Trends in Computational and Applied Mathematics, **27** (2026), e01883
Sociedade Brasileira de Matemática Aplicada e Computacional
Online version ISSN 2676-0029 www.scielo.br/tcam
ORIGINAL ARTICLE
doi: 10.5540/tcam.2026.027.e01883

Social Inequality and Access to Healthcare: An Analysis Using Unsupervised Machine Learning in Greater São Paulo

T. N. VILCHES

Received on August 1, 2025 / Accepted on January 9, 2026

ABSTRACT. The social determinants of health are demographic and socioeconomic characteristics that influence the population's lifestyle and access to healthcare, including the comprehensive and universal care provided by Brazil's Unified Health System (SUS). This study aims to understand how population characteristics are associated with the availability of health services in Greater São Paulo. We applied an unsupervised machine learning technique to analyze how municipalities within the São Paulo Metropolitan Region cluster based on demographic and socioeconomic indicators. We then used the resulting clusters to evaluate the number of healthcare facilities available in each. Our findings indicate that peripheral cities in the region tend to share certain features, such as a higher proportion of Black and Brown populations and lower average income (up to one minimum wage). These clusters tend to have fewer healthcare facilities. In contrast, the cluster consisting solely of São Caetano do Sul, which has the highest proportion of White residents and a large share of individuals earning more than two minimum wages, has, relative to its population, the highest number of healthcare facilities, tripling the number found in the second-ranking cluster (Cluster 2). Through a machine learning approach, our results highlight the structural inequality present in the São Paulo Metropolitan Region and the disparities in access to healthcare, particularly in its peripheral areas.

Keywords: epidemiology, k-means, public health.

1 INTRODUCTION

The socioeconomic and demographic characteristics of a population are, in general, determinants for understanding the development of diseases within it. In fact, these characteristics are called "social determinants of health", typically including income, education, race/color, sex, age, income distribution, and even population density [2]. It is well known that the development of chronic and/or infectious diseases can be associated with these socioeconomic and demographic characteristics, which directly impact people's lifestyles and their access to basic healthcare [4]. Ferreira et al. [7], for example, studied the incidence of cancer in males and concluded that although cancer incidence appears lower in more vulnerable groups, mortality is higher in these

São Paulo State University, Institute of Biosciences, Biodiversity and Biostatistics Department, R. Prof. Dr. Antônio Celso Wagner Zanin, 250, Distrito de Rubião Junior, Botucatu, SP – E-mail: thomas.vilches@unesp.br – https://orcid.org/0000-0002-4980-7564

groups, indicating difficulties in diagnosis. Furthermore, the survival rate of the most vulnerable was lower for all cancers studied.

Brazil's Unified Health System (Sistema Único de Saúde, SUS) offers a service that is universal and comprehensive in nature, with the expectation that access to healthcare is evenly distributed across the Brazilian population. However, according to Bittencourt et al. [1], this access has been "overrun" by a system of healthcare commercialization. Even so, it is possible to observe that access to healthcare has been increasing over the years, for example, in the execution of diagnostic exams for prostate and breast cancer, with the proportion of people seeking these services depending on factors such as individuals' education levels [13].

It is important that authorities reinforce access to public health services in areas where the population has less access to private health plans and services. However, this is not exactly what happens in practice, as the distribution of healthcare facilities in Brazilian cities can be found through DATASUS, in the National Registry of Healthcare Facilities (Cadastro Nacional de Estabelecimentos de Saúde, CNES) [6].

To investigate socioeconomic inequality and its relationship with access to healthcare in the São Paulo Metropolitan Region (Greater São Paulo), a socioeconomic and demographic database was analyzed using an unsupervised machine learning technique, and the results were compared with the distribution of healthcare facilities in the region. The following section presents the methodology used, followed by the results and discussion.

2 METHODS

2.1 Data

For the execution of this work, two different databases were used. Covering the period from January 2015 to December 2019, the National Registry of Healthcare Facilities (Cadastro Nacional de Estabelecimentos de Saúde, CNES), sourced from DATASUS; and demographic data from the 2010 census of the Brazilian Institute of Geography and Statistics (Instituto Brasileiro de Geografia e Estatística, IBGE) [5, 6, 9]. By the time this study was conducted, 2022 census was not fully available.

Data were used only if the healthcare facility was located in one of the 39 cities of Greater São Paulo, according to CETESB [3]. These cities are: São Paulo, Arujá, Barueri, Biritiba-Mirim, Caieiras, Cajamar, Carapicuíba, Cotia, Diadema, Embu das Artes, Embu-Guaçu, Ferraz de Vasconcelos, Francisco Morato, Franco da Rocha, Guararema, Guarulhos, Itapecerica da Serra, Itapevi, Itaquaquecetuba, Jandira, Juquitiba, Mairiporã, Mauá, Mogi das Cruzes, Osasco, Pirapora do Bom Jesus, Poá, Ribeirão Pires, Rio Grande da Serra, Salesópolis, Santa Isabel, Santana de Parnaíba, Santo André, São Bernardo do Campo, São Caetano do Sul, São Lourenço da Serra, Suzano, Taboão da Serra, and Vargem Grande Paulista.

The data extracted from IBGE were used to characterize the cities according to socioeconomic and demographic features, such as the population's age distribution, race/color distribution, income distribution, and sex distribution. Table 1 shows the variables used in each database.

Table 1: Variables used in the database.

Database	**Variable**	**Groups**
IBGE Census 2010	Age distribution	0 to 4 years, 5 to 9 years, 10 to 14 years, 15 to 17 years, 18 to 19 years, 20 to 24 years, 25 to 29 years, 30 to 34 years, 35 to 39 years, 40 to 49 years, 50 to 59 years, 60 to 69 years, 70 years or older
IBGE Census 2010	Race/color distribution	Black, White, Brown (Pardo), East Asian, Indigenous, and Not declared
IBGE Census 2010	Literacy rate	-
IBGE Census 2010	Income distribution	Up to half minimum wage, from half to one minimum wage, 1 to 2 minimum wages, 2 to 5 minimum wages, 5 to 10 minimum wages, 10 to 20 minimum wages, more than 20 minimum wages, no income

2.2 Exploratory Analysis

Initially, through Bartlett's sphericity test [16], it was verified that the correlation matrices of the sociodemographic variables from IBGE are statistically different from the identity matrix (which would represent uncorrelated variables). Pearson correlation was used to analyze the association between variables [8], and principal component analysis (PCA) was applied to the data [8], aiming to create uncorrelated variables and reduce the dimensionality of the problem for cluster analysis. PCA is an unsupervised technique which, if a new observation is added to the dataset, must be rerun; that is, it is not used for predictions [8]. PCA was employed to enhance the visualization of the clusters. The specific details regarding it will be presented in the Results section to provide context for the analysis.

Using Z-score standardized data, a cluster analysis of the 39 cities in Greater São Paulo was performed using the k-means technique. In this technique, the researcher defines the number k of clusters to be constructed, and the algorithm seeks to find k centroids in the n-dimensional space (where n is the number of variables used) that minimize the sum of distances between each cluster centroid and the observations belonging to that cluster [8].

An alternative for selecting the number of clusters is the silhouette method. This metric evaluates the quality of clustering by measuring the cohesion (similarity to its own cluster) compared to the separation (difference from the nearest neighboring cluster) for each observation. A silhouette

value closer to 1 indicates a better clustering allocation [15], meaning that the observation is well-matched to its own cluster and poorly matched to neighboring ones. In this work, the silhouette method and the `stats` package of the R software were used for the cluster analysis [12].

The characteristics of the groups found by the clustering technique were analyzed through the standardized difference between means. Let X_{ij} be the mean of variable j in cluster i, and X_j and s_j be the mean and standard deviation of variable j among all 39 municipalities of interest. The standardized difference between means is given by

$$\Delta\mu = \frac{X_{ij} - X_j}{s_j}. \tag{2.1}$$

This means that the further $\Delta\mu$ is from zero, the further the cluster is from the global mean; when positive, the cluster's mean is higher than the global mean (of the 39 municipalities), and when negative, it is lower.

3 RESULTS

Using the normalized data, principal component analysis was performed. To determine the number of principal components to retain, we applied the Kaiser criterion [8], which suggests keeping only components with eigenvalues greater than 1. Eigenvalues represent the total amount of variance from the original data explained by each component, serving as a measure of their informational weight. They are calculated based on the correlation matrix of the standardized variables. The rationale is that a principal component should explain at least as much variance as a single standardized variable. Components with eigenvalues below this threshold provide less information than an original variable and are therefore discarded. Figure 1 shows the eigenvalues that are calculated and the proportion of variance that is explained by each principal component. For visualization purposes of the clusters, only principal components 1 and 2, which already explain 70% of the data variance, were used.

A correlation analysis between the principal components and the original variables shows that principal component 1 is associated with age-related variables, being positively correlated with the proportion of younger individuals, the Black and Brown (Pardo) population, and the low-income population (up to two minimum wages), while the second principal component 2 is positively and strongly correlated with adults aged 25 to 39 years and the literacy rate.

Using the k-means clustering technique, a cluster analysis was performed among the municipalities of Greater São Paulo, simulating 50 different initial conditions for the position of the centroids. The number of clusters was chosen as $k = 6$, which corresponds to a local maximum of the average silhouette width measure (see Figure 2A), since empirically two clusters, suggested by the global maximum, did not provide a good description of the data. The clustering quality was visually assessed using the first two principal components. Figure 2B shows a distinct separation among clusters, even within this reduced two-dimensional projection.

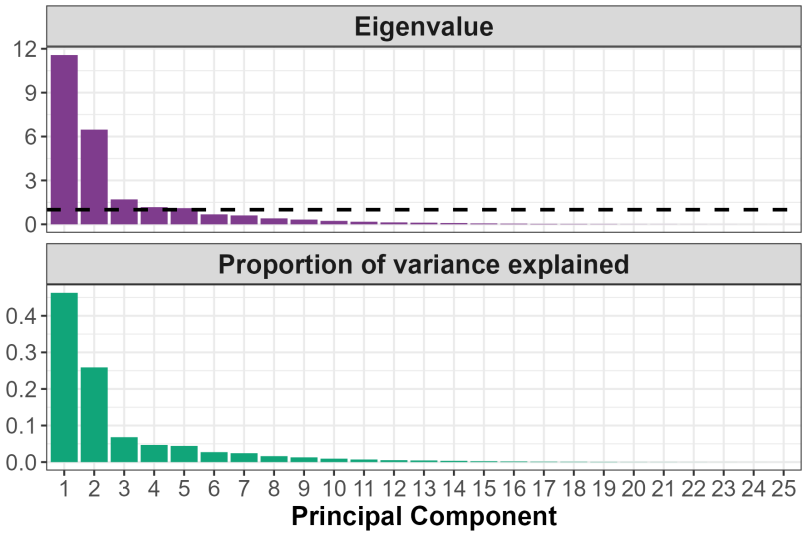

Figure 1: The eigenvalues associated with each Principal Component and the proportion of the variance of the data that is explained by each Principal Component.

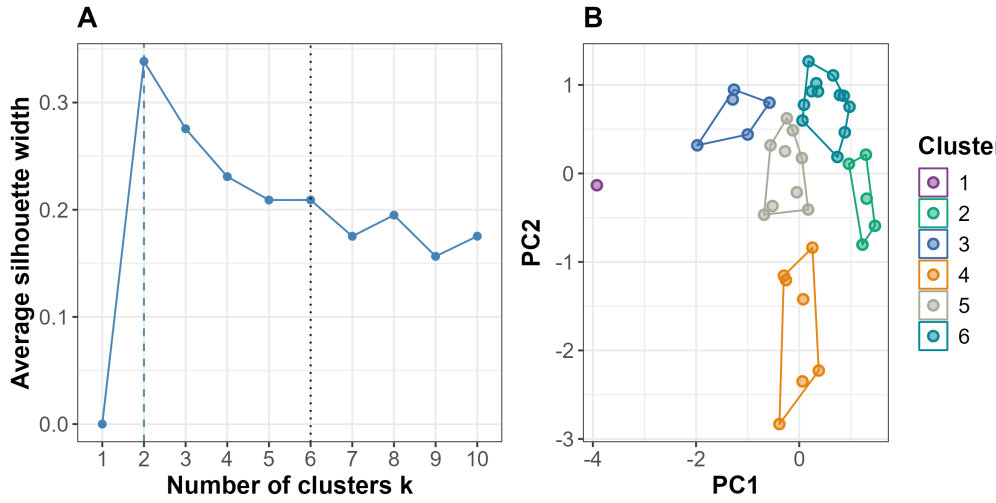

Figure 2: In A: average silhouette width in function of the number of clusters. The dashed line shows the global maximum and the dotted line indicates the local maximum that was used in the analysis. In B: the two-dimensional projection of the observed data using the Principal Component Analysis performed.

Figure 3 shows the standardized difference of cluster means relative to the overall mean. Cluster 1, consisting solely of the city of São Caetano do Sul, is associated with a population over 40 years old and a proportion of individuals with income greater than two minimum wages above the Greater São Paulo average. It is also the cluster with the highest proportion of White population. Cluster 3 is associated with a younger, Brown (Pardo) population, and at the same time with income lower than two minimum wages. Clusters 1, 2, and 6 are associated with income greater than two minimum wages; however, it is noted that in Cluster 6, the proportion of people earning more than 20 minimum wages is actually higher than in the other income groups, while in Clusters 1 and 2 the lower income brackets have a greater proportion, decreasing as income increases. The city of Santana de Parnaíba, where the Alphaville neighborhood is located, for example, is in Cluster 6.

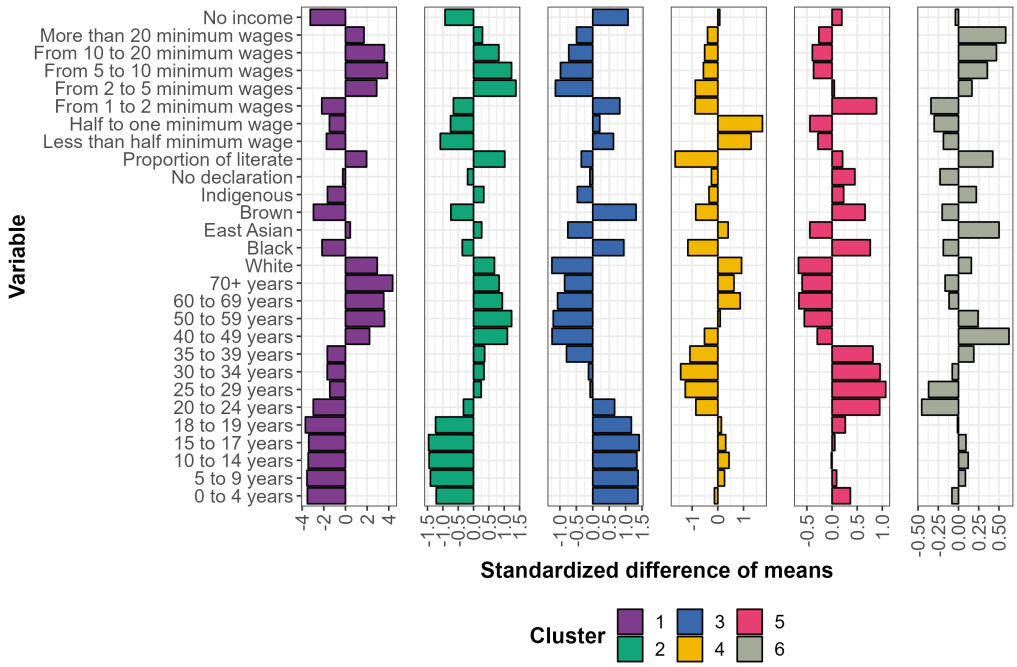

Figure 3: Standardized difference between the cluster means and the overall mean for each original variable in the study.

Figure 4 shows the geographic distribution of the clusters. It is noted that Cluster 4, associated with lower income (up to one minimum wage), is distributed in the cities farther from the city center (peripheral areas).

Figure 5 shows the number of healthcare facilities per group found by the k-means method per 100,000 inhabitants, considering the population data from the 2010 IBGE census. It can be seen that Cluster 1 has, proportionally, an extremely high number of facilities, reaching almost three times the number of Cluster 2, the second largest one. Cluster 1, as mentioned, is composed

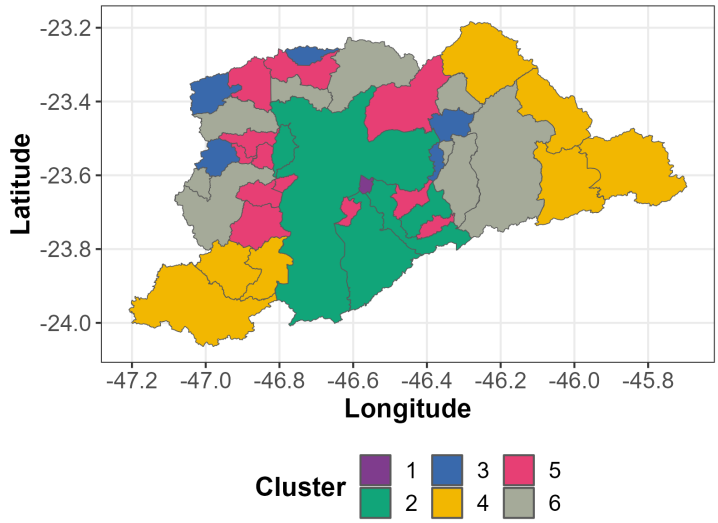

Figure 4: Geographical distribution of the clusters identified through the k-means method.

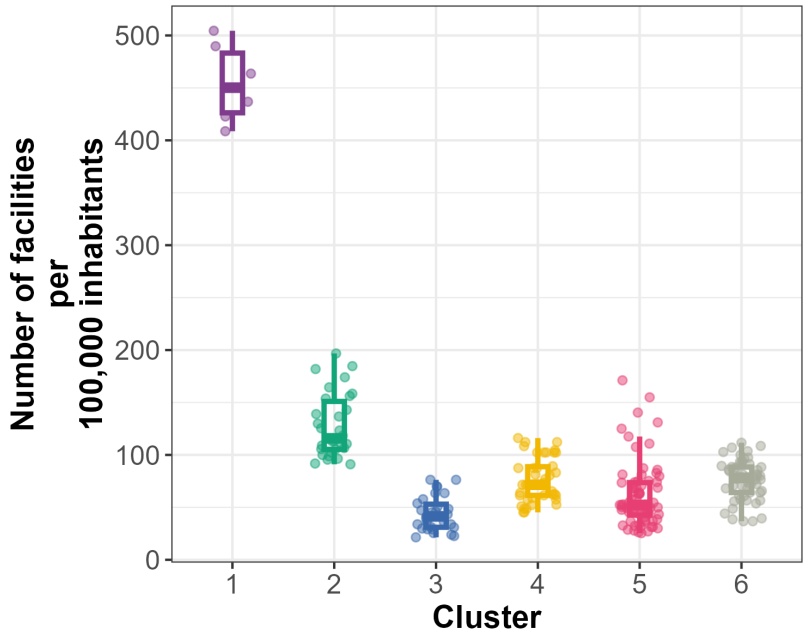

Figure 5: Number of healthcare facilities per 100,000 inhabitants in the different clusters identified. Each point represents a municipality in a year between 2015 and 2020.

solely of the city of São Caetano do Sul, associated with a higher proportion of people earning above five minimum wages and a White population.

4 DISCUSSION

This work presents an analysis of the 2010 demographic census data [9], focusing especially on the so-called social determinants of health. The World Health Organization estimates that 30% to 55% of disease occurrence is related to socioeconomic risk factors [11].

Through PCA, it was possible to reduce the dimensionality of the dataset, with Principal Component 1, which explains the largest proportion of data variance, being associated with a higher proportion of young population, low income, and a higher proportion of Black and Brown (Pardo) population. The analyses also showed that this proportion of Black and Brown individuals in the population is positively and strongly correlated with the proportion of people earning up to two minimum wages, confirming the findings of Santos [14].

The Z-score standardized data were used to perform a cluster analysis, which resulted in six distinct clusters. The clusters found presented a geographical organization; for example, Cluster 4, associated with a higher proportion of people with income up to one minimum wage, is located on the outskirts of the São Paulo Metropolitan Region, farther from the city center, reinforcing the results found by Lima et al. [10] which show that the urban health index decreases in the outskirts of São Paulo city, with special attention to the East Zone and the far South. This also emphasizes the geographical characteristic of income distribution occurring in the region [18].

It is important to note that there is evidence of a relationship between the number of healthcare facilities and the sociodemographic characteristics of the cities. The city of São Caetano do Sul, which was selected as a separate cluster by the k-means technique, also has the highest number of healthcare facilities (per 100,000 inhabitants). This result is supported by Tomasiello *et al.* [17], who analyzed the distance between individuals and healthcare facilities, stratified by race and income. They found that Black individuals have lower access to high-complexity healthcare facilities compared to White individuals, as these facilities are generally concentrated in specific areas. While their approach relies on analyzing high-spatial-resolution data within specific municipalities, ours applies machine learning techniques to lower-resolution data (municipality level) across a Metropolitan Area. Both approaches are valid and complementary.

It is important that authorities understand the demand and availability of services, especially considering that part of the population cannot afford private service costs. Such disparities in access to services can generally deepen the occurrence of diseases, for example infectious ones, and delay the diagnosis of chronic diseases whose late diagnosis can worsen the patient's condition, such as cancer [7, 13].

This work provides an analysis of the socioeconomic and demographic characteristics of the metropolitan region of São Paulo city, reinforcing disparities in access to health services. Despite results somewhat previously discussed by other authors, the use of unsupervised machine learning techniques highlights disparities among the cities and, when associated with the geo-

graphic distribution and number of healthcare facilities, reinforces the relationship between the metropolitan periphery, difficulty of access to healthcare, and socioeconomic and demographic characteristics. It is, therefore, evidence to support authorities' work toward better distribution of services and resource allocation, providing a direct strategy to target groups of municipalities more affected by socioeconomic disparities and limited healthcare access, such as Cluster 3.

Acknowledgments

The authors declare no conflict of interest. The author thanks Professor Adâmara Felício for the fruitfull discussions. All the results were generated and interpreted by the authors. Generative AI was used only to improve the English writing and enhance the clarity and flow of the text.

Data availability

The data were processed and uploaded in a relational database and can be found at `https://github.com/thomasvilches/TCAM_2026_healthcare_access`.

Associate editor: Luiz Rafael Santos

REFERENCES

[1] L.D.J. Bittencourt, K.D.S.O. Santana & D.S.M. Santos. Saúde da população negra na atenção primária: incompreensão que legitima iniquidade em tempos de Covid-19. *Saúde em Debate*, **47**(137) (2023), 31–41. doi:10.1590/0103-1104202313702. URL `http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-11042023000200031&tlng=pt`.

[2] P.M. Buss & A. Pellegrini Filho. A saúde e seus determinantes sociais. *Physis: revista de saúde coletiva*, **17** (2007), 77–93.

[3] CETESB. Municípios que fazem parte Região Metropolitana de São Paulo » Licenciamento Ambiental (2014). URL `https://cetesb.sp.gov.br/licenciamentoambiental/licenca-previa-documentacao-nescessaria/municipios-que-fazem-parte-regiao-metropolitana-de-sao-paulo/`. Accessed on 2023-10-16.

[4] W.C. Cockerham, B.W. Hamby & G.R. Oates. The Social Determinants of Chronic Disease. *American Journal of Preventive Medicine*, **52**(1) (2017), S5–S12. doi:10.1016/j.amepre.2016.09.010. URL `https://linkinghub.elsevier.com/retrieve/pii/S0749379716304408`.

[5] G.C. Coelho Neto & A. Chioro. Afinal, quantos Sistemas de Informação em Saúde de base nacional existem no Brasil? *Cadernos de Saúde Pública*, **37** (2021), e00182119.

[6] DATASUS. Transferência de Arquivos (2023). URL `https://datasus.saude.gov.br/transferencia-de-arquivos/`. Accessed on 2023-10-16.

[7] M.D.C. Ferreira, I. Arroyave & M.B.D.A. Barros. Desigualdades sociais em câncer no sexo masculino em uma metrópole da região Sudeste do Brasil. *Revista de Saúde Pública*, **57**(1)

(2023), 38. doi:10.11606/s1518-8787.2023057004712. URL https://www.revistas.usp.br/rsp/article/view/214004.

[8] L.P. Fávero & P. Belfiore. "Manual de análise de dados: estatística e modelagem multivariada com Excel, SPSS e Stata". Elsevier Brasil (2017).

[9] IBGE. Censo 2010 (2010). URL https://www.ibge.gov.br/estatisticas/sociais/populacao/9662-censo-demografico-2010.html?edicao=10411&t=downloads. Accessed on 2023-10-10.

[10] O.A.C.P.d. Lima, E. Kruger & M. Tennant. São Paulo urban health index: measuring and mapping health disparities. *Revista Brasileira de Epidemiologia*, **25** (2022), e220005.

[11] Organização Mundial da Saúde. Social determinants of healh (2024). URL https://www.who.int/health-topics/social-determinants-of-health. Accessed on 2025-02-05.

[12] R Core Team. Stats (2024). URL https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html. Accessed on 2024-06-15.

[13] E.F.D.S. Santos, C.N. Monteiro, D.B. Vale, M. Louvison, M. Goldbaum, C.L.G. Cesar & M.B.D.A. Barros. Social inequalities in access to cancer screening and early detection: A population-based study in the city of São Paulo, Brazil. *Clinics*, **78** (2023), 100160. doi:10.1016/j.clinsp.2022.100160. URL https://linkinghub.elsevier.com/retrieve/pii/S1807593222033610.

[14] J.A.F. Santos. Desigualdade racial na transmissão intergeracional da herança de classe social. *Sociologias*, **24**(59) (2022), 328–360.

[15] D.M. Saputra, D. Saputra & L.D. Oswari. Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method. In "Sriwijaya international conference on information technology and its applications (SICONIAN 2019)". Atlantis Press (2020), p. 341–346.

[16] S. Tobias & J.E. Carlson. Brief report: Bartlett's test of sphericity and chance findings in factor analysis. *Multivariate behavioral research*, **4**(3) (1969), 375–377.

[17] D.B. Tomasiello, J.P.B. Vieira, J.P.F. Parga, L.M. Servo & R.H. Pereira. Racial and income inequalities in access to healthcare in Brazilian cities. *Journal of Transport & Health*, **34** (2024), 101722.

[18] United Nations Human Settlement Programme. São Paulo: A tale of two cities (2010). URL https://unhabitat.org/sites/default/files/download-manager-files/Sao%20Paulo%20A%20tale%20of%20two%20cities.pdf. Accessed on 2025-02-05.

How to cite

T.N. Vilches. Social Inequality and Access to Healthcare: An Analysis Using Unsupervised Machine Learning in Greater São Paulo. *Trends in Computational and Applied Mathematics*, **27**(2026), e01883. doi: 10.5540/tcam.2026.027.e01883.