# Interpretable machine learning learns spatial features to understand socio-economic inequality

## Introduction

Regional inequalities have become one of the most fundamental social and economic issues of our time. Traditional inequality issues arose with the stratification of socio-economic classes and relations characterized by income concentration. As research progressed and cities grew, studies in this area began to address issues of inequality in people's lives, such as satisfactory public services, accessibility to life's needs and educational opportunities[1]. Despite significant efforts in research and practice to measure and mitigate inequality, the optimal equality of amenity services and commuting choices in cities remains systematically biased and poorly understood[2][3]. In this study, we examine inequality as a result of complex interactions between demographic and spatial characteristics of urban areas. Capturing this mechanism through supervised machine learning classification can help measure and explain the presence of inequality and inform policy and planning to promote equitable cities[1].

The amenities and human activities are heterogeneous and dynamic, leading to high variations in socio-economic patterns[4]. Due to the varying pace of dynamic urban components such as the commuting activities of residents and the evolution of the number of amenities, we argue that their interactions are subtle and nonlinear. To understand the inequalities created by intertwined urban spatial features, it is crucial to utilize machine learning to capture the variability and nonlinearity of the interactions between these spatial components[5].

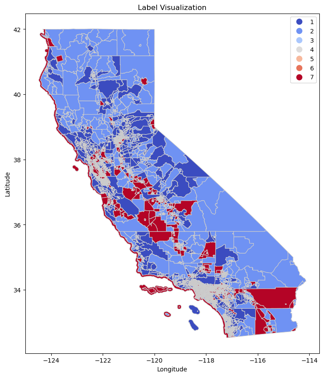
## This paper utilizes five machine learning classification models to identify and measure inequality in cities. An attempt is made to explore the differences in the performance of each model in terms of overall classification accuracy, various socio-demographic categories and the reasons behind them, and feature importance, with the hope of drawing conclusions about inequality due to the interaction of socio-spatial features in California from the interpretation of the comparative results of the models.

## 2.Method

### 2.1Dataset

#### 2.1.1 Label

To compare the features of different urban areas, we collected socio-economic public data including per capita income and race–ethnicity data from the US Census 2018– 2022 (5 years) American Community Survey (ACS) at census tract level of spatial aggregation (United States Census Bureau, 2023). We focused on the three largest race–ethnicity groups as determined by self-identification in the Census: White, Black or African American, and Asian. The race that accounts for greater than 50% of people in a census tract reported in the Census data is considered the race label of this census tract. We similarly classified the census tracts as low-income or high-income based on whether the per capita income of the census tract is higher than the median of the California or not. We assign the label of a census tract, as such, specific census tracts in California are labeled by one of six socio-economic labels.(label1 : Rich White; label2 : Poor White; label3 :Rich Black; label4:Poor Black; label5: Rich Asian; label6: Poor Asian). While processing the data on labels, we found that 4,162 census tracts did not fall into any of the six categories. This is due to the fact that California is supposed to be a cluster of immigrants of multiple races, so mixed multiracial is common. In order to categorize more efficiently and accurately, this paper will remove other observations that do not fall within these 6 categories. A total of 4967 samples were left for model training and testing.



#### 2.1.2 Feature Selection

We use spatial variables as predictor variables. We used the osmnx package for python to crawl the poi of various amenities in California at the Open Street Map website. and using R studio, we counted the number of each type of amenity in each census tract and used this as a spatial explanatory variable for the model. These facilities cover all aspects of residents' lives and are representative. In addition, the population inflow data of within county, from different county within same state, from different state, and from abroad at four scales in each census tract are also selected as predictor variables. In this paper, average commuting time and commuting mode are also used as predictor variables of spatial-transportation.

|  |  |  |
| --- | --- | --- |
| Amenity | Category | Total Number |
| Sustenance | bar', 'biergarten', 'cafe', 'fast food', 'food court', 'ice cream', 'pub', 'restaurant' | 55347 |
| Entertainment | 'arts center', 'casino', 'cinema', 'community center', 'conference center’, ‘events venue’, ‘exhibition center', 'gambling’, ‘music venue', 'planetarium', 'public bookcase', 'theatre' , 'social center' | 4826 |
| Health Center | 'baby hatch', 'clinic', 'dentist', 'doctors', 'hospital', 'nursing home', 'pharmacy', 'social facility’, ‘veterinary' | 9969 |
| Financial | 'atm', 'bank' | 6416 |
| Education | 'library', 'kindergarten’, ‘school' | 18307 |

### 2.2Algorithm

We trained four state-of-the-art classification models and compared their results, including Random Forest (RF), Support Vector Machine(A common approach when using SVM for multi-categorization tasks is to use a one-to-many strategy. Specifically, for a multi-categorization problem with N categories, we train N binary classifiers, each of which is used to distinguish one category from all other combinations of categories.), Decision Tree, Boosted Decision Tree, and Bootstrap Aggregation(Bagging). In the model training step, we first split the total sample data into 80% training and 20% testing data.

## Finding

*Detailed Value and figures show on the Appendix*

### 2.1Accuracy evaluation

SVM has an accuracy of 0.698 and performs the best of all the models, which indicates that SVM usually performs well when dealing with linear and nonlinear classification problems. However, the prediction of this model only contains two groups, "Rich-White" and "Poor-White". Even when I adjusted the model parameters from a linear kernel function to a radial basis function kernel, this did not solve the problem, which should have little to do with the model complexity. The problem is mainly due to the imbalance of data, the sample size of "Rich-White" and "Poor-White" categories in the dataset is much higher than the sample size of other categories, which have fewer samples, so the model is more inclined to predict these categories and ignores the other categories.

Random Forest and Gradient Boosting Classifier have an accuracy of 0.696, which is a better performance. They are both integrated learning algorithms, the former improves the prediction performance by combining multiple decision trees, and the latter minimizes the loss function by iteratively training the decision trees, which usually has high accuracy and robustness. Bagging: with an accuracy of 0.663, it performs moderately well. Bagging is also an integrated learning method, which constructs multiple models and then averaging the prediction results to reduce variance, but does not perform as well as the other two ensemble methods. Decision tree has the worst performance with an accuracy of 0.591. This may be due to the fact that decision trees are prone to overfitting and may not be stable enough, especially in high-dimensional datasets or in the presence of more noise.

Browsing through the classification reports of all the models shows that the F1-score for the category "Rich-White" is the highest among all the models, which is greater than 0.7, which indicates that the classifier performs the best for this category due to the fact that it has the largest number of samples in this category." The F1-score for the category "Poor-White" is the second highest among all the models, with values in the range of 0.49-0.60, which indicates that the classifier performs better for this category, also because of the sufficient number of samples. From this we can conclude that in California the predominantly purebred white population is the majority of the census tract.

Browsing through the classification reports of all the models shows that the F1-score for the category "Rich-White" is the highest among all the models, which is greater than 0.7, which indicates that the classifier performs the best for this category due to the fact that it has the largest number of samples in this category." The F1-score for the category "Poor-White" is the second highest among all the models, with values in the range of 0.49-0.60, which indicates that the classifier performs better for this category, also because of the sufficient number of samples. From this we can conclude that in California the predominantly purebred white population is the majority of the census tract. When the model is transformed from SVM to Decision Tree, the F1 of the three categories "Poor-Black", "Rich-Asian", and "Poor-Asian" increases from and is no longer equal to 0. This shows that when facing the imbalance in the amount of data in each category, the overall classification performance of Decision Tree is better than SVM. Decision Tree is a model based on local optimization, where each node is classified based on the data distribution and features of the current node without the need of global optimization process. In contrast, SVM is usually trained based on a global optimization objective and may be affected by unbalanced data distribution. Decision Tree is more adaptive to unbalanced data because it can adaptively divide the data based on its distribution and features to better handle unbalanced data. Among all the Tree Models, Random Forest has the worst F1-score for each category. This may be due to overfitting of the model, sample imbalance, and poor parameter selection.

### 2.2Feature Importance

SVM itself does not provide a direct feature importance assessment. Therefore, we come to analyze the feature importance of four tree-based models. In all four models, the importance of commuting mode was very high, especially for the variable of the percentage of workers commuting by car. It is worth noting that in the gradient descent boosted tree model, the importance of all variables except commuting mode is very, very low. In the other three tree-based models, the importance of the various types of in-migration variables, and commute time in categorization should not be overlooked, in contrast to the number of amenities, except for the number of sustenance, which did not play a significant role in categorization. This may be due to the fact that the number of basic arts, culture, entertainment, education, healthcare, and financial facilities are more evenly distributed in California and do not show strong spatial inequality. Second, we can also conclude that the distribution of the number of restaurants is an important indicator of socio-spatial inequality in California. This is because restaurants are closely related to the prosperity of local economic activities and the prosperity of residents' lives.

## Conclusion

In this paper, we label individual California census tracts with reference to income and racial proportions, and train five supervised machine learning classification models to classify individual census tracts according to various dimensions of spatial feature. This paper illustrates the spatial nature of social inequality in the urban realm, shifting the discussion of urban inequality from socio-economic variables to a focus on urban space, particularly commonly used facilities and transportation. Urban social research on inequality tends to take space for granted, implicitly utilizing it in its analyses, but fails to adequately consider the impact and power of space in reproducing social inequalities or as a tool for social actors. A focus on spatiality can reveal how social inequalities are reproduced in multiple ways in urban space[7].

Machine learning captures the various facility systems and commuting patterns and demographic characteristics of cities and their interactions; inequality exists if machine learning models characterized by commuting activities and facility characteristics can accurately predict socioeconomic conditions in different areas. In other words, if inequality exists, then commuting activity and amenity characteristics can play an important role in models that identify high-income/low-income areas and areas dominated by populations of which races. Therefore, the predictive performance metrics of machine learning models can be used to measure the degree of inequality. A high predictive performance of the model indicates a higher degree of socioeconomic inequality in the city. Therefore, we use the F1 score, which measures the predictability of machine learning models, as a measure of the degree of urban inequality[5]. Based on the results we know that the F1 scores of the models in the two white neighborhoods are very high, and from this we can infer that racial inequality is an important manifestation of spatial social inequality in California, and that inequality between races is more prominent than inequality between incomes.

## Reference

[1]Acemoglu, D., & Robinson, J. (2009). Foundations of societal inequal- ity. Science, 326(5953), 678–679.

[2]Balland, P. A., Jara-Figueroa, C., Petralia, S. G., Steijn, M. P. A., Rigby, D. L., & Hidalgo, C. A. (2020). Complex economic activities concentrate in large cities. Nature Human Behaviour, 4, 248–254.

[3]Mirza, M. U., Xu, C., Bavel, B. V., van Nes, E. H., & Scheffer, M. (2021). Global inequality remotely sensed. Proceedings of the National Academy of Sciences, 118(18), e1919913118.

[4]Han, S. Spatial stratification and socio-spatial inequalities: the case of Seoul and Busan in South Korea. Humanit Soc Sci Commun 9, 23 (2022).

[5]Fan, Chao & Xu, Jeffrey & Natarajan, Barane & Mostafavi, Ali. (2023). Interpretable machine learning learns complex interactions of urban features to understand socio‐economic inequality. Computer-Aided Civil and Infrastructure Engineering. 38. 10.1111/mice.12972.

[6]Chen, Y., Koch, F., & Reicher, C. (2023). Developing a two-level machine-learning approach for classifying urban form for an East Asian mega-city. Environment and Planning B: Urban Analytics and City Science.

[7]Joseph, L. (2008). Finding Space Beyond Variables: An Analytical Review of Urban Space and Social Inequalities. Spaces for Difference: An Interdisciplinary Journal, 1(2).