Machine Learning in Satellite Imagery and Other Geotagged Data Sources for Health Monitoring in Low- and Middle-Income Countries

Summer Research Internship, 2023
Department of Primary Care and Population Health
Stanford University, School of Medicine

Vedant Zope
Indian Institute of Technology Kharagpur
vedantzope@kgpian.iitkgp.ac.in

Soham Tripathy
Indian Institute of Technology Kharagpur
soham@kgpian.iitkgp.ac.in

Kushagra Parmeshwar
Indian Institute of Technology Kharagpur
kush@kgpian.iitkgp.ac.in

Abstract

This research project pioneers the integration of machine learning and satellite imagery to monitor health indicators in low- and middle-income countries (LMICs). Utilizing a comprehensive dataset from MOSAIKS and the Google Earth Engine API, we have developed a model that predicts critical health indicators at over 120k geographical locations. Our innovative approach addresses the data gaps prevalent in traditional health surveys, employing machine learning regression models to expedite computations on standard hardware. The project has achieved an MCRMSE score of 11.20, demonstrating the potential of this method for health monitoring in LMICs. This work contributes significantly to public health, showcasing the potential of satellite imagery and machine learning in delivering valuable insights into health conditions within resource-constrained settings.

1 Introduction

The fusion of satellite imagery and machine learning techniques has emerged as a potent instrument for monitoring and interpreting various global phenomena. Its application in the realm of public health, particularly, holds the potential to transform the surveillance and evaluation of health indicators in low- and middle-income countries (LMICs). This project is at the forefront of this transformation, harnessing the capabilities of satellite imagery and machine learning to monitor and predict key health indicators in LMICs. The ultimate objective is to equip policymakers and organizations with the necessary insights to effectively make informed decisions and allocate resources.

Monitoring health indicators in LMICs presents a myriad of challenges. The limited access to reliable, up-to-date health data, compounded by resource constraints, hinders the acquisition of comprehensive insights into the health status of these populations. Traditional data collection and analysis methods, which rely heavily on surveys and manual data entry, are not only time-consuming but also prone to data gaps and inaccuracies. Therefore, the need for innovative, efficient, and accurate approaches to health indicator information in LMICs is more critical than ever.

In response to this need, our project leverages a rich dataset derived from Demographic and Health Surveys (DHS) conducted in 59 countries over a decade. This dataset is enriched by incorporating satellite imagery data from MOSAIKS and the Google Earth Engine API. Integrating these diverse data sources allows us to derive a comprehensive set of features and delve into their relationship with health indicators.

Our approach primarily relies on machine learning regression models to forecast critical health indicators, including $Mean\_BMI$, $Median\_BMI$, $Unmet\_Need\_Rate$, $Under5\_Mortality\_Rate$, $Skilled\_Birth\_Attendant\_Rate$, and $Stunted\_Rate$. In contrast to deep neural networks, which may demand substantial computational resources, we concentrate on ensemble regressors. These models expedite computations and facilitate predictions on standard computers, striking an optimal balance between prediction accuracy and computational efficiency. This equilibrium ensures the accessibility and feasibility of our models for deployment in LMICs.

Through meticulous preprocessing and model training, we have achieved promising results, as evidenced by an MCRMSE (Mean Column-Wise Root Mean Squared Error) score of 11.20083. These findings underscore the potential of satellite imagery and machine learning in providing valuable insights into health conditions and informing decision-making processes in LMICs.

This report offers a detailed analysis of our methodology, dataset, experimental setup, and results. We also discuss the limitations, challenges, and opportunities for future research in this field. By illuminating the potential of satellite imagery and machine learning in the context of public health, this project contributes to the burgeoning body of knowledge in data-driven approaches for health monitoring and resource allocation.

In the subsequent sections, we provide an overview of related work, describe our methodology, present the dataset, discuss the experimental setup, and analyze the results. The report concludes with a summary of our findings, implications for practice, and future research directions.

2 Datasets

In this section, we will explore the various datasets that form the backbone of our project. These datasets, each with its unique characteristics and information, have been instrumental in shaping our approach and methodology.

We will delve into the details of each dataset, discussing their origins, the type of data they contain, and how they have been utilized in our project

2.1 Google Earth Engine Dataset

The dataset used in this project is a comprehensive compilation of data from Demographic and Health Surveys (DHS) conducted in 58 countries, combined with various satellite imagery sources. The dataset, both training and test sets, was meticulously curated by the Geldsetzer lab at Stanford Medicine.

The primary component of the dataset is the `gee_features.csv` file. This file contains extracted features from Google Earth Engine (GEE) and keys to match it with other types of data, such as country names (DHSCC), cluster numbers (DHSCLUST), and year of survey (DHSYEAR). It's important to note that certain column names in this file are not predictive features but rather identifiers and metadata. The predictive features are extracted from different sources of satellite images and other public data.

The dataset encompasses 11,959 numerical features for 120,984 DHS community reports, representing a total of 58 countries. These features include numerical meteorological and geographical data for each DHS community from Google Earth.

The `training_label.csv` file serves as the label dataset for the training set. It can be linked to the features in `gee_features.csv` using the DHSID. The project aims to predict six key health indicators: Mean BMI, Median BMI, Unmet Need Rate, Under5 Mortality Rate, Skilled Birth Attendant Rate, and Stunted Rate.

The distribution of the DHS survey points across different countries is visualized in the heatmap shown in Figure 1. This heatmap provides a global view of the number of data points available per country. The color intensity in each country corresponds to the number of data points available, with darker colors indicating a higher number of data points. This visualization helps to highlight the geographical distribution of the data and reveals the countries with the most comprehensive data coverage. It's important to note that the distribution of data points is not uniform across all countries, which reflects the varying availability and accessibility of health data in different regions. This uneven distribution poses a challenge for the prediction models, as countries with fewer data points may not be as accurately represented in the model's predictions.

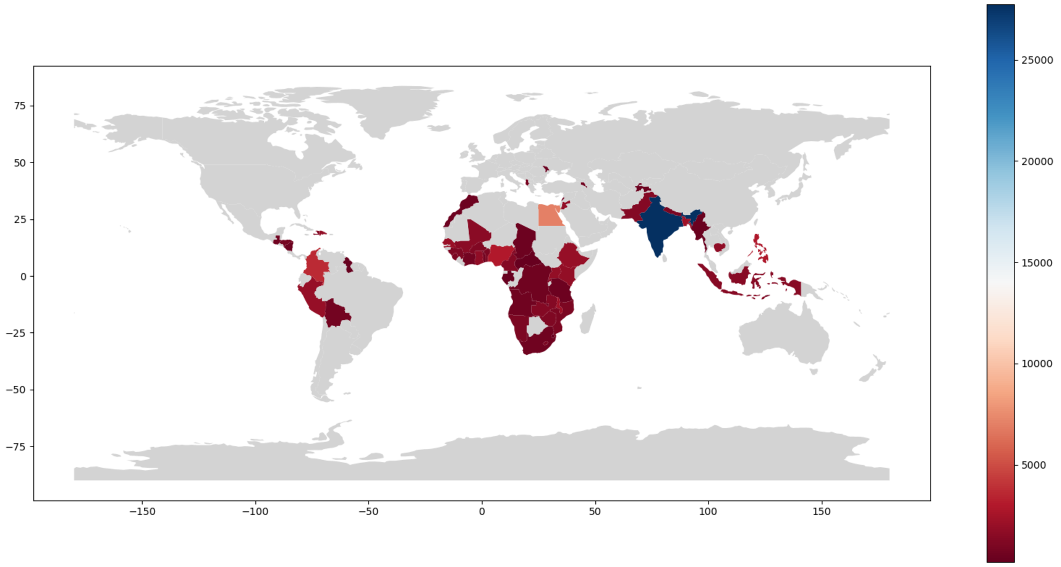


Figure 1: DHS Survey Points per Country

2.2 MOSAIKS Satellite Image Data

In an effort to enhance our dataset and potentially improve the predictive power of our models, we have explored the use of external datasets. One such dataset that we have considered is the MOSAIKS satellite image data. This dataset provides a wealth of geospatial information about global land areas, which could be instrumental in understanding and predicting health indicators.

The MOSAIKS data is structured as a global grid with a resolution of 0.01 x 0.01 degrees, derived from Planet imagery. Due to the vast size of the complete dataset, which spans multiple terabytes, we strategically request custom subsamples of the imagery that align with our project's needs. We supply a list of locations of interest via the "File Query" tool, and the MOSAIKS API allocates each input latitude and longitude coordinate to the nearest point on the global grid.

For each point on the grid, the MOSAIKS data provides 4000 precomputed features. These features encompass a wide range of environmental and geographical factors. The effectiveness and impact of incorporating this data on our model's performance will be discussed further in the report.

2.3 Handling Large Datasets

The initial challenge in our project was managing multiple large datasets, which exceeded memory limits. To address this, we implemented two strategies. First, we converted the dataset from CSV to Parquet format. This conversion significantly improved memory efficiency and enabled faster data access and processing. Second, we downcasted the variables to the lowest possible datatype without losing precision. This further reduced the memory footprint of the dataset and facilitated efficient data handling.

3 Data Preprocessing

Data preprocessing is a crucial step in any machine learning project. It involves cleaning and transforming raw data into a format that can be easily understood and utilized by machine learning algorithms. In this section, we will discuss the various preprocessing techniques we have employed to prepare our dataset for model training.

3.1 Handling Missing Values

A significant challenge in our dataset was the presence of missing feature and label data. To address this, we first dropped the data which had missing label data, as these instances lacked the necessary ground truth for training and evaluation.

Next, we examined the DHS data rows with the highest proportion of missing data. We decided to drop all rows where data had a higher proportion of missing information than a certain threshold(50%), as these instances could potentially introduce noise and inaccuracies into our models.

Finally, we addressed the remaining missing feature values through the application of the K-nearest neighbors (KNN) imputation technique. KNN imputation estimates missing values based on the values of neighboring data points. We set the number of neighbors parameter to 30, a value determined based on the dataset's characteristics. Importantly, this imputation was performed country-wise, ensuring that the imputed values were contextually relevant and accurate.

3.2 Understanding Labels

The labels in our dataset represent key health indicators derived from the Demographic and Health Surveys (DHS). These indicators provide valuable insights into the health status of populations in low- and middle-income countries. The labels are as follows:

- Mean_BMI: This represents the average Body Mass Index in a given community. It is a measure of body fat based on height and weight that applies to adult men and women.
- Median_BMI: This is the median Body Mass Index in a given community, providing a measure that is not skewed by outliers.
- Unmet_Need_Rate: This represents the percentage of women who want to stop or delay childbearing but are not using any method of contraception.
- Under5_Mortality_Rate: This is the probability per 1,000 that a newborn baby will die before reaching age five, if subject to current age-specific mortality rates.
- Skilled_Birth_Attendant_Rate: This is the percentage of births attended by skilled health personnel.
- Stunted_Rate: This is the percentage of children under 5 years of age who suffer from stunting (low height-for-age).

We have generated plots for each label, showing the label value versus the number of instances. These plots provide a visual representation of the distribution of each health indicator in our dataset. The distribution of these health indicators is shown in Figure 2.

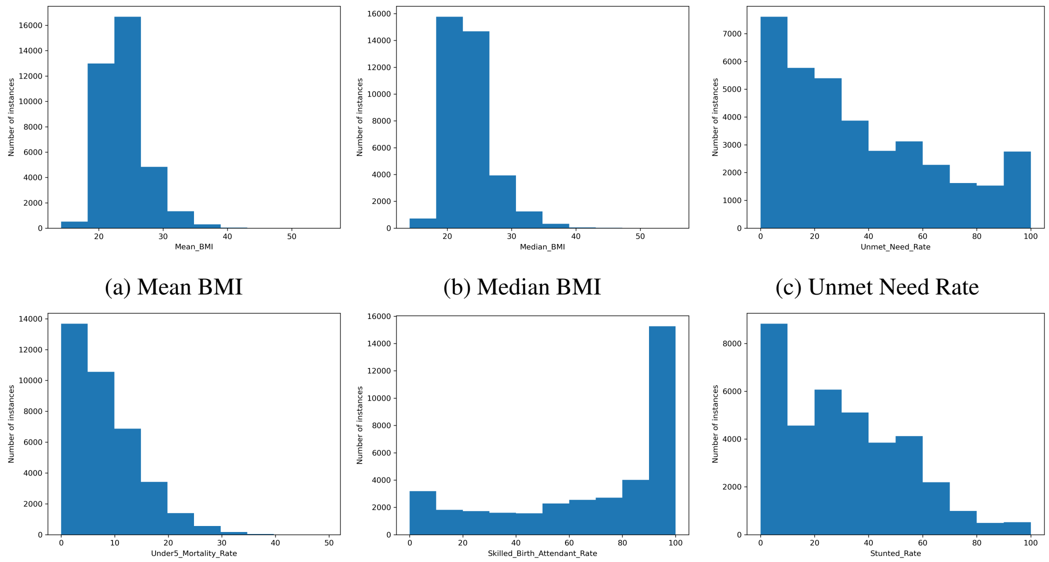
3.3 Train-Test Split

In contrast to the conventional approach of separating train and test cohorts by country, we adopted a unique strategy for creating a train/dev/test split at the community level. For any given country, 80% of its data was allocated to the train set, while the remaining 20% was evenly split between the dev set and the test set. This ensured that all countries were represented in the train, dev, and test cohorts, providing a more comprehensive and representative sample for our machine learning models.

3.4 Country-wise Segregation

To facilitate regional analysis, we assigned each country in the dataset to a specific region based on predefined country sets representing distinct geographical areas. These regions include East Asia & Pacific, South Asia, Central Asia, North Africa & Middle East, Sub-Saharan Africa, Europe & Central Asia, and Latin America & Caribbean.

We implemented this by creating mappings between country codes and unique numerical identifiers, as well as between regions and their corresponding numerical identifiers. This allowed for efficient representation of countries and regions using numeric values. Additionally, we performed necessary data adjustments, such as replacing certain values in the dataset.



(a) Mean BMI

(b) Median BMI

(c) Unmet Need Rate

(d) Under5 Mortality Rate

(e) Skilled Birth Attendant Rate

(f) Stunted Rate

Figure 2: Label value versus number of instances for each health indicator

By applying this process to each row in the dataset, we assigned the appropriate target region to each country. The resulting region information was stored in a dedicated column, facilitating subsequent regional analyses.

This categorization enabled us to analyze countries based on their respective regions, providing valuable insights into regional patterns and characteristics. It enhanced our understanding of the data and enabled us to derive meaningful conclusions about the health indicators and associated factors within different regions.

3.5 Feature Selection

Feature selection plays a crucial role in developing accurate predictive models. We employ a combination of techniques to shortlist the most relevant features. First, we analyze the correlation matrix to identify features with high collinearity with the target variables, by using a threshold to shortlist the features. Additionally, we leverage the XGBoost algorithm by training the model on these features and then ranking the features based on their importance. This process helps us prioritize the most influential features and improve the predictive power of our upcoming models.

4 Experiments

4.1 Evaluation method

The primary evaluation metric used in this project is the Mean Column-wise Root Mean Squared Error (MCRMSE). This metric provides a measure of the average error of our model's predictions across all target variables. The Root Mean Squared Error (RMSE) for each target variable is calculated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{1}$$

where $\hat{y}_i$ is the predicted value and y_i is the original value for each instance i. The MCRMSE is then the average of the RMSEs for each predicted column.

In addition to MCRMSE, we also utilize other evaluation metrics to assess the performance of our models. These include:

1. **Coefficient of Determination (R-Squared)**: This metric provides a measure of how well the variations in the predicted values can be explained by the model. It is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{2}$$

where $\bar{y}$ is the mean of the original values.

2. **Mean Absolute Error (MAE)**: This metric calculates the average of the absolute differences between the predicted and actual values. It is less sensitive to outliers compared to RMSE and is defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3}$$

These additional metrics provide a more comprehensive understanding of the model's performance, taking into account different aspects of the prediction errors.

4.2 Experimental Details

Our experiments were conducted using an ensemble of three machine learning models: XGBoost, LightGBM, and CatBoost. These models were chosen for their high performance on structured data and their ability to efficiently handle high-dimensional datasets.

The models were trained using the following parameters:

Model	Learning Rate	Max Tree Depth	Min Data in Leaf
XGBoost	0.3	6	1
LightGBM	0.1	-1 (No limit)	20
CatBoost	0.03	6	1

Table 1: Model HyperParameters

The ensemble model was created using Bayesian Model Averaging (BMA), a technique that calculates the weights of each model in the ensemble based on their performance. The weights were determined based on the inverse of the validation scores, with models that had lower validation scores (indicating better performance) being assigned higher weights.

The models were trained on a dataset consisting of the top 120 features, which were selected through a combination of correlation analysis and feature importance ranking from the XGBoost model. The target variables were six key health indicators: Mean BMI, Median BMI, Unmet Need Rate, Under5 Mortality Rate, Skilled Birth Attendant Rate, and Stunted Rate.

The training process was conducted on the Kaggle platform, using a Kaggle notebook with 30GB of RAM and four cores of CPU. The training of the ensemble model took approximately 4-5 minutes. This relatively short training time allowed us to iterate quickly on different model configurations and hyperparameters.

It's important to note that we are still in the process of fine-tuning the hyperparameters of our models to further improve their performance. Future work will involve more extensive hyperparameter tuning and potentially the exploration of other model architectures and training techniques.

5 Approach

Our current best approach employs an ensemble of three models: XGBoost Regressor, LightGBM Regressor, and CatBoost Regressor. The ensemble model combines the predictions of these three models using weights determined through Bayesian Model Averaging (BMA). BMA is a technique that calculates the weights of each model in the ensemble based on their performance. The better a model performs, the higher weight it is assigned.

In our approach, we calculate the BMA weights based on the inverse of the validation scores. This means that models with lower validation scores (indicating better performance) are assigned higher weights. The weights are then normalized so that they sum to 1.

We use the top 120 features for our models. The order of training is decided in such a way that the later trained models benefit from the earlier predictions that have been used as features. This is achieved through a technique known as regression chaining, where the output of one model is used as an input feature for the next model in the chain. This approach allows us to capture complex dependencies between the target variables and improve the overall predictive power of our models.

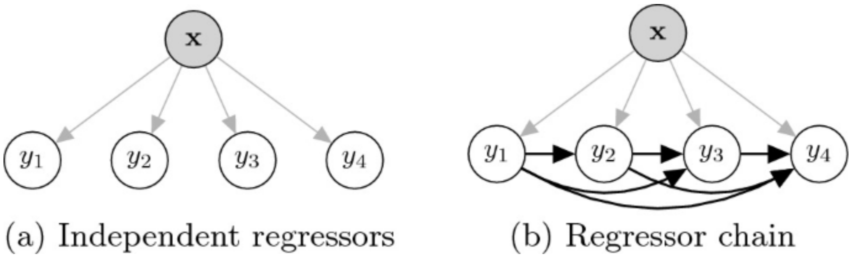


Figure 3: Independent Regression Models vs Regression chained Models

5.1 Chaining Order

The order in which the models are trained is a crucial aspect of regression chaining. In our approach, we have chosen the following order: $Mean\_BMI$, $Median\_BMI$, $Stunted\_Rate$, $Skilled\_Birth\_Attendant\_Rate$, $Under5\_Mortality\_Rate$, and $Unmet\_Need\_Rate$.

This order was chosen based on the understanding of the relationships between the health indicators. The $Mean\_BMI$ and $Median\_BMI$ are fundamental health indicators that can influence other health outcomes. For instance, malnutrition (indicated by low BMI) can lead to stunting in children, hence the $Stunted\_Rate$ follows the BMI indicators.

The $Skilled\_Birth\_Attendant\_Rate$ is placed after $Stunted\_Rate$ because skilled birth attendance can influence both child stunting and mortality rates. The $Under5\_Mortality\_Rate$ follows next as it can be influenced by all the preceding indicators.

Finally, the $Unmet\_Need\_Rate$ is placed last as it is a more complex indicator that can be influenced by a variety of factors, including the other health indicators. This order allows us to capture the complex dependencies between the health indicators and improve the overall predictive power of our models.

6 Results

The results of our models are presented in the table below. The table shows the Mean Column-wise Root Mean Squared Error (MCRMSE) for each model, which is the metric we used to evaluate our models' performance. Lower MCRMSE values indicate better performance.

Our Ensemble Regression Chain model performed the best, with an MCRMSE of 11.2. This model leverages the power of ensemble learning and regression chaining, which allows us to capture complex dependencies between the target variables and improve the overall predictive power of our models.

The Multioutput Stacking Regressor and Tensorflow Decision Forests models also performed well, with MCRMSEs of 11.201 and 11.38, respectively. These models demonstrate the effectiveness of ensemble learning and decision tree-based models in handling our complex, high-dimensional dataset.

The models that incorporated data imputation techniques, namely KNN and MICE, did not perform as well as expected. The Ensemble Regression Chain model with KNN imputation had an MCRMSE of 11.86, while the one with MICE imputation had an MCRMSE of 12.18. These results suggest that while data imputation can help deal with missing data, it may not always lead to improved model performance. The imputation process may introduce noise or distort the underlying data distribution, which could negatively impact the model's ability to learn.

Model/Approach	**MCRMSE**
Ensemble Regression Chain	11.2
Multioutput Stacking Regressor	11.201
Ensemble Regression Chain(+KNN Imputation)	11.36
Tensorflow Decision Forests	11.38
XgBoost top 120 features	12.004
Ensemble (+PCA)	12.12
Ensemble Regression Chain(+MICE Imputation)	12.18
CatBoost top 120 features	12.263
Fully Connected Neural Network Regression	12.249
Country Wise XGBoost models	12.27
Tabnet Regressor	12.60

Table 2: Model Performance

Overall, our results suggest that our approach of using ensemble learning and regression chaining is effective for predicting health indicators from satellite imagery and other geotagged data sources. However, there is still room for improvement, particularly in the areas of data preprocessing, feature engineering, and model tuning. Future work could explore different data imputation techniques, feature selection methods, and model architectures to further improve performance.

7 Conclusion

In this project, we have explored the potential of machine learning and satellite imagery to predict key health indicators in low- and middle-income countries. Our approach has demonstrated promising results, with an ensemble of three powerful boosting algorithms: XGBoost, LGBMRegressor, and CatBoostRegressor, achieving an impressive score of 11.20.

Key highlights of our work include:

- **Regression Chaining:** One of the strengths of our approach is the use of regression chaining, which allows us to take into account the interactions among the target variables. This is particularly important as the health indicators we are predicting are not independent of each other. By incorporating the predictions of earlier models as features in subsequent models, we are able to capture these complex dependencies and improve the overall predictive power of our models.
- **Handling Missing Values:** Our work faced challenges due to the presence of missing values in the dataset. Although we employed the K-nearest neighbors (KNN) imputation technique to address this issue, the method tends to overfit the data, limiting its effectiveness within our framework. In future, we plan to refine our imputation strategies by developing targeted techniques specific to regions with limited available data.
- **Dimensionality Reduction:** The application of Principal Component Analysis (PCA) for dimensionality reduction and feature extraction was limited due to the presence of missing values and its inability to capture non-linear relationships between features. As a future direction, we aim to incorporate Gaussian process latent variable models (GPLVM) for dimensionality reduction, which can better handle non-linear relationships.
- **MOSAIKS Satellite Image Data:** We have already explored the MOSAIKS satellite image data but plan to integrate it more effectively into our model. One of the challenges we faced with the MOSAIKS data was the spatial merging due to the resolution of the data. We aim to refine this process to align the MOSAIKS data with our existing dataset.
- **Feature Engineering:** We also plan to enhance our feature engineering and preprocessing techniques to understand temporal trends in the data better.

In conclusion, our project represents a significant step forward in the application of machine learning and satellite imagery for health monitoring in low- and middle-income countries. We believe that with further refinement and exploration of advanced techniques, we can unlock even greater improvements in prediction accuracy and contribute to the global health monitoring efforts.

References

[1] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[2] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[3] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qi Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30*, pages 3146–3154. Curran Associates, Inc., 2017.

[4] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems 31*, pages 6638–6648. Curran Associates, Inc., 2018.

[5] Google Earth Engine Team. Google earth engine: A planetary-scale geo-spatial analysis platform, 2015.

[6] The dhs program - quality information to plan, monitor and improve population, health, and nutrition programs.

[7] Grigorios Tsoumakas and Ioannis Vlahavas. Random k-labelsets: An ensemble method for multitarget regression. In *Proceedings of the 18th European Conference on Machine Learning (ECML 2007)*, pages 406–417. Springer, 2007.

[8] David H Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.