Project:

Gov Sananga - Mapping and Analysis of Companies in Estrutural City with Machine Learning and OSINT

Author: Roberto Moreira Diniz.

Advisor: Simone de Araújo Góes Assis.

Institution: IESB University Center.

Course: Data Science and Artificial Intelligence.

City: Brasília,Brazil.

Year: 2024

The "Gov Sananga" project seeks to map and analyze companies in Estrutural City using technologies such as machine learning, data scraping, and OSINT. The research aims to collect detailed information on economic activities in the region, which suffers from the informality of the labor market. Data collection will be done through Google Street View images and Google APIs to provide a database of local companies, offering insights for public policies that promote formalization and economic development.

1. Introduction.

The informal labor market has grown significantly in recent years, especially in low-income urban and peripheral areas. According to data from the International Labor Organization (ILO), informality accounts for a substantial portion of economic activity in developing countries, generating jobs without guarantees of social protection, regulation, or formal labor rights. This is especially evident in areas such as the Estrutural city in Brazil.

1.1 Choosing a project name.

Sananga is a traditional eyewash indigenous people use to bring clarity of vision. The name symbolizes the project's intention to clarify and bring to light the reality of informal businesses. The focus is to help formalize and sustainably grow these economic activities. Just as Sananga promotes healing, the project seeks to transform the local economy by guiding public policies.

Mapping companies such as those in Cidade Estrutural through unconventional mechanisms can be of great value when compared to official databases in understanding the growth of the informal market and its impacts. The lack of structured data on these economic activities makes it difficult to formulate public policies and government actions that can support the formalization and grofth of these companies. In this context, technologies such as machine learning, data scraping, and OSINT (Open Source Intelligence) can play a fundamental role in collecting and analyzing information, contributing to a better understanding of the local economy. The "Gov Sananga" project aims to map and analyze formal and informal companies in Cidade Estrutural, using advanced machine learning techniques, scraping tools, and open data. TThesechnologies aim to create a detailed and robust database that allows the visualization and analysis of these economic activities, contributing to developing more assertive public policies.

2. Theoretical Framework

2.1 Machine Learning and Image Analysis in Real Scenarios

The use of machine learning to analyze visual and spatial data has been widely applied in various sectors, providing valuable insights into social, environmental, and economic dynamics. Wang et al. (2022) conducted a study that explores how Google Street View (GSV) images, in conjunction with machine learning techniques, can be used to predict changes in real estate prices. This study highlights the practical application of computer vision and urban image analysis to capture information about the quality of the urban environment, infrastructure, and the socioeconomic profile of areas. In real scenarios, this approach allows for monitoring urban development and making decisions based on accurate data, such as managing public housing and urban development policies. In addition, applying machine learning to visual data offers an efficient way to interpret variables that affect real estate values, such as the presence of green areas, road quality, and proximity to essential services.

Another relevant example is the study by Cai et al. (2022), which explored the use of deep learning and Google Street View images to analyze the impact of drivers' visual environment on traffic accidents. In this case, the image segmentation technique was used to identify and quantify elements such as trees, buildings, and roads, which were later correlated with traffic safety. The research revealed that the presence of trees and the visual complexity of the environment can directly influence driver behavior and the frequency of accidents. This type of application of machine learning for road safety analysis is a clear example of how the technology can be used in urban planning and transportation management scenarios, providing detailed information on the factors that affect road safety.

2.2 Technologies for Analysis of Markets and Urban Environments.

Visual data analysis, such as images captured by services like Google Street View, has been integrated with advanced machine learning techniques to study urban environments and their relationship with different economic sectors. Wang et al. (2022) demonstrate that the use of these technologies can assist in the analysis of the real estate market by providing a more accurate assessment of the infrastructure and quality of the environment. This analysis is highly relevant in real-world scenarios, such as in the evaluation of zoning policies, improvements in urban infrastructure, and land and property pricing. Similarly, Cai et al. (2022) applied machine learning techniques, such as object detection, to measure the impact of the urban environment on traffic safety. The results indicate that city planning can be improved by analyzing visual data that identifies environmental elements that influence driver behavior. The

solutions proposed from studies like this can be applied to guide urban planners in creating safer infrastructures, using insights generated by these visual data analyses.

2.3 Impact of Technologies on Urban Planning and Security

These examples highlight the potential of computer vision and machine learning technologies in real-world sectors such as urban planning and road safety. Using deep learning algorithms to process and interpret visual data captured in cities offers an opportunity to significantly improve public policy management, particularly in areas related to housing, transportation, and security. In addition, automating these analyses reduces costs and improves the efficiency of data collection and analysis, facilitating evidence-based decision-making.

In short, using technologies such as machine learning, computer vision, and visual data analysis consolidates itself as a powerful tool for transforming real scenarios and providing data-based solutions that contribute to the development of intelligent and safer cities.

2.4 Informality in the Labor Market and Companies in Low-Income Communities

In the article by Badaoui et al. (2023) explores the relationship between informality, self-employment and heterogeneous managerial skills in developing countries. They suggest that many individuals opt for informal self-employment due to the lack of adequate managerial skills to compete in the formal sector. The decision to maintain businesses in the informal sector is not only a matter of survival, but also a choice based on human capital and market access limitations.

Informality in the self-employment sector is associated with individuals with limited managerial skills who prefer to avoid the fixed costs of formality. Informal firms tend to be less productive and more volatile, but they offer greater flexibility to workers who lack the skills needed to compete in highly regulated environments. The low requirement for formalization and the flexibility of self-employment are attractive to workers in low-income communities, where formal education is scarcer and formal employment opportunities are limited. Many micro and small businesses choose to remain informal due to the costs and bureaucracy involved in formalization. However, Badaoui et al. (2023) point out that these firms end up trapped in a cycle of low productivity, as the lack of access to formal credit and government programs impedes sustainable growth. The lack of institutional and infrastructural support also forces many informal entrepreneurs to remain on the margins of the formal economy.

Impact of Informality in Low-Income Communities In low-income communities, labor market informality is even more prevalent due to difficulties in

accessing quality education and formal employment opportunities. Badaoui et al. (2023) argue that informality, in these contexts, becomes a natural outlet for those with few qualifications or access to capital. However, this perpetuates inequality and limits economic growth, since these companies and workers do not have access to mechanisms that boost competitiveness in the formal market.

2.5 The study by Bosch and Esteban-Pretel (2012), Job creation and destruction in the context of informal markets.

This study, focusing on transitions between the formal and informal sectors in developing economies, examines how employment volatility affects workers and firms in the informal sector, highlighting barriers to formalization.

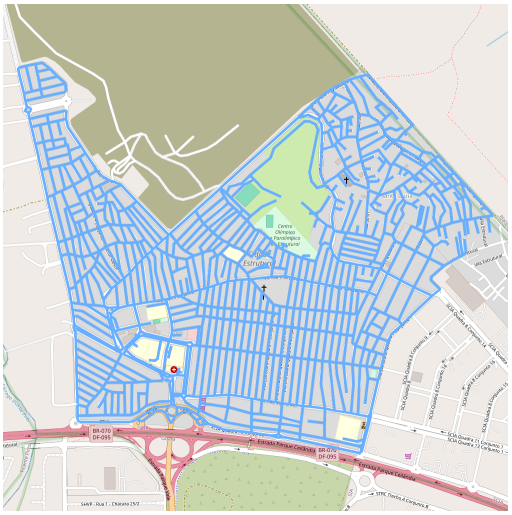
The informal labor market has significantly higher turnover rates than the formal sector. Bosch and Esteban-Pretel (2012) point out that the separation rate for informal jobs is three times higher than that for the formal sector. Job instability is a central feature of the informal sector, especially in low-income communities with greater economic vulnerability. The authors note that many workers transition from the informal to the formal sector during economic expansion. However, in times of crisis, the informal sector acts as a "buffer," absorbing unemployed workers who cannot find jobs in the formal market. This dynamic highlights the role of informality as an informal regulator of the labor market, but it also reveals the instability and precariousness associated with these jobs.

Bosch and Esteban-Pretel (2012) suggest that public policies aimed at reducing the costs of formality can increase the transition of workers to the formal sector. However, caution is needed, as increased formalization can, paradoxically, increase wage inequality since the informal sector tends to absorb less qualified and lower-paid workers. Policies that seek to integrate these populations into the formal market need to consider both workforce training and the creation of a more flexible regulatory environment for small businesses.

3. Metodologia de Pesquisa

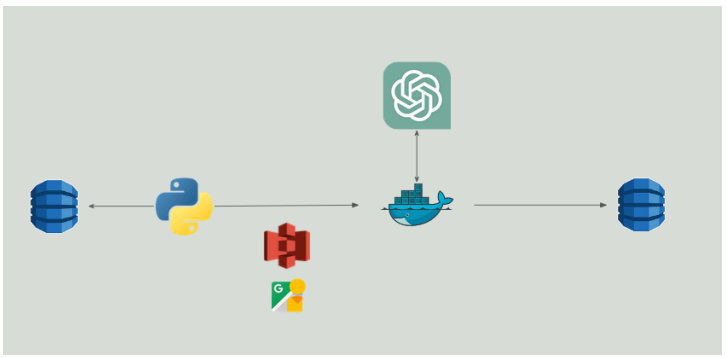
• **Data Collection**: The OSMnx library was used to obtain valid coordinates, and the data generated was in Geopandas format. After cleaning the data, it was organized into a DataFrame (df). The coordinates were grouped based on start and end nodes, and a mathematical algorithm was employed to order these coordinates efficiently. This is crucial because when using the Google Street View API, it is necessary to provide the camera position. An additional calculation will be performed based on the

north orientation and the following coordinate, which will ensure a more accurate capture of storefronts for analysis.



```
In [3]: df
Out[3]:
       start_node    end_node                 coordinates                                  name
0        364895814   364895870  (-15.7846454, -47.9953589)          SCIA Quadra 1 - Setor Leste
1        364895814   364895870  (-15.7826915, -47.9951213)          SCIA Quadra 1 - Setor Leste
2        364895814  1558095208  (-15.7846454, -47.9953589)                    Avenida Comercial
3        364895814  1558095208  (-15.7846809, -47.9956666)                    Avenida Comercial
4        364895815  1558095159  (-15.7814874, -47.9949775)                                  NaN
...            ...         ...                         ...                                  ...
5824  10860455442  2506309838  (-15.7721571, -48.0035443)  Quadra 8 Conjunto 4 Vila Estrutural
5825  11909206374  1558099654    (-15.7849849, -48.00124)                                  NaN
5826  11909206374  1558099654  (-15.7846467, -48.0012524)                                  NaN
5827  11909206374  1558099672    (-15.7849849, -48.00124)                                  NaN
5828  11909206374  1558099672  (-15.7850312, -48.0012383)                                  NaN
```

• **Análise:** As imagens serão analisadas por meio da aplicação do GPT, que irá estruturar informações sobre as lojas com fachadas voltadas para a rua. O prompt utilizado para essa operação já foi testado com sucesso, assim como a integração com a API do Google Street View, garantindo a extração eficiente de dados comerciais relevantes.



Expected Results:

• **Detailed Mapping:** Creation of a robust database that includes both formal and informal companies in Estrutural City. The mapping will allow a more accurate analysis of the region's economic activity.

• **Scalable Tool:** Implementing an infrastructure based on K8s, Terraform, and Google APIs will provide a scalable tool that can be replicated in other regions that present similar challenges in mapping the informal economy.

• **Economic Insights:** The analysis of the collected data will provide essential insights that can be used to support studies on the informal market and assist in creating public policies that seek to formalize and support informal companies, contributing to the economic and social development of these regions.

3.2. Description of the Analysis Model to be Implemented

The analysis model will use a combination of machine learning and a GPT API to identify and classify businesses. The API will process images captured by Google Street View, using image segmentation to detect storefronts. The model will be adjusted to identify economic sectors, such as salons, markets, and bars, to help profile stores for future analysis and strategic decision-making.

To execute the model, the GPT API will be used for textual analysis and company categorization. This API will be executed within a Docker container, which will be orchestrated in a Kubernetes (K8s) cluster. The K8s-based infrastructure will ensure scalability and efficient resource management for image and data processing.

This setup will allow the processing of large volumes of data and adapting to the project's needs, automating image analysis and company categorization and ensuring a robust infrastructure for future expansions.

3.3. Description of the Dictionary of Variables to be Selected in the Model, the Database, and the Information Collection Method (Updated)

Selected Variables:

- **Image ID:** Unique identifier for each image captured from Google Street View, which will be used for reference and tracking during analysis.
- **Image Address in S3:** Path of the image stored in Amazon S3, which will be used to access and analyze the images. Example: s3://bucket-name/path/to/image.jpg.
- **Geographic Coordinates:** Latitude and longitude of the commercial establishments.

- **Store Name:** Name captured from the storefronts, which can be used later for NLP (Natural Language Processing) analysis, such as classifying the type of business or detecting linguistic patterns associated with the economic sector.
- **Start and End Node:** Start and end points that represent the geographic positioning of the stores on the main roads of Cidade Estrutural. These nodes will be used to organize and order the coordinates, ensuring a coherent spatial analysis and facilitating navigation through the mapped areas.

By including the image address in S3, accessing and processing visual data becomes more efficient, integrating cloud storage with subsequent analysis.

3.4. Description of the Analyses to be Performed

Exploratory Data Analysis (EDA) will investigate the distribution of companies, types of businesses, and location patterns in Estrutural City, using graphs, tables, and spatial analysis to identify areas with the highest concentration of informal activities. The study will focus on using this information and services to create a tool that allows monitoring and generating metrics about the ecosystem, facilitating decision-making based on informal activities in the region.

Bibliography:

• Wang, L., Wu, X., & Fan, Y. (2022). The effect of environment on housing prices: Evidence from the Google Street View. Journal of Forecasting, 41(4), 1-18.

• Cai, Q., Abdel-Aty, M., Zheng, O., & Wu, Y. (2022). Applying machine learning and Google Street View to explore the effects of drivers' visual environment on traffic safety. Transportation Research Part C, 135, 103541.

• Bosch, M., Esteban-Pretel, J. (2012). Job creation and job destruction in the presence of informal markets. Journal of Development Economics, 98, 270–286.

• Badaoui, E., Strobl, E., Walsh, F. (2023). Informality, self-employment, and heterogeneous managerial ability: A model for developing countries. Journal of International Development.