International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

Socio-Economical Status of India using Machine Learning Algorithms

V. Balasankar, P. Suresh Varma

Abstract*: Data is everywhere and lots of data is openly available to people. We can analyze this data to find the hidden and unnoticed information to use it purposefully. One important source of information is census data and it provides data related to the people living in a country. Analyzing such data is useful for knowing the socio economic status of the country. Data mining and machine learning techniques can be used to analyze such large volumes of data. In this work Indian census 2011 is analyzed and identified the socio economic status of different states of India. To identify the social status of each state we studied literacy rate, categories of workers in different fields, gender wise working population. To identify economical status like people living below poverty and above poverty we used clustering techniques of machine learning. At first we pre-processed the data and later correlation based feature selection was applied, and on that result k-means and k-mediods clustering methods were implemented independently. Finally the clusters are evaluated to see the performance using confusion matrix. The final results show that k-mediod has better performance than K-means.

***Keywords:* Machine Leaning, k-means, K-monodies, Socio-Economical, poverty**

I. INTRODUCTION

Governments often want to know the socio economic conditions of each region, so it collects data from various sources and analyzes this data. Inequalities in income are one of the main problem that government want to solve. By reducing income inequalities we can achieve balanced social development; It also increases national income and provides political stability in the country. For this purpose India-districts-census data was used to study the socio economic developments in the specific regions. India-districts-census contains large volumes of data with many attributes. But many of these attributes are not filled and may not contain a value. Such sparse data is a common in many applications. The dimensionality of objects may be very high like in census data. Such sparse data need to be handled properly in order to improve the query processing and minimizing the storage, it is important to find active attributes and their correlations with other objects there by predicting regional economy .So many data mining and machine learning techniques have been implemented to clean , Analyze data and to predict the socio economic status of people in a region and cluster have been created based on the poverty and the state wide poverty were presented.

Algorithms like cluster analysis, correlation, decision trees, K-means, are studied on census and other related economic data. In this paper we have implemented K-means and kmediods clustering algorithm on census data to find the economic status of people in India. Socio economic indicators like poverty, gender distribution in work, literacy rate for male and female, percentage of people working in different sectors, are measured.

Section II presents literature review and section III presents the detailed methodology for classification section IV presents results final section presents the conclusion.

II. LITERATURE SURVEY

Author (s) can send paper in the given email address of the journal. There are two email address. It is compulsory to send paper in both email address. Ronit nirel et.al[2009] described about sample surveys and censuses. The author described the new trends in census preparation methodologies. Population census is a very important official statistical data, on which many analyses are built. it should be defined properly and accurate data about individuals place of residence and other demographic characters need to be collected. It is a complex task and a costly procedure. There are many types of errors and one of the important errors was coverage error which is further divided in to under coverage and over coverage. This study tells the eligibility status of an individual. Due to advancements in technology there is big influence on census preparation with GPS, Mobile phones and advancements in internet. Regarding data collection there are two methods are there one is record linkage procedure which uses information from administrative sources and the other one is sample surveys . Estimation of coverage errors by using sample survey, use of sample data to evaluate statistical adjustment of census counts were described. A different type of census that is based on continuous sample survey called rolling survey implemented by France was also described. Finally sample surveys in conjunction with census and usage of long form and short forms as a data collection was described.

C.Hakim [1984] reviewed the new trends in census data dissemination in particular data dissemination in Great Britain.

Manuscript published on January 30, 2020.
* Correspondence Author
V. Balasankar*, Associate Professor, Department of CSE, Aditya college of engineering, Surampalem, Andhra Pradesh, India.E-MailL balasankar.v@gmail.com
Dr. P. Suresh Varma, Professor, Department of CSE, Adikavi Nannaya University,Andhra Pradesh, India

Author discussed about many key issues like importance of census analysis by different people and the users of these analysis and the need for summaries and commentaries of census tables, guides and indexes for fast searching of the content to help the different users to understand the census statistical tables, consultations with users for census data design, data dissemination centres like census regional office, central government departments, independent research institutes, universities,

commercial agencies and Also discussed about census by products and the need for census based area classification, publishing sample data for public use and the related confidentiality issues, alternatives to the census like national surveys were discussed. Even though there are developments in census data dissemination. And alternative sources of information, the basic published volumes are sufficient for general analysis.

Bin sheng et.al [2010] presented the importance of data mining technique to analyze census data. Census data contains rich set of information and many meaningful insights are hidden in it. CART (Classification and regression trees) is a decision tree based classification model, using this model census data was analysed and predictions were made on socio economic conditions of the people. Decisions trees of classification gives high accurate prediction results, and the advantage is the results are easy to understand. CART has sound theoretical foundation , it has three phases like tree building, tree pruning, and tree estimation. Using c++ language CART model was implemented on fifth census data of Chengy and Laixi. GINI index was used as a evolution function in this work. Even though CART is a recursive algorithm, non-recursive version was used to improve the performance. The main aim of this work was classifying the residents in Chengy and Laixi into four categories like poor, general, better and best. The results showed that middle level per capita income people were more in this area. Based on the results local authority can plan for economic development.

Jian ming et.al [2018] presented the implementation of artificial neural networks on economical and technical data of mining enterprise. Generally mining enterprise data is multi-dimensional and nonlinear. One of the important indicators of mining enterprise is mineral products sales price data. Due to some technical limitations in the environment the geological data was lost and the author reconstructed this data using artificial neural networks and geo statistics. Here back propagation algorithm of artificial neural network model was used to predict the mineral product price. Neural network was created with a single hidden layer and three –layered neural network with 5input neurons and 1 output neuron was built. For reconstruction of geographic data artificial neural networks and geo statistics were used. The predictions showed that the model is strong and prediction accuracy is high, for geological missing data the prediction results and interpolation were reliable.

Maria Beatriz Bernabe Lacranca et.al.[2004] described a statistical classification approach on subset of data from the XII house hold and population census data of metropolitan zone of the maxico valley(MZMV) to present the properties of the population data. The main idea was classification of zones in MZMV region. in this work K-Means procedure was implemented to create the clusters. To analyse this population data cluster analysis was applied. To represent the area of study 57 variables were used which came from correlation analysis using k means the number of clusters maintained going from 8 to 70 range. K means needs the n value for the number of clusters to be formed in advance and it also needs k value for k initial centroid in advance. The statistical R software was used to classify 4925 census zones. Zones belong to clusters having greater size, lowest average in all variables, zones with lowest population ,zones mean values very close to the global means, high income areas of city were generated and the zones were analysed This work was mainly provides the analysis about the relationship between the population and transportation trips in MZMV area.

Jose cazal et al[2015] Explored predictive models for economic system. It was observed that selection of what data to be included and what to be discarded is a big challenge .There are some variables that are complex and many factors affect these variables. Data mining techniques scientifically proven and gives reliable results in analyzing, predicting socio economic studies. In this work data mining techniques were proposed as a valid option which is better than traditional econometrics methodology. SEMMA model was used in data mining which stands for Simple, Explore, Modify, Model and Assess. Here based on type of data set (time series or cross sectional) predictive algorithms were chosen. Experiments were done on two cultures namely data modelling culture (here data is generated by stochastic given model) and algorithmic model culture (which assumes the data as complex and unknown). Author did analysis and prediction using both traditional econometrics with E-views tool and data mining techniques with EMMA tool. The results show that data mining techniques were more effective and efficient than the techniques of econometrics.

Joon heo et al[2014] proposed user driven economic data analysis by using a mobile app. The user sends the data and the analysis parameters through mobile phone to the server. The server uses big data and mathematical algorithms to perform analysis on given parameters. Normally Big data analysis is done by only few companies because they can afford for it and user-oriented analysis were generally neglected. This framework contains two entities one is server and the other one is user application running in the mobile phone. Stock data of 8 countries over a period of 33 years were used for this work. This data was first cleaned and next it was processed. Android SDK was used to build this mobile app. At the server side three algorithms were implemented (minimum spanning tree, principal component analysis and clustering). User selects economic parameters from the mobile and transmits it to the server and the server sends the results back to the mobile device.

Sharath R et al[2016] studied and analysed socio-economical conditions from US household data. In this work the size of dataset is huge i.e., 3.5 million household information. The major conclusion was income of individuals decides various aspects of life like education, health, standard of living, household decisions, and economic status and so on. In this work five modules were implemented to study the importance of income domain.

The five modules are gender distribution in occupation (male vs female), education-salary relation (higher degrees and their income vs professional degrees and their income), economic hierarchies to find the economic classes using classifiers, Benford's law of US income (outlines the frequency distribution in many datasets), mean and median of income distribution across all the states. These economic hierarchy predictions are useful in many areas like planning houses for poor and middle class people, better pension plans for retired people, planning for various welfare programs to poor and so on. To conduct this work five tools are used,

they are Hadoop, Java1.7, Python, R, and Pig. The data was first normalised by eliminating less important attributes, null values. After normalisation economic hierarchies were created using K-means and predictions of those economic classes are done using classifiers. Three classifiers were applied to improve the efficiency .All of them performed nearly same on the dataset. The accuracy by those classifiers were as follows:

Table 1: Classifiers and Accuracy

Classifiers	Accuracy
Naïve Bayes	48.13%
C4.5	51.3%
Boosted C5.0	53.7%

Finally using relevant attributes demographic graphs were plotted. Octavio juraz Espinosa et al[1999] described visual techniques for input/output data. These visualisations are necessary because the data was represented in matrix form. So data navigation becomes a problem because of screen limitation. Here the techniques were designed based on user tasks like querying about interaction of two sectors, labelling data points, matrix area magnification, comparing two industries based on the goods they produced, comparing two sectors based on the co efficient values, finding patterns for matrix, modifying values and re-computing the matrix. In this work better visualisation techniques were created to represent the economic input output data. The main need of analysis is to study the interdependencies between industries in regional economy. The data was maintained in four different matrices like make matrix(commodities produced by industries), total requirements matrix(direct and indirect interactions between industrial sectors),use matrix(inter industrial activities and commodity inputs for industrial production),direct matrix(based on use matrix and total output). All these matrices are displayed with the help of a window using a pixel for each cell , colours were assigned based on its category. Each window is partitioned into three. First one is a large window which contains matrix, second the bottom part contains the detailed information and third left window represents the data to be visualised. Apart from these matrices economic IO data was also represented as geographical information. This way the visualisation makes it easy to perform analysis on economic input output data.

Yin Cai et.al.[2010] studied online analytical mining for analyzing regional economic data. Regional economic data is gathered from statistical data. As per this work old statistical data was maintained in the form of word or excel format such a data is structured in non hierarchical manner. This creates a problem in analysis and research. The regional economic data contains large number of industries in various regions. In this work the data was analysed with the help of online analytical mining (OLAM) it was designed with the help of Data warehouse, Analytical processing and data mining. The are three main components namely data layer, middle layer, client layer. data layer was built with MS SQLServer 2005 as data warehouse. Middle layer built with MS Analysis service and created data cubes, metadata, OLAP engine, data mining engine. At the client layer web application was developed using OLAP service and different operations like data cube browsing, rotation, slicing, and drilling can be done. Here zhenjiong province economic data was used and multidimensional cubes are created and clustering algorithm was implemented. Final results showed that manufacturing, building trade, wholesale and retail markets are more developed than other trades.

Avery sandborn et.al.[2016] described the relationship between special features derived from high resolution satellite imagery and Census data of Accra Ghana. Special features are the metrics that analyze pixel groups for describing geometry, orientation, patterns of objects in an image. Such special features can be used to find housing conditions and living standards in a city. To see the association between demographical variables and special features five special features (panTex,Line support regions, histograms of oriented gradients, Fourier transform ,local binary patters) were selected and extracted from image and then related to census variables. This method was proposed as a alternative methodology for census data. The special feature and spectral information normalized difference vegetation index.(NDVI).were computed and correlated to census data and there exists a high correlation between LBP and census derived variables. final results shows that the special features can be used to find the socio economic conditions of the population.

Ferdin joe j et.al[2011]. Differentiated Horizontal, Vertical representation of data set with a new representation called Hover(Horizontal over Vertical). In many applications sparse data is common issue. This sparse data is not managed well in both Horizontal and Vertical models in the context of performance, storage, and query processing .The author worked with a new method called Hover on Census data and ecommerce data in sparse form. Hover representation contains two steps generating correlation table and generating subspaces the sub spaces are generated from the correlation table in a Heuristic manner. In this work the author used rapid miner tool and measured subspaces, space usage, execution time, running time in a systematic way and the and the changes in parameters with the schema changes are analysed and differences are observed.

Zhuang et.al.[2014] dealt with the analysis of regional economic indicators. Traditional economic methods are not effective in finding the factors influencing economy. In this work the analysis on key factors of regional economy were done with the help of k-means and CADD algorithms. High –dimensional data is sparse in nature, when processing such data based on distance and density the clustering was not efficient. In order to improve efficiency of clustering weighted CADD was proposed by the author. This partition the cluster based on adaptive density reachable ideas.

Here the natural, economic attributes of the regional economic data was clustered using k-means but the results were not ideal, so CADD algorithm was used which reduced the total dimensions. Finally comparative analysis was done on Chinese regional economics.

III. METHODOLOGY

In this work we studied the socio economic status of various states and union territories of India.

We have taken India-district-census 2011 data set for this work. This data set contains 640 rows and 118 attributes of different districts from 28 states and 7 union territories. Except three columns all the columns contain numerical values.

The data in this data set contains values in aggregated form and the usage of this data is limited. Form this data we want to analyze socio economical condition of India. The proposed work is divided into 2 main modules

1) Analysis of social conditions
2) Economical conditions like measuring people living in bellow and above poverty.

Analysis of social conditions includes the following sub modules.

1) Finding different categories of workers state wise.
2) Finding gender wise working group for all states.
3) Finding gender wise literacy rate for all states.

3.1 social economical condition

Different categories of workers

From this data set various working categories are found. Like percentage of people working in Cultivation, Agriculture, Household and Others from the total working population state wise. And all these categories are plotted on graph state wise

Gender wise working group

Here from the data set we have extracted total males and females state wise, and total male workers and female workers state wise, and found the percentage of males and females in different working area. This tells us the employment conditions for males and females. It was found that the percentage of male and female workers in LAKSHADWEEP it is 46.24,11.4 respectively, and this is the minimum value in this working group when compared with all other states. The average working group of India is 54.37 and 31.08 for male and females respectively.

Gender wise literacy rate

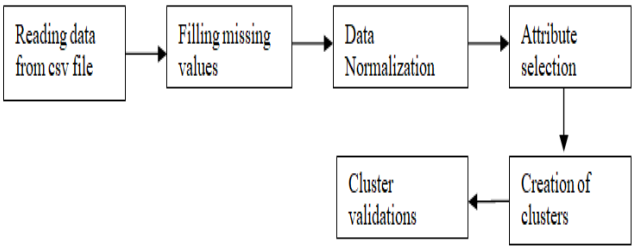
Here we analyzed state wise male and female literacy rates.

Here we found male and female literacy rate by calculating percentage of male and female literates from the total population of male and total population females state wise respectively. This gives us male and female literacy rate for each state. It was found that the percentage of male and female workers in BIHAR it is 58.23,41.94 respectively, and this is the minimum value in this working group when compared with all other states. The average literacy rate of India is 73.31,61.68 for male and females respectively.

3.2 Economical conditions Of the people

The objective this module is to find the poverty in India,and finding population above poverty line and bellow poverty line. Here on the dataset we applied K- means, kmediods clustering algorithms. The dataset is a open dataset and it contains partial data about 28 states and 7 union territories. With the available data and with some assumptions clusters were created for states above and below poverty.

The method for creating clusters is represented with the help of flow chart as given below.



In this work python has been used to study socio economical conditions. All the required modules are imported for filling missing values, normalizing data, for statistical calculations and plotting data.

In this module fist data set was verified for missing values. And the missing values are filled with mean value of that attribute.

The data was scaled using min max scaling which transforms features in the given range here we have scaled in the range of 0 to 9.

Creating clusters using K-means and K-mediods

This experiment was divided into two steps:
1) Finding most correlated or significant variables
2) Creating clusters using k means and kmediods

The first step is common for both k means and kmediods that is finding most correlated or significant variables. Here we have used Pearson-correlation-coefficient method on this data set. This has given the most correlated 12 features out of 118 attributes. There is 0ne important feature called 'Households_with_TV_Computer_Laptop_Telephone_mobile_phone_and_Scooter_Car' which talks about the living standards of people.

For these 12 features we have calculated mean values for all 640 districts of 28 states and 7union territories. From these mean values we find class label. This class label identifies poverty status. As the dataset does not provide class label for each district the following method was used.

1) first we find the difference between maximum and minimum mean values of all districts.
2) 20% of this difference is added to the minimum mean value and placed in variable called pvalue.
3) finally we compared mean value each district with the pvalue . If the district's mean value is greater than pvalue then the class label is 1 other wise it is 0.

After this K means and kmediods clustering techniques are applied independently to find the state population bellow and above poverty line.

And these clusters are compared with the class label that was calculated in the previous step using confusion matrix and the results are presented. The entire procedure is detailed here under. Pearson-correlation coefficient method :

This method is a measure of strength which tells us how two variables are linearly associated with each other. In our data set we have total 118 columns and 640 rows. We have one attributecalled 'Households with TV Computer Laptop Telephone mobile phone and Scooter Car' which is an important attribute that we have used to find most important features that can influence the poverty condition.

Pearson-correlation-coefficient is defined as

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (y_i - \overline{y})^2}}$$

With this we found attributes. Based on these 12 highly correlated features we find the clusters using K-means.

k means clustering Implementation:

The first two rows of the dataset are chosen as the initial centroids and found the euclidean distance from the first row to all the other rows and stored in a variable Temp1 and found euclidean distance from second row to all other rows and stored in a variable temp2. And tested whether temp1 is minimum or temp2 is minimum .if temp1 is less than temp2 then our prediction is true and its value is represented as 1 and we are comparing this with actual result which is in class label. if it is false its value is 0 then we are incrementing false negative otherwise we increment false positive than current c1 value is put into x and y is the current row value. Now we are taking mean of x and y that become the new centroid . After this c1 is updated with nc1 which is a new centroid applied to test the remaining rows. Else if temp1 greater than temp2 we increment true positive by 1. If the actual value of classlabel is 1, else we increment true negative. We move c2 and current row in x and y variables respectively and taking mean value of x y we will get the new centroid. We repeat all these steps until previous centroid and the newly found centroid are equal. The objective function is defined as

K means= $\sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2$

K means [14] algorithm is defined as follows :

1. Take data points and cluster them in to k groups where k is already defined.
2. From the dataset choose *k* points at random as cluster centers.
3. Assign each data point to their closest cluster center according to the *Euclidean distance* method.
4. Calculate the mean value of all data points in each cluster now this become the new centroid for the cluster.
5. Repeat steps 2, 3 and 4 until the same data points are assigned to each cluster in consecutive rounds.

Finally we are calculating the actual mean values of each state and showing the same on pie chart.

K MEDIODS: K-mediod is also called as partitioning around mediods method(PAM). It is a partitional algorithm for clustering and it is a extenction to K-means algorithm. It is more robust noise and outliers than K-means.

In k-means method mean values are choosen as centriods for the cluster,whereas in K-medios the representative data points are choosen as the mediods. K-Mediod is interapretable as one of the data points is selected as mediod.

A mediod is nothing but any object or data point of the cluster, whose average dissimilarity to all other bojects or data points in the cluster is minimal.

Here the basic idea is first compute the k representative objects and call them as mediods. After set of mediods are found each representative object of the data set is assigned to one of the nearest mediod. Which means object i is put into cluster c_i when the mediod mc_i *is nearest than any other method* m_j

The algorithm proceeds in two steps

1) Build: This is a initialization step, select k random data points out of n points as the initial medicos.

2)Swap: if the objective function can be minimized by swapping mediod with any other data point, than swapping is performed. This is repeated until the objective function can no longer be minimized.

K mediods algorithm is implemented as follows

Input:

K:Represents the number of clusters

D: A data set consists of n objects

Output:

A set of K clusters

Method:

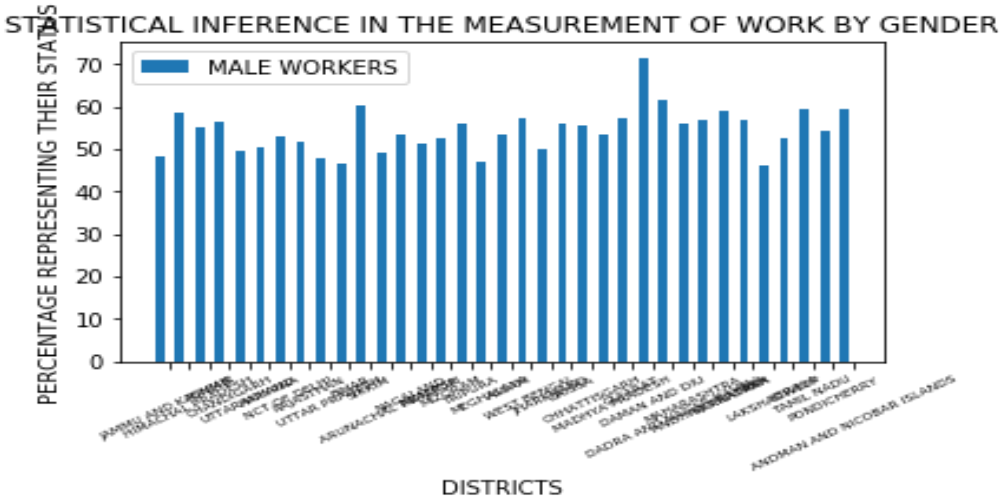
1. The algorithm starts with random selection of the K representative objects as medoid points out of n data points.
2. Associate each data point to the nearest mediod by using any common distance metric methods.
3. Randomly select any non-medoid object $O_{random.}$
4. Compute total cost S of swapping current medoid with object O_{random} .
5. If S>0, swap current medoid with the O_{random}.
6. Repeat these steps until there is no change in the medoid.

IV. EXPERIMENTAL RESULTS

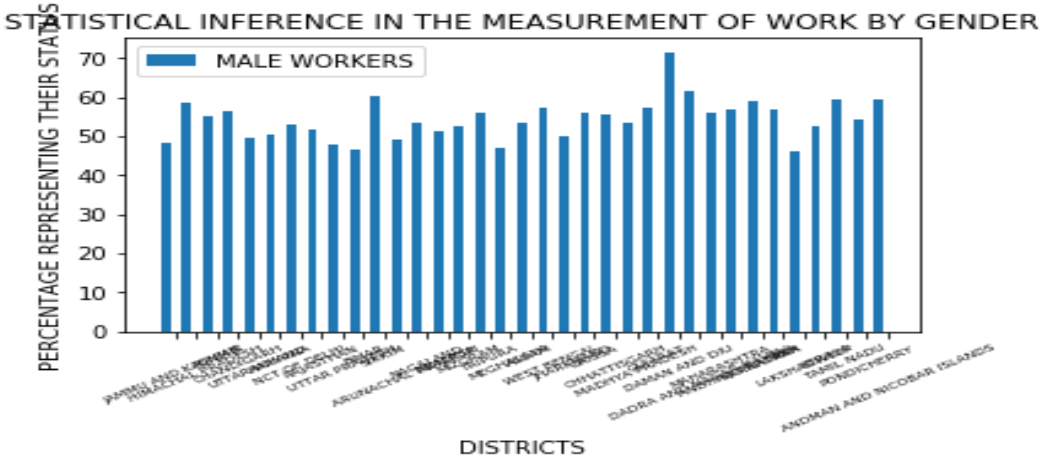
The social status of the states in India are represented as different categories of workers, gender wise working group and gender wise literacy rate. The results are as follows. The percentage of male workers

Are presented in 1(a), The percentage of female workers

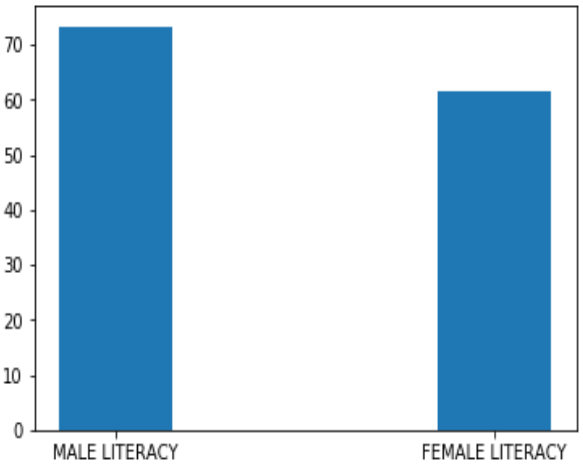
Are presented in 1(b),1(c) represents country wise male versus female literacy rate.



1(a) Percentage of male workers state wise



1(b) Percentage of female workers state wise



1(c) Male versus female literacy rate country wide

2) Finding different categories of workers state wise.
3) Finding gender wise working group for all states.
4) Finding gender wise literacy rate for all states.

For experimental purpose Indian districts 2011 census data is taken. The data set contains different states and districts information. It contains total 640 rows and 118 columns. The data set was first analyzed for missing values, after that the missing values are filled with the mean values of the respective feature. The following diagrams represent this issue. Fig. 2 (a) Represents missing values in the original dataset. Fig 2(b)Represents data after filling missing values.

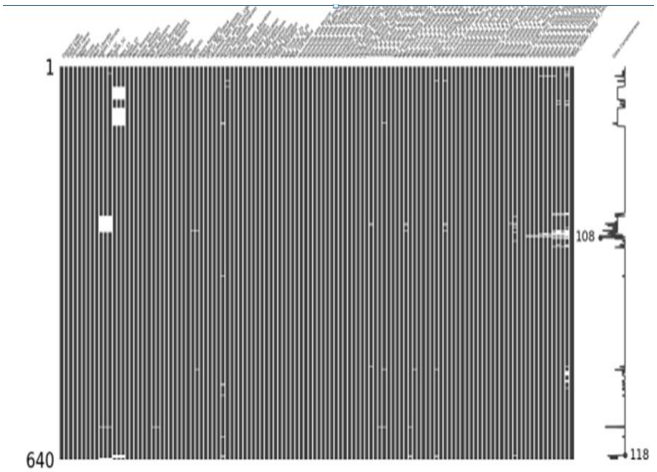


Fig.2.(a) represents missing values for some columns

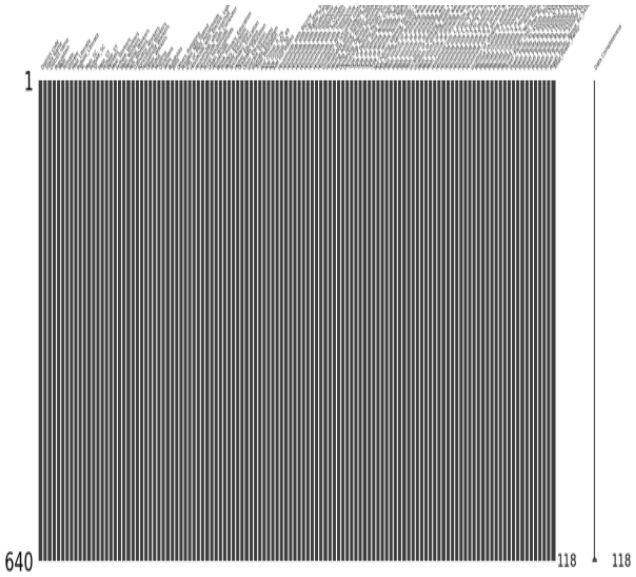


Fig.2.(b) represents data after filling missing values

After this Pearson correlation is found with respect to an important feature called 'Households with TV Computer Laptop Telephone mobile phone and Scooter_Car' which describes the economic status of Households . After this the following 12 fields are more correlated to this feature. Only these features are used for cluster creation.

1)Other_Workers, 2)LPG_or_PNG_Households,
3)Households_with_Internet, 4)Households_with_Computer
5)Urban_Households, 6)Graduate_Education,
7)Households_with_Car_Jeep_Van
8)Households_with_Radio_Transistor9)Households_with_Scooter_Motorcycle_Moped
10)Households_with_TV_Computer_Laptop_Telephone_mobile_phone_and_Scooter_Car,
11)Households_with_Telephone_Mobile_Phone_Both
12)Ownership_Rented_Households'
The above 12 fields are more correlated and they are important for calculating poverty.

K-Means clustering is applied on the data set and the data is made into two clusters namely bellow poverty and above poverty. The following Figures represents state wise poverty percentage.fig.3(a) shows the actual poverty in all over India and fig.3(b) shows the experimental results for poverty in India with k-means.

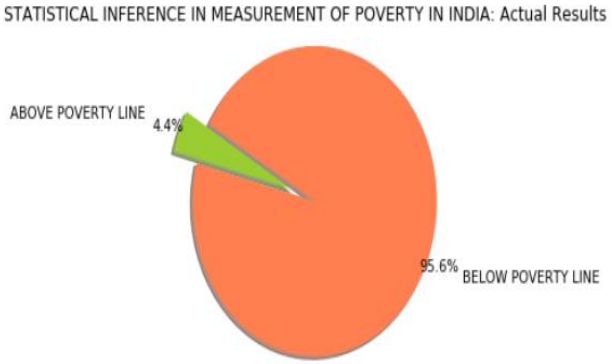


fig.3(a) The actual poverty in India

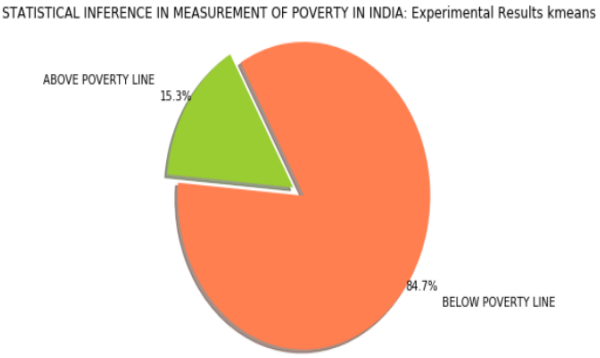


Fig. 3(b) The experimental results for poverty in India with K-means

fig.4(a) shows the actual poverty state wise in bar diagram fig.4(b) shows the experimental values of different states in bar diagram.

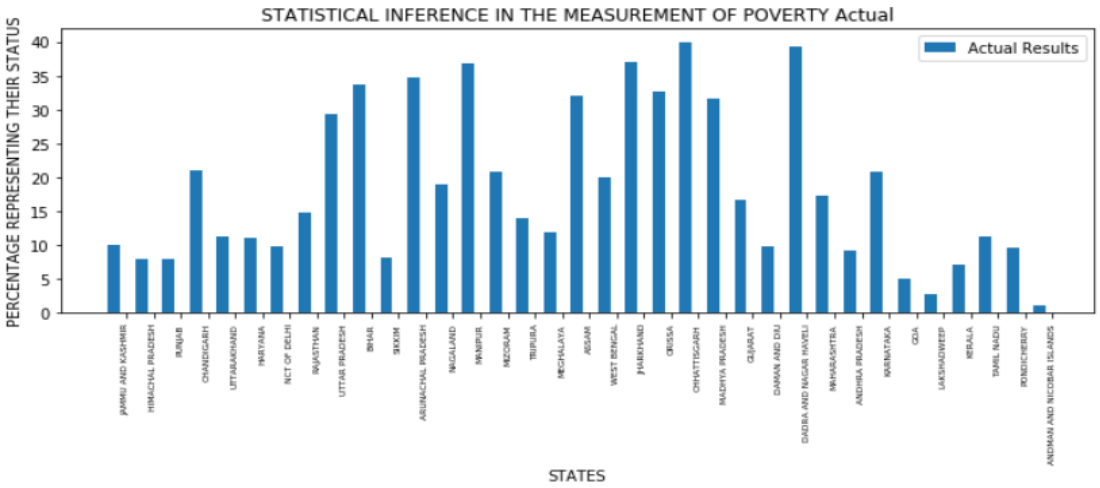


Fig. 4(a) Actual Poverty State Wise

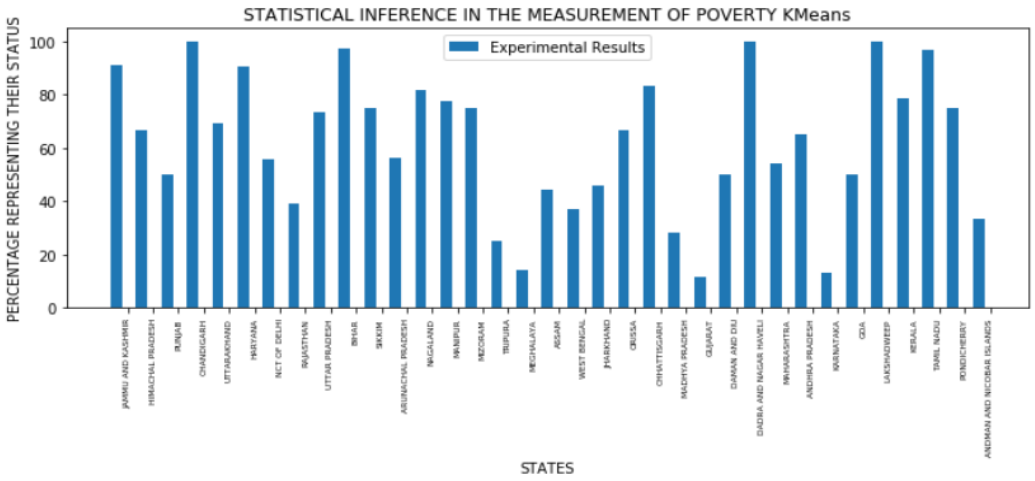


Fig. 4(b) experimental values of different states in bar diagram

When Kmediods algorithm applied on the 12 highly correlated features the data is made into two clusters namely bellow poverty and above poverty. The following Figures represents state wise poverty percentage.fig.5(a) shows the actual poverty in all over India and fig.5(b) shows the experimental results for poverty in India with K-modiods.

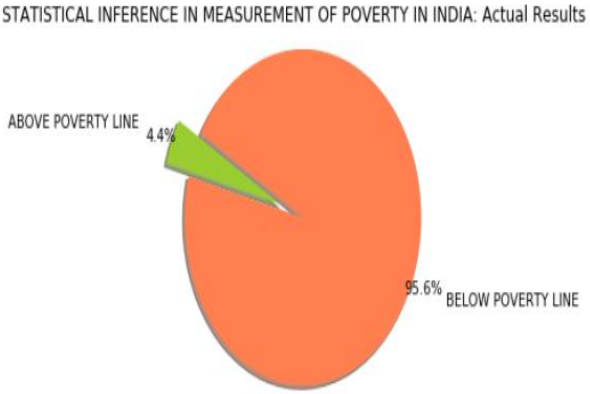


Fig.5.(a) The actual poverty in India

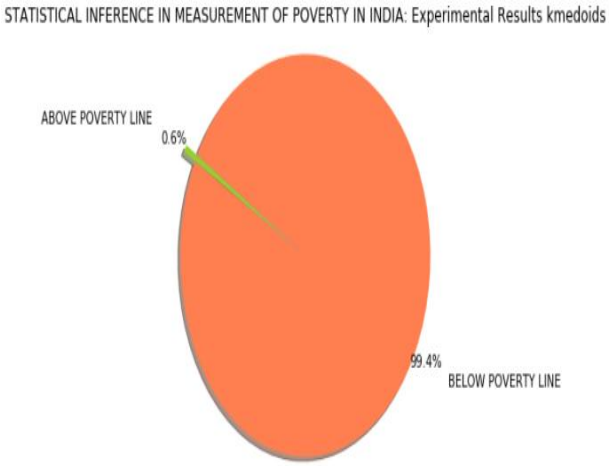


Fig.5.(b) The experimental results for poverty in India with Kmeans

fig.6(a) shows the actual poverty state wise in bar diagram fig.6(b) shows the experimental values of different states in bar diagram.

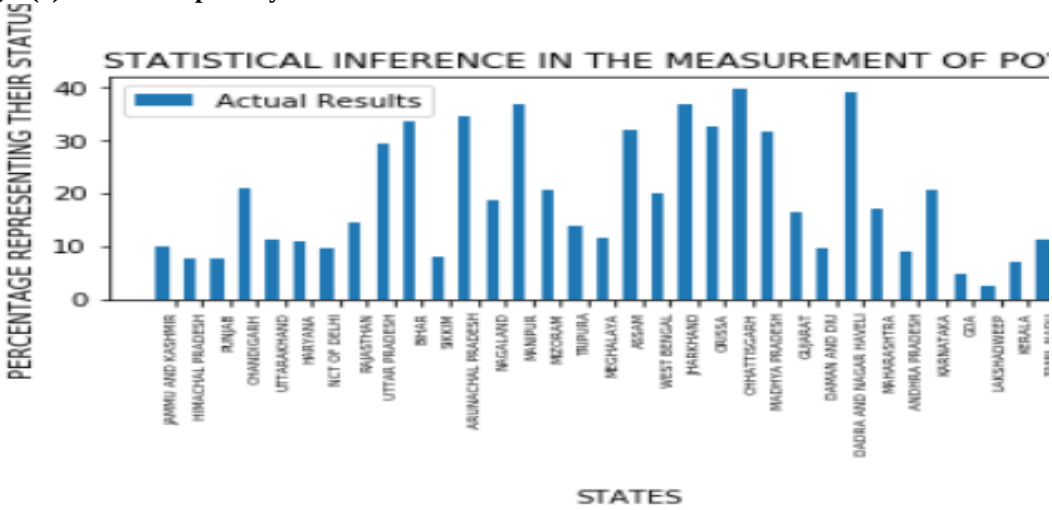


Fig.6.(a) Actual poverty state wise

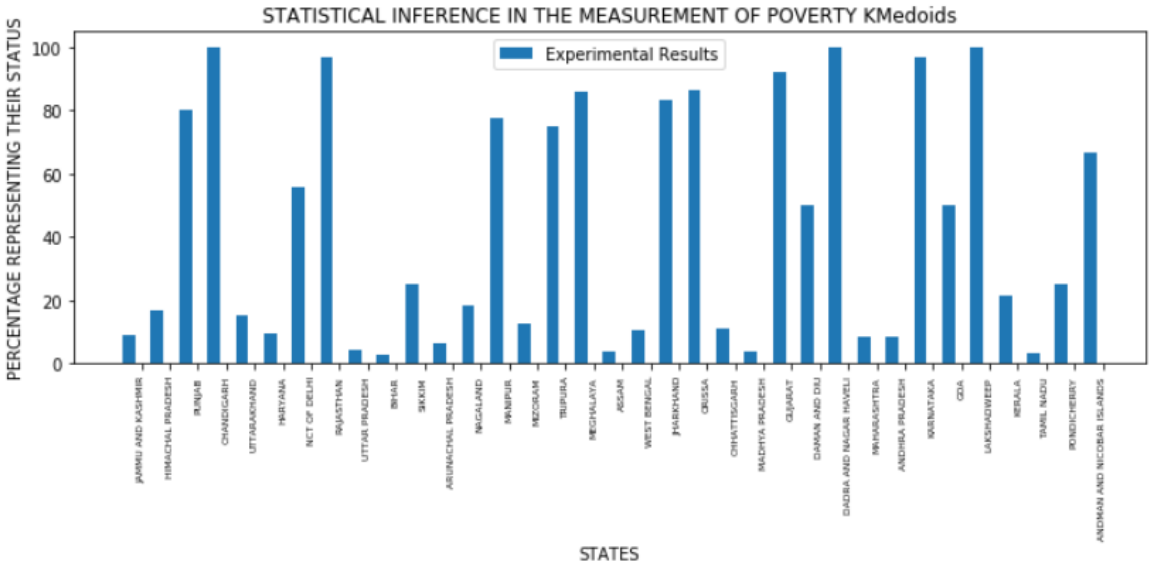


Fig.6. (b) Experimental values of different states After K-mediods

The following table shows the compassion of k-mediods and k means with respect to performance measures like Error rate, Precision, Recall, F-measure, Sensitivity, Specificity, and Accuracy.

Table 2: K-Means and K-mediods Analysis

Measures of performances	K-means (in %)	K-mediod (in %)
Error rate	20	4
Precision	0	100
Recall	84	100
F-measure	0	100
Sensitivity	0	15
Specificity	84	100
Accuracy	81	97

As comparative analysis, the Kmediod performance is better than Kmeans for this dataset. The table 2 shows the detailed analysis about k-means and k-mediods that the Kmediod is giving better accuracy, specificity, recall, precision-measure and less error rate when compared to Kmeans. But for the sensitivity parameter, K-means is giving 0 whereas Kmediod is giving 15.

V. CONCLUSION

We have discussed the importance of studying and analyzing census data with the help of data mining and machine learning. The new trends in census preparation and census data dissemination procedures and data mining and machine learning methodologies to analyze and predict socio economic status of different areas are studied. This provided a brief idea about data mining and machine learning techniques applications on census and house hold data. The first step in this process is data cleaning, and finding highly correlated attributes that represents the financial status of the people. After that k-means and k-mediods techniques were applied for clustering data to find people living below poverty and above poverty. In this experiment it was found that k-mediods is giving more accuracy, precision, recall, f-measure and less error rate when compared to k means algorithm.

ACKNOWLEDGMENT

We would like to thanks the Director and management of Aditya college of engineering, College, Surampalem for give the assistance and support for this work. We are very much thankful to R&D department of Adikavi Nannaya University, Andhra Pradesh, India for accepting and give their cooperation for this work. We are also thanking to Dr. T.PanduRanga Vital for valuable suggestions towards this research work.

REFERENCES

1. Ronit Nirel and Hagit Glickman “Sample Surveys and Censuses”,2009 Elsevier, Vol. 29A,issue PA,p.g 539-565.
2. C .Hakim, “Data dissemination for the population census”,Social Science Information Studies (1984), vol 4,issue 4,p.g 273-282.
3. Bin Sheng,Sun Gengxin,“Data Mining in census data with CART”, ICACTE 2010 - 2010 3rd International Conference on Advanced Computer Theory and Engineering, Proceedings (2010),vol 3,p.g 260-264.
4. Jian Ming, Lingling Zhang, Jinhai Sun,”Analysis models of technical and economic data of mining enterprises based on big data analysis”,2018 3rd IEEE International Conference on Cloud Computing and Big Data Analysis, ICCCBDA 2018,p.g 224-227.
5. Maria Beatriz Bemabe Loranc,Ramiro Lsez,”Application of Nonsupervised Classification to Population Data”,2004 1st International Conference on Electrical and Electronics Engineering, ICEEE (2004), p.g 182-187.
6. John Justus C ,”Predictive models of economic systems based on data mining”,2015 International Workshop on Data Mining with Industrial Applications,DMIA 2015,Part of the ETyC 2015 (2016),p.g 61-65.
7. Joon Heo,Okyu Kwon,”User-driven Economic Data Analysis Framework”, 2014 IEEE 10th World Congress on Services, p.g 263-264.
8. Sharath R , Krishna Chaitanya S, Nirupam K N, Sowmya B J and Dr K G Srinivasa,”Data Analytics to predict the Income and Economic Hierarchy on Census Data”,2016 International Conference on Computational Systems and Information Systems for Sustainable Solutions,CSITSS 2016 p.g 249-254.
9. Octavia Juarez Espinosa, Chris Hendrickson Ph.D., James H. Garrett Jr. Ph.D,”Visualization of economic input-output data”,Proceedings of the International Conference on Information Visualisation (1999),isssue 1999-January,vol 1,p.g 496-501.

10. Ying Cai, Jiangping Chen, Xiaoqing Fan, Zhenguo Yu,"Study on the Regional Economic Data Analysis and Mining Platform Base on OLAM",2010 Second International Workshop on Education Technology and Computer Science,ETCS 2010 (2010),vol 3,p.g 817-820.
11. Avery Sandborn and Ryan N. Engstrom,"Determining the Relationship Between Census Data and Spatial Features Derived From High-Resolution Imagery in Accra, Ghana",IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2016),vol 9,issue 5,p.g1970-1977.
12. Ferdin Joe J,Dr. T. Ravi,John Justus C"Classification of Correlated Subspaces Using HoVer Representation of Census Data",2011 International Conference on Emerging Trends in Electrical and Computer Technology, ICETECT 2011 (2011),p.g 906-911.
13. Zhuang Cheng,"Regional Economic Indicators Analysis Based on Data Mining",Proceedings - 2014 5th International Conference on Intelligent Systems Design and Engineering Applications, ISDEA 2014 (2014),issue 2, p.g 726-730.
14. Vital, T. P., Lakshmi, B. G., Rekha, H. S., & DhanaLakshmi, M. (2019). Student Performance Analysis with Using Statistical and Cluster Studies. In Soft Computing in Data Analytics (pp. 743-757). Springer, Singapore.

AUTHORS PROFILE

V. Balasankar completed his M.Tech in Computer Neural Networks from Gokul Engineering College, A.P., he has 14 years of teaching experience. He is currently working as Associate Professor in Department of Computer Science and Engineering, Aditya College Of Engineering And Technology (ACOE) Surampalem A.P, India. he is a member of CSI. He has published 4 papers in reputed international journals and 1 in conference. His main research interests are Computer Networks, Machine Learning, Deep Learning and GIS

Dr. Suresh Varma Penumatsa is a Professor in the Department of Computer Science and Rector of Adikavi Nannaya University, Rajahmundry, India. He is having 21 years of teaching experience. His areas of interest include communication networks, image processing, speech processing and Machine Learning. He has published several research papers. He is a life member of ISTE, SMORSI, ISCA and IISA.