2nd National Conference on Engineering Applications of Emerging Technology in association with International Journal of Scientific Research in Science, Engineering and Technology | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
[https://ijsrset.com | doi : https://doi.org/10.32628/IJSRSET]

Representational Analysis of Indian Census Data

Sathvik R G[1], Srinarayana Tantry[1], Vidhi P Jain[1], Vyasa Krishna K[1], Kamleshwar Kumar Yadav[2]
[1]Student, [2]Assistant Professor
Department of Computer Science & Engineering Global Academy of Technology, Bengaluru, Karnataka, India

ABSTRACT

There is an abundance of data that is available in print format, that was collected and compiled for many centuries since the dawn of the printing press. A vast majority of this data, even if available on the internet, is in a very archaic form and not easily comprehensible for the average person. As a result of this, essential information tends to remain unknown by the vast majority of people. With the advancement of mass computation techniques and visualization tools that have become popular in recent years, circumstances are beginning to change and there is renewed enthusiasm by both the experts and the general public in digging up past data. In our paper, we seek to represent and fulfil the desires of those data enthusiasts and try to contribute to this growing movement. The initial step of this project is to extract data from the Indian census, through which data was collected every ten years starting from 1871. The latest census undertaking was conducted in the year 2011. Details from the National Family Health Survey (NFHS) and other government surveys will also be considered as a part of the dataset. We will use statistical methods and Machine Learning algorithms on the data that we gathered, which will enable us to achieve the main objective of the project. The purpose is to identify parameters and predict the outcomes of various policies and incentives taken by the government, which will help us arrive at conclusions on the effectiveness of different approaches to solving the socio-economic problems of society. This will lead us to the resolution of our initial objective, which is to represent our results in a format that can be easily understood and make this information accessible through appealing charts and graphs, by using Data Visualization techniques.

Keywords: Indian Census, Data Visualization, Machine Learning

I. INTRODUCTION

The Indian Census is the most comprehensive single source of statistical data on India's inhabitants and their many characteristics. This solid, time-tested process has been producing a true wealth of facts every ten years since 1872, which was when the first census was conducted non-synchronously in different parts of India. For academics and professors in demographics, economics, anthropology, sociology, statistics, and many other subjects, the Indian Census has been a fascinating source of data. The decennial census, which has become one of the most important instruments for understanding and analyzing India, clearly demonstrates the country's tremendous diversity.

Data visualization is the graphical representation of data and information. Data visualization tools, as well as visual elements such as graphs, charts, and maps, are used to make it easier to recognize and evaluate patterns, anomalies, and characteristics of the data. From art and advertising to tv and films, everything in our life is visual. Data visualization is another sort of visual art that keeps our focus on the data.

Machine learning is a subfield of artificial intelligence that enables software to increase its accuracy rate without being specifically designed to do so. Machine learning approaches employ historical data as an input to foresee new output values. Machine learning approaches are built on the foundation of models. In machine learning, emulation is a fundamental notion.

II. LITERATURE SURVEY

The paper by V Balasankar and P Suresh Varma [1] applies K-Medoids and K-means clustering methods to the 2011 Indian Census data, after pre-processing the dataset. Data is ubiquitous, and people have access to a lot of it. We can use this data to look for hidden and overlooked information and put it to good use. Census data is an essential source of information since it contains information about the people who live in a certain country. Evaluating such data is important for assessing the country's socioeconomic standing. When dealing with enormous amounts of data, data mining, and machine learning methods are frequently used. They describe the need for such an exercise quite well and state how useful it is for governments to work on such data. They have represented their output in the form of simple Bar graphs and Pie Charts. We seek to expand the scope of this endeavor by collecting data from many more previously conducted Census and also reason with our results.

The paper by Dierdre Bevington-Attardi & Michael Ratcliffe [2] has leveraged modern developments in statistical mapping and data visualization to publish the results of the US national census. The authors used a variety of ways to visualize the large amount of data from the US Census Bureau's archives, which has been a pioneer in the field of data collection. The work done by the Census Bureau of the US includes data visualization galleries, interactive data tools, business and demographics map, and much more. This paper provides a good example of the utility such endeavors to governments and business agencies alike.

The paper by Naina Bharadwaj, Prachi Mishra, and Prajyoti Dsilva [3] talks about an interactive large data visualization tool for the Census dataset should comprise seven main modules and their accompanying functions. This paper seeks to overcome the lack of interactivity constraint by improving the interactive visualization process with more relevant procedures for manipulating the generated visuals based on numerous attributes. Data gaps and discontinuities were also taken into account for visualization. The reliability factor for massive data sources has been introduced. It also illustrates why, in compared to existing visualization tools, the Census dataset requires new capabilities and modules. Every module's operation and interactions have been documented using a diagram.

The paper by Himani Rani and Dr. Gaurav Gupta [4] contains and teaches us on data mining, and it walks us through numerous techniques and datasets to get effective results. There has been an upsurge in the volume of data produced in recent years. Big data has provided a new perspective on how data is stored, handled, and used in various industries. This

is owing to the rise in wages, as well as the increased accessibility and reach of technology. Data visualization aids in the organization, comprehension, and correlation of data in a graphical or preferred format. They assist us in deciphering data behaviour and hidden patterns in a dataset. Data visualization is significant because it provides visual images that help people understand the information in a dataset. Data visualization isn't always the best way to analyze data. Companies, researchers, and others who use these tools can gain insights into their data and discover trends by employing various visualization approaches based on the data. This may aid in making quick decisions, which will benefit the user in multiple ways. The paper imparts knowledge about data mining in databases, which is a technique for extracting necessary information from raw data. The paper has borrowed ideas from multiple other papers, with the usage of algorithms such as ID3, CART, Multilayer Perceptron, Naïve Bayes and C5.5. Tools such as Weka, and RapidMiner were made use of, and the datasets of Heart Diseases, Lung Cancer, and Stock Market were worked on. They have also filtered attributes and used MATLAB, WEKA, and python to implement the model. This has led to producing efficient outputs, with certain shortcomings like complexity being high, and less accuracy on k-means clustering algorithms. Prediction analysis is the process of combining machine learning methods such as clustering and classification. To improve the accuracy of prediction analysis, this specified problem can be solved in the future.

The paper by Dr.Aniruddha S Rumale and Ms. Aishwarya Bhagwat [5] includes data visualization techniques. Data visualization is an information technology that can help to find incorrect data points and trouble in data. The exploration of data has a lot of uses like data processing and other applications in the health industry. For the most effective results, data visualization gives quicker data exploration and supplies more efficient results where the algorithms are used. Information Visualization concentrates mainly on data deficient 2D or 3D objects, there are measured representations of non-figurative information into a forcible screen. To design data with the assistance of diagrams and also the data is typically logical or special, we need scientific visual techniques like charts and graphs, etc. governing standards of perception. There's a powerful motivation to keep asking about low-cost and innovative ways of imagining the details that are sent in detail. The visual scheme should take pleasure in the hand-crafted methods where the facility assists designers to make changes in their practice. Recognition should be ready to present a range of information and will be interactive and permit effective communication. The presentation of data is known as data visualization. Static data visualization and interactive data visualization are the two types of data visualization. In an interactive data visualization, the users can give the format of data and how it should be displayed is designed by the users. Some common data visualization techniques are line plots, area plots, histograms, bar charts, pie charts, box plots and scatter plots.

The paper by A.Dinesh Kumar, R.Pandi Selvam, and K.Sathesh Kumar [6] examines the prediction algorithms and data processing tools employed in educational data processing, as well as future insights into upper prediction algorithms to be identified and new data processing tools to be familiarized with in order to predict students' performance, which aids teachers and institutions in extending their study level. Knowledge Discovery in Databases (KDD) is a huge data

warehouse analysis and modelling that is done automatically. KDD is a method for extracting legitimate, new, valuable, and intelligible patterns from large and complex data sets in a controlled manner. The heart of the KDD process is data mining (DM), which involves inferring algorithms that examine the data, create the model, and uncover previously unknown patterns. The model is used to comprehend, analyze, and predict phenomena based on data. Educational data mining is a distinct topic of data mining study. Educational data mining, or the application of data mining techniques to educational data, might be a fascinating study subject (EDM). EDM analyses data collected by educational institutions, such as student performance prediction, learning analytics, grouping students based on their performance, and making recommendations to students. EDM examines data generated by any type of information system that supports learning or education in schools, colleges, universities, and other academic or professional learning institutions that offer both a traditional teaching style and easy learning. The basic goals of educational data mining are to predict students' achievements and to model pupils. These two challenges are inextricably linked to the educational setting.

The paper by Mohini Chakarverti, Nikhil Sharma, and Rajiva Ranjan Divivedi [7] examines numerous methodologies given by many writers in order to comprehend the most recent developments in prediction analysis. Feature extraction and classification are two- step approaches in the predictive analytic techniques. The various categorization algorithms are evaluated in terms of particular factors and their outputs are compared. Data mining is a pattern for evaluating data and a method for extracting useful information. Various data mining tools are available for analyzing various sorts of information in data mining. Making judgments, analyzing the market basket, manufacturing control, customer retention, scientific discoveries, and education systems are just a few of the applications that data processing is used for. Clustering in this approach is applied to a similar cluster of data rather than the same style of data. Clusters are created by looking for similar patterns in a computer file. While categorizing genes with similar functions and in population obtain insight into structures that may be inherited in biology for the purposes of developing plant and animal taxonomies. Clustering in geology is frequently used in cities to identify related homes and land areas. Information clustering may be used to categorize all materials available on the Internet in order to discover new theories. The unsupervised data clustering classification approach forms clusters, and objects like these that are in the same cluster and are quite similar to one another are distinct. Cluster analysis is a common topic in data mining that is used for knowledge discovery. Clusters are a collection of distinct classes that contain the data items.

III. PROPOSED WORK

The existing system uses data from the 2011 Census of India to map literacy and poverty rates in different states of India. Data on social conditions such as worker categories by state, work categories by gender, and literacy levels were recorded and presented. However, current models and visualization methods do not take into account all the socio-economic parameters available in the Indian census.

The United States Census has a rich collection of data which has been used by many scholars to visualize on a scale not available for other nations.

In particular, the paper titled "Data visualization at the US census bureau– an American tradition "[reference number] by Dierdre Bevington-Attardi & Michael Ratcliffe stands out in contrast to other works in terms of detail and presentation.

Thus, in our pursuit to visualize larger and varied amounts of data in the Indian framework, different techniques have been used. In our project, we have incorporated live visualization techniques like race charts to make the presentation of information more interesting. In addition to the race charts, line graphs and column graphs will be displayed on the same data, to get a better picture of the dataset.

Machine Learning algorithms will be incorporated to work on predicting the future based on current trends. Based on the results of our model, conclusions regarding the better performance of some states over the other can be ascertained.

IV. METHODOLOGY

As shown in Fig 4.1, we understand that the data is produced from reality and recorded, mapped through surveys, reviews, etc. Gathering and measuring information from Indian Census Data and other different sources is the next step. This process of data collection should be monitored well. Data processing involves the Cleaning and Organizing of Raw Data to make it suitable for the machine learning model. Choosing certain attributes, and filtering out unwanted data comes in the process of data exploratory analysis. Implementing supervised and unsupervised learning techniques, analysis, and enhancement for the best outputs and filtering out the models that do not give efficient outputs takes place in the model and algorithmic phase. We then use this output Dataframe, visualize it using multiple techniques and tools, which is visually appealing, communicates the existence of certain aspects of the economy, etc. to make decisions in terms of government policies, see patterns and variation, what policies and acts lead to those patterns and variations to then add or remove according to the needed initiative to finally get the enhanced data product.

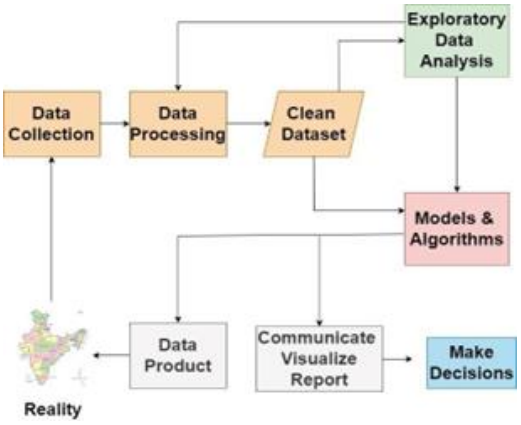

Fig. 4.1 Architecture Diagram

V. RESULTS

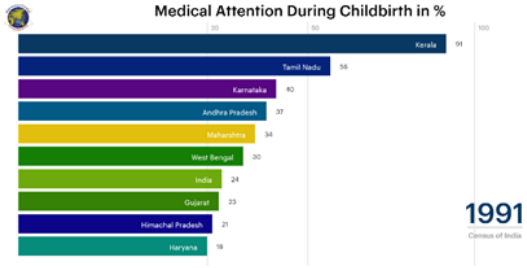

Fig. 5.1 Snippet of Bar Race Chart of year 1991

Figures 5.1, 5.2, and 5.3 are snippets of the bar race chart while showing the details about Medical Attention during Childbirth in % in various Indian states from the years 1991-2013.

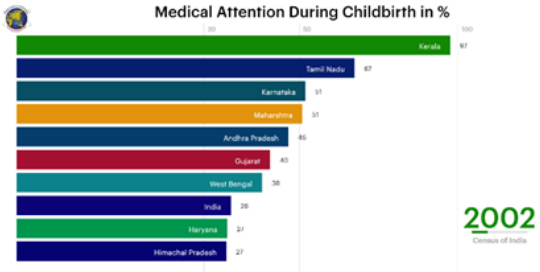

Fig. 5.2 Snippet of Bar Race Chart of year 2002

Fig. 5.3 Snippet of Bar Race Chart of year 2013

Fig 5.4 is a scatter-plot with a regression line for representing total HDI (Human Development Index) from the years 1990-2020.

Data related to fertility rate, birth rate, and death rate are visualized and predicted using machine learning regression algorithms like Decision Tree Regression, SVM regression, Linear Regression, Polynomial Regression, and Ridge and Lasso regression.

Fig 5.6 shows a bar chart for fertility rate from the years 1971-2012 and in Fig 5.7, the fertility rate for the years 2013-2022 is predicted using all the previously mentioned machine learning regression algorithms.

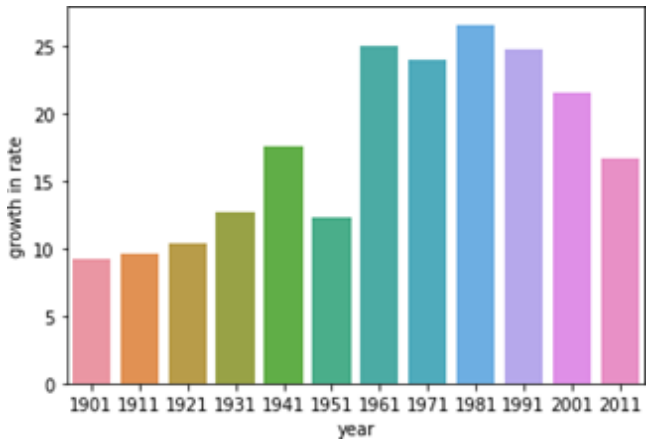

Fig. 5.5 Total literacy rate of India from the years 1901- 2011 (growth in % rate)

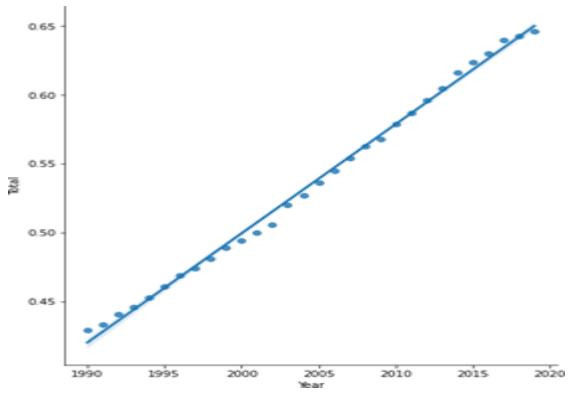

Fig. 5.4 Total HDI of India from the years 1990-2020 with a regression line

Fig 5.8 shows a bar-chart of birth rates from the years 1971-2012 and in Fig 5.9, the birth rate for the years 2013-2022 is predicted using all the previously mentioned machine learning regression algorithms.

Fig 5.10 shows a bar-chart of death rates from the years 1971-2012 and in Fig 5.11, death for the years 2013- 2022 is predicted using all the previously mentioned machine learning regression algorithms.

Fig 5.5 displays a chart showing the literacy rate of India from the years 1901-2011 and Fig 5.12 shows a chart comparing male and female literacy rates from the years 1901-2011. In Fig 5.13, literacy rates for every ten years until 2050 are predicted.

Fig 5.14 displays a chart representing the availability of drinking water state-wise throughout the country taken from 1991, 2001, and 2011 Indian censuses. Fig 5.15 represents the same for the country overall. In Fig 5.16, the availability of drinking water for the years 2021 and 2031 is predicted.

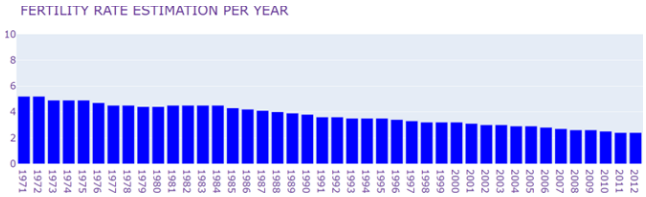

Fig. 5.6 Total fertility rate from the years 1971-2012

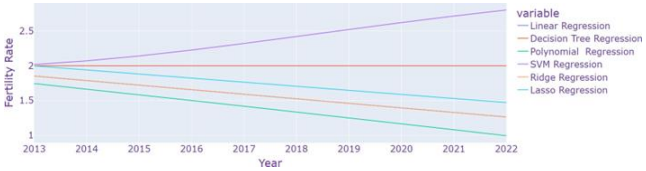

Fig. 5.7 Prediction of total fertility rate for the years 2013-2022

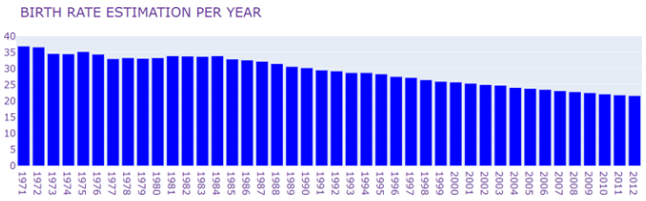

Fig. 5.8 Birth rate from the years 1971-2012

Fig. 5.9 Prediction of birth rate for the years 2013-2022

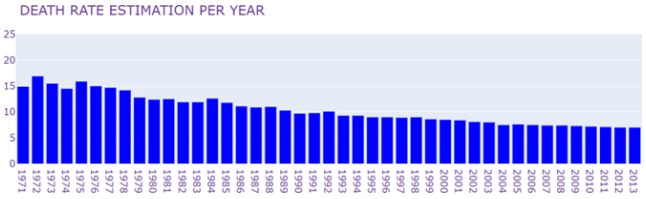

Fig. 5.10 Death rate from the years 1971-2013

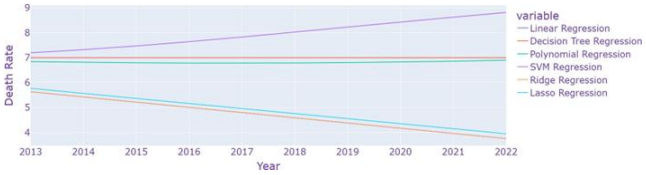

Fig. 5.11 Prediction of death rate for the years 2013-2022

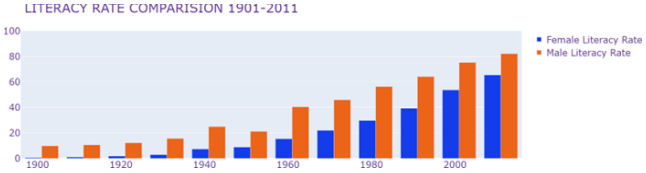

Fig. 5.12 Comparing male and female literacy rate from the years 1901-2011

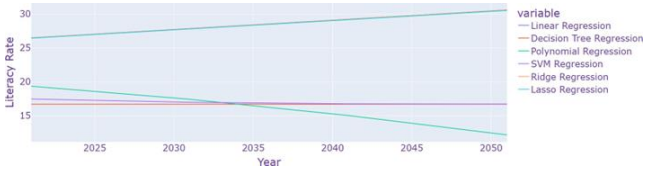

Fig. 5.13 Prediction of literacy rate for every ten years until 2050

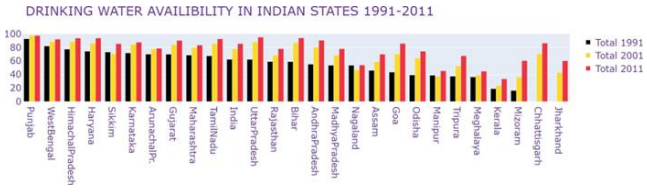

Fig. 5.14 Availability of drinking water in Indian states from 1991,2001 and 2011 censuses

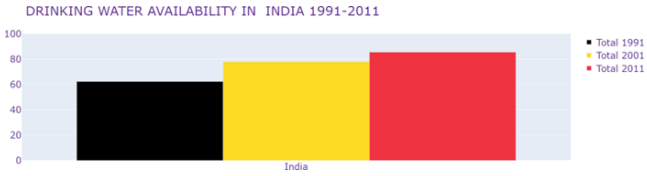

Fig. 5.15 Availability of drinking water in India from 1991,2001 and 2011 censuses

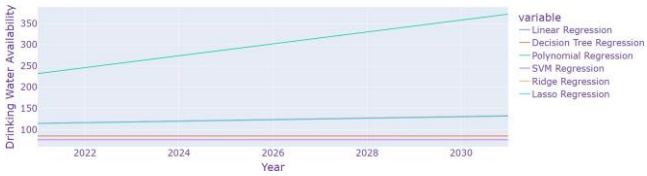

Fig. 5.16 Prediction of availability of drinking water for the years 2021 and 2031

VI. CONCLUSION

The differences in the social and economic development between different regions and societies are well known. The reason for these differences can be manifold and sometimes, not very clear. We attempt to map and correlate government policy to human development. Governments, both at local and central levels play a detrimental role in the fate of the people over which they govern. This is why finding the factors for good development in some regions and poor

performance in others will give us an opportunity to study and uniformly implement good schemes. Techniques like Data Visualization, which will be used extensively in our research and presentation of facts, will aid us in representing our conclusions and demonstrating our case in an orderly manner, in a way that can be understood by any layman. Further on, work to research socioeconomic conditions needs to be promoted and popularized, especially in emerging economies, so that maximum effort can be put to effectively improve the overall condition of the population.

VII. REFERENCES

[1]. Balasankar, P. Suresh Varma."Socio-Economical Status of India using Machine Learning Algorithms", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277- 3878, Volume-8 Issue-5, January 2020

[2]. Dierdre Bevington-Attardi & Michael Ratcliffe (2015) Data visualization at the US census bureau– an American tradition, Cartography and Geographic Information Science, 42:sup1, 63-69, DOI: 10.1080/15230406.2015.1060149

[3]. Naina Bharadwaj, Prachi Mishra, Prajyoti Dsilva. "Interactive Data Visualization for Census Data", International Journal of Computer Applications (0975 – 8887) Volume 149 – No.4, September 2016

[4]. Himani Rani, Dr. Gaurav Gupta. "Prediction Analysis Techniques of Data Mining: A Review ", International Journal of Computer Science and Mobile Computing, A Monthly Journal of Computer Science and Information Technology ISSN 2320–088X IMPACT FACTOR: 6.199

[5]. Dr. Aniruddha S Rumale, Ms. Aishwarya Bhagwat2."Data Visualization Techniques ", International Journal of Research Publication and Reviews ISSN 2582-7421

[6]. A.Dinesh Kumar, R.Pandi Selvam, K.Sathesh Kumar."Review on Prediction Algorithms in Educational Data Mining", International Journal of Pure and Applied Mathematics Volume 118 No. 8 2018, 531-537

[7]. Chakarverti, Mohini and Sharma, Nikhil and Divivedi, Rajiva Ranjan and Divivedi, Rajiva Ranjan, Prediction Analysis Techniques of Data Mining: A Review (March 11, 2019). Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE) 2019, Available at SSRN: https://ssrn.com/abstract=3350303 or http://dx.doi.org/10.2139/ssrn.3350303