Journal of Green Engineering (JGE)

Volume-10, Issue-3, March 2020

Statistical Analysis of Socio-Economic and Environmental factors in India using Green Cloud

[1]S. S. Saranya, [2]Sharmin Kantharia, [3]Atharva Hajare

[1]*Assistant Professor,Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kancheepuram, Tamil Nadu, India.*
E-mail: saranya.ss@ktr.srmuniv.ac.in.
[2]*Student,Department of Computer Science and Engineering, SRM Institute of Science and Technology,Kancheepuram, Tamil Nadu, India.*
E-mail: sharminkantharia26@gmail.com
[3]*Student, Department of Computer Science and Engineering,SRM Institute of Science and Technology, Kancheepuram, Tamil Nadu, India.*
E-mail:atharva.hajare@gmail.com

Abstract

Census is a process carried out by countries all over the world, to collect information related to housing, population, healthcare, education, etc, after fixed intervals of time. Census data is used to find out how demographics have changed over a period of time. In developing countries such as India, census data is widely utilized for policy evaluation and formation. In India, a complete census has been carried out every ten years, since the year 1881. The most recent census was conducted in 2011. It collected multiple features such as age, sex ratio, disability status, highest educational level attained, religion, housing facilities, and many more. With the advancement of Big Data in everyday life, we can now work on such large-scale data and apply the concepts of Data Mining algorithms to get a better understanding of the demographics of the country, thereby allowing us to gain better insights for the development of the country. This work is implemented in Green Cloud.

Keyword: Socio-Economic, Census Data Mining, Statistical Analysis, Exploratory Data Analysis, Visualization, Descriptive Statistics

Journal of Green Engineering, Vol. 10_3, 713-733.

Alpha Publishers

1 Introduction

A census involves the acquisition and recording of information associated with individuals in a nationwide population along with data on housing and other prevalent domains such as agriculture, business and traffic. A census, is thus an economic tool that allows us to gather and work with national data and helps us improve and develop policies and plans that can build a nation by refining the existing conditions of such policies in the fields of healthcare, education, employment, economic growth, poverty, etc. Given the widespread requirement of Data mining, we can easily utilize the algorithms and tools of the trade to achieve the goals mentioned above. Data Mining is a process that involves transforming raw data into useful information. It is part of the knowledge discovery process, where in we find interesting patterns and new, previously unknown information. Despite the present rate of technological advancements in India, we are yet to successfully apply data mining techniques using census data. For a developing country such as India, it is crucial that socio-economic issues, such as overpopulation, discrimination, unemployment, be addressed. It is therefore crucial to find relevant trends or associations between the different features that make up the census. These socio-economic trends will allow us to better understand the existing situation of India, one issue at a time. Taking on some major domains, this paper focuses on trends on disability, education and housing.

2 Literature Survey

Sheng Bin et al. [1] used Concept Hierarchy as a pre-processing method on Census data of two cities, namely Chengyang and Laixi. The main aim of their paper was to apply concept ladder on the board construction cost attribute. They make apply the Dynamic Hierarchy Adjustment Algorithm on the selected attribute for data generalization. It includes top-down big nodes promotion and bottom-up small nodes merging. This is useful to make the data more meaningful, easier to interpret and improve the quality of mining and the patterns that are obtained. A concept chain of command basically defines a sequence of mappings from lower-level concepts to their higher level. It organizes the concepts in the form of a tree, a lattice, a directed acyclic graph, etc. The experiment was performed using C++ and handled 49 items (name, age, sex, birth date, housing construction cost, etc.) with 15000 records whose data was stored in the SQL server. According to the results, Chengyang, which is an economic development zone, has the

following distribution of housing construction costs given as, Households (in percent):Housing Construction Costs (in Yuan). 49.43%:0- 10,000 Yuan, 19.64%:10,000-30,000 Yuan, 14.84%:30,000- 100,000Yuan, 1.25%: 100,000-300,000 Yuan. Similarly, Laixi, a purely agricultural are had the following distribution, given as Households (in percent):Housing Construction Costs (in Yuan). 68.8%:0-10,000 Yuan, 8.12%:10,000-30,000Yuan, 8.71%:30,000-100,000 Yuan, 4.48%:100,000-300,000Yuan. Thus, based on the estimates, it was inferred that residents areas.

Manan Chawda et al. [2] implemented the ARIMA model for the selected dataset. ARIMA or AutoRegressive Integrated Moving Average is used for time series data sets. Linear Regression and Decision Tree Algorithm were used for prediction. Monte Carlo simulation was used on the regression equations to find out the risk and uncertainty in the forecasting model. The focus of the paper was to examine the fields of education and healthcare, and to predict the future statistics, with the aim to aid the municipal corporation. The model had three simple steps: Obtaining the data, performing data pre-processing, applying the algorithms and verification of the predicted values against actual values. The paper com-pared literacy rate against twelve parameters, these included net enrollment ratio, dropout rate, pupil teacher ratio, student classroom ratio, girls enrollment ratio, female teachers etc. The paper compared life expectancy against five parameters. These are birth rate, death rate, sex ratio, infant mortality rate and total fertility ratio. In both sectors, education and health care, linear regression provided better accuracy (61.52% for education and 90.97% for health care) compared to decision tree regression algorithm (35.30% for education and 82.83% forhealthcare),proof that range of all parameters gives the birth death ratio with decreasing with an error of minimum ratio.

Bin Sheng et al. [3] used Classification and Regression Trees (CART) to analyze the census data and use it for classification of inhabitants. Classification tree is used because it gives discrete value, while regression tree is used for continuous and numerical value as the end result. Tree pruning was done in the implementation to prevent any overfitting. CART is a dual decision tree modeling algorithm, with minimal cost complexity and GINI index is used as the evaluation function to split the tree, the techniques are non-parametric in nature. The model is in four stages which includes census data which is extracted and application of algorithm and verification of predicted values against real time values. In this paper the Decision- Tree-Based sorting model CART is used to analyze the census data in Chengyang and Laixi, c1assify the inhabitants, then evaluate the results, and finally discuss the important implication for using Data Mining in ballot data. It is

observed that 70.453% of the residents in Chengyangand Laixi belong to the class 1 general, 0 (poor) - 19.473%, class 3(best)have very small section residents 8.7% and 1.373%. These results are used in area planning and construction, for example, amusement.

Neil Hernandez-Gress et al. [4], analyzed the Socio Demo-graphic factors that affect patients having Type II Diabetes Mellitus (T2DM). It was estimated that in 2010, 285 Million people were diagnosed with Diabetes Mellitus (DM) and 90% of those people had Type II DM. This paper analyses Genetic and epigenetic factors along with societal factors. The socio-demographic profile of patients was analyzed and classified by region. In Theran, Iran, patients were classified on the basis of a diet point of view as well as social, biological and cultural factors. The risk profile was male, non-smokers, low physical activity, consumption of legumes, eggs, fish, etc. In Eastern England, it was from a physical activity point of view. The risk profile was male, young, lower socio-economic class, low Body Mass Index (BMI), etc. In Lublin. Portland classification was on the basis of residence point of view. Risk profile included lower income, high number of persons with disabilities, high BMI, etc. in Spain it was based on socio-economic inequalities. The risk profile included low socio-economic status (especially women), obesity, etc. In Saudi Arabia, local women were less prone to diabetic 2 as a result of exercise while the men relied on medication. In general, factors such a low income, socio-economic conditions, education and exercise play a crucial role in the diagnosis of diabetic 2. The methods used were Cross-sectional study, eloquent and analytic value, degeneration analysis and Principle Component Analysis.

Esmeralda Florez Ramos et al. [5], describes how open data plays a key role for governments to strategize and deal with challenges in the future. It tries to answer questions on the impact of open data on innovation, which data is used to assess the progress of a country and which countries are effectively utilizing open data. The analysis performed on the connection between the pointed of open data readiness and the open data impact, along with comparing the rank positions of the countries and considering other indicators such as level of ICT progress of a country, precision and political rights and civil liberties of individuals. The data used is global level data primarily from the historical dataset of Open Data Barometer (ODB), World Wide Web Foundation (2013 to 2016), ICT Development Index (IDI) 2016, Freedom in the world status (2016 to 2017), and several secondary datasets such as Corruption Perceptions Index (CPI) and Gross National Income (GNI) from different sources. Several analysis were conducted on trends such as Regions and GNI-based, Open data readiness, accomplishment and measures of impact, ODB rank and countries level of freedom, transparency

and ICT development, etc. Results suggested that success on opendata of countries is based on good levels of ICT development, freedom and interest of becoming more transparent. However, there are indications that countries with low ICT development do not profit from open data.

DhwaniSondhi et al. [6] used the tool Waikato Environ-ment of Knowledge Analysis (WEKA) to find information related to gender inequality in different spheres, from the key survey of the Raigarh district from Maharashtra.The author used and applied the seven steps of the knowledge discovery process such as data cleaning, data amalgamation, data assortment, data adaptation, data withdrawal, outline study and knowledge perception. The dataset was downloaded from open government data platform. The data set has eight variables and attributes which include total male population, female inhabitants, literate male population, cultured female population, scheduled caste population, scheduled tribe pop-ulation etc. The data is fed into the tool and different kind of visualizations are carried on the data set - these include cluster visualization, scatter plot visualization etc. Clustering technique is implemented via the EM clustering method whose results clearly show that its accuracy is quite highand that such kind of exercises can be done for analyzing ballot datasets. These visualizations give a larger insight in gender dissimilarity of Raigarhdistrict.

IvarsGutmanis et al. [7], presents results on the im-plications of economic growth on the environment in the United States from 1970 to 2000. This is achieved through traditional input output analysis, which allows various projection, each associated with specific assumptions of policy to be pursued. Two population assumptions were chosen, namely U.S ballot Series B and ballot Series E, incorporated with two economic growth assumptions fixed with the labour productivity of different levels. The results showed four vital scenarios high inhabitants and high profitable growth, low population and low growth and twointermediary cases. The core of the model used in this study is the 185-sector University of Maryland Interindustry Forecasting Model which uses standard key-in output equations AX + Y = X, where, X: column vector of total. outputs, Y: column vector of final demands and A: 185 order matrix of key-in output coefficients. The results elaborated on Waste generation, Abatement, Waste generation assuming alternate production technologies, generous to pollution by sectors of economy, Trends of residual abatement costs and impact of technology on cost trends of pollutant control. The paper drew conclusions such as increase in population level bears less influence the environmental quality than the increase in the standard of living. If waste treatment is not intensify, even

low-population low-growth rate state yields significant increases in current pollution levels by the year 2000.

Sharath R et al. [8] performed data analytics on the US Census Bureau dataset to predict income and economic hierarchy in United States of America. Dataset is inserted into Hadoop for processing and PIG is used for Map Reduce, along with this tools such as Python, R and JAVA1.7is used. Naive Bayes classifier is applied on the dataset which produces an output showing Gender distribution in Occupation using a bar graph representation. Naive Bayes and C4.5 are used later and data is visualized using different form of visualization such as Education salary relationship using box plot. Optimization has been performed on the dataset to increase efficiency of time complexity, Level - 1 Normalization and Level - 2 Normalization has been carried on the dataset. Statistical analysis and a variety of algorithms are used to derive insights from the population represented by the selected census. The issues of economic inequality in the society, gender inequality in the domain of income and their root causes determine the relationship between income and education. The mean and median income distribution across the states is depicted. A cataloging into economic classes that has been used to forecast categories of economic class people belong and hence, their standard of living and social status.

Chin Jui Chang et al. [9] performed data mining on Taiwans population census to gain an insight in the disad-vantaged social classes of Taiwan. The methods used for the process include, association rule mining, clustering algo-rithms and decision tree. A correlation was plotted to find attributes with higher correlation and relationships among data items of data columns. For association rule mining- Generalized Induction rule was used due to better and independent rules, along with Apriori algorithm. For Classi-fication C5.0 and CART were used, C5.0 due to its nature of output column display and the partitioning method used was information gain ratio, while CART used Dispersion as its partitioning method. Results concluded that female led single parent families consisted of single-women having higher educational attainment diploma, bachelors degree, masters degree or doctorate and divorced women running single- parent families had lower educational attainment at the elementary, senior high or senior high vocational school level. Male led single parent families had highest educational attainment of men running single-parent families as junior high elementary school when they are divorced or separated. Highest educational attainment of men running single-parent families was junior high school and senior high school (23.29%) when they are widowers. Widowed elderly people were illiterate or reached elementary school as highest educational attainment and were singles. Elderly people with a spouse had elementary

school as their highest enlightening ability. The highest enlightening attainment of single aborigines of Ami origin was elementary school or senior high vocational school and that of wedded aborigines of Ami origin was basic school.

OgochukwuC.Okeke et al. [10] focused on the geo-spatial distribution of population in Nigeria. The paper looks at dissimilar methods such as Meta Learning, Bagging and Stacking. Bagging is an area of prognostic mining that combines predicted classifications from multiple models. Meta erudition combines predictions from multiple models. Stacking is used to combine predictions from numerous models, where the types of model are very different. The authors make use of decision tree learning which is a method that approximates discrete - valued target function, in which the learned func- tion is represented by a decision tree. Decision tree algorithm was used to predict basic attributes of population from the census, along with Structured System Analysis and Design Methodology. Hidden information is extracted from large census data and geographical information systems (GIS).The result includes prognostic attributes of the population to give geo-spatial distribution in Nigeria. For example, on the basis of marital status, sex, employment and unemployment, etc.The model is able to predict the wealth of the nation.

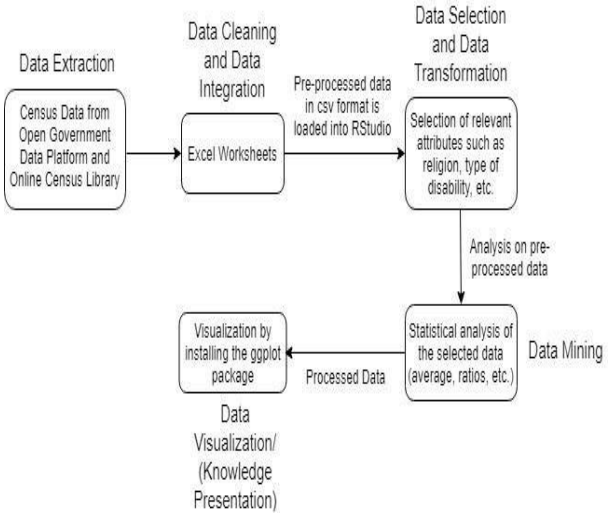

Fig. 1 Architecture - Knowledge Discovery Process

3 Architecture

3.1 Data Mining

Data Mining is defined as the process of analyzing data in large datasets and extracting previously unknown, potentially useful and novel patterns. These patterns or trends serve as knowledge, which then allows us to make decisions and even predictions.

3.2 Knowledge Discovery in Databases(KDD)

Simply applying the techniques of data mining does not provide us with useful patterns or knowledge. It is part of a larger process known as the Knowledge Discovery Process. The following steps are involved in the KDD process:-

•Data Cleaning - Tidying up data by removing noisy, unclean and missingdata.

•Data Integration - Combining data from severalsource.

•Data Selection - Based on the requirements of the tasks or knowledge to be extracted, relevant data is retrieved from thedatabases.

•Data Transformation - Consolidating data using sum-marizing or aggregatingfunctions.

•Data Mining - Analyzing and extracting interesting patterns from thedata.

•Pattern Evaluation - Patterns mined in the previous step are now evaluated andinterpreted.

•Knowledge Presentation - Results of are visualized through graphs orcharts.

The architecture used in this paper, follows the steps involved in Knowledge Discovery, however, it has been customized for census data, as shown in Figure 1.

4 Implementation

4.1 Algorithms Used

Descriptive analysis of involves basic statistical testing along with the visualization of results. In this paper, data from several datasets has been subjected to dissimilar types of descriptive statistics such as compute of

Frequency and determine of Central Tendency. By using the Measures of Frequency, numerical attributes have been used to gather the count and percentages of various results for each State and Union Territory. Similarly, using the Measures of Central Tendency, the mean or in this case, the national average of India and the different states,have been found.

4.2 Tools Used

The dataset was found online, on the open government platform. It was downloaded and imported to MSExcel. The data is then cleaned and aggregated. Once cleaned, it is uploaded into the RStudio environment and required packages for EDA are installed. The dplyr package is installed to perform statistical analysis or testing. Several functions such as filter, select, summarize, are provided under it. Hence, using these functions through R programming, certain results were found. The ggplot2 package allows to visualize the results, thereby allowing us to explain them in a simpler manner.

5 Disability

According to the 2011 census, India has 26 Million dis-abled people in the country, i.e., 2.1% of the inhabitants has some kind of a disability. In 2016, after years of meticulous support by disability rights activists, the Indian Parliament expanded the number of disabilities covered under the law from 7 to 21. Rare conditions such as intellectual palsy and autism, among others were incorporated. The existence and acceptance of a medical condition as a disability is not enough. In order to include the disabled population in the social fold of the country, it is important to create special policies for them and accept them as part of the society. In order to do that, we need to primarily recognize where the disabled population stands right now, so that, a way can be found to uplift them.

The following datasets are used for analysis:-

•Allotment of disabled workers by Sex, Economic Status and dwelling.

•Population presence learning institution by Age, Sex and Type of learning Institution.

•Disabled population by nature of disability, learning level and Sex.

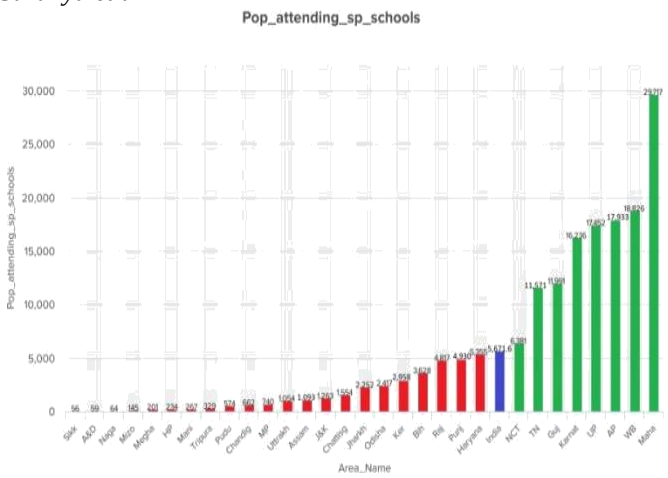

Fig. 2. Disabled population attending schools for the differently-abled.

5.1 Results

•Disabled population attending schools for the differently abled (Fig. 2.) - On an average each State/Union Territory has 5671.067 disabled people attending special schools.Maharashtra has the highest number of disabled people attending special schools at 11571, which is double the national average. However, it also has a disabled illiterate population of 958849, which is 82 times more than the population attending school and studying. Sikkim has the lowest number of disabled people attending special schools at 56 which is less than a hundredth of the national average. It also has a disabled illiterate population of 9911, which is 180 times more the population attending school and studying. Sikkim being a small state has a very high ratio of disabled people not attending special schools.

•Disabled working population (Fig. 5.) - The National average for the disabled working population is 2,78,411 disabled persons, in India. Accordingly, Uttar Pradesh has the Maximum number of disabled working popu-lation, with 14,46,393 disabled persons. Lakshadweep consists of the least number of disabled working pop-ulation at only 321 disabled persons. The values above are absolute values, hence they vary with size and population of each State/Union Territory. In the next section, we see the trend in terms of the ratios, which levels the playing field in between the smaller and larger states.

•Working vs Non-working disabled population (Fig. 4.) - The national average of working disabled population to non-working disabled population is 57.08%, meaning for every 57 disabled worker there are 100 disabled non-workers in India. Nagaland has the best ratio with 107.98%, meaning for every 100 disabled non-worker there are nearly 108 disabled workers in Nagaland. This indicates that, a disabled person in Nagaland is more likely to be employed in Nagaland than in Maharashtra. Lakshdweep has the worst ratio with 24.80%, i.e., for every 24 disabled worker there are 100 disabled non-worker. Sikkim however, being the worst in disabled population attending schools has the 2nd highest work ing to non-working ratio at 96.21%, i.e., for every 96 disabled worker there are 100 disabled non-worker.

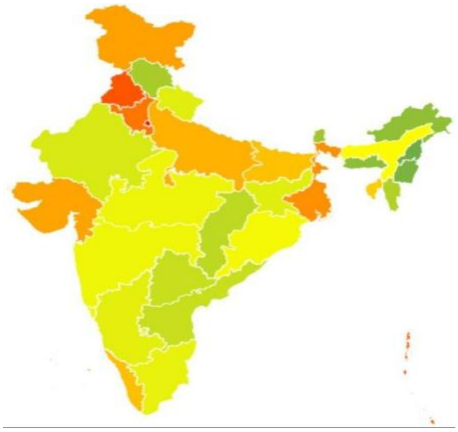

Fig. 3. Female to male ratio of disabled workers.

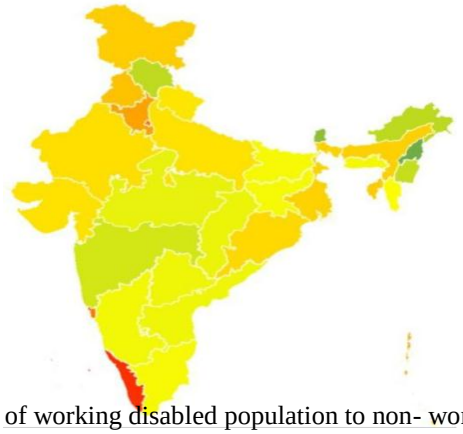

Fig.4. Ratio of working disabled population to non- working disabled population.

•Female to male disabled workers (Fig. 3.) - Thenational average of female to male disabled workers is 37.72%, implying that for every 37 female disabled worker there are 100 male disabled workers. The highest ratio in Nagaland at 70.51%, i.e., for every 70 female disabled worker there are 100 male disabled workers. The worst is National Capital Region at 14.98%, indicating that for every 14 female disabled worker there are 100 male disabledworkers.

6 Education

India is said to have one the largest youth populations, with nearly 50% of the population under the age of 25. This suggests that India has a great potential for human resource and hence education acts as one of the most important pillars, if the country seeks to be a future super power. Hence, it is important to look at the current state of the education and accordingly form policies for the future, wherever the country is lagging behind in education

.

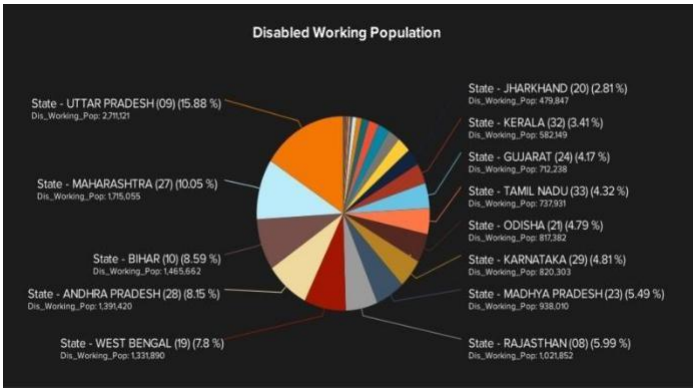

Fig. 5. Disabled working population

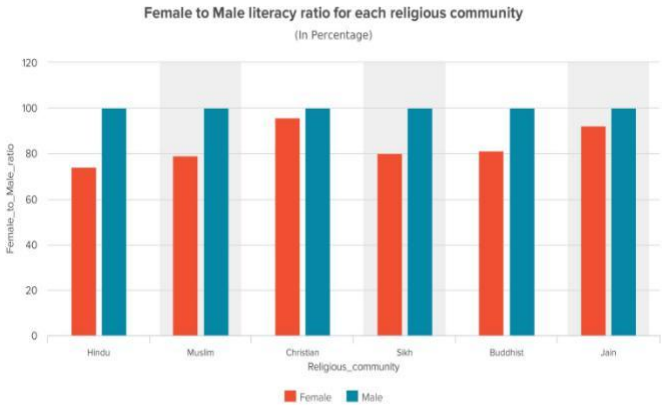

Fig. 6. Female to Male literacy ratio for each religious community

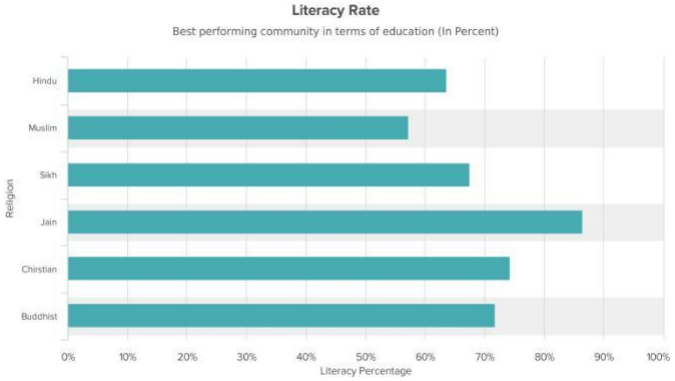

Fig. 7. Best performing community in terms of education.

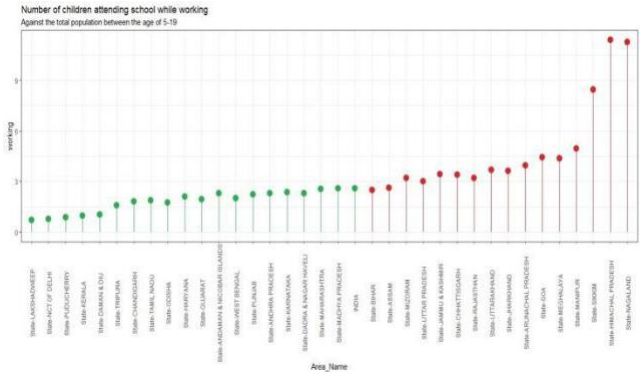

Fig.8. Number of children attending school while working against the total population between the age of 5-19.

The following datasets are used for analysis:-
• Population Age 5-19 Attending Educational Institution ByEconomicActivityStatusAndSex
• Education Level By Religious Community And Sex For Population Age 7 And Above 2011 (India and States/UTs)
• Migrants By Place Of Last Residence,Age,Sex,Reason For Migration And Duration Of Residence-2011(India andStates/UTs)

6.1 Results

•Female to Male literacy ratio for each religious commu-nity (Fig. 6.) - According to the data, the best female to male literacy ratio is from the Christian community at 95.9%, i.e., for every 100 literate male Christian members, there are nearly 96 literate women. This is followed by the Jain community with 92.2%. The least performing community is the Hindu Community with 74.2%, i.e., for every 100 literate male Hindu members, there are nearly 74 literate women.

•Best performing community in terms of education (Fig. 7.)-The Jain community has the highest ratio,at 86.4%, i.e.,from every 100 Jain members 86 are literate and can read and write. The least performing community is the Muslim community with 57.2%, implying that, for every 100 Muslim members,only 57 are literate, andcan read andwrite.

•Number of children attending school while working (Fig. 8.) - The national average is 3.19% meaning for every 100 students between the age group of 5-19, attending educational institutions, there are 3 students who work throughout the year. The best performing region is National Capital Region at 0.70% indicating that there is approximately 1 individual from 100 school attending children who attends school while working. The worst performing state is Himachal Pradesh at 11.4% implying that, for every 100 school attending children there are 11 children who work from a period of few days to a whole year.

•Working to Non-working ratio of children attending school (Fig. 9.) - The national average is 3.8%meaning, for every 100 students in school that are not working, only 3 are working, which is quite a low number. The lowest is from Lakshadweep with 0.826%, i.e., for 100 non-working children only 1 child is working while attending school. The worst ratio is from Nagaland at 17.7%, implying that, for 100 non-working school going children from Nagaland,18 are working.

•Migration of people from each state for education

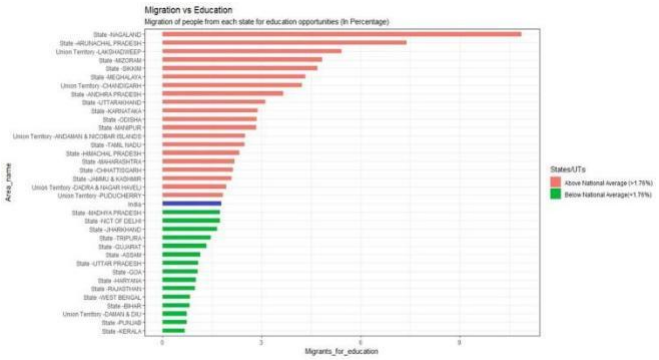

Fig. 9. Working to Non-working ratio of children attending school.

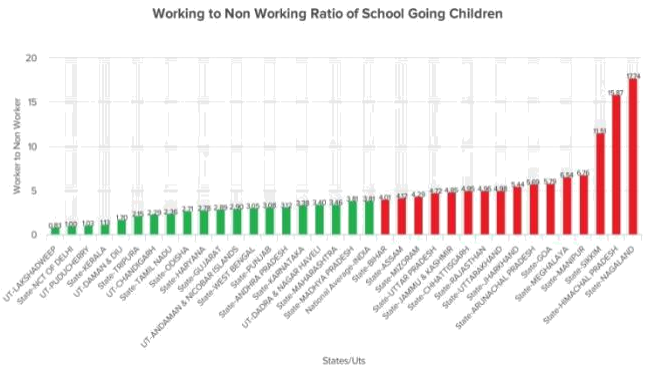

Fig. 10. Migration of people from each state for education opportunities.

opportunities (Fig. 10.) - The national average is 1.76%, indicating that approximately 2 people migrating from each State/UT, do it for educational pirposes. The high-est ratio of migrants from a state was Nagaland, with 10.9% of migrants moving out for the reason of better educational institutions. The lowest was Kerala with 0.66% of migrants moving for educational purposes. Kerala having the highest literacy rate in the country has very few migrations to pursue education in other states.

7 Housing

According to a report by Ministry of Housing and Urban Poverty Alleviation, the housing shortage in urban India is nearly 18.78 million households. This indicates that there are 18 million people without a house to live in. The housing statistics list which houses have electricity,cleanwater, toilet facilities, etc. Housing for all or affordable housing is an indicator of the economic well being of the country. The following datasets are used for analysis:-

•Number Of Households Availing Banking Services And Number Of Households Having Each Of The Specified Assets

•Households Classified By Source and Location of Drinking Water and Availability of Electricity and La-trine(Total/SC/ST)

•Households By Availability Of Type Of LatrineFacility

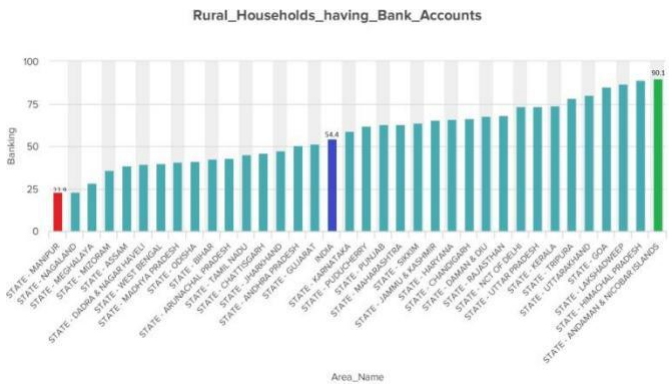

Fig. 11. Banking vs Households - number of rural households having bank accounts thus making them part of the nancial system of the country.

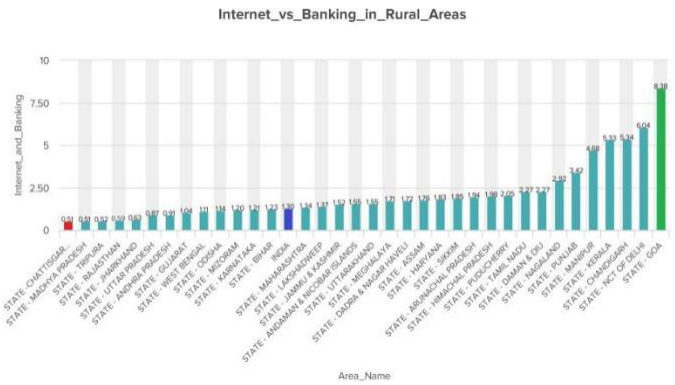

Fig. 12. Banking vs Internet - number of households having bank accounts against internet connectivity.

7.1 Results

•Banking vs Households (Fig. 11.) - This analysis de-scribes how many rural households have bank accounts and are thus them making part of the financial system of the country. The national average of the rural economy is 54.4%, indicating 45 rural households have bank account for at least one person, from a sample of 100 rural households. The highest ratio is from Andaman and Nicobar with 90% rural households having a bank account, while the lowest is Manipur with only 22.9% rural households having a bank account.

•Banking vs Internet (Fig. 12.) - The analysis is done on households having banking against internet connec-tivity. The national average is 1.30%, i.e., forevery 100 citizen having access to internet via computer 1 person has a bank account. The best state is Goa, with 8.38% of people having access to banking along with internet connectivity, while the worst performing state is Chhattisgarh having a minute ratio 0.51%, meaning only 1 person will have access to banking along with internet connectivity from 200 rural households.

•Unclean source of water vs no electricity and latrine facilities (Fig. 13.) - The ratio of rural households having no access to clean water with no latrine and no electricity was found in this analysis.The national average is 17.4% meaning 17.4 rural households with tap water from untreated source don't even have latrine and electricity facilities. The worst performing state is Uttar Pradesh with 66.8% of households with tap water from untreated source and lack of latrine and electricity facilities, while Lakshadweep is the best performing state with 0 houses having tap water from untreated source along with no latrine and no electricity.

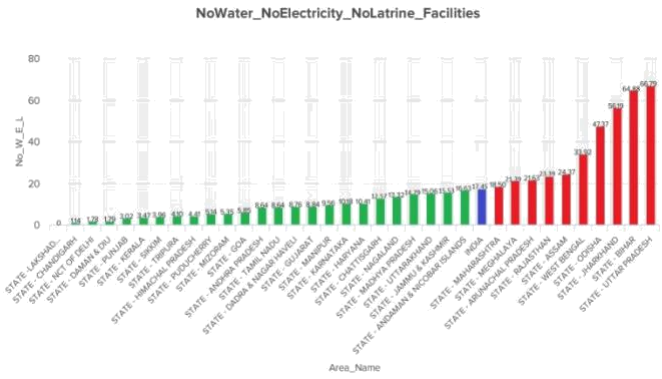

Fig.13. Unclean source of water vs no electricity and latrine facilities

- The ratio of rural households having no access to clean water with no latrine and noelectricity.

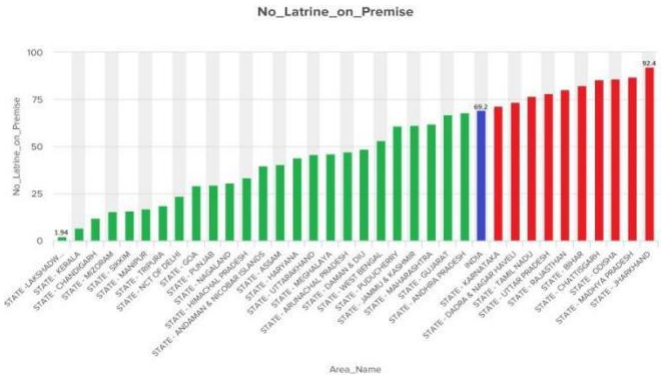

Fig.14. Latrine facilities - the percentage of the rural population not having a latrine on the premises.

• Latrine facilities (Fig. 14.) - The analysis was done to find the percentage of the rural population not having a latrine on the premises. The best performing state is Lakshadweep with 1.94% of rural households not having latrine on the premise, implying that only 2 houses out of 100 rural households in Lakshadweep don't have a latrine on the premise. In comparison to this, Jharkhand has 92.4% of rural households that have no latrine on premise, implying that they either defecate in the open or use public latrines. The national rural average is 69.2% i.e., nearly 70 households out of 100 don'thavealatrineonthepremise.

8 Conclusion

From the above results we gain insights into 3 different socio-economic factors that impact the country. These in-sights permit the creation of better targeted policies, which in turn helps the nation to move forward as a whole, without leaving behind any section of a society. With the next census in 2021, this data can be used by the government to better understand the effects of old policies and hence formulate new ones.

References

[1] Sheng Bin , Gengxin Sun, "The Preprocessing in Census Data with Concept Hierarchy", 2nd International Conference on Computer Engineering and Technology, 2010.

[2] Manan Chawda, Rutuja Rane, Srikanth Giri, "Demographic Progress Analysis of Census Data Using Data Mining", Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018.

[3] Bin Sheng , Sun Gengxin, "Data Mining in Census Data with CART', 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), 2010.

[4] Neil Hernandez-Gress , Diana Canales, "Socio-Demographic Factors and Data Science Methodologies in Type 2 Diabetes Mellitus Analysis", International Conference on Computational Science and Computational Intelligence (CSCI), 2016.

[5] Esmeralda Florez Ramos, "Open Data Development of Countries:Global Status and Trends", ITU Kaleidoscope: Challenges for a Data Driven Society (ITU K), 2017.

[6] Dhwani Sondhi, "Application of Data Mining in Census Data Analysis using Weka', International Journal of Engineering Trends and Technology (IJETT), vol. 52, no. 3, 2017.

[7] Ivars Gutmanis, "Environmental Implications of Economic Growth in the United States, 1970 to 2000: An Input-Output Analysis", Proceedings of the IEEE Conference on Decision and Control and 11th Symposium on Adaptive Processes ,1973.

[8] R .Sharath, S.Krishna Chaitanya, K N .Nirupam, B.J.Sowmya,K.G Srinivasa,"Data Analytics to predict the Income and Economic Hierarchy on Census Data", International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), 2016.

[9] Chin-Jui Chang , Shiahn-Wern Shyue, "A study on the application of data mining to disadvantaged social classes in Taiwans population census", Expert Systems with Applications,vol. 36,no.1,pp.510-518, 2009.

[10]Ogochukwu C.Okeke , Boniface C.Ekechukwu, "Using Data-Mining Technique for Census Analysis to Give GeoSpatial Distribution of Nigeria", IOSR Journal of Computer Engineering,vol.14,no.2,pp1-5,2013.

Biographies

S.S.Saranya is an Assistant Professor in Department of Computer Science and Engineering S.R.M Institute of Science and Technology, Kattankulathur campus, Chennai, India. Currently she is pursing Ph.D.(CSE) in B.S Abhur Rehaman Institute of Science and Technology, Vandalur ,Chennai . She has over eight years of experience in Teaching. Her research interest is Security and Privacy on Big Data Analytics, Network Security, Internet of Things.

Sharmin Khathariya B.Tech,Dept of CSE Student of SRM IST. Area of Interest : Data mining,Big data analytics

Atharva hajare B.Tech,Dept of CSE Student of SRM IST. Area of Interest : Data mining,Big data analytics