# CLASSIFICATION OF POVERTY LEVELS USING MACHINE LEARNING

Dedeepya Poreddy*
*Department of Computer Science and Engineering*
*GITAM School of Technology, Bengaluru, Karnataka, India*

Elidhandla Vijaya Vardhan Reddy
*Department of Computer Science and Engineering*
*GITAM School of Technology, Bengaluru, Karnataka, India*

S Venkatesh Prasad
*Department of Computer Science and Engineering*
*GITAM School of Technology, Bengaluru, Karnataka, India*

K Ashika Reddy
*Department of Computer Science and Engineering*
*GITAM School of Technology, Bengaluru, Karnataka, India*

Dr. Mylara Reddy C
*Assistant Professor*
*Department of Computer Science and Engineering*
*GITAM School of Technology, Bengaluru, Karnataka, India*

***Abstract* – Poverty has become a tenacious root cause of many socio-economic problems. One of the major reasons for poverty in India is high population rate. In order to reduce poverty, the government has to lay down few best policies. Usually before designing the policies, a survey is conducted by considering only the direct parameters like income or consumption levels. But only direct parameters are not sufficient to categorize household into levels of poverty. Hence, we classify the poverty levels using indirect parameters like education, number of adults in household, house condition and many more parameters. Our approach is designed in such a way that (i) a subset of features is extracted which are important to classify poverty class (ii) inspect how the extracted subset of features affect the class, and at last (iii) we use few machine learning models and check which model performs better. By using Proxy Means Test (PMT) we examine poverty classes within a multidimensional feature space, instead of examining poverty classes within a classically used single dimension perspective.**

**Keywords – Multidimensional feature space, feature engineering, poverty classification, machine learning, feature selection.**

## I. INTRODUCTION

Poverty has features which vary according to geographical location. For example, if a person is said to be poor in any country, he might not be poor in other countries. Classifying poverty into different levels is time consuming, tough and costly. People who are poor usually do not have documentation of their savings or expenses. Historically, poverty was measured considering only income, which neglected other assets and costs. But, we classify poverty levels using numerous variables including income, assets, education, health and many more, this process is called Proxy Means Test (PMT).

The PMT is based on the surveys conducted by the Government. Given that household income in developing countries is often difficult and expensive to measure accurately, the methodology relies on household assets and other indicators or proxies to estimate household welfare[1].

Classifying poverty has complications, (1) Identifying poverty (2) Creating an index for measuring poverty[2]. First problem can be solved using income which is a classical approach, but second is little difficult one. In order to solve the second complication researches proposed suggested poverty measurement indices, one among them is Multidimensional Poverty Index (MPI)[3-5]. Machine learning models are used to find out target from datasets that are labelled using MPIs. Bigdata is being used along with machine learning to classify poverty levels in many developing countries.

In our approach, the following steps are employed, (i) Poverty classification using multidimensional concept, which uses various household characteristics which were extracted using PMT; (ii) feature engineering, which helps to group features into a specific poverty class; (iii) dividing poverty into four classes instead of classifying them into poor or non-poor.

## II. DATA

The dataset is taken from Kaggle website. The data is collected using PMT. PMT is a tool for targeting poverty, which uses visible characteristics of household, when income is not present. It consists of four classes (extreme poverty, moderate poverty, vulnerable and non-vulnerable which are denoted by 1, 2, 3 and 4 respectively, here 1,2,3,4 are target values) and various household characteristics. The folder extracted consists of train.csv and test.csv files with 9557 rows and 23856 rows respectively.

Now when we look at poverty level distribution it is very imbalanced as shown in figure 1.

Even missing values are found in many columns in dataset. Figure 2 states that there are 5911 people who own a house and 961 people who own a house but still pay instalments. This fact can be used during data pre-processing.

Correlation is also performed. Correlation determines the relationship or association of two variables.

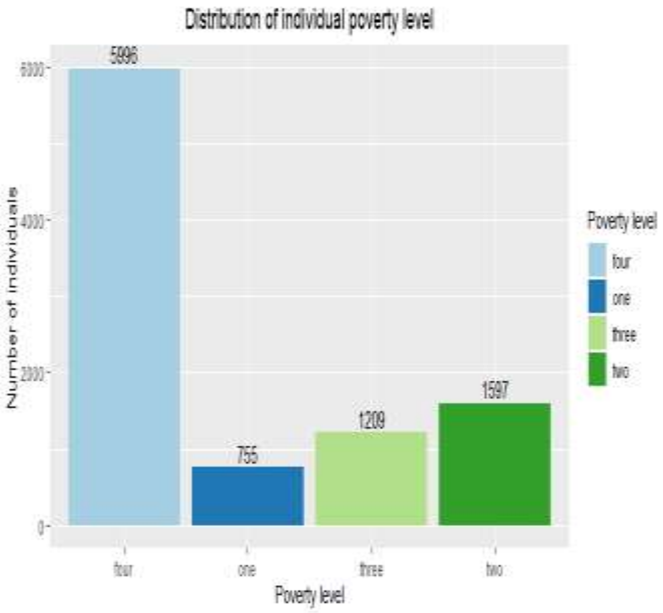

Figure. 1 Distribution of poverty level

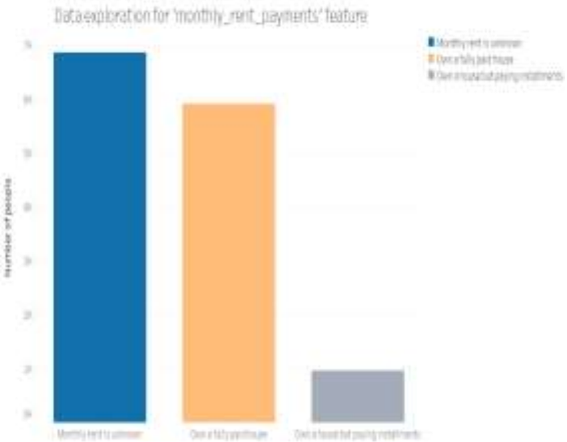

Figure. 2 Data exploration of ‘monthly_rent_payments’ feature

Duplicate columns are removed from dataset. The columns that are unimportant have been eliminated. Machine learning models are used to handle all these complications.

## III. METHODOLOGY

We proposed a framework using following steps and methodologies, and those are data cleansing, feature engineering, feature construction, features selection, machine learning modelling, model optimization and finally validation is also done.

- *Data cleansing*: It is a process of removing unwanted observation, fixing structural errors, managing unwanted outliers and handling missing data.
- *Feature engineering*: It is the process of extracting features from raw data. The extracted features can be used to improve the performance[6].
- *Feature construction*: It involves transforming a given set of features to generate a new set of powerful features which can then be used for prediction. Engineering a good feature space is a prerequisite for achieving high performance in any machine learning task[7].
- *Feature Selection*: Feature selection is also known as variable selection attribute selection or variable subset selection, is a process of selecting a subset of relevant features for use in model construction[8].
- *Machine Learning Modelling*: In this step we will apply data against different machine learning techniques.
- *Model Optimization*: Model optimization is the process of extracting the best performance from a machine learning model by tuning the hyper-parameters through cross-validation. This is necessary because the best model hyper-parameters are different for every dataset. The reduction of a algorithm to its most efficient form by removing unused portions of code and improving the speed.
- *Validation:* Validation is a process where a model which is already trained is evaluated with testing dataset.

Figure 3 represents general framework for classification of poverty levels.

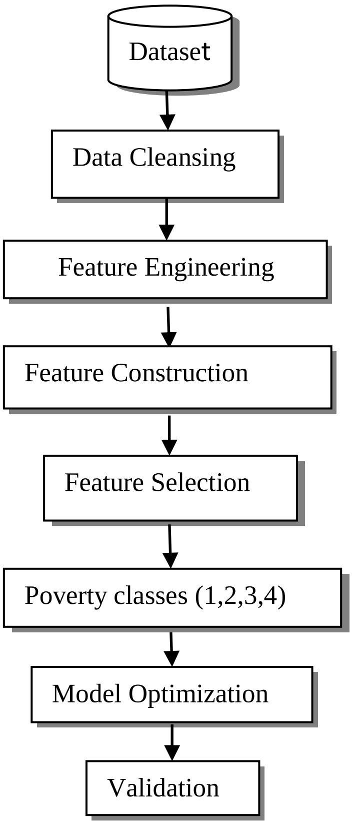

Figure 3: General framework for classification of poverty levels

## IV. CLASSIFICATION

In order to incorporate the individual data into household data, we used aggregation so that a consistent model can be build.We then divided whole dataset into four classes extreme poverty, moderate, vulnerable and non-vulnerable classes. Later, we applied data to various classifiers like LinearSVC, GaussianNB, MLPClassifier, LogisticRegressionCV, RidgeClassifierCV, LinearDiscriminantAnalysis, KNeighborsClassifier and RandomForestClassifier. It is observed that mean F1 score of Random forest classifier is high. Since the data is small, it was tough for the classifiers to perform well in separation of less populated classes. The non-vulnerable class(class 4) is over-represented compared to other classes (shown in figure. 1). And it is even hard to separate class 4 from all other classes. Therefore, we used an overall F1 score and the scores are shown in figure 4. We also calculated confidence using 5-fold cross validation and it is shown in figure 5.

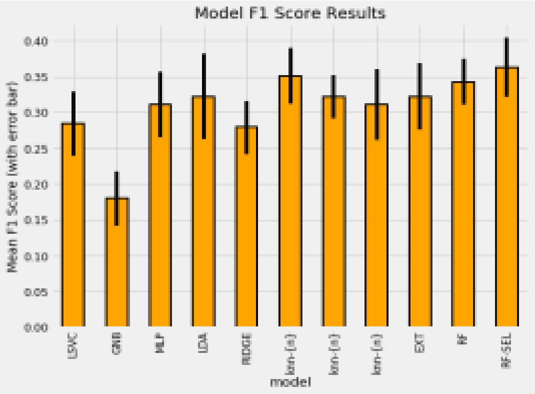

Figure 4: F1 scores results

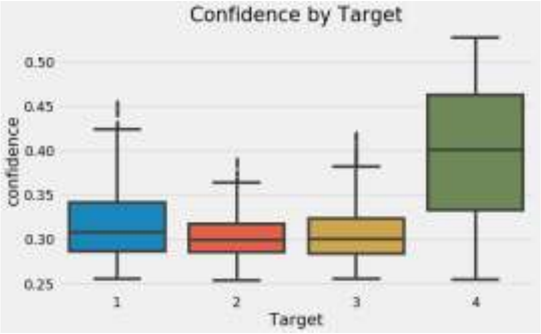

Figure 5: Boxplot representing confidence

## V. DISCUSSION OF RESULTS

We chose few features from combined features that divides each class the best instead of combined selection of features for all classes. For example, if a person is not educated he falls under extreme poverty class, while well educated comes under non-vulnerable class. Similarly different features contribute differently to classes. Hence, we can say that poverty is a multi-dimensional concept. We evaluated the classification performance using F1 score and the confusion matrix shown in figure 6. We had to use F1 score on test data because the access to actual labels of text data was not provided. We have divided the data into the ratio of 3:1 i.e., 75% of data is for training and 25% of data for testing. The outcomes show that the class individuals are mixed strongly, with an exception of non-vulnerable class. Figure 7 shows the difference between train label distribution and predicted label distribution.

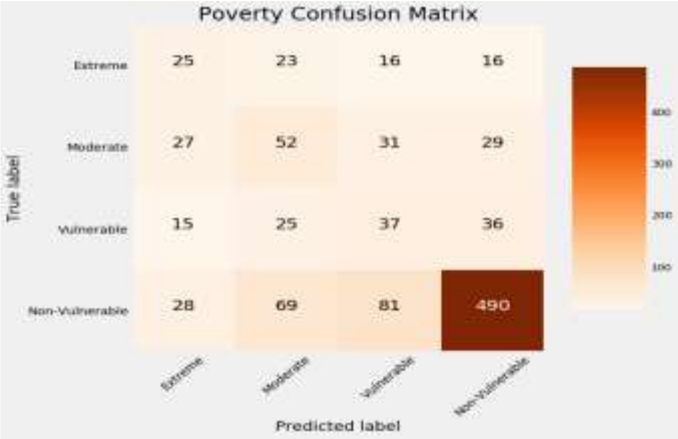

Figure 6: Confusion matrix

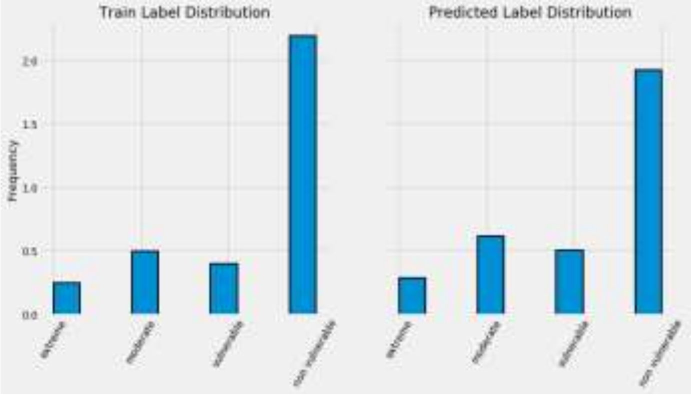

Figure 7: Train label and predicted label distribution

## VI. CONCLUSION

In this paper, we have classified each class of poverty. We applied different classifiers and found out that Random Forest performed well. We also observed that only a single pre-defined set of features are not sufficient to classify levels of poverty, instead we need different features for different classes. Classification performance is completely based on the data. The data that we took was not balanced and missing data was present, where these contribute adversely to the performance of classification. Feature selection worked out well for classifying the data.

## REFERENCES

[1] An assessment of Proxy Means Test made by Australian Government “Targeting the poor”

[2] Poverty: An Ordinal Approach to Measurement,by Sen, Econometrica, vol. 44, no. 2, p. 219, 1976.

[3] S. Alkire and M. E. Santos, “Multidimensional Poverty Index,” Oxford Poverty Hum. Dev. Initiat., no. July, pp. 18, 2010.

[4] S. Alkire and S. Seth, “Multidimensional Poverty Reduction in India between 1999 and 2006: Where and How?,” World Dev., vol. 72, pp. 93108, 2015.

[5] N. Nari and N. Quinn, “Alkire-Foster Method The Global MPI Policy Use Public Communication The Global Multidimensional Poverty Index,” no. November, 2017.

[6] Blog: Wikipedia, “Information about Feature Engineering”,(URL: https://en.m.wikipedia.org/wiki/Feature_engineering)

[7] Blog: Wikipedia, “Information about Feature Construction”,(URL: http://sifaka.cs.uiuc.edu/~sondhi1/survey3.pdf)

[8] Blog: Wikipedia, “Information about Feature Selection”,(URL: https://en.m.wikipedia.org/wiki/Feature_selection)