ANALYSIS AND FORECAST OF LITERACY RATES IN INDIA

Vaidehi Nimje[1], Aboli Kulkarni[2], Prajakta Kulkarni[3], Ishwari Mahajan[4], Geetha. R. Chillarge[5]
[1]B.E. Student, [2]B.E. Student, [3]B.E. Student, [4]B.E. Student, [5]Assistant Professor
[1]Department of Computer Engineering,
[1]MMCOE, Pune, India

Abstract : The literacy rate is one of the strategic indicators of a country's economic condition as improved literacy rate leads to augmentation of a country's human capital. Study, analysis and forecast of various features affecting a country's literacy has to be carried out to plan improvements in education system so as to increase literacy rate. In this paper, we aim to review the methods to study the trends and patterns in literacy rates over the years, predict the literacy rates based on the available enrollment and population data and present the recommendations to improve the literacy rates. The proposed machine learning model will have to analyze the statistics of literacy based upon multiple factors like Gender, Population, Social classes and categories, Religious communities, Location (urban/rural) and enrollment and dropout rates, etc. It will then give a detailed insight into the trends in literacy in every discipline mentioned above, in the form of visual interactive interpretations (graphs and charts) and perform prescriptive analytics for a targeted region.

IndexTerms **- Literacy rate, predictive analysis, prescriptive analysis, data analytics, visualization, machine learning.**_

I. INTRODUCTION

Literacy has always been an important topic of interest for the world. Every country has the aim to achieve full literacy rate as it depicts the country's development. Although literacy rate has increased up to a great extent now, there is a need to know the areas that are still lagging behind and work for their improvement. A country's economy gets enhanced when its population has higher literacy levels. Better literacy rates lead to increase in educational and employment opportunities for the people. Hence the people are pulled out of poverty and chronic underemployment. In this increasingly complex and rapidly changing technological world, it is important that individuals continuously expand their knowledge and learn new skills in order to keep up with the pace of the world.

India being a developing country has much to do in the sector of education so as to achieve a 100% literacy rate. Being a diverse and huge nation that ranks second in terms of population, India still lags behind when it comes to literacy despite the efforts taken in the field of education. There is a need for detailed, directed analysis of the trends observed around every state of India in terms of education and literacy, get patterns over the years and implement different and target specific measures to improve literacy in all the zones.

So, study and analysis of literacy data is required to establish or contribute to an education system for collection, organization and utilization of education data so as to provide a timely and informed basis for helping planning and management of education services across the nation.

To enable this, data analytical and machine learning techniques can be applied over the huge data related to the country's literacy collected over the years. Data analytics (DA) is the process of examining large data so as to draw conclusions about the information they contain and Machine learning is a branch of artificial intelligence (AI) that gives systems the ability to learn automatically and improve from previous experience without being explicitly programmed.

Data Visualization of our analyzed literacy data will be helpful to convey the important insights to the end users such as common people, government policy makers and NGO workers, etc. though our system. Data visualization is the representation of data in a graphical or pictorial format which enables decision makers to see analytics presented visually. Visualization helps to grasp difficult concepts or identify new patterns in the big data. Interactive visualization is a step ahead in which technology is used to drill down into charts and graphs to get more detailed, interactively changing overview of data. Because of the way the human brain processes information, using pictorial representations like charts or graphs to visualize large amounts of complex data is an easier and better alternative than analyzing spreadsheets or reports. Data visualization can convey concepts in an easier manner. It helps to identify areas that need improvement or attention.

Further, a predictive analysis can be applied over the data to get a forecast of the literacy rates of the states in India in a chronological order. It will give us a vision of the future which will help in planning and improving the educational scenario all over the nation. Some of the algorithms that can be applied are Multiple regression, Random Forest, XG boost. The patterns and differences observed in different states can be compared and conclusions upon the rate of development of different regions can be obtained. A prescriptive analysis can be performed upon our data. The intention of prescriptive analytics is to suggest what action to take to curb a future problem or to get better future prospects. In case of the regions showing poor growth in literacy, the system can suggest measures undertaken by other regions that show relatively better growth. In this way a complete analysis of the literacy statistics of the nation can be carried out.

II. LITERATURE SURVEY

1. FORECASTING OF LITERACY RATE USING STATISTICAL AND DATA MINING METHODS

In this research paper, the principle feature for forecasting literacy is considered to be Population of a region. Hence, the projection of future population trends is carried out first and then predictive algorithm is used to forecast literacy rate. The total Population, male population and female population of the state of Chhattisgarh is projected using a statistical method called logistic curve method. This method is used when the growth rate of population due to births, deaths and migrations happen under normal situation and it is not subjected to any extraordinary changes like natural disasters, war or epidemic, etc. The growth curve characteristics of living things in limited space and economic opportunity is followed by population. The curve obtained when the population of a region is plotted with respect to time, under normal condition looks like S-shaped curve that is known as logistic curve. From these projected populations, literacy rate is forecasted using a data mining method of multiple regressions for the state. The multiple linear regression is used to explain the relationship between one dependent variable which is continuous and two or more independent variables. Here, literacy rate is the dependent variable which is predicted using multiple regression upon the independent variable- male population and female population which are obtained from projections.

2. LITERACY RATE ANALYSIS

In this paper 5 different algorithms are given that can be used for analyzing the data.

1.IDE3 - Decision tree algorithm (Iterative Dichotomiser 3. It is implemented in serial and this is based on Hunt's algorithm. The tree is constructed in two phases same as other decision tree algorithms - tree growth and tree pruning. The attributes accepted in building a tree are only categorical attributes.

2. C4.5 - The improvement of IDE3 algorithm is C4.5 algorithm. By replacing the internal node with a leaf node Pruning takes place in C4.5 thereby reducing the error rate. C4.5 accepts both continuous and categorical attributes in the process of building a Decision tree. It has an enhanced tree pruning method that decreases the misclassification errors due to too much details in the training data set. To determine the best splitting attribute the data is sorted at each and every node of the tree. Gain ratio impurity method is used to evaluate the splitting attribute.

3. CART (Classification and regression trees) - CART builds both classifications and regressions trees. The classification tree construction is based on binary splitting of the attributes. For building the decision tree CART uses both numeric and categorical attributes. CART has in built feature that deals with missing attributes.

4. SLIQ (Supervised Learning In Ques) - It is a fast, scalable decision tree algorithm that can be implemented in serial as well as parallel pattern. SLIQ is not based on Hunt's algorithm which is used for decision tree classification. It partitions a training data set recursively using breadth-first greedy strategy which is integrated with pre-sorting technique during the tree building phase. In the pre-sorting technique sorting at decision tree nodes is eliminated and replaced with one-time sort. To determine the best split point list data structure for each attribute is used. Disadvantage of SLIQ is that it uses list data structure which is memory resident which imposes memory restrictions on data.

5. SPRINT (Scalable Parallelizable Induction of decision Tree algorithm) - Same as SLIQ SPRINT uses one time sorting of the data items and advantage is that it has no restriction on the input data size. SPRINT uses two data structures- attribute list and histogram which is not memory resident and which makes SPRINT suitable for large data set and in this way it removes all the data memory restrictions on data. It handles both continuous and categorical attributes.

In the proposed approach the algorithm C4.5 was used because of following advantages:
1. Handling of both the continuous and discrete attributes - To handle continuous attributes, C4.5 algorithm uses a concept of threshold by creating one threshold and then splits the list into those whose attribute value is above the threshold and those below or equal to it.
2. One of the advantages is handling of training data with missing attribute values - C4.5 allows to mark attribute values as "?" for missing values. Missing attribute values are not considered in gain and entropy calculations.
3. Handling attributes with differing costs.
4. Pruning of trees after creation is an important advantage because - C4.5 goes back through the entire tree once it has been created and this algorithm attempts to remove that branches which do not help by replacing them with leaf nodes.

3. EDUPAD- A Tablet Based Educational System for Improving Adult Literacy in Rural India

Several methodologies have been introduced by UIS for improving literacy. Overall rise in adult literacy rates is one of its effects. This has led to reduction in the illiterate population. Global education agenda is that everyone should have primary education.
Most of India's population is situated in rural areas. The rural areas of India are often deficient of education. Our literacy rate doesn't meet the international standards. Lack of primary education is one of the major hinderance to literacy.
Foundation for the National Literacy Mission was laid by Prime Minister of India, Mr. Rajiv Gandhi. Many projects have been implemented which make use of technology to reach masses. One such example is television sets were used to teach languages with the help of subtitles. Another is children were taught English with the help of cell phone games. But the

problem lies with illiteracy of adults. So, the paper has proposed tablet based educational system, called EduPad, which can considerably reduce the literacy problem faced in rural areas. It uses interactive way than the conventional class room system which we use traditionally.
EduPad is used to reduce the rural adult illiteracy using advancements in technology. It looks after the lack of adequate infrastructure in rural India.
The device proposed here is an interactive Tablet, which teaches multiple languages. It consists of developing interactive educational software which can run on the tablet. The software helps the user to learn how to write and spell the alphabets. Initially the software teaches alphabets. After which it moves onto words and then sentences.
Research has proved access to primary education depends on the wealth of the family. Families which are poor cannot afford to send their children to school. Children of families having minimal land and monetary holdings are forced to do manual labor which leads to these children becoming illiterate. This increases the rate of illiteracy. Because of projects undertaken by government of India and NGOS literacy rate has increased. Literacy opens door for opportunities and improved standard of living. Illiteracy traps people in poverty and problems.
The EduPad based educational system will be helpful to the illiterate people of rural India to become literate. It is an interactive and enjoyable method. People can get educated without affecting their day to day life.
Hence it can be inferred from this paper that formulating new and innovative techniques to overcome the adult illiteracy challenge faced by Indian population have to be initiated in rural areas so as to attain complete literacy. Most of the illiterate people of rural area rely on manual labor for their living. They cannot take out time for education. So, the technological solution like EduPad can be a convenient method for rural India to become literate. It is easy and people can use it to learn anything at anytime and anywhere. Government needs to formulate such innovative and effective policies in the targeted regions of the country that considerably lag behind other states in terms of literacy rates so as to bring about their development and country's overall progress.

4. WHEN WILL INDIA ACHIEVE UNIVERSAL ADULT LITERACY: Status and Prospects

This paper inspects the rankings and differences among sub-population groups which are determined by gender, caste, location and across the states with respect to literacy. Also, the expected literacy rates in India with a modelling of simulation exercise while considering different policy interventions are discussed in this paper. In this respect, it is observed that the performance shown by India in literacy rate is relatively poor. The progress in the literacy rate especially during the last decade has slowed down when compared to that of the previous decade. The gender gaps, rural-urban differences, social group disparities and regional variations across states continue to persist. The measures taken for improving the rate of literacy through adult education programs yielded very poor results. The simulation exercise in this paper has shown that it is impossible to achieve 100% literacy rate in the near future for the country unless there is a policy intervention by the means of adult education programs. This simulation exercise shows the need for regeneration of the National Literacy Mission (NLM) and rebuild adult literacy programs of TLC and PLP.
This paper consists of Section I which presents the status of literacy levels in India. Section II contains details about the expected literacy in India through simulation exercise. Section III discusses on policy initiatives and issues with respect to adult education programs.
The simulation exercise prospects that India can achieve 95% of literacy rate by the year 2050 if all the states carefully implement the adult literacy programs for at least 5 years period from 2012 for the age group 15-35 years covering all those illiterates in the group. Moreover, if these programs are extended to the 35-60 years age group and is implemented for at least 5 years period from 2012, India can achieve 100% literacy rate by 2050.
Thus, it can be said that achieving the goal of 100% literacy rate in the country depends on its policy intervention through adult literacy programs, careful implementation and their coverage. Unless India improves its literacy levels remarkably, it will remain one of those poor performing countries at the global level in terms of HDI and ranking of countries based on it. Moreover, the population below poverty line and illiterates require the enabling of skills of literacy to cope up with the ever- changing economic system in the context of emerging knowledge based economy and globalization.

III. DATA SOURCES

The literacy related data is obtained from India's census website, Census 2011 website and government's open data platform. The data is very vast and heterogeneous in nature. It gives literacy rates over the years, statistical information regarding features like population, sex ratio, age, etc. and data based on various categories such as religious backgrounds, social classes and communities, etc.

IV. PROPOSED METHODOLOGY

4.1 Pre-Processing
Data on which we are going to work comes from different sources and display heterogeneous nature of parameter selection, time range and data representation. Hence, this vastly varying data needs to be analyzed and transformed to a standard format before making predictions. Preprocessing of data involves:

4.1.1. Sampling: Determination of the number of time periods (year) for which data will be evaluated.

4.1.2 Outlier Detection: Another consideration while making predictions is to find the presence of extreme observation called outliers if any and handle them so that they don't hamper accuracy of the system.
4.1.3 Interpolation or Extrapolation: When intercensal data are available, values for intermediate years between censuses can be found using linear interpolation or extrapolation. This is done by calculating the average annual percentage change.

4.2 Data Visualization
Summarization of the large data that is available in the form of spreadsheets will be done to represent the information in an interactive pictorial format (charts, graphs, etc,) to learn about the trends observed in literacy rates over the years based upon various aspects such as gender, location- rural-urban, population, social classes, etc. in a detailed overview of each for every state of India. This will help to easily convey important insights obtained by analyzing the datasets and obtaining patterns and trends which can help predict future literacy scenarios.

4.3 Feature Selection
It is intended to find relationship between the literacy rate and demographic features like population growth, sex ratio, age that have major effect upon the literacy rate. Feature selection for the purpose of prediction has to be carried out in this phase.

4.3 Predictive analysis
The different algorithms that will be tested for prediction are as following:

1. Multiple regression
   In this method, literacy rate (dependent variable) is predicted as a function of one or more independent variables. The value of the dependent variable (Y) is predicted using mathematical equation for a straight line on the basis of the independent variables (X_i).

$$Y = a + b_1X_1 + b_2X_2 + \cdots + biX_i + e$$

Where, a – Y-intercept, which is the expected value of Y when X=0
b_i – Regression coefficient for X_i, which is the amount by which Y changes for unit change in X_i
e – Error term, the difference between actual value of Y and its predicted value

2. Random forest

The **random forest** algorithm is a type of additive model that makes predictions by combining decisions from a collection of base models.

$$g(x)=f_0(x)+f_1(x)+f_2(x)+...$$

here, the final model g is the sum of simple base models f_i. Each base classifier f_i is decision tree. This technique of making use of multiple models to obtain better predictive performance is called model ensembling. In random forests, a different subsample of the data is used by each base model and are constructed independently. Hence for analyzing literacy datasets, decision trees having different attributes can be utilized and a random forest can be generated out of them that predicts literacy rate more accurately.

4.4 Accuracy Measure

Root mean square error is used to measure the difference between value predicted by model and the actual value of the dependent variable. If we represent actual data by At and forecasted value as Ft and n representing number of forecasts made then root mean square error (RMSE) is given by:

$$RMSE = \text{sqrt}\ (1/n\ .\ \sum(At - Ft)^2)$$

4.5 Prescriptive analysis

A region-specific prescriptive analysis will be performed over the states of India. The literacy patterns in various aspects like gender, age, etc. will be observed to make apt conclusions and suggestions. The poorly performing regions will be given suggestions based upon the comparative study of measures and policies that helped other regions to improve literacy. The suggested measures will aim to help improve literacy rates of those regions.

V. CONCLUSION

Looking at the performance of India in education and literacy and studying the progress rate in the same, it is evident that significant measures have to be undertaken for improving current scenario. A technological approach of a complete analytical model will thus be of help. From the study of various techniques to perform data visualization, predictive analysis and prescriptive analysis and the available data, a system for detailed, summarized as well as targeted analysis of literacy rates in states of India can be implemented which will benefit the government and common masses.

REFERENCES

[1] Swati Jain, Nitin Mishra. 2017. Forecasting of Literacy Rate using statistical and data mining methods. International Journal of Advanced Computational Engineering and Networking, 2015
[2] Tarun Verma, Sweety Raj, Mohammad Asif Khan, Palak Modi. 2012. Literacy Rate Analysis. International Journal of Scientific & Engineering.
[3] Rajesh Kannan Megalingam, Ananthakrishnan P. Rajendran, Abhiram T. Solamon, Deepak Dileep. 2012. EduPad- A Tablet Based Educational System for Improving Adult Literacy in Rural India. IEEE International Conference on Technology Enhanced Education (ICTEE)
[4] Venkatanarayana Motkuri. When Will India Achieve Universal Adult Literacy: Status and Prospects. MPRA Paper No. 48061.
[5] Dr. Sudhir B. Jagtap. 2013. Census Data Mining and Data Analysis using WEKA. International Conference in "Emerging Trends in Science, Technology and Management-2013, Singapore.
[6] Literacy Rates Continue to Rise from One Generation to the Next , 2017, UNESCO.
[7] Census of India Website, censusindia.gov.in
[8] Census 2011India, www.census2011.co.in
[9] Open Government data platform, event.data.gov.in
[10] Wikipedia, https://en.wikipedia.org/wiki/Demographics_of_India