Journal of Computational Social Science
https://doi.org/10.1007/s42001-024-00309-z

RESEARCH ARTICLE

Early student dropout detection in Indian secondary education with special reference to selected districts in Tamil Nadu: a machine learning-based survival analysis approach

Raghul Gandhi Venkatesan[1,2] · **Bagavandas Mappillairaju**[2]

Received: 6 February 2024 / Accepted: 6 July 2024

Abstract

Education is crucial for individual growth and national development. India, with its ambitious School Education Vision 2030, aims to overcome persistent challenges in achieving universal education. This study examines the complex issue of student dropout, specifically focusing on the secondary level in Tamil Nadu, by analyzing the demographic profiles of 846 students. Machine Learning classification approaches such as Logistic Regression, Support Vector Machine, Multi-Layer Perceptron, and Random Forest demonstrate impressive performances, with Random Forest standing out as a powerful tool for accurate prediction. In dropout prediction, survival analysis approaches, specifically the Random Survival Forest (RSF) model, outperform the Weibull model. Through variable importance analysis, age and attendance are found to be significant factors, emphasizing their critical role in predicting dropout events. This study pioneers the integration of survival analysis and machine learning-based classification in the Indian educational context, contributing to the improvement of dropout prediction models. The combined approach enhances the accuracy of dropout prediction and temporal understanding. Despite its cohort-specific focus, the study provides valuable insights for future research and interventions, supporting inclusive education in India by integrating essential characteristics in predictive models.

Keywords Dropout · Determinants · Machine learning · Survival analysis · Secondary education

Extended author information available on the last page of the article

Published online: 08 August 2024

Introduction

Education lays the foundation for individual and national progress. In an attempt to provide the highest level of education to every child and address the existing inadequacies of the educational system, India has launched the ambitious School Education Vision 2030. One of the goals of this strategy is to increase the number of students in schools from 25 crore in 2010 to 30 crore by 2030 [1]. Though school enrollment has shown recent improvement, student dropouts remain a challenging problem. This is especially true at the secondary level, where the school dropout rate is high at 12.61% compared to 1.45% at the primary level in the academic year 2021–2022 [2]. Therefore, India needs to address the complex problem of school dropout irrespective of the increase in enrollment to attain the standards of universal education.

India has the second-highest adolescent population in the world and is struggling with a drop in school attendance [3, 4]. The focused geographical location of our study, Tamil Nadu, has a relatively high dropout rate of 4.5% as a result of a decrease in school strength from upper primary to secondary education [2]. A large percentage of children in this state, Tamil Nadu, drop out of school after finishing Class 10 due to various reasons such as lack of interest in school, health problems, lack of guidance, child marriage, child labor, and academic failures [5]. The implementation of universal education is seriously hampered by this trend.

The transformative potential of secondary education in facilitating upward socioeconomic mobility and conferring various social benefits is significant [6]. Completion of secondary education not only paves the way for higher education but is also increasingly crucial for active participation in the formal labor market [7]. As such, it becomes imperative to understand and address the challenges that lead to a high rate of student dropout, hindering the nation's progress toward the goal of universal education [1].

Student dropout stands out as one of the most complicated and serious issues for both students and schools. The primary causes behind students abandoning their studies are diverse, ranging from factors related to the child, family, society, and the school environment [8, 9]. Furthermore, it is critical to examine temporal information regarding these factors. It is obvious that student dropout is influenced by a wide range of factors, and any investigation in this setting must be capable of dealing with multiple factors and temporal data [10]. Addressing this multifaceted problem while upholding high academic standards is a daunting task for educational institutions [11]. Usually, there were two options for taking action: before the event had transpired and after the "dropout" had occurred. After the event, correcting each problem one by one is a time-consuming procedure; however, if dropouts are predicted in advance, mass retention can be achieved at a lower cost in terms of both time and resources. In the complex social ecosystem of India, prediction or early detection is difficult and requires sophisticated tools to compute and linearize the numerous patterns.

The prediction of early student dropout using data has emerged as a pertinent problem in education, explored through various learning environments in and around the country [12, 13]. Initial attempts at addressing this Student Dropout Prediction

(SDP) problem involved the application of classification algorithms, proving successful on both national and global scales [14–16]. Because of its broad nature, this subject can be addressed from several perspectives, allowing for an extensive choice of analyses. Two important questions emerged from this subject: (1) Who are the most likely dropouts? (2) When is the dropout bound to occur?

Generally, to approach the first question, traditional machine learning models have been employed, and more specifically, a systematic review revealed the most frequently used algorithms [16, 17]. To address the issue of student dropout in the Indian context, Nangia et al. compared classification methods and found that Gradient Boosting (GB) had the highest prediction accuracy. Their study, based on a dataset of 17,359 students from Maharashtra, utilized the student Data Capture Format (DCF) of the Unified District Information System for Education (UDISE) to identify students at risk of dropping out of secondary education [12, 18].

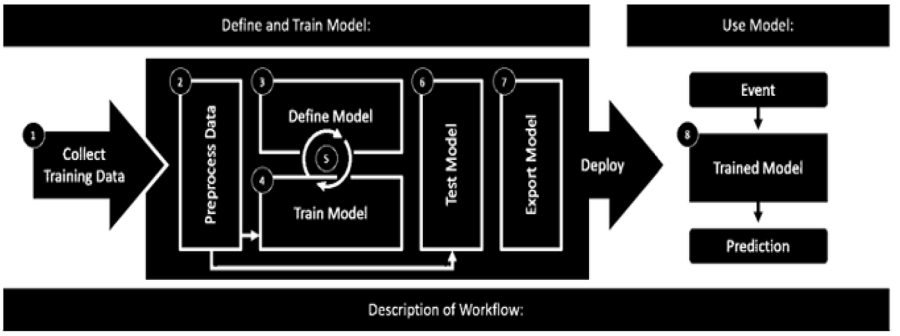
Survival analysis has helped answer the second question. To evaluate dropout patterns, survival analysis, a statistical branch that assesses the influence of predictors on the time until an event, has been used. The impact of specific factors on student dropout and the likelihood of a student completing secondary education are research questions in this context. Several works included the Cox Proportional Hazard (CPH) model in their formulation to handle the SDP problem [10, 19–25]. The Kaplan-Meier estimator and CPH were utilized by Pachas et al. to examine the effects of curriculum design in a computer science programme [24]. Ameri et al. utilized CPH and Time-Dependent Cox (TD-Cox) models to detect early student dropout [10]. Pachas et al. presented a methodology to determine dropout likelihood and timing utilizing a dataset from 655 students at a Peruvian university [22]. However, there is minimal application for machine learning-based survival analysis models such as Conditional Survival Forest (CSF), Random Survival Forest (RSF), and Multi-Task Logistic Regression (MTLR) [26–28]. Only a few studies reported on some of these strategies' applicability to the SDP problem [25, 29, 30].

Many investigations have been conducted in recent decades to determine the primary causes of dropouts in India [9, 12, 31–34]. Irrespective of all the efforts, dropouts remain a problem which necessitates the use of advanced statistical approaches, such as the Survival analysis approach, to detect students at risk of dropping out. This study seeks to bridge this gap by integrating machine learning models including Logistic Regression, Support Vector Machine, Naive Bayes, K-Nearest Neighbour, Multi-Layer Perceptron, and Random Forest, along with survival models such as Weibull, CPH, and RSF. Upon leveraging academic, socio-economic, and temporal information from students in the southern and northern districts of Tamil Nadu, our analysis aims to provide computational support to academic managers. Our goal is to identify at-risk students quickly and to assess the influence of academic and socio-economic factors in the Indian educational setting. In doing so, we contribute to the existing body of knowledge by not only evaluating various algorithms for machine learning and survival analysis methodologies but also by providing a holistic understanding of the factors influencing student dropout in the Indian educational landscape. Applying such statistical approaches to a social problem is an interdisciplinary approach which seeks to offer insights that can aid in the development of targeted interventions to reduce dropout rates and promote inclusive education in India.

Methods

Data

The primary aim of this study is to identify students at risk of dropping out, which requires a detailed and accurate dataset of student characteristics and academic performance. Data were collected from Class 9 students in the Tiruvallur and Madurai districts of Tamil Nadu. We utilized the Student Data Capture Format provided by the UDISE to ensure standardization and reliability of the data [18]. Enrollment data for the academic year 2019–2020 were gathered through a structured format, which included direct submissions from schools to the UDISE system. This method ensures comprehensive coverage and uniformity in the data collected, enabling accurate comparisons and analyses. The dataset comprises 846 entries, each representing an individual student. Key variables collected include demographic information such as age, gender, and social category; socio-economic status such as parent's occupation and education, below poverty line (BPL) status; and educational factors such as disability status, attendance records, and academic performance. The framework of analysis is shown in Fig. 1.



Description of Workflow:

1	2	3	4	5	6	7	8
Collect sample data from UDISE	Data cleaning & Feature selection	1-Machine Learning Classification. 2- Survival Analysis Prediction	Model is trained based on the given data.	Modify the model until it perform well against training data	Test the model with hold out set from collected data.	Model Evaluation using Balanced Accuracy, F1 score, ROC-AUC for Machine Learning. C-index and IBS for Survival Analysis	Expose the trained model. Model will return a prediction accordingly.

Fig. 1 Methodological flowchart of dropout classification and prediction

Feature analysis

For the prediction model to be effective, the choice of associated factors is essential. To determine how strongly variables were correlated, Cramer's V matrix was used

[35]. This statistical measure is particularly suited for categorical data, providing a robust assessment of the strength of association between variables. The Cramer's V value results range from 0 to 1, with higher values indicating stronger associations between variables. To calculate Cramer's V values, contingency tables were created for each pair of variables under consideration to perform Chi-square tests, and the obtained statistics were incorporated [36]. The Cramer's V value matrix was then investigated to discover variables having significant relationships with dropout status.

Cramer's V is calculated using the following formula:

$$V = \sqrt{\frac{\frac{\chi^2}{N}}{min\,(k-1, r-1)}}$$

where:

- $\chi 2$ is the chi-squared statistic from the chi-squared test of independence.
- N is the total number of observations (the sample size).
- k is the number of columns in the contingency table.
- r is the number of rows in the contingency table.

The selected variables were then listed and briefly described, highlighting their relevance to the study's objective. It was ensured that the chosen variables represented a diverse set of characteristics, including demographic details, socio-economic factors, attendance, and academic performance, to construct a comprehensive predictive model for potential dropouts among class 9 students.

Data preprocessing

The dataset used for this study exhibits a favorable condition with no missing data, and all variables except age are categorical. We used one-hot encoding and ordinal encoding to ensure that the categorical nature of the variables is preserved while allowing for their integration into the model training process [37].

Following feature selection and encoding, the dataset was separated into training and testing sets. This divide, which was frequently done in a stratified way, permitted successful model training and evaluation. The training set was used to train the model, while the testing set was used to assess the model's performance on unknown data. This stage ensured the model's ability to generalize to new instances and allowed for an unbiased evaluation of its prediction abilities.

Machine learning methods

For the classification task, we employed Logistic Regression, Random Forest, K-Nearest Neighbour, Support Vector Machine, Naïve Bayes, and Multi-Layer Perceptron. The choice of the algorithm was based on its suitability for handling cat-

egorical data and its proven effectiveness in similar educational prediction contexts [12, 16].

Logistic Regression predicts the probability of y, $p(y = 1|X)$ based on the logistic function [38]:

$$p(y = 1|X) = \frac{1}{1 + \exp\left(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p\right)},$$

Where y is the dropout ("yes" or "no") and X represents the vector of input features such as age, parents' occupation, etc., and β_p are the learned coefficients.

Random Forest involves an ensemble of decision trees, each providing a vote for the most likely class. The final prediction is determined by majority voting [39]:

$$y = mode(\widehat{y_1}, \widehat{y_2} \ldots \widehat{y_t}),$$

where y is the dropout ("yes" or "no") and each $\widehat{y_t}$ is the prediction of dropout ("yes" or "no") from the *t*-th tree based on your variables such as age, attendance, academic performance, etc.

K-Nearest Neighbors classifies a new instance based on the majority label among the nearest *k* points [40]:

$$y = mode(k\ nearest\ y_i),$$

where y is the dropout ("yes" or "no") and nearest neighbors are calculated based on the similarity in variables such as social category, disability, etc.

Support Vector Machine constructs a hyperplane in high-dimensional space that best separates the classes, with the decision rule given by [41]:

$$y = sign(w.X + b),$$

where y is the dropout ("yes" or "no") and X includes the predictors such as BPL status, religion, etc.

Naïve Bayes uses Bayes' theorem, assuming independence among predictors, to calculate the probability of dropout [42]:

$$p(y|X) = \frac{p(X|y)\, p(y)}{p(X)},$$

where $_e$ ach $p(y|X)$ reflects the probability of observing the predictors given dropout status.

Multi-Layer Perceptron, a type of neural network, uses layers of neurons, each applying a nonlinear activation function to the weighted sum of inputs [43]:

$$y = \sigma\,(W.X + b),$$

where y is the dropout ("yes" or "no") and X encompasses all the study variables, transformed appropriately for input into the network.

The dataset following feature selection and encoding in the preprocessing phase was utilized to train the classification models. The training set was input into each algorithm, and model parameters were systematically optimized through grid search to maximize predictive performance. Additionally, k-fold cross-validation was employed to ensure robust training and to minimize the risk of overfitting [25].

The classification model was evaluated on the independent testing set to assess its performance in predicting potential dropouts. In general, the ability of the optimal model to generalize data across all types while providing the most accurate findings was phenomenal. However, in the presence of skewed classes, as in our case, the model gets biased towards the most prevalent class, producing overly optimistic results. Balanced accuracy was found to be the better statistic for evaluating the performance of imbalanced datasets [44].

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{True\ Positive}{True\ Positive + False\ Negative} + \frac{True\ Negative}{True\ Negative + False\ Positive}\right)$$

In addition, the F1 score is the harmonic mean of Precision and Recall and is a better measure than accuracy for cases where you have an uneven class distribution [45].

$$F1\ score = 2\left(\frac{Precision * Recall}{Precision + Recall}\right)$$

and Area Under the Curve (AUC) is a performance measurement for binary classifiers that were calculated to assess the model's efficacy. It is a probability curve that plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. AUC is calculated by using the below formula [46]:

$$TPR = \frac{Number\ of\ True\ Positive}{Number\ of\ True\ Positive + Number\ of\ False\ Negative}$$

$$FPR = \frac{Number\ of\ False\ Positive}{Number\ of\ False\ Positive + Number\ of\ True\ Negative}$$

Survival analysis methods

Survival analysis was carried out using Weibull, CPH, and RSF. These techniques helped us model the time until a possible dropout and were appropriate for time-to-event data. Maximum likelihood estimation (MLE) was used to estimate the parameters of the Weibull survival model [47]. With the help of this approach, we can effectively model the hazard function by obtaining reliable estimates for the shape and scale parameters. Using relevant statistical methods such as the log-likelihood ratio test or the Akaike Information Criterion (AIC), the goodness-of-fit of the Weibull model was assessed. These assessments provided evidence of the models' adequacy in representing the observed survival data [48, 49].

The model is defined by its hazard function [50]:

$$h\left(t\right) = \alpha\, \lambda\ \left(t\right) t^{\alpha - 1}$$

where $\lambda\ \left(t\right) = \ exp\left(\beta_{\ 0} + \beta_{\ 1} x_1 + \beta_{\ 2} x_2 + \ldots\ + \beta_{\ p} x_p\right)$

- α (shape parameter): Determines the nature of the hazard over time. If α>1, the hazard increases over time; if α<1, it decreases.
- λ (scale parameter): Affects the time scale of the data.
- t: Time at which the event of interest (e.g., dropout) occurs.

The survival function, which describes the probability of surviving past time t, can be derived from the hazard function and is given by:

$$S\left(t\right) = exp(-(\lambda\ \left(t\right) t)^{\alpha})$$

The CPH model was employed in this study to investigate the influence of covariates on the risk of dropout. This model is based on the assumption that the hazard rate, which signifies the risk of dropout at any given time, can be deconstructed into two fundamental components: a baseline hazard function that varies with time and a set of covariates. The model's hazard function is expressed as [50]:

$$h\left(t|X\right) = h_0\left(t\right) exp\left(\beta_{\ 1} x_1 + \beta_{\ 2} x_2 + \ldots\ + \beta_{\ p} x_p\right)$$

- $h0(t)$: Baseline hazard function, which is the hazard when all covariates are zero.
- $x_1,\ x_2 \ldots\ldots\ x_p$:Represents the covariates or predictors included in the model.
- $\beta 1$, $\beta 2$, …, βp: Coefficients representing the log-relative hazard of the covariates.

The model estimates coefficients associated with each covariate, revealing the direction and magnitude of their impact on the hazard of dropout. A positive coefficient indicates an increased hazard, suggesting a higher risk of dropout for individuals with higher values of the covariate. Conversely, a negative coefficient implies a reduced hazard, indicating a lower risk of dropout [10, 25].

As used in previous studies, the RSF machine learning technique was utilized to capture complex linkages and interactions within the dropout data [25, 26]. With the use of an ensemble learning technique, a forest of decision trees was created, each trained on a bootstrap sample of the data and taking into account the outcomes of time and censoring status. It also offers insights into variable importance, which helps identify the critical components impacting dropout. With this method, intricate survival patterns within the dataset were captured with flexibility, making it a strong substitute for parametric models. The survival function is given by:

$$S\left(t|X\right) = 1 - \frac{1}{B}\sum\nolimits_{b=1}^{B} I(t_b \leq\ t, \delta_{\,b} = 1)$$

where:

- B : Number of bootstrap samples.
- t_b : Dropout time or censoring for the b-th bootstrap sample.
- $\delta_{\,b}$: Indicator variable, where 1 indicates dropout and 0 indicates censoring.

All these models were trained using the selected features and the time-to-event data. The training process involved estimating coefficients that characterize the relationship between the features and the hazard function, considering censored data points where the dropout event had not occurred by the end of the observation period.

The concordance index (C-index) was used to evaluate the predicted accuracy of all survival models. This metric assesses the model's ability to rank individuals based on expected survival times. A higher C-index suggests greater discriminating power [51].

$$\text{C-index} = \frac{\sum_{i<j} I\left(T_i < T_j\right) I\left(\widehat{S}\left(T_i\right) < \widehat{S}\left(T_j\right)\right)}{\sum_{i<j} I\left(T_i < T_j\right)}$$

where:

- $I\left(T_i < T_j\right)$ is an indicator function that returns 1 if the actual survival time T_i of the subject *i* is less than the survival time T_j of subject *j*. This function checks if the observed time until the event for subject *i* is indeed less than that for subject *j*.
- $I\left(\widehat{S}\left(T_i\right) < \widehat{S}\left(T_j\right)\right)$ is another indicator function that returns 1 if the predicted survival probability $\widehat{S}\left(T_i\right)$ (i.e., the model's prediction of survival at time T_i) for subject *i* is less than the predicted survival probability $\widehat{S}\left(T_j\right)$ for subject *j*. This evaluates if the model predicts that subject *i* is more likely to experience the event earlier than subject *j*, aligning with their actual times.

The Integrated Brier Score (IBS) was used to assess the overall accuracy of survival probabilities. This statistic takes into account both the projected survival probability and the actual outcomes over time. Lower IBS values indicate higher model calibration and accuracy [52, 53].

$$\mathrm{IBS}\,(t) = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left[\left(Y_i\,(t) - \widehat{S}\,(t|X_i)\right)^2\right]$$

where:

- n is the number of students in the study.
- $Y_i\,(t)$ is the observed status of the *i*-th individual at time *t*. It is typically 1 if the event has occurred before or at time *t*.
- $\widehat{S}\,(t|X_i)$ is the predicted probability of surviving until time *t* for the *i*-th individual, given their covariates X_i.

Statistical analysis

A complete and efficient analysis was carried out in this study through the strategic integration of R and Python tools, each chosen for its specialized capabilities. R was crucial in the early experimental phase, offering a stable framework for generating descriptive statistics and parametric survival modelling with the survival and flexsurvreg packages [54, 55]. Python, which is well-known for its adaptability, was used for advanced machine learning classification and survival analysis. Python allows the construction of algorithms ranging from Logistic Regression and Random Forest to complicated survival models such as Weibull and RSF by leveraging packages such as Scikit-Learn, Scikit-Survival, and Lifelines [56–58]. The use of both R and Python not only optimized each analytical phase but also ensured the easy integration of outputs, encouraging a comprehensive understanding of the factors impacting future dropouts. Ethical factors such as data privacy guidelines and student information confidentiality were upheld to perform a rigorous analysis of potential dropout patterns among class 9 students.

Results

Our study investigated the demographic profiles of 846 students, revealing pertinent insights into the factors associated with academic persistence. Of the total cohort, 790 students (93.4%) were classified as non-dropouts, while 56 students (6.6%) were identified as dropouts. The age distribution among non-dropouts indicated a majority in the 16 and 17-year age groups, constituting 45.8% and 51.5%, respectively. Conversely, the dropout group exhibited a notable shift, with 75% falling in the 18- and 19-year age categories. Gender distribution unveiled a higher representation of males among both non-dropouts (58.5%) and dropouts (62.5%).

Parental education levels depicted an interesting pattern, with a significant percentage of fathers and mothers having completed secondary education. While 48.1% of fathers of non-dropouts had secondary education, only 5.4% of dropout fathers fell into this category. A similar trend was observed in mothers' education, with 7.1% of dropout mothers having completed secondary education. The occupational status of parents showcased diverse employment scenarios. Government or private service

and business were prominent among fathers, whereas mothers engaged in a spectrum of occupations including government/private service, business, daily wages, and unemployment.

Socioeconomic status, gauged by Below Poverty Line (BPL) status, disclosed a higher representation of BPL among dropouts (98.2%), implying a potential link between economic challenges and academic discontinuation. The community distribution indicated the majority belonged to the Most Backward Class (MBC) for both non-dropouts (81.3%) and dropouts (78.6%). Religion remained predominantly Hindu (99.4%) among non-dropouts, while among dropouts, a significant proportion identified as Christian (3.6%).

The medium of instruction appeared to influence academic persistence, as 66.1% of dropouts received education in English compared to 37.1% of non-dropouts. Notably, a proportion of non-dropouts had disabilities (0.4%), underlining the need for tailored support. Attendance and academic performance revealed distinctions, with a lower attendance rate and a higher percentage of below-average academic performance among dropouts. These results highlight the significance of focused interventions for the at-risk student population listed in Table 1 and provide insight into the complex interactions between sociodemographic determinants influencing academic achievement.

After analyzing demographic characteristics, we further explored the relationships between these variables and dropout rates using Cramer's V matrix. This statistical tool measures the strength of association between categorical variables, providing insights into which factors are most predictive of dropout risks. Table 2 below presents these associations, highlighting key variables with a substantial impact on student dropout rates.

Among the variables examined, age stands out with a Cramer's V of 0.7189, highlighting its significant influence on dropout patterns, a finding that echoes the literature suggesting older students face increased dropout risks due to external pressures such as employment or familial responsibilities. Attendance is another critical predictor with a robust Cramer's V of 0.4595, supporting the theory that consistent school attendance is pivotal for academic persistence. Academic performance also shows a noteworthy association (Cramer's $V = 0.2107$), underlining its role in dropout occurrences.

Conversely, variables such as gender, religion, community, and disability show little to no association with dropout likelihood, indicated by low or zero Cramer's V values, suggesting limited predictive power in our model. Socioeconomic factors like being below the poverty line (BPL) have a mild association (Cramer's $V = 0.1026$), hinting at economic challenges influencing dropout rates. Parental education and occupation also demonstrate moderate associations with dropout, with the father's and mother's education levels showing Cramer's V values of 0.2059 and 0.1998, respectively, indicating the influence of family background on educational continuity.

This detailed examination not only identifies key predictors such as age, attendance, and academic performance but also underscores the need to consider socioeconomic factors in constructing predictive models of dropout. These insights are crucial for developing targeted interventions that address the multifaceted nature of dropout risks and ensure resources are allocated efficiently to support at-risk students.

Table 1 Demographic characteristics of students by dropout status

Variables	Non-Dropout ($n=790$)	Dropout ($n=56$)
Age		
16	362 (45.8%)	0 (0%)
17	407 (51.5%)	14 (25%)
18	19 (2.4%)	20 (35.7%)
>18	2 (0.3%)	22 (39.3%)
Gender		
Male	462 (58.5%)	35 (62.5%)
Female	328 (41.5%)	21 (37.5%)
Father's Education		
Primary	410 (51.9%)	53 (94.6%)
Secondary	380 (48.1%)	3 (5.4%)
Mother's Education		
Primary	405 (51.3%)	52 (92.9%)
Secondary	385 (48.7%)	4 (7.1%)
Father's Occupation		
Government/Private	250 (31.6%)	2 (3.6%)
Business	250 (31.6%)	21 (37.5%)
Daily wages	290 (36.8%)	33 (58.9%)
Mother's Occupation		
Government/Private	168 (21.3%)	1 (1.7%)
Business	194 (24.6%)	17 (30.4%)
Daily wages	215 (27.2%)	24 (42.9%)
Un-employed	213 (26.9%)	14 (25%)
Below Poverty Line		
Yes	635 (80.4%)	55 (98.2%)
No	155 (19.6%)	1 (1.8%)
Community		
BC	54 (6.8%)	4 (7.1%)
MBC	642 (81.3%)	44 (78.6%)
SC	94 (11.9%)	8 (14.3%)
Religion		
Hindu	785 (99.4%)	54 (96.4%)
Christian	5 (0.6%)	2 (3.6%)
Medium of Instruction		
English	293 (37.1%)	37 (66.1%)
Tamil	497 (62.9%)	19 (33.9%)
Disability		
Yes	3 (0.4%)	-
No	787 (99.6%)	56 (100%)
Attendance		
Low	65 (8.2%)	30 (53.6%)
Medium	55 (7%)	20 (35.7%)
High	670 (84.8%)	6 (10.7%)
Academics		
Below Average	263 (33.3%)	41 (73%)
Average	266 (33.7%)	13 (23.2%)
Above Average	261 (33%)	2 (3.5%)

Table 2 Association between covariates and dropout using Cramer's V test

Covariates	Cramer's V value
Father's Occupation	0.1505
Father's Education	0.2059
Mother's Occupation	0.1203
Mother's Education	0.1998
Age	0.7189
Gender	0
Religion	0.0421
Medium	0.1387
Community	0
Disability	0
Below Poverty Line	0.1026
Attendance	0.4595
Academics	0.2107

Table 3 Performance metrics of classification techniques for dropout prediction

Techniques	Balanced accuracy	F1 score	ROC-AUC
Logistic Regression	0.891	0.878	0.891
Random Forest	0.934	0.930	0.934
K-Nearest Neighbour	0.846	0.799	0.846
Support Vector Machine	0.891	0.878	0.891
Naïve Bayes	0.869	0.850	0.869
Multi-Layer Perceptron	0.891	0.878	0.891

Table 4 Performance metrics of survival techniques for dropout prediction

Metrics	Weibull	Cox	Random survival forest
C-index	0.931	0.864	0.977
IBS	0.161	0.163	0.151

Table 3 provides a comprehensive overview of the performance metrics for various classification techniques employed to identify individuals at risk of dropout. Balanced accuracy, F1 score, and ROC-AUC were utilized as evaluation measures. The results reveal notable differences in performance metrics. Logistic Regression, Support Vector Machine, and Multi-Layer Perceptron demonstrate commendable efficacy with a balanced accuracy of 0.891, an F1 score of 0.878, and a ROC-AUC of 0.891, showcasing a balanced trade-off between sensitivity and specificity. Random Forest stands out as the top-performing technique with a balanced accuracy of 0.934, an F1 score of 0.930, and a ROC-AUC of 0.934, indicating its superior ability to accurately predict both dropout and non-dropout students. Overall, these results highlight how each technique performs differently in the setting of dropout prediction, with Random Forest emerging as the most robust classifier in this scenario. Detailed confusion matrices and ROC curves are provided in Appendix A.

Continuing our investigation into dropout prediction, we delved into the performance of survival techniques. The Concordance Index (C-index) and Integrated Brier Score (IBS) served as pivotal metrics for evaluating the effectiveness of these techniques in predicting the time of dropout, as shown in Table 4. The Weibull parametric

model exhibited commendable performance with a C-index of 0.931, underscoring its ability to accurately rank the survival times of students. However, the Integrated Brier Score of 0.161 suggested a moderate level of prediction error, indicating some variability between predicted and observed survival times.

On a positive note, the Random Survival Forest (RSF) model demonstrated exceptional performance, surpassing the Weibull model with an impressive C-index of 0.977. This elevated C-index highlights the superior discriminatory power of the RSF model in ranking survival times, indicating a more precise prediction of the time when students might drop out. Furthermore, the RSF model exhibited a slightly improved IBS of 0.151, suggesting a reduction in prediction error compared to the Weibull model.

Figure 2 displays an average survival function estimated from a Random Survival Forest model. The plot suggests that the survival probability starts high and gradually decreases in a stepwise fashion as time progresses. The steps in the survival curve correspond to the time points where the cumulative hazard increases due to the occurrence of the dropout. The relatively smooth and gradual decline without abrupt drops suggests that the dropouts are spread out over the timeline rather than clustered.

Examining the variable importance scores, age appears to be an influential factor with a high importance score. This highlights the significant impact of age on predicting the timing of dropout events, aligning with our earlier findings that highlighted age as a crucial determinant associated with dropout patterns. Attendance is also a significant predictor with an importance score of 0.01. This reinforces our earlier observation that low attendance rates significantly contribute to the risk of dropout. The significance of attendance emphasizes its role as a key indicator for identifying students at risk of discontinuation.

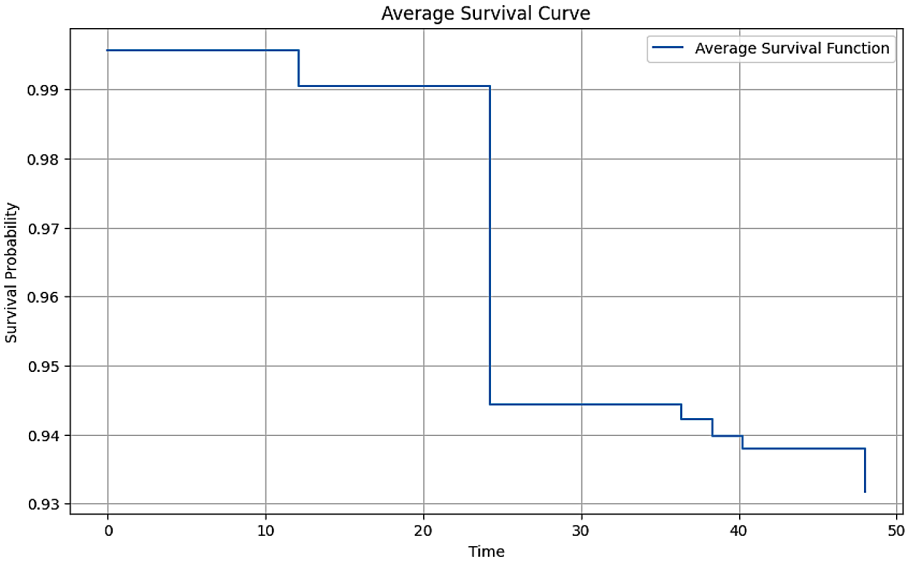


Fig. 2 Average survival probability over time from random survival forest analysis

Other variables, including parental education, parental occupation, BPL, medium, and academic performance, exhibit comparatively lower importance scores. While these variables may not carry as much weight individually, their collective contribution cannot be discounted. It suggests that in combination, these factors play a role in shaping the dropout landscape, highlighting the complex interplay of socio-demographic variables in influencing academic persistence.

In summary, the RSF model variable importance results provide a more detailed view of the factors that contribute to dropout prediction, as shown in Fig. 3. Age and attendance appear as significant factors, emphasizing their critical role in identifying students at risk of dropping out. The aggregate impact of other variables, while individually less prominent, highlights the multidimensional nature of dropout dynamics, emphasizing the importance of a holistic approach to intervention and support techniques for at-risk students.

Discussion

Education is an essential component of human growth and development, influencing individuals and contributing to nations' development. Despite notable strides in enrollment rates, India is still struggling with the complicated issue of student dropout, especially at the secondary level [1]. The persistent issue of school dropout poses a problem to the realization of India's School Education Vision 2030, which envisions universal education for all. According to this objective, this study aims to determine the likelihood of dropouts by carefully analysing demographic informa-

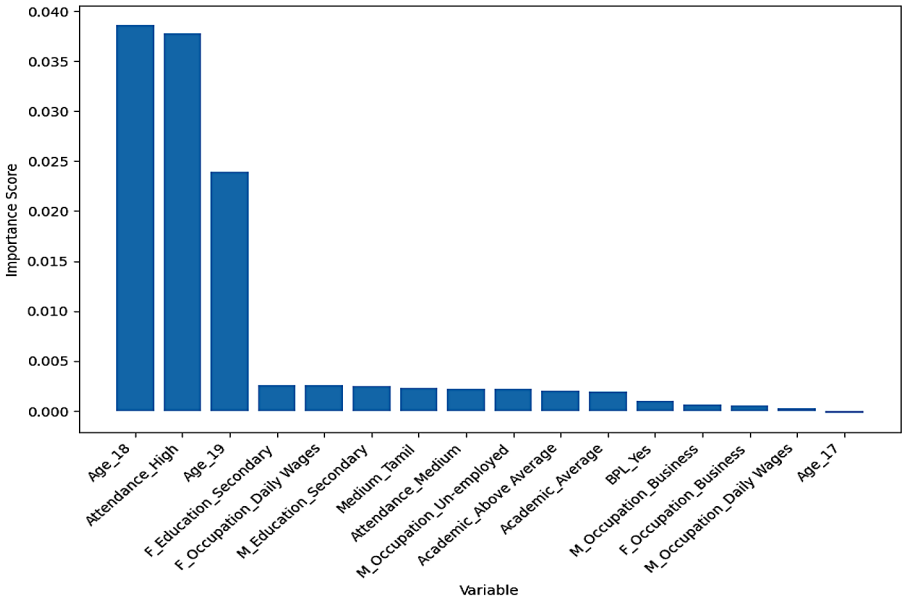


Fig. 3 Variable importance analysis from random survival forest

tion obtained from the various schools in the districts of Tiruvallur and Madurai, Tamil Nadu.

The examination of demographic characteristics of 846 students gave interesting insights into the factors associated with academic persistence. In contrast to non-dropouts, the majority of dropouts were between 18 and 19, implying that age is a key predictor determining dropout trends. Parental education levels and occupational statuses have distinctive patterns, emphasizing the relevance of family backgrounds in academic persistence. Socio-economic characteristics, notably Below Poverty Line (BPL) status, are found to have a significant relationship with academic discontinuation, highlighting economic challenges as potential contributors.

Beyond demographic findings, the predictive modelling phase sought to build a strong framework for identifying future dropouts. The insights derived from Cramer's V matrix are instrumental in identifying the key factors influencing student dropout rates. Notably, the variable 'Age' exhibits a Cramer's V of 0.7189, indicating a strong association with dropout likelihood. The robust relationship shown in the literature between age and dropout trends is consistent with the idea that students are particularly vulnerable during important transitional periods. This confirms that age is a strong and universal predictor across diverse educational contexts [34, 59, 60].

Additionally, attendance shows a Cramer's V of 0.4595, emphasizing its significant role in educational continuity. Consistent with the arguments presented by Gubbels et al., (2019) [8], our results suggest that monitoring attendance rigorously could serve as an early indicator for identifying students at risk of dropping out. Educational institutions could benefit from implementing targeted interventions focused on improving attendance among these students.

Our identification of academic performance as the major indicator is consistent with findings from previous research [12, 31, 34, 61]. Education research frequently examines the relationship between academic performance and attendance rates as predictors of dropout risk. To add a quantitative aspect to this well-established qualitative relationship, our study specifically quantified the strength of these associations using Cramer's V values.

Findings from previous studies are consistent with the socioeconomic dynamics found in our study, which include parental education levels and BPL status. In dropout prediction models, these characteristics consistently emerge as influential contributors. The validation of these socioeconomic indicators as predictors emphasizes their importance across varied geographically distinct educational settings, emphasizing the need for targeted interventions addressing economic inequality to successfully minimize dropout risks [31–34, 60, 62].

Contrastingly, variables such as 'Gender' and 'Disability' showed minimal associations with dropout rates, with Cramer's V values of 0, indicating these factors may not significantly predict dropout within our study context. This highlights the importance of focusing on the more predictive variables when designing dropout prevention strategies. The practical implications of these findings are substantial. Integrating the statistically significant predictors from our Cramer's V analysis can enable more effective prediction and prevention of dropouts. This approach not only optimizes resource allocation but also supports timely interventions.

Machine Learning classification approaches such as Logistic Regression, Support Vector Machine, Multi-Layer Perceptron, and Random Forest showcased impressive performances. A study by Nangia et al. explored dropout prediction using machine learning in a similar educational context [12]. In our study, Logistic Regression demonstrated exceptional metrics including a balanced accuracy of 0.891, an F1 score of 0.878, and an ROC-AUC of 0.891. These metrics surpass the performance reported by Nangia et al., where the weighted accuracy of Logistic Regression was found to be 0.63.

The strong metrics of Random Forest, with a balanced accuracy of 0.934, an F1 score of 0.930, and a ROC-AUC of 0.934, establish it as a powerful tool for accurate prediction. These findings not only validate the effectiveness of machine learning in educational prediction but also establish the better performance of certain algorithms in dealing with the multifaceted nature of dropout patterns, which aligns with the study done by Emanuel Marques Queiroga, indicating the robustness of Random Forest in diverse educational contexts [63].

Moving on to survival analysis, the Weibull parametric model performed well with a c-index of 0.931, indicating an accurate ranking of student survival times. However, the RSF model performed better than the Weibull model, as seen by its slightly better Integrated Brier Score of 0.151 and higher c-index of 0.977. Age and Attendance were found to be particularly significant factors based on the variable importance analysis conducted using the RSF model. With high importance scores, these variables demonstrated their significant impact on predicting dropout events. Other variables, while individually exhibiting relatively low importance scores, collectively contributed to shaping the dropout landscape, emphasizing the complex interplay of socio-demographic variables.

In comparison with earlier research which stated that RSF was ineffective in dropout prediction, our result showed a significant difference in the RSF models' performance. Our study showed that RSF performs better than the semi-parametric and parametric approaches [25]. In the context of predicting student dropouts, the RSF model is particularly advantageous due to its ability to handle censored data, a common challenge in dropout analysis where the event (dropout) may not have occurred yet for all subjects during the observation period. This model's non-parametric nature allows it to adapt flexibly to the underlying data structure without assuming a specific statistical distribution for survival times [26]. Additionally, RSF can account for a mixture of categorical and continuous variables and evaluate the importance of each predictor in the presence of censoring, which is vital for educational data that typically features a mix of demographic and academic performance indicators. Furthermore, RSF's ensemble approach helps improve prediction accuracy and robustness by aggregating results from multiple decision trees, reducing the variance and avoiding overfitting. This makes it particularly effective in educational settings where dropout predictors can interact in complex ways that simpler models might fail to capture. This discrepancy highlights the sensitivity of model performance to the specific context and characteristics of the dataset. Differing demographic profiles, socioeconomic conditions, or other contextual characteristics within the datasets may be the cause of the different results. Furthermore, the quality and quantity of data that is accessible affects how effective machine learning models are, particularly compli-

cated ones like RSF. Our study might have benefitted from a dataset that aligns more closely with the strengths of RSF, leading to its superior performance.

Furthermore, a useful finding from this study and Gutierrez-Pachas et al. is the continuous observation that machine-learning survival models outperform classical survival models. The comparable results of both studies point to a larger trend suggesting that machine learning techniques, which can capture complicated patterns and nonlinear relationships, are better suited to handle the complexities involved in dropout prediction. The evidence supporting the effectiveness of advanced modelling techniques in the field of educational predictive analytics is strengthened by our study's alignment with Gutierrez-Pachas's findings about the higher performance of machine learning survival models over conventional ones [25].

Although previous research has mostly concentrated on survival analysis or machine learning-based classification separately, our work is the first in the Indian context to combine the two approaches in a school setting [12, 25, 64]. The predicted accuracy and temporal knowledge of dropout occurrences are improved by this novel method, which combines the advantages of survival models and classification algorithms. This methodological development advances the field of educational predictive analytics.

While this study focuses on a specific cohort, it is important to consider the generalizability of the findings. Generalizability refers to the extent to which the results of a study can be applied to other settings, populations, or times. In our case, the cohort under study is drawn from the districts of Tiruvallur and Madurai, Tamil Nadu. Although this provides a rich context for our findings, it also limits the direct applicability to other populations without further validation. To enhance the understanding of how these findings might generalize to broader populations, future research should aim to replicate this study in diverse settings. For instance, similar studies could be conducted in different regions or among populations with varying socioeconomic statuses. This would help to identify if the observed patterns hold across different contexts and enhance the robustness of the conclusions.

Dropout rates can be influenced by a variety of factors, including but not limited to regional economic conditions, educational policies, and cultural attitudes towards education [65]. For example, regions with higher socioeconomic instability may experience higher dropout rates due to financial pressures on students and their families [66]. Variations in the quality and accessibility of educational facilities can significantly impact student retention [67]. Regions with well-funded schools and ample resources tend to have lower dropout rates. Cultural attitudes towards education and gender roles can also play a crucial role [68]. By considering these regional variations, we can better understand the nuanced interplay between context-specific factors and educational outcomes. Future research could incorporate these variables to provide a more comprehensive analysis of dropout rates and their determinants across different regions.

Another significant limitation is the potential for biases in data collection. The demographic information and other data used in this study were obtained from the DCF from various schools, similar to UDISE. While UDISE is a comprehensive and widely used data source, it may still be subject to reporting biases or inaccuracies [69]. These biases could impact the reliability of the study's findings and the robustness of the predictive models. Recognizing this, future studies should con-

sider employing multiple data sources and triangulating data to enhance accuracy and reliability.

Measurement errors in variables such as attendance and academic performance, which are critical predictors in our models, can also pose limitations. Inaccurate or inconsistent recording of these variables can lead to erroneous conclusions about their impact on dropout rates. It is crucial to ensure rigorous data collection methods and validation checks to minimize such errors [70].

The cross-sectional nature of the data limits the ability to draw causal inferences. While the study identifies associations between various factors and dropout rates, it cannot definitively establish causality. Longitudinal studies would be beneficial in tracking students over time to better understand the causal pathways leading to dropout events [71].

These limitations have important implications for policy and practice. Policymakers and educators should be cautious in generalizing the findings beyond the studied cohort and should consider regional variations and potential data biases when designing interventions. There is a need for tailored, context-specific strategies that address the unique challenges faced by different regions and demographic groups. Additionally, efforts should be made to improve data collection processes to ensure more accurate and comprehensive data, which can lead to more reliable research outcomes and better-informed policy decisions.

Conclusion

Our comprehensive study on student dropout rates in Indian secondary education, with special reference to selected districts in Tamil Nadu, underscores the urgency and complexity of this issue, which is pivotal to achieving the goals outlined in India's School Education Vision 2030. By identifying specific patterns and key determinants of dropout, our research highlights critical areas needing immediate and targeted interventions. Our investigation into predictive modeling techniques, such as Logistic Regression, Support Vector Machine, Multi-Layer Perceptron, and Random Forest, has demonstrated substantial accuracy in identifying at-risk students. These models have proven their utility in educational settings, offering robust tools that can be integrated into future educational analytics systems. A novel aspect of our research was the combination of survival analysis with machine learning approaches, specifically through the use of the Random Survival Forest model. This model, in particular, outperformed traditional survival analysis models, showing superior predictive precision. The integration of these methodologies enhances our ability to accurately predict dropout times, providing educational administrators and policymakers with a powerful tool to implement timely and effective interventions.

The study identifies economic challenges, age-related vulnerabilities, and academic performance as significant predictors of dropout rates. Targeted interventions that address these specific areas are essential for reducing dropout rates. Our findings advocate for a holistic approach to intervention, considering the complex interplay of socio-demographic factors. This approach should not only focus on individual predictors but also on how these factors interact within the broader educational and social

context. Strategies that are inclusive and consider the diverse needs of all students can lead to more equitable educational outcomes. To enhance the generalizability of our findings, future research should expand the range of demographic factors and geographical areas studied. This expansion would help to ensure a more comprehensive understanding of dropout phenomena across different contexts, which is crucial for developing universally applicable and effective educational strategies. Longitudinal studies are recommended to track the effectiveness of interventions over time. Such studies can provide valuable insights into the long-term impacts of policies and practices, helping to refine and adjust strategies to maximize their effectiveness. Integrating the insights from our study into educational policy and practice can significantly accelerate progress toward inclusive education and reduce student dropout rates. By aligning intervention strategies with the detailed understanding provided by our research, stakeholders can make informed decisions that contribute to sustainable educational development in India.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42001-024-00309-z.

Acknowledgements We would like to thank Mr. Saravanaraj K, Biostatistician, CMC, Vellore, and Ms. Supriya, Research Scholar, Translational Medicine and Research, SRM Institute of Science and Technology for the meticulous and insightful proofreading of this manuscript.

Author contributions RGV and BM conceptualized the study, RGV performed data analysis, RGV and BM finalised the results. RGV and BM wrote the main manuscript text, reviewed and finalised the manuscript.

Data availability The data set that was analyzed during the study is not publicly available due to the ethical concerns. However, the corresponding author can provide the data upon reasonable request.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Venkatesan, R. G., & Mappillairaju, B. (2023). Detection of hotspots of school dropouts in India: A spatial clustering approach. *PLoS One*, *18*(1), e0280034.
2. Department of School Education and Literacy M of E. District Information System for Education [Internet]. Government of India (2023). https://www.education.gov.in/sites/upload_files/mhrd/files/statistics-new/udise_21_22.pdf.
3. India, R. G. (2001). *Census of India 2001*. Provisional Popul Total New Delhi.
4. Banerji, M., & Mathur, K. (2021). Understanding school attendance: The missing link in schooling for all. *Int J Educ Dev*, *87*, 102481.
5. Arthi, D., Rajalakshmi, M., & Ganapathy, K. Reasons for school dropouts in suburban areas near Villupuram district: A retrospective study. J Curr Res Sci Med [Internet]. 9900; https://journals.lww.com/jcsm/fulltext/9900/reasons_for_school_dropouts_in_suburban_areas_near.19.aspx.
6. Boudet, A. M. M., Petesch, P., Turk, C., & Thumala, A. (2012). On norms and agency: Conversations about gender equality with women and men in 20 countries. World Bank Washington, DC.
7. Colclough, C., Kingdon, G., & Patrinos, H. (2010). The changing pattern of wage returns to education and its implications. *Dev Policy Rev*, *28*(6), 733–747.

8. Gubbels, J., van der Put, C. E., & Assink, M. (2019). Risk factors for School Absenteeism and Dropout: A Meta-Analytic Review. *Journal of Youth and Adolescence*, *48*(9), 1637–1667.
9. Venkatesan, R. G., Karmegam, D., & Mappillairaju, B. (2024). Exploring determinants of school dropout across regions in India: a comprehensive meta-analysis [Internet]. Journal of Computational Social Science. Springer Nature Singapore; https://doi.org/10.1007/s42001-024-00285-4.
10. Ameri, S., Fard, M. J., Chinnam, R. B., & Reddy, C. K. (2016). Survival analysis based framework for early prediction of student dropouts. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. pp. 903–12.
11. Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, *45*(1), 89–125.
12. Nangia, S., Anurag, J., & Gambhir, I. (2020). A machine learning approach to identify the students at the risk of dropping out of secondary education in India. In: Soft Computing and Signal Processing: Proceedings of 2nd ICSCSP 2019 2. Springer; pp. 557–69.
13. Colak Oz, H., Güven, Ç., & Nápoles, G. (2023). School dropout prediction and feature importance exploration in Malawi using household panel data: Machine learning approach. *J Comput Soc Sci*, *6*(1), 245–287.
14. Prenkaj, B., Velardi, P., Stilo, G., Distante, D., & Faralli, S. (2020). A survey of machine learning approaches for student dropout prediction in online courses. *ACM Comput Surv*, *53*(3), 1–34.
15. de Oliveira, C. F., Sobral, S. R., Ferreira, M. J., & Moreira, F. (2021). How does learning analytics contribute to prevent students' dropout in higher education: A systematic literature review. *Big Data Cogn Comput*, *5*(4), 64.
16. Venkatesan, R. G., Karmegam, D., & Mappillairaju, B. (2023). Exploring statistical approaches for predicting student dropout in education: A systematic review and meta-analysis. *J Comput Soc Sci*. ;1–26.
17. Mduma, N., Kalegele, K., & Machuve, D. (2019). A survey of machine learning approaches and techniques for student dropout prediction. .
18. Education, M. (2017). Student Data Capture Format (U-DISE) [Internet]. https://educationforallinindia.com/wp-content/uploads/2022/09/NIEPA-SDMIS-UDISE-2016-17-data-capture-format.pdf.
19. Juajibioy, J. C. (2016). Study of university dropout reason based on survival model. *Open J Stat*, *6*(5), 908–916.
20. Bani, M. J., & Haji, M. (2017). College Student Retention: When Do We Losing Them? arXiv Prepr arXiv170706210. .
21. da Costa, F. J., Bispo, M., de Pereira, S., & de C, R. (2018). de F. Dropout and retention of undergraduate students in management: a study at a Brazilian Federal University. RAUSP Manag J. ;53:74–85.
22. Pachas, D. A. G., Garcia-Zanabria, G., Cuadros-Vargas, A. J., Camara-Chavez, G., Poco, J., & Gomez-Nieto, E. (2021). A comparative study of WHO and WHEN prediction approaches for early identification of university students at dropout risk. In: 2021 XLVII Latin American Computing Conference (CLEI). IEEE; pp. 1–10.
23. Pan, F., Huang, B., Zhang, C., Zhu, X., Wu, Z., Zhang, M., et al. (2022). A survival analysis based volatility and sparsity modeling network for student dropout prediction. *PLoS One*, *17*(5), e0267138.
24. Gutierrez-Pachas, D. A., Garcia-Zanabria, G., Cuadros-Vargas, A. J., Camara-Chavez, G., Poco, J., & Gomez-Nieto, E. (2022). How do Curricular Design Changes Impact Computer Science Programs? A case study at San Pablo Catholic University in Peru. *Educ Sci*, *12*(4), 242.
25. Gutierrez-Pachas, D. A., Garcia-Zanabria, G., Cuadros-Vargas, E., Camara-Chavez, G., & Gomez-Nieto, E. (2023). Supporting decision-making process on higher education dropout by analyzing academic, socioeconomic, and equity factors through Machine Learning and Survival Analysis Methods in the latin American context. *Educ Sci*, *13*(2), 154.
26. Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. .
27. Yu, C-N., Greiner, R., Lin, H-C., & Baracos, V. (2011). Learning patient-specific cancer survival distributions as a sequence of dependent regressors. *Advances in Neural Information Processing Systems*. ;24.
28. Wright, M. N., Dankowski, T., & Ziegler, A. (2017). Unbiased split variable selection for random survival forests using maximally selected rank statistics. *Statistics in Medicine*, *36*(8), 1272–1284.
29. Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network. *Bmc Medical Research Methodology*, *18*(1), 1–12.

30. Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N. A., Trollor, J., et al. (2020). A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports*, *10*(1), 20410.
31. Singh, R., & Mukherjee, P. (2017). Diverging Pathways: When and Why Children Discontinue Education in India.
32. Prakash, R., Beattie, T., Javalkar, P., Bhattacharjee, P., Ramanaik, S., Thalinja, R., et al. (2017). Correlates of school dropout and absenteeism among adolescent girls from marginalized community in north Karnataka, south India. *Journal of Adolescence*, *61*, 64–76.
33. Marphatia, A. A., Reid, A. M., & Yajnik, C. S. (2019). Developmental origins of secondary school dropout in rural India and its differential consequences by sex: A biosocial life-course analysis. *Int J Educ Dev*, *66*, 8–23.
34. Paul, R., Rashmi, R., & Srivastava, S. (2021). Does lack of parental involvement affect school dropout among Indian adolescents? Evidence from a panel study. *PLoS One*. ;16(5).
35. Bhapkar, V. P. (1965). *Categorical data analogs for some multivariate tests*. North Carolina State University. Dept. of Statistics.
36. Bhapkar, V. P. (1968). On the analysis of contingency tables with a quantitative response. *Biometrics*. ;329–338.
37. Cerda, P., Varoquaux, G., & Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. *Machine Learning*, *107*(8–10), 1477–1494.
38. Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics*, *9*(4), 705–724.
39. Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.
40. Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, *4*(2), 1883.
41. Kecman, V. (2005). Support vector machines–an introduction. *Support vector machines: Theory and applications* (pp. 1–47). Springer.
42. Berrar, D. (2019). Bayes' Theorem and Naive Bayes Classifier.
43. Delashmit, W. H., & Manry, M. T. (2005). Recent developments in multilayer perceptron neural networks. In: Proceedings of the seventh annual memphis area engineering and science conference, MAESC. p. 33.
44. Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In: 2010 20th international conference on pattern recognition. IEEE; pp. 3121–4.
45. Dalianis, H., & Dalianis, H. (2018). Evaluation metrics and evaluation. *Clin Text Min Second use Electron Patient Rec*. ;45–53.
46. Melo, F. (2013). Area under the ROC Curve BT - Encyclopedia of Systems Biology. In: Dubitzky W, Wolkenhauer O, Cho K-H, Yokota H, editors. New York, NY: Springer New York; pp. 38–9. https://doi.org/10.1007/978-1-4419-9863-7_209.
47. Austin, P. C. (2017). A tutorial on multilevel survival analysis: Methods, models and applications. *International Statistical Review*, *85*(2), 185–203.
48. Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370.
49. Demler, O. V., Paynter, N. P., & Cook, N. R. (2015). Tests of calibration and goodness-of-fit in the survival setting. *Statistics in Medicine*, *34*(10), 1659–1680.
50. Kleinbaum, D. G., & Klein, M. (1996). *Survival analysis a self-learning text*. Springer.
51. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, *247*(18), 2543–2546.
52. Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N. (2010). Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures. Epidemiology [Internet]. ;21(1). https://journals.lww.com/epidem/fulltext/2010/01000/assessing_the_performance_of_prediction_models__a.22.aspx.
53. Zhou, H., Wang, H., Wang, S., Zou, Y., & SurvMetrics (2022). An R package for Predictive Evaluation Metrics in Survival Analysis. *R J*, *14*(4), 1–12.
54. Therneau, T., & Lumley, T. (2013). R survival package. *R Core Team*. .
55. Jackson, C. H. (2016). Flexsurv: A platform for parametric survival modeling in R. *Journal of Statistical Software*. ;70.
56. Davidson-Pilon, C. (2019). Lifelines: Survival analysis in Python. *J Open Source Softw*, *4*(40), 1317.
57. Pölsterl, S. (2020). scikit-survival: A Library for Time-to-event analysis built on Top of scikit-learn. *J Mach Learn Res*, *21*(1), 8747–8752.

58. Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.
59. Chouhan, B. S. (2017). Factors affecting the School Dropout in Rajasthan: A study based on NSSO Data. *J Cult Soc Dev*, *30*, 56–61.
60. Patel, R., Singh, A. K., Chandra, M., Khanna, T., & Mehra, S. (2018). Is Mother's Education or Household Poverty a Better Predictor for Girl's School Dropout? Evidence from Aggregated Community Effects in Rural India. Educ Res Int. ;2018.
61. Singh, R., & Mukherjee, P. (2018). 'Whatever she may study, she can't escape from washing dishes': gender inequity in secondary education–evidence from a longitudinal study in India. Compare [Internet]. ;48(2):262–80. https://doi.org/10.1080/03057925.2017.1306434.
62. Mitra, S., Mishra, S. K., & Abhay, R. K. (2022). Out-of-school girls in India: a study of socioeconomic-spatial disparities. GeoJournal [Internet]. ;7. https://doi.org/10.1007/s10708-022-10579-7.
63. Queiroga, E. M., Batista Machado, M. F., Paragarino, V. R., Primo, T. T., & Cechinel, C. (2022). Early Prediction of At-Risk students in secondary education: A Countrywide K-12 Learning Analytics Initiative in Uruguay. *Inf*, *13*(9), 1–25.
64. Weybright, E. H., Caldwell, L. L., Xie, H., Wegner, L., & Smith, E. A. (2017). Predicting secondary school dropout among South African adolescents: A survival analysis approach. *SOUTH AFRICAN J Educ*. ;37(2).
65. Hunt, F. (2008). *Dropping out from School: A Cross Country Review of the literature. Create pathways to Access*. Research Monograph, No. 16. ERIC.
66. Vadivel, B., Alam, S., Nikpoo, I., & Ajanil, B. (2023). The Impact of Low Socioeconomic Background on a Child's Educational Achievements. Educ Res Int. ;2023.
67. Al-Shamsi, A., Al-Hawari, M. A., Aderibigbe, S. A., & Omar, M. (2023). In H. M. K. Al Naimiy, M. Bettayeb, H. M. Elmehdi, & I. Shehadi (Eds.), *Impact of Service Quality on Student Retention in UAE Higher Education Institutions BT - Future trends in Education Post COVID-19* (pp. 205–219). Springer Nature Singapore.
68. Du, H., Xiao, Y., & Zhao, L. (2021). Education and gender role attitudes. *Population Economics*, *34*(2), 475–513.
69. Ministry of Human Resource Development G of I. UDISE+ (2019). p. 90.
70. Blackwell, M., Honaker, J., & King, G. (2017). A Unified Approach to Measurement Error and Missing Data: Overview and applications. *Sociol Methods Res*, *46*(3), 303–341.
71. Savitz, D. A., & Wellenius, G. A. (2023). Can cross-sectional studies contribute to causal inference? It depends. *American Journal of Epidemiology*, *192*(4), 514–516.

Authors and Affiliations

Raghul Gandhi Venkatesan[1,2] · Bagavandas Mappillairaju[2]

✉ Bagavandas Mappillairaju
bagwandm@srmist.edu.in

Raghul Gandhi Venkatesan
raghulgandhivenkatesan@gmail.com

[1] Department of Mathematics, Faculty of Engineering and Technology, SRM Institute of Science & Technology, Kattankulathur, Tamil Nadu 603203, India

[2] Centre for Statistics, SRM Institute of Science & Technology, Kattankulathur, Tamil Nadu 603203, India