Empirical Statistical Analysis and Cluster Studies on Socio-Economic Status (SES) Dataset

V. Balasankar 1, Dr. Suresh Suresh Varma Penumatsa 2, Dr. T. Pandu Ranga Vital3,*

1Research Scholar, Department of CSE, Adikavi Nannaya University, Rajahmundry, Andhra Pradesh, India. Email- balasankar.v@gmail.com

2Professor, Department of CSE, Adikavi Nannaya University, Rajahmundry, Andhra Pradesh, India. Email- vermaps@yahoo.com

3Associate Professor, Department of CSE, Aditya Institute of Technology and Management, Tekkali, Srikakulam, Andhra Pradesh, India. Phone: +91 9491936489 Email- vital2927@gmail.com

Abstract. Socio-economic status (SES) levels and conditions are extremely influential variables in the study of a particular area of society or any society. Social factors, for instance, the position of caste, religion, marital status, education levels, give good assessment results for us about a person’s goals and the method of achieving their objectives. Generally economic status of any family is needy upon the social factors, for instance, the size of the family, educators in family and levels, and the level of the friendly environment in the family. SES with machine learning (ML) especially cluster analysis is important to identify important features or dimensions of the SES dataset, evaluate the rankings of dimensions and dimensional reductions. In this research, we collected 1742 samples (household information) as per socio-economic ratios and area (rural and urban) wise ratios with good questionnaires between 2018 and 2019 from Rajamahandravaram, East Godavari District, AP, India. We conduct the statistical analysis and cluster analysis for identifying the important factors of SES levels and their problem analysis. In cluster analysis, we apply k-means, hierarchal clustering (HC), and hierarchal with principal component analysis (PCA). The good projection results related to HC and PCA-HC specifies passemens of SES class values.

1. Introduction

The word 'socio' originates from the word 'social' and denotes to the people, and the ways (level) they fit into the community or residential area in which they live. It reflects how well they are instructed, educated, and working a job and so on. Economic denotes to the money related situation of the individuals inside society and incorporates, earning capabilities, own house and the resources, and so on. As such, it is a sociological and economical total assessment measures of an individual's work insight and a person's or family's economic and social situation according to other people, based on occupation, education, income, and other factors. The assessment of the family's SES is measured using the income of the household, education in house, earning members in the house, other sources of income, and assets of that house. Mainly, SES levels are three that are low, middle, and high SES. Some researchers are
described as the SES indicator low, middle, upper-middle, and high as per indicator variables education, income, occupation, and wealth [1] [11]. SES levels are differentiated in rural and urban areas. Some of the research works are confined to the linkages between urban and rural provincial establishment as it is most parts accepted that these two zones are autonomous to one another with the distinct limit in financial, political, and social aspects. Regardless, whether urban and rural provincial zones, one can't fulfill all the requirements of the individuals without depending upon others. It is thusly relying upon one another to address different issues and angles which are unavoidable. The mutuality relationship between urban and rural settlement is important. In such a manner, the collaboration is perceived through elements of components in these settlements by factors for the flow of people (migration), merchandise and enterprises flow, thoughts and data flow, culture, and assets flowing from one place to another. So, communication like the progression of individuals, merchandise and enterprises, ideas, and assets; to help livelihoods, that is required instead of two separate and isolated socio-economic entities [2] [13].

The improvement of positive mental states is seen as a continuous interaction between the child's intrinsic inequalities, the family qualities, and the more and wider environment. The family's financial status seems to assume a characteristic part in deciding the positive mental and psychological conditions of the youth, with this hypothesis [3] [4]. Poverty and neediness as the socio-economic development whereby the assets accessible to society are utilized to fulfill the needs of the meager few while the many don't have even their fundamental needs met [5].

SES with machine learning (ML) research is vital to predicting and preventions related to SES factors like poverty, education, administration, and so on, in that particular area. We can analyze SES assessments with unsupervised ML models like k-Means, hieratical, and hybrid cluster analysis with PCA (principal component analysis) for effective and vital results. In this research, we analyze the SES dataset of Rajamahandravaram using unsupervised cluster analysis like k-means and hieratical clustering with PCA analysis.

The main Objectives are based on the above problem that the present study was carried out with the following objectives.

1. To find out the impact of SES on the people's lifestyles of Rajamahandravaram using statistical analysis.
2. To examine the SES levels using cluster analysis with visualizations.
3. To explore the impact of SES on the resilience of education factors with cluster visualizations.
4. To find out the correlation among SES feature attributes.
5. To find out the ranking of attributes and analyze the projections using the HC-PCA algorithm.

The organization of the paper as follows.

In section 2, we describe in detail various research works related to SES and ML. In section 3, we describe the proposed model, and k-means, HC, and PCA algorithms and their mechanisms. In section 4, we analyze SES attributes statistical values, conduct the cluster analysis, and projects the high ranging attributes using HC-PCA algorithms.

2. Literature Survey

In this section, we reviewed numerous reputed journal papers and various works on SES from reputed articles, and some of the paper works of different authors are presented. Conger et al. (2009) [6] reviewed SES related to individual development and family processes. In this review, they discussed the integrationist model for family life and SES, incorporates social selection and social causation positions. Mainly, they focused on concepts that are 1. New Millennium Economic Climate, 2. Social class and SES measurements, 3. SES underlying assumption studies and family life 4. SES predictor variables quality of relations, sustainability, and stability 5. Empirical Findings like Couple Relationships and SES, 6. Romantic Relationships and FSM (Family Stress Model), 7. Child Development and Parenting relationships, and SES, 8. FSM and Children’s Development, 9. The IM (Investment Model) and Development children, 10. Predicting from the Individual to SES, 11. The Interactionist Perspective. They concluded with some suggestions that are 1. The SES QAS
(Quantitative Assessment), 2. Research Methodology Improvements and 3. Current Theoretical Frameworks (CTF) Extensions and Elaborations. Schäfer et al. (2012) [7] researched age-related health issues related to SES that there are 3 psychosocial socioeconomic disadvantages, maltreatment, and social isolation. For this, they experimented with 1037 members from New Zealand. In this research, they found that study members suffered from 3 age-related health issues at age of 32 that are high inflammation levels, metabolic risks, and major depression. Lampos et al. (2016) [8] studied SES classification using Facebook user's language and behavior. At first, we detail a 3-way characterization task, where clients are delegated having an upper, center, or then again lower financial status. A nonlinear, generative learning approach utilizing a composite Gaussian Process bit gives essentially better grouping precision (75%) than a serious direct other option. By transforming this assignment into a double order – upper versus medium and lower class – the proposed classifier arrives at the exactness of 82%. Every day, the feeling and assessment of various individuals over the world are reflected as short messages utilizing microblogging stages Ghazouani et al. (2019) [9] reviewed and researched Twitter user’s classification according to SES levels. In this, they mainly focused on the topics like “Tweet Gathering and Analysis, Classification Methods of SES, Data Corpuses and Resources, Inference Methods for SES Evaluation on Twitter, Regression Methods, Demographic Attributes, Temporal Information, Social Network Relations, Spatial Information, SES Features and Indicators, and Evaluation of Socioeconomic Status”.

SES is a significant social and economic perspective broadly concerned. Evaluating individual SES can help related associations in settling on an assortment of policy choices. Customary methodology experiences a very significant expense in gathering enormous scope SES-related review information. With the pervasiveness of advanced cells, cell phone information has become a novel information hotspot for anticipating singular SES with minimal effort. In any case, the assignment of foreseeing singular SES on cell phone information likewise proposes some new difficulties, including inadequate individual records, scant express connections, and restricted named tests, indifferent in earlier work confined to territorial or household unit situated SES expectation [10]. The social determinants of well-being writing regularly SES as a critical factor in representing ladies' height. a setup marker of human government assistance at the populace level—utilizing customary relapse. Notwithstanding, this writing comes up short on a precise distinguishing proof of the prescient influence of SES just as the conceivable non-direct connections between the proportions of SES (material abundance, education, and occupation) in anticipating variety in ladies' tallness [11]. Zhao et al. (2020) [12] evaluated the predictive power of SES. In this research, they used demographic health surveys from 66 middle and low countries 1,273,644 women individual samples from 1994 to 2016. They assessed this data with 7 ML algorithms and concluded with the best ML performer NN. The NN was performed for SES prediction with 31.52 variances or MSE value.

3. Methodology and Models

3.1 Proposal Model

The figure1 shows the proposed model of the Socio-economic system with statistical and cluster studied models. In this, we collected the information from each house of rural and urban areas of the Rajahmundry constitution, district of East Godavari, A.P, India. We have gathered all the information with a good questionnaire and store the necessary information in the secondary storage section. After that, we extract the information with features and classes into a data set as *.csv format. Some of the information is plotted on the Rajamahandravaram (Rajahmundry) map with longitude and latitude measures. The targeted class attribute contains four classes that are rich, upper-middle-class, middle class, and poor. For this investigation, we extract the feature attributes as per household information that are personal-data, Socio-status, Economical-status, Living-status, Health-wealthy status, and so on. In this, we have to use 1742 household records with 49 attributes and construct *.csv format file and stored into the secondary storage section as well as plot the data points on the Rajamahandravaram map with
full details for the convenience of visualization. We have to conduct the statistical analysis of the SES data set with statistical parametric equations like mean, median, max, and min and generate the reports for analysts. Cluster analysis is important to analyse the data set for the identification of important factors related to the target class values, dimensional reduction, and ranking of the attributes. In clustered studies, we choose the k-means and HC clusters using PCA. Before that, we conducted the pre-processing, the data set using the min-max method for better analysis.

3.2 Data Set Description

Rajahmundry renamed as Rajamahandravaram is one of the major consistencies of East Godavari district in Andhra Pradesh, India. We gather information about each house from this constitution area of rural and urban. Nearly, we collected the 1742 samples as per socio-economic ratios and area wise ratios with good questionnaires between 2018 and 2019. Some of the data is plotted on the Rajamahandravaram Map using longitude and latitude values. The figure 2 shows location details and detailed information about plotted houses clicking on that point of more details button. For this experiment, we used 48 feature attributes and one class that is the status (rich, poor, middle, and upper-middle classes) described in table 1.

| Dataset               | Data Type   | Description                                                                 |
|-----------------------|-------------|------------------------------------------------------------------------------|
| Area(R-0/U-1)         | Discrete (Integer) | House hold from Rural (0) or Urban (1)                                         |
| Family Size           | Continues (Integer) | Total members in House, range is 1 to 16                                       |
| Male Size             | Continues (Integer) | Total male members in House, range is 1 to 8                                   |
| Female Size           | Continues (Integer) | Total female members in House, range is 0 to 8                                 |
| below 18              | Continues (Integer) | Total members less than 18 age in House, range is 0 to 5                      |
| above 18              | Continues (Integer) | Total member >18 age in House, range is 0 to 12                               |
| married people        | Continues (Integer) | Total married members in House, range is 0 to 8                               |
| No. of children       | Continues (Integer) | Number of children in House, range is 0 to 2                                  |
| No. of literates      | Continues (Integer) | Number of literates in House - range is 0 to 12                               |
| High Qualification    | Discrete (Integer) | Qualification of house members range 0 to 5                                  |
| No. of Workers        | Continues (Integer) | Number of workers in House - range is 0 to 8                                  |
| Child work below 15   | Continues (Integer) | Number of Child workers in House - range is 0 to 1                            |
| Occupation (0-5)      | Discrete (Integer) | Occupation of house hold members range is 0 to 7                              |

0-very low 3-moderate 5- very high
| Major Work       | Discrete (Integer) | Work Category 0 to 5 0-very low 3-moderate 5-very high |
|------------------|--------------------|--------------------------------------------------------|
| Ration Cards    | Discrete (Integer) | Ration Cards of house 1-white and 2-pink               |
| Health cards    | Discrete (Integer) | Health cards range is 0 to 1, 0-noyes                  |
| No. of Diseased people | Continues (Integer) | Number of Diseased members in House range is 0 to 2 |
| No. of Handicapped | Continues (Integer) | Number of Handicapped members in House range is 0 to 1 |
| Bikes           | Continues (Integer) | Number of Bikes in House range 0 to 3, 0 for none 3 for ‘3 or more’ |
| Cars            | Continues (Integer) | No. of Cars in House range 0 to 3, 0 for none 3 for ‘3 or more’ |
| Others          | Continues (Integer) | No. of other Vehicles in House range 0 to 3, 0 for none 3 for ‘3 or more’ |
| Own house (1) or rental (0) | Discrete (Integer) | Status of House 0 for ‘Rental’ 1 for ‘Own’ |
| Land (cents)    | Continues (Real)   | Agriculture Land in cents range is 0 to 2000.0         |
| Gold            | Continues (Real)   | Gold in grams in House range is 0 to 1500.0            |
| Annual Income   | Continues (Real)   | Annual income of house range is 27000.0 to 15000000.0  |
| Income from Govt| Continues (Real)   | Income from Govt. of house range is 0 to 14800000.0    |
| Income from pension | Continues (Real) | Income from Pension of house range is 0 to 40000.0     |
| Social status (0/1/2/3/4) | Discrete (Integer) | Social Status 1 for ‘ST’ 2 for ‘SC’ 3 for ‘OBC’ 4 for ‘OC’ 0 for ‘none’ |
| Nearest Hospital in Km | Continues (Real) | Hospital distance range 1.0 to 5.0                     |
| Nearest Primary School in Km | Continues (Real) | Primary School distance in KMs range 1.0 to 5.0       |
| Nearest High School in Km | Continues (Real) | High Schools distance in KMs range 1.0 to 10.0        |
| Nearest College in Km | Continues (Real) | College distance in KMs range 2 to 14.0                |
| Nearest University | Continues (Real) | University distance in KMs range 30.0 to 45.0         |
| Addicted persons to smoke drinks | Discrete (Integer) | Habits of drink and smoke in house 0-None 1-Partial 2-Addicted 3-Extreme |
| Building model  | Discrete (Integer) | Building model range is 0 to 5 0-low level 3-moderate 5-high level |
| Water sources   | Continues (Real)   | Water Facilities 0-none or poor 1-moderate 2-good facility 3-very good |
| Toilets facilities 1/0 | Discrete (Integer) | Toilet facilities 0 for ‘no’ and 1 for ‘yes’ |
| Electricity/1/0 | Discrete (Integer) | Electricity facilities 0 for ‘no’ and 1 for ‘yes’ |
| TV 1/0          | Discrete (Integer) | TV facilities 0 for ‘no’ and 1 for ‘yes’               |
| Fridge 1/0      | Discrete (Integer) | Fridge facilities 0 for ‘no’ and 1 for ‘yes’           |
| Air Condition 1/0 | Discrete (Integer) | Air Condition facilities 0 for ‘no’ and 1 for ‘yes’    |
| Heater 1/0      | Discrete (Integer) | Heater facilities 0 for ‘no’ and 1 for ‘yes’           |
| Computer 1/0    | Discrete (Integer) | Computer facilities 0 for ‘no’ and 1 for ‘yes’         |
| Fuel for cooking 1/0 | Discrete (Integer) | Fuel for cooking 0 for ‘Non-Gas’ and 1 for ‘Gas’       |
| Income status in past 5 years (2/0/1) | Discrete (Integer) | 0-Decremental Income 1-Remain same 2-Increment |
| Internet 1/0    | Discrete (Integer) | Internet facility 0-No and 1-Yes                      |
| Migrated family or not | Discrete (Integer) | Migrated from other places 0-No and 1-Yes             |
| SES Levels (1-4) | Discrete (String) | Levels 1-Poor 2-middle 3-upper middle 4-Rich          |

![Figure 2: Plotting data points according to latitude and longitude on Rajamahandraravaram map](https://example.com/image)

### 3.3 Plotting Data Points
The figure shows the Rajamahandravaram map where it contains the collected household data points that plotted on map with location points (blue coloured bubble points) with latitudes and longitudes. We established this map with attach with our data set using latitudes and longitudes values. Moreover, in this we provide statistical analysis results of the data set, individual data, bar charts and pi charts of the attributes, searching and upload data record facilities, defined queries projections and so on. The local cite information with IP address that is 13.233.164.180.

3.4 k-Means algorithm

K-Means is an apportioning clustering technique which displaces objects by moving starting with one cluster then onto the next beginning from an initial partitioning [14][15]. The motivation of the cluster examination is to segment n observations into K groups in which every perception has a place with the cluster with the closest mean. It is one of the simplest unsupervised ML that solve the problems with clustering. The K-Means is a transformative and evolutionary method that picks up its name from its technique for activity [16] [17]. The k-means algorithms as follows

step 1: The model chooses K data points as the centres of initial cluster using mean values
step 2: Each data point in the data is allocated to the nearer cluster group using Euclidean distance between each data point and each centre of the cluster.
step 3: Each cluster is determined by cluster average data points.
Step 4: Steps 2 and 3 repeats until the clusters converge.

3.5 Hieratical Clustering

Agglomerative HC has been the predominant approach to deal with embedded group of clusters. The main aim is to focus toward useful models and methods that both efficient and effective [18]. It is regularly useful to recognize strategy, including a minimization standard and the objective structure of a 2-way tree speaking to the fractional request on subsets of the power set; instead of an execution, which identifies with the detail of the calculation utilized. Similarly, as with numerous other multivariate strategies, the items to be classified have mathematical estimations on cluster factors or characteristics. A mathematical system of this sort isn't the one in particular which can be utilized to plan grouping calculations. Reasonable elective types of capacity of a rectangular cluster of qualities are not conflicting with review the issue in mathematical terms [19][20].

3.6 Principal Component Analysis (PCA) Algorithm

PCA is a symmetrical linear transformation that moves the information to new coordinate system to such an extent that the best variance by any projection of the information comes to lie on the primary arrange (first principal component (PC)), the second most noteworthy difference lies on the subsequent coordinate (second PC), and etc. PCA used to lessen measurements of information absent a lot of loss of data. It is utilized in AI in frequently.

1. Get the mean vector, in this calculation every data point value is subtracted by the mean value and divided by the Standard deviation (SD). Below equation shows mean vector value.

\[
Z = \frac{\text{Factor Value} - \text{Mean Value}}{\text{Standard Deviation (SD)}}
\]

In this, every variable value is changed in normalized value (same scale value).

2. Construction of Covariance Matrix using assembled samples from mean datapoint value matrix.

This progression is to perceive how the components of the informational collection are transforming from the mean with respect to each other. At times, the components are very related or associated so they contain abundance or excess information [21]. Thusly, in order to recognize these relationships, we compute the covariance model network. The covariance grid is a \( n \times n \) symmetric lattice with n
measurements that has as entrances the covariance identified with each and every likely pair of the underlying components.

The covariance \( n \times n \) Matrix is

\[
\begin{bmatrix}
\text{Cov}(x_0, x_0) & \text{Cov}(x_0, x_1) & \ldots & \text{Cov}(x_0, x_n) \\
\text{Cov}(x_1, x_0) & \text{Cov}(x_1, x_1) & \ldots & \text{Cov}(x_1, x_n) \\
\vdots & \vdots & \ddots & \vdots \\
\text{Cov}(x_n, x_0) & \text{Cov}(x_n, x_1) & \ldots & \text{Cov}(x_n, x_n)
\end{bmatrix}_{n \times n}
\]

3. Register the Eigen values and Eigenvectors.
4. Register the premise vectors.
5. Correspond to each feature data point as a LC (linear combination) of principle vectors.

4. Experimental Setup and Result Analysis

In this section, we discuss the data set attributes empirical statistical analysis. Furthermore, we apply unsupervised ML models k-means and HC cluster on SES Dataset and analyse their abilities.

4.1 Statistical Analysis

We collected the data from rural and urban areas of the Rajahmundry constitution, East Godavari District, A.P., India. For this, collected sampling data is as per ratios of socio-economic status. The rural area samples are 946, and urban area samples are 796 (Total 1742). As per the statistical analysis of the household dataset, some of the houses contain the average 4 to 5 where the mean value is 4.381 and Std. Dev is 1.467. Some of the houses have only one member (min value is 1), and some of the houses contain 16 (max value). Each house contains at least one male person (min value male persons in a house is 1) and a maximum of 8 male persons as well as on average 2 to 3 persons per one house. On the other hand, the female persons’ min value is 0, and the max value is 8, and mean and SD values are 1.975 and 0.776 respectively, which means every house contains on an average one to two females. As per statistics some good conditions that very fewer child workers, average young generation 2 to 3 people in every house and average 1 to 2 workers in each house. Another good thing, the number of diseased people and the number of disabled persons is very less percentage that the mean values are 0.066 and 0.024 respectively. Table 2 shows the detailed statistical analysis of each attribute of the data set.

Table 2: Detailed statistical analysis of Rajahmundry SES data set

| Attribute                  | Min | Max | Mean | SD  | Yes | No |
|----------------------------|-----|-----|------|-----|-----|----|
| Family Size                | 1   | 16  | 4.381| 1.467|     |    |
| Male                       | 1   | 8   | 2.406| 0.773|     |    |
| Female                     | 0   | 8   | 1.975| 0.776|     |    |
| below 18                   | 0   | 5   | 1.592| 0.882|     |    |
| above 18                   | 0   | 12  | 2.802| 1.183|     |    |
| married people             | 0   | 8   | 2.037| 0.375|     |    |
| No. of children            | 0   | 2   | 0.131| 0.362|     |    |
| No. of literates           | 0   | 12  | 0.738| 0.826|     |    |
| No. of Workers             | 0   | 8   | 1.693| 0.693|     |    |
| Child. work                | 0   | 1   | 0.006| 0.079|     |    |
| No. of Disabled people     | 0   | 2   | 0.066| 0.268|     |    |
| No. of Handicapped         | 0   | 1   | 0.024| 0.152|     |    |
| land(cents)                | 0   | 2000| 140.576| 202.602|     |    |
| Gold(grams)                | 0   | 1500| 34.788| 51.792|     |    |
| Annual Income              | 27000| 15000000| 331504.6| 467944.2|     |    |
| Income from Govt.          | 0   | 1480000| 24055.68| 117628.8|     |    |
| income from pension        | 0   | 40000| 408.726| 1351.704|     |    |
| Income from private        | 0   | 15000000| 308868| 467484.8|     |    |
| Hospital in Km             | 1   | 6   | 3.637| 0.836|     |    |
| Pri mary School in Km      | 1   | 5   | 2.846| 1.901|     |    |
| High School in Km          | 1   | 10  | 3.832| 1.899|     |    |
| College in Km              | 2   | 14  | 6.866| 2.057|     |    |
| University in Km           | 30  | 45  | 35.065| 4.185|     |    |

4.1.3 Social status

| Attribute                | Type | Value |
|--------------------------|------|-------|
| Ration Cards             | white| 660   |
| Fuel for cooking         | Gas  | 1575  |
| Social status            | ST   | 31    |
|                           | SC   | 190   |
|                           | BC   | 896   |
|                           | OC   | 625   |
| Addicted persons         | None | 736   |
| to smoke and drinking in House | Partial | 826 |
|                          | Addicted | 158 |
|                          | Extreme | 22   |
## Some other Types of Attribute Statistics

| Attribute                  | Type                  | Value |
|---------------------------|-----------------------|-------|
| Literacy and Educators Houses | None or Below 10th    | 296   |
|                           | 10th Standard         | 428   |
|                           | Inter Level or ITI    | 386   |
|                           | Degree Level          | 272   |
|                           | Technical Degree or Other | 249 |
|                           | P.G. level            | 101   |
|                           | Professional or Ph.D. Level | 10 |
| Occupation                | Major Work Occupation in House |       |
|                           | No Work or Very less  | 6     |
|                           | Seasonal Workers      | 463   |
|                           | Average or Daily wagers | 497 |
|                           | Permanent Low salary  | 345   |
|                           | Permanent Middle Salary | 364 |
|                           | Permanent High Salary | 62    |
|                           | Business or Organizers | 5     |
| Having Bikes in House     | None                  | 939   |
|                           | One                   | 671   |
|                           | Two                   | 118   |
|                           | More Than Two         | 14    |
| Having Cars in House      | None                  | 1607  |
|                           | One                   | 128   |
|                           | Two                   | 5     |
|                           | More Than Two         | 2     |
| Having Other Traveling Recourses | None | 1615  |
|                           | One                   | 121   |
|                           | Two                   | 5     |
|                           | More Than Two         | 1     |
| Target Class (SES levels) | Rich                  | 73    |
|                           | Middle class          | 794   |
|                           | Upper Middle class    | 526   |
|                           | Poor                  | 349   |

An important thing for the economic status that fully depends on annual income for each house and its resources that are from the public, private, asserts, and work, and so on. As per statistics, the annual income min value is 27000/- and the max value is 80,000,000/-. The main income resources from private, government, or pension schemes. The detailed analysis is shown in table 1. The educational and health resources are also available within the distance of every house. The figure shows the SES level samples that are rich, middle-class, upper-middle-class, and poor as per the ratios of Rajamahandhravaram. The Rich and Poor positions are in the 4th and 3rd, and the middle class and upper-middle-class occupy the first and second positions in the area. Figure 3 shows the counting number of SES status levels were low to high as per ratios.

### 4.2 Matrix (Heat Map):

Figure 4 shows the Correlation heat map of each attribute of the Rajamahandhravaram SES dataset. The correlated values are specified with the colour that the indicator mentioned in the figure. The values are measured between -1 and 1. If the cell colour is a dark red that its indication is neutrally correlated with each other. The dark blue indicates the zero correlation. The attribute electrical is correlated marginal with all other attributes. The relationship between attributes in the dataset is vital reasons to evaluate the fitness of the data set that one variable value is fully dependent on other variables or weak associations with others. Sometimes the variable relations are peculiar in that one unknown variable value depends on two or more variable values. The correlation values are useful for modelling and analysing the data in a better way. The correlation is defined using two attributes or variables of statistical value relationships. The correlation value is between -ve (negative) and +ve.
(positive) values. +ve value of correlation represents that change of attributes or variable movement in the same direction. -ve value of correlation represents that change or relationship of attributes or variable movement in the reverse direction. Neutral or zero correlation represents that both variables are divergent or unrelated. In some situations, two or more of the variables in the dataset are related very rightly named as multicollinearity that it impacts more and more on the performance of some algorithms like linear regression. In this situation, we can remove highly correlated attributes from the experimental dataset for improving model performance.

Figure 4: Correlation Matrix Analysis of Rajamahandhravaram SES Dataset

4.3 k-Means Clustering

In the k-means cluster analysis, we give the value 4 for the number of clusters that clusters are cluster 0, cluster 1, cluster 2 and cluster 3. The total data set size is 1742 (no. of houses information). The cluster 0 contains 485 instances, the cluster 1 is dealing with 480, the cluster 2 contains highly 547 instances and the cluster 3 contains 230 instances. The table 2 shows the k-means unsupervised ML algorithm centroids of each cluster. In this we describe some of the attributes centroids related to each cluster and full data set also. The total data set concened with the target class attribute SES level value Middle-class, and cluster 0 and 3 are related to Upper-middle. The cluster 0 is related to class attribute value poor and cluster 1 is related to middle. The detailed cluster centroids details are shown in the table 3.

Table 3: K-Means Cluster Centroids of Rajahmundry SES data set

| Attributes                        | Full Data (1742) | Cluster 0 (485) | Cluster1 (480) | Cluster2 (547) | Cluster3 (230) |
|----------------------------------|------------------|----------------|----------------|----------------|----------------|
| Family Size | 4.3812 | 4.4763 | 3.8438 | 4.4863 | 5.0522 |
| Male | 2.4059 | 2.4474 | 2.1750 | 2.4516 | 2.6913 |
| Female | 1.9753 | 2.0289 | 1.6687 | 2.0347 | 2.0739 |
| below 18 | 1.5924 | 1.4515 | 1.4063 | 1.6399 | 2.1652 |
| above 18 | 2.802 | 3.0289 | 2.4354 | 2.8867 | 2.887 |
| married people | 2.0373 | 2.0804 | 1.9812 | 2.0329 | 2.0739 |
| No. of children | 0.1309 | 0.1134 | 0.1021 | 0.1225 | 0.2478 |
| No. of literates | 0.7382 | 1.1567 | 0.1854 | 0.5978 | 1.3435 |
| High Qualification | 0 | 3 | 0 | 0 | 0 |
| No. of Workers | 1.6935 | 1.7464 | 1.4708 | 1.7623 | 1.8826 |
| Child works below 15 | 0.0063 | 0 | 0.0146 | 0.0055 | 0.0043 |
| Occupation (0-5) | 2 | 4 | 1 | 2 | 2 |
| Major Work | 1 | 5 | 1 | 2 | 2 |
| Ration Cards | 1 | 2 | 1 | 1 | 2 |
| Health cards | 1 | 0 | 1 | 1 | 0 |
| No. of Diseased people | 0.066 | 0.0577 | 0.0604 | 0.0878 | 0.0435 |
| No. of Handicapped | 0.0235 | 0.0227 | 0.0333 | 0.0201 | 0.013 |
| Bike | 1 | 1 | 0 | 1 | 1 |
| Own house (1) or rental (0) | 1 | 1 | 0 | 1 | 1 |
| land(cents) | 140.5758 | 210.2268 | 28.1104 | 102.9068 | 318.0 |
| Gold | 34.7878 | 60.4396 | 9.0156 | 21.9481 | 65.0174 |
| Annual Income | 331504.6 | 586360.8 | 9104.67 | 222734.9 | 554608.7 |
| Income from Govt | 24055.68 | 72288.66 | 395.8333 | 6133.455 | 14347.83 |
| income from pension | 408.7256 | 154.6392 | 420.8333 | 420.8333 | 926.087 |
| Income from private | 308868 | 514080.4 | 89654.17 | 213515.5 | 560400 |
| Nearest Hospital in Km | 3.6366 | 3.3897 | 3.7229 | 3.7623 | 3.6783 |
| Nearest Primary School in Km | 3.8462 | 3.8536 | 3.9958 | 3.7331 | 3.787 |
| Nearest High School in Km | 3.8318 | 3.8309 | 3.7125 | 3.8062 | 4.1435 |
| Nearest College in Km | 6.8657 | 6.7711 | 6.9354 | 6.9744 | 6.6609 |
| nearest university | 68.0654 | 69.066 | 68.4104 | 68 | 65.3913 |
| Building model | 1.0327 | 1.6866 | 0.25 | 0.9452 | 1.4957 |
| Water sources | 1 | 2 | 1 | 1 | 1 |
| TV | 1 | 1 | 0 | 1 | 1 |
| Fridge | 1 | 1 | 0 | 1 | 1 |
| Air Condition | 0 | 1 | 0 | 1 | 1 |
| Heater | 0 | 1 | 0 | 0 | 1 |
| SES Levels | Middle | Upper- Middle | Poor | Middle | Upper- Middle |

Figure 5: k-Means Cluster Analysis (clusters vs SES classes) of Rajamahendravaram SES Dataset
Figure 5 shows the cluster analysis between cluster number and target classes. In this plot, the X-axis specifies the SES status and Y-axis indicates the cluster number. The clusters form different colors and describes related SES status. The cluster 0 (color blue) contains most of instances from upper-middle class very less instances from middle and rich classes. The cluster 1 (red color) is constructed with more poor and less middle-class instances. The cluster 2 color is green that it is constructed with most of middle-class elements and very less elements of poor. The cluster 3 is combination of middle and upper-middle class elements.

4.4 Hierarchical Cluster (HC) Attribute (Column) wise Examination on SES Dataset:

Figure 6 shows the HC column-wise complete link clustering. In this, the attributes closure relations of SES data set are analysed one to other. The Euclidian distance is the measurement of the HC construction. The complete HC network is constructed with 7 clusters. Each cluster is represented with each color shown in figure 6. The cluster 1 is constructed with the pairs of attributes like occupation and Major Work with height 0.4, married People and income with height 3.93 status, female and below 18 with height 0.334. In the C1 cluster, highly correlated and related attributes are female and below-18-age.

The cluster 2 is constructed with the pair of attributes like male and above-18-age with 0.33 height, this pair is connected to the social-status with height of 0.71. In this, the connected attributes in chain and increase the height values. In cluster 2 highly correlated attributes are male and above-18-age. The cluster 3 is constructed with 12 essential attributes. In this cluster, highly correlated attributes Diseased-people and income-from-Govt with height 0.109 and next chain connection attributes are car, others and so on. The cluster 3 and cluster 4 are constructed with single attributes that are gold and nearest-university. The cluster 6 is constructed with two attributes that are income-private and annual-income with height 0.88. The cluster 7 has 3 attributes that are nearest-hospital and family-size with 0.64 connected to
nearest-college attribute with height 0.912. As per analysis, the cluster 6 contains highly correlated attributes than other clustered attributes and cluster 7 contains non-correlated or low-correlated attributes where distance is very high than other cluster attributes. The overall cluster analysis is with complete linkage, maximum depth is 8, the height ratio is 14%, and distance measurements are calculated with Euclidian distance formula.

Figure 7 describes the resulting HC Attribute Analysis on attributes Target class, Annual Income, and High Qualification. The X-axis represents the Annual income (AI x 1e+06), Y-axis represents the High Qualification that the range is 0 to 6, and the target class data points represented with colour bubbles (red-poor, green-rich, blue-middle, and orange-upper-middle). As per the analysis, income is related to higher education impacts on SES levels that poor class attribute data points decrease inversely proportional to the higher education. Most of the data points in higher education levels 3,4,5 and 6 contain middle, upper-middle, and rich classes only. The highest degree of Ph.D. (level 6) is pursued by rich and upper-middle classes only. In this analysis, we observe that poverty is related to education

4.5 HC Row wise Analysis on Rajahmundry SES Dataset:

Figure 8 describes the complete link HC with row-wise class attribute values like Rich ‘class, Poor-class, Middle-class, and Upper-Middle-class classes. In this, the target class values (Rich, Poor, Middle, and Upper-Middle of SES dataset) are elaborated with 8 clusters in closure relations at the 24 distance. As per feature attribute values and target class values, the closure relation clusters are determined and analysed using Euclidean distance measures. The distance is at 3.38 in cluster 5 that the values are rich. C1 cluster related to rich-class, C2 cluster specifies the upper-middle-class values, C3, and C4 clusters represent the rich and middle-class independently. The C5 contains all 4 values that most of the data points are middle and upper-middle classes, as well as the C6 cluster, which covers middle and upper-middle-class values. C7 and C8 clusters dedicate to the rich class attributes.
4.6 HC without PCA Projections Analysis:

Figure 9 depicts the HC projections without PCA according to attributes concerning Annual-Income, land in cents, having a bike, Gold, and Building models. The blue elements indicate the middle-class, the red elements specify the poor data points, the green-coloured elements describe rich, and the orange-
coloured data points specify the upper-middle-class. Most of the rich data points (green coloured dots) have high Fraction of Annual-Income, land in cents, having a bike, Gold, and Building models. These components are related to the HC target class values like rich, poor, middle, and upper-middle classes. The projections are described in Figure 9 without implementing PCA. These five dimensions are the first 5 listed ranks in the experiment according to SES classes with an accuracy of 99.13%.

4.7 PCA-HC Projections Assessments on SES Dataset:

Figure 10 demonstrates the HC cluster projections with PCA according to attributes concerning Annual-Income, land in cents, having a bike, Gold, and Building models. The blue elements indicate the middle-class, the red elements specify the poor data points, the green-coloured elements describe rich, and the orange-coloured data points specify the upper-middle-class. Most of the rich data points (green coloured dots) have high Fraction of Annual-Income, land in cents, having a bike, Gold, and Building models. These components are related to the HC target class values like rich, poor, middle, and upper-middle classes. The projections are described in Figure 9 implementation of PCA-HC. These five dimensions are the first 5 listed ranks in the experiment according to SES classes with an accuracy of 99.93%.

5. Conclusion

Statistical analysis and Machine learning are very essential and crucial current days for analysing the Socio-Economic Status problems. In this research work, we collected and analysed Rajamahandarvaram SES dataset using statistical models and identify the important features of the SES. As well as, we conducted the cluster analysis like k-means, HC, and HC-PCA on SES dataset and get good visualization results from this experiment.

References

[1] López-Goyburu, P., & García-Montero, L. G. 2018. The urban-rural interface as an area with characteristics of its own in urban planning: A review. Sustainable cities and society, 43, 157-165.
[2] Chapman, B. P., Fiscella, K., Kawachi, I., & Duberstein, P. R. 2010. Personality, socioeconomic status, and all-cause mortality in the United States. American journal of epidemiology, 171(1), 83-92.
[3] Wang, M. T., Henry, D. A., & Degol, J. L. 2020. A development-in-sociocultural-context perspective on the multiple pathways to youth's engagement in learning. In Advances in Motivation Science (Vol. 7, pp. 113-160). Elsevier.
[4] Perrotta, C., & Williamson, B. 2018. The social life of Learning Analytics: cluster analysis and the 'performance'of algorithmic education. Learning, Media and Technology, 43(1), 3-16.
[5] Mishra C. P. 2012. Nexus of poverty, energy balance and health. Indian journal of community medicine : official publication of Indian Association of Preventive & Social Medicine, 37(2), 71–78.
[6] Conger, R. D., Conger, K. J., & Martin, M. J. 2010. Socioeconomic Status, Family Processes, and Individual Development. Journal of marriage and the family, 72(3), 685–704.
[7] Schäfer, I., Hansen, H., Schön, G. et al. 2012 The influence of age, gender and socio-economic status on multimorbidity patterns in primary care. first results from the multicare cohort study. BMC Health Serv Res 12, 89.
[8] Lampos, V., Aletras, N., Geyti, J. K., Zou, B., & Cox, I. J. 2016, March). Inferring the socioeconomic status of social media users based on behaviour and language. In European Conference on Information Retrieval pp. 689-695. Springer, Cham.
[9] Ghazouani, D., Lancieri, L., Ounelli, H., & Jebari, C. 2019, September. Assessing socioeconomic status of Twitter users: A survey. In Proceedings of the International Conference on Recent
Advances in Natural Language Processing (RANLP 2019) (pp. 388-398).

[10] Danese A, Moffitt TE, Harrington H, et al. 2009 Adverse Childhood Experiences and Adult Risk Factors for Age-Related Disease: Depression, Inflammation, and Clustering of Metabolic Risk Markers. Arch Pediatr Adolesc Med. 2009;163(12):1135–1143.

[11] Daoud, A., Kim, R., & Subramanian, S. V. 2019. Predicting women's height from their socioeconomic status: A machine learning approach. Social Science & Medicine, 238, 112486.

[12] Zhao, T., Huang, H., Yao, X., & Fu, X. 2020. Predicting individual socioeconomic status from mobile phone data: a semi-supervised hypergraph-based factor graph approach. International Journal of Data Science and Analytics, 9(3), 361-372.

[13] Bose, E., & Radhakrishnan, K. 2018. Using unsupervised machine learning to identify subgroups among home health patients with heart failure using telehealth. CIN: Computers, Informatics, Nursing, 36(5), 242-248.

[14] Qi, J., Yu, Y., Wang, L., & Liu, J. 2016, October. K*-means: An effective and efficient k-means clustering algorithm. In 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom) (pp. 242-249). IEEE.

[15] yakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. 2018, April). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In IOP Conference Series: Materials Science and Engineering (Vol. 336, No. 1, p. 012017). IOP Publishing.

[16] Meng, Y., Liang, J., Cao, F., & He, Y. 2018. A new distance with derivative information for functional k-means clustering algorithm. Information Sciences, 463, 166-185.

[17] Jothi, R., Mohanty, S. K., & Ojha, A. 2019. DK*-means: a deterministic k-means clustering algorithm for gene expression analysis. Pattern Analysis and Applications, 22(2), 649-667.

[18] 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

[19] HC1. Murtagh F. 2011 Hierarchical Clustering. In: Lovric M. (eds) International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg.

[20] HC2. Murtagh, F., & Contreras, P. 2017. Algorithms for hierarchical clustering: an overview, II. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7(6), e1219.

[21] Vital, T. P., Satyanarayana, L. V., & Murthy, G. S. N. (2020). Empirical Analysis on Uddanam CKD with Unsupervised MLs Including PCA. In Computational Intelligence in Pattern Recognition (pp. 571-587). Springer, Singapore.