J. Electrical Systems 20-11s (2024): 2930-2943

[1] Anatoli Nachev

Identifying Drivers of Social Marginalisation Using Decision Tree Analysis: A Data-Driven Approach

Abstract: **-** This article presents a study that employs machine learning techniques to investigate the factors influencing poverty as a primary driver of social marginalisation in the United States. Addressing gaps in previous research, we propose a decision tree-based approach to develop a set of predictive models aimed at identifying and quantifying the factors affecting poverty. To this end, we train models using microdata from IPUMS USA, collected from censuses and surveys, containing a wide range of socio-economic variables. By grouping the variables into three categories: predisposing, socio-demographic, and socio-economic, we build three sets of predictive models used to assess the significance of factors associated with these variables. Through variable sensitivity analysis and Variable Effect Characteristic (VEC) analysis to evaluate the values of these variables, we propose a methodology for empirically assessing various factors related to poverty. The proposed methodology would aid in making informed decisions when developing policies and setting priorities in this area.

Keywords: classification, decision trees, machine learning models, poverty, supervised learning.

I. INTRODUCTION

The study of social inclusion factors discussed here is of great importance, as it helps to understand and address the challenges and obstacles faced by marginalized and disadvantaged groups, ultimately promoting a more just and inclusive society. The examination of factors leading to poverty, which is our focus, helps identify the root causes of social exclusion, such as limited access to education, healthcare, employment, and housing. By understanding these causes, targeted interventions can be developed to address the specific barriers that prevent individuals from fully participating in society.

Essentially, poverty involves more than just insufficient financial resources; it also encompasses a lack of access to opportunities, resources, and services essential for a minimum standard of living and full societal participation.

Poverty is defined differently depending on the country or context. In the United States, the federal government defines poverty using a metric called the "poverty threshold" or "poverty line" [9], [10]. This measure is based on specific income levels that vary according to family size and composition. The thresholds are adjusted each year for inflation and are used to assess eligibility for various social programs and to evaluate economic conditions.

Previous research in the field, particularly those using the same data as we do, evaluates the impact of specific factors on poverty, such as age and child poverty [11], [12], [13]; poverty by gender, race, or ethnicity [11], [14], [15]; urban inequality [17], and others. The literature also indicates that the most common approach for data analysis is statistical, including descriptive statistics [11], [12], [13]; correlation analysis, LQ measurement, OLS regression [11], [16], [15]; Gini coefficient [17], etc.

To assess the risk factors associated with poverty, this study employs predictive models based on decision trees, supplemented by sensitivity analysis and analysis of the variable effect characteristics (VEC) of the variables.

The paper is organized as follows: Section II provides an in-depth overview of the methodologies used for data analysis; Section III details the data utilized in the study; Section IV presents and discusses the experimental results; and Section V concludes the study.

II. METHODS

A. Data Analysis Framework

Successfully conducting a data analysis project requires a methodology that determines the correct sequence of steps and actions necessary to achieve the desired results. The one used here is *CRoss Industry Standard Process for Data Mining* (CRISP-DM), which is the framework for data science and machine learning that has proven itself among the research community to be practical and reliable [1]. It consists of 6 main stages, illustrated in Fig. 1, which are carried out sequentially, but some of them allow backtracking to previous stages for the purpose of

[1] Business Information Systems, School of Business and Economics, University of Galway, Ireland.
Email: anatoli.nachev@universityofgalway.ie

corrections, specifying details that have been neglected, or adaptation of some parameters. The stages can be outlined as follows:

The *business understanding* step focuses on understanding the business problem in all its aspects, including defining it in both business and data mining terms. The step also formulates success criteria for the project from both a business and data mining perspective.

The *data understanding* phase emphasizes gathering, describing, and exploring the data. This step includes applying univariate statistical analysis to examine each individual data variable and bivariate statistical analysis to investigate the dependencies between pairs of variables. Common techniques used in this phase include correlation analysis, Chi-square tests, T-tests, and Z-tests to run significance tests and explore variable relationships.

The *data preparation* step prepares the provided data for modelling. This involves handling N/A and null values, imputing missing data, treating outliers, rescaling data, e.g. standardising continuous variables. Additionally, it includes feature selection and dimensionality reduction, which is essential for optimising model parameters. This step also deals with feature transformation and extraction to create additional attributes from the existing ones.

The *modelling* step involves partitioning the data, establishing strategies for model testing, training models using selected machine learning or statistical modelling algorithms, tuning their hyperparameters, and estimating their performance using relevant metrics.

The *evaluation* step deals with evaluating the results against the success criteria defined at the business understanding step, reviewing the previous steps and the work completed, and determining if deployment is necessary.

The *deployment step*, not applicable in this case, involves deploying the model in a business environment and providing maintenance, if needed.

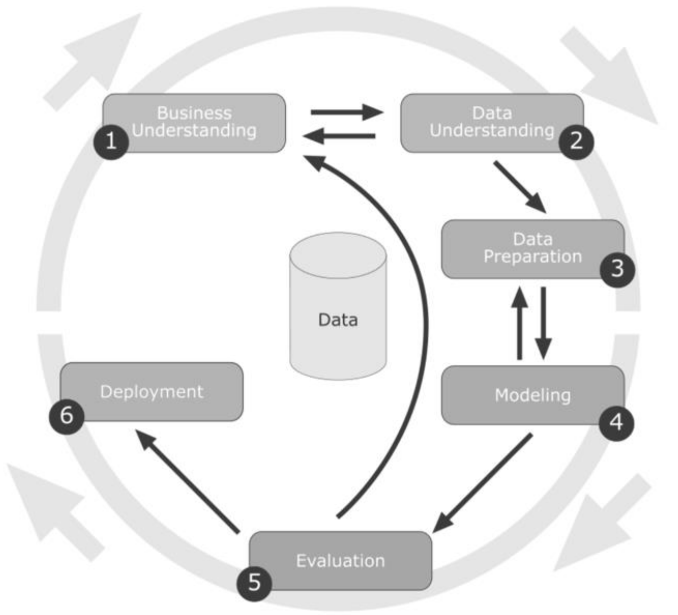

Figure 1. CRISP-DM framework.

B. Decision Trees

Breiman et al. [2] developed decision tree methods for both classification and regression. They introduced the CART (Classification And Regression Trees) methodology and the related C4.5 procedure. In this context, we focus on classification trees since our data-mining task involves binary classification.

The two key concepts in classification trees are recursive partitioning and tree pruning. Recursive partitioning involves iteratively splitting the training data into subsets and then further dividing each subset along the tree branches. For cross-sectional data, let x represent the vector of predictor variables x_1, x_2, ..., x_p, and y denote the dependent categorical variable. The predictor variables x can be continuous, ordinal, or binary. The aim of the initial step in recursive partitioning is to divide the p-dimensional space of the training data into two distinct half-spaces. This is done by selecting a variable, e.g., x_i, as the splitting one and choosing a value, s_i, as the split point for that variable. Data points in the training set with $x_i \leq s_i$ are placed in one partition, while those with $x_i > s_i$ in the other. This division forms the root node of the tree and creates two branches, which leads to the next two nodes representing the split partitions (Fig. 2).

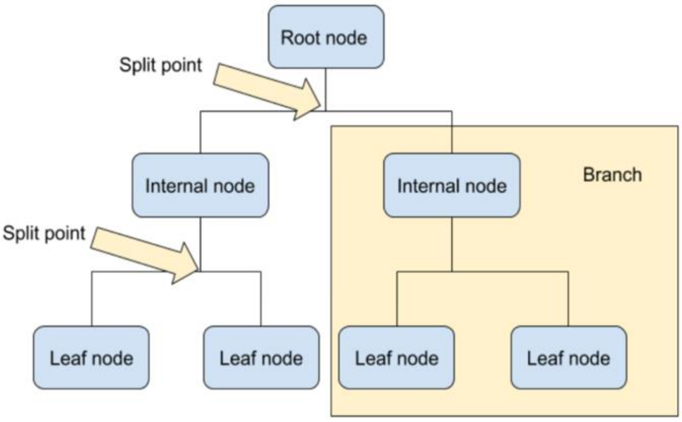

Figure 2. Example of a decision tree.

The process then continues recursively, further splitting each of the two partitions into smaller ones. Each split creates a new node in the decision tree. This process stops when the sub-partitions become as homogeneous or "pure" as possible, meaning that the data points in each partition predominantly belong to a single class. These pure nodes are known as terminal or leaf nodes. Achieving absolute purity may not always be possible due to the distribution of the data points.

Two common metrics used to measure the impurity of partitions are the *Gini impurity* and *Information Gain* (IG). Let *m* represents the number of classes of *y*. The Gini impurity for a child partition *A* resulting from a split in the tree is defined as:

$$Gini(A) = 1 - \sum_{k=1}^{m} p_k^2 = \sum_{i \neq k} p_i\, p_k \tag{1}$$

where p_k is the proportion of observations in A that belong to class *k* (*k=1, 2…, m*). For binary classifiers, the Gini impurity ranges from 0 (indicating all data points belong to the same class) to 0.5 (indicating data points are equally distributed between the two classes). For multi-class classification, the upper limit is (*m-1*)/*m*, where all *m* classes are equally represented.

The other widely used impurity measure is Information Gain, based on the concept of *entropy* from the information theory. Entropy (H) is defined as:

$$H(A) = -\sum_{k=1}^{m} p_k \log_2(p_k) \tag{2}$$

In binary classification, entropy ranges from 0, where all data points belong to a single class (a pure partition), to 1, where the classes are equally distributed. Information gain measures how much entropy is reduced after dividing a dataset based on an attribute. Constructing a decision tree involves choosing the attribute that provides the greatest information gain, meaning it results in the most uniform branches. The process starts with splitting the dataset by a predictor variable, calculating the entropy for each resulting branch, and then combining these entropies proportionally to get the total entropy for the split. The information gain is then calculated by subtracting this total entropy from the entropy before the split (3).

$$IG_{parent} = H_{parent} - \ AVERAGE(H_{children}) \tag{3}$$

To identify the best split variable at a node, the tree-building algorithm assesses each predictor variable one by one. It examines all possible splits and selects the one that results in the highest purity among the resulting sub-partitions. This optimal split becomes the root node of the tree, specifying the splitting variable and its value. The process is repeated recursively, adding sub-nodes, and growing the tree until further splits no longer enhance partition purity. The outcome is *fully grown tree*, which often achieves 100% accuracy on scoring the training dataset, but the accuracy typically decreases on the validation dataset due to overfitting. This issue can be mitigated by pruning the tree, a technique used by the C4.5 and CART algorithms. C4.5 prunes the tree using the training set, while CART prunes the fully grown tree using the validation set by removing the weakest branches that do not significantly reduce the error rate. A useful strategy to identify the optimal tree is to use the *complexity parameter* (*cp*), which will be discussed in the following sections.

Decision trees are effective for evaluating variable importance as the most significant ones appear at the top of the tree. Consequently, this classification technique doesn't require a separate variable selection before model building, as this is automatically handled by the splitter selection mechanism. Another advantage of decision trees is their non-linear nature, making them well-suited for non-linear tasks, especially when distinguishing between classes can be achieved with horizontal and/or vertical splits in the data space. Additionally, decision trees are robust to outliers.

On the downside, decision trees require a large dataset to train effectively and build a reliable model. They are also relatively computationally intensive, as growing the tree involves testing multiple potential splitter variables, split values, and sorting.

III. Dataset

The data we use are available from the IPUMS database [3] that provides access to US Census samples from 2021 to 2022. The initial dataset consists of 87 variables and 985,790 records.

Following the data understanding and preparation stages as defined by the CRISP-DM framework, the variables were reduced to 47 to address multicollinearity. The variables were then semantically categorized into three groups: 14 representing predisposing factors, 18 socio-demographic factors, and 14 socio-economic factors.

The *predisposing* category contains information about census region, population density, metropolitan or farm status, sales of farm products if any, race, Hispanic origin, cognitive difficulties, ambulatory difficulties, independent living difficulties, self-care difficulties, vision difficulties, and hearing difficulties.

The *socio-demographic* category represents householder couple type, number of families in the household, number of subfamilies in the households, multigenerational households, sex, age, marital status, citizenship status, years in the US, migration status, educational attainment, public or private school, employment status, usual hours worked per week, and language spoken.

The *socio-economic* category contains information about the total personal income, social security income, welfare income, retirement income, mortgage status, mortgage monthly payment, annual property insurance cost, monthly gross rent, annual electricity cost, annual gas cost, annual water cost, annual home heating fuel cost, any health insurance coverage, house acreage, house value, number of rooms, and number of bedrooms.

The target (dependent) variable POV_BIN indicates the presence or absence of poverty. It is imbalanced in terms of class label distribution, containing 90.2% absence of poverty (class 1) values and 9.8% presence of poverty (class 0) values. To address this imbalance, we performed under-sampling to create a balanced training partition. This technique reduces the number of samples in the majority class to equal the number of samples in the minority class.

Since we are dealing with a classification task, we can apply Chi-square test to each predictor variable to determine its relevance to the dependent one. Variables with significant p-values are considered important for predicting the target variable and can be selected for building the classification model. While p-values from the Chi-square test help identify significant associations, Cramer's V values measure the strength of these associations. The significance within each category of variables will be discussed in the following section.

IV. Experiments And Analysis

The aim of the research is to experimentally assess the factors of poverty using machine learning models based on decision trees. We used the R language and environment [7] [8] to develop three sets of decision tree models trained on data from the three categories of variables. After the CRISP-DM modelling phase, we set aside 20% of the data exclusively for testing. They contain records that preserve the original unbalanced ratio between the two classes. The remaining 80% was split 2:1 for training and validation purposes. To mitigate biases arising from lucky random dataset compositions, we used 5-fold cross-validation and averaged the results of 10 runs of each model. Performance evaluation for each model uses metrics such as prediction accuracy (Acc); sensitivity or true positive rate (TPR); specificity or true negative rate (TNR), precision, and F1 score. We also performed ROC analysis [4] on the models and calculated the area under the curve (AUC) metric to assess the overall performance of the model, regardless of the threshold chosen to map the predicted probabilities into class labels.

Table 1 presents the performance indicators for each of the three classifier models. The results indicate that, in general, the predisposing models have the worst predictive performance, which is an indication that these factors play the least significant role in predicting presence or absence of poverty. Next in significance is the group of socio-demographic factors, outperforming the predisposing ones. Evidently, the socio-economic factors are the most significant in predicting the presence or absence of poverty. It should be emphasized that this ranking evaluates the three groups of factors, without considering the role of each individual factor in the groups, something we discuss in the following sections. One of the techniques used to evaluate individual factors is the sensitivity analysis [5] that examines how the range of values of each variable affects the change in predictions. We also conducted a Variable Effect Characteristic (VEC) analysis [6] for non-binary variables, highlighting the impact of different variable values on the prediction outcomes.

Table I. Performance Metrics of Decision Tree Models Trained by Three Categories of Variables.

	Predisposing	*Socio - emographic*	*Socio - economic*
Acc	58.3%	70.3%	80.3%
AUC	0.586	0.730	0.848
TPR	52.8%	69.4%	71.5%
TNR	63.6%	71.2%	90.1%
Precision	58.1%	71.4%	88.9%
F1	55.3%	70.4%	79.3%

A. Predisposing Factors

Analysing the predisposing data, we conducted a hybrid feature assessment approach using statistical Chi-square tests combined with embedded feature assessment method based on decision trees. This allows capturing different aspects of the feature significance, leading to more detailed observations and correct conclusions. The Chi-square test using Cramer's V measure ranks how strong the relationships are between each predictor variable and the dependent variable. Results presented in Fig 3 show that the race plays a significant role in shaping poverty in the US, reflecting historical, social, and economic inequalities. Following closely in significance are factors related to physical, mobility, cognitive, and vision difficulties, along with whether an individual resides in a metropolitan or rural area.

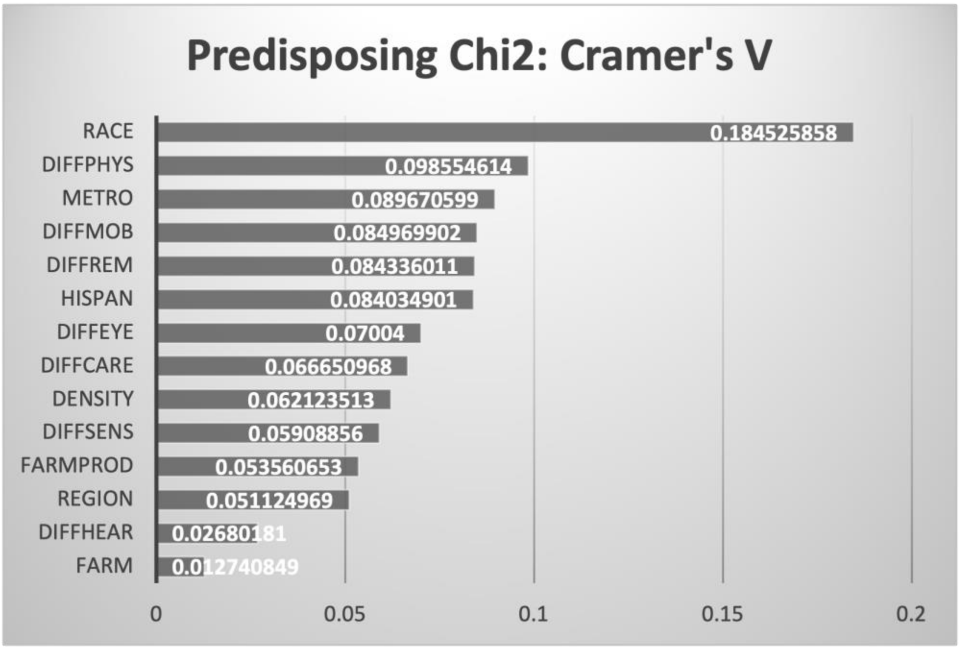

Figure 3. Chi-square test of predisposing factors.

Next experiments aimed to build decision tree model trained by the predisposing data. After constructing the fully grown tree, we optimized its structure by pruning branches, using the tree complexity parameter (cp) as a guide. The cp values represent thresholds for the minimum reduction in error needed for a split to be included in the tree. The cp value is related to the number of splits for that tree, hence its size. Larger cp values lead to smaller, simpler trees, while smaller values result in more complex trees. Fig. 4 shows how the cp affects the relative prediction error when scoring the validation set. It is apparent that cp=0.002934272 minimizes the error, indicating that the optimal tree size is 9. That tree, called by Breiman et al. [2] minimum error tree is presented in Fig. 5.

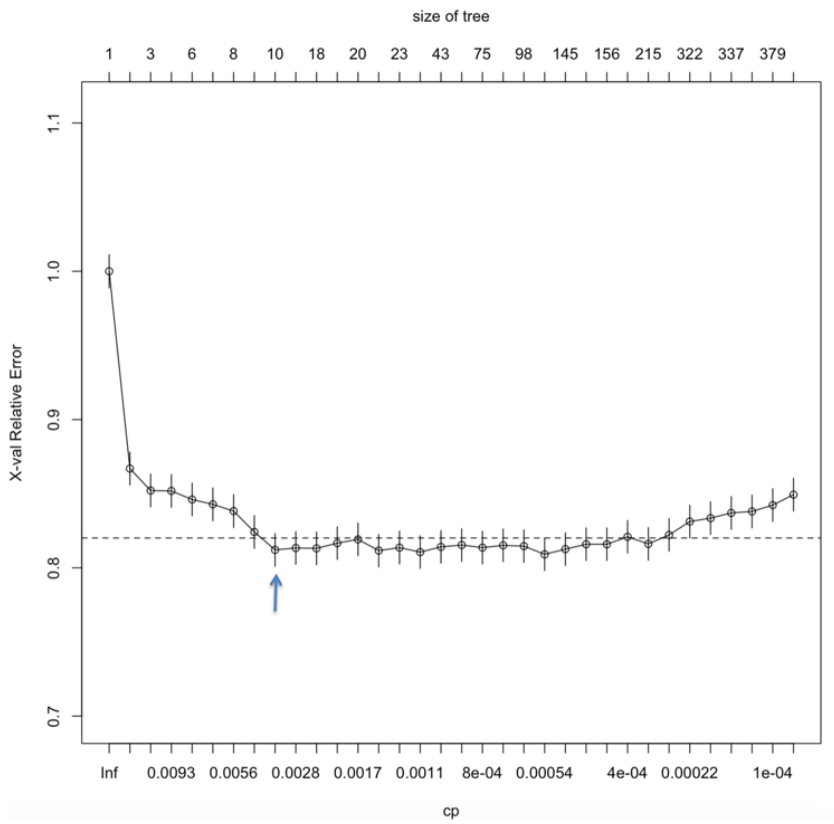

Figure 4. Pruning predisposing tree, guided by the tree complexity parameter (*cp*). Minimal error tree has size 9 and *cp*= 0.002934272

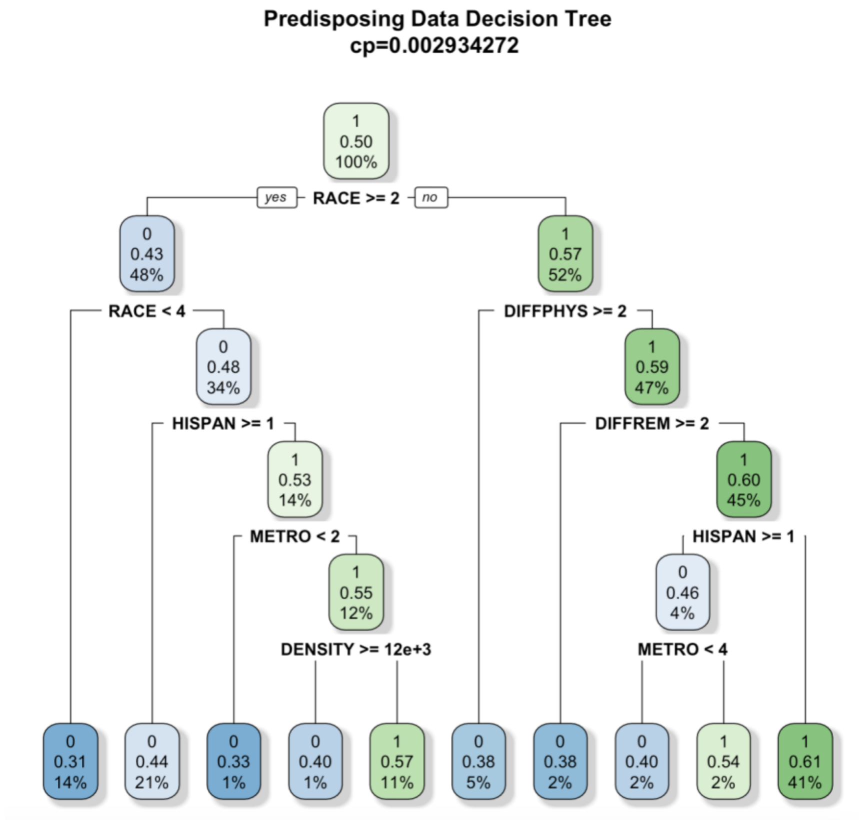

Figure 5. Minimal error predisposing tree

By analyzing the tree and its nodes, we can infer the significance of the predisposing factors. The closer a node is to the root, the more significant it is. According to the tree structure, the most significant factor is race appearing twice, once at the root and once just below it, along with Hispanic origin (HISPAN), which is related to race. Physical and cognitive difficulties also appear close to the root. Table 2 summarizes the reported factor significances, measured as a percentage, and ranked in descending order.

Table III. Variable Significance Score of the Predisposing Tree

RACE	40
HISPAN	16
DIFFPHYS	11
DENSITY	11
METRO	5
REGION	5
DIFFREM	5
DIFFCARE	3
DIFFMOB	3
FARMPROD	1

The result from this embedded method exhibits a ranking of factors that closely aligns with the Chi-square statistical method, though there are minor differences in their order.

Overall, the most significant predisposing factors appear to be related to race and minorities. This is caused by income inequality of minority groups compared to white Americans. This income gap is a significant contributor to higher poverty rates among these groups. There are other contributors related to race, such as residential segregation and social protection policy. Minority communities are more likely to live in economically disadvantaged neighbourhoods with higher crime rates, limited access to quality services like healthcare and education and underfunded programs for social welfare assistance.

The VEC analysis of the race factor, as shown in Fig. 6, also supports this finding. The figure presents values and their contributions to poverty, with the following assignments: 1 represents white; 2 - black/African American; 3 - American Indian or Alaska Native; 4 - Chinese; 5 - Japanese; 6 - other Asian; and 7 - other. Value 2, which corresponds to the black/African American group, shows the highest contribution to poverty, reflecting its status as the largest minority group.

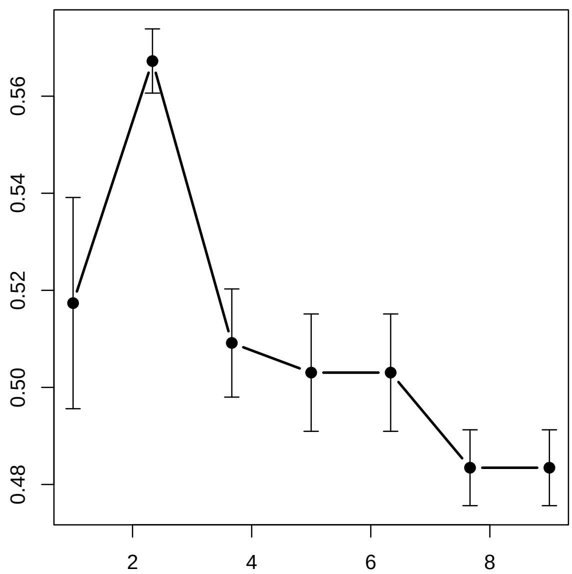

Figure 6. VEC analysis of RACE

B. Socio-Demographic Factors

In analyzing the socio-demographic data, we conducted Chi-square tests with Cramer's V measure to rank the factors according to their significance in relation to poverty. As shown in Fig. 7, there are no distinctly dominant factors contributing to poverty within this group. Instead, the top three factors are associated with the relationship type of household partners (whether they are in same-sex or opposite-sex relationships). Following these are factors related to employment, marital status, and education. Next are set of factors pertaining to the characteristics of the inhabited property.

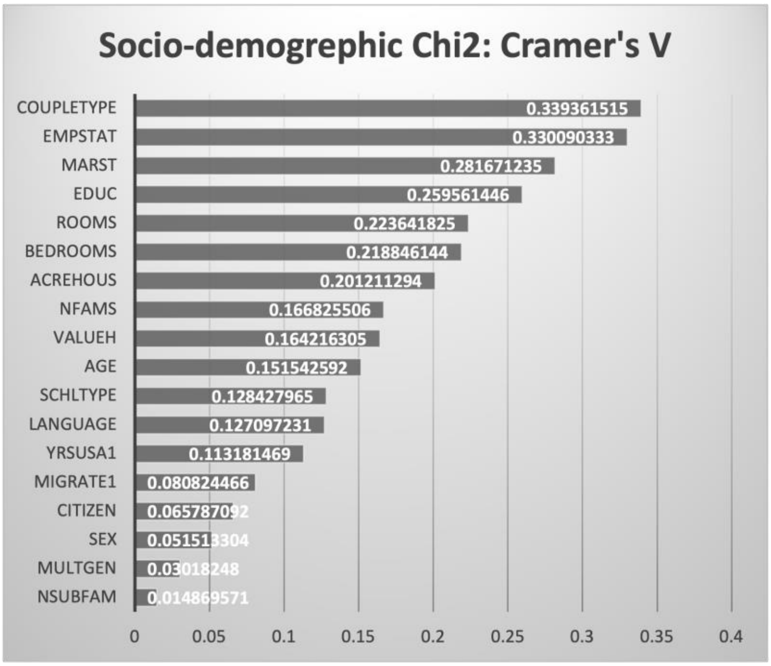

Figure 7. Chi-square test of socio-demographic factors.

Following the previous experiments, our next goal was to create an optimized decision tree-based model to derive an embedded ranking of the socio-demographic variables.

We began by training a fully-grown tree, followed by pruning to reduce its size. This was accomplished by using the complexity parameter (cp) to construct a minimal error tree, as outlined by Breiman et al. [2]. Fig. 8 indicates that a cp value of 0.001255966 achieves the minimal relative error, resulting in a minimal error tree of size 30, as shown in Fig. 9.

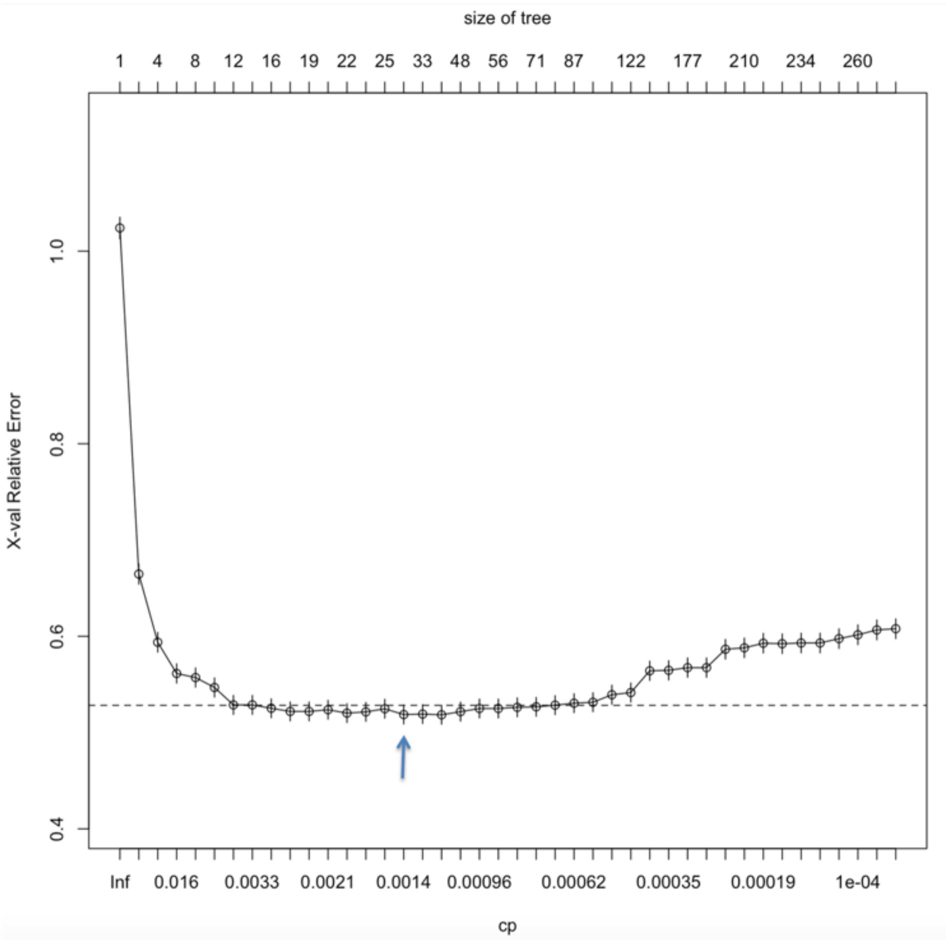

Figure 8. Pruning socio-demoraphic tree, guided by the tree complexity parameter (*cp*). Minimal error tree has size 30 and *cp*= 0.001255966

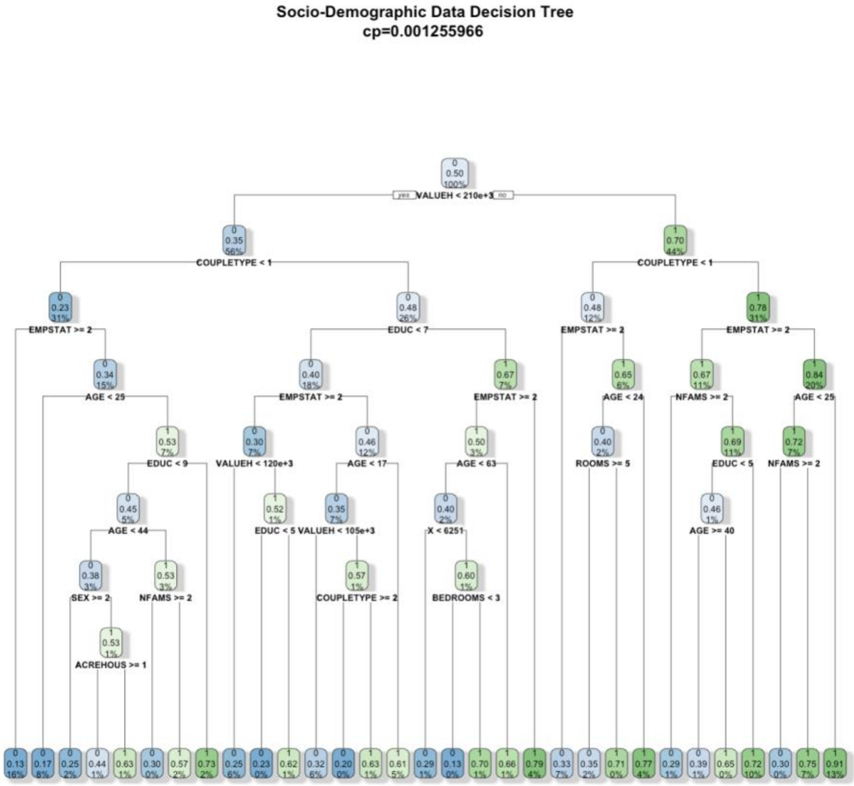

Figure 9. Minimal error socio-demographic tree.

The nodes nearest to the root of the tree represent the most significant factors affecting poverty, which include property value, type of couple relationship, employment status, education, and age. Table 3 provides further details on their significance as reported by the model. It can be noticed that the ranking of poverty factors made by the tree somewhat changes that made by the Chi-square test. In particular, the most significant variable with 20% influence is the property value, which is also root of the tree. It can serve as an important indicator of poverty when looking at economic and social conditions in a given community or region. High property values often indicate higher wealth among homeowners, while low property values may reflect limited wealth accumulation, making it difficult for residents to build financial security. Next in significance is the employment status (12%), which has a significant impact on an individual's income, financial stability, and access to basic resources. Equally significant (12%) is the education, which has an impact on an individual's capabilities, economic prospects, and overall quality of life. This includes access to job, career opportunities, and the pay gap.

Table IIIII. Variable Significance Score of the Socio-Demographic Tree

VALUEH	20
EMPSTAT	12
EDUC	12
COUPLETYPE	11
MARST	10
AGE	10
ROOMS	5
BEDROOMS	5
SCHLTYPE	5
MULTGEN	4
ACREHOUS	4
NFAMS	1
LANGUAGE	1

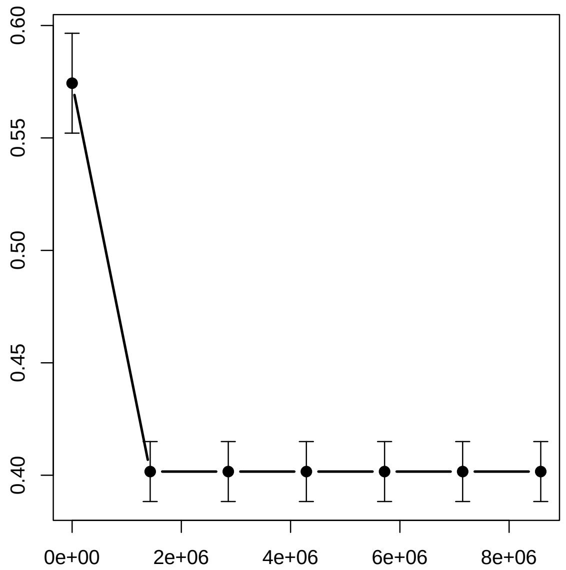

Figure 10. VEC analysis of VALUEH.

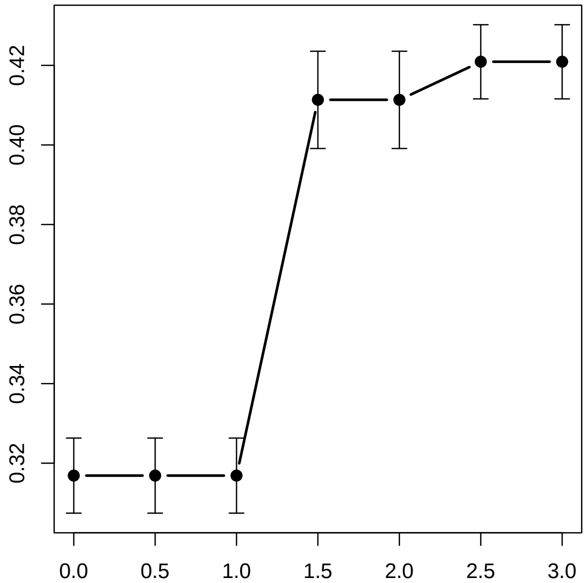

Figure 11. VEC analysis of EMPSTATUS.

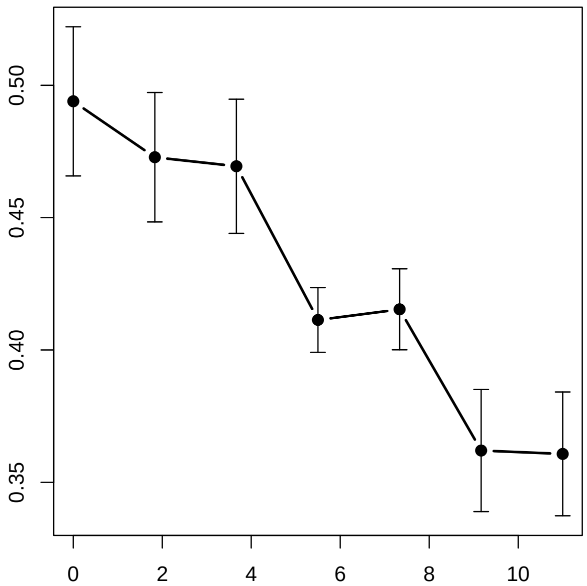

Figure 12. VEC analysis of EDUC.

Performing a VEC analysis of the property value variable (Fig. 10), it is observed that poverty peaks among the cheapest properties and decreases with higher property values, becoming virtually non-existent for properties valued above $600,000.

The VEC analysis of the employment status (Fig. 11) indicates that poverty levels are low for those classified as value 1 (employed), higher for value 2 (currently unemployed but seeking work), and highest for value 3 (not in the labour force).

The VEC applied to the education variable (Fig. 12) shows that value 0 (no education) is associated with the highest poverty, which gradually decreases with values from 1 to 10, reflecting the degree of education.

C. Socio-Economic Factors

To analyze the socio-economic factors, we used the same method as previously described: beginning with a Chi-square test and Cramer's V measure, then constructing a decision tree model to assess the significance of the variables within this group. These variables generally pertain to household or individual income and expenditure. The Chi-square test ranks the socio-economic variables as shown in Fig. 13. Apparently, no single leading factor emerges distinctly; instead, several variables show similar levels of importance. The most significant are the usual working hours weekly, followed by mortgage amount, property insurance, total income, and others.

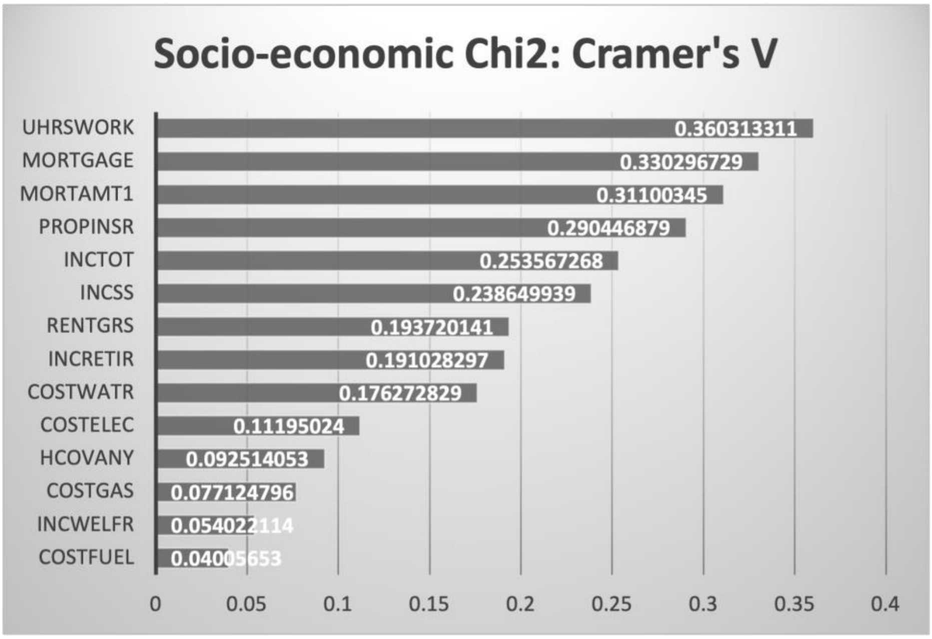

Figure 13. Chi-square test of socio-economic factors.

The decision tree was first trained as a fully grown tree and then pruned to achieve minimal prediction error on the validation dataset. The pruning process was guided by the complexity parameter (cp), as illustrated in Fig. 14. With a cp of 0.001520913, the pruned tree, which has 12 internal nodes, minimizes the error. This represents the minimum error tree (Fig. 15), as outlined by Breiman et al. [2].

Analyzing the tree's structure provides insight into the significance of socio-economic factors. This is further illustrated by Table 4, which displays the model's ranking of these factors. Notably, half of the tree's nodes, six in total, feature the variable INCTOT, which represents the household's total income. One of these nodes is even the root itself. It's clear that, according to the model, this is the most significant factor as it directly influences an individual's or household's ability to meet basic needs and achieve financial stability. Insufficient income often forces people to choose between essential needs like housing, food, healthcare, and education, leading to poor living conditions, malnutrition, lack of medical care, and other hardships associated with poverty.

Another significant factor identified by the model as influencing poverty is UHRSWORK—typical hours worked per week. This factor directly affects an individual's or household's income levels, job stability, and access to benefits. Insufficient working hours, whether due to part-time work, underemployment, or job instability, can result in lower income and increased vulnerability to poverty.

Other factors of decreasing importance are amount of mortgage, income from social services or retirement, property insurance, etc.

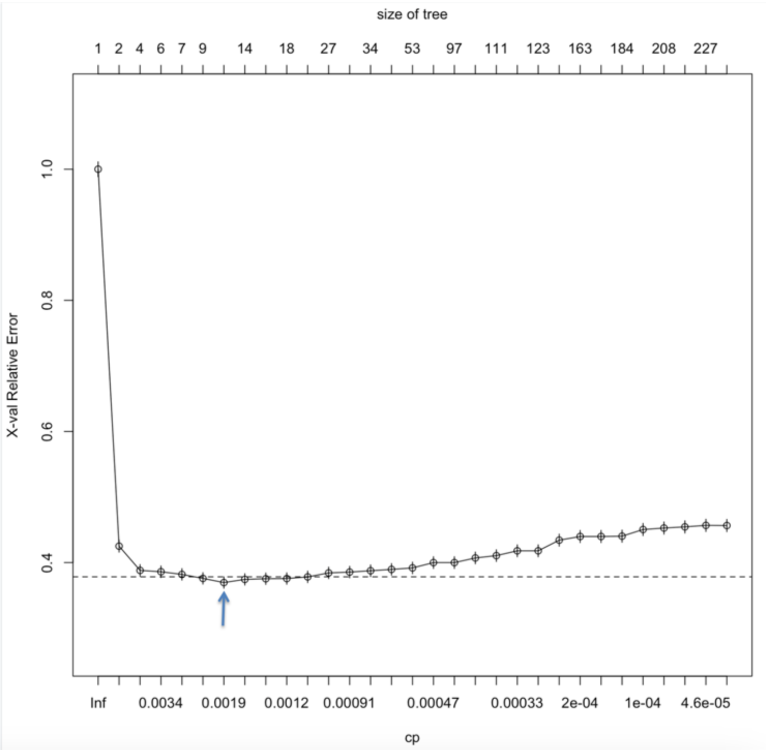

Figure 14. Pruning socio-economic tree, guided by the tree complexity parameter (*cp*). Minimal error tree has size 12 and *cp*= 0.001520913

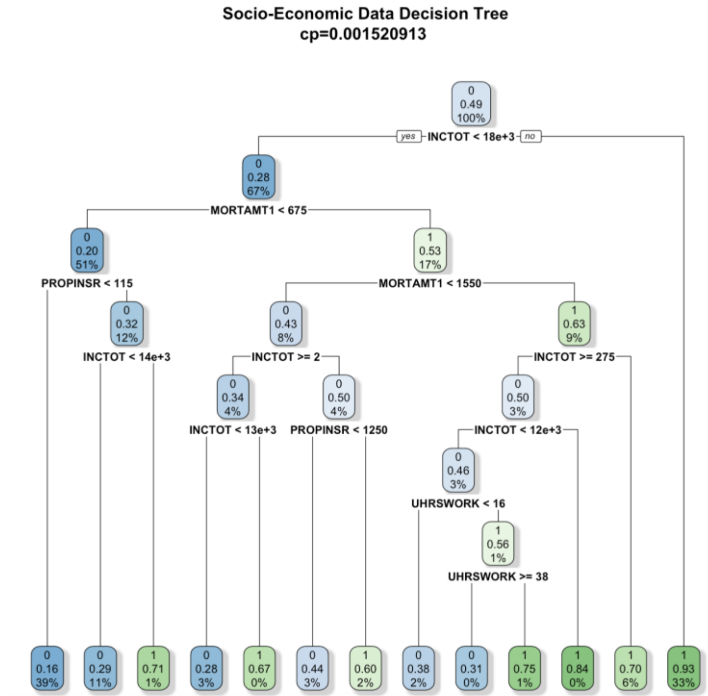

Figure 15. Minimal error socio-ecnomic tree.

Table IVV. Variable Significance Score of the Socio-Economic Tree

INCTOT	47
UHRSWORK	22
MORTAMT1	8
INCSS	8
MORTGAGE	6
INCRETIR	6
PROPINSR	3

The VEC analysis of the total income variable (Fig. 16) reveals that poverty is most prevalent among the lowest income levels and decreases as income increases. Poverty is virtually non-existent for households with annual incomes above approximately $180,000.

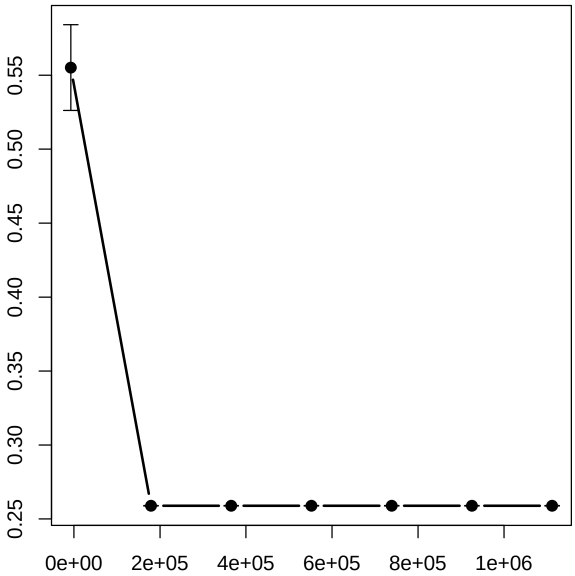

Figure 16. VEC analysis of EDUC.

V. Conclusions

This article presents a study that uses machine learning techniques to build decision tree models aimed at exploring the factors influencing poverty. The models were trained using data collected from surveys conducted in the United States during the 2021-2022 period. The data includes variables grouped into three categories: predisposing, socio-demographic, and socio-economic. By examining the models and the role of these three groups of variables, we draw conclusions about the factors associated with these variables, specifically their significance in determining the presence or absence of poverty. These conclusions are empirically derived, data-based, and therefore practically valid.

The study's results indicate that socio-economic factors have a dominant influence on poverty, as they affect an individual's or household's economic well-being and their ability to meet basic needs such as housing, food, healthcare, and education. The results show that socio-demographic factors are next in significance, highlighting the importance of characteristics like employment status, marital status, education, age, and others.

The models also reveal that predisposing factors have the least significance concerning poverty. Among these are characteristics such as race, physical and cognitive difficulties, regions, social environment, and others. Although these factors are the least significant, they still impact poverty and should not be ignored.

This study aims to demonstrate the potential of modeling and machine learning in extracting socially significant information and knowledge from available data, which could help in developing appropriate priorities and policies through informed decision-making.

References

[1] Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, CRISP-DM 1.0 - Step-by-step data mining guide, CRISP-DM Consortium, 2000.

[2] Breiman, L., Friedman, J., Olshen, R.A., Stone, C. Classification and Regression Trees, Wadsworth, Belmont, CA, 1984

[3] Ruggles S., Flood, S., Sobek M., Backman D., Chen A., Cooper G., Richards S., Rodgers R., and Schouweiler M. *IPUMS USA: Version 15.0* [dataset]. Minneapolis, MN: IPUMS, 2024.

[4] Fawcett, T., An introduction to ROC analysis, Pattern Recognition Letters 27(8), pp. 861–874., 2005.

[5] Kewley, R., Embrechts, M., Breneman, C. Data strip mining for the virtual design of pharmaceuticals with neural networks. IEEE Transactions on Neural Networks, vol. 11 (3), pp. 668 – 679, 2000.

[6] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J., Modeling wine preferences by data mining from physicochemical properties, Decision Support Systems, 47(4), 2009

[7] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2023.

[8] R Core Team, "R: A language and environment for statistical computing." [Online] Available: http://www.R-project.org/ (URL).

[9] Creamer, J., Shrider, E., Burns, K., and Chen, F., U.S. Census Bureau, Current Population Reports, P60-277, Poverty in the United States: 2021, U.S. Government Publishing Office, Washington, DC, September 2022.

[10] United States Census Bureau. How the Census Bureau measures poverty. https://www.census.gov/topics/income-poverty/poverty/guidance/poverty-measures.html , 2021.

[11] Thiede, B.C. and Brooks, M.M., Child poverty across immigrant generations in the United States, 1993–2016: Evidence using the official and supplemental poverty measures. *Demographic Research*, *39*, pp.1065-1080, 2018.

[12] Sharma, M., 2023. Poverty and gender: Determinants of female-and male-headed households with children in poverty in the USA, 2019. *Sustainability*, *15*(9), p.7602.

[13] Hawkins, J., The rise of young adult poverty in the US. Issue Brief, 2019.

[14] Davila, A., Mora, M.T. and Hales, A.D., Income, earnings, and poverty: A portrait of inequality among Latinos/as in the United States. In Latinas/os in the United States: Changing the face of America (pp. 181-195). Boston, MA: Springer US, 2008.

[15] Ortega, L., 2022. The Effect of Race/Ethnicity, Health Status and Socioeconomic Variables on Poverty in the US. The Park Place Economist, 29(1), p.7, 2022

[16] Lichter, D., Zhenchao, Q. and Crowley, M., Race and poverty: Divergent fortunes of America's children. *Focus*, *24*(3), pp.8-16, 2006.

[17] Glaeser, E.L., Resseger, M. and Tobio, K., Urban inequality. *Justice for All*, pp.98-121, 2015.