INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Intelligent socio-economic status prediction system using machine learning models on Rajahmundry A.P., SES dataset

V Balasankar[1]*, Suresh Varma Penumatsa[2], PanduRanga Vital Terlapu[3]

1 Research Scholar, Department of CSE, Adikavi Nannaya University, Rajahmundry, AP, India
2 Professor, Department of CSE, Adikavi Nannaya University, Rajahmundry, AP, India
3 Associate Professor, Department of CSE, Aditya Institute of Technology and Management, Tekkali, Srikakulam, Andhra Pradesh, India

OPEN ACCESS

Received: 30.08.2020
Accepted: 20.09.2020
Published: 13.10.2020

Editor: Dr. Natarjan Gajendran

Citation: Balasankar V, Penumatsa SV, Terlapu PRV (2020) Intelligent socio-economic status prediction system using machine learning models on Rajahmundry A.P., SES dataset. Indian Journal of Science and Technology 13(37): 3820-3842. https://doi.org/10.17485/IJST/v13i37.1435

***Corresponding author**.

balasankar.v@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Background: Developing economic and social systems and assuring the efficiency of economic and social processes is the major task for the government of any country. Predictable machine learning (ML) models are used for analyzing data sets that allow more efficient enterprise management. Now a day, the research on Socio-Economic Status (SES) and Machine Learning (ML) is very crucial to find socio-economic inequalities, and take further actions that are preventions, protections, and suppressions. **Objectives:** The main objective of this research is to understand the Socio Economic System issues and predicting SES levels on particular area like Rajahmundry, AP, India using statistical analysis and machine learning methodologies. **Methods:** In this, we analyze the data that is collected from Rajahmundry (Rajamahandravaram) ,Andhra Pradesh, India with 48 feature attributes (dimensions), and one target four class attribute (poor, rich, middle, upper-middle). The SES levels like poor, rich, middle, and upper-middle classes are predicted by 5 ML algorithms. **Findings:** In this paper, we conduct the statistical analysis of each attribute, and analyze and compare the performance accuracies using confusion matrix, performance parameter (classification accuracy, Precision, Recall, and F1) values and receive operating characteristic (ROC) under AUC values of five efficient ML algorithms like Naïve Bayes, Decision Trees (DTs), k-NN, SVM (kernel RBF) and Random Forest (RF). We observed that the RF algorithm showed better results when compared with other algorithms for the Rajahmundry AP SES dataset. The RF algorithm performs 97.82% of classification accuracy (CA) and time is taken for model construction 0.41 seconds. The next superior performed ML model is DTs with 96.67% of CA and 0.16 seconds for model construction. **Novelty:** Comprehensive analysis indicates that the novel AP SES Dataset with empirical statistical analysis gives the good results and predicts the SES levels with RF model is very effective.

Keywords: Machine Learning; socio-economic status; Rajahmundry; household; poverty

1 Introduction

The Socio-Economic System (SES) at the regional or provincial level refers to the way economic and social factors influence each other in households and local community units. These systems significantly affect the pollution, deforestation of the environment, contamination, catastrophic events, and production of energy and use [(1)]. Through telecoupled frameworks, these affiliations can extend to worldwide [(2)]. Local economies, environmental hazards, and food insecurity are some of the problems, which impact SES. The household family size is one of the important socio-economic elements, that family unit is very small or nuclear, that traditionally two guardians and their kids living under the same rooftop. Previously, families frequently hindered more distant family individuals, for example, grandparents. With the move in the number of individuals under one rooftop, there has been an expansion of indirect energy consumption. The individuals have been moving towards single individual families as our cultural standards develop. More households mean more energy used to do things like power more number of TVs and utilize more lights. It likewise implies more geological land space [(3)], which has been a move in networks over the globe. Education, agriculture, industries, and grate administration are vital roles in that area [(4)]. The word, socio originates from the social and denotes to individuals and the level they fit into their residential areas or communities in which they live. It reflects how well those educated, have a job, and so on. Economic denotes the financial or money related positions of the people in society, and also their regular earnings, own house and own assets, and so on. It is a combination of so many factors like education, occupation, family economic and social position, and so on. Society Built by the principle of difference [(5)]. In all social associations, there is a social separation of the populace based on different variable factors. One of the main factors is gender separation in culture. The lead of genders is canalized in various ways directly from the family. Religion is also a worldwide organization [(6)]. The Statistical analysis and probabilistic value estimations are things on the SES to preventions and predictions for the development of a particular area, state, or country. Using some of the indicators or markers assumes the poverty levels or other SES levels of that area. Machine Learning is a tool to predict or assessments future results. Machine learning with socio-economic research is an important topic in this decade to identify the economic and social levels of human as well as living status predictions.

East Godavari district is one of the northern coastal areas of A.P., India, being the division of the flush Godavari delta; horticulture and aquaculture are significant economies for this District. With the ongoing discoveries of hotspots for oil and Natural gas, it expanded its pace in the industrial part too. It is home for two significant fertilizer 135 industrial facilities and scarcely any gas-based force plants and petroleum treatment facilities. Presently it is the One of the biggest oil and Gas Hub in India [(7)]. Rajamahandravaram (formally known as Rajahmundry) is one of the consistencies of the East Godavari district in Andhra Pradesh, India [(8)]. In this research, we gathered the household information from this place in so many dimensions and plotted it on the Rajamahandravaram map. Mainly, we have to analyze on socio-economic status in multidimensional indicators. Those are living, education, occupation, job status and satisfaction, income, resources, assets and liabilities, agriculture, business, health, happiness, diseases and disabilities, family size for each house with statistical analysis, and machine leering methods. Machine learning and socioeconomic status research are very crucial these days to predict poverty levels of a particular area, social and economic inequalities. As per ML results, the government or organization will decide to preventions and protection about SES levels. In this research, we used popular ML methods to predicted SES levels of sampling data of Rajamahandravaram.

This research work is related to Rajamahandravaram SES with an empirical dataset. As per the review of the literature, very little work has been reported on Rajamahandravaram socio-economic studies using machine learning (ML) for the prediction of SES levels. Statistical analysis and less work on machine learning (ML) for prediction classes of SES levels are done. This data set is novel, No attempt is found to design ML models on the Rajamahandravaram SES data set. Further analysis, Section 2 gives other research works descriptions in detail relevant to SES analysis with machine learning algorithms with different data sets with various MLs. Section 3 provides the proposed model and materials that different ML algorithms are analyzed. Section 4 and section 5 provides detailed comparative experimental result analysis and conclusion of the work and future work proposals.

In this, we have to describe descriptions of the researcher's views about socio-economic status and ML. We reviewed reputed journals related to this topic and some of the papers are presented in this section. Socio-economic status is a multidimensional problem that has a variety of definitions. For some authors, it is measured by income, while other researchers include also health, happiness, education, social status, peace, and political rights into the picture. However, what connects all researchers in their work in the field of identification of factors, classification of the population according to different views of Socio-economic status (SES), and prediction of future Socio-economic status levels.

SES is a potential social determining factor of health and education. [(9)] determined the inequalities in income effect related to health status and SES in the USA statewide. For this, they collected information on people's income, self-rated health, education, and some other issues. Statewide income inequalities were computed with the Gini coefficient. They conclude that income distribution inequality was related to health independent of the effect of household income. [(10)] examined SES relation

with an independent contribution of income, occupation, and education to some set of risk factors like smoked, BP, BMI, and cholesterol-related to heart diseases. For this, they choose 2380 people from Stanford and used the forward selection model. They conclude that higher education might be the best SES indicator of good strong health. [11] analyzed the relationship between SES and COPD. For the experiment, they collected 11,042 people's SES, lung function, and demographic data between the age group of 35-95 from 5 countries Argentina, Chile, Bangladesh, Uruguay, and Peru. For the relation between SES and COPD, they used PCA (principal component analysis) and Multivariable alternating (MVAL) logistic regression methods. Overall COPD preponderance was 9.2%, laying out 1.7% to 15.4% across sites. As per their analysis, lower education, lower composite SES index, and lower household income were related to COPD.

Machine Learning (ML) and deep learning related to SES research is vital role in the prediction inequalities, levels of SES, or area wise SES**.** Some of the researchers researched SES levels problems with GPU systems, demographic data, image maps, and so on using ML and DL. [12] estimate the SES of French users of Twitter. They take a horizontal approach to the SES problem and investigate different methods to infer the SES of examples of web-based social media users. They propose various data assortment and a combination of creep able data, expertly annotated data, or open census information for the prediction. [13] The main motive of their study was the prediction of MSW (municipal solid waste) based on demographic and SER variables of 220 municipalities in Ontario, Canada. For this experiment, they used two algorithms that are DTs (Decision Trees) and NN (Neural Networks). As per the results and conclusions, the performance and accuracies of MLs are good, that the ML models can predict the SES level performances. The NN models had the best accurate values with 72% of the variation in the data. The outcomes showed that given adequate socio-economic factors, the ML methods can develop models with high accuracy for waste prediction applications. The SES factors are of key significance during all periods of wildfire management that incorporate restoration, prevention, and suppression. [14] described SES drivers of wildfire occurrence in central Spain. GLM and ML Maxent methods predicted wildfire occurrence during the 1980s and during the 2000s to recognize changes between every period in the SES drivers influencing wildfire occurrence. Creating social and economic frameworks and guaranteeing the effectiveness of social and financial procedures is one of the significant tasks for the government of any nation. Predicting ML models utilized for analyzing enormous data permit effective enterprise management. [15] analyzed on predicting Ukraine's GDP utilizing the ARIMA ML model and use a twofold exponential smoothing model. In the review of literature, very little work has been reported towards the socio-economic system with statistical and ML models for predicting SES levels in the different areas in the whole world. In the review of literature, very little work has been reported towards the socio-economic system with statistical and ML models for predicting SES levels in the different areas in the whole world.

1.1 Contributions/ motivations of the work

The following is the consignments of this work

- In this research, we collected the household information from Rajahmundry, Andhra Pradesh, India using a good questionnaire. The data sampling is using ratios of SES levels like rich, above middle class, middle class, and poor.
- We compose the data set (*.csv) using 49 attributes including class attribute.
- In this paper, we apply 5 reputed ML models like Naïve Bayes, DTs (Tree), k-NN, SVM (kernel RBF), and Random Forest (RF) as well as past and recent SES levels detection research works. As per comparison, the RF model is superior to others.
- This research is very useful in socioeconomic systems for the researchers, analysts and administrative employees and government, and so on.
- This research leads or helps to auto-detection SES level applications like mobile apps.
- We will extend this research with COVID-19 effects on the SES of the Rajahmundry area.

1.2 Organization of the paper

The paper organized as following points

- Section 2 describes the introduction and literature descriptions in detail relevant to SES research work.
- Section 3 outlines the proposed model, along with the structure of the proposed model. The detailed architecture describes the mathematical and algorithmic structure.
- Section 4 presents the details about the experimental setup and analysis of the simulated results. In this, we have been analyzed ML models with Rajahmundry AP SES Dataset as well as compare the results of ML models.
- Section 5 concludes the work with some future directions.

2 Materials and Methods

In this section, we describe the detailed model of the experimental setup and its working process step by step. And it describes experimental materials like metrics and measurement equations. Mainly, it focuses on ML algorithms and their setups and working process, and also describes measuring performance tools like confusion matrix, ROC, and so on.

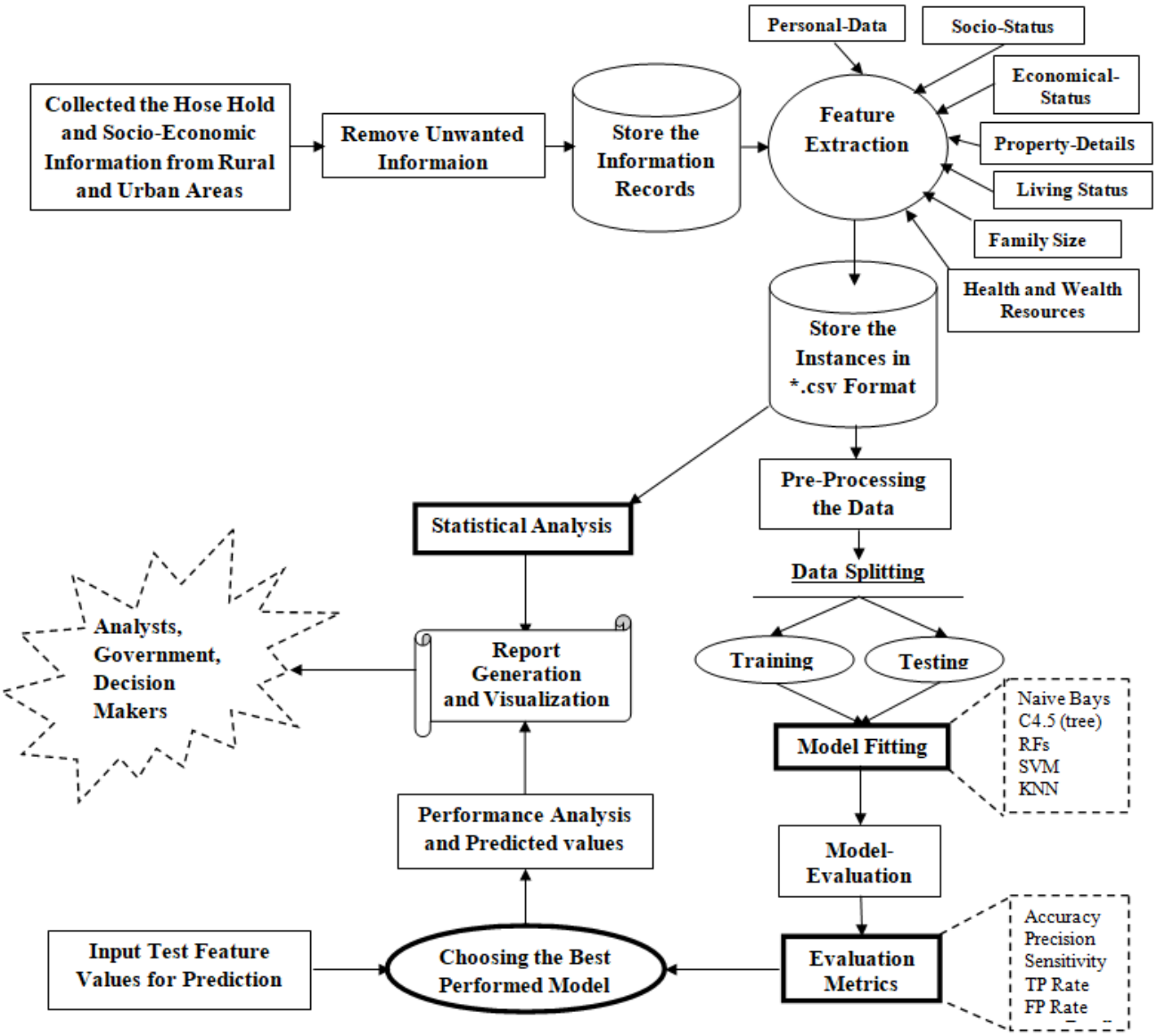

Fig 1. Socio economic level predicting system

2.1 Socio economic level prediction model

The Figure 1 shows the proposed model of poverty predicting system with Household data set. 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. The predicted 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 constructed 1742 records of information with 49 attributes *.csv files and stored into the secondary storage section.

Using this information, construct the statistical analysis reports for analysts and decision-makers to prevent actions about poverty. In another hand, the data is pre-processed by pre-processing algorithms like PCA (principal component analysis) and split the data set into training and testing parts (80% of Train and 20% of Test) for applying Machine Learning algorithms. Mainly, we use popular ML algorithms like Naïve-Bays, Decision Trees, Random Forest Trees, k-NN (k-Nearest Neighborhood), and SVM (Support Vector Machines). After designing the models, we evaluate the models with evaluated metrics like Accuracy (AC), TP Rate, FP Rate, F1, and AUC (using ROC). As per comparison, choose the best-performed ML model for predicting unknown input feature attribute values. Lastly, we will send the performance results, predicting values and visualization graphs to the analysts and decision-makers

2.2 Dataset description

Rajahmundry renamed as Rajamahandravaram is one of the major consistency 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). The Table 1 describes detailed data set 49 attributes include class attribute.

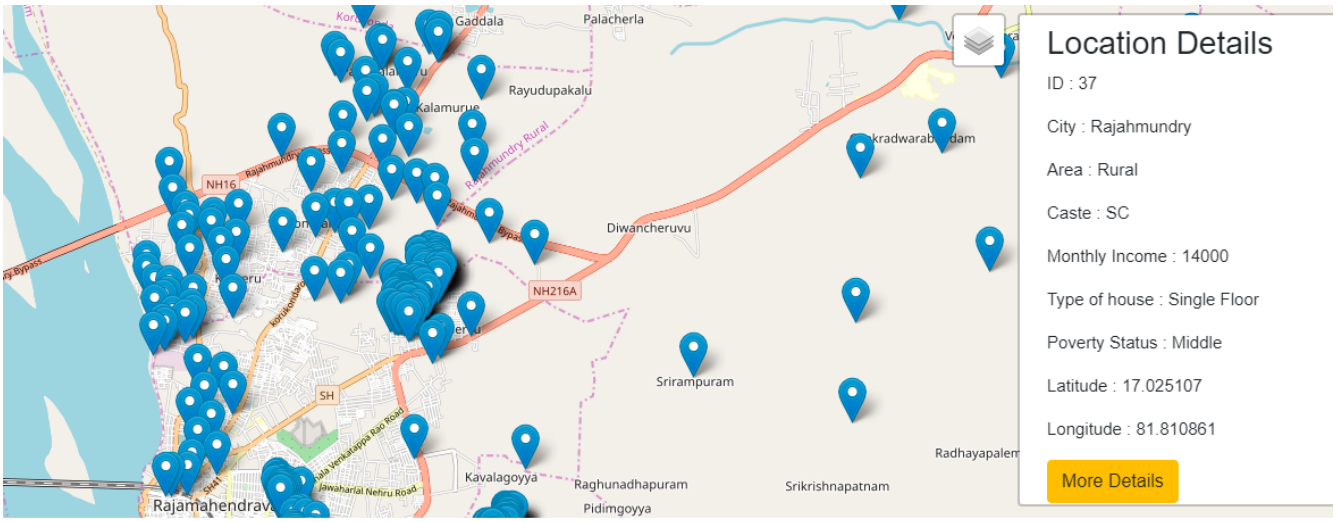

Fig 2. Visualization data points on Rajamahandravaram map

Table 1. Data set attributes description

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 than18 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 0-very low 3-moderate 5- very high

Continued on next page

Table 1 continued

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 or none 4-moderate 7- very high
Major Work	Discrete (Integer)	Work Category 0 to 5 0-very low 3-moerate 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-no 1-yes
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)	No. 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 1480000.0
income from pension	Continues (Real)	Income from Pension of house range is 0 to 40000.0
Income from private	Continues (Real)	Income from private or own of house range is 0 to 15000000.0
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 6.0
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	Discrete (Integer)	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’
electricity1/0	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 Condition1/0	Discrete (Integer)	Air Condition facilities 0 for ‘no’ and 1 for ‘yes’
Heater1/0	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

2.3 Machine learning models

2.3.1 Naïve Bayes (NB) classification

It expects that the presence of an unambiguous aspect of a class is autonomous of every other aspect[16]. As per Bayes theorem, the contingent probability is given by the Equation

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{2}$$

It is the most successful algorithm for many applications such as text document classification, spam filtering, Recommender system, etc.

Working of Naive Bayes Algorithm in SES Problem

NB classifier model for the SES level probabilities:

Step1: Firstly we compute the SES data set class levels prior probabilities.

Step2: Find likelihood with each attribute for each class in SES

Step3: Bayes Formula is computed using feature attributes of SES and computer the posterior probabilities.

Step4: find the superior probability as per input to class which is high probability.

For streamlining posterior and prior probabilities utilize the two tables' probability and frequency tables. Both of these tables will assist us with calculating the probabilities of posterior and prior. All features of SES are in frequency table.

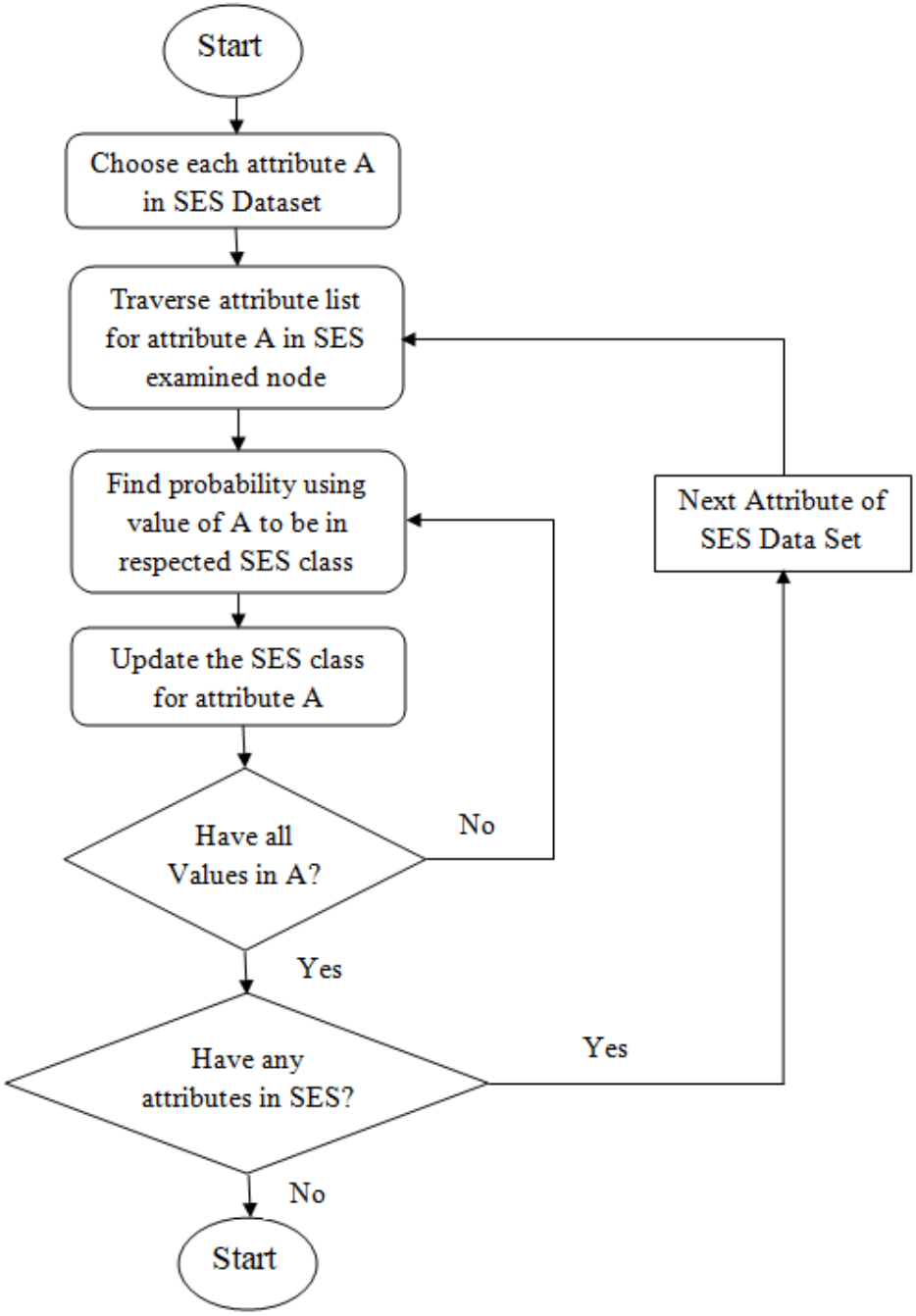

Fig 3. Bayes classification process for SES data set

2.3.2 Support Vector Machine (SVM):

Another incredible supervised ML model is SVM that can be used for both regression and classification issues. The Figure 4 shows the analysis of data using SVM. The numbers of characteristics 'n' are spoken to on the n-dimensional space with each component depicted by the estimation of a specific coordinate. An information component comprising n characteristics is plotted on this n-dimensional space. The point is to find a hyper plane that classifies and increases the edge in an n-dimensional space[17].

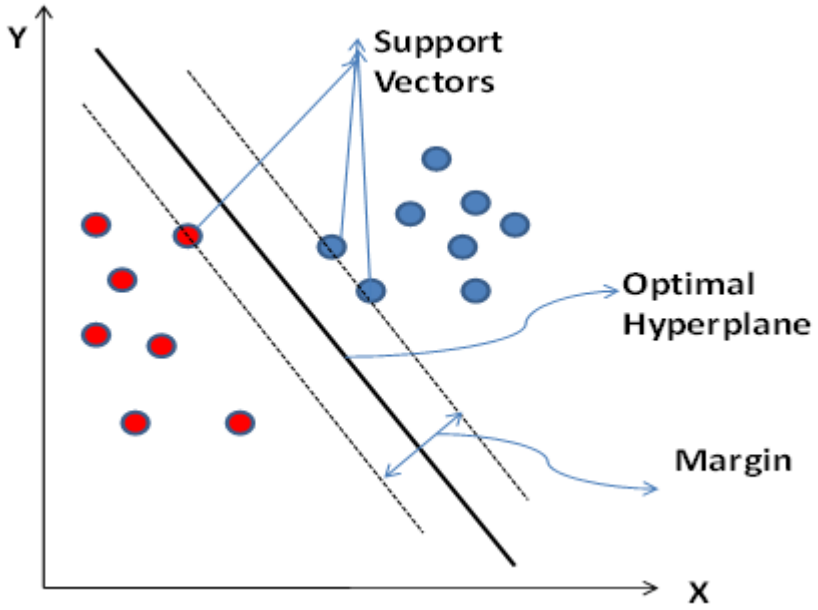

Fig 4. SVM classifier analysis

2.3.3 K-Nearest neighbors' (k-NN) classification

The k-NN is a non-parametric supervised algorithm method suitable for both classification and regression. It considers the k closest data points in the training examples. The output differs based on the fact that KNN is used for classification or regression. The output predicts the class to which a data point belongs based on how closely it matches with the k nearest neighbors. This is one of the instance-based learning, or lazy learning algorithms[18]. This algorithm uses the distance function to calculate the close approximate with the K Nearest Neighbors. For continuous variables, Euclidean, Manhattan, and Minkowski distance measures are used and hamming distance for categorical variables shown in equations (3-5).

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \tag{3}$$

$$\text{Mahattan Distance} = \sum_{i=1}^{k} |x_i - y_i| \tag{4}$$

$$\text{Minkowski Distance} = \left(\sum_{i=1}^{k} (|\, x_i - y_i)^q\right)^{1/q} \tag{5}$$

Working of KNN algorithm for SES dDataset

K-nearest neighbors (KNN) model utilizes ' similarity of features' to estimate the estimations of new information or data which further implies that the new data points will be allotted a value on how tightly matches the data points in the set of training. The Figure 5 shows the classification process for the SES dataset in detail. We can comprehend its working with the assistance of following algorithm

Step 1 – Give the SES data set of training and testing.

Step 2 – Initialize the K value that it can be any number.

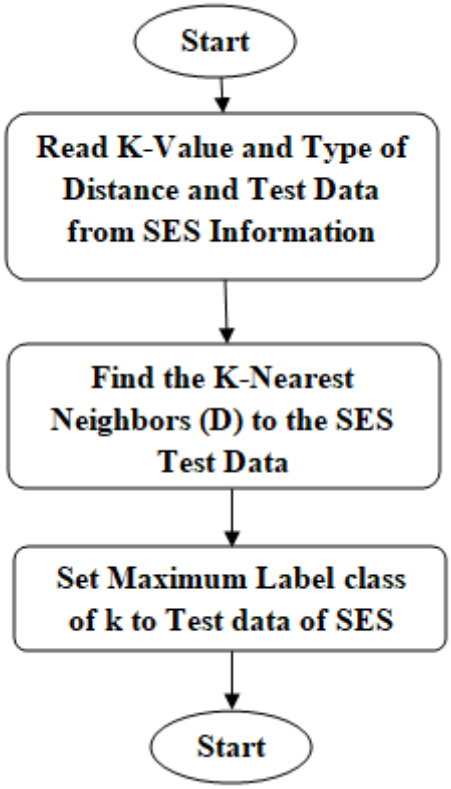

Fig 5. k-NN classification process for SES sata set

Step 3 – For each data point in the test information do the accompanying –

i. Calculate the distance train and test data points using Hamming, Manhattan or Euclidean methods. (Euclidean distance is used in the experimental set up for SES data set)

ii. Sort them in order of ascending.

iii. We will pick the top rows as per value of K from the arranged data set.

iv. Now, it will allot a class to the test point dependent on the most recurrent class of these data rows.

Step 4 – End

2.3.4 Decision Tree Algorithm

DTs model is one of the supervised learning algorithms. In contrast to other supervised ML models, the DTs can be utilized for solving both classification and regression problems, but most researchers used this model for classification issues. It is a tree-organized classifier, where intermediate nodes describe the features of dataset. The decisions rules are designed with and leaf nodes are described with results. DTs classify the data points by arranging them down the tree from the root to some terminal node, with the leaf node giving the order of the model. Every node in the tree goes about as an experiment for some feature attributes, and each edge plunging from the node relates to the potential responses to the experiment. This procedure is recursive in nature and is recurrent for each sub tree rooted at the new node.

Decision Tree algorithm working with SES Data set

In Decision Trees, for anticipating a class name for a record we start from the base of the tree. We look at the estimations of the root characteristic with the record (genuine dataset) property. Based on correlation, we follow the branch relating to that worth and hop to the following hub. For the following hub, the calculation again contrasts the quality worth and the other sub-hubs and move further. It proceeds with the procedure until it arrives at the leaf hub of the tree. The Figure 6 shows Decision Trees Classification process for SES Data Set. The total procedure can be better comprehended utilizing the beneath calculation:

Step-1: Begin the tree with the root hub, says S, which contains the total SES dataset.

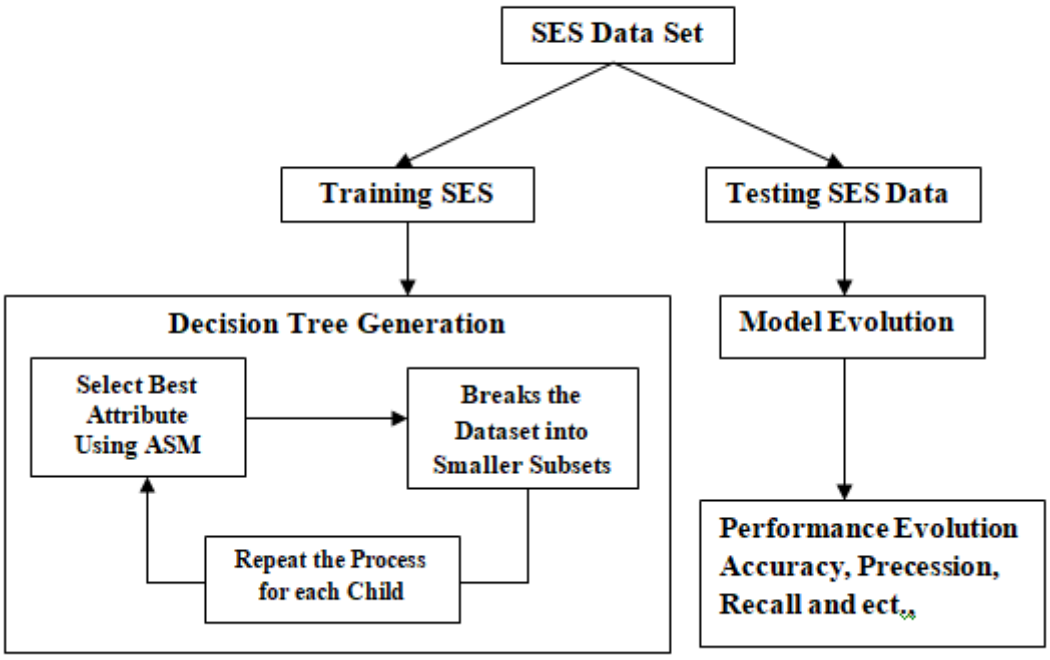

Fig 6. Decision trees classification process for SES data set

Step-2: Find the best characteristic in the SES dataset utilizing Attribute Selection Measure (ASM).

Step-3: Divide the S into subsets that contains potential qualities for the best properties.

Step-4: Generate the choice tree hub, which contains the best trait.

Step-5: Recursively settle on new choice trees utilizing the subsets of the dataset made in step-3. Proceed with this procedure until a phase is arrived at where you can't further arrange the hubs and called the last hub as a leaf hub.

2.3.5 Random Forest (RFs) algorithm

RF is a supervised ML models for classification that is ensemble learning model. The basic reason of this model is that building a little decision-tree with small set of features is a computationally modest procedure. On the off chance that we can construct smaller trees in large number, parallel constructed trees in weak, we would then be able to join the trees to frame a single, averagely strong learner or taking the vote in major. The Figure 7 shows the Random Forest Classification process for SES Data Set. The RF classifier, if numbers of trees are higher in forest then it gives the high performed accurate results.

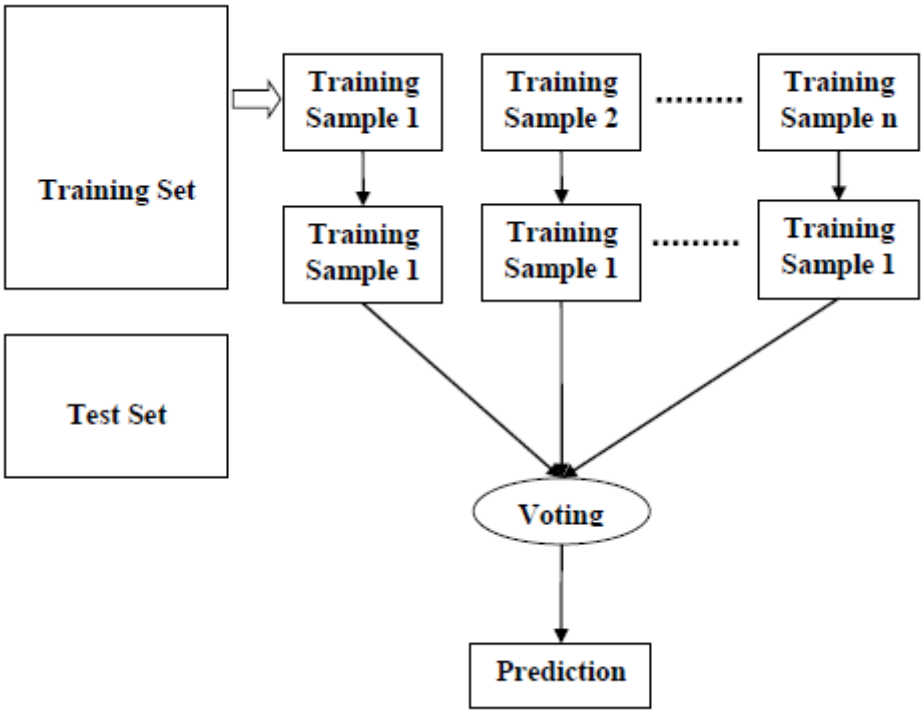

Fig 7. Random Forest Classification process for SES data set

Working of Random Forest Algorithm for SES Data Set
Step 1: Choose randomly n features from the SES total feature Set.
Step 2: As per decision trees, choose best splitting tree for the root node.
Step 3: Predict the result utilizing these trees for decisions.
Step 4: Calculate the target votes using each decision tree predictions.
Step 5: The objective or target with the most prominent vote is considered as the last prediction of the SES Data Set.

2.4 Confusion Matrix

In this, we represent the 4 class problem that is Middle, Poor, Rich, and upper-middle. Table 2 shows the confusion matrix for the Socio-Economic Status with 4 class problems. The accuracy is calculated by the diagonal of the confusion matrix. The confusion matrix is constructed using actual or true values and Predicted values [19].

Table 2. Confusion Matrix for 4 class problem

Classifier		Actual or True Values				
	Class	Middle	Poor	Rich	U-middle	∑(Total)
Predicted Values	Middle(M)	M-M	M-P	M-R	M-U	T5
	Poor(P)	P-M	P-P	P-R	P-U	T6
	Rich(R)	R-M	R-P	R-R	R-U	T7
	U-middle(U)	U-M	U-P	U-R	U-U	T8
	∑(Total)	T1	T2	T3	T4	Total(T)

2.5 Performance parameters

Performance parameters results give the performance of data set [20]. We calculated the performance parameters like TPR-True Positive Rate-Recall-Sensitivity, Probability of Detection, Power, FNR-False Negative Rate, Miss Rate, FPR-False Positive Rate,

Fall Out, Probability of False Alarm, SPC-Specificity, Selectivity, True Negative Rate (TNR), PPV-Positive Predictive Value, Precision, FOR-False Omission Rate, LR+-Positive Likelihood Ratio, LR—Negative Likelihood Ratio, ACC-Accuracy, FDR-False Discovery Rate, NPV-Negative Predictive Value, DOR-Diagnostic Odds Ratio, F1Score 6 to 17 respectively.

$$TPR = \frac{\sum \text{True Positive}}{\sum \text{Condition Positive}} \tag{6}$$

$$FNR = \frac{\sum \text{False Nagative}}{\sum \text{condition Positive}} \tag{7}$$

$$FPR = \frac{\sum \text{False Positive}}{\sum \text{condition Negative}} \tag{8}$$

$$SPC \text{ or } TNR = \frac{\sum \text{True Nagative}}{\sum \text{Condition Negative}} \tag{9}$$

$$\text{Pr evalence} = \frac{\sum \text{Condition Positive}}{\sum \text{Total Population}} \tag{10}$$

$$PPV \; or \; PRC = \frac{\sum \text{True Positive}}{\sum \text{Predicted Condition Positive}} \tag{11}$$

$$FOR = \frac{\sum \text{False Negative}}{\sum \text{Predicted Condition Negative}} \tag{12}$$

$$\text{Accuracy}(ACC) = \frac{\sum \text{True Positive} + \sum \text{True Negative}}{\sum \text{Total Population}} \tag{13}$$

$$FDR = \frac{\sum \text{False Positive}}{\sum \text{Predicted Condition Positive}} \tag{14}$$

$$NPV = \frac{\sum T\text{rue Negative}}{\sum \text{Predicted Condition Negative}} \tag{15}$$

$$DO\,R = \frac{LR+}{LR-} \tag{16}$$

$$F_1 \text{ score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

3 Results and Discussion

In this, we have to analyze the statistical analysis results and machine learning models classification accuracies in detail.

3.1 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 social and economical 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 on average 4 to 5 members 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 max values are 8 and mean and SD values are 1.975 and 0.776 respectively which means every house contains on 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 handicapped people are very less percentage that the mean values are 0.066 and 0.024 respectively.

Table 3. Detailed statistical analysis of Rajahmundry SES data set

Attributes Min, Max, Mean and Standard Deviation Statistics					Yes or No Attributes Statistics		
Attribute	Min	Max	Mean	SD	Attribute	No	Yes
Family Size	1	16	4.381	1.467	Health cards	689	1053
Male	1	8	2.406	0.773	Own House	466	1276
Female	0	8	1.975	0.776	Toilets facilities	168	1574
below 18	0	5	1.592	0.882	electricity	0	1742
above 18	0	12	2.802	1.183	TV	442	1300
married people	0	8	2.037	0.375	Fridge	1134	608
No. of children	0	2	0.131	0.362	Air Condition	1481	261
No. of literates	0	12	0.738	0.826	Heater	1523	219
No. of Workers	0	8	1.693	0.693	Computer	1421	321
Child. work	0	1	0.006	0.079	Other Type of Attributes Statistics		
No. of Diseased people	0	2	0.066	0.268	Attribute	Type	Value
No. of Handicapped	0	1	0.024	0.152	Ration Cards	white	660
land(cents)	0	2000	140.576	202.602		Pink	1082
Gold(grams)	0	1500	34.788	51.792	Fuel for cooking	Gas	1575
Annual Income	27000	15000000	331504.6	467944.2		other	167
Income from Govt.	0	1480000	24055.68	117628.8	Social status	ST	31
income from pension	0	40000	408.726	1351.704		SC	190
Income from private	0	15000000	308868	467484.8		BC	896
Hospital in Km	1	6	3.637	0.836		OC	625
Primary School in Km	1	5	2.846	1.901	Addicted persons to smoke and drinking in House	None	736
High School in Km	1	10	3.832	1.899		Partial	826
College in Km	2	14	6.866	2.057		Addicted	158
University in Km	30	45	35.065	4.185		Extreme	22

Some other Types of Attribute Statistics					
Attribute	Type	Value	Attribute	Type	Value
Literacy and Educators Houses	None or Below 10th	296	Occupation Major Work Occupation in House	No Work or Very less	6
	10th Standard	428		Seasonal Workers	463
	Inter Level or ITI	386		Average or Daily wagers	497

Continued on next page

Table 4 continued

	Degree Level	272		Permanent Low salary	345
	Technical Degree or Other	249		Permanent Middle Salary	364
	P.G. level	101		Permanent High Salary	62
	Professional or Ph.D. Level	10		Business or Organizers	5
Having Bikes in House	None	939	Having Cars in House	None	1607
	One	671		One	128
	Two	118		Two	5
	More Than Two	14		More Than Two	2
Having Other Traveling Recourses	None	1615	Target Class (SES levels)	Rich	73
	One	121		Middle class	794
	Two	5		Upper Middle class	526
	More Than Two	1		Poor	349

Very important thing for the economic status that it is fully depends on annual income for each house and their resources that are from public, private, asserts and work, and so on. As per statistics annual income min value is 27000/- and the max value is 8000000/-. The income sources from private, government or pension schemes. The detailed analysis is shown in the Table 3. The educational and health resources are also available within the distance of every house.

3.2 Experimental setup

In this section, we analyze accuracy values of ML algorithms k-NN, DTs, SVM, RF and NB in detailed. For this, we used confusion matrix for each algorithm.

3.2.1 K-Nearest neighbor

The k-NN model classifies correctly 1643 instances out of 1742. The remaining 99 instances are classified incorrectly by this model. The total accuracy (CA) value is 0.94316 (94.4%). The F1-score is 0.94312 and the precision value is 0.94341. The time taken for the construction of the model is 0.29 seconds. The Figure 8 shows the confusion matrix of the KNN model. In this, we used that k-value is 5, and the distance calculation method is Euclidean. As per the analysis, the prediction class "poor" is very accurate (0.975 or 97.5%) than other classes where only 9 instances are incorrectly classified out of 353 poor class instances. In the next positions two and three occupied by upper-middle predicted class with 94.4% accuracy and 93.7% accuracy of middle-class relatively. In the rich-class predictor, 62 instances are classified correctly and 10 instances are going to upper-middle-class premises, so the accuracy is 86.2% only

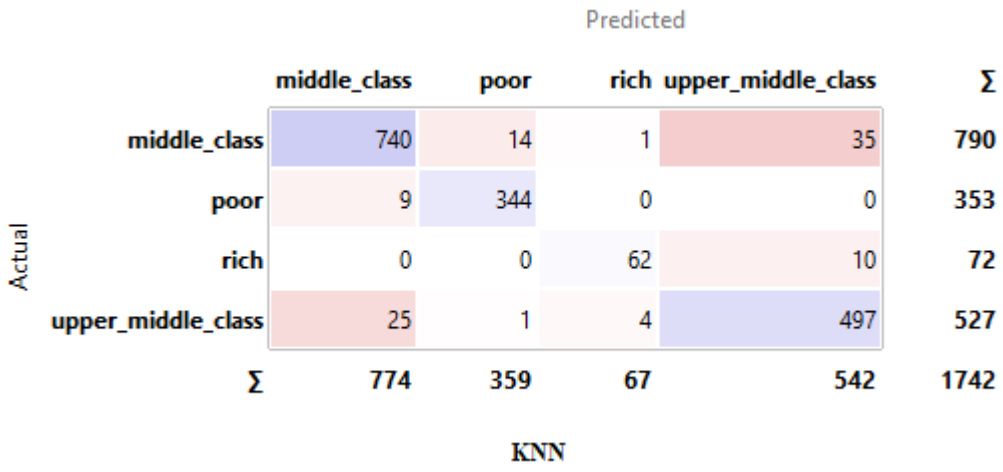

Fig 8. Confusion matrix of the KNN model

3.2.2 Decision Tree (DTs):

The C 4.5 model classifies correctly 1684 instances out of 1742. The remaining 58 instances are classified incorrectly by this model. The total accuracy (CA) value is 0.9667 (96.67%). The F1-score is 0.96659 and the precision value is 0.96672. The time taken for the construction of the model is 0.18 seconds. The Figure 9 shows the confusion matrix of the DTs model. As per the analysis, the prediction class "Middle-class" is slightly more accurate (0.97468 or 97.47%) than the target class "poor" where the poor class classifies 344 instances correctly out of 353 instances(accuracy value is 0.974504(97.45%)). The target "rich" class accuracy is 0.875(87.5%) and the target class "upper-middle" accuracy is 0.962(96.2%). In the upper-middle, out of 527 instances 507 instances are classified correctly and remain 19 instances are classified as "middle-class" and one as in rich incorrectly. The DTs algorithm is somewhat good that it classifies three target classes (poor, middle, and upper-middle) out of four target classes with above 96% of accuracy.

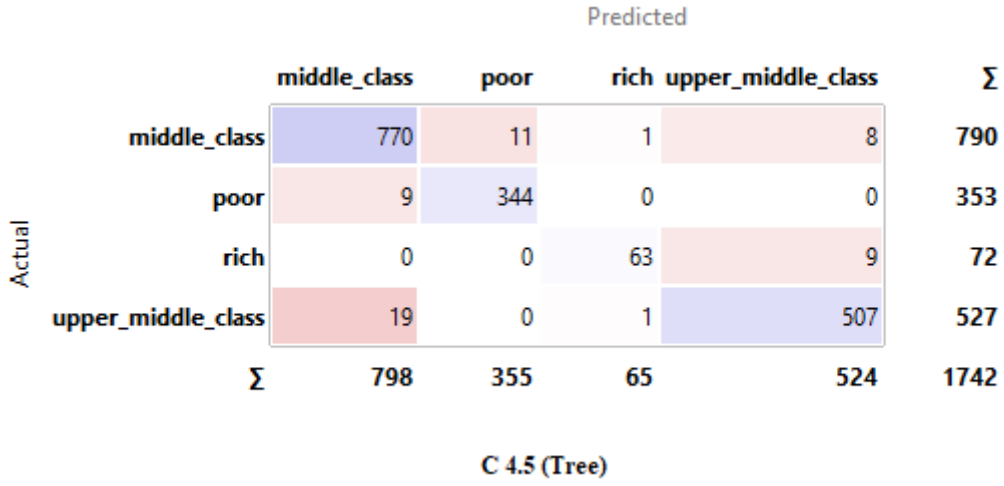

Fig 9. Confusion matrix of the DTs model

3.2.3 SVM

The Support Vector Machine with kernel radial bios function (RBF) model classifies correctly 1647 instances out of 1742. The remaining 95 instances are classified incorrectly by this model. The total accuracy (CA) value is 0.94546 (94.6%). The F1-score is 0.94545 and the precision value is 0.94822. The time taken for the construction of the model is 0.16 seconds. The Figure 10 shows the confusion matrix of SVM with the kernel RBF model. In this model, we used the kernel RBF that expression is exp(-g|x-y|^2) where g =0.1, numerical tolerance is 0.001, Cost(C) value is 1.0 and regression loss epsilon is 0.1 and number of iteration limit is 100. As per the analysis, the target class "poor" is very accurate (0.9887 or 98.87%) than other classes where only 4 instances are incorrectly classified (3 instances in "middle" and 1 in upper-middle) out of 353 instances. In the next positions two and three occupied by upper-middle predicted class with 97.91% accuracy and 90.88% accuracy of middle-class relatively. In the rich-class predictor, 64 instances are classified correctly and 8 instances are classified incorrectly (7 in upper-middle and 1 in middle). So, the accuracy is 88.88% only.

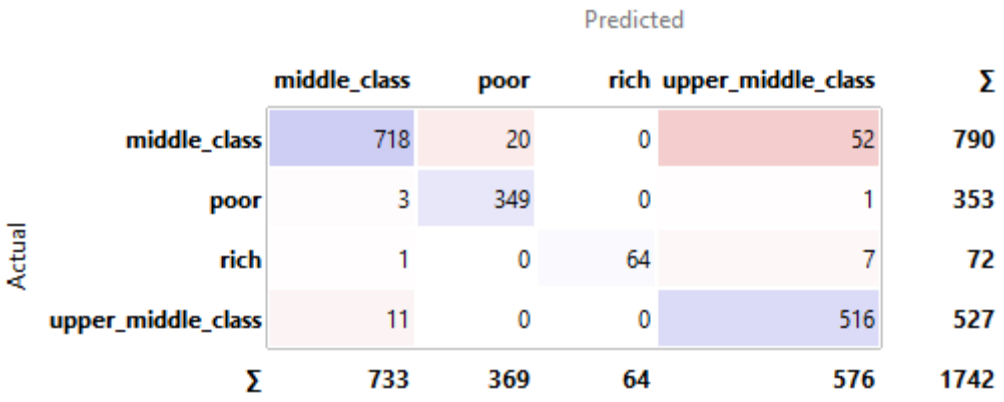

Fig 10. Confusion matrix of the SVM model

3.2.4 Random forest

The RF's model classifies correctly 1704 instances out of 1742. The remaining 38 instances are classified incorrectly by this model. The total accuracy (CA) value is 0.97818 (97.81%). The F1-score is 0.9782 and the precision value is 0.9781. The time taken for the construction of the model is 0.41seconds. The Figure 11 shows the confusion matrix of the RF model. As per the analysis, the prediction class "poor" is slightly more accurate (0.9887 or 98.87%) than other target class "middle", "upper-middle" and "rich" classes. The rich target class classifies 71 instances correctly out of 72 instances (accuracy value is 0.9861(98.61%)). The target "upper middle" class accuracy is 0.98102(98.1%) and the target class "middle" accuracy is 0.9708(97.08%). In middle, out of 790 instances 767 instances are classified correctly and remain 23 instances are classified 11 as "poor" and 12 as in "upper-middle" incorrectly. The RF algorithm is somewhat good that it classifies three target classes (poor, rich, and upper-middle) out of four target classes with above 98% of accuracy.

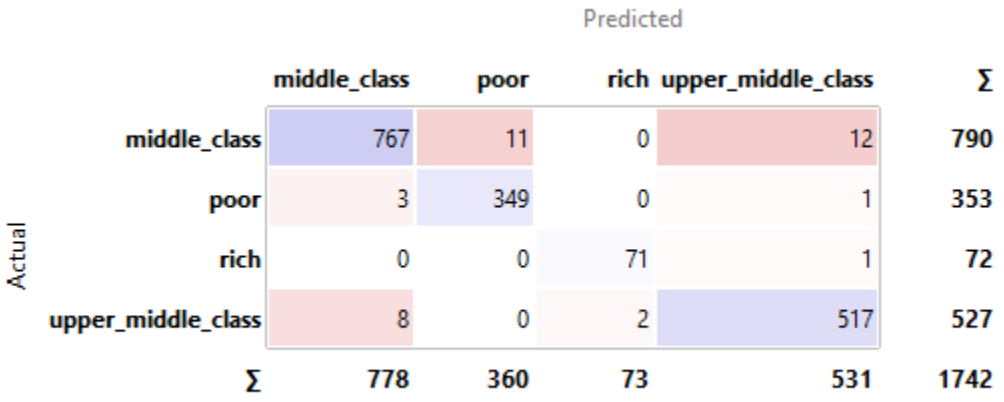

Fig 11. Confusion matrix of the RF model

3.2.5 Naïve Bayes (NB)

The Naïve Bayes model classifies correctly 1514 instances out of 1742. The remaining 228 instances are classified incorrectly by this model. The total accuracy (CA) value is 0.86912 (86.91%). The F1-score is 0.87333 and the precision value is 0.89096. The time taken for the construction of the model is 0.12 seconds. The Figure 12 shows the confusion matrix of the Naïve Bayes model. As per the analysis, the prediction class "poor" is more accurate (0.991501 or 99.15%) than other target class "middle", "upper-middle" and "rich" classes. The rich target class classifies 67 instances correctly out of 72 instances (accuracy value is 0.93055 (93.05%)). The target "upper middle" class accuracy is 0.853889943(85.38%) and the target class "middle" accuracy is 0.818987342 (81.9%). In the middle, out of 790 instances 647 instances are classified correctly and remain 163 instances are classified 45 as "poor" and 98 as in "upper-middle" incorrectly. The NB algorithm is not so good that it classifies three target classes (middle, rich and upper-middle) out of four target classes with below or equal 93% of accuracy compared to other used ML algorithms.

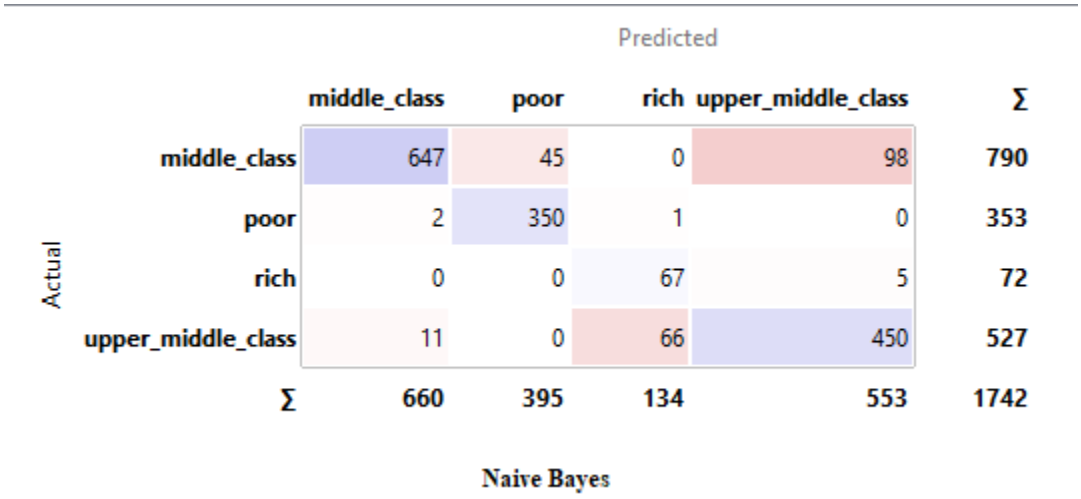

Fig 12. confusion matrix of the Naïve Bayes model

3.3 Receiver Operating Characteristic (ROC) curves

The ROC curve constructed with specificity (FP Rate) and Sensitivity (TP Rate) measures with 0 to 1 values. The Figure 13 shows the targeted class "middle" by utilizing the ROC curves with experimental models. In this analysis, the experimental models k-NN, DTs (Tree), SVM, Random Forest, and Naïve Bayes on targeted class middle AUC values are 0.9937, 0.9884, 0.9862, 0.9989 and 0.9794 respectively. Each model ROC curves represents each color. All models AUC values are above 0.97, so all the models are efficient and effective for predicting the middle-class unknown values. In this, the RF model is the more efficient performer to predict target class "middle" than other models. The light green specifies the RF performed ROC curve shown in the Figure 13.

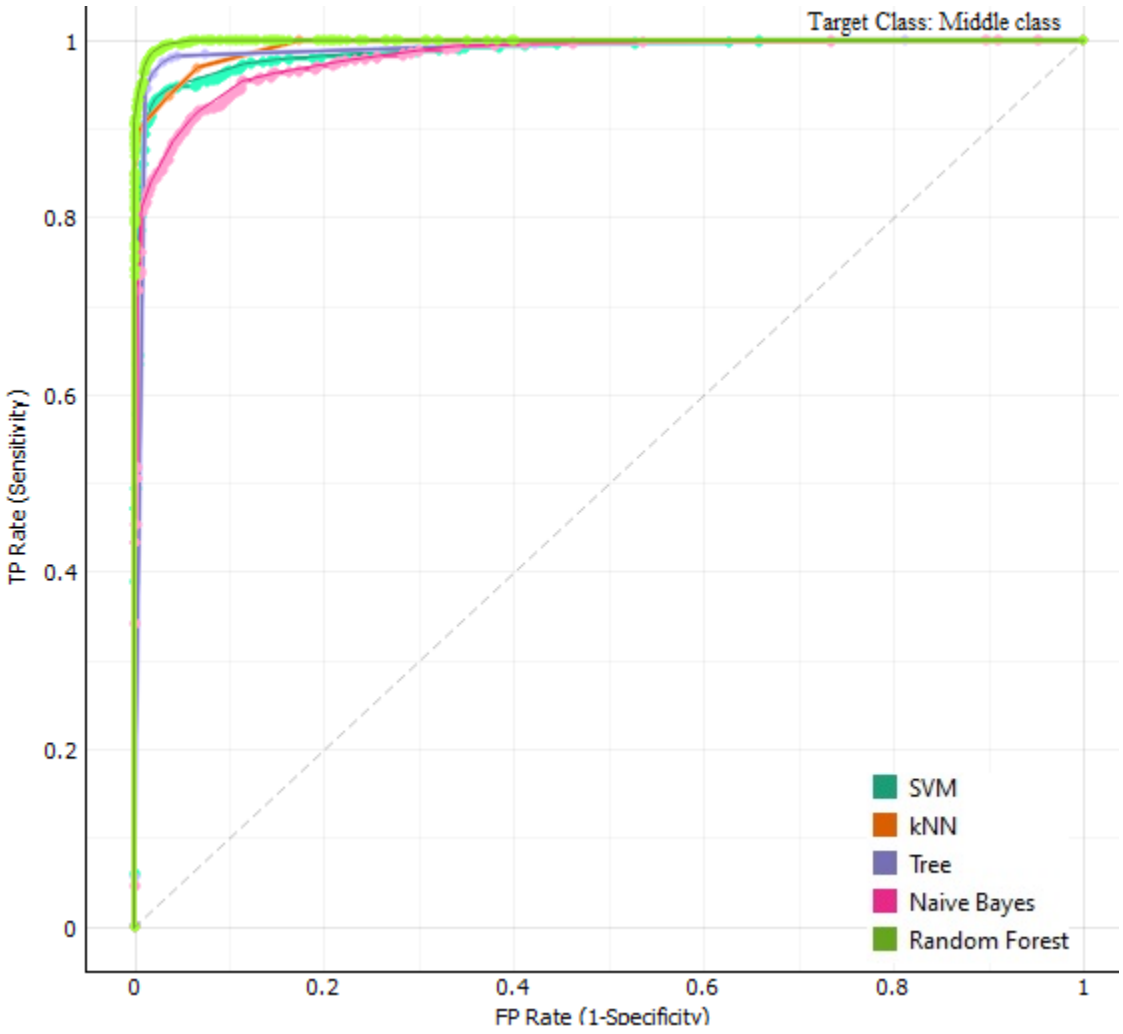

Fig 13. ROC Curves of 5 ML Models on target class "Middle-class"

The Figure 14 shows the targeted class "poor" by utilizing the ROC curves with experimental models. In this analysis, the experimental models k-NN, DTs (Tree), SVM, Random Forest, and Naïve Bayes on targeted class middle AUC values are 0.9992, 0.9927, 0.9992, 0.9998 and 0.9952 respectively. Each model ROC curves represents each color. All models AUC values are above 0.99, so all the models are efficient and effective for predicting the target class "poor" with unknown values. In this, the RF model is the more efficient performer to predict target class "poor" than other experimental models. The light green specifies the RF performed ROC curve shown in the Figure 14.

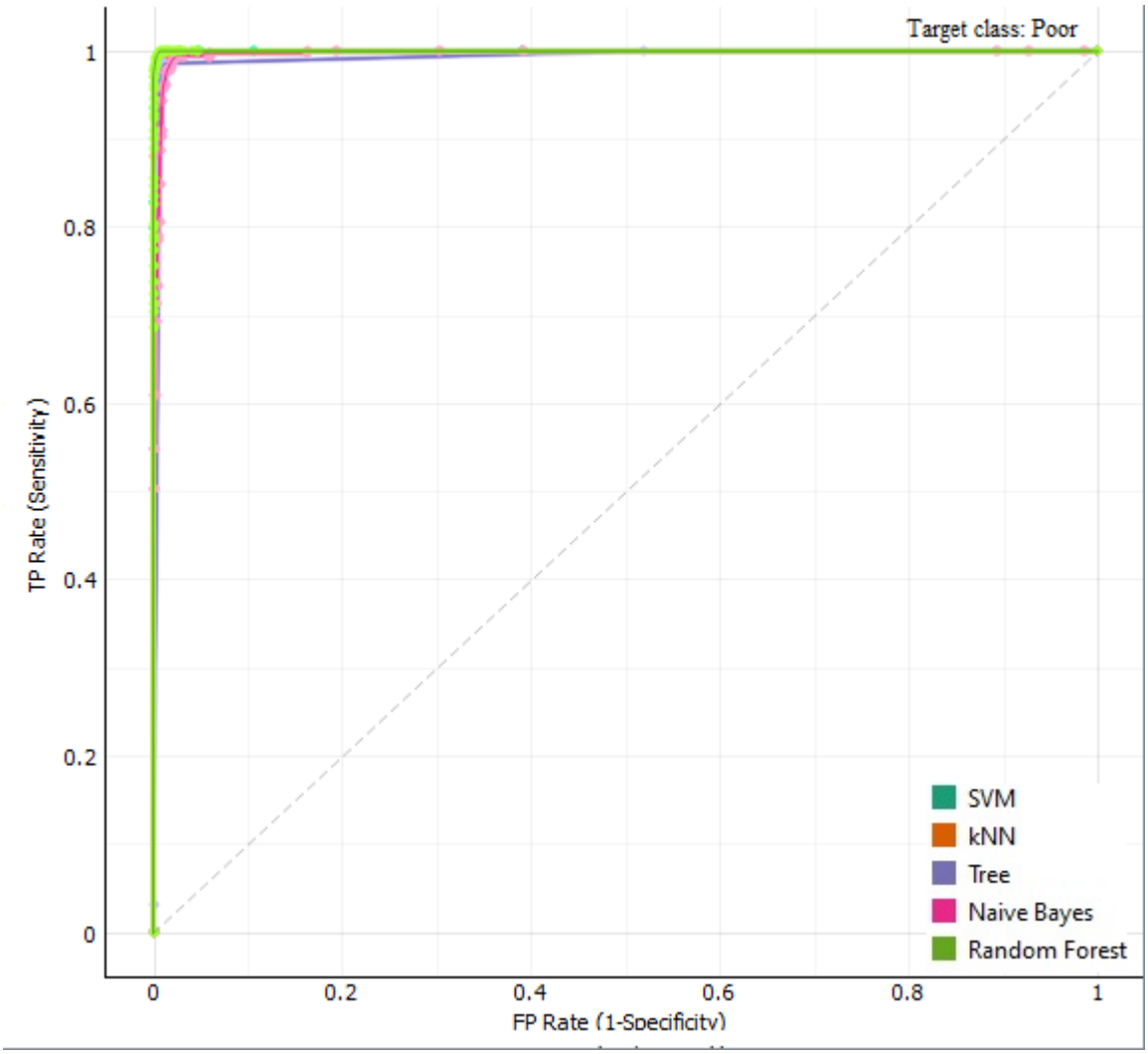

Fig 14. ROC Curves of 5 ML Models on Target class "Poor"

The Figure 15 shows the targeted class "rich" by utilizing the ROC curves with experimental models. In this analysis, the experimental models k-NN, DTs (Tree), SVM, Random Forest, and Naïve Bayes on targeted class middle AUC values are 0.9985, 0.9874, 0.99903, 0.99993 and 0.98745 respectively. Each model ROC curves represents each color. All models AUC values are above 0.98, so all the models are efficient and effective for predicting the target class "rich" with unknown values. In this, the RF model is the more efficient performer to predict target class "poor" than other experimental models. The light green specifies the RF performed ROC curve shown in the Figure 15.

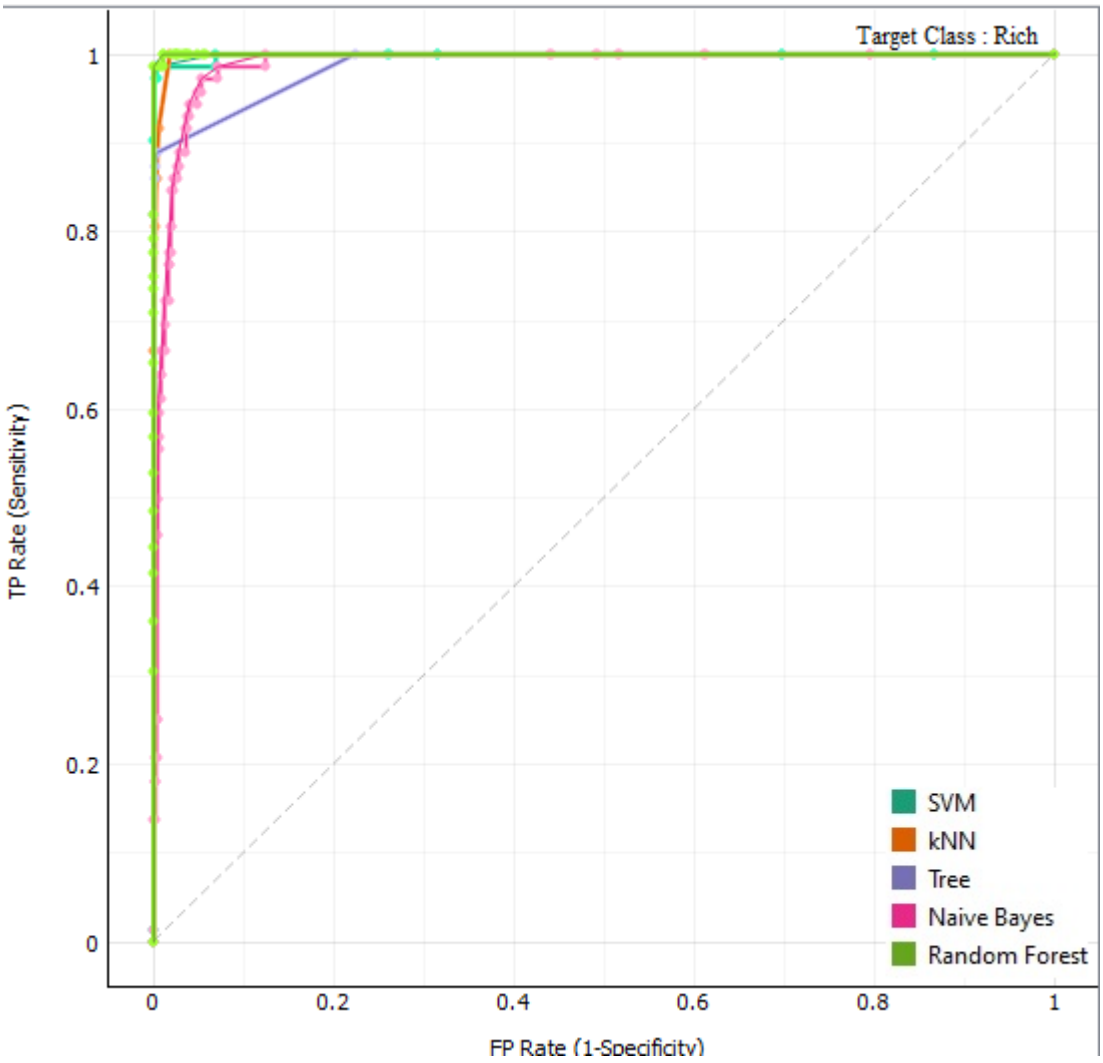

Fig 15. ROC Curves of 5 ML Models on Target class "Rich"

The Figure 16 shows the targeted class "rich" by utilizing the ROC curves with experimental models. In this analysis, the experimental models k-NN, DTs (Tree), SVM, Random Forest, and Naïve Bayes on targeted class middle AUC values are 0.9985, 0.9874, 0.99903, 0.99993 and 0.98745 respectively. Each model ROC curves represents each color. All models AUC values are above 0.98, so all the models are efficient and effective for predicting the target class "rich" with unknown values. In this, the RF model is the more efficient performer to predict target class "poor" than other experimental models. The light green specifies the RF performed ROC curve shown in the Figure 16.

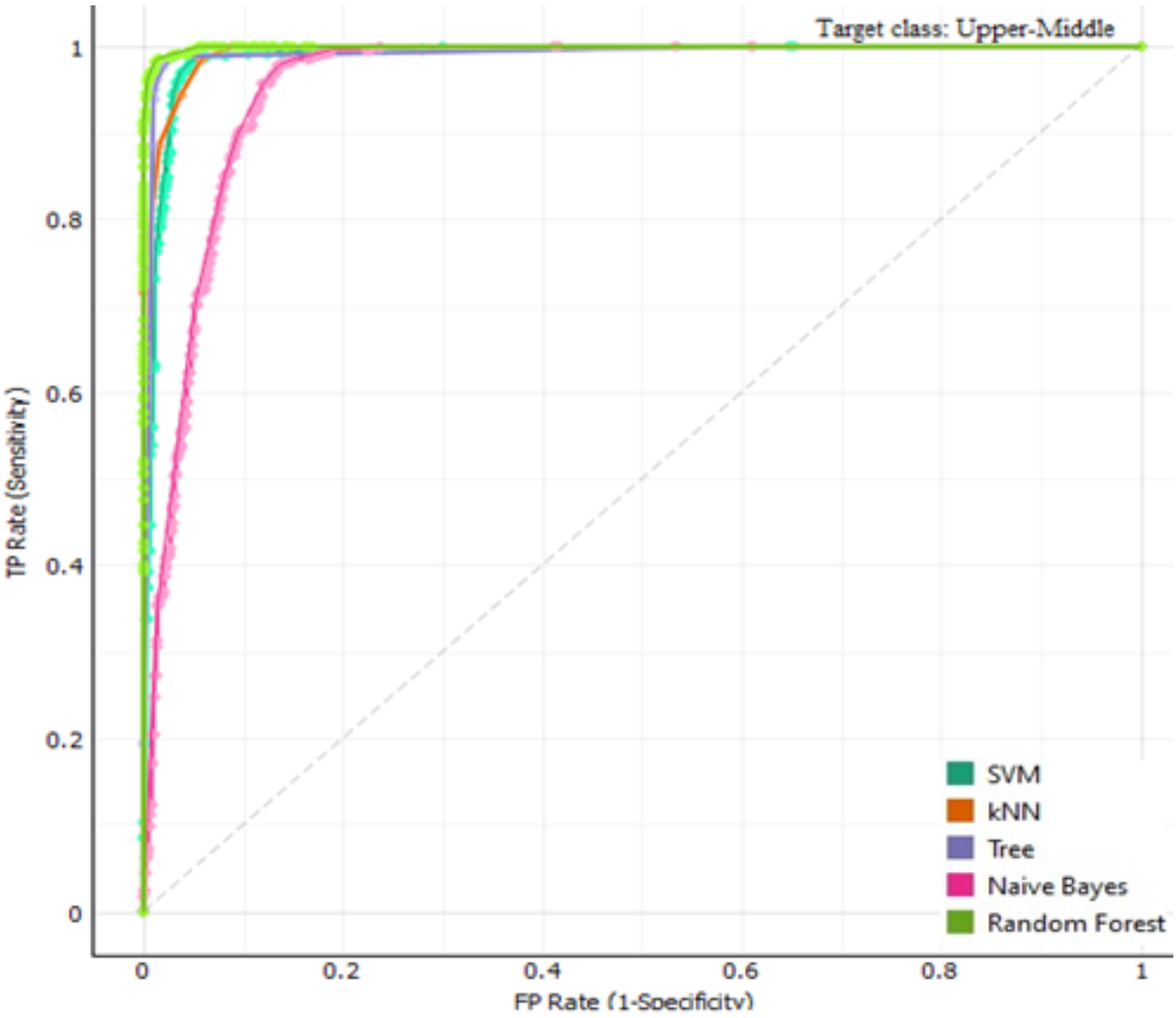

Fig 16. ROC Curves of 5 ML Models on Target class "Upper-Middle"

3.4 Experimental ML algorithms comparative analysis

Table 5. Accuracy parameter values of Rajahmundry SES data set

Model	AUC	CA	F1	Precision	Recall	Time Taken
k-NN	0.9954	0.94316	0.94312	0.94341	0.94316	0.29 sec.
DTs (Tree)	0.9905	0.96670	0.96659	0.96672	0.96671	0.19 sec.
SVM	0.9907	0.94546	0.94545	0.94822	0.94546	0.16 sec.
Random Forest	0.9992	0.97818	0.9782	0.9781	0.97818	0.41 sec.
Naive Bayes	0.9765	0.86911	0.87333	0.89096	0.86911	0.12 sec.

The Table 5 shows all performance parameter values of each ML model and the time taken for building a relative model. In this analysis, one of the main observations that the RF algorithm is the best model than other experimental ML models where all the parameters of performance that CA, AUC, and F1 values are greater than others. But, it takes a lot of time those

0.41 seconds for building the model. This is the highest time than comparative other experimental models. The second high performed model is the DTs model and it takes 0.16 seconds only for model building. Naïve Bayes takes the lowest time taken for building the model but it occupies the last position that the second-lowest time taken for model building.

This analysis analyzed using bar-chart diagram in detail. The Figure 17 shows the comparative analysis of the ML models. As per analysis, the classification accuracy (CA) represents in pink color and AUC represent in blue color. High values of CA (Classification Accuracy) and AUC values (0.976 and 0.999) are indicated by the random forest. The second highest AUC value is (0.9954) indicated by the KNN model as well as the second highest CA value (0.966) is indicated by Tree (DTs) model.

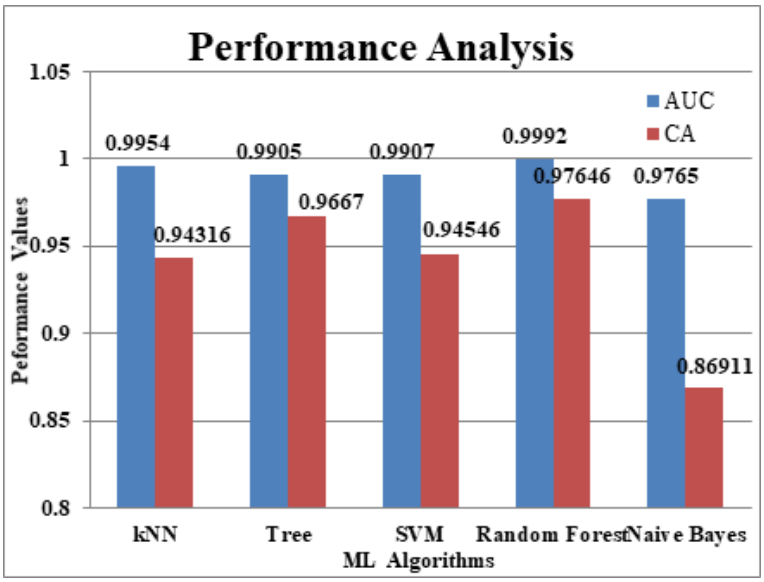

Fig 17. Comparative Analysis AUC and CA of ML Models

The Figure 18 shows the time taken for built the ML models. In this, the Naïve Bayes model takes 0.12 seconds, but the CA value is 0869 that it is least performed model comparative other. So, it is considered only time based for the predictions. In other hand, the Random forest model takes the highest time (0.41 sec.) for built the model and accuracy point of view it is in fist position. So, the RF is very better model to predict the SES levels based on accuracy only. The DTs is moderate model and somewhat good where the model is built within 0.19 seconds third position and the accuracy is 0.966 in second position. As per analysis, accuracy and time based DTs is the better than all experimental ML models.

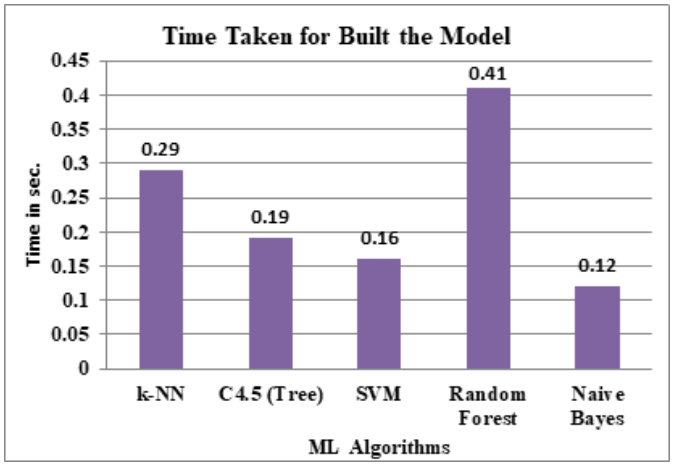

Fig 18. Comparative model building time analysis of ML models

4 Conclusion

Analysis and prediction of socio-economic status research work are very useful for analysts, organizations, and government. Good sampling statistical results represented economic features, social standards, and SES levels of the Rajamahandravaram consistency area. This useful work had described SES with machine learning representation. As per the comparative study, the Random Forest ML model was the best for predicting SES levels of Rajahmundry SES data set where accuracy (CA) value is 0.976, and the AUC value is 0.999. Further, we will take and elaborate overall East Godavari district samples and working with GPS data using Deep Learning concepts for more accurate performance values. And also conduct the research work on before the COVID-19 and after the COVID-19 for this area.

Acknowledgement

We would like to thank Prof. M. Jagannadha Rao Vice Chancellor Adikavi Nannaya University, AP, India who supported this research, and we would like to thank R & D Cell Adikavi Nannaya University for providing necessary materials and technical supports obtained for this research. We would like to thank the people who gave the house hold information by themself for this research.

References

1) Qureshi MI, Qayyum S, Nassani AA, Aldakhil AM, Abro MMQ, Zaman K. Management of various socio-economic factors under the United Nations sustainable development agenda. *Resources Policy*. 2019;64. Available from: https://dx.doi.org/10.1016/j.resourpol.2019.101515.
2) Dou Y, da Silva R, McCord P, Zaehringer J, Yang H, Furumo P, et al. Understanding How Smallholders Integrated into Pericoupled and Telecoupled Systems. *Sustainability*. 2020;12(4):1596–1596. Available from: https://dx.doi.org/10.3390/su12041596.
3) Eberstadt N, Verdery AM, Zeng Y, Wang Z, Wang F, Shen K, et al. China's Changing Family Structure: Dimension and implications. AEI Paper & Studies, 79. Corpus ID: 219939979 . 2019. Available from: https://www.aei.org/research-products/report/chinas-changing-family-structure-dimensions-and-implications/.
4) Kumar A, Sharma A. Socio-Sentic framework for sustainable agricultural governance. *Sustainable Computing: Informatics and Systems*. 2018. Available from: https://doi.org/10.1016/j.suscom.2018.08.006.
5) Shah R, Zimmermann R. Multimodal analysis of user-generated multimedia content. and others, editor;Springer International Publishing. 2017. Available from: https://doi.org/10.1007/978-3-319-61807-4.
6) Goodwin Y, Strang KD. Socio-Cultural and Multi-Disciplinary Perceptions of Risk. *International Journal of Risk and Contingency Management*. 2012;1(1):1–11. Available from: https://dx.doi.org/10.4018/ijrcm.2012010101.
7) Dahdouh-Guebas F, Collin S, Seen DL, Rönnbäck P, Depommier D, Ravishankar T, et al. Analysing ethnobotanical and fishery-related importance of mangroves of the East-Godavari Delta (Andhra Pradesh, India) for conservation and management purposes. *Journal of Ethnobiology and Ethnomedicine*. 2006;2(1). Available from: https://dx.doi.org/10.1186/1746-4269-2-24.
8) Ranjith S, Shivapur AV, Kumar PSK, Hiremath CG, Dhungana S. Water quality evaluation in term of WQI river Tungabhadra. *International Journal of Innovative Technology and Exploring Engineering*. 2019;(8):247–253. Available from: https://doi.org/10.35940/ijitee.I1051.0789S219.
9) Kennedy BP, Kawachi I, Glass R, Prothrow-Stith D. Income distribution, socioeconomic status, and self rated health in the United States: multilevel analysis. *BMJ*. 1998;317(7163):917–921. Available from: https://dx.doi.org/10.1136/bmj.317.7163.917.
10) Winkleby MA, Jatulis DE, Frank E, Fortmann SP. Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. *American Journal of Public Health*. 1992;82(6):816–820. Available from: https://dx.doi.org/10.2105/ajph.82.6.816.
11) Grigsby M, Siddharthan T, Chowdhury M, Siddiquee A, Rubinstein A, Sobrino E, et al. Socioeconomic status and COPD among low- and middle-income countries. *International Journal of Chronic Obstructive Pulmonary Disease*. 2016;11:2497–2507. Available from: https://dx.doi.org/10.2147/copd.s111145.
12) Abitbol JL, Karsai M, Fleury E. Location, occupation, and semantics based socioeconomic status inference on twitter. In: and others, editor. IEEE International Conference on Data Mining Workshops (ICDMW). IEEE. 2018;p. 1192–1199. Available from: https://doi.org/10.1109/ICDMW.2018.00171.
13) Kannangara M, Dua R, Ahmadi L, Bensebaa F. Modeling and prediction of regional municipal solid waste generation and diversion in Canada using machine learning approaches. *Waste Management*. 2018;74:3–15. Available from: https://dx.doi.org/10.1016/j.wasman.2017.11.057.
14) Vilar L, Gómez I, Martínez-Vega J, Echavarría P, Riaño D, Martín MP. Multitemporal Modelling of Socio-Economic Wildfire Drivers in Central Spain between the 1980s and the 2000s: Comparing Generalized Linear Models to Machine Learning Algorithms. *PLOS ONE*. 2016;11(8). Available from: https://dx.doi.org/10.1371/journal.pone.0161344.
15) Dehtiarova YV, Yevdokimov YURI. Data Mining Methods and Models for Social and Economic Processes Forecasting. 2018. Available from: https://doi.org/10.21272/mer.2018.80.03.
16) Saritas MM. Performance analysis of ANN and Naive Bayes classification algorithm for data classification. *International Journal of Intelligent Systems and Applications in Engineering*. 2019;7(2):88–91. Available from: https://dx.doi.org/10.18201/ijisae.2019252786.
17) Karimi F, Sultana S, Babakan AS, Suthaharan S. An enhanced support vector machine model for urban expansion prediction. *Computers, Environment and Urban Systems*. 2019;75:61–75. Available from: https://dx.doi.org/10.1016/j.compenvurbsys.2019.01.001.
18) Liu ZG, Zhang Z, Liu Y, Dezert J, Pan Q. A new pattern classification improvement method with local quality matrix based on K-NN. *Knowledge-Based Systems*. 2019;164:336–347. Available from: https://doi.org/10.1016/j.knosys.2018.11.001.
19) Düntsch I, Gediga G. Confusion matrices and rough set data analysis. *Journal of Physics: Conference Series*. 2019;1229(1):012055–012055. Available from: https://dx.doi.org/10.1088/1742-6596/1229/1/012055.
20) Hicks LA, Wheeler N, Sánchez-Busó L, Rakeman LJ, Harris RS, Grad HY. Evaluation of parameters affecting performance and reliability of machine learning-based antibiotic susceptibility testing from whole genome sequencing data. *PLOS Computational Biology*. 2019;15(9). Available from: https://dx.doi.org/10.1371/journal.pcbi.1007349.