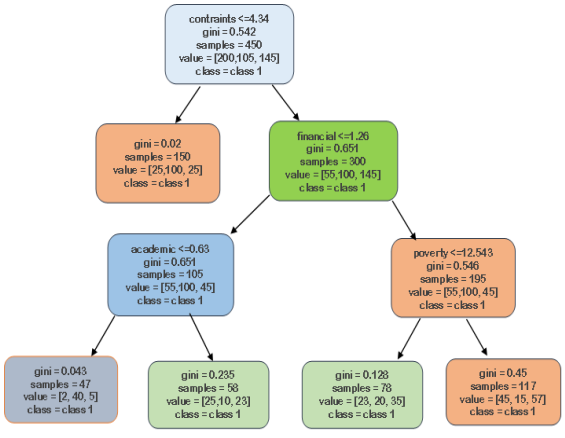

Figure 7. The final classification results

RESULTS AND DISCUSSIONS

The current study explores the challenges that students experienced for leaving the schools and how the pandemic impacted their study experience. The findings revealed that the challenges faced by the students varied in terms of type and extent. Their greatest challenge was linking their learning environment at home, while their least challenge was technological literacy and competency [26]. Based on the students' responses, their challenges were also found to be aggravated by the pandemic, especially in terms of poverty, poor academic performance, early marriage, Alcoholism of the Parents, Negative Attitude towards schooling and Peer Influence [27]. These reasons are explained below.

- **Poverty:** Poverty plays a major role in students' dropout. The research found that it has 54 per cent of impact on the leavening of school students from schools UNESCO also records its concern on the situation analysis on the Effects of and Responses to COVID-19 on the Education Sector in Asia in its report. According to this study, economic contraction and rising unemployment are the major reason for leaving schools [28]. Though schools were started in online mode after October 2020, many fail to buy a smart phone due to their poor financial situation. Issues related to the challenges of accessing digital tools by students and teachers were unaddressed and the researchers found that more than 54 per cent of the students stayed home and not willing or not sent to schools when schools reopen.
- **Poor Academic Performance:** According to World Bank report, the pandemic is amplifying the global learning crisis that already existed. It also increased the percentage of dropouts of primary school-age children in low- and middle-income countries living in learning poverty to 63 percent from 53 percent [29]. Students last their interest in studies and memorizing capacities which leads to poor academic performances. Moreover, the hard looking of the class teachers are not entertained by the students of this generation. They enjoyed

writing test in online mode of just copying the answers from the text books and felt difficult in memorizing the subject and understanding the concepts. 21 per cent of the respondents of the study left the schools because of their poor academic performances.

- **Early Marriage:** One of the devasting effect of Covid-19 is early marriage of girl children. It has impacted the lives and livelihoods of people across rural India. With incomes under threat, families are increasingly likely to get their daughters married before they come of age [30]. Many parents could no longer educate their children on their pandemic-affected meager income. A report entitled as 'COVID-19: A threat to progress against child marriage' expressed that the school closures, increased economic insecurity and job losses, parental deaths due to the Covid-19 and the interruption of support services for families are main reason for girl children's marriage in pandemic period [31]. There is a risk of domestic violence for the girls who remain home at the closure of schools in Covid-19. Moreover, child marriage increases the risk of early and unplanned pregnancy, thereby increasing the risk of maternal complications and mortality. It isolates girls from family and friends and exclude them from participating in their communities, taking a heavy toll on their mental health and well-being, the report added.

CONCLUSION

As the Covid-19 Pandemic draws to a close the education system worldwide, this research made an analysis on school dropouts during covid-19 using machine learning algorithms in the southern state of Tamilnadu, India. Data was collected in Theni district of Tamilnadu using formal questioner for the study and collected data is cleaned using a data cleaning mechanism. These data were clustered using K-means algorithm. K value is calculated using Canopy algorithm and is classified using Random Forest classification algorithm. It was realized and proved that due to the economic and social constrains many students left the schools. This study explicitly proves that Poor academic performance, early marriage, Alcoholism of the Parents, Negative Attitude towards schooling and Peer Influence are also influences the drop out ratio. So, to increase the national student's enrolment ratio in schools, the law makers have to take necessary action at root level. Teachers need to find innovative adopted techniques and practices and should design it in such a way to match it with students' interest and preferred learning styles. As the education system is passing through a highly stressful time, there is an urgent demand for counseling of students on regular interval and to assist them in every possible manner. Moreover, this study gives insight into the perspective of students regarding new mode of teaching-learning during COVID-19 outbreak.

REFERENCES

1. Census of India – 2011, obtained from http://www.dataforall.org/dashboard/censusinfoindia_pca/
2. Cathy Li and Farah Lalani, "The COVID-19 pandemic has changed education forever. This is how", World Economic Forum, 29 Apr 2020
3. Kyungmee Lee, Mik Fanguy, Brett Bligh and Xuefei Sophie Lu, "Adoption of online teaching during the COVID-19 Pandemic: a systematic analysis of changes in university teaching activity", Educational Review 2021, pp. 1-24, 2021, https://doi.org/10.1080/00131911.2021.1978401
4. Dr. Naziya Hasan and Dr. Naved Hassan Khan, "Online Teaching-Learning During Covid-19 Pandemic: Students' Perspective", The Online Journal of Distance Education and e-Learning, October 2020 Volume 8, Issue 4, https://www.researchgate.net/publication/344932812
5. Jobin Joy and M. Srihari, "A Case study on the School dropout Scheduled Tribal students of Wayanad District, Kerala", Research Journal of Educational Sciences, Volume 2, Issue 3, pp. 1-6, 2014.
6. Yawen Dai, Guanghui Yuan, Zhaoyuan Yang , and Bin Wang, "K-Modes Clustering

Algorithm Based on Weighted Overlap Distance and Its Application in Intrusion Detection", Hindawi Scientific Programming, Vol. 2021, Article ID 9972589, 9 pages, https://doi.org/10.1155/2021/9972589.

7. Azzah Al-Maskari, Thurayya Al-Riyami and Siraj K. Kunjumuhammed, "Students academic and social concerns during COVID-19 pandemic", Education and Information Technologies, Vol. 27, pp. 1–21, 2022. https://doi.org/10.1007/s10639-021-10592-2
8. Noor S. Sagheer and Suhad A. Yousif, "Canopy with k-means clustering algorithm for big data analytics", AIP Conference Proceedings 2334, 070006, 2021. https://doi.org/10.1063/5.0042398
9. Premsagar Dandge and Prof. A.K. Gupta, "Efficient Seed and K Value Selection in K-Means Clustering Using Relative Weight and New Distance Metric", International Journal of Innovative Science and Research Technology, Vol. 2, No.6, 2017 , pp. 34 – 39. ISSN No: - 2456 – 2165
10. Chunhui Yuan and Haitao Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm", Multidisciplinary Scientific Journal, Vol. 2, No.16, 2019, pp. 226 – 235. doi:10.3390/j2020016
11. Mohiuddin Ahmed, Raihan Seraj and Syed Mohammed Shamsul Islam, "The k-means Algorithm: A Comprehensive Survey and Performance Evaluation", Electronics, Vol. 9, No.1295, 2020, pp. 1-12.
12. https://www.mygov.in/covid-19/
13. Manish Sisodia, http://timesofindia.indiatimes.com /articleshow/77716857.cms? utm_source= contentofinterest&utm_medium=text&utm _campaign=cppst]
14. DIVYA TRIVEDI, https://frontline.thehindu.com/ dispatches/10-million-girls-at-risk-of-dropping-out-of-school-because-of-the-covid-19-pandemic-says-rte-forum-policy-brief/article33662229.ece], January 25, 2021
15. By Jenelle Babb and Natalie Buchanan, https://thediplomat.com/2020/11/covid-19-leaves-millions-of-girls-at-risk-of-school-dropout-in-asia-pacific/, November 05, 2020
16. ABP News Bureau |Updated : 02 Jul 2021 12:52 PM (IST) https://news.abplive.com/ education/dropout-rate-at-secondary-school-level-in-india-is-more-than-17-claims-study-1466998
17. Khan, S. Canopy approach of image clustering based on camera fingerprints. Multimed Tools Appl (2022). https://doi.org/10.1007/s11042-022-12463-5
18. https://www.census2011.co.in/census/district/46-theni.html
19. "District Census Handbook - Theni" - Retrieved 28 March 2022.
20. A Jenifer Jothi Mary and L Arockiam, A Study on Basic Concepts of Big Data. International Journal of Emerging Trends in Computing and Communication Technology, Volume 1, Issue 3, August 2015.
21. Uma, K. and M. Hanumanthappa, Data Collection Methods and Data Preprocessing Techniques for Healthcare Data Using Data Mining. International Journal of Scientific & Engineering Research, Volume 8, Issue 6, pp. 1131- 1136, 2017
22. Noor S. Sagheer and Suhad A.Yousif, "Canopy with k-means Clustering Algorithm for Big Data Analytics", Proc. Of the Fourth International Conference of Mathematical Sciences 2334, 070006-1–070006-4;
23. R.Suganya, M.P., P.Nandhini, "Algorithms and Challenges in Big Data Clustering", International Journal of Engineering and Techniques, Volume 4, Issue 4, 2018.
24. Lakshmanaprabu, S.K., Shankar, K., Ilayaraja, M. et al., "Random forest for big data classification in the internet of things using optimal features", International Journal of Machine Learning and Cyber. Volume 10, pp. 2609–2618, 2019. https://doi.org/10.1007/s13042-018-00916-z
25. Chaudhary A, Kolhe S, Kamal R (2016) An improved random forest classifier for multi-class classification. Inf Process Agric 3(4):215–222
26. Sagar Chakraborty and Arpan Ghoshal,

"Prediction and Analysis of Dropout Secondary Students in India Using Machine Learning Classification Algorithms", Abstracts of 1st International Conference on Machine Intelligence and System Sciences, 2021. DOI: 10.21467/abstracts.120ISBN:978-81-954993-2-8

27. Vaibhav Singh Makhloga, Kartikay Raheja, Rishabh Jain and Orijit Bhattacharya, "Machine Learning Algorithms to Predict Potential Dropout in High-Schools", EasyChair preprint, July 21, 2020
28. India Case study: Situation Analysis on the Effects of and Responses to COVID-19 on the Education Sector in Asia, October, 2021, pp. 1-58. ISBN 978-92-806-5249-9 (UNICEF)
29. Laura Moscoviz and David K. Evans. 2022. "Learning Loss and Student Dropouts during the COVID-19 Pandemic: A Review of the Evidence Two Years after Schools Shut Down." CGD Working Paper 609, pp.1-28.
30. Madhumita Paul, DownToEarth, 8th March, 2021, https://www.downtoearth.org.in/news/economy/international-women-s-day-10-million-more-girls-at-risk-of-child-marriage-due-to-covid-19-warns-unicef-75813
31. United Nations Children's Fund, "Covid-19 A Threat to Progress Against Child Marriage", 2021, pp. 1-32.