VIT

Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

J Component report

Programme : B.Tech

Course Title : Foundations of Data Analytics

Course Code : CSE3505

Slot : F2

Title: Estimation of citizens below poverty line

Team Members: Ansh Goel 20BCE1798

Jesmine Akhter 20BCE1945

Khushi Mattu 20BCE1189

Uditi Gupta 20BCE1445

Faculty: Dr. Sheik Abdullah

Sign:

Date: 18/11/2022

CSE 3505
Foundations of Data Analytics
J Component

Estimation of citizens below poverty line

By
Ansh Goel 20BCE1798
Jesmine Akhter 20BCE1945
Khushi Mattu 20BCE1189
Uditi Gupta 20BCE1445

INDEX

List of tables

List of Figures

1. Abstract

Through this project, we aim to provide a better classification method for the given problem, which takes into consideration the number of family members, gender of family members, age of family members, number of rooms in the house, presence of sanitation facility, years of schooling, etc. We would be using a data set of the population in Costa Rica and classify them in the following categories: extreme poverty, moderate poverty, vulnerable households, and non-vulnerable households.

In the future we would work with data sources, exploratory data analysis through visualization, model development, fine-tuning, approaches to tackle data imbalance problems, performance metrics, and visualization of results.

2. Introduction

2.1 Objective and goal of the project

We plan to perform Exploratory Data Analysis to understand the data and impactful variables.

After understanding the data, we will build a prediction model to perform classification to segregate households into four levels :

- Extreme poverty households
- Moderate poverty households
- Vulnerable households
- Non-vulnerable households

We aim to provide a better classification method for the given problem, which takes into consideration the number of family members, gender of family members, age of family members, number of rooms in the house, presence of sanitation facility, years of schooling, etc.

We would be using a data set of the population in Costa Rica and classify them in the following categories: extreme poverty, moderate poverty, vulnerable households, and non-vulnerable households.

2.2 Problem Statement

It is crucial for the government or banks to pinpoint the appropriate households in need of assistance for their social welfare programmes in a state or a neighborhood.

People in economically disadvantaged communities have been found to lack the

knowledge or be unable to produce the required paperwork, such as income and spending records, to demonstrate their eligibility for assistance. Poverty is a major hurdle our country, and world in the general, faces today. But to be able to help these people stuck in the vicious cycle of poverty, one has to first categorize whether a person lives in poverty or not. This presents a major problem for government agencies and leads to wastage of resources or incorrect distribution of resources to people enrolled in these poverty alleviation programs.

The classification is done exclusively over the person's income but there are a number of factors that remain unanalysed. Through our project we plan to take into account a number of factors such as number of family members, gender of family members, age of family members, number of rooms in the house, presence of sanitation facility, years of schooling, etc. to classify people into various levels of poverty. For that we have used various algorithms like naive bayes classification , Random Forest, KNN Classification, Gradient Boosting etc.

2.3 Motivation

Poverty is a major hurdle our country and world in the general face today. It's difficult to get out of the vicious cycle of poverty once you've been ensnared in it. Governments of various countries try to elevate this situation by providing numerous social programs.

But mostly these programs are not of great help as they are not able to correctly determine the target audience, i.e the people who actually need the help granted by these governments. The classification based solely on income and expense records is found to be ineffectual.

2.4 Challenges

The major challenges faced were :

- The selection of important features for the prediction. The dataset used for prediction has 140 independent variables, directly using this entire set for prediction can be quite time consuming and cumbersome.

- The dataset selected has 4 different target values : 1 = extreme poverty, 2 = moderate poverty, 3 = vulnerable households, and 4 = non vulnerable households. But the training dataset has a class imbalance where 62.73% of data belongs to the 'Non-vulnerable household' class. This may lead to inaccurate predictions for the other classes and thus, needs to be dealt with.

2.5 Data set Description

The train.csv and test.csv files in the data folder include 9557 rows and 23856 rows,

respectively. The 'target' column, which establishes the poverty level, is absent from the test.csv. As a result, the 3.08 MB train.csv data set is utilized alone. There are 143 columns in all. One individual is connected to each record. The above-mentioned data source's URL contains the descriptions of 143 columns.

Below are some of the columns' descriptions.
Target: Indicates level of poverty 1 represents the poorest households, 2 the poorest households, 3 the vulnerable families, and 4 the non-vulnerable households.
Idhogar: A household's individual identification number.
This column serves to identify individuals who live in a single home.
v2a1: Each family pays a monthly rent. rooms: the number of rooms in the house. escolari: years of schooling etc.

3. Literature Survey

For proceeding with our research, we have surveyed various research papers in the related field. Below are the notable ones among them:

The dataset used for the stated analysis was Costa Rican Household data provided by Inter-American Development Bank . The data was pre-processed using missing value imputation by mode or 0, dealing with categorical columns.Prediction Algorithms Used: K-Means Clustering Decision Tree,Multinomial Logistic Regression and Gradient Boosting For the model evaluation Accuracy Score for the ensemble model, Residual Deviance and AIC(Akaike Information Criteria) for Multinomial logistic regression was taken into consideration. Variable Importance Plot generated to discover the effective predictors. Accuracy of the model is about 92.64% observed for the validation set. The proposed model works well for the training set but needs to be tested on the testing set provided.[1]

Data Preprocessing is done using missing value imputation, interpolation,Binarization: transforming 4-class problem into 4 different binary problems and SMOTE to tackle class imbalance.Emphasizes on understanding the features that most affect each target class.Random Forest has been proposed for conducting the above analysis. Evaluation was doing using F! Macro score. The proposed model fails to identify a fair decision boundary between the less populated classes: “Extreme”, “Moderate”, “Vulnerable”[2]

Simulations are made on varying imbalance levels: 50%, 40%, 30%, 20%, 10%, 5%, and 1% for both approaches. Performance evaluation has been done for real world dataset after introducing simulated noise predictors. Analyzing class specific variable importance. Analyzing influence of sample size: observing performance for both methods

with varying sample size Analyzing influence of effect size : observing performance for both methods for varying effects[3]

Classification Methods used are Logistic Regression, Random Forest, LightGBM and Overfitting Control Measure is Cross Validation.It was clearly observed that LightGBM outperformed all other classification models in terms of F1 Score and Accuracy. Logistic Regression didn't perform well because as the data is not linear it was not able to generate the accurate equation. The next best alternative is Random Forest Classifier which had better results when compared to Logistic Regression.[4]

A clustering approach to identify multidimensional poverty indicators for the bottom 40 percent group

The Davies Bouldin metric measures the variation between points within the cluster and the distance between clusters. In each cluster, this metric determines which other group has the highest ratio between the average intracluster distance of points in two clusters to the distance between. **Average within Centroid Distance** metric is measured by calculating the average distance per point from a centroid point within a cluster. **Sum of squares** metric divides the number of data points in a group by the number of data points in each cluster. This is called squared, and the values of all the clusters are summed. Cosine Similarity > Euclidean Distance > Correlation Similarity > Dice Similarity It is shown that the **Cosine Similarity is the best distance technique** based on the lowest score obtained.A clustering model-based K-Means Algorithm with Cosine Similarity measure is developed to form clusters of B40 group.The evaluation found k = 8 to be the best k value for the model. By employing the descriptive statistics method, three dimensions have been established: Education, Living Standards and Employment with seven indicators: literacy, highest education level and grade, sanitation, housing, access to television services, assets and work.[5]

Classification of Poverty Levels Using k-Nearest Neighbor and Learning Vector Quantization Methods

Classification Methods used : K-nearest neighbor algorithm, Learning vector quantization algorithm. Based on the results and discussion, it can be concluded that the accuracy of the classification by using the amount of training data with the value of parameters k = 4, $\alpha = 0.01$ and 300 iterations values obtained highest accuracy in the k-nearest neighbor (k-NN) amounted to 93.52%, while highest accuracy on learning vector quantization (LVQ) amounted to 75.93%
In terms of the performance of both methods of classification, k-NN method is faster in the process of running the program when compared to LVQ.It can be concluded that the k-NN method is better compared to LVQ in relation to the issues of poverty level classification.[6]

Machine Learning Approach for Bottom 40 Percent Households (B40) Poverty Classification

Identifying the best machine learning models for classifying the B40
population using Naive Bayes, Decision Tree and k-Nearest Neighbors algorithm
Several data pre-processing task using WEKA such as data cleaning, feature engineering, normalization, feature selection: Correlation
Attribute, Information Gain Attribute and Symmetrical Uncertainty Attribute and sampling methods using SMOTE.
Each classifier is then optimized using different tuning
parameter with 10-Fold Cross Validation for achieving the optimal values before the performance of the three classifiers are
compared to each other.[7]

Multivariate random forest prediction of poverty and malnutrition prevalence

The viability of both contemporaneous prediction and sequential nowcasting of malnutrition and poverty prevalence using a set of features drawn from open access data sources, in conjunction with random forest methods. Multivariate, joint prediction of multiple malnutrition and poverty indicators modestly improved predictive performance of some prevalences in sequential nowcasting, but not in contemporaneous prediction

However, we found the performance of our model deteriorated considerably when predictions were assessed at the level of individual surveys in our nowcasting framework, as measured by r^2, due largely to poor performance on initial, small sample size, surveys, alongside a less extreme drop in performance when assessing contemporaneous results at this scale. Data from eleven USAID Feed the Future (FTF) priority countries: Bangladesh, Ethiopia, Ghana, Guatemala, Honduras, Kenya, Mali, Nepal, Nigeria, Senegal, and Uganda was used in this paper.[8]

Global Poverty- A First Estimation of its Uncertainty

Using the findings for the DAD method, the number of people living in conditions of absolute poverty in the world's developing countries in 1990 is estimated to stand between 502 and 2823 million people, while in 2015 the interval is between 128 and 1788 million. We further show that MDG1 was achieved at a confidence level of 80%, considerably lower than the typical benchmark of 95%. Moreover, we find that at 95% confidence level a 5.19% reduction in global poverty is identified by the DAD method (baseline scenario). However, given the considerable poverty reduction identified by the CBN method in the same period, we conclude that the inability of DAD to identify substantial poverty reduction in the MDG1 period says more about the uncertainties of the method per se, rather than about the actual evolution of global poverty.[9]

Statics and analysis of targeted poverty Alleviation information integrated with big data mining algorithm

Classification methods used - Maslow's Demand Level Random forest model, Logistics Model & Waterfall Model.The essence of targeted poverty alleviation is that the government effectively identifies poor families and members, investigates the causes and

extent of poverty, and provides practical and effective assistance in order to fundamentally break down poverty barriers. The logistic algorithm, random forest algorithm, and newly proposed waterfall model in data mining are discovered through the research. The newly proposed waterfall model has the advantages of high sample reuse rate, effective overfitting prevention, and no requirement for massive data.[10]

Poverty Prediction using Random Forest based Machine Learning Technique

Random Forest, MPI based classification.The accuracy of Random Forest algorithm is compared with MPI.The accuracy of the proposed method is always higher. the accuracy of the proposed method is 100% as compared to previous method whose value is 92.85%.

Poverty Classification Using Machine Learning: The Case of Jordan

Logistic regression, KNN, decision trees, Support vector machine, Naive Bayes, Adaboost.The final machine learning classification model has achieved an accuracy that aligns
with the acceptable accuracy in the scientific literature.In terms of the robustness of the final model, Jordan has undergone many political, economic and social changes that had a direct impact on the pattern of poverty, and at other times indirectly. These changes were reflected in the data obtained from the field surveys over the different years. Furthermore,since the proposed model is originally based on these data, we conclude that the model.[11]

Poverty Level Prediction Based on E-Commerce Data Using K-Nearest Neighbor and Information-Theoretical-Based Feature Selection

3rd International Conference on Information and Communications Technology (ICOIACT), 2020

An e-commerce data from the largest e-commerce company. This dataset contains several advertisement data namely motorcycles, cars, lands,
apartments, and houses. In this dataset, there are 96 features, and 1 label.

Predict the poverty level based on an e-commerce dataset using K-Nearest Neighbor and Information Theoretical
Based Feature Selection.
Data Cleansing
Normalization
Information Theoretical Based Feature Selection (CIFE, MRMR, DISR) [12]

Root mean squared error (RMSE) and R^2 (R-Square). R^2 is utilized for vector variances that can be predicted, such as if R^2 = 1 then the
regression model is said to be correct, and vice versa. The RMSE is used to measure the difference in error between the actual and predicted vectors.
The lower RMSE value indicates less difference between the actual and the predicted value.

The information theoretical based feature selection calculates the redundancy value of each feature or between features. The results of the regression in the KNN method and the
information-theoretical based feature selection are very relevant and excellent as a result of the R-Square.
The R-Square or the accuracy value reaches 0.35 in the CIFE feature selection experiment. From our experiments, it can be concluded that ecommerce
data coupled with KNN methods and information-theoretical based feature selection can be used to predict poverty levels in an area. [13]

Household poverty classification in data-scarce environments: a machine learning approach

NIPS 2017 Workshop on Machine Learning for the Developing World

Data used comes from national household expenditure or income surveys, using the most recent survey available. Uses data from the 2015 Zambia Living Conditions Measuring Survey (LCMS), which
was carried out by the Zambian Central Statistical Office.

Methods used are variable selection, fitting the selected model, and translating the model into scorecard format.

Using a random 2:1 train:test split, the model is trained to process and then evaluated for predictions on the held-out data. Analyzed using box plot. [14]

Methodology leads to reasonable separation between poor and non-poor households
nationally and across several breakdowns (by consumption deciles, sub-national regions, and urban, rural locations). The results are also very close to those from a full logistic regression model with no variable selection, showing that limiting the additive model to only 10 variables does not impair
performance.

Classification of Poverty Levels Using k-Nearest Neighbor and Learning Vector Quantization Methods

International Journal of Computing Science and Applied Mathematics, Vol. 2, No. 1, March 2016

Source of data used is the data targeted households Documenting Social Protection Program in 2011 by taking a sample of the data as much as 216 households
consisting of four 14 criteria/poverty indicators.

Classification Methods used :
K-nearest neighbor algorithm

Learning vector quantization algorithm

Accuracy = (Number of correct predictions)/ (Total predictions)

Based on the results and discussion, it can be concluded that the accuracy of the classification by using the amount of training data with the
value of parameters k = 4, α = 0.01 and 300 iterations values obtained highest accuracy in the k-nearest neighbor (k-NN) amounted to 93.52%, while highest accuracy
on learning vector quantization (LVQ) amounted to 75.93%
In terms of the performance of both methods of classification, k-NN method is faster in the process of running the program when compared to LVQ.
It can be concluded that the k-NN method is better compared to LVQ in relation to the issues of poverty level classification.

4. Proposed Methodology- Diagram

The following flowchart represents the system design of the proposed system:

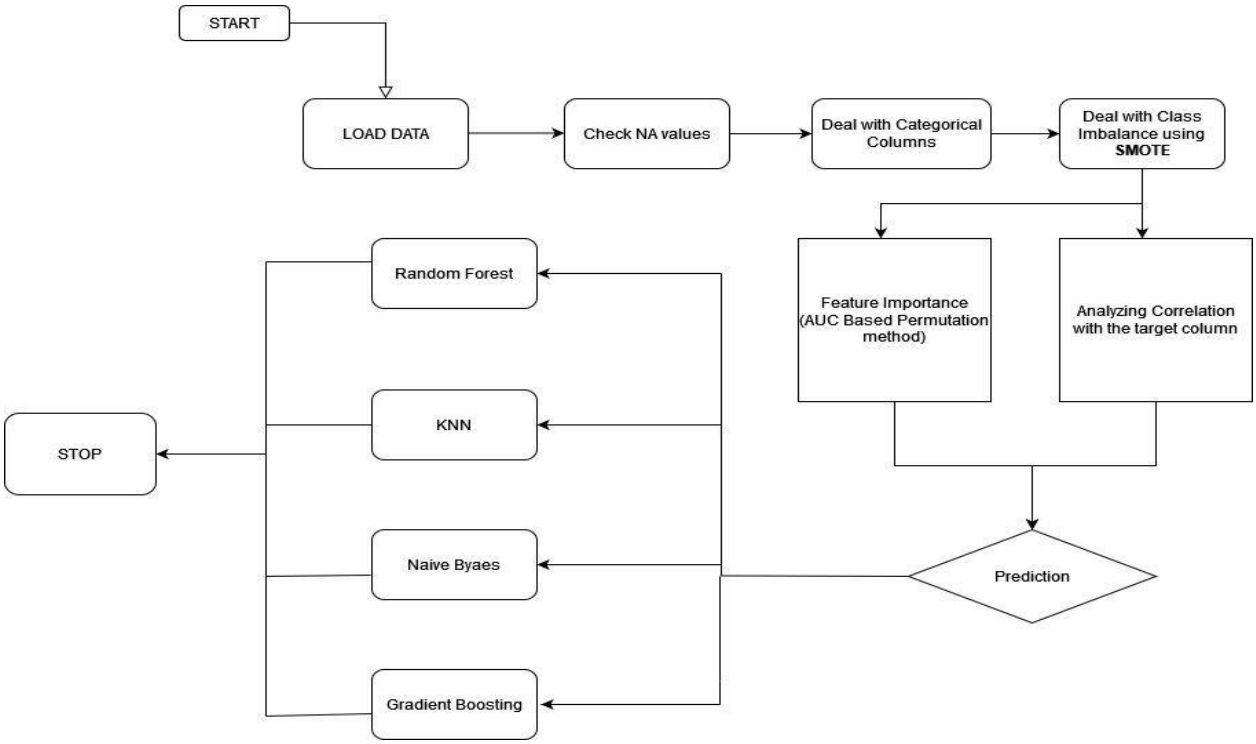

Figure 1. Proposed system

5. Implementation

5.1 Loading Dataset

Load the dataset using head(train)

5.2 Checking NA values

We found a total of 22140 NA values in 5 different columns namely, v2a1, v18q1, rez_ecs, meaneduc, SQRmeaned. In case of v2a1, v18q1, meaneduc and SQRmeaned, we are replacing the missing values with the median values of the variable. We are dropping the column "rez_ecs" as it's missing more than 80% of its values.

v2a1	v18q1	rez_esc	meaneduc	SQBmeaned
<db1>	<db1>	<db1>	<db1>	<db1>
6860	7342	7928	5	5

Table 1. Checking for Na values

For 'rez_ecs': can drop column
For 'v2a1','v18q1','SQBmeaned','meaneduc': Replace NA with median value

5.3 Dealing with Categorical Columns

We found 5 columns with categorical values namely, Id, idhogar (household level identifer), dependency (dependency rate), edjefe (Years of education for male head) and edjefa (Years of education for female head). We are dropping the 'Id' column as it is the unique identifier for each column and has no use in prediction. 'Idhogar' has a huge number of unique values so we are using label encoding for this column. Lastly for 'dependency', 'edjefe' and 'edjefa' we are replacing "No" with 0 value and "Yes" with median value of respective columns.

Id column not needed for classification, though keep for right now
'idhogar ' column: Household level identifier (too many unique values with small frequencies, implementing label encoding)
'Dependency': dependency rate (replace 'no' with zero, 'yes' with median value)
'Edjefe': Years of education for male head of family (replace 'no' with zero, 'yes' with median value)
'Edjefa': Years of education for female head of family (replace 'no' with zero, 'yes' with median value)

5.4 Dealing with Class Imbalance(SMOTE)

The dataset used for prediction showcases class imbalance. About 63% of the training data belongs to the "Non-vulnerable household" target class. To resolve this problem, we are applying SMOTE (Synthetic Minority Over-sampling Technique) thrice. In the first iteration, we oversample class "Extreme Poverty" class i.e. class with the smallest

percentage of records in the training dataset. In the second iteration, we oversample the "Vulnerable Household" class and lastly, in the third iteration, we oversample the "Moderate Poverty" class.

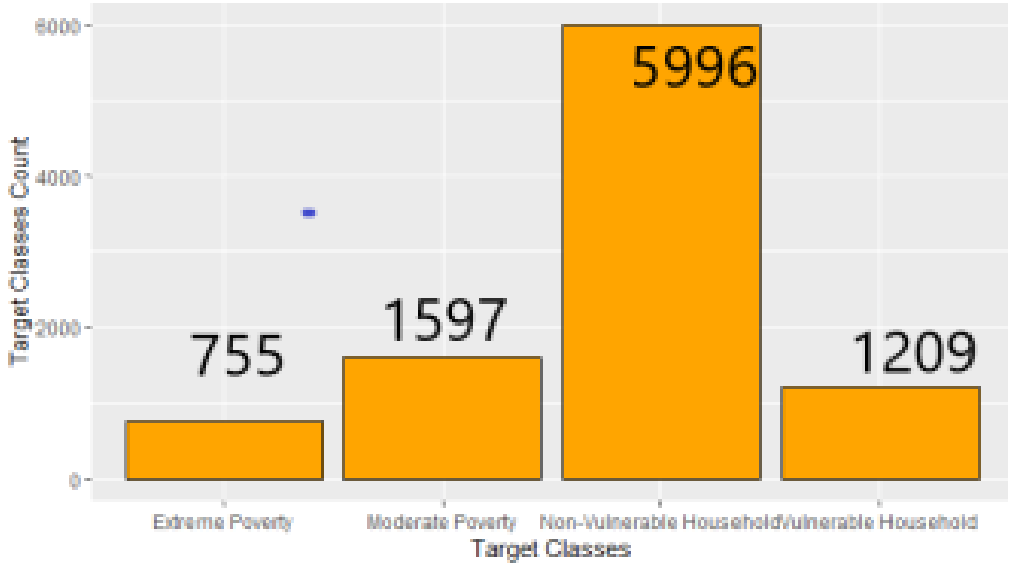

Figure 2: Dealing with Class Imbalance(SMOTE)

Undersampling would decrease the proportion of your majority class until the number is similar to the minority class. At the same time, oversampling would resample the minority class proportion following the majority class proportion.

Var1	**freq**
<fct>	**<int>**
1	**15855**
2	**3782**
3	**2905**
4	**14453**

Table 2. balanced_train

Var1	**freq**
<fct>	**<int>**
1	**27030**

2	**6393**
3	**31955**
4	**24677**

Table 3. balanced_train1

Var1	**freq**
<fct>	**<int>**
1	**29133**
2	**31965**
3	**34059**
4	**26310**

Table 4. balanced_train2

5.5 Statistical Analysis used:

5.5.1 Feature-Importance (AUC-based Variable Importance)

Due to oversampling, the number of observations in the training dataset has crossed 100000, so we took a small proportion of the dataset and used it to calculate VIM (Variable Importance Measure).

We are using the Random Forest model and RSME as the loss function here.

Obtained VIM graph over 50 permutations :

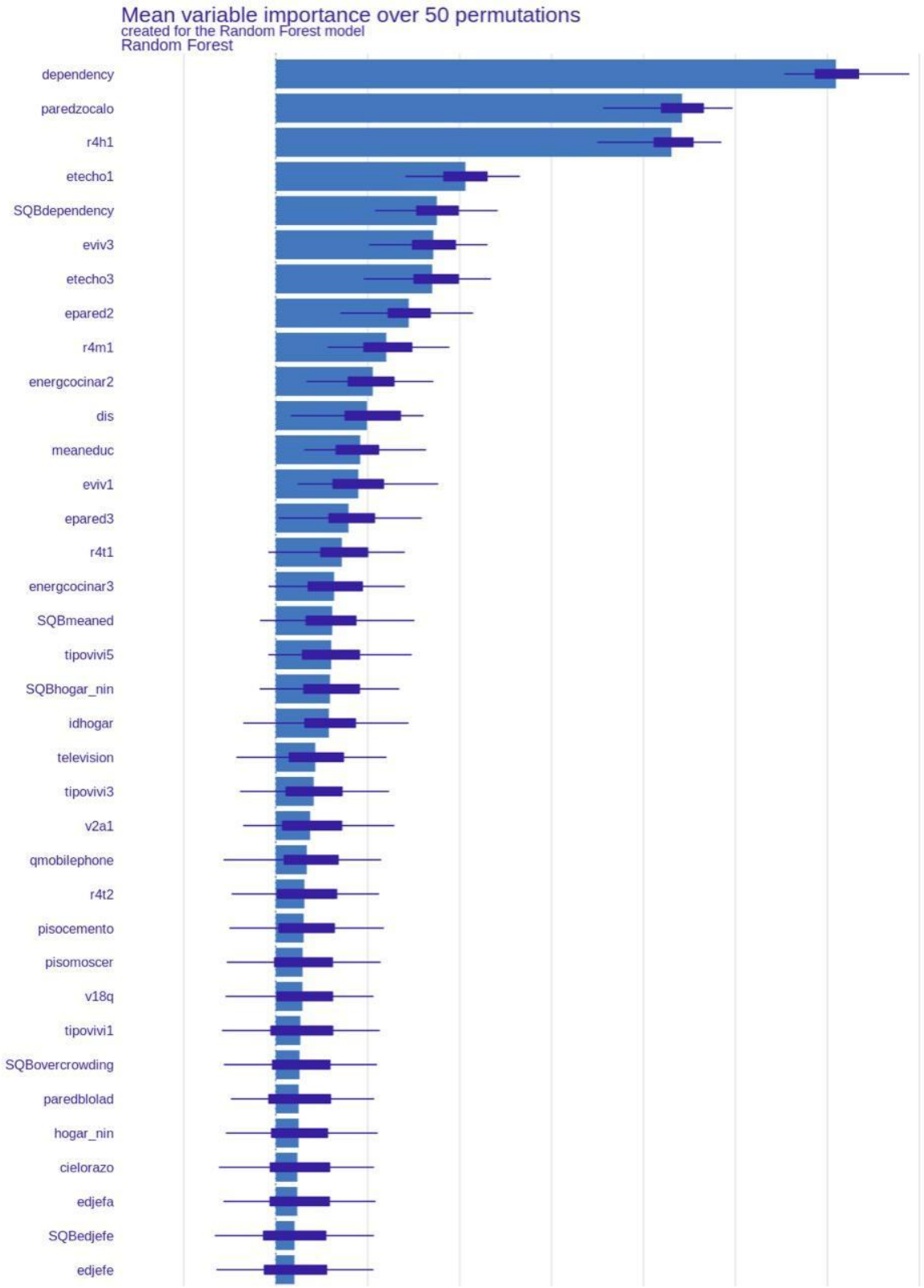

Figure 3: Feature importance

Dependency, paredzocalo, r4h1, etecho1 have highest mean so we opted for these features in our implementation.

Columns selected for prediction according to VIM :

- dependency = Dependency rate
- paredzocalo =1 if predominant material on the outside wall is socket (wood, zinc or absbesto
- r4h1 = Males younger than 12 years of age
- etecho1 =1 if roof are bad
- SQBdependency = **can ignore same as dependency, just squared**
- eviv3 =1 if floor are good
- etecho3 =1 if roof are good
- epared2 =1 if walls are good
- r4m1 = Females younger than 12 years of age
- energcocinar2 =1 main source of energy used for cooking electricity
- dis =1 if disable person
- meaneduc = average years of education for adults (18+)
- eviv1 = 1 if floor are bad
- epared3 =1 if walls are good
- r4t1 = persons younger than 12 years of age
- energcocinar3 =1 main source of energy used for cooking gas
- SQBmeaned = meaned squared
- tipovivi5 = 1 other(assigned, borrowed)
- SQBhogar_nin = hogar_nin squared

5.5.2 Correlation

Next, we tried to find important features using correlation of each feature with the Target column. We considered columns with correlation coefficients greater than 0.3 or columns with correlation coefficients lesser than -0.3.

Greater than 0.3:

	Target
	<db1>
Target	**1.00**
epared3	**0.4750542**
etecho3	**0.4558925**
escolari	**0.3595028**
v18q	**0.3562209**
paredblolad	**0.3441004**
SQBescolari	**0.3404588**
television	**0.3078663**

Table 5. Greater than 0.3

Less than -0.3:

	Target
	<db1>
r4t1	**-0.5082399**
etecho1	**0.4441216**
paredzocalo	**-0.4425884**
r4m1	**-0.4179358**
epared2	**-0.4012238**
r4h1	**-0.23776451**
hogar_nin	**-0.3596533**
estadocivil1	**-0.3217847**

Table 6. Less than -0.3

Columns selected for prediction according to VIM :

- epared3 (=1 if walls of the house are in good condition)
- etecho3 (=1 if roof of the house is in good condition)
- escolari (Years of schooling)
- v18q (Owns a tablet)
- paredblolad (=1 if predominant material on the outside wall is block or brick)
- SQBescolari (Escolari squared) : can ignore
- television (=1 if the household has TV)
- r4t1 (Number of persons younger than 12 years of age) : can ignore, already taking count of female and male separately
- etecho1 (=1 if roof are bad)
- paredzocalo (=1 if predominant material on the outside wall is socket (wood, zinc or absbesto))
- r4m1 (Number of females younger than 12 years of age)
- epared2 (=1 if walls are regular)
- r4h1 (Number of males younger than 12 years of age)
- hogar_nin (Number of children of age 0 to 19 in household)
- estadocivil1 (=1 if less than 10 years old)

5.6 Prediction

5.6.1 Prediction using VIM feature set :

- **Random Forest**

  Accuracy: 0.969
  95% CI: (0.9645,0.9731)
  No Information Rate: 0.2899
  P-Value [Acc > NIR]: <2.2e-16

- **KNN Classification**

  Accuracy: 0.8119
  95% CI: (0.8022,0.8213)
  No Information Rate: 0.2902
  P-Value [Acc > NIR]: <2.2e-16

- **Naive Bayes Classification**

  Accuracy: 0.6757
  95% CI: (0.6642,0.6871)
  No Information Rate: 0.3806
  P-Value [Acc > NIR]: <2.2e-16

- **Gradient Boosting**

  Accuracy: 0.853
  95% CI: (0.8441,0.8615)
  No Information Rate: 0.3051
  P-Value [Acc > NIR]: <2.2e-16

5.6.2 Prediction using Correlation feature set :

- **Random Forest**

  Accuracy: 0.9154
  95% CI: (0.9083,0.9222)
  No Information Rate: 0.2814
  P-Value [Acc > NIR]: <2.2e-16

- **KNN Classification**

Accuracy: 0.7711
95% CI: (0.7606,0.7813)
No Information Rate: 0.2895
P-Value [Acc > NIR]: <2.2e-16

```
[ ]  plot(model_knn)
```

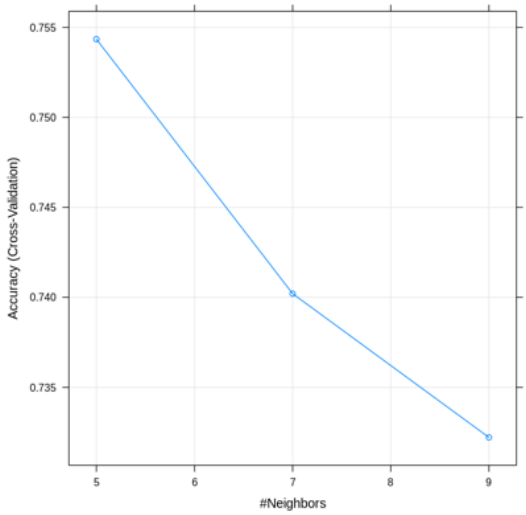

Figure 4: KNN

- **Naive Bayes Classification**

Accuracy: 0.6517
95% CI: (0.6399,0.6634)
No Information Rate: 0.3445
P-Value [Acc > NIR]: <2.2e-16

- **Gradient Boosting**

Accuracy: 0.806
95% CI: (0.7961,0.8156)
No Information Rate: 0.3143
P-Value [Acc > NIR]: <2.2e-16

5.7 Visualization Analysis

5.7.1 Class Distribution

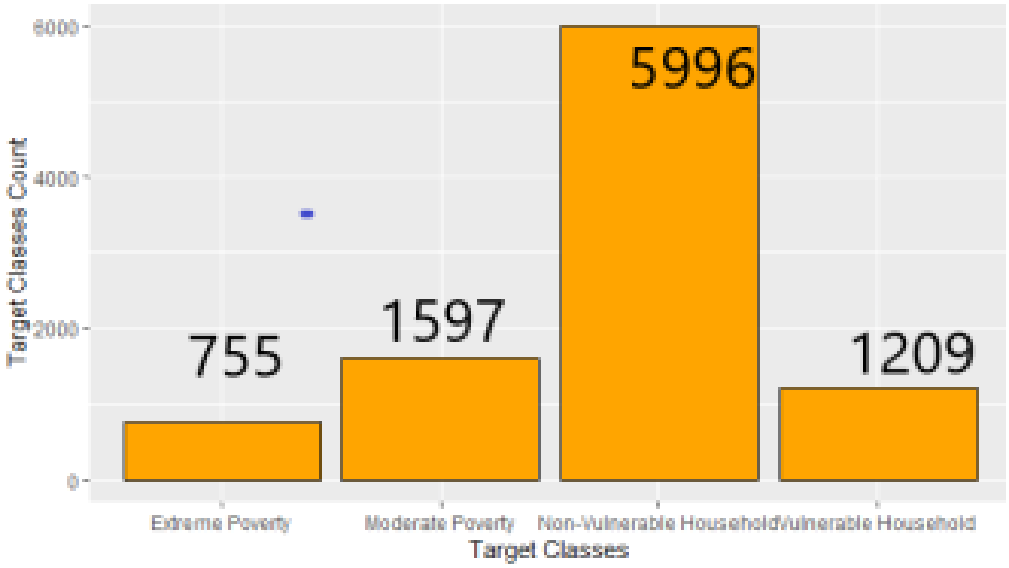

Figure 5: Class Distribution

5.7.2 Elbow Method to find optimal number of k_neighbours

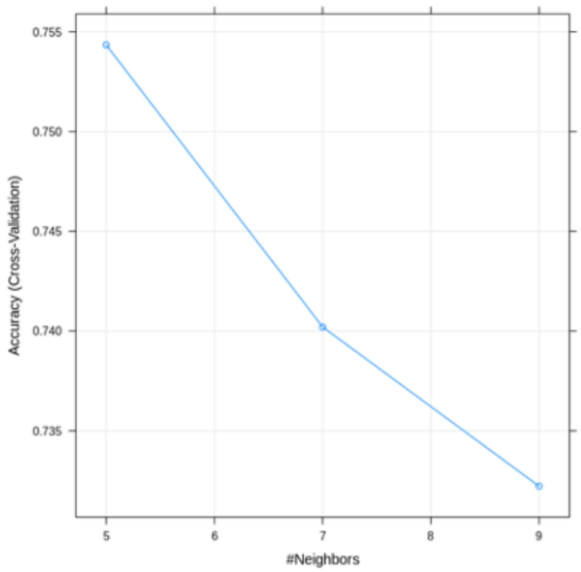

Figure 6: Elbow Method to find optimal number of k_neighbours

6. Experimental results & Discussion

We can see from the table above that the accuracy of each model used varies depending on the feature set. We can also conclude that the accuracy of each model employed in the VIM feature set is higher than that in the Correlation Feature Set.

Models used	Case 1 - VIM Feature Set	Case 2 - Correlation Feature Set
Random Forest	96.90%	91.54%
KNN Classification	81.18%	77.11%
Naive Bayes Classification	67.57%	65.17%
Gradient Boosting	85.30%	80.60%

Table 7. Experimental results

We can observe that random forest is able to provide the best results in both the cases seconded by gradient boosting, this may be because of the fact that these are ensemble techniques which combine the decisions from multiple models to improve the overall performance. Also we can say that the VIM feature set provides results for the test dataset because it took into consideration the area under the ROC curve for all the possible feature set and then selected the one with the best results.

7. Conclusion and Future Work

We would want to do the following in the future:

1. interact with the data set's sources and conduct exploratory analysis using data visualization
2. practice model development in order to obtain better models with higher accuracies.
3. fine-tune the present models in order to acquire even higher accuracies
4. devise new techniques to addressing the data imbalance issues that have arisen
5. conduct performance metrics and see the outcomes

8. References

[1] Sheng, W. A. N. G., Yu, Z. H. A. O., & Yiwei, Z. H. A. O. (2020). Costa Rican Poverty Level Prediction. *IETI Transactions on Social Sciences and Humanities*, *7*, 171-176.

[2] Mohamud, J. H., & Gerek, O. N. (2019, April). Poverty level characterization via feature selection and machine learning. In *2019 27th Signal Processing and Communications Applications Conference (SIU)* (pp. 1-4). IEEE.

[3] Janitza, S., Strobl, C., & Boulesteix, A. L. (2013). An AUC-based permutation variable importance measure for random forests. *BMC bioinformatics*, *14*(1), 1-11.

[4]Venkatramolla, S. K. (2019). Machine learning and data science for a household-specific poverty level prediction task.

[5] Abdul Rahman, M., Sani, N. S., Hamdan, R., Ali Othman, Z., & Abu Bakar, A. (2021). A clustering approach to identify multidimensional poverty indicators for the bottom 40 percent group. *Plos one*, *16*(8), e0255312.

[6] Santoso, S., & Irawan, M. I. (2016). Classification of Poverty Levels Using k-Nearest Neighbor and Learning Vector Quantization Methods. *IJCSAM (International Journal of Computing Science and Applied Mathematics)*, *2*(1), 8-13.

[7] Sani, N. S., Rahman, M. A., Bakar, A. A., Sahran, S., & Sarim, H. M. (2018). Machine learning approach for bottom 40 percent households (B40) poverty classification. *Int. J. Adv. Sci. Eng. Inf. Technol*, *8*(4-2), 1698.

[8] Aulia, T. F., Wijaya, D. R., Hernawati, E., & Hidayat, W. (2020, November). Poverty Level Prediction Based on E-Commerce Data Using K-Nearest Neighbor and Information-Theoretical-Based Feature Selection. In *2020 3rd International Conference on Information and Communications Technology (ICOIACT)* (pp. 28-33). IEEE.

[9] Kshirsagar, V., Wieczorek, J., Ramanathan, S., & Wells, R. (2017). Household poverty classification in data-scarce environments: a machine learning approach. *arXiv preprint arXiv:1711.06813*.

[10] Santoso, S., & Irawan, M. I. (2016). Classification of Poverty Levels Using k-Nearest Neighbor and Learning Vector Quantization Methods. *IJCSAM (International Journal of Computing Science and Applied Mathematics)*, *2*(1), 8-13.

[11] Browne, C., Matteson, D. S., McBride, L., Hu, L., Liu, Y., Sun, Y., ... & Barrett, C. B. (2021). Multivariate random forest prediction of poverty and malnutrition prevalence. *PloS one*, *16*(9), e0255519.

[12] Moatsos, M., & Lazopoulos, A. (2021). Global poverty: A first estimation of its uncertainty. *World Development Perspectives*, *22*, 100315.

[13] Gao, M., Li, L., & Gao, Y. (2022). Statistics and Analysis of Targeted Poverty Alleviation Information Integrated with Big Data Mining Algorithm. *Security and Communication Networks*, *2022*.

[14] iska nhi mil rha hai , i mean bahut research paper aarhe hai

[15] Alsharkawi, A., Al-Fetyani, M., Dawas, M., Saadeh, H., & Alyaman, M. (2021). Poverty classification using machine learning: The case of jordan. *Sustainability*, *13*(3), 1412.