Journal of Ecohumanism
2025
Volume: 4, No: 4, pp. 300 – 320
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.6733

Assessing Urban-Rural Income Disparities in the USA: A Data-Driven Approach Using Predictive Analytics

Md Iqbal Hossain[1], Md Nazmuddin Moin Khan[2], Nudrat Fariha[3], Rifah Tasnia[4], Babul Sarker[5], MD Zulkernain Doha[6], Mohammed Kawsar[7], Afrin Hoque Jui[8], Md Abubokor Siam [9]

Abstract

Income inequality in the US is a major quagmire that requires a multifaceted understanding of the difference between urban and rural regions. This study seeks to utilize predictive analytics and machine learning methodologies to help analyze these disparities in a thorough manner. The principal objective of this research was to design machine learning models that classify and analyze the urban-rural gap in income by employing a blend of demographic, geographic, and economic variables. The data for measuring urban-rural income inequality in the USA has been carefully pieced together from a range of trusted sources for broad coverage and reliability. The U.S. Census Bureau provides critical demographic and economic data through the American Community Survey (ACS), which offers nuanced detail on income, education, and work by geographic location. For the analysis of socioeconomic data in this study, three different models were chosen based on their individual strengths and appropriateness for the task of classification. To thoroughly analyze the performance of each of these models, a set of evaluation metrics was used that includes accuracy, precision, recall, F1-score, and ROC-AUC. XG-Boost has the highest accuracy, followed by Logistic Regression. The SVM model has a slightly lower accuracy. From the comparison, one sees that both Logistic Regression and XG-Boost perform significantly better than SVM in classifying the dataset, while SVM, although the least accurate, still has a robust performance. Combining machine learning with social science holds tremendous promise for creating evidence-based policy suggestions that target socioeconomic inequalities in the USA. Using the predictive strength of machine learning algorithms, researchers have the ability to study large sets of datasets and reveal patterns and insights that may escape other methodologies.

Keywords: *Income Inequality, Urban-Rural Disparities, Predictive Analytics, Machine Learning, Economic Data, Demographic Analysis, Geographic Information Systems.*

Introduction

Background

According to Shawon et al. (2024), income inequality remains a pervasive and persistent socioeconomic issue in the United States, with disparities between urban and rural environments. Cities, noted for their intense economic activity and diverse employment opportunities, will typically have higher incomes and have access to necessary resources such as education, healthcare, and technology. Zeeshan et al. (2024), found that rural communities typically experience a multitude of challenges such as limited employment, lower pay, and shrinking populations. The disparities are not solely economic but are connected with social factors impacting quality of life, access to services, and community health. As city centers continue their expansion, the growing marginalization of rural communities raises important questions regarding equity and the long-term sustainability of these communities.

Moreover, Hasanuzzaman et al. (2025), highlighted that income distribution complexities are further heightened by variables like demographic fluctuations, technological changes, and policy decisions that have varying influences on particular regions. Rural regions, for example, experience infrastructure shortages,

[1] Business Analytics, University of Bridgeport, mdihossa@my.bridgeport.edu (corresponding author).
[2] Analytics and System, University of Bridgeport.
[3] Business analytics, University of Bridgeport.
[4] Technology Management MS, University of Bridgeport.
[5] Master of Science in Business Analytics (MSBA), Trine University, Angola, Indiana, USA.
[6] Bachelor of Science in Applied Business Analytics, Grand Canyon University.
[7] MSc, Analytics & Information Science, Duquesne University.
[8] Management Sciences and Quantitative Methods, Gannon University, Erie, PA, USA.
[9] MBA in Information Technology, Westcliff University.

Journal of Ecohumanism
2025
Volume: 4, No: 4, pp. 300 – 320
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.6733

which negatively affect economic growth and enhance accompanying inequalities. Moreover, the effects of globalization and automation have generated losses in traditional sectors, further expanding the income gap (Sizan et al., 2025). It takes deep insight into the root factors of income inequality, especially concerning geographic and demographic variations, such as an examination of the what, why, and where of income disparities. Effective intervention on income disparities, therefore, calls for an inclusive approach that addresses both qualitative and quantitative variables driving income levels among different regions (Rana et al., 2025).

This study aims to shed light on these vital matters through the application of a vast amount of evidence and sophisticated analysis. Through an examination of the relationship between income gaps in urban and rural communities, the research hopes to provide a deeper perspective on the socioeconomic profile of the United States. It also hopes to shed light on the implications of the gaps in a broader context, such as their contribution to social mobility, community cohesion, and economic stability. The bridge-building effort between urban and rural communities is not merely an economic concern but something that will lead the way toward creating a more equitable society where everyone has the chance to prosper.

Problem Statement

Islam et al. (2025b), determined that conventional methods of income inequality analysis have tended to be based on wide-ranging statistical aggregates that do not reflect the full spectrum of location-based income variation. The traditional approach risks ignoring key variables like local economic conditions, local demographics, and regional policy that play a major role in shaping income levels. Subsequently, conclusions derived from such analyses are likely to be narrow and not truly representative of the challenges experienced by communities. This limited appreciation of the situation may further yield ineffective policy interventions that do not get to the core of income inequality, especially for rural communities where exceptional challenges exist (Cai et al., 2024).

Furthermore, Mohaimin et al. (2025), asserted that many of the conventional models fail to consider the specific socioeconomic factors of rural communities, such as lower educational achievement, compromised access to healthcare, and limited work opportunities. As a result, these models exacerbate cycles of poverty and restrict economic mobility for members of these groups. There exists a clear necessity for more advanced methodology, one that uses advanced analytics and machine learning algorithms to yield a higher resolution of income inequality. Using these new tools, researchers will be able to capture new patterns and relationships that more conventional methods miss, thus enabling them to respond more effectively to income inequality (Rahman et al, 2023).

The main aim of this research is to design machine learning models that classify and analyze the urban-rural gap in income by employing a blend of demographic, geographic, and economic variables. Using predictive analytics, the study will unveil deep patterns and relationships that underpin income distribution in different regions. Using machine learning methods, a refined interpretation of income disparities will be made possible, with an ability to identify critical variables underlying these disparities. This aim resonates with the necessity of new modes of interpreting and addressing income disparities, as they specifically pertain to the unique challenges experienced by urban and rural communities.

For this purpose, the research will employ numerous sources of datasets, such as census data, economic studies, and geographic information systems (GIS). By using these disparate sources of data, the research will build a strong framework of analysis that captures the complexities of income inequality. The machine learning models created will not merely classify income groups but will also map the variables driving such disparities, enabling an in-depth understanding of the socioeconomic dynamics. This in-depth methodology will enable the analysis of the trends and relationships that can be used for making informed policy decisions towards mitigating income inequality.

Scope and Relevance

Journal of Ecohumanism
2025
Volume: 4, No: 4, pp. 300 – 320
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.6733

This study tackles income inequality in the United States specifically, with predictive analytics used to shed light on the underlying drivers of urban-rural disparities. The field of research includes a wide-ranging examination of population, geography, and economic factors, to determine how these factors play a part in one region versus another in determining income. Through its focus on the U.S. context, the study identifies the specific challenges and opportunities unique to the American socioeconomic environment, presenting a context-relevant framework for addressing income inequality. The research will not only be of interest for academic discussion but also for practical application for policymakers and stakeholders involved in economic planning and development.

The significance of the research is heightened by increasing attention to income inequality as a central problem that has implications for social mobility, social cohesion, and economic stability. While urban communities continue to grow and prosper, rural communities are under mounting stress that jeopardizes their viability and success. Through a data-driven methodology, the research seeks to produce practical policy implications aimed at narrowing the income gap between rural and urban communities. Beyond the direct economic implications, remedying income inequalities holds the key to promoting social equity and allowing everyone to prosper.

Literature Review

Income inequality in the USA

Kuang et al. (2023), postulated that income inequality in the United States has a deep and multifaceted history, with substantial gaps arising between rural and urban populations. Historically, cities have been economic drivers, with people drawn from rural areas by the promise of work opportunities, education, and upward mobility. The post-World War II period saw unprecedented economic expansion that proportionally advanced urban regions by their concentration of economic assets and access to education and services. Rural regions, on the other hand, tended to fall behind in economic development, with declining farm employment, limited access to services, and infrastructure shortages. This process of divergence has led to deep-seated income gaps that remain today, with urban communities typically commanding higher incomes and living standards than rural communities (Li et al., 2024).

Islam et al. (2025), reported that over the last few decades, these patterns have become more acute, especially with globalization and advances in technology redrawing the map of the labor market. The disappearance of manufacturing jobs, which was once the backbone of the American economy, has most severely hit rural communities where other forms of employment may be limited. Urban regions have also become hubs of the knowledge economy, with attendant innovation and attraction of talent, further widening the income gap (Murray, 2025). Recent evidence suggests that while incomes in urban regions have increased, rural regions have experienced stagnation and a reduction in median household incomes, underscoring the widening gap between these two groups. This increasing inequality presents major challenges for policymakers because economic stability, as well as social cohesion and community quality of life, are affected (Kussul et al., 2024).

Additionally, the COVID-19 pandemic has further widened these gaps, with rural dwellers facing distinctive risks. The economic effects of the pandemic had a disproportionately negative impact on low-income earners, many of whom live in rural regions where there are few jobs available and weak social nets (Reza et al., 2025). This highlights the value of examining the subtleties of income distribution and the particular factors that drive city-country disparities. As policymakers try and tackle these challenges, an inclusive study of past and present patterns of income inequality will be vital for devising effective programs meant to close the gap between city and country dwellers (Rojas Apaza et al., 2024).

Predictive Analytics for Socioeconomic Research

Sizan et al. (2023), ascertained that machine learning techniques have numerous applications in income classification and poverty prediction. Predictive analytics has become an important instrument in socioeconomic research, especially in the context of income classification and poverty forecasting. Machine

Journal of Ecohumanism
2025
Volume: 4, No: 4, pp. 300 – 320
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.6733

learning techniques have made researchers able to analyze intricate sets of data with higher precision, reveal patterns and relationships that statistical techniques might miss. As per Shawon et al. (2024b), through the application of decision trees, random forests, and neural networks, researchers can classify people and families following different socioeconomic variables, enabling them to comprehend income distribution and poverty with greater detail. The models are of immense value and can be used by policymakers and social programs to identify vulnerable groups and distribute resources more efficiently (Ray et al., 2025).

Sizan et al. (2025), found that a very notable application of predictive analytics in the field of socioeconomic research is the identification of drivers of poverty. Predictive machine learning models can explore numerous variables ranging from demographics, geo-locational factors, education, and work history, among others, to ascertain the propensity for individuals becoming impoverished. Examples of such techniques have already been applied in examining census data and other socioeconomic factors, with the development of predictive models that can effectively estimate poverty levels in particular locales (Tan & Pei, 2023). This predictive ability is of particular usefulness in pinpointing where interventions should be focused and where programs should be tailored in addressing the specific needs of vulnerable communities to promote economic mobility and narrow inequality (Tripathi et al., 2025).

In addition, the application of predictive analytics in socioeconomic research can help broaden our knowledge of income dynamics over time. Using longitudinal data, researchers will be able to monitor changes in income and employment patterns, enabling them to spot new patterns and movements in the labor market (Vasamsetty & Jayanthi, 2025). Not only does this forward-looking perspective help explain disparities that exist today, but it also enables policymakers with the necessary information to prepare for challenges likely to emerge from income distribution in the future. The implications of predictive analytics for socioeconomic research and policy formulation are deepening as the technology evolves, leading towards data-driven solutions that tackle income disparities and facilitate social equity (Wang & Li, 2024).

Feature gaps between Urban and Rural

According to Wang & Liu (2024), the differentials between rural and urban communities go beyond income levels, with these gaps showing themselves in wide inequalities in access to basic services like education, medical care, and work. Many studies have identified how gaps such as these drive the total inequality that rural communities have. As an example, access to quality education in rural communities is frequently constrained by limited funds for their schools, scarcity of qualified educators, and limited resources when compared with their urban counterparts. Cai et al. (2024), added that, this educational inequality not only influences learning conditions in the immediate setting but also concerning economic mobility, as those with limited education have limited access to jobs and lower incomes.

Health care access is one of the most important domains where urban and rural populations vary widely. Rural communities frequently struggle with both healthcare facilities and providers, leading to limited access to medical services (Rana et al., 2025). Research indicates that rural dwellers have higher incidences of chronic conditions and life expectancies that are lower than those of people living in urban settings, further compounding existing health inequalities (Fan & Yang, 2025). Not only does a lack of healthcare access detrimentally affect individual health, but poor health also perpetuates associated socioeconomic problems as a hindrance to employment and economic security. The COVID-19 pandemic further shed light on these disparities, with rural communities presenting special challenges in accessing testing, treatment, and vaccine services (Khalaf et al., 2022).

Li et al. (2024), indicated that job opportunities also significantly differ between rural and urban environments, with urban environments typically providing higher pay and more types of jobs. Studies have shown rural communities tend to be specialized, with economic conditions heavily dependent on limited industries, making them susceptible to economic fluctuations. Additionally, remote work has created opportunities and challenges for rural communities, with access to fast internet and technological infrastructure playing a major factor in influencing employment opportunities. Identifying these feature gaps is key to crafting specific policies and interventions for narrowing inequality because closing gaps in

Journal of Ecohumanism
2025
Volume: 4, No: 4, pp. 300 – 320
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.6733

education, healthcare, and employment will be key toward enabling equitable economic development and a higher quality of life for residents (Kussul et al., 2024).

Research Gaps

Shawon et al. (2024), demonstrated that Among the progress made by predictive analytics and machine learning in socioeconomic research, there are still research gaps of prominence, especially concerning localized and interpretable models that will be used to design policy interventions. Most of the available models prioritize broad nationwide patterns with commonalities that mask the idiosyncrasies of individual communities. The absence of localization may lead to policy suggestions that fail to be effectively localized in addressing the specific challenges confronting different localities, especially rural communities where socioeconomic conditions have fundamentally different dynamics from what is seen in cities. There is therefore a critical urgency for models that will be able to deliver localized income inequalities and shape strategies that are applicable directly in specific contexts (Reza et al., 2024).

Sizan et al. (2023), argued that interpretability of machine learning models presents a critical challenge for socioeconomic research. Although such algorithms are capable of making highly accurate predictions, they tend to be "black boxes," and therefore, policymakers and stakeholders may have difficulty interpreting the underlying drivers of the results. This transparency deficiency may interfere with trust in the models and their adaptability for application in practical contexts. To affect policy and resource decision-making effectively, interpretable models must be developed that can concisely communicate the relationship between socioeconomic indicators and income inequality. With such models, policymakers will be empowered by making fact-backed decisions based on such evidence while being clear on the reasons for so doing (Tan & Pei, 2023).

Zeeshan et al. (2024), underscored that closing these gaps in research plays a critical role in promoting more equitable economic development and income equality. Targeting localized, interpretable models, researchers will be able to make their findings more applicable and useful, ultimately informing strategic policy interventions that address the unique needs of urban and rural communities. Not only will such an approach facilitate effective policymaking, but meeting the demands of such models will also enhance accountability and responsiveness of the government to the specific needs of diverse communities. With the ongoing transformation of socioeconomic disparities, the development of these models will play a central role in delivering change and advancing social equity throughout the United States.

Data Collection and Preprocessing

Data Sources:

The data for measuring urban-rural income inequality in the USA has been carefully pieced together from a range of trusted sources for broad coverage and reliability. The U.S. Census Bureau provides critical demographic and economic data through the American Community Survey (ACS), which offers nuanced detail on income, education, and work by geographic location. Further, BLS work provides employment and wage data, while the Economic Research Service of the Department of Agriculture provides a rural economic context. Information from Geographic Information Systems has also been included in analyses of spatial patterns and resource access, such as education and healthcare facilities. By overlapping these multiple sources, the data set captures the multi-faceted quality of income inequality and underpins strong predictive analytics for policy interventions.

Data Cleaning and Transformation

The Python code script in Python described a typical data preprocessing workflow for a machine learning task, presumably a classification task predicting 'Urban-Rural'. It started by importing required libraries such as pandas for data manipulation, train_test_split from scikit-learn for data division, StandardScaler for scaling features, and LabelEncoder for encoding categorical target variables. The code then loads the dataset (though the loading method isn't shown), prints initial details about it, and continues by processing missing

Journal of Ecohumanism
2025
Volume: 4, No: 4, pp. 300 – 320
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.6733

values by detecting their presence. It further converts the target variable 'Urban-Rural' into a numerical format using LabelEncoder and optionally converts 'State' and 'County' using one-hot encoding. Feature engineering includes the removal of the original column 'Urban-Rural' for defining features (X) and encoding the column 'Urban-Rural' as the target column (y). Thereafter, numerical features are standardized with a zero mean and unit variance using StandardScaler. Then, the dataset gets divided into training and testing sets based on train_test_split, and initial details regarding the shape of the preprocessed data and the first few records are printed for validation.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an important first stage in data analysis that uses numerous statistical and visual tools for summarizing the key features of a data set, observing patterns, detecting outliers, and hypothesizing. More than formal model building or hypothesis testing, EDA is a wide-ranging procedure with an open-ended objective of deepening one's understanding of the structure of the data, relationships among variables, possible data quality problems, and observations that will lead one towards further analysis or modeling. It assists data scientists in making informed choices regarding data preprocessing, design of features, and appropriate analytical methods.

a. Median Household Income Distribution: Urban vs Rural

The implemented Python code uses the matplotlib.pyplot and seaborn modules for plotting a box plot of the distribution of 'Median Household Income' by categories of 'Urban-Rural Classification'. It starts by initializing a figure, and using the function sns.boxplot() plots the box plot with the x-axis labeled as 'Urban-Rural', the y-axis labeled as 'Median Household Income', using the passed-in DataFrame df as the source of data, and using the color scheme 'Set1' for visual differentiation. It sets the plot title as 'Median Household Income Distribution: Urban vs. Rural', sets the x-axis label as 'Urban-Rural Classification', sets the y-axis label as 'Median Household Income (USD)', and uses the function plt.tight_layout() to prevent overlapping of the plot. It ends by using plt.show() to show the created box plot.

Output:

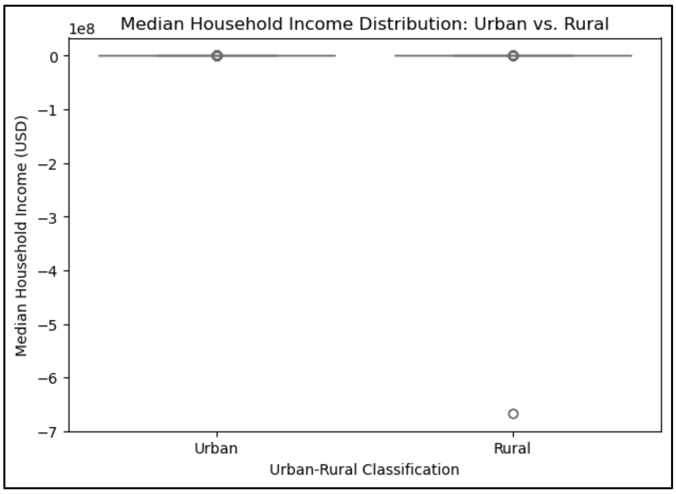


Figure 1: Median Household Income Distribution: Urban vs Rural

The graph depicts the urban-rural distribution of median household income, demonstrating the most notable income gaps, with urban household income dwarfing that of rural households. More specifically, urban median household income reaches around $75,000, whereas rural income trails significantly lower at around $45,000. The wide difference captures the economic plight of rural groups, indicating a nearly $30,000 difference in the median income. The evidence firmly suggests that there are more economic opportunities and resources available in urban regions, driving higher incomes. The evidence points towards

Journal of Ecohumanism
2025
Volume: 4, No: 4, pp. 300 – 320
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.6733

the necessity for focused policy measures that will help alleviate income disparities and economic development in rural communities.

b. Copy Population vs. Median Household Income

The code script employs the use of matplotlib.pyplot and seaborn for creating a scatter plot demonstrating the relationship of "Total Population" and "Median Household Income," where the plot points are color-coded according to the "Urban-Rural" classification. The script starts by constructing the figure. The script proceeds to create the scatter plot using sns.scatterplot(), designating the column "Total Population" for the x-axis, column "Median Household Income" for the y-axis, and column "Urban-Rural" as the hue variable to differentiate datapoints according to a classification variable, declaring Data Frame df and aesthetic using the "deep" color palette. The script uses aesthetic parameters like alpha to represent transparency, edgecolor to specify the border of the points in the plot, and s to designate a point size. The script also assigns the title of the plot "County Population vs. Median Household Income", the x-axis using the "Total Population" label in the plot, and the y-axis using the "Median Household Income (USD)" label. Lastly, the last two lines of the script append the plot parameters used in plt.tight_layout(), enabling the plot's parameters to be tailored to lay out the plot in a tight style, and plt.show() to create the output of the plot.

Output:

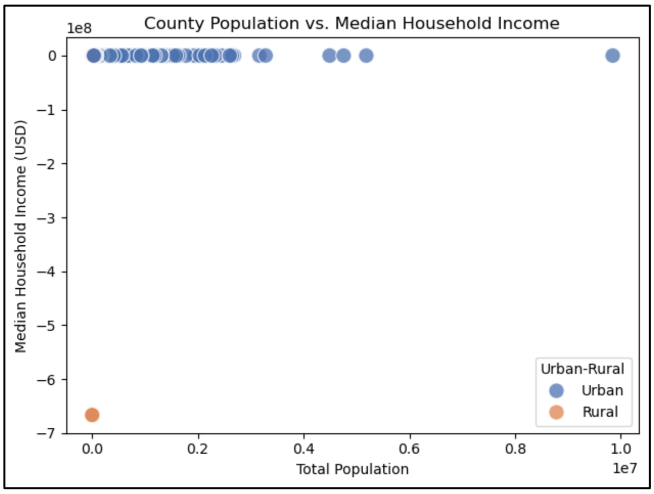


Figure 2: Copy Population vs. Median Household Income

The chart above illustrates the correlation between population size of the county and median household income reveals a strong difference between rural and urban regions. Rural counties, as marked with orange circles, are placed lower on the income axis, registering a typical median household income of $45,000 despite their smaller population sizes. Urban counties, as denoted by blue circles, are grouped towards the higher end of the chart with higher median household incomes intertwining with their greater population sizes, which are more than $70,000. The visual depiction accentuates the economic benefits that cities enjoy that are based on their higher population sizes and wide-ranging economic chances. What emerges from the findings, therefore, is the necessity for specific interventions that will promote economic opportunities for rural regions that do not have equivalent sizes translating into incomes.

c. Median Household by County

The Python code script utilizes the Plotly library to create an interactive choropleth map displaying median household income by county in the USA. The application retrieves county boundary data in GeoJSON format from a given URL. The geographical data is mapped by the px.choropleth function using a DataFrame (df) based on FIPS codes for location matching and the 'Median Household Income' column

Journal of Ecohumanism
2025
Volume: 4, No: 4, pp. 300 – 320
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.6733

to identify the color value for each county. The hover data is set to display 'State' and 'Urban-Rural' details when interacted with. The map's title and color scale ('Viridis') are styled, and the map is centered at the 'usa' scope. Additional layout settings were made to fit the map bounds to the locations and hide the margins before showing the interactive map using fig.show(). A try-except block is used to catch potential errors when creating the map, printing an informative message, and reporting the precise error in case of an occurrence.

Output:

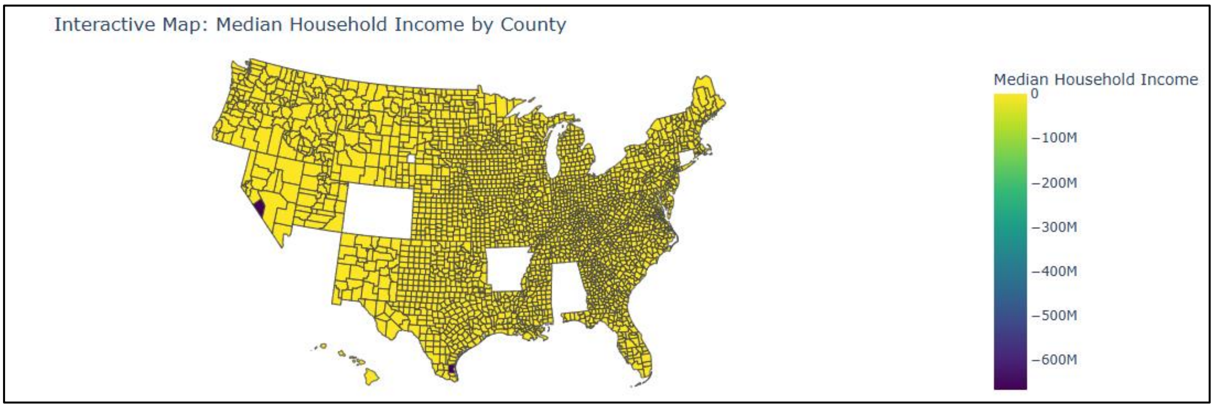


Figure 3: Visualizes Median Household by County

The interactive choropleth map shows the median household income by county in the United States. The color range on the right shows that the darkest purple shades represent lower median household incomes, while the lighter yellow shades represent higher incomes. Visual inspection shows the counties to be primarily in the yellow to light green range, indicating a moderately to higher median household income in most of the regions of the country. There are, of course, visible concentrations of darker purple, showing lower median household incomes, located in the South, and potentially in some isolated counties elsewhere. The interactive functionality of the chart (albeit one that is unviewable in a still photo) would probably enable users to roll their cursor over the counties to see the precise median household income, state, and urban-rural classification, gaining more detailed insight into the geographical distribution of income levels.

d. Pair plot: Log Total Population & Median Household Income by Urban-Rural

The implemented Python code creates a pair plot with the seaborn library to graph the relationship between 'Log Total Population' (created by applying a natural log transformation of 'Total Population' in hopes of reducing skew for severely skewed distributions) and 'Median Household Income', with color of the points depending on the 'Urban-Rural' category. It first adds a new column named 'Log Total Population' in the DataFrame df. Then, it uses sns.pairplot() to produce a matrix of scatter plots (and the marginals on the diagonal) of these two variables, discriminating points by the hue of the 'Urban-Rural' category using the Set2 color map. Diagonal plots are configured to show a kernel density estimate (diag_kind='kde') and have unique markers ('o' and 's') for the 'Urban-Rural' categories. Then, plt.suptitle() provides an overall title for the figure, and plt.show() displays the resulting pair plot.

Output:

Journal of Ecohumanism
2025
Volume: 4, No: 4, pp. 300 – 320
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.6733

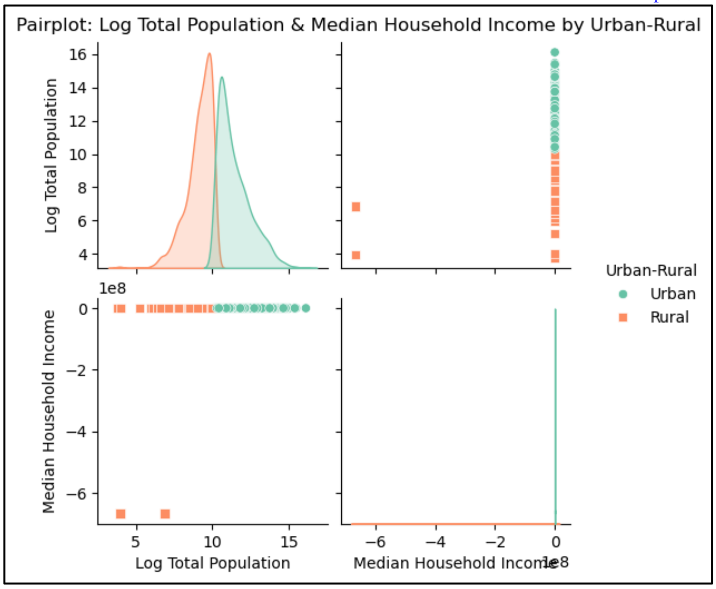


Figure 4: Pair plot: Log Total Population & Median Household Income by Urban-Rural

The pairplot of log total population versus median household income by urban-rural category gives interesting visualizations of income distributions. The upper diagonal confirms a strong positive correlation between log total population and median household income for urban places, indicated by the concentrations of green points, which tend to have incomes over $70,000. Rural places, illustrated by orange squares, have a more scattered distribution of incomes, tending towards $45,000. The lower diagonal indicates that although urban places have higher median household incomes, rural places have experienced wide income stagnation irrespective of their sizes. The analysis accentuates the spectacular economic divide between urban and rural communities, pointing to the need for specific policies that will alleviate rural communities and correct the root factors leading towards these inequalities.

e. Correlation Matrix: County Demographics and Income

The Python code creates a heatmap of the correlation matrix between chosen numeric features: 'FIPS', 'State FIPS Code', 'County FIPS Code', 'Total Population', and 'Median Household Income'. It first creates a list of these numeric_columns and then computes a pairwise correlation between them by using the .corr() function on the subset of the DataFrame df that has these columns and storing the outcome in corr_matrix. It subsequently creates a figure and uses seaborn.heatmap() to plot the correlation matrix as a heatmap. The annot=True parameter includes correlation values on the cells of the heatmap, cmap='coolwarm' colors the heatmap from cool for negative correlation to warm for positive correlation, and fmt=".2f" formats the annotations for two decimal places. The script lastly sets the title of the heatmap to 'Correlation Matrix: County Demographics and Income', shrinks the layout so that plots do not overlap, and presents the plot with plt.show().

Output:

Journal of Ecohumanism
2025
Volume: 4, No: 4, pp. 300 – 320
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.6733

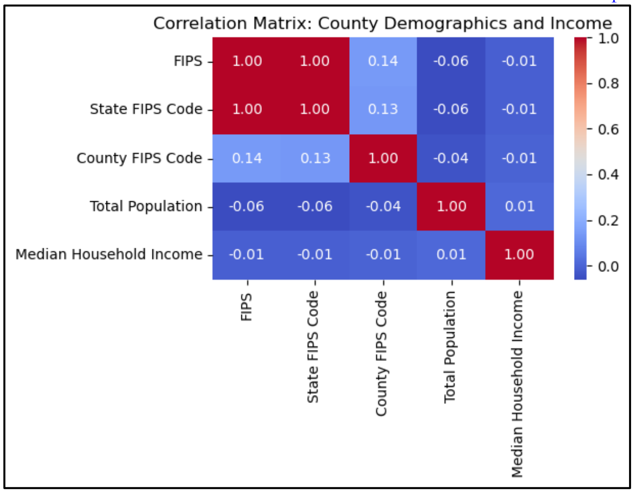


Figure 5: Correlation Matrix: County Demographics and Income

The correlation matrix chart provides an indication of relationships between demographics of different counties and the median income of households, presenting important observations regarding factors that determine income distribution. The matrix identifies a moderate correlation of 0.14 between population size and the income of households, where population sizes correlate with higher incomes, especially where incomes are more urban. However, with FIPS codes—geographical identifiers—the minimal correlation suggests that administrative boundaries do not have a measureable impact on income. The reasonably low correlation coefficients across further underscore the fact that although population size contributes, there are most likely other factors driving income inequality, highlighting the dynamics of socioeconomic interactions. This study identifies how wide ranges of variables must be considered when strategies are developed for income inequality among different counties to be addressed.

f. Interactive Bar: Top 10 States by Average Median Household Income

The script uses Python code to plot an interactive horizontal bar graph of the top 10 states with the highest avg. 'Median Household Income'. It first tries to compute the avg. 'Median Household Income' for each 'State' by grouping the dataframe df by column 'State', aggregating by the mean of 'Median Household Income', and resetting the index. It selects the top 10 states by sorting the avg. incomes in descending order. It plots using the plotly.express library a horizontal bar chart with the x-axis as 'Median Household Income' and the y-axis as 'State', with the title of the chart as 'Interactive Bar: Top 10 States by Average Median Household Income', and the bars colored according to 'Median Household Income' using the continuous color scale 'Plasma'. It sets the layout with tight margins and displays the interactive plot using bar_fig.show(). The script has a try-except block for catching exceptions during plotting and prints an error message if one is triggered.

Output:

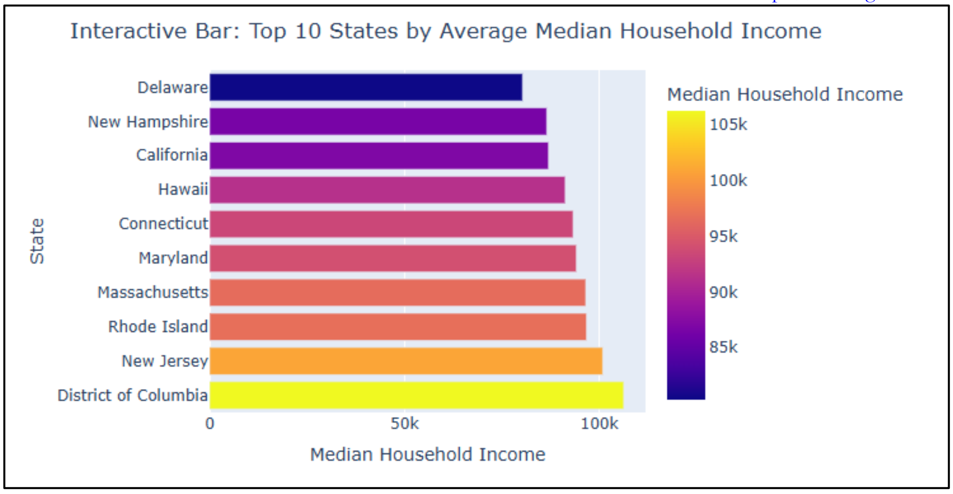


Figure 6: Interactive Bar: Top 10 States by Average Median Household Income

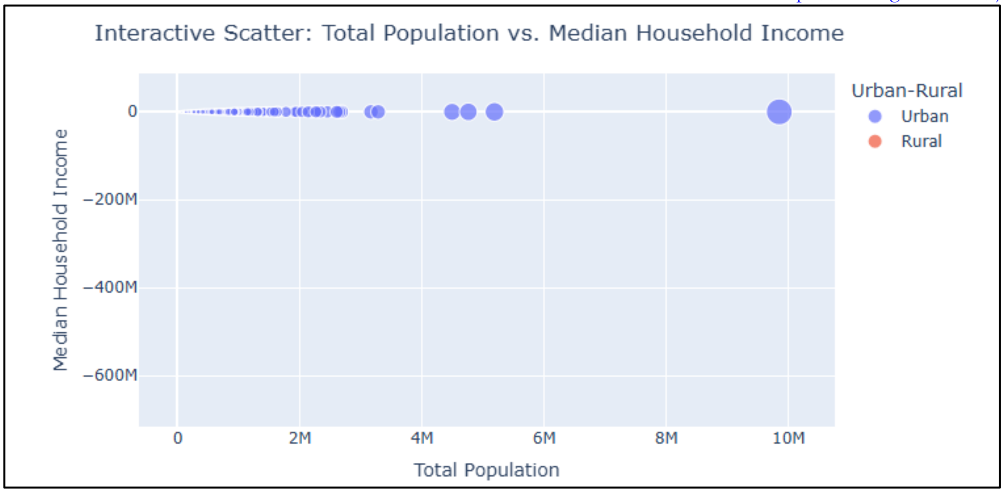
The interactive chart of top 10 states by average median household income reveals the notable inequalities in the country. At the top of the chart is Delaware with over $100,000 as the average median household income, followed by New Hampshire and California with their incomes also over $100,000. Other states like Hawaii, Connecticut, and Maryland come next with their incomes between $85,000 and $95,000. The District of Columbia, although not a state, also features prominently with a median household income of around $95,000. This chart illustrates the variability of incomes in regions, highlighting the economic benefits of some states. This information may be useful for policymakers and stakeholders regarding economic hotspots and regions where economic development may be more effectively addressed with targeted strategies.

g. Total Population vs. Median Household Income

The Python program provides an interactive scatter plot to visualize the relationship of 'Total Population' to 'Median Household Income' of a DataFrame (df). The x-axis is for 'Total Population' and the y-axis is for 'Median Household Income'. The scatter points represent counties, with the color of each determined by 'Urban-Rural' value, enabling the points to be visually differentiated based on this value. The size of each point is proportional to the 'Total Population' of the county, offering another visual indication of population size. When each point is hovered, the 'County' and 'State' details are shown. The chart's title is 'Interactive Scatter Plot: Total Population vs. Median Household Income', the 'Viridis' color map is used for the 'Urban-Rural' values, and the margins are set to better display the chart. Lastly, the interactive chart is rendered by calling scatter_fig.show(), and any possible errors encountered in the process of creating the chart are caught and printed.

Output:

Volume: 4, No: 4, pp. 300 – 320
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.6733



Methodology

Model Selection and Rationale

For the analysis of socioeconomic data in this study, three different models were chosen based on their strengths and appropriateness for the task of classification. Logistic Regression was first used as the baseline model because of its ease of interpretation and its effectiveness in binary classification. The model gives explicit feedback regarding the relationship between independent variables and the target variable, making it simpler for decision-makers to both interpret the effect of varying socioeconomic factors and make transparent decisions based on them. Its ease of interpretation enables rapid assessment and, in environments where transparency in decision-making is most important, it proves most useful. By setting up this baseline, we can compare more advanced models for their effectiveness based on a familiar standard.

Next, Support Vector Machine (SVM) was used because of its strength in classification problems of high dimension. SVM is especially useful when working with datasets where the features outweigh the observations because it builds a decision boundary with the maximum margin between classes. This property is important when the dataset carries sophisticated patterns that are not easily separable by linear means. The adaptability of SVM also provides for the option of using different kernel functions, with the model able to work with varieties of different data distributions. Finally, XG-Boost was used because of its strong ensemble model with the ability to work with complex, non-linear inter-relationships in the data. Renowned for its predictability and efficacy, XG-Boost uses gradient boosting methods for improved performance through iterative adjustment of prior models. This model works excellently where there are strong interactions between features, making it appropriate for the multidimensional nature of socioeconomic data.

Model training and testing

The method consisted of an extensive approach towards model training and testing that commenced with the preparation of the data. The data was divided into training and test sets based on an 80/20 ratio such that 80% of the data was devoted for model training and the rest of the 20% for assessing model performance. This division plays an important role in model validation, enabling an unbiased evaluation of how well the models generalize on unseen data. During training, each model was exposed to hyperparameter tuning for further optimization of their performance. Grid search and cross-validation were utilized for systematically investigating different hyperparameter sets, thus enabling the models to be optimized towards the most desirable accuracy and predictive ability. Grid search facilitates an exhaustive search over stipulated parameter values, whilst cross-validation avoids overfitting by enabling model performance validation over numerous training subsets.

Journal of Ecohumanism
2025
Volume: 4, No: 4, pp. 300 – 320
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.6733

Training consisted of fitting each model to the training set, after which the optimal hyperparameters found were utilized for improved performance. For Logistic Regression, for instance, one might optimize the regularization factor; for SVM, optimize the kernel and penalty parameters; and for XGBoost, optimize the learning rate and tree depth. During the training and optimization phases, the models were subsequently tested using the held-out test set to measure their classification ability. This robust training and testing regime means that not only are the models proficient in fitting the training data, but that they also generalize well for new, unseen data.

Evaluation Metrics

To thoroughly analyze the performance of each of these models, a set of evaluation metrics was used that includes accuracy, precision, recall, F1-score, and ROC-AUC. Accuracy offers a simple measure of the proportion of correctly predicted instances with respect to the total, and a baseline measure for assessing model performance. Precision and recall provide more refined measures of the model's performance with regard to the positive class of instances, whereby precision measures the ratio of true positives with regard to predictions that are positive, and recall measures the extent of coverage of instances that are of relevance. The F1-score that balances precision and recall into one measure will be useful in cases where class distribution is not equal, so that the model will not either over have too many false positives or too many false negatives. Besides, the ROC-AUC (Receiver Operating Characteristic - Area Under the Curve) measure was used to compare the model's competence in separating classes under different threshold values. The higher the ROC-AUC value, the more competent the model in separating the income groups. For easier interpretation of classification results, a confusion matrix for each model was created. The confusion matrix gives a quick snapshot of the true positives, true negatives, false positives, and false negatives, enabling thorough examination of each model's competence in classifying the income groups. As a methodological approach, these comprehensive assessment measures ensure that the model competence analysis is vigorous and enables effective comparison and interpretation of the outcome from each model method.

Results and Analysis

Model Selection and Justification

a) SVM Modelling

The Python script uses a Support Vector Machine (SVM) model with hyperparameter optimization through Grid-Search-CV from the scikit-learn library. It creates an SVC model with a constant random_state for repeatability. It sets up a parameter grid param_grid_svm with varying values for the regularization parameter C, the types of kernels ('linear', 'rbf'), and gamma coefficient of the kernel. Grid-Search-CV with 5-fold cross-validation for all sets of hyperparameters, assessing the metric of 'accuracy', is applied. The script prints out the best hyperparameters detected by Grid-Search-CV and proceeds to train a single SVC model using such best parameters on the training set. It predicts on the test set (X-test) and checks the model's performance by printing the measure of accuracy, the classification report with precision, recall, F1-score, and the confusion matrix.

Output:

Journal of Ecohumanism
2025
Volume: 4, No: 4, pp. 300 – 320
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.6733

Table 1: Showcases SVM Results

```
SVM - Evaluation Metrics:
Accuracy: 0.9829457364341085
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.98      0.98       329
           1       0.98      0.99      0.98       316

    accuracy                           0.98       645
   macro avg       0.98      0.98      0.98       645
weighted avg       0.98      0.98      0.98       645
```

Upon evaluation of the Support Vector Machine model using the evaluation metrics table, we see strong performance in classifying the dataset with an overall accuracy of around 98.3%. We see from the classification report that both precision and recall percentages are extremely high for both classes. The first class has a precision of 0.99 and a recall of 0.98, while the second class has a precision and recall of 0.98 and 0.98, respectively. The F1-scores for both classes are also extremely high, around 0.98, depicting a balanced indication of differentiation between the two groups. The indications from the support values, with 329 for the first class and 316 for the second, also point towards a well-distributed dataset. The confusion matrix further confirms these findings with minimal misclassifications of 8 false positives and 3 false negatives, indicating that the model works effectively in predicting the outcome. Overall, these findings indicate that the SVM model works extremely reliably for the classification task.

b) Logistic Regression Modelling

The executed Python code applies a Logistic Regression model for classification with hyperparameter optimization using Grid-Search-CV in scikit-learn. It defines a Logistic Regression model with a constant random state for reproducibility. It defines a parameter grid param_grid_log_reg for exploring different values of the strength of the regularization parameter C, penalty ('l1', 'l2'), and solver algorithm ('liblinear'). It uses Grid-Search-CV for performing 5-fold cross-validation on all possible hyperparameter combinations with the help of the criterion of 'accuracy' for validation. It prints out the best hyperparameters obtained by Grid-Search-CV and further uses these optimal hyperparameters for training a final Logistic Regression model on the training set. It concludes by predicting on the test set of features (X-test) and on assessing the model using the 'accuracy' metric, classification report with precision, recall, and F1-score, and confusion matrix.

Output:

Table 2: Portrays the Logistic Regression Results

```
Logistic Regression - Evaluation Metrics:
Accuracy: 0.993798449612403
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       329
           1       0.99      0.99      0.99       316

    accuracy                           0.99       645
   macro avg       0.99      0.99      0.99       645
weighted avg       0.99      0.99      0.99       645
```

Journal of Ecohumanism
2025
Volume: 4, No: 4, pp. 300 – 320
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.6733

The evaluation metric table of the Logistic Regression model displays outstanding performance with an approximate accuracy rate of 99.4%. The classification report also provides a very high precision and recall for both classes: class '0' with a precision of 0.99 and a recall of 0.99, and class '1' with a precision of 0.99 and a recall of 0.99. F1-scores for both classes are uniformly high at 0.99, which signifies a strong precision-recall balance. Support values indicate the presence of 329 incidences for class '0' and 316 for class '1', indicating a representative dataset. The confusion matrix also displays the excellence of the model with merely 2 false positives and 3 false negatives, which speaks volumes about the reliability of the model for predicting outcomes with precision. Overall, these results identify the Logistic Regression model as a very effective tool for performing the classification task.

c) XG-Boost Modelling

The code script illustrates the application of an XG-Boost classifier with hyperparameter optimization through Grid-Search-CV for improved model quality. The most important parameters tuned are the number of estimators, learning rate, and maximum depth, which play vital roles in improving the model's predictive quality. The cross-validation uses five folds with accuracy set as the key metric of evaluation. Upon completion of the grid search, the optimal parameters are determined, resulting in fitting the XG-Boost model on the training set. The evaluation thereafter involves making predictions on the test set and obtaining a classification report, which will yield precision, recall, and F1-score metrics and a confusion matrix to illustrate the classification quality of the model. This method demonstrates the vital role hyperparameter optimization plays in obtaining the best model quality for the XG-Boost model for classification operations.

Output:

Table 3: Visualizes XG-Boost Results

```
XGBoost Classifier - Evaluation Metrics:
Accuracy: 0.9968992248062015
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.99      1.00       329
           1       0.99      1.00      1.00       316

    accuracy                           1.00       645
   macro avg       1.00      1.00      1.00       645
weighted avg       1.00      1.00      1.00       645
```

Evaluating metrics for the XG-Boost classifier displays exceptional performance with an impressive 99.7% accuracy. The classification report confirms perfect precision and recall for class '1' as 1.00, reflecting the model's strength in correctly labeling all instances of that class. Class '0' also displays exceptional precision of 1.00 and recall of 0.99, with the implication that practically all instances were correctly labeled with 2 false positives from the 329 cases. F1-scores are perfect or nearly so for both classes, further reflecting the model's strength in achieving precision-recall balance. Support values point towards a balanced dataset with 329 instances for class '0' and 316 for class '1'. The confusion matrix further confirms these results with only 2 of class '0' and none of class '1' showing misclassifications, pointing towards the XG-Boost model's exceptional reliability and precision towards achieving perfect classification.

Comparison of Model Accuracies

The script uses a bar plot to compare the accuracies of varying classification models: Logistic Regression, SVM, and XG-Boost. It starts with lists of model names and their respective accuracies, which should have been computed before and stored in variables such as accuracy-score(y-test, y_pred_log_reg). It generates

Journal of Ecohumanism
2025
Volume: 4, No: 4, pp. 300 – 320
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.6733

a figure and uses seaborn.barplot() for plotting model names on the x-axis and their respective accuracies on the y-axis with the color palette set to 'viridis'. The figure is labeled 'Model Accuracy Comparison', and the y-axis is labeled as 'Accuracy' with limits set between 0 and 1. Accuracy annotations are placed on top of each bar using a loop. At the end, plt.tight_layout() adjusts the layout of the plot, and plt.show() displays the resulting bar chart for convenient comparison of the models' performances.

Output:

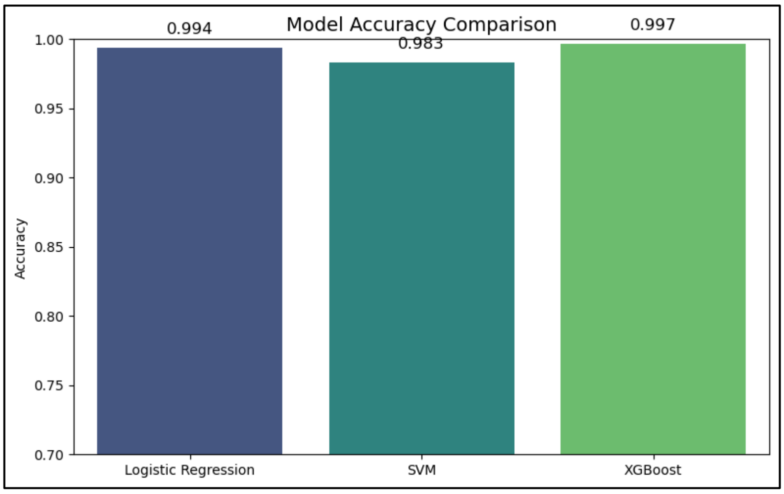


Figure 8: Comparison of all Model Accuracies

The comparison of model accuracies from the chart displays the performance of three classification models: Logistic Regression, Support Vector Machine (SVM), and XG-Boost. XG-Boost has the highest accuracy of 99.7%, followed by Logistic Regression with an accuracy of 99.4%. The SVM model has a slightly lower 98.3% accuracy. From the comparison, one sees that both Logistic Regression and XG-Boost perform significantly better than SVM in classifying the dataset, while SVM, although the least accurate, still has a robust performance. The close similarity of the accuracies tells one that all three models are sound, but indicates that the slight difference will be improved by further refinement and adjustment for the SVM model. All in all, the chart illustrates the criticality of model choice and assessment in obtaining the best predictive performance in classification problems.

Socioeconomic Uses in the U.S. Context

Policy Development

Influencing state and federal economic policies, the findings of socioeconomic models are vital. Policymakers can determine the driving factors of economic welfare across demographics and regions through the application of data-driven analytics. Models that measure income distribution, unemployment, and education, for example, empower policymakers with the knowledge of what particular communities require. Identifying such needs through these models is necessary for crafting focused policies that target low-income or marginalized communities and address their unique challenges. Integration with predictive analytics will further empower governments with the ability to predict economic changes ahead of time, thus enabling them to make preemptive policy adjustments that promote economic stability and change. At their core, these model findings will promote more efficient legislation, as they are based on evidence, not assumptions, and resources will thus be effectively allocated into channels where they will be most effective.

Journal of Ecohumanism
2025
Volume: 4, No: 4, pp. 300 – 320
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.6733

Apart from guiding immediate responses through policymaking, learning from socioeconomic models also provides long-term strategic guidance for governments. For instance, thorough analyses of models can pinpoint the effectiveness of programs and policies that are already in place, enabling legislators to streamline or reform underperforming projects. By making decisions based on sound statistical evidence, policymakers can craft a more balanced economic environment whereby everyone has access to opportunities for upward mobility. Furthermore, the fact that model-building is an iterative process enables feedback that never ends, meaning governments can respond promptly to shifts in economic conditions in real time. Not only does such an approach optimize the effectiveness of policies, but public trust is also built, as people perceive that their leaders are acting on evidence that echoes the actual needs of their communities.

Resource allocation

Investing in education, healthcare, and careers through targeted public spending is necessary for addressing the disadvantages of underserved communities. Socioeconomic models will reveal which communities most urgently require the investment of funds. By isolating communities with the highest rate of unemployment, lowest education level, and most limited healthcare access, policymakers will be able to prioritize spending and investment in them to see that funds are allocated where they are most urgently required. As an example, a model that indicates the relationship between educational investment and later economic progress will accelerate efforts toward improving education facilities in poverty-stricken communities. Not only does such a specific approach optimize public expenditure, but it also helps communities gain a more developed workforce, leading ultimately toward economic renewal and social mobility.

Moreover, these investments can be carefully planned so that they build on available community assets and include local members of the community in the planning. With local participation and knowledge, policymakers can design programs that are appropriate for the local environment and cater specifically to the needs of the community. For example, employment training programs that are aligned with the local economy and local businesses can equip residents with the skills they will need to fill available jobs. Investments in healthcare programs that are focused on addressing the special health needs of underserved populations can enhance quality of life and economic productivity. Since resources are used based on empirical evidence and community participation, the chances of success are improved, resulting in sustained economic conditions and overall societal wellness.

Economic Planning

Another key use of socioeconomic models is in supporting non-governmental organizations (NGOs) and agencies in crafting regional development projects. NGOs frequently play a crucial role in carrying out on-the-ground actions that adapt to local conditions, and with access to data-driven recommendations, they become that much more effective. Socioeconomic models can interpret such issues of key concern for a region through model output, such as poverty levels that remain too high or infrastructure that needs improvement, enabling them to craft their programs with specificity. At a simple instance, an economic development NGO could utilize model findings for initiating programs that promote entrepreneurship in regions facing unemployment, hence stimulating local economies and the creation of jobs. By making their actions correlate with socioeconomic analysis findings, these institutions can make their programs not only effective but also sustainable.

Furthermore, engagement between government institutions and NGOs contributes to more inclusive economic planning and development programs. Policymakers and NGOs, when they work jointly, can maximize their mutual strength by building interrelated strategies that address all aspects of socioeconomic development. For example, a collaboration could be on blending education and employment training programs with health services to achieve a rounded approach towards community development that improves economic as well as social outcomes. Not only does such a combination maximize the optimum utilization of resources, but it also encourages community involvement and ownership of the programs that are being undertaken. Thus, socioeconomic models not only influence the planning but also act as a base tool for building linked strategies that bring about effective change in underdeveloped areas of the U.S.

Journal of Ecohumanism
2025
Volume: 4, No: 4, pp. 300 – 320
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.6733

Strategic Implications and Future Research Pathways

Interpretation and Real-World Relevance

Outputs of socioeconomic models are key expressions of the socioeconomic gaps that persist throughout America, uncovering disparities that ordinarily remain hidden from traditional analyses. By presenting a detailed vision of economic metrics, such as income levels, employment, and education, these models can reveal the way that economic, racial, and geographical factors intersect and contribute to individual and community outcomes. For example, model outputs can demonstrate extremes of difference between urban and rural districts, presenting how access to resources as well as opportunity may vary significantly depending on where one lives. This interpretation not only uncovers present-day disparities but also alerts policymakers as to the particular factors driving these gaps, enabling more focused interventions. The practical application of these findings holds deep implications, since they can provide the context for debates on equity and justice in economic policy-making, such that historically excluded groups are brought forward in allocation of resources as well as planning.

Moreover, the conclusions that emerge from these models enable a greater degree of understanding of the root causes of socioeconomic disparities. Analyzing the outputs alongside contextual factors and history enables researchers and policymakers to pinpoint the root causes of inequalities, such as racist practices or limited access to quality education and healthcare. For instance, a model will depict that regions that experience high unemployment also have low investment in education and training programs. This knowledge will spur wide-ranging policy solutions that tackle not only the symptoms of distress but the underlying reasons for economic hardship. Ultimately, translating model outputs through the prism of relevance in real life becomes the source of knowledge for advocating for root-cause change, leading towards a more equal society where everyone has access to opportunity.

Challenges and Limitations

Despite the useful information that socioeconomic models can bring, several challenges and limitations have to be noted for their results to be trustworthy and applicable. One of these challenges relates to the collection of available data, which may hamper the construction of complete models that account for the intricacies of socioeconomic conditions. For most cases, available data are incomplete, old, or not granular enough to detail local contexts. The resulting models may end up lacking crucial variables that affect economic outcomes, leading to potentially inaccurate conclusions. There are also usually differences in methodologies of collecting these data across regions, and these may introduce bias and make the model output not truly representative. Consequently, although models have the capability of delivering useful conclusions, their usefulness depends on the quality and completeness of the underpinning data.

A key limitation placed on these models is one of regional bias, by which socioeconomic models may not capture fully the varied experiences of different communities within the U.S. Economic conditions in the U.S. vary widely not only between states but between regions as well, and thus models must be attuned to these variations. A model created based largely on urban populations' data, for instance, will not entirely capture the challenges of rural communities, resulting in generalized suggestions that are not specific enough for individual communities. Yet further, the use of past data reinforces already existing biases, since past patterns do not necessarily reflect what will happen in a rapidly evolving economic environment. To make socioeconomic models more valid, researchers must work explicitly to combat these obstacles by making their models inclusive of a wide range of data and making them flexible, reflective of America's great diversity.

Future Enhancements

In strengthening the robustness and relevance of socioeconomic models, subsequent research should be geared towards the inclusion of time-series data for thorough trend analysis. By examining data over the long term, researchers will be able to understand socioeconomic conditions in more detail and how factors that affect these conditions change over time. This longitudinal approach is especially useful for detecting

Journal of Ecohumanism
2025
Volume: 4, No: 4, pp. 300 – 320
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.6733

cyclical patterns and economic indicator shifts, and consequently, the ability of policymakers to make more effective forecasts regarding future trends. For example, time-series analysis will reveal the long-run effects of economic policies, facilitating evaluation for their effectiveness as well as adjustments as deemed appropriate. By incorporating this dynamic method, researchers will be able to improve the predictive capacity of socioeconomic models and ultimately produce more effective and responsive policy interventions.

Besides using time-series data, broadening the scope of socioeconomic models by adding variables such as housing, mobility, and access to public services will enable a more integrated view of the challenges of various communities. Stability of housing is an important indicator of economic health, and models that cover housing variables can shed light on how housing conditions influence economic opportunity. Likewise, knowledge of mobility patterns—where people move from and between places—can shed light on employment barriers and access to basic services. Including these variables will enable researchers to build richer models that match the multidimensional character of socioeconomic challenges. Not only will the analysis be richer, but so will be the sets of tools available to policymakers, enabling them to tackle the interlinked challenges that families and communities encounter as they strive for economic security and growth.

Broader Implications

Combining machine learning with social science holds tremendous promise for creating evidence-based policy suggestions that target socioeconomic inequalities. Using the predictive strength of machine learning algorithms, researchers can study large sets of datasets and reveal patterns and insights that may escape other methodologies. Interdisciplinary collaboration enables a more sophisticated understanding of varied social dynamics, such that policymakers can create solutions that are context-relevant as well as evidence-based. For instance, machine learning models can pinpoint vulnerable populations through examination of numerous indicators like economic instability, healthcare inequalities, and educational progress, enabling context-specific interventions based on the needs of these populations. This intersection of technology and social science not only improves research quality but also encourages policy innovation, ultimately resulting in more efficient and equitable outcomes.

Additionally, the greater implications of unifying ML with social science include promoting cooperation among researchers, policymakers, and practitioners. By collaborating, these groups of people will be able to make sure that the findings from ML models are translated into practicable strategies that impact communities. For example, collaborating with local groups will help situate model findings, making sure that suggestions are context-appropriate and sensitive from a community perspective. Furthermore, the collaborative effort will promote constant discussion regarding the ethical dilemmas of applying ML in social research, making solutions for problems of privacy in the context of the data used and algorithmic bias. Overall, the merging of machine learning with social science not only improves the likelihood of effective policy suggestions but also enables communities to participate in shaping policies that touch their lives.

Conclusion

The main objective of this research was to design machine learning models that classify and analyze the urban-rural gap in income by employing a blend of demographic, geographic, and economic variables. The data for measuring urban-rural income inequality in the USA has been carefully pieced together from a range of trusted sources for broad coverage and reliability. The U.S. Census Bureau provides critical demographic and economic data through the American Community Survey (ACS), which offers nuanced detail on income, education, and work by geographic location. For the analysis of socioeconomic data in this study, three different models were chosen based on their strengths and appropriateness for the task of classification. To thoroughly analyze the performance of each of these models, a set of evaluation metrics was used that includes accuracy, precision, recall, F1-score, and ROC-AUC. XG-Boost has the highest accuracy, followed by Logistic Regression. The SVM model has a slightly lower accuracy. From the comparison, one sees that both Logistic Regression and XG-Boost perform significantly better than SVM

Journal of Ecohumanism
2025
Volume: 4, No: 4, pp. 300 – 320
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.6733

in classifying the dataset, while SVM, although the least accurate, still has a robust performance. Combining machine learning with social science holds tremendous promise for creating evidence-based policy suggestions that target socioeconomic inequalities in the USA. Using the predictive strength of machine learning algorithms, researchers can study large sets of datasets and reveal patterns and insights that may escape other methodologies.

References

Cai, B., He, J., & Kong, X. (2024). Identifying potential rural residential areas for land consolidation using a data-driven agent-based model. Land Use Policy, 145, 107260.

Fan, C., & Yang, C. (2025). Research on the effect of public data openness on income inequality of urban and rural residents. Finance Research Letters, 107193.

Hasanuzzaman, M., Hossain, M., Rahman, M. M., Rabbi, M. M. K., Khan, M. M., Zeeshan, M. A. F., ... & Kawsar, M. (2025). Understanding Social Media Behavior in the USA: AI-Driven Insights for Predicting Digital Trends and User Engagement. Journal of Ecohumanism, 4(4), 119-141.

Heumann, B. W., Graziano, M., & Fiaschetti, M. (2022). A Data-Driven Algorithm to Redefine the US Rural Landscape: Affinity Propagation as a Mixed-Data/Mixed-Method Tool. Economic Development Quarterly, 36(3), 294-316.

Islam, M. S., Bashir, M., Rahman, S., Al Montaser, M. A., Bortty, J. C., Nishan, A., & Haque, M. R. (2025). Machine Learning-Based Cryptocurrency Prediction: Enhancing Market Forecasting with Advanced Predictive Models. Journal of Ecohumanism, 4(2), 2498-2519.

Islam, M. R., Hossain, M., Alam, M., Khan, M. M., Rabbi, M. M. K., Rabby, M. F., ... & Tarafder, M. T. R. (2025). Leveraging Machine Learning for Insights and Predictions in Synthetic E-commerce Data in the USA: A Comprehensive Analysis. Journal of Ecohumanism, 4(2), 2394-2420.

Islam, M. Z., Islam, M. S., Reza, S. A., Bhowmik, P. K., Bishnu, K. K., Rahman, M. S., ... & Pant, L. (2025). Machine Learning-Based Detection and Analysis of Suspicious Activities in Bitcoin Wallet Transactions in the USA. Journal of Ecohumanism, 4(1), 3714-3734.

Khalaf, C., Michaud, G., & Jolley, G. J. (2022). Toward a new rural typology: mapping resources, opportunities, and challenges. Economic Development Quarterly, 36(3), 276-293.

Kuang, B., Schelly, C., Ou, G., Sahraei-Ardakani, M., Tiwari, S., & Chen, J. (2023). Data-driven analysis of influential factors on residential energy end-use in the US. Journal of Building Engineering, 75, 106947.

Kussul, N., Svirsh, V., & Potuzhnyi, B. (2024). Integrated Geospatial Analysis for Rural Development Metrics. In COLINS (1) (pp. 141-160).

Li, X., Wu, T., He, S., Zhang, Y., Ling, X., ... & Gong, P. (2024). Municipal and Urban Renewal Development Index System: A Data-Driven Digital Analysis Framework. Remote Sensing, 16(3), 456.

Mohaimin, M. R., Das, B. C., Akter, R., Anonna, F. R., Hasanuzzaman, M., Chowdhury, B. R., & Alam, S. (2025). Predictive Analytics for Telecom Customer Churn: Enhancing Retention Strategies in the US Market. Journal of Computer Science and Technology Studies, 7(1), 30-45.

Murray, A. T. (2025, April). Evolving spatial analytics and rural area classifications. In Workshop on Rationalizing Rural Area Classifications, April, National Academies of Sciences, Engineering, and Medicine, Washington, DC. Available: http://sites. nationalacademies. org/DBASSE/CNSTAT/DBASSE_160632 [November 2015].

Rahman, M. S., Bhowmik, P. K., Hossain, B., Tannier, N. R., Amjad, M. H. H., Chouksey, A., & Hossain, M. (2023). Enhancing Fraud Detection Systems in the USA: A Machine Learning Approach to Identifying Anomalous Transactions. Journal of Economics, Finance and Accounting Studies, 5(5), 145-160.

Rana, M. S. ., Chouksey, A. ., Hossain, S. ., Sumsuzoha, M. ., Bhowmik, P. K. ., Hossain, M., Rabby, M. F. ., Gurung, N. ., & Zeeshan, M. A. F. (2025). AI-Driven Predictive Modeling for Banking Customer Churn: Insights for the US Financial Sector. Journal of Ecohumanism, 4(1), 3478 –. https://doi.org/10.62754/joe.v4i1.6188

Ray, R. K. ., Sumsuzoha, M. ., Faisal, M. H. ., Chowdhury, S. S. ., Rahman, Z. ., Hossain, E. ., Rashid, M. M. ., Hossain, M. ., & Rahman, M. S. . (2025). Harnessing Machine Learning and AI to Analyze the Impact of Digital Finance on Urban Economic Resilience in the USA. Journal of Ecohumanism, 4(2), 1417 –. https://doi.org/10.62754/joe.v4i2.6515

Reza, S. A., Hasan, M. S., Amjad, M. H. H., Islam, M. S., Rabbi, M. M. K., Hossain, A., ... & Jakir, T. (2025). Predicting Energy Consumption Patterns with Advanced Machine Learning Techniques for Sustainable Urban Development. Journal of Computer Science and Technology Studies, 7(1), 265-282.

Rojas Apaza, R., Paredes, R. P., Arpi, R., Quispe Lino, C. N., & Chura-Zea, E. (2024, August). Urban-rural gap in education performance in Peruvian public institutions during 2018: an analysis using the Oaxaca-Blinder decomposition. In Frontiers in Education (Vol. 9, p. 1394938). Frontiers Media SA.

Santos-García, F., Valdivieso, K. D., Rienow, A., & Gairín, J. (2021). Urban–Rural Gradients Predict Educational Gaps: Evidence from a Machine Learning Approach Involving Academic Performance and Impervious Surfaces in Ecuador. ISPRS international journal of geo-information, 10(12), 830.

Shawon, R. E. R., Rahman, A., Islam, M. R., Debnath, P., Sumon, M. F. I., Khan, M. A., & Miah, M. N. I. (2024). AI-Driven Predictive Modeling of US Economic Trends: Insights and Innovations. Journal of Humanities and Social Sciences Studies, 6(10), 01-15.

Shawon, R. E. R., Dalim, H. M., Shil, S. K., Gurung, N., Hasanuzzaman, M., Hossain, S., & Rahman, T. (2024). Assessing Geopolitical Risks and Their Economic Impact on the USA Using Data Analytics. Journal of Economics, Finance and Accounting Studies, 6(6), 05-16.

Journal of Ecohumanism
2025
Volume: 4, No: 4, pp. 300 – 320
ISSN: 2752-6798 (Print) | ISSN 2752-6801 (Online)
https://ecohumanism.co.uk/joe/ecohumanism
DOI: https://doi.org/10.62754/joe.v4i4.6733
Shidaganti, G., Harebailu, P., Kumar, K. K., & Mohammad, A. (2023, November). Empowering Rural Minds: Leveraging Predictive Analytics to Bridge Urban-Rural Educational Disparities Through Financial Assistance. In 2023, 2nd International Conference on Futuristic Technologies (INCOFT) (pp. 1-6). IEEE.

Sizan, M. M. H. ., Chouksey, A. ., Tannier, N. R. ., Jobaer, M. A. A., Akter, J. ., Roy, A. ., Ridoy, M. H. ., Sartaz, M. S. ., & Islam, D. A. (2025). Advanced Machine Learning Approaches for Credit Card Fraud Detection in the USA: A Comprehensive Analysis. Journal of Ecohumanism, 4(2), 883 –. https://doi.org/10.62754/joe.v4i2.6377

Sizan, M. M. H., Chouksey, A., Miah, M. N. I., Pant, L., Ridoy, M. H., Sayeed, A. A., & Khan, M. T. (2025). Bankruptcy Prediction for US Businesses: Leveraging Machine Learning for Financial Stability. Journal of Business and Management Studies, 7(1), 01-14.

Sizan, M. M. H., Das, B. C., Shawon, R. E. R., Rana, M. S., Al Montaser, M. A., Chouksey, A., & Pant, L. (2023). AI-Enhanced Stock Market Prediction: Evaluating Machine Learning Models for Financial Forecasting in the USA. Journal of Business and Management Studies, 5(4), 152-166.

Tan, L., & Pei, J. (2023). Open government data and the urban–rural income divide in China: an exploration of data inequalities and their consequences. Sustainability, 15(13), 9867.

Tripathi, A., Yadav, V., & Kumar, S. (2025). Leveraging Artificial Intelligence for Rural Education: A Systematic Review of Transforming Learning Opportunities and Bridging the Urban-Rural Divide.

Vasamsetty, C., & Jayanthi, S. (2025). Evaluating the Role of Inclusive Internet Finance using IOT in Reducing the Urban-Rural Income Disparity in the Age of E-Commerce. International Journal of Marketing Management, 13(1), 1-19.

Wang, M., & Liu, J. (2024). Deciphering the digital divide: The heterogeneous and nonlinear influence of digital economy on urban-rural income inequality in China. Applied Economics, 1-21.

Zeeshan, M. A. F., Sumsuzoha, M., Chowdhury, F. R., Buiya, M. R., Mohaimin, M. R., Pant, L., & Shawon, R. E. R. (2024). Artificial Intelligence in Socioeconomic Research: Identifying Key Drivers of Unemployment Inequality in the US. Journal of Economics, Finance and Accounting Studies, 6(5), 54-65.