**Literacy Rate Analysis Using Machine Learning**

**Preetham Kasarla1 . Ganesh Kore2**

1. **ABSTRACT**

Literacy is an important indicator of the socioeconomic development of a country; therefore, it is important to continuously monitor and understand the factors affecting literacy rates [1]. This study presents a comprehensive survey of literacy rates with the use of machine learning techniques, and a large dataset of demographic, socioeconomic, and educational variables, spanning several decades [2]. Uses machine learning algorithms including regression, classification, and clustering models to explore relationships between the difficulty between these variables and literacy rate [3], and examines the influence of socio-economic factors. Regression analysis reveals the strength and direction of these associations, providing a deeper understanding of the determinants of literacy [4]. Classification models are then used to characterize individual literacy status, reveal important determinants of literacy, and help identify at-risk populations These predictive models can assist policymakers in targeting policies encouragement of literacy [5]. Furthermore, the application of clustering methods to cluster areas or countries with similar literacy rate trends sheds light on geographic patterns and helps to identify areas in need of attention. The findings of this study provide valuable insights for policymakers, educators, and social planners to develop evidence-based strategies to improve literacy levels [6]. By harnessing the power of machine learning, this research advances our understanding of multidimensional literacy and contributes to global efforts to promote education and narrow the gap between literacy and the number of writings [7].

**Keywords**

Literacy rate, Machine learning

1. **INTRODUCTION**

The literacy rate is an important indicator of both a country’s economic prosperity and its human capital development [1]. The connection between literacy and the well-being of a nation is well established, and higher literacy rates tend to mean greater access to education and opportunities for its citizens. This paper begins a comprehensive analysis of a wide range of effects of literacy rates [2].

The literacy rate in India has always been a cause for concern, attracting the attention of governments, policymakers, researchers, and activists since independence [3]. The importance of literacy in the country’s life and economic growth is not overstated. This is an important indicator of a country’s development, reflecting not only its educational standards but potential for economic growth and social welfare [4]. This paper explores the multifaceted topic of literacy in India in detail, which aims to shed light on the various determinants of literacy rates in its various forms.

Algorithms and automation have emerged as the architects of this new information age, having a profound impact on everything from information delivery to delivery and consumption [5]. These technological advancements have ushered in a brave new world where tracking user actions, data mining, and profiling have become the norm, powered by computational wizardry and machine learning prowess [6]. However, with the ease of producing and disseminating content, we find ourselves navigating a terrain rife with both opportunity and peril, where misinformation and advertising vie for our attention in increasingly sophisticated ways. It is in this context that a pressing need arises for a robust understanding of modern media tools [7].

This study endeavors to bridge this educational gap by designing, implementing, and evaluating an AI literacy course tailored specifically for university students [8]. The primary objective is to ascertain whether students from diverse academic disciplines can develop a robust conceptual understanding of AI through this course [9]. With an initial pool of 4000 students, 120 enthusiastic volunteers willingly embarked on this educational journey, dedicating themselves to a comprehensive 7-hour course [11].

Top of Form

1. **LITERATURE SURVEY**

**SURVEY DETAILS:**

Mansift Kaur B.A. (Honours) Economics University of Delhi, Sri Guru Tegh Bahadur Khalsa College, New Delhi, India, (2020) have developed An Econometric Analysis of Literacy Rates in Different States of India and Factors Stimulating Them [1].

Vaidehi et al. (2018) have developed Analysis and Forecast of Literacy Rates in India [2].

Teemu et al.(2019), have developed Media Literacy Education in the Age of Machine Learning [3].

Viktoriya et al. (2021) have developed Addressing AI and Data Literacy in Teacher Education: A Review of Existing Educational Frameworks [4].

Siu-Cheung Kong et al. (2021) have developed Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds [5].

Akpofure et al. (2020) have developed Application of Machine Learning Methods to Predict Student Performance: A Systematic Literature Review [6].

Massachusetts et al. (2021) have developed Developing Middle School Students’ AI Literacy [8].

Davy Tsz Kit Ng et al. (2021) have developed Conceptualizing AI Literacy: An Exploratory review [9].

Susanna Lavantesi et al. (2021) have developed Machine Learning and Financial Literacy: An Exploration of Factors Influencing Financial Knowledge in Italy [10].

Khritish Swargiary et al. (2022) have developed Literacy rate in India in 2022 [11].

Chun-Wei Chiang et al. (2022) have developed Exploring the Effects of Machine Learning Literacy Interventions on Laypeople’s Reliance on Machine Learning Mode [12].

Irene Lee et al. (2021) have developed Indian Literacy Analysis Using Machine Learning Algorithms [13].

Mohammed Ashfaq et al. (2021) have developed an Analysis of Literacy Rates in Karnataka [12].

Chun-Wei et al. (2021) have developed Using digital story writing as a pedagogy to develop AI literacy among primary students [14].

**TABLE:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Year** | **MODEL USED** | **PARAMETERS** | **MERITS** | **DEMERITS** |
| 2018[2] | Machine Learning Models (Predictive Analysis Model and Classification Model) | * Learning Rate * Mean Absolute Error * Root Mean Squared Error * Evaluation Metrics | * Data-Driven Decision Making * Comprehensive Analysis * Visual Interpretations * Predictive Analysis * Prescriptive Analysis * Government Benefits | * Data Quality and Availability * Complexity of Factors * Model Complexity * Ethical Concerns * Cultural and Contextual variation |
| 2020[1] | Multiple Linear Regression Model | * Literacy Rate * Net Attendance Ratio (NAR) * Government Expenditure on Education * Government Policies | * Relevance and Importance * Data-Driven Analysis * Policy Implications * Awareness and Education * Focus on Net Attendance Ratio (NAR) | * Simplistic linear regression model * Causation vs Correlation * Data Quality * Simplistic view of Education * Sole focus on Government Policy |
| 2020[5] | Logistic Regression, Decision Trees, SVM, KNN, Naïve Bayes, Gradient Boosting | Independent Variables (   * Demographic data, * Academic History * Attendance * Study Habits   -Dependent Variables (   * Performance) | * Improved Student Retention and Performance * Informed Decision-Making * Tailored Learning * Efficiency in Resource Allocation * Contribution to Educational Research | * Data Privacy and Ethics * Bias and Fairness * Overreliance on Predictive Models * Limited Contextual Understanding * False Positives and Negatives |
| 2021[6] | N/A | * Inter-relation of AI * Multi-level data Literacy | * Holistic Understanding * Real-world Relevance * Cross-disciplinary Learning * Cognitive Skills Development * Long-terms Benefits | * Curriculum Overload * Teacher Preparedness * Rapidly Changing Landscape * Resource Constraints * Assessment Difficulties * Equity Concerns |
| 2019[3] | K- Means Clustering,  Data Mining | * Evaluation Metrics * User demographics * Content Interactions | * Enhanced Media Literacy * Practical Skills * Real-world Application * Critical Thinking * Awareness of Ethical Issues | * Complexity * Lack of Relevance * Fast-paced Technological Change * Resource Requirements * Unintended Consequences |
| 2021[8] | Parametric Models (Linear and Logistic regression),  Decision Trees,  Random Forests,  Gradient Boosting | * Financial literacy level of Individuals * Demographic Factors * Learning rate * Number of Estimators * Random seed | * Innovative Approach * Complex Data Handling * Holistic Insights * Performance Comparison * Policy Implications * Practical Relevance * Encouraging Integration | * Limited Generalization * Data Availability * Overfitting Concerns * Interpretability * Data Collection Challenges * Algorithm Complexity * Potential Bias |
| 2012[12] | Quantitative and Qualitative analysis,  Comparative analysis | * Literacy Rate * Decadal Growth * Co-efficient of variation * Gender, Regions * Benchmark Literacy Rate | * Comprehensive Analysis * Data-Driven Insights * Identification of Successes * Focus on challenges | * Limited Contextual Information * Lack of Causality Analysis * Simplistic Analysis of Social Groups * Limited Policy Recommendations |
| 2021[11] | Linear Regression,  Decision Tree,  Random Forest, ANN and Data Visualization | * Literacy rate * GDP * Geographical Parameters (States, Districts) * MSE, R-Squared * React, Angular, Vue.js * Flask, Django | * Data-Driven Insights * Visualization * Identifying Patterns and trends * Prediction * Policy Implications * Comparative Analysis | * Data Quality * Model Complexity * Overfitting * Data Privacy and Ethics * Interpretability * Assumptions and Limitations |
| 2021[7] | N/A | * AI Literacy * Teaching Methods * Ethical Considerations * Discourse Contribution | * Clarity in Understanding * Foundation for Education * Interdisciplinary Perspective * Future Career Readiness * Setting a Discourse | * Rapidly Evolving Field * Complexity * Lack of Consensus * Resource Intensive * Cultural and Contextual Variation |
| 2021[9] | Data Analysis and Visualization | * Workshop Focus * Curriculum * Outcomes * Challenges and Opportunities * Accessibility | * Informed and Critical Users of AI * Early Exposure to AI * Ethics Education * Career Readiness | * Age Appropriateness * Depth vs. Breadth * Sustainability * Technical Depth * Online challenges |

1. **RESEARCH OBJECTIVE**

* **Predictive Modelling:** Develop machine learning models to predict literacy rates for regions, countries, or specific demographics based on various socio-economic, educational, and cultural factors.
* **Early Warning System:** Develop a machine learning-based early warning system to detect regions or demographics at risk of declining literacy rates.
* **Data Quality Assessment:** Evaluate the quality and reliability of literacy rate data sources, and develop machine learning models to detect and address data anomalies or inconsistencies.

1. **METHODOLOGY**

Analysing literacy rates using involves collecting and preprocessing data, selecting appropriate features, choosing a suitable machine learning algorithm, training the model, and evaluating its performance.

**ARCHITECTURE:**

Data Preprocessing

Data Collection

Business Understanding

Data Understanding

Data Modelling

Data Preparation

Deployment and Maintenance

Visualization and Reporting

Results Evaluation

**STEPS:**

**1. Data Collection:** Collecting data from various sources.

**2. Data Preprocessing:**

- Normalizing the data**.** And splitting the dataset into training and testing sets.

**3. Business Understanding:**

- Outlining the questions, we will answer along with why they are relevant and important.

**4. Data Understanding:**

- Understanding the whole data and providing stats about it.

**5. Data Preparation:**

- Showing how we clean and prepare the data.

**6. Data Modelling:**

-  If we create a model, in this section we would put it in of the analysis.

**7. Results Evaluation:**

**-** Evaluating the results.

**8. Visualization and Reporting:**

- Creating visualizations, such as scatter plots, bar charts, or heatmaps, to present the results of our analysis. And prepare a report summarizing our findings, including the model's performance, feature importance, and any actionable insights.

**9. Deployment:**

- If the model proves valuable, will consider deploying it as part of an application or system that can provide real-time literacy rate predictions or recommendations.

**10. Maintenance and Updates:**

- Regularly update the model with new data to ensure it remains accurate over time. Because, machine learning models can degrade in performance if not maintained.

1. **DATASET DESCRIPTION**

This dataset is designed for analysis and prediction of literacy rates in different countries. This dataset aims to provide data and information that can be used to achieve several important objectives like understanding literacy trends, forecasting future literacy trends, etc. This information can be used for evidence-based policymaking, program development, and efforts to improve educational outcomes and reduce literacy disparities.

Here, we are using only 1 dataset for analysing literacy rates. That is District and state-wise primary & secondary school education data (2015-2016) dataset [15].

District and state-wise primary & secondary school education dataset (2015-2016) has 8 attributes. Here are some operations done on some attributes:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Attributes** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| STATCD | 36 | 18.5 | 10.535654 | 1 | 9.75 | 18.5 | 27.25 | 36 |
| DISTRICTS | 36 | 18.888889 | 15.947732 | 1 | 8 | 13.5 | 27.75 | 75 |
| BLOCKS | 36 | 203.25 | 220.756928 | 1 | 35.75 | 135.5 | 306.25 | 971 |
| VILLAGES | 36 | 16520.5 | 21166.73977 | 10 | 1038.5 | 8686 | 23341 | 98470 |
| CLUSTERS | 36 | 2289.833333 | 2645.161756 | 7 | 121.5 | 1575.5 | 3865 | 10594 |
| TOTPOPULAT | 36 | 58626.33333 | 165584.0557 | 64 | 1348.25 | 14651 | 60570.75 | 991348 |
| GROWTHRATE | 36 | 19.990556 | 11.455856 | -0.47 | 13.88 | 18.91 | 22.6375 | 55.5 |
| SEXRATIO | 36 | 930.888889 | 79.126581 | 618 | 904.25 | 946.5 | 975.75 | 1084 |
| OVERALL\_LI | 36 | 78.309444 | 8.257752 | 63.82 | 70.9375 | 78.48 | 86.2875 | 93.91 |
| MALE\_LIT | 36 | 84.985 | 6.468443 | 73.39 | 80.1925 | 85.92 | 90.6125 | 96.11 |
| FEMALE\_LIT | 36 | 71.091389 | 10.581109 | 52.66 | 60.4475 | 71.25 | 79.925 | 91.98 |
| AREA\_SQKM | 36 | 94501.11111 | 102321.9097 | 32 | 9926.75 | 54578 | 140320 | 342239 |

**Target variables:** The target variables in this dataset are “OVERALL\_LI” (which shows overall literacy rate of different states),” MALE\_LIT” (which shows male literacy rate of different states) and “FEMALE\_LIT” (which shows female literacy rate of different states).

1. **RESULTS AND DISCUSSIONS**

Provide an overview of the data used, including timeline and geographical area.

Describe the variables involved, such as literacy rate, age, sex, and region. Provide a summary of basic statistics, such as average literacy rate, standard deviation, minimum value, and maximum. Create a visual graph such as a histogram or box plot to show the distribution of literacy rates. If available, examine how literacy rates have changed over time. Use imagery or timeline analysis to visualize trends. Assess gender inequality by examining gender-specific literacy rates. Calculate and present gender equality indicators to measure gender equality in literacy. Assess gender inequality by examining gender-specific literacy rates. Calculate and present gender equality indicators to measure gender equality in literacy. Discuss the socio-economic implications of the observed literacy rates. Emphasize the potential benefits of improving literacy rates, such as improved employability, improved health, and reduced poverty. Based on the research, identify policy recommendations to improve literacy rates. Consider ways to address gender and community differences, as well as promote education at different age levels. Acknowledge any limitations of the study, such as data limitations, potential bias, or missing variables. Discuss the potential impact of these limitations on the validity of the findings. Summarize the main findings and implications of the literacy rate survey. Restate the importance of addressing literacy rates for socio-economic development. Suggest areas for future research, such as conducting in-depth research in specific areas or examining the effectiveness of specific educational interventions.

1. **CONCLUSION AND FUTURE SCOPE**

Literacy studies have shown positive trends in many industries around the world. Governments, institutions, and communities have made significant efforts to improve access to education, increasing literacy rates. There is a gender gap in literacy rates in many parts of the world, and women and girls tend to have lower literacy rates compared to men and boys Addressing this gap should be especially in future efforts. Literacy rates vary widely between regions and countries. Socioeconomic development, access to quality education, and cultural factors play important roles in these changes. With the increasing importance of digital technologies, digital literacy is becoming an integral part of all literacies. Future literacy rate research should include digital literacy research. Simply increasing literacy rates is not enough; A good education is also important. Future research should examine the provision of quality education in greater detail to ensure that individuals are not only literate but also with appropriate skills.

Conduct more in-depth research at the regional or national level to identify specific factors that contribute to changes in literacy rates. This can help policymakers better target interventions. Examine how factors such as gender, socioeconomic status, and ethnicity interact with literacy rates. This will provide a nuanced understanding of the challenges faced by marginalized groups. Conduct comprehensive studies to monitor changes in literacy rates over time, to better evaluate the effectiveness of literacy programs and programs. Given the increasing importance of digital skills, future studies should include an assessment of digital literacy in addition to traditional literacy rates. Examine the impact of the COVID-19 pandemic on literacy rates, as waste of information in education is likely to impede literacy progress. Investigate strategies to mitigate these effects.

1. **REFERENCES**

**[1]** Mansift Kaur B.A. (Honours) Economics University of Delhi, Sri Guru Tegh Bahadur Khalsa College (2020) *An Econometric Analysis of Literacy Rates in Different States of India and Factors Stimulating Them* [Reference Research Paper](http://www.aiirjournal.com/uploads/Articles/2020/05/4563_04.Mansift%20Kaur.pdf)

**[2]** Vaidehi Nimje, Aboli Kulkarni, Prajakta Kulkarni (2018), *Analysis and Forecast of Literacy Rates in India* [Reference Research Paper](http://www.jetir.org/papers/JETIR1812965.pdf)

**[3]** Teemu Valtonen University of Eastern Finland, Finland Matti Tedre University of Eastern Finland, Finland Kati Mäkitalo University of Oulu (2019), *Media Literacy Education in the Age of Machine Learning* [Reference Research Paper](https://files.eric.ed.gov/fulltext/EJ1227158.pdf)

**[4]** Siu-Cheung Kong, William Man-Yin Cheung (2021), *Evaluation of an artificial intelligence literacy course for university students with diverse study backgrounds* [Reference Research Paper](https://www.sciencedirect.com/science/article/pii/S2666920X21000205)

**[5]** Akpofure AEnughwureNigeria Maritime University, Okerenkoko (May 2020), *Application of Machine Learning Methods to Predict Student Performance: A Systematic Literature Review* [Reference Research Paper](https://www.researchgate.net/publication/341674171_Application_of_Machine_Learning_Methods_to_Predict_Student_Performance_A_Systematic_Literature_Review)

**[6]** Viktoriya Olari Computing Education Research Group Freie Universität Berlin Berlin, Germany viktoriya.olari@fu-berlin.de Ralf Romeike Computing Education Research Group Freie Universität Berlin Berlin, Germany (October 2021), *Addressing AI and Data Literacy in Teacher Education: A Review of Existing Educational Frameworks* [Reference Research Paper](https://www.researchgate.net/profile/Viktoriya-Olari/publication/355415488_Addressing_AI_and_Data_Literacy_in_Teacher_Education_A_Review_of_Existing_Educational_Frameworks/links/6214a950eb735c508ae92eac/Addressing-AI-and-Data-Literacy-in-Teacher-Education-A-Review-of-Existing-Educational-Frameworks.pdf)

**[7]** Davy Tsz Kit Ng, Jac Ka Lok Leung, Samuel Kai Wah Chu, (2021*) Conceptualizing AI literacy: An exploratory review* [Reference Research Paper](https://www.sciencedirect.com/science/article/pii/S2666920X21000357)

**[8]** Susanna Lavantesi and Giulia Zacchia, (13 March 2021), *Machine Learning and Financial Literacy: An Exploration of Factors Influencing Financial Knowledge in Italy* [Reference Research Paper](https://www.mdpi.com/1911-8074/14/3/120)

**[9]** Irene Lee MIT STEP Lab Cambridge, Massachusetts Safinah Ali MIT Media Lab Cambridge, Massachusetts (2021*), Developing Middle School Students’ AI Literacy* [Reference Research Paper](https://dl.acm.org/doi/pdf/10.1145/3408877.3432513)

**[10]** Chun-Wei Chiang Purdue University West Lafayette, Indiana, USA, Ming Yin Purdue University West Lafayette, Indiana, USA (2022) *Exploring the Effects of Machine Learning Literacy Interventions on Laypeople’s Reliance on Machine Learning Models* [Reference Research Paper](https://dl.acm.org/doi/pdf/10.1145/3490099.3511121)

**[11]** Shubhi Jain, Sakshi Bindal, Ruchi Goel, Gaurav Aggarwal (2021) *Indian Literacy Analysis Using Machine Learning Algorithms* [Reference Research Paper](https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119752134.ch14)

**[12]** Mohammed Ashfaq AhamedInstitute for Social and Economic Change (2021) *Analysis of Literacy Rates in Karnataka* [Reference Research Paper](https://www.researchgate.net/publication/255968491_ANALYSIS_OF_LITERACY_RATES_IN_KARNATAKA_-_REFLECTIONS_FROM_CENSUS_DATA)

**[13]** Khritish Swargiary and [Kavita Roy](https://www.researchgate.net/profile/Kavita-Roy?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19)Gauhati University (2022*) Literacy rate in India in 2022* [Reference Research Paper](https://www.researchgate.net/publication/363173220_LITERACY_RATE_IN_INDIA_IN_2022)

**[14]** Chun-Wei Chiang and Wing Yin (2021), *Using digital story writing as a pedagogy to develop AI literacy among primary students Reference Research Paper*   
**[15]** District and state-wise primary & secondary school education data, 2015-2016 [Dataset Link](https://www.kaggle.com/datasets/rajanand/education-in-india)

**[16]** Literacy of Indian Districts from Government of India 2011 Census Data [Dataset Link](https://www.kaggle.com/datasets/satyampd/india-literacy-data-district-wise)

**[17]** Youth and Adult Literacy Rates  [Dataset Link](https://www.kaggle.com/datasets/thedevastator/youth-adult-literacy-rates-in-2019)

**[18]** India's state-wise literacy data from RBI publication in Annual report  [Dataset Link](https://www.kaggle.com/datasets/gokulrajkmv/indian-statewise-data-from-rbi)

**GITHUB LINK:**

[**Click here**](https://github.com/PreethamKasarla/DAFE_FinalResearchWork_TeamG9_AIML-A/upload)