Data Science Program

Capstone Report - Spring 2024

Mapping Deprived Areas in Low- and Middle-Income Countries (LMIC)

Daqian Dang
Akhil Bharadwaj

supervised by
Dr. Amir Jafari

Abstract

In low- and middle-income countries (LMICs), understanding the spatial distribution of deprived areas is crucial for effective policymaking, resource allocation, and targeted interventions aimed at reducing poverty and improving socio-economic conditions. However, traditional methods of identifying deprived areas often rely on outdated or limited data sources, leading to incomplete or inaccurate assessments. To address this challenge, our project proposes the development of a comprehensive mapping framework for identifying deprived areas in LMICs using geospatial analysis and machine learning techniques. Leveraging satellite imagery, census data, and other geospatial datasets, our approach aims to create high-resolution maps that capture key indicators of deprivation, such as access to basic services, infrastructure, and socio-economic status. In this study, we demonstrate to map deprived areas in two African cities- Lagos (Nigeria) and Nairobi (Kenya). Contextual features and reference files were extracted at 10-meter spatial resolution and aggregated to a 100-meter grid. By integrating advanced spatial analytics and machine learning algorithms, we seek to enhance the accuracy and granularity of deprivation mapping, enabling policymakers and stakeholders to better target resources and interventions to areas in greatest need. The resulting maps will provide valuable insights into the spatial distribution of deprivation, facilitate evidence-based decision-making, and support efforts to achieve sustainable development goals in LMICs.

Contents

Introduction

The prevalence and growth of slums within low- and middle-income countries (LMICs) have emerged as a considerable challenge, particularly in the context of the accompanying economic crises. Slums, often characterized by inadequate housing, overcrowding, and lack of access to clean environments, represent some of the most visible manifestations of urban poverty and inequality. The pandemic has exacerbated those conditions highlighting the vulnerabilities of slum dwellers to health and economic shocks. The economic downturn triggered by the pandemic has further strained the limited resources available for communities, pushing more individuals into conditions of poverty and precarious living situations. This situation presses the urgent attention for focused interventions and support to address the challenges faced by slum residents, not only to combat the immediate impacts of the pandemic but also to improve their long-term resilience and well-being.

The accurate mapping of deprived areas, including slums, is crucial for monitoring progress towards the Sustainable Development Goals (SDGs), which aims to make cities and human settlements inclusive, safe, resilient, and sustainable. Detailed mapping provides essential data that can inform policymaking and resource allocation, ensuring that interventions are targeted effectively to address the needs of the most vulnerable populations. It also facilitates the monitoring of changes over time, enabling stakeholders to access the impact of interventions and adjust strategies accordingly as necessary. Furthermore, accurate mapping supports the identification of gaps in service provision and infrastructure, guiding efforts to improve access to essential services such as clean water, healthcare, and education. Therefore, the task of mapping deprived areas is not just a technical challenge but a critical step towards achieving equity and inclusion in urban development.

Problem Statement

This capstone project's primary objective is to map deprived areas in low- and middle-income countries by developing a comprehensive geospatial analysis framework. This framework will be enriched with machine learning and deep learning models, mainly focusing on African cities, specifically Lagos, Nigeria and Nairobi, Kenya. This project will utilize open geospatial data sources, including Google Maps Engine and OpenStreetMap. The best-performing model will be analyzed based on results obtained by using the same model in other cities and ranking them based on performance metrics such as F1 score. This model should be able to generalize to other cities for identifying deprived areas.

Related Work

The integration of big data and geospatial analysis is revolutionizing urban planning and environmental research, presenting a myriad of opportunities alongside formidable challenges. A review of recent literature indicates common difficulties such as managing massive and noisy datasets, integrating varied data sources, addressing privacy issues, and requiring advanced technical know-how and cross-disciplinary efforts. Lee and Kang (2015) [1] stress the need for advanced algorithms to process and store the ever-growing and diverse geospatial big data. Robinson et al. (2017) [2] highlight the challenges of big data in cartography, focusing on synthesizing large data volumes into coherent and comprehensible visual formats. Lehner et al. (2020) [3] discuss the hurdles in distilling useful insights from sprawling, unstructured datasets and the integration challenges within urban environments. Koldasbayeva et al. (2023) [4] explore geospatial modeling in environmental research, tackling issues like imbalanced datasets and spatial autocorrelation, as well as the importance of accurately quantifying uncertainty in predictions. Runfola et al. (2024) [5] demonstrate the application of deep learning to estimate socioeconomic factors from satellite imagery, addressing the variance in geographic scope. Collectively, these works underscore the necessity for sophisticated data management, analytical approaches, and interdisciplinary collaboration to fully exploit geospatial big data to improve urban living conditions and environmental studies.

In the project under discussion, a multifaceted approach is employed, starting with an in-depth geospatial analysis of Lagos, Nigeria, and Nairobi, Kenya. Factors such as urban infrastructure (e.g., health, schools, and employment opportunities) [6], population density, and other significant geographic elements are explored to understand their impact on urban deprivation. Following this analysis, the project will integrate these geospatial insights with sophisticated machine learning algorithms. Machine learning methodology is crucial in this project, as it facilitates the processing of extensive datasets and enables the identification of intricate patterns in urban environments. These approaches significantly enhance the precision and depth of deprivation classification in cities like Lagos and Nairobi, leveraging the strengths of each method to achieve a more comprehensive analysis. However, challenges such as having imbalanced data with fewer data points for deprived regions [7] can affect our model's performance. This imbalance poses a significant obstacle to model performance, highlighting the need for solutions to address data disparities and ensure a more equitable representation of urban areas. In user-driven earth observation-based slum mapping [8] Owusu mentions the need for two main elements which are a key to successful slum-based mapping involving a.) contextualizing slums or understanding of local context and user requirements. T. Stark [9] talks about the challenges with imbalanced dataset while using deep learning models and handles them using augmentation techniques like rotation and affine transformations and implements transfer learning for slum mapping. Another challenge is that slum labels change over time in a particular area subject to social, political and economic changes so Fisher [10] introduces a concept called uncertainty and quantifies it to help us in understanding and detecting slum area changes over a given period.

Solution and Methodology

In the realm of data science and geographic information system (GIS), the integration of covariate, contextual, and pixel features analysis and spatial data extraction offers a powerful approach to understanding urban dynamics, environmental factors, population density, and socio-economic conditions. This section details a comprehensive project aimed at harnessing this approach within Lagos, Nigeria, focusing on extracting and analyzing data from three main raster files to study covariate, contextual, and pixel features across Lagos's geographic areas.

Essential Python Libraries for Data Pipeline build and Geospatial Analysis

Python libraries have become essential tools for building efficient data pipelines and performing geospatial analysis. Simplifying the transition from raw data to actionable insights, these libraries offer specialized functionalities that cater to various aspects of data handling and analysis. This brief overview introduces pivotal Python packages such as Geopandas, GeoWombat, Rasterio, Pandas, and PyCaret, which streamline data processing, enhance machine learning workflows, and facilitate the management of geospatial data. Together, they equip data analysts with the means to effectively tackle data science challenges and construct robust data pipelines.

Geopandas: It is an open-source Python project designed to bridge the gap between the capabilities of pandas—a staple for data manipulation and analysis—and geospatial data operations. By extending pandas to support spatial data types, Geopandas allows for the easy execution of sophisticated spatial operations (e.g., spatial joins, overlays) directly within the familiar DataFrame structure. This integration effectively democratizes geospatial analysis, making it accessible to data scientists and researchers familiar with pandas but less so with specialized GIS software. Geopandas relies on other foundational geospatial libraries like Fiona for file access and Shapely for geometric operations, thus standing on the shoulders of the broader open-source geospatial ecosystem to provide a high-level interface for spatial data science.

GeoWombat: It is tailored for processing geospatial and raster data at scale, facilitating the efficient handling of satellite imagery datasets like Landsat and Sentinel, as well as generic RGB data formats. By extending the capabilities of xarray—a library designed for multi-dimensional arrays—GeoWombat enables the streamlined processing of large datasets, incorporating spatial attributes and metadata essential for geospatial analysis. This makes it particularly valuable for environmental monitoring, agricultural analysis, and urban planning, where satellite data offers critical insights over time and space.

Rasterio: It simplifies the reading, writing, and processing of raster data in Python, catering to geospatial professionals' need to work efficiently with satellite imagery, digital elevation models, and other forms of rasterized data. By providing a more Pythonic interface to GDAL (Geospatial Data Abstraction Library), Rasterio makes raster data manipulations more accessible and intuitive. Its capabilities are crucial for tasks that involve raster data transformations, such as reprojecting images, analyzing raster data statistics, and visualizing geographical information.

Pandas: It is a foundational Python library for data manipulation and analysis, known for its powerful data structures like DataFrames and Series. These structures facilitate handling and

analyzing structured data, supporting operations ranging from data cleansing and transformation to complex aggregations and pivot tables. Pandas is indispensable in the data science workflow, serving as the backbone for data exploration and analysis across diverse data science and financial analysis applications.

PyCaret : It stands out as an open-source, low-code machine learning library that streamlines the transition from data to insights. It offers an end-to-end ML workflow, automating tasks from data preprocessing and feature engineering to model selection and tuning. PyCaret is designed to expedite the machine learning project lifecycle, making it an excellent tool for both novices and experts looking to enhance productivity. Its low-code approach democratizes access to advanced ML techniques, enabling users to focus more on problem-solving and less on the intricacies of algorithm implementation.

Data Pipeline Build and Processing

Covariate Features: Our mission is to utilize a dataset derived from the multi-band raster file "lag_covariate_compilation_53bands.tif," which has a 100-meter resolution, to evaluate the performance of predictive models when dealing with extensive and varied datasets. This analysis is designed to reveal unique patterns and trends across regions, providing critical insights to support interventions and policy decisions aimed at reducing urban deprivation. A significant part of our analysis involves assessing the importance of the collected data, specifically how each of the 53 covariate features contributes to understanding and addressing urban challenges.

The "Lagos_Slum_reference.gpkg" GeoPackage file meticulously delineates Lagos with a high-resolution 10-meter grid, from which we exclusively extract slum labels. This extraction process leverages the precise geometric points within each grid cell to ensure an accurate association of slum labels with the corresponding raster values from the 53 bands in the "lag_covariate_compilation_53bands.tif" file. These geometric points serve as critical reference points, facilitating the integration of slum labels with the raster data to construct a comprehensive portrayal of the urban environment.

This intricate process yields a CSV file "lagos_centroid.csv", where each row represents a grid cell in Lagos, identified by its centroid and associated with a slum label to underscore deprived areas. Each column within this CSV file corresponds to one of the 53 raster bands, together with the slum label, creating a dataset that meticulously maps the complex terrain of Lagos.

Contextual Features: Contextual features offer a deep dive into the spatial attributes that define the geographical landscapes under examination. By analyzing the spatial arrangement, orientation, and structural properties captured in images, we can discern invaluable information on a variety of attributes, such as building counts, density, climate risk factors, housing quality, extreme temperatures, population density, and even the differentiation between urban and rural areas, along with nocturnal lighting. These insights enable us to unravel the complex interplay between the natural environment and human-made structures, providing a comprehensive view of any given area.

Our methodology initially involves resampling the original 10-meter resolution of the 144 contextual TIFF files to a more comprehensive 100-meter resolution. This resampling process is crucial for aligning the data's spatial resolution with the geometric points in the 'A100mGrid_Lagos' GPKG file, enhancing the accuracy of our geographical analysis. Following this, our goal is to retrieve geometric points and harness a broad spectrum of contextual feature information, encapsulated within 144 unique contextual values. This process kicks off with the execution of the "extract_context_by_point.py" script. Utilizing the 'A100mGrid_Lagos' GPKG file and 144 individual resampled contextual feature TIFF files as inputs, this script outputs a directory named "contextual_features_extraction". This directory contains 144 CSV files, each meticulously documenting contextual feature values aligned with their corresponding geometric points—a crucial step towards our comprehensive geographical analysis.

The next phase of our approach is the integration of these 144 discrete CSV files into a single, cohesive dataset. By running the "merging_contextual_feature.py" script, we achieve seamless consolidation of this vast array of contextual data into one unified CSV file, aptly named 'merged_contextual_features.csv'. This aggregation process not only simplifies the data structure but also sets the stage for an extensive examination of the spatial characteristics of Lagos.

Delving into the specifics, our project explores eleven sophisticated contextual features, employing techniques such as Fourier Transforms, Gabor Filters, Histogram of Oriented Gradients (HOG), and others to capture the essence of Lagos's spatial complexity. This exploration results in the creation of 144 TIFF images, each painting a unique picture of the city's contextual landscape—from detecting image patterns with Fourier Transforms and identifying edges with Gabor Filters to describing shapes with HOG.

Finally, by pinpointing centroids within predefined areas of Lagos, we extract crucial information from these images, culminating in the collection of 144 CSV files. Each file represents distinct contextual features, paving the way for the last stage of our project. We merge the data from these files, focusing on geometric data points and feature values, and integrate them based on their geometric identifiers. This process yields a unified dataset, ready for deep analysis and the generation of actionable insights, thus encapsulating our rigorous exploration of Lagos through the lens of contextual features.

Sentine-2 Satellite Pixel Features: Sentinel-2 satellite imagery, provided by the European Space Agency (ESA) under the Copernicus program, is a pivotal resource in geospatial analysis due to its accessibility, high resolution, and versatility. The imagery is freely available, making high-quality satellite data accessible to a wide range of users. With a resolution of 10-meter, Sentinel-2 captures detailed imagery suitable for various applications, from environmental monitoring to urban planning. One of the significant advantages of Sentinel-2 is its capability to create mosaic images for areas with cloud coverage, thereby enhancing data usability. The satellite's extensive coverage is ideal for scalable applications across diverse geographic extents. Moreover, the integration with Google Earth Engine allows for the creation of cloud-free mosaics over large areas anywhere on the globe, facilitating easy access to and analysis of large datasets. Sentinel-2 imagery includes data across four spectral bands—red, blue, green, and near-infrared—each providing

unique insights into the Earth's surface. This combination of features establishes Sentinel-2 as an invaluable tool in remote sensing, offering comprehensive data for many analytical applications.

Initially, the 10-meter resolution RBGN TIFF file has been resampled to a more encompassing 100-meter resolution, ensuring a uniform scale for further analysis. The following process is to harness the four-band RBGN raster data from Sentinel-2 to conduct a thorough examination of specific zones within Lagos. Our focus centers on the extraction and detailed analysis of pixel data across the satellite's spectral bands to derive insights into the physical characteristics of the area.

The process involves running the "extract_by_points_lag_bgrn.py" script, which extracts geometric points and their corresponding RBGN values from the resampled TIFF image. The inputs for this operation are the 'A100mGrid_Lagos' GPKG file and the resampled TIFF image itself. The outcome of this script is the "lagos_bgrn.csv" file, which organizes the extracted geometric points alongside the four RBGN values into a structured dataset. This CSV file effectively represents the spectral data for each location within Lagos, setting the stage for an in-depth analysis of the physical characteristics revealed by the Sentinel-2 imagery. Through this meticulous extraction process, we aim to delve into the rich spectral data provided by Sentinel-2, enhancing our understanding of the physical landscape of Lagos.

Data Combination Process: The primary goal of the final step is to set out with the ambitious goal of merging diverse datasets—specifically, slum labels, 144 contextual features, 53 covariate band values, and four-band Sentinel-2-pixel data (RBGN)—into a singular, cohesive dataset. By linking this data through a shared 'geometry' column corresponding to specific geographic locations, we aim to create a robust foundation for detailed spatial analysis. This integrated dataset is designed to enable a multifaceted examination of Lagos, facilitating a deeper understanding of its urban fabric and environmental conditions.

The integration process begins with the preparation of individual CSV files, which includes verifying the integrity of the data and ensuring that the formats are compatible for merging. Central to this process is the use of the 'geometry' column, which serves as a pivotal linkage point that ensures the spatial congruence of the combined datasets. By meticulously aligning data based on geometric points, we maintain the spatial accuracy essential for subsequent analysis. This step involves executing the "combined_data.py" script, which takes as its inputs the lagos_centroid.csv, merged_contextual_features.csv, and lagos_bgrn.csv.

The result of this intricate merging process is the creation of the final_output_lagos.csv file, a comprehensive dataset that integrates slum labels, contextual features, covariate band values, and RBGN values, all precisely aligned by their geographic points. This consolidated file stands as a critical asset for conducting advanced modeling and analysis. It provides a panoramic view of Lagos's complex urban and environmental dynamics, laying the groundwork for insightful research and effective policy formulation aimed at addressing the city's multifarious challenges.

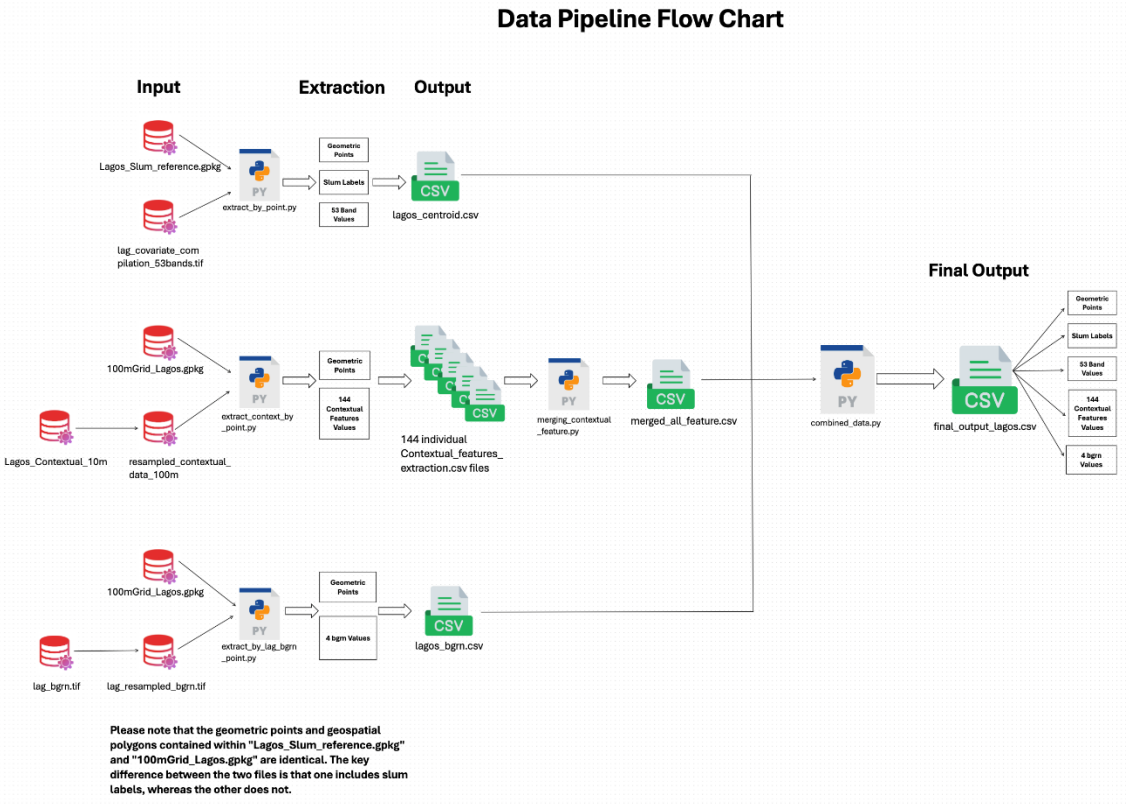

Figure 1: A Workflow Chart about Data Pipeline Processing

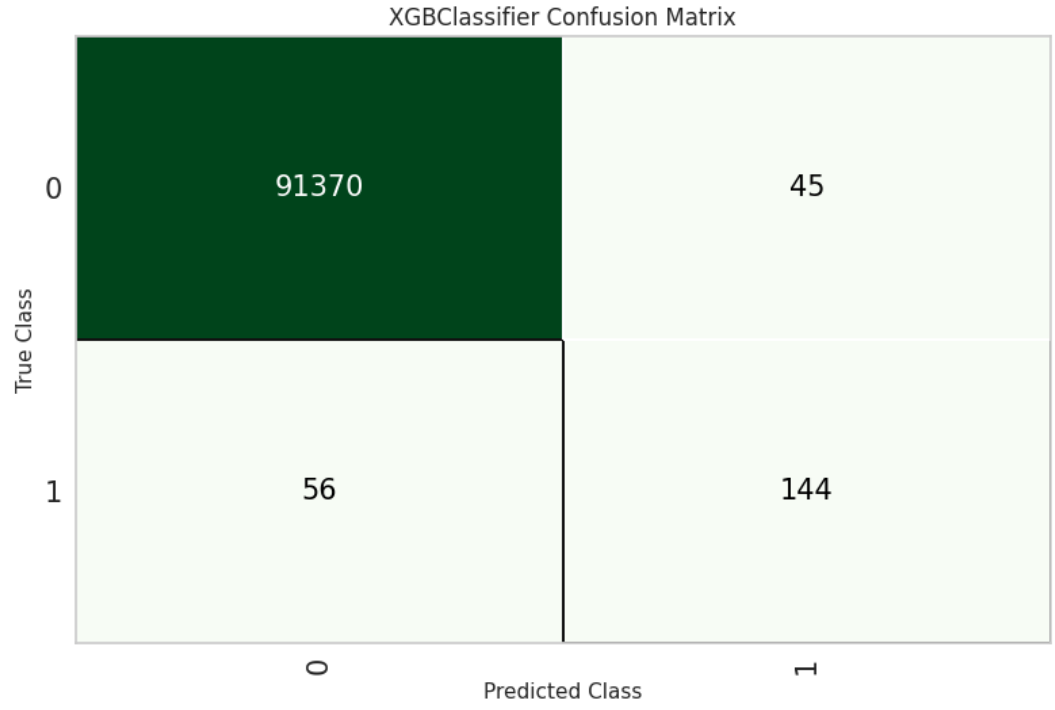

Figure 2: XGBoost classifier confusion matrix (0 represents non-deprived and 1 is for deprived)

Results and Discussion

The dataset used for training for Lagos contained 305, 381 rows with 204 columns that had 4 features for pixel values obtained from Sentinel-2 images, 53 covariate features and 144 contextual features, longitude and latitude. There slum labels contained values 0, 1, 2 and 3 (see Fig 3). The label values 1 and 2 were combined as a single slum value 1 that indicates deprived area. While 0 as indicated in Fig4. was considered as a non-deprived area. Label values 3 were ignored as the person involved in labelling the data was unsure of. For training purposes, a python package called PyCaret was used. A default train test split of 70% for training and 30% for testing was used.

Value	Classification
0	This is no slum. This is the baseline we created above when we created a new column called Slum.
1	This is a slum (95% coverage). Verified as far possible on Google Street View and by Adenike. This label made sure that there would be no interference from roads or anything green for example if it was used for training data.
2	This is 50% slum. We included this as we had previously discounted any areas, we thought would affect the model. It was decided that slums do contain areas of water, green, 'built' roads, and formal-looking buildings (rather than just irregular layouts). We knew the locations of the slums but weren't able to say that this would be the best data if used for training.
3	This was used by Alex in areas she was unsure of. We had over 300,000 features so 3 was a way of putting a 'post-it note' on an area for Alex to ask for verification.

Fig 3: Description of target variable Slum

Data tables

Model	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC
Logistic Regression	0.9977	0.9337	0.085	0.3953	0.1399	0.1393	0.1825
Random Forest Classifier	0.9987	0.9993	0.44	0.9263	0.5966	0.596	0.6379
K Neighbors Classifier	0.9978	0.5034	0	0	0	0	0
XGBoost	0.9989	0.9995	0.595	0.8322	0.6939	0.6933	0.7031
Multi-Layer Perceptron	0.9978	0.5671	0	0	0	0	0
Decision Tree	0.9981	0.7596	0.52	0.5683	0.5431	0.5421	0.5427
Ridge Classifier	0.9978	0.5	0	0	0	0	0
Quadratic Discriminant Analysis	0.9733	0.9849	0.995	0.0753	0.14	0.1365	0.27
Ada Boost Classifier	0.9986	0.9962	0.53	0.7413	0.6181	0.6174	0.6261
Gradient Boosting Classifier	0.9981	0.8009	0.265	0.6625	0.3786	0.3778	0.4183
Extra Trees Classifier	0.9986	0.9993	0.405	0.9205	0.5625	0.5619	0.6101
Light Gradient Boosting Machine	0.9938	0.8028	0.435	0.1608	0.2348	0.2324	0.262
Dummy Classifier	0.9978	0.5	0	0	0	0	0
Support Vector Machine	0.9965	0.5043	0.01	0.0161	0.0123	0.0107	0.011
Naive Bayes	0.0687	0.9138	0.995	0.0023	0.0046	0.0003	0.0116

Table 1: Classical modelling results on test set using imbalanced data for Lagos.

Model	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC
Logistic Regression	0.918	0.9566	0.885	0.0231	0.045	0.0409	0.1353
Random Forest Classifier	0.9982	0.9988	0.605	0.5789	0.5917	0.5908	0.5909
K Neighbors Classifier	0.9392	0.5583	0.17	0.0062	0.0121	0.0079	0.0219
XGBoost	0.999	0.9994	0.73	0.7807	0.7545	0.754	0.7544
Multi-Layer Perceptron	0.9774	0.9742	0.855	0.0773	0.1417	0.1383	0.2531
Decision Tree	0.9971	0.769	0.54	0.383	0.4481	0.4467	0.4534
Ridge Classifier	0.9661	0.973	0.98	0.0594	0.112	0.1083	0.237
Quadratic Discriminant Analysis	0.9832	0.9871	0.98	0.1134	0.2033	0.2002	0.3305
Ada Boost Classifier	0.9936	0.9956	0.81	0.2278	0.3557	0.3534	0.4276
Gradient Boosting Classifier	0.9938	0.9985	0.94	0.2517	0.397	0.395	0.4847
Extra Trees Classifier	0.9987	0.9992	0.75	0.6977	0.7229	0.7223	0.7227
Light Gradient Boosting Machine	0.9987	0.9992	0.745	0.6835	0.7129	0.7123	0.7129
Dummy Classifier	0.9978	0.5	0	0	0	0	0
Support Vector Machine	0.8674	0.8263	0.785	0.0128	0.0252	0.021	0.0895
Naive Bayes	0.0713	0.6147	0.995	0.0023	0.0047	0.0003	0.0118

Table 2: Classical modelling results on test set using balanced data for Lagos.

Table 2 shows the confusion matrix for XGBoost model that was trained on the data obtained from Lagos after oversampling the training set using SMOTE.

Graphs

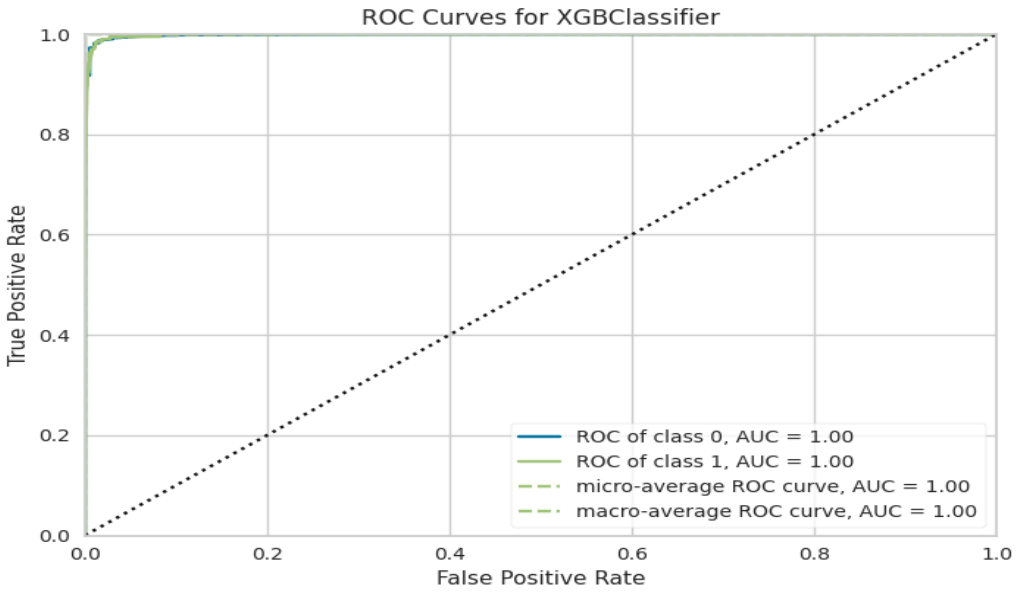

Figure 4: XGBoost classifier ROC curve

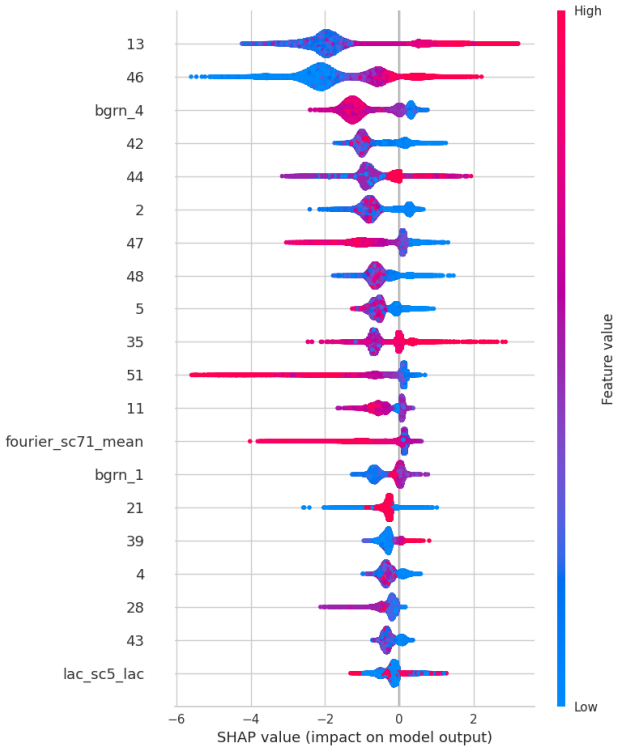

Figure 5: SHAP values for XGBoost classifier showing impact of features on model output

In Figure 5, this is what the covariate features stand for:

Covariate Feature	**Description**
13	ph_dist_cultivated_2015
46	uu_bld_den_2020
42	sh_dist_mnr_pofw_2019
44	sh_ethno_den_2020
2	fs_dist_school_2020
47	uu_impr_housing_2015
48	fs_dist_hf_2019
5	in_dist_waterway_2016
35	ses_m_lit_2018
51	ph_dist_riv_network_2007
11	ph_dist_art_surface_2015
21	ph_grd_water_2000
39	ses_preg_2017
4	in_dist_rd_intersect_2016
28	ph_slope_2000
43	sh_dist_pofw_2019

Table 3: Covariate Feature description based on Figure 5

As shown in Figure 5, the image presents a visualization of test results from an XGBoost model. With an F1 score of 0.75, the model demonstrates a reasonably good level of predictive accuracy. The F1 score, an aggregate measure of precision and recall, indicates a reasonable balance between the two metrics. Within the scope of satellite imagery analysis, the model reliably differentiates various land features or conditions, such as areas of vegetation, bodies of water, urban developments, or other relevant categories. Nonetheless, it is important to recognize that despite the robustness of an F1 score of 0.75, opportunities for enhancement remains and deep learning techniques can increase the score further, as the score falls short of the ideal 1, which signifies a perfect equilibrium of precision and recall.

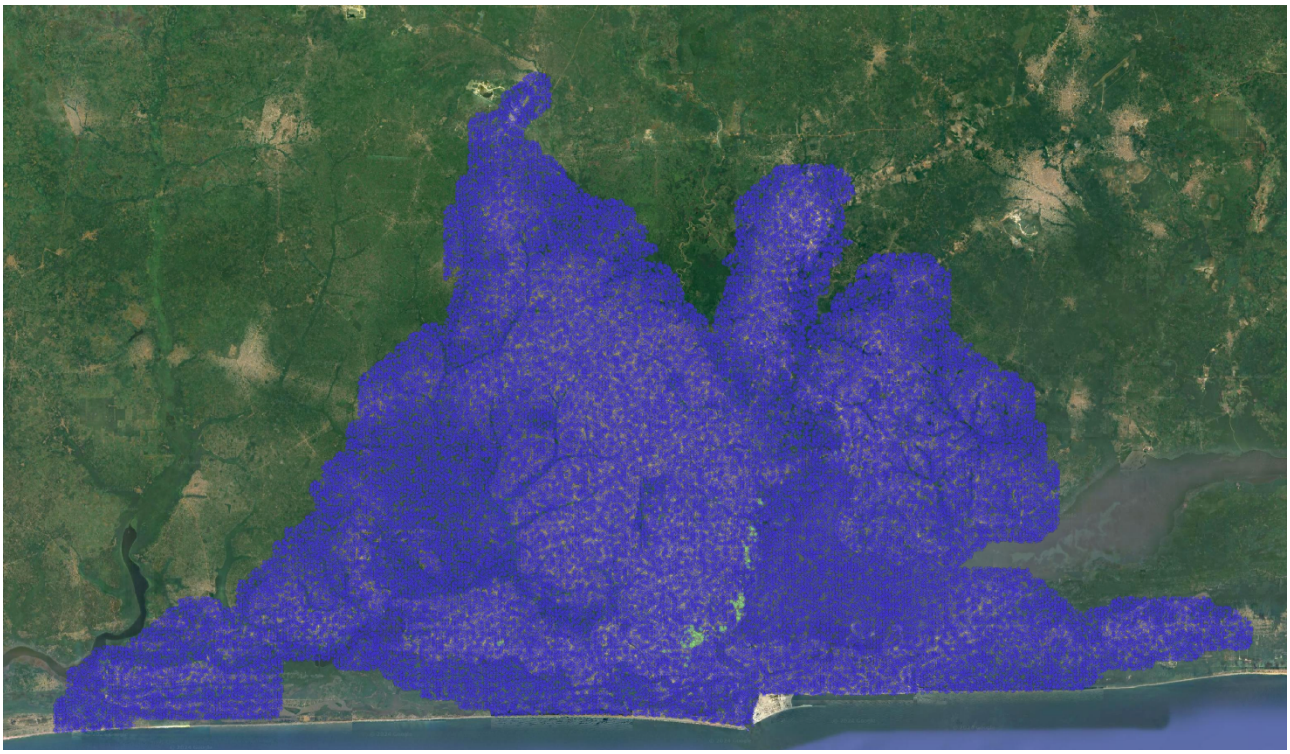

Figure 6: A test set result for Lagos after mapping the points on QGIS software

In Figure 6, the points in blue represent non-deprived areas and the green points represent deprived areas. XGBoost classifier was used which obtained an F1 score of 0.75 to generate predictions on the test set for Lagos. These predictions were then exported to a CSV file and were imported in QGIS software. The coordinates (longitudes and latitudes) were then used to map the points on QGIS software by adding a layer and using the EPSG:4326 (WGS 84) coordinate reference system. Google Earth satellite image was added as a layer and the points were classified based on the prediction labels obtained from the modelling results into deprived and non-deprived which were mapped on QGIS software.

Discussion

The primary challenges in this capstone project stemmed from complexities within the data pipeline. Due to the diverse formats of geospatial data—GeoPackage (.gpkg) for reference and GeoTIFF (.tif) for features—the task of creating a robust data pipeline proved particularly demanding. Initially, while the GeoPandas library was useful in extracting slum labels from the reference file and features from the TIFF files, it fell short in effectively combining these features based on coordinates. The adoption of the GeoWombat library addressed this issue, facilitating compatibility with different formats of geospatial data. This enhancement in our data pipeline allowed not only the extraction of features and coordinates but also the integration of features extracted from both GPKG and TIFF files along with their corresponding coordinates. It is crucial,

however, to note that this data pipeline's effectiveness is contingent upon the consistency and stability of the data formats used.

Furthermore, resampling the 100-meter resolution GPKG files to a finer 10-meter resolution introduces additional challenges. This resampling is pivotal as it lays the groundwork for developing a secondary data pipeline tailored for a 10-meter resolution dataset. This new pipeline is vital for conducting more detailed analyses and applying sophisticated machine learning and deep learning techniques. A dataset with higher resolution will capture more intricate spatial features, potentially boosting the accuracy and broadening the applicability of our models across various urban environments. Despite its challenges, this step is critical for enhancing the scope and efficacy of our predictive models.

In terms of performance, classical modeling yielded decent results with an F1 score of 0.75. These results could be further improved by incorporating deep learning models, such as CNNs. Moreover, experimenting with resampling at 10-meter resolution for classical models might also enhance outcomes. Ultimately, the optimal model should be capable of predicting slums in diverse cities, thereby allowing models trained in one city to be effectively applied in predicting slum labels in other urban areas. This adaptability is essential for scaling the impact of our models globally.

Conclusion

This capstone project represents a significant step forward in the quest to better understand and address urban poverty, particularly through the lens of slum mapping in low- and middle-income countries. By developing a comprehensive geospatial analysis framework, this project has laid the groundwork for informed interventions that can significantly enhance the quality of life for slum dwellers. Despite facing challenges such as diverse data formats and the need for high-resolution data, the adoption of tools like GeoWombat and the strategy of resampling have enabled the creation of a more robust and efficient data pipeline.

The use of machine learning model has demonstrated promising results, with a classical modeling approach yielding an F1 score of 0.75. There remains potential for further improvements, particularly through the application of advanced deep learning techniques and refining data resolution. The scalability of the models developed allows for their application beyond the initial focus cities of Lagos and Nairobi, offering a pathway to generalize this approach to other urban settings globally. This capability is crucial for contributing to the global efforts towards achieving the Sustainable Development Goals, particularly in making cities inclusive, safe, resilient, and sustainable.

In conclusion, while challenges remain, the progress made in this capstone project provides a hopeful outlook for the future of urban development and poverty alleviation. It underscores the importance of detailed and accurate mapping of deprived areas, not only as a technical achievement but as a fundamental aspect of social justice and equitable urban planning.

Bibliography

1. Lee, J.-G., & Kang, M. (2015). Geospatial Big Data: Challenges and Opportunities. *Big Data Research*, *2*(2), 74–81. https://doi.org/10.1016/j.bdr.2015.01.003
2. Robinson, A. C., Demšar, U., Moore, A. B., Buckley, A., Jiang, B., Field, K., Kraak, M.-J., Camboim, S. P., & Sluter, C. R. (2017). Geospatial big data and cartography: research challenges and opportunities for making maps that matter. *International Journal of Cartography*, *3*(sup1), 32–60. https://doi.org/10.1080/23729333.2016.1278151
3. Kovacs-Györi, A., Ristea, A., Havas, C., Mehaffy, M., Hochmair, H. H., Resch, B., Juhasz, L., Lehner, A., Ramasubramanian, L., & Blaschke, T. (2020). Opportunities and Challenges of Geospatial Analysis for Promoting Urban Livability in the Era of Big Data and Machine Learning. *ISPRS International Journal of Geo-Information*, *9*(752).
4. Koldasbayeva, D., Tregubova, P., Gasanov, M., Zaytsev, A., Petrovskaia, A., & Burnaev, E. (2023). Challenges in data-based geospatial modeling for environmental research and practice. *arXiv.Org*. https://doi.org/10.48550/arxiv.2311.11057
5. Runfola, D., Stefanidis, A., Lv, Z., O'Brien, J., & Baier, H. (2024). A multi-glimpse deep learning architecture to estimate socioeconomic census metrics in the context of extreme scope variance. *International Journal of Geographical Information Science*, 1–25. https://doi.org/10.1080/13658816.2024.2305636
6. Engstrom, R., Owusu, M., Nair, A., Jafari, A., Thomson, D., & Kuffer, M. (2023). Evaluating the ability to use contextual features to map deprived areas 'slums' in multiple cities. *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*. https://doi.org/10.1109/igarss52108.2023.10282954
7. Owusu, M., Engstrom, R., Thomson, D., Kuffer, M., & Mann, M. L. (2023, November 8). *Mapping deprived urban areas using open geospatial data and machine learning in Africa*. MDPI. https://www.mdpi.com/2413-8851/7/4/116/htm
8. Owusu, M., Kuffer, M., Belgiu, M., Grippa, T., Lennert, M., Georganos, S., & Vanhuysse, S. (2021). Towards user-driven earth observation-based slum mapping. *Computers, Environment and Urban Systems*, *89*, 101681. https://doi.org/10.1016/j.compenvurbsys.2021.101681
9. T. Stark, M. Wurm, H. Taubenböck and X. X. Zhu, "Slum Mapping in Imbalanced Remote Sensing Datasets Using Transfer Learned Deep Features," *2019 Joint Urban Remote Sensing Event (JURSE)*, Vannes, France, 2019, pp. 1-4, doi: 10.1109/JURSE.2019.8808965.
10. Fisher, T., Gibson, H. W., Liu, Y., Abdar, M., Posa, M., Salimi-Khorshidi, G., Hassaïne, A., Cai, Y., Rahimi, K., & Mamouei, M. (2022). Uncertainty-Aware interpretable deep learning for slum mapping and monitoring. *Remote Sensing*, *14*(13), 3072. https://doi.org/10.3390/rs14133072