Educational Administration: Theory and Practice
2024, 30(6), 5448-5454
ISSN: 2148-24036
https://kuey.net/

Research Article

A Machine Learning Predictive Analysis on the Educational Out-migration from the Northeast India

*N. Jayenta Meitei[1], Md. Baharuddin Shah[2]

[1]Assistant Professor, Department of Geography, Y.K. College, Wangjing, Manipur
[2]Md. Baharuddin Shah, Professor, Department of Geography, G.P. Women's College, Imphal, Manipur

*Corresponding Author: N. Jayenta Meitei
Email: jayenta.kak@gmail.com

Citation: N. Jayenta Meitei, et.al (2024). A Machine Learning Predictive Analysis on the Educational Out-migration from the Northeast India, *Educational Administration: Theory and Practice*, 30(6) 5448-5454
Doi: 10.53555/kuey.v30i6.10959

ARTICLE INFO

ABSTRACT

This study adopts a machine learning-based predictive analysis to forecast the trend of educational out-migration from the Northeast region of India during 2011–2031. Using Census of India migration data (1981–2011) as the data source, the research employs Linear, Quadratic, Exponential and ARIMA statistical models. Due to the delay of census exercise post 2011, there is an evident gap of recent migration trends in India which needs to be fulfilled either by periodical surveys or interpolation methods. This study is an attempt to fill up this gap of recent data in respect of educational migration from the Northeast region, which has already revealed prevalence of high intensity of educational out-migration stream. Based on the analysis, it is predicted that the volume of this migration stream shall continue to increase with declining intensity. The findings highlight critical policy implications for addressing educational infrastructure gaps within the region and managing migration-driven pressures in major urban educational hubs.

Keywords: Machine learning, Educational migration, Predictive models, Northeast India

1. Introduction

Machine learning (ML) and predictive analytics have emerged as transformative tools across multiple domains enabling data-driven decisions that enhance efficiency, accuracy and strategic planning. Across diverse sectors like migration studies, education, healthcare and organizational management, machine learning has become a transformative analytical approach for uncovering hidden patterns, predicting human and institutional behaviours, and optimizing decision-making. Scholars highlight that predictive analytics has evolved from simple regression models to complex hybrid and ensemble methods capable of handling high-dimensional nonlinear data structures (Nayak and Soy, 2024; Vatti et al., 2024). From population mobility forecasting to educational outcome prediction and healthcare resource management, the integration of artificial intelligence (AI) and machine learning marks a paradigm shift toward data-driven governance and strategic planning (Hussain, 2021; Avinash, G. et al., 2025).

One crucial area of research to apply ML-based predictive analysis is India's Northeast region (NER), which has been a significant source of out-migration for work and education in the interstate migrant flows over the past few decades (Lyndem & De, 2004; Shimray & Usha, 2009; Chyrmang, 2011; McDuie-Ra, 2012; Lusome & Bhagat, 2020). Compared to the all India average, the Northeast states had a notably higher share of educational migration (Mistri & Sardar, 2022).

Yet, recent studies on trends and patterns of migration have been greatly limited by the delay of Census exercise in the country post 2011. In this regard, this study seeks to fill the gap of recent trends and magnitude of educational out-migration from the Northeast states of India during the period 2011-2031. It may be noted that the Northeast region of India comprises eight states of Assam, Arunachal Pradesh, Sikkim, Meghalaya, Nagaland, Manipur, Mizoram and Tripura.

2. Literature Review

This review synthesizes recent empirical studies to illustrate the expanding scope and methodological convergence of ML-based predictive models across socio-economic, educational, and health contexts. In the context of population studies, traditional models such as the Gravity and Radiation models have long been used to explain migratory patterns. However, these models often fail to capture the dynamic and nonlinear nature of contemporary migration. Hussain et al. (2021) advanced this field by integrating ML algorithms into reverse migration modeling, emphasizing the increasing trend of urban-to-rural migration in Malaysia. Their systematic literature review identified Decision Tree, Random Forest, and Linear Regression as the most effective models for predicting population shifts while minimizing forecast error. Educational forecasting has evolved from descriptive statistics to intelligent modeling systems that anticipate global student flows. Yang et al. (2020) proposed a hybrid Feature Selection–Differential Evolution–Support Vector Regression (FSDESVR) model to forecast Taiwanese outbound student mobility.

Deepa and Kumar (2024) evaluated multiple classifiers—Decision Tree, Random Forest, and XGBoost—and found ensemble models to outperform linear classifiers, achieving over 90% accuracy in predicting engagement and adaptability. Similarly, Bharath Kumar et al. (2025) utilized Random Forest to classify engagement levels in online education with 100% accuracy, highlighting its applicability in adaptive learning systems. Avinash et al. (2025) conducted an extensive study on time series forecasting of bed occupancy in Indian mental health hospitals**,** utilizing six algorithms—Support Vector Regression (SVR), Random Forest, XGBoost, Gradient Boosting, K-Nearest Neighbors, and Decision Tree. With 866 weekly observations spanning 2008–2024, the study demonstrated that ML-based time series models outperform classical approaches like ARIMA by incorporating lag selection, hyperparameter tuning, and error metrics such as RMSE, MAE, and MAPE.

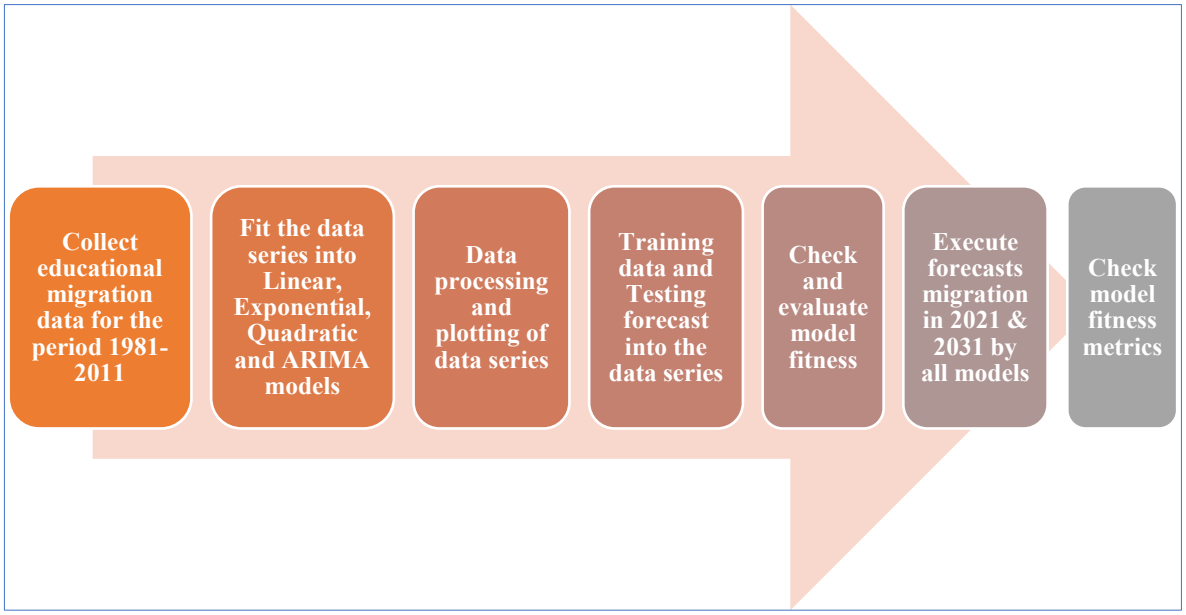
The Autoregressive Integrated Moving Average (ARIMA) model, developed by Box and Jenkins (1970), has long been a cornerstone of time series forecasting across economics, healthcare, climatology, and migration studies. ARIMA combines autoregression (AR), differencing (I), and moving average (MA) components to model temporal dependencies and non-stationarity, making it highly effective for univariate forecasting tasks. Hyndman and Athanasopoulos (2018) noted that ARIMA remains a benchmark method for short-term forecasting because it captures autocorrelation structures without requiring exogenous predictors. Similarly, Zhang (2003) observed that although nonlinear methods like neural networks have gained prominence, ARIMA often performs competitively, particularly when the data exhibit linear trends and seasonal patterns. In the field of migration and population forecasting**,** ARIMA has been instrumental in modeling long-term demographic shifts. Singh and Das (2022) demonstrated that ARIMA provides accurate forecasts for intercensal migration data when seasonality and trend components are properly differenced, making it valuable for policy formulation in population planning. Applications in public health and epidemiology have also demonstrated the robustness of ARIMA. For instance, Ceylan (2020) utilized ARIMA to forecast the spread of COVID-19 cases, finding it capable of producing reliable short-term projections under conditions of rapidly changing data. In environmental science, Kumar and Jain (2010) applied ARIMA to predict river discharge, emphasizing its strength in handling noisy hydrological data through differencing and residual diagnostics. Despite its utility, researchers acknowledge ARIMA's limitations in handling highly nonlinear or multivariate datasets. To address these issues, hybrid models combining ARIMA with machine learning techniques have emerged. For example, Khashei and Bijari (2011) proposed a hybrid ARIMA–Artificial Neural Network (ANN) model that achieved higher accuracy by integrating ARIMA's linear forecasting ability with ANN's nonlinear adaptability.

3. Methodology

The present study adopts a predictive analytical research design incorporated with a machine learning (ML) approach. The steps of the ML-based forecasting are illustrated in Figure-1. In order to achieve the objective of the research, this study employs a machine learning approach in application of selected statistical models, i.e. Linear Regression, Quadratic Regression, Exponential Regression and Auto-regressive Integrated Moving Average (ARIMA) models. Machine learning predictive analysis involves three basic exercises namely "split", "train" and "test" to be executed in a predictive or forecasting model. In the present study, this is done by splitting the data series into two parts, by training the model itself to learn the historical patterns from one part of the data series, and by testing a forecast performance based on the remaining part of the data series. This is how the best performing forecast model is constructed on the basis of model fitness metrics so that future predictions can be made.

All the statistical analyses and modeling of this research are done with Python package version 3.13. The Python libraries used in this study include NumPy, Pandas, Itertools, Statsmodels, Scikit-learn and Matplotlib.

Figure-1: Flowchart of workflows of the ML-based predictive analysis of educational out-migration



3.1 Data Sources

This study uses the stock of educational out-migrants from the Northeast states of India for the period 1981-2011. The source of this data is D-3 tables of the Census of India 1981, 1991, 2001 and 2011. Based on the information given on 'reasons of migration' and 'last usual place of residence', inter-state educational out-migrants (migration for the reason of education) whose last usual place of residence is any of the eight Northeast states of India are extracted from the Census tables for analysis. The period 1981-2011 is considered because the Census began collecting information on the 'education' reason of migration since 1981 while 2011 Census was the last one conducted so far in the country.

Table-1: Stock of educational out-migrants from the Northeast states of India

Year of census	Intercensal migrants	Total stock of migrants
1981#	--	18,000
1991	17,884	26,327
2001	26,659	30,173
2011	32,737	40,478

Data source: Computed from D-3 tables of Census of India 1981, 1991, 2001, 2011.
#Census was not conducted in the state of Assam in Northeast India due to socio-political disturbances in 1981. Hence, the total stock of migrants is estimated by interpolation.

3.2 Model fitting

Statistical modeling is done to forecast aggregate stock of migrants in 2021 and 2031 based on total stock of migrants in the last four census years. Only four past data points are in hand to fit into the decadal forecast model (see Table-1). Considering the scarcity of data, the study adopts four different forecast models so that the best fitted one can be chosen. The four fitted models are below:

(i) Linear Regression

$$Y = 18052.50 + 712.80\times(X_{year}-1981) + \varepsilon \quad \text{..... (i)}$$

(ii) Quadratic Regression

$$Y = 18547.00 + 564.45\times(X_{year}-1981)^2 + \varepsilon \quad \text{..... (ii)}$$

(iii) Exponential Regression

$$Y = \exp(9.84 + 0.03\times(X_{year}-1981)) + \varepsilon \quad \text{..... (iii)}$$

(iv) ARIMA(p,d,q)

$$\Delta y_t = 0.00 + 0.203\times(\Delta y_{t-1}) + 1.00\times(\varepsilon_{t-1}) + \varepsilon_t \quad \text{..... (iv)}$$

where,

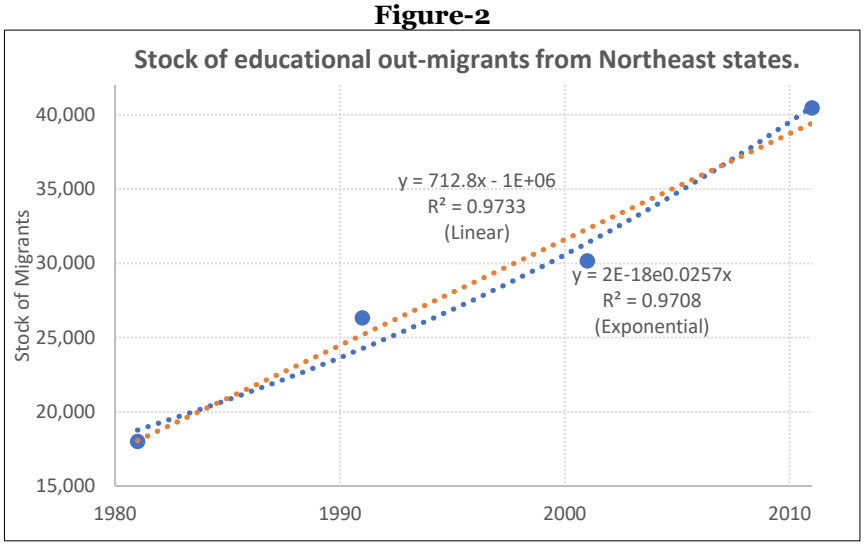
'p' stands for order of the autoregressive (AR) term;
'd' stands for order of differencing;
'q' stands for order of the moving average (MA) term;

'*Y*' or '*y*' stands for stock of migrants;
't' stands for time index i.e. census year;
'ε' stands for error term.

4. Machine Learning

For decadal forecast of migrants, there are four historical data points i.e. 1981, 1991, 2001 and 2011 to study and learn from (see Figure-2). This data series follows a steady rising trendline, which satisfactorily fits either a linear or an exponential curve. As mentioned before, machine learning requires a training and testing exercise by splitting the data series into two parts so that it can identity the trends and patterns to satisfactorily forecast future trends.

Figure-2



Source: Computed from Census of India 1981, 1991, 2001 & 2011.

The *train_test_split* function under *scikit-learn* or *sklearn* python library is used to perform supervised machine learning while fitting a forecast model. The data series is split into training and testing data in 80:20 ratio. Thus, the model is trained on data from 1981-2001 while tested into the data point in 2011. This is performed in all models i.e. linear, exponential, quadratic and ARIMA models (see Figure-3). The models predict different test predictions in 2011. Only exponential regression model can predict very close to the actually observed test data. Other remaining models tend to predict lower than the test data. The accuracy of machine learning test prediction can be evaluated such metrics like RMSE, R^2 and p-value (see Table-2). Exponential and quadratic models can provide best fit with lowest errors and highest R^2 scores. The negative R^2 score of ARIMA model indicates that it fails to capture the trend well.

Figure-3

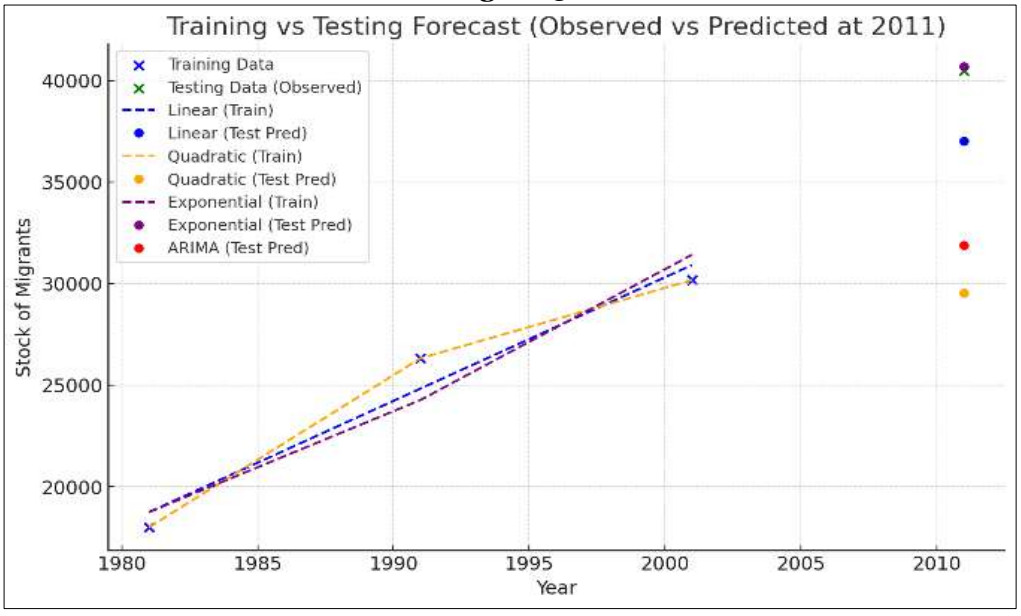


Table-2: ML models training and testing results

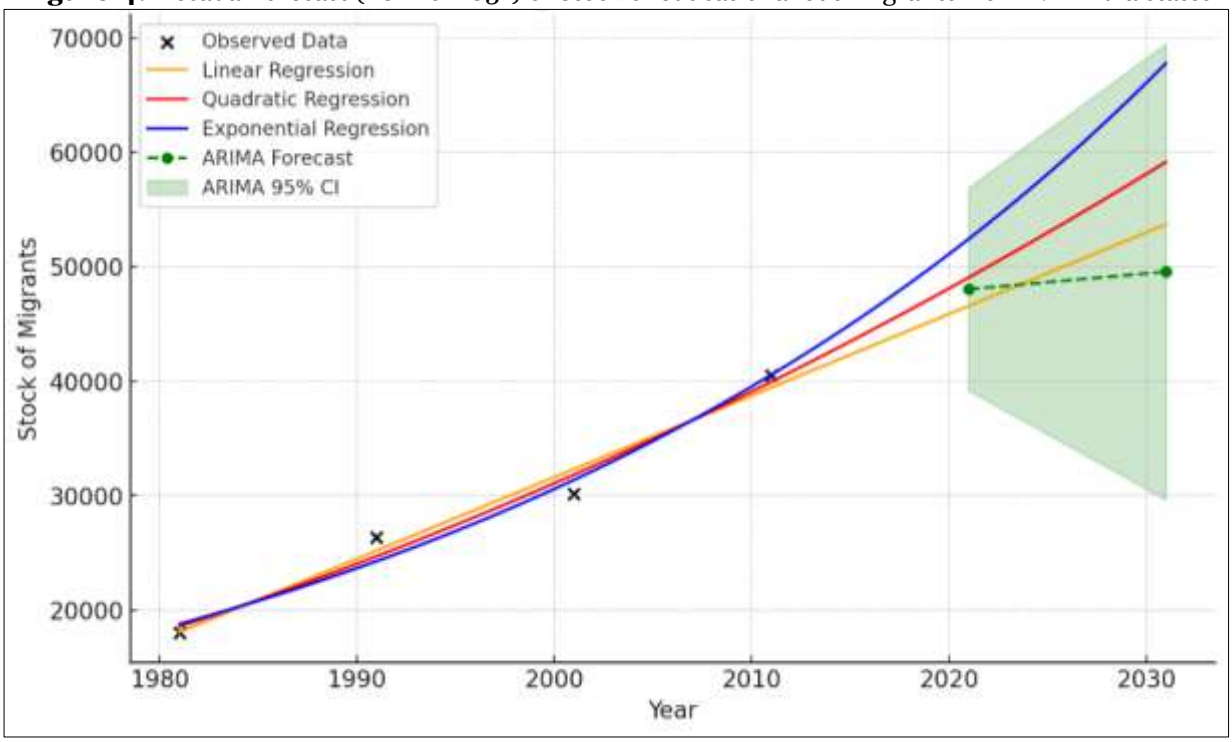
Year	Observed/ test value	Linear reg. test predict	Quadratic reg. test predict	Exponential reg. test predict	ARIMA reg. test predict
2011	40,478	37,006	29,538	40,685	31,887
RMSE		1319	1223	1252	9875
R^2		0.973	0.977	0.976	-0.495
p-value		-	-	-	0.456

Source: Computed from Census of India 1991, 2001 & 2011.

5. Findings

Four models' forecasts of the decadal stock of educational out-migrants from the Northeast India states in 2021 and 2031 are shown in figure-4 and table-3. While exponential regression predicts the highest stock of migrants, ARIMA predicts the lowest stock. Across the models, the predicted stock of migrants ranges from 46,564 (linear) to 52,407 (exponential) in 2021 and from 49,550 (ARIMA) to 67,748 (exponential) in 2031. All three regression models march upward at different speeds (exponential fastest, linear slowest). ARIMA is more conservative and comes with wide uncertainty bands especially by 2031 because there are very few historical data points to learn from.

Figure-4: Decadal forecast (2021 & 2031) of Stock of educational out-migrants from NE-India states



As far as model fitness is concerned, Linear, Quadratic and Exponential regressions fit the data series very well ($R^2 \approx 0.97–0.98$, low RMSE). Quadratic regression gives the best overall fit with highest R^2 and lowest RMSE. ARIMA(1,1,1) model performs poorly here (negative R^2, very high RMSE) because the dataset is too short for a stable time-series model. Ljung-Box test indicates residuals white-noise. With only 4 observations, ARIMA is not appropriate. It's designed for denser time series. That's why the ARIMA forecast is nearly flat and has very wide intervals. Quadratic and exponential models fit slightly better than linear model in-sample. However, with so little data, the extra curvature can overstate long-horizon growth. That's visible in 2031 where exponential shoots to 67,748 mark. Linear model is the most conservative and stable extrapolation. It under-fits slightly in-sample but avoids compounding curvature risk. The best three models forecast can be considered in such scenario that Linear forecast (46,564) is the baseline, Quadratic forecast is for moderate acceleration and Exponential forecast is for high-growth scenario.

Table-3: Forecast of educational out-migrants stock from Northeast India states

Models	Forecasted stock of migrants		Model fitness metrics		
	2021	2031	R^2	RMSE	L-B test p-value
Linear Regression	46564	53692	0.97	1319.31	
Quadratic Regression	49036	59131	0.98	1223.13	
Exponential Regression	52407	67748	0.98	1252.63	
ARIMA(1,1,1)	48021	49550	-0.49	9875.56	0.33 (lag-1)

Source: Computed from Census of India 1981, 1991, 2001 & 2011.

6. Conclusions

The three regression models viz. linear, exponential and quadratic, altogether predict steady rise of the stock of educational out-migrants from the Northeast states of India at different speeds based on historical decadal stocks during 1981-2011. Across the models, the prediction ranges from 46,564 migrants (linear) to 52,407 (exponential) in 2021 and from 53,692 (linear) to 67,748 (exponential) in 2031.

The forecast pattern is logical**,** reflecting ongoing socio-economic dynamics. Together, this model captures a smooth long-run trend in migration while smoothing out annual fluctuations. The pattern suggests continued educational out-migration growth, consistent with long-term social and economic trends driving students to move out of the region for higher education. The model is statistically sound for medium- to long-term forecasting of student migration trends. It reliably projects a steady rise in educational out-migration, though policymakers should consider uncertainty in the upper and lower forecast bounds for planning and policy formulation. It also suggests an apparent vacuum of educational environment in the Northeast states, which needs to be fulfilled for the migrant students to stay around (Lyndem & De, 2004). On the other hand, the ever-rising wave of out-migrants are destined for the educational hubs in the country like Delhi, Mumbai, Bengaluru etc., which are expected to accommodate them with state-of-the-art facilities and living amenities (Shimray & Usha, 2009; Mistri & Sardar, 2022).

References

1. Avinash, G., et al. (2025). Time series forecasting of bed occupancy in mental health facilities in India using machine learning. *Scientific Reports*, 15(2686), 1–12. https://doi.org/10.1038/s41598-025-86418-9
2. Bharath Kumar, G., et al. (2025). A predictive approach to student performance in online learning using data analytics. *Journal of Computational Analysis and Applications*, 34(6), 175–186.
3. Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. San Francisco, CA: Holden-Day.
4. Ceylan, Z. (2020). Estimation of COVID-19 prevalence in Italy, Spain, and France. *Science of the Total Environment, 729*, 138817. https://doi.org/10.1016/j.scitotenv.2020.138817
5. Chyrmang, R. (2011). Magnitude of Migration from North-Eastern Region of India. In S. Irudaya Rajan (Ed.), *Migration, Identity and Conflict – India Migration Report 2011*. New Delhi: Routledge.
6. Deepa, P. S., & Kumar, M. (2024). Machine learning for predicting student engagement and adaptability in online courses. *Journal of Computational Analysis and Applications*, 33(6), 1629–1638.
7. Hussain, N. H. M., et al. (2021). Machine learning of the reverse migration models for population prediction: A review. *Turkish Journal of Computer and Mathematics Education*, 12(5), 1830–1838.
8. Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (2nd ed.). Melbourne, Australia: Texts.
9. Khashei, M., & Bijari, M. (2011). A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Applied Soft Computing, 11*(2), 2664–2675. https://doi.org/10.1016/j.asoc.2010.10.015
10. Kumar, U., & Jain, S. K. (2010). Time series models (ARIMA) for forecasting river flow: A case study of the Savitri River, India. *Journal of Earth System Science, 119*(4), 491–500. https://doi.org/10.1007/s12040-010-0042-6
11. Lusome, R., & Bhagat, R. B. (2020). Migration in Northeast India: Inflows, outflows and reverse flows during pandemic. *The Indian Journal of Labour Economics*, 63, 1125-1141.
12. Lyndem, B. and De, U. K. (Eds.) (2004). *Education in North-East India: Experience and Challenge*. New Delhi: Concept Publishing Company.
13. McDuie-Ra, D. (2012). The 'NorthEast' Map of Delhi. *Economic and Political Weekly*, 47(30):69-77.
14. Mistri, A and Sardar, S. S. (2022). Student Migration from the North-East India: Level, Trend, Pattern and Challenges. *Demography India*, 51(1):40-62.
15. Nayak, A. & Soy, A. (2024). Machine Learning Applications in NLP: An In-Depth Review of Techniques and Trends. *Journal of Computational Analysis and Applications (JoCAAA), 33*(2),1423–1431. Retrieved from: https://www.eudoxuspress.com/index.php/pub/article/view/2556

16. Shimray, U. A., & Usha Devi, M. D. (2009). Trends and patterns of migration: Interface with education – A case of the North-Eastern Region. (*Social and Economic Change Monographs*, No. 15). Bangalore: Institute for Social and Economic Change (ISEC).
17. Singh, M., & Das, R. (2022). Forecasting population migration trends using ARIMA time series modeling: A case of Northeast India. *Journal of Population Research, 39*(3), 543–560.
18. Vatti, P.R., Vatti, V.R. & Sen, K. (2024). Advancing Data Science with System Intelligence: A Machine Learning Approach to Predictive Data Engineering. *Journal of Computational Analysis and Applications (JoCAAA), 33*(08), 2344–2360. Retrieved from https://www.eudoxuspress.com/index.php/pub/article/view/2086
19. Yang, S., et al. (2020). Forecasting outbound student mobility: A machine learning approach. *PLOS ONE*, 15(9), e0238129. https://doi.org/10.1371/journal.pone.0238129
20. Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing, 50*, 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0