Infrastructure Proximity and Socioeconomic Outcomes: A Machine Learning Approach to Urban Planning and Social Justice in the Netherlands

Domonkos B. Toth (13567039)
Politics, Psychology, Law and Economics
Data Science and Artificial Intelligence Minor
University of Amsterdam

6012B0419Y: Machine Learning

Dr. Stevan Rudinac, Shuai Wang

December 12, 2024

Word Count: 1179

Introduction

The intricate relationship between urban infrastructure and socioeconomic outcomes represents a critical intersection of public policy consideration with economic development and social justice. In a time of increasing urbanization and widening inequality, understanding the interaction of these areas have become more urgent than ever. The COVID-19 pandemic has further exposed and exacerbated infrastructure-related inequalities, revealing how access to healthcare facilities, green spaces, and digital infrastructure can dramatically impact community resilience and individual outcomes during crises. Moreover, according to recent studies, nearly 70% of the world's population will live in urban areas by 2050, making the equitable distribution of infrastructure resources a pressing global challenge. This holds especially true in the already highly developed countries of Western Europe, including the Netherlands which forms the center of this research.

Thus, this paper aims to investigate how the spatial distribution of infrastructure influences socioeconomic indicators across neighborhoods, while also paying careful attention to potential differences among demographically similar communities. Considering that marginalized communities often lack critical infrastructure, such as schools or hospitals, have fewer businesses and leisure opportunities, while also often experiencing persistent patterns of environmental injustice, there is a high need for considering the impact of these factors on the life of the inhabitants and their future outcomes. For instance, historical redlining practices in the United States continue to influence contemporary patterns of infrastructure development, creating intergenerational cycles of disadvantage that demand innovative analytical approaches to spot and understand. The unprecedented availability of data nowadays present a great opportunity to implement new approaches to analyze these problems and aid decision makers to come up with solutions in order to meaningfully address them.

However, these methods presents both an opportunity and a challenge for urban researchers and policymakers. Many municipal governments now collect vast amounts of real-time data on infrastructure usage, maintenance, and performance. Social media platforms generate continuous streams of user sentiment and behavior patterns. Satellite imagery provides detailed temporal snapshots of urban development. Traditional statistical methods, while valuable, struggle to process and derive meaningful insights from this exponentially growing diverse data sources and their volumes. This is the place where machine learning algorithms can take the spotlight by offering a compelling solution to these analytical challenges and providing actionable insights for policy makers on which they can address intricate societal problems.

2. Preprocessing and Methodology

2.1 Dataset Description

Our analysis is based on an elaborate dataset consisting of 17,681 geographic units with information on 122 variables, hierarchically structured according to Dutch administrative levels. We combined demographic information with six different proximity datasets that capture the accessibility of several urban amenities: supermarkets and commercial facilities, healthcare institutions, educational facilities, public transportation nodes, cultural establishments, and recreational areas. Such a broad approach enables us to analyze in detail the complex interrelations between accessibility and socioeconomic factors.

2.2 Data Integration and Cleaning

Our preprocessing methodology adopted a thorough strategy to ensure the quality and reliability of the data. We started by merging several datasets on 'DistrictsAndNeighbourhoods', which was the primary key in order to keep the hierarchical structure of the geographic units. During our cleaning, we removed 27 columns that had more than 40% missing values since those were not good enough for reliable statistical inference. We have standardized all data to float64 and int64 formats for computational accuracy, followed by standardized missing value representation with NaNs. Quality was ensured by the different verification functions after each and every step in the preceding steps.

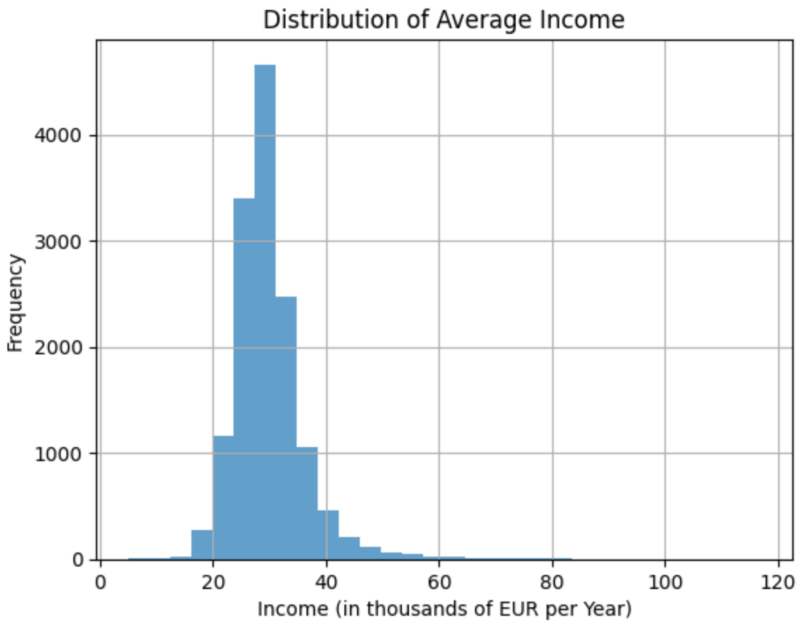
3. Exploratory Data Analysis

3.1 Education Distribution

The level of education is relatively evenly spread out. Medium education level has the highest share, at an estimated 15.5 million. It is followed by high education, with about 11.6 million people, and by low education, which estimates to 10.0 million persons. From this information, it would appear that this is a population that leans toward the higher ends of educational attainment.

3.2 Income Patterns

When analyzing the income distribution pattern, we found a right-skewed normal distribution among neighborhoods. Most annual incomes range from 20,000 to 40,000 EUR, and the peak of the distribution lies between 28,000 and 30,000 EUR. One important feature of this distribution is its long right tail, extending beyond 60,000 EUR, which indicates that there are some high-income neighborhoods without an abundance of extreme outliers.



3.3 Accessibility Correlations

Our analysis showed a strong correlation between income levels and proximity to amenities. Educational facilities had a moderate negative correlation with distance, indicating that neighborhoods closer to schools have higher average incomes. Healthcare access was similarly related, with higher-income areas generally having better access to medical services. Commercial amenities had a weaker but still notable correlation, with the relationship being most pronounced within the first few kilometers of distance.

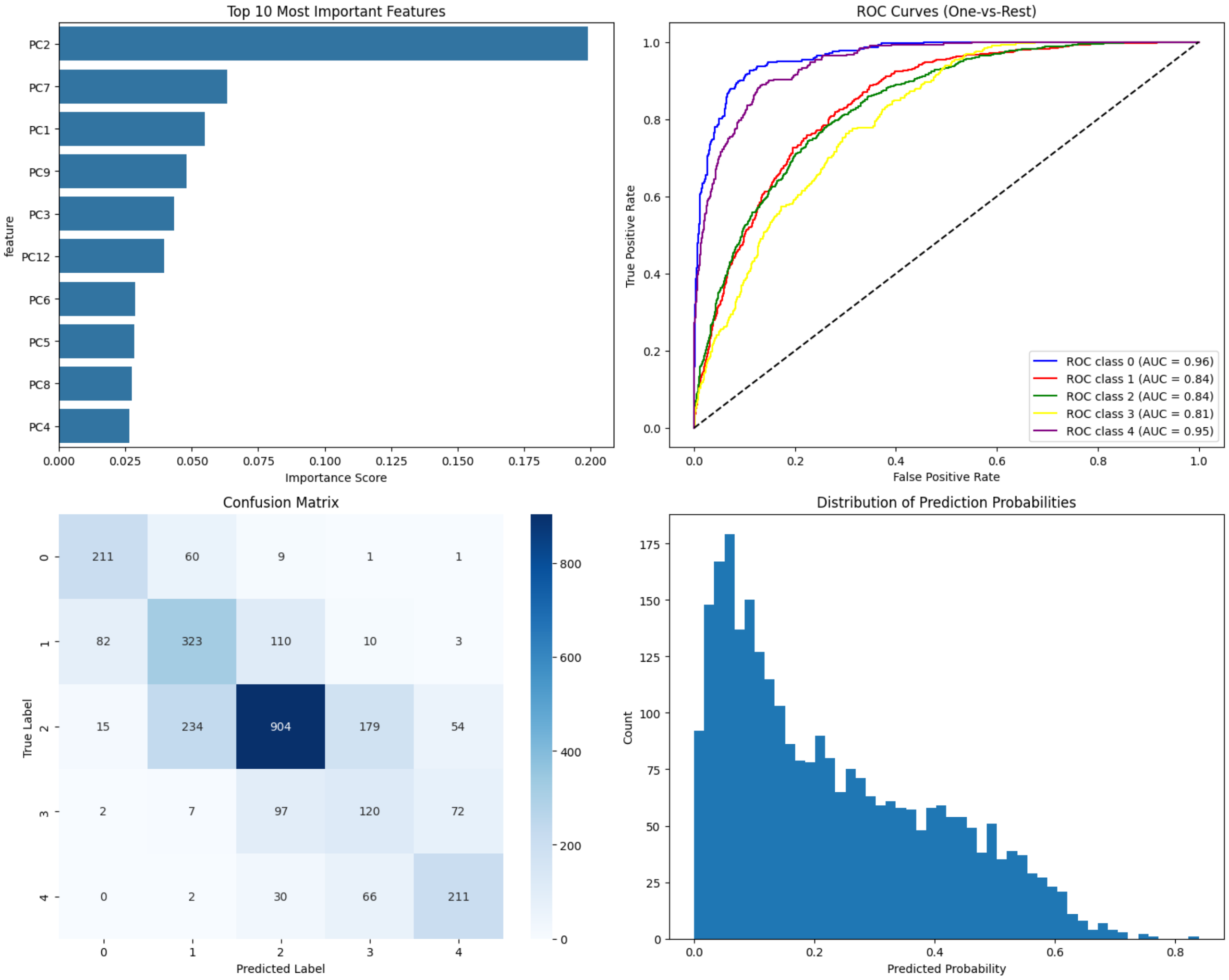
4. Model Development and Results

4.1 Methodology

In the development of these relationships, we implemented and evaluated three different classification algorithms: Random Forest, K-Nearest Neighbors (KNN), and Linear Support Vector Classification (LinearSVC). Each model was integrated into a comprehensive pipeline that incorporated median imputation for handling missing values, standard scaling for feature normalization, and SMOTE techniques for addressing class imbalance issues.

4.2 Model Performance

The Random Forest classifier emerged as our most effective model, achieving an impressive overall F1-score of 0.639. This performance significantly outpaced both the KNN (0.556) and LinearSVC (0.472) models. Our Random Forest implementation demonstrated robust performance across multiple metrics, achieving an accuracy of 0.631, precision of 0.657, and recall of 0.631. When examining class-specific performance, we found varying levels of effectiveness across different income categories. Both categories 0 and 2 obtained outstanding results from the model at



True Label \ Predicted Label	0	1	2	3	4
0	211	60	9	1	1
1	82	323	110	10	3
2	15	234	904	179	54
3	2	7	97	120	72
4	0	2	30	66	211

0.71 F1-score each. Performance is quite similar with a score of 0.65 for category 4, whereas category 1 is quite average at an F1-score of 0.56. Category 3 is the worst performing with a score of 0.36.

4.3 Feature Importance

Through our feature importance analysis, we identified several key predictors that drove model performance. Principal components analysis revealed that PC2, PC7, and PC25 emerged as the most influential features, suggesting that our dimensional reduction effectively captured important patterns in the underlying data. The optimal Random Forest configuration utilized 102 trees with a maximum depth of 14, which we determined through careful hyperparameter tuning.

5. Discussion and Conclsuions

In fact, our results reflect strong predictive capabilities for the two extreme income categories and flag certain challenges in the middle-income class classification. The ROC curve analysis yielded strong discriminative ability, ranging from 0.81 to 0.95 in AUC value for the various income categories. Such results indicate that, though machine learning can well predict neighborhood income categories using the input of accessibility metrics, the linkage between urban accessibility and income itself is complex and multi-faced. We recommend that future research explore additional features and alternative modeling approaches to improve classification accuracy for middle-income categories, as this remains an area with potential for improvement.