Maternal and child health monitoring in low and middle income countries (LMICS) using Machine Learning and Deep learning on satellite imagery

IIT Internship Program

Jaya Kishnani
Department of Electronics and Electrical Communication Engineering
Indian Institute of Technology, Kharagpur
jayakishnani@kgpian.iitkgp.ac.in

Vanshika Sahu
Department of Aerospace engineering
Indian Institute of Technology, Kharagpur
vanshikasahu2kgpian.iitkgp.ac.in

Rishan Joshua D'Silva
Department of Aerospace engineering
Indian Institute of Technology, Kharagpur
joshdslv@kgpian.iitkgp.ac.in

Abstract

Monitoring maternal and child health in LMICS can be very challenging due to limited healthcare infrastructures and missing records. Satellite imagery enables in tracing and monitoring earth's resources, ecosystem and events. Thus it can be advantageous for mapping healthcare facilities, population distribution and migration, tracking various health indicators, environmental changes, disease outbreak etc. It also helps in disaster management, forecasting and planning. In this paper we are using various machine learning and deep learning models to predict six health indicators using satellite data from Google earth engine and Demographic and Health Survey (DHS). Our future work will include incorporating more satellite data LANDSAT 8[1] and ICEYE.

1 Introduction

Assessing demographic patterns is critical for policymakers towards sustainable development which affects the ecosystem on earth. However the traditional approaches for measuring the required information are time-consuming, inaccurate and expensive, and majorly relies on incomplete data. For example, in Africa, a significant proportion of countries, specifically 34%, have not yet conducted the agricultural survey for a period exceeding 15 years. Additionally, in half of all African nations, comprehensive surveys capturing the livelihoods of the entire population are conducted only once in every six years[2]. This lack of reliable or development data hinders the effective implementation of policies, programs, and research focused on sustainable development.

Satellite imagery provides images with higher resolution that caters to potential solution to remove data scarcity and unreliability. Every single one of these satellites provides an unparalleled opportunity to obtain images that can assist in evaluating the progress of sustainable development objectives [3]. These objectives encompasses various areas, including eradicating hunger, enhancing healthcare and overall wellness, as well as constructing resilient and sustainable communities. ML models that uses satellite imagery provides more spatial and temporal information of the location, which gives policymakers additional data points and accurate information.

In our project we tried to implement various machine learning not limited to Randomforest, Ridge regressor etc and deep learning algorithms such as ResNet18 [4]on the data collected from Demographic and Health survey (DHS) from 59 countries to predict various heath indicators. We also used the data collected by google earth engine to match other features of the particular country collected by its satellite images and incorporated features extracted from MOSAICKS.

2 Related Work

In the past, the researchers have evaluated human development using census data or household surveys [5]. A census is a systematic process of collecting and recording demographic, social, economic, and other relevant information about the population of a country or a specific geographic area. It is usually conducted at regular intervals, such as every ten years, and aims to provide a comprehensive and accurate snapshot of the population's characteristics and distribution.

Geostatistical techniques hold significance in geo-temporal estimation, their effectiveness is constrained by the necessity of geo-temporal data. In other words, they depend on having tabulated data for numerous geographic regions, spanning both time and space. In the context of human development, this kind of data is frequently absent, hindering the practical application of geostatistical approaches ([6], [7], [8]). Innovative ideas and essential technologies have been introduced in the field of intelligent hyperspectral remote sensing satellite systems[9] since 2011. It gathers tailored real-time data and facilitate interaction between satellites and ground stations, broadening applications to individuals. Yet, challenges encompass privacy, societal values, and legal frameworks for data sharing and distribution[2].

Lately in past few decades, applying machine learning and deep learning models have been significantly used to predict various demographic, socio-economic conditions and health indicators. Unsupervised deep feature extraction for remote sensing image classification [10], unsupervised feature learning for Aerial Scene Classification [11], Tile2Vec: Unsupervised Representation Learning for Spatially Distributed Data[12] etc. Apart from these many digitally traced data has also been collected to incorporate more training data for analyzing international migration flows that includes tracking of IP addresses via various means such as logging in to email accounts, twitter data etc.[13].

Massive open-source contributions that are light-weighted and easy to use frameworks are also developed. MOSAICKS was introduced by Global Policy lab to avail satellite data for wide range applications for addressing global issues by making algorithms radically simpler and robust[14]. *Orbuculum* is an innovative and rapidly evolving platform designed with the specific intent to empower GIS and Earth Observation (EO) researchers by offering a unique avenue for monetizing their machine learning models. *IDinsight* also provides efficient way to implement MOSAICKS using *Microsoft's planetary computer API.*

3 Approach

3.1 Dataset Preprocessing

Obtained a comprehensive dataset containing health indicators, and satellite image features from the geefeatures dataset, it includes the extracted features from google earth engine (GEE) and keys to match it with another type of data. We cleaned the dataset, imputed missing values using mean, median, regression techniques. Removed duplicates, performed feature engineering for meaningful variables, excluded non-predictive columns. Used Random Forest Regressor, bypassing feature scaling due to its nature.

3.2 Exploratory Data Analysis (EDA)

Performed thorough EDA with pandas profiling, visualized using histograms, scatter plots, heatmaps, and box plots. Employed statistical analysis and hypothesis testing to reveal health indicator insights and associations.

- **Linear or non-linear data?**
  By observing the correlation between various features we concluded that data is non-linear since the correlation values are between -1 to 1 i.e. not close to either -1 or 1. So we tried to implement **Kernel PCA**[15] (Principal Component Analysis) for dimensionality reduction with **radial basis function**[16] as its kernel. Due to limited computation, we observed and eliminated correlated features from large dataset to enhance training efficiency.

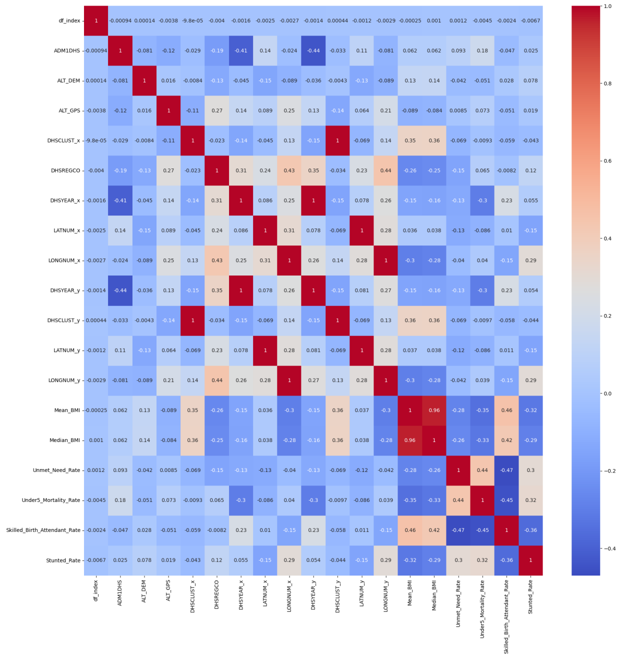

Figure 1: Heatmap - Data correlation

3.3 Model Selection and Training

3.3.1 Machine Learning models

We tried to include and experiment with dataset provided to us. We tuned the parameters of our model using **grid search cross validation** and incorporated more data that boosted our performance.

- **Random Forest Regressor:** After our initial training with Ridge regressor, catboost regressor and other linear and non-linear models. We found that ensemble techniques gave significant boost in performance. We started training with decision trees but tried to include multiple decision trees using random forest to improve overall accuracy of the model and prevent overfitting.
- **XGBoost Regressor (Extreme gradient boosting):** We also implemented XGBoost due to its parallel processing power. It implements regularization to control overfitting, handles missing and sparse data well.

We also decided to include **LightGBM** and **CatBoost** regressor for this task, but currently there is no support for it to support multi regression tasks.

3.4 Model Tuning and Evaluation

We tuned the parameters of our model using Grid Search CV package from scikit-learn library. The we evaluated our model using the evaluation metrics RMSE and R-squared. The high R-squared value and low RMSE value gave us the insight about the model that fits the data well.

4 Experiments

This section contains the following.

4.1 Data

The training dataset used training label merged with the gee_features.csv. The training_label was missing quite a huge chunk of data in outputs as well as the inputs which as a whole reduced the accuracy. The data is processed by initially removing columns that have missing values comparable to the dataset. To tackle the problem of filling in the missing data, country-specific mean of the data was implemented. The country code though available in the 'CCFIPS' and 'DHSCC' columns, had some missing values. Therefore a new column was created - 'ID' that was created from the first 2 characters of the 'DHSID' column. The mean of all the columns was created and grouped by the country code since it is true that healthcare and other facilities are common to a specific geographical boundary. Once all the missing values are taken care of, all the categorical columns such as 'DHSYEAR', 'DHSREGNA', 'DHSCLUST', 'SOURCE' and 'URBAN_RURA' were converted to numerical data. The column that contributed the most was the 'URBAN_RURA' as the accuracy improved significantly with the utilization of this categorical data. Though 'SOURCE' may not seem to contribute to the learning model intuitively, it increased the accuracy nonetheless. The categorical data was then normalized by dividing by the max value of the column. 'ID' was converted to numerical data to include the country classification in the learning.

The MOSAIKS satellite image data was incorporated along with the training data to increase the accuracy. The data included the mosaik features of each country and not every specific data point as the website omitted multiple data points while providing the features along with changing the order of the data. It was observed that using the initial 15 columns of the mosaiks features provides better accuracy than including the first 100 columns of the features.

4.2 Evaluation method

- **Root Mean Squared Error (RMSE):** RMSE is the square root of the MSE and provides a measure of the average absolute difference between the predicted and actual values. RMSE is useful as it is in the same unit as the target variable and allows for easier interpretation.
- **R-squared Score:** R-squared is a statistical measure that represents the proportion of the variance in the target variable that is predictable from the input features. It provides an indication of how well the model fits the data. Higher R-squared values indicate a better fit.

4.3 Experimental details

We have majorly trained two models XGBoost and Random forests and included more data to improve the models performance. The below table contains parameters after using grid search to find best estimators of the hyperparameters of the model. We trained our model using parallel processing (i.e. the CPU cores).

Model name	Model Parameters
Random forest	max_depth=20, min_samples_leaf=4, min_samples_split=4, n_estimators=300, oob_score=True
XGBoost	max_depth = 20, n_estimators = 300, learning_rate = 0.02, nthread = 4

4.4 Results

*The current score of our team **TrioML_T46** on the leaderboard is **6th** and the score is **11.05028** .* The table below includes the results of the above parameters for the entire dataset(including features from Google Earth Engine (GEE) and MOSAICKS that we prepared. We finally trained on the whole dataset but we tested on 20% of the dataset.

Model name	RMSE	R-squared	Leaderboard score
Random forest	7.877039781519336	0.7773653740846921	11.05028
XGBoost	2.9656673233233146	0.9607971488595587	11.55542

5 Analysis

Our best possible accuracy was achieved with Random Forest regressor. We also trained deep learning models but due to computational limitations we tried to fine tune our models and preprocess the dataset by including more features to train. However, it was observed that while training more data (100% rather than the random split of 80%) boosted the accuracy of our predictions by improving RMSE and R-squared parameters.

6 Conclusion

We tried to include more data from *MOSAICKS* and *microsoft's planetary computer* to improve our predictions by lowering root mean squared error(RMSE). We primarily trained and fine tuned hyperparameters over our inital model *random forest regressor* and refined our dataset with other sources. As a result we significantly boosted our score, jumping to 6th position from 19th on leaderboard. In future we will try to load features from Microsoft's planetary computer to incorporate more data (This time due to limited access to computational power and API access issues we couldn't include it). Further we saw that ensemble techniques works better after preprocessing the non-linear and skewed data. We'll also try to incorporate *Kernel PCA* for dimensionality reduction of the data.

7 Team contributions

- **Vanshika Sahu** - Did the preprocessing , feature engineering and model training for the given dataset on ridge and random forest regression. Conducted research analysis and contributed in report writing.
- **Jaya Kishnani** - Explored different models such as XGBoost, LightGBM and ResNet-18 models to be used and contributed in data analysis.Tried to include features via Microsoft's Planetary Computer using LATNUM and LONGNUM columns of gee_features dataset. Conducted research analysis and contributed in report writing.
- **Rishan Joshua D'Silva** - Implemented data filtering and preprocessing and fine-tuned Random Forest Model to improve performance significantly. Tried to implement other models such as XGBoost and ResNet-18. Included the MOSAICKS data. Conducted research analysis and contributed to report writing.

References

[1] D.P. Roy, M.A. Wulder, T.R. Loveland, Woodcock C.E., R.G. Allen, M.C. Anderson, D. Helder, J.R. Irons, D.M. Johnson, R. Kennedy, T.A. Scambos, C.B. Schaaf, J.R. Schott, Y. Sheng, E.F. Vermote, A.S. Belward, R. Bindschadler, W.B. Cohen, F. Gao, J.D. Hipple, P. Hostert, J. Huntington, C.O. Justice, A. Kilic, V. Kovalskyy, Z.P. Lee, L. Lymburner, J.G. Masek, J. McCorkel, Y. Shuai, R. Trezza, J. Vogelmann, R.H. Wynne, and Z. Zhu. Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment*, 145:154–172, 2014.

[2] Amna Elmustafa African Institute for Mathematical Sciences, Amna Elmustafa, African Institute for Mathematical Sciences, Erik Rozi Stanford University, Erik Rozi, Stanford University, Stanford UniversityView Profile, Yutong He Stanford University, Yutong He, Gengchen Mai Stanford University, and et al. Understanding economic development in rural africa using satellite imagery, building footprints and deep models: Proceedings of the 30th international conference on advances in geographic information systems, Nov 2022.

[3] Marshall Burke, Anne Driscoll, David B. Lobell, and Stefano Ermon. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535):eabe8628, 2021.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[5] Sara Randall and Ernestina Coast. Poverty in african households: the limits of survey and census representations. *The Journal of Development Studies*, 51(2):162–177, 2015.

[6] Emily Aiken, Suzanne Bellue, Dean Karlan, Chris Udry, and Joshua E. Blumenstock. Machine learning and phone data can improve targeting of humanitarian aid, Mar 2022.

[7] Amarpreet Singh, Mohammad S. Obaidat, Sandeep Singh, Alok Aggarwal, Kamaljeet Kaur, Balqies Sadoun, Manoj Kumar, and Kuei-Fang Hsiao. A simulation model to reduce the fuel consumption through efficient road traffic modelling. *Simulation Modelling Practice and Theory*, 121:102658, 2022.

[8] Andrew J. Tatem. Worldpop, open data for spatial demography, Jan 2017.

[9] Lingli Zhu, Juha Suomalainen, Jingbin Liu, Juha Hyyppä, Harri Kaartinen, and Henrik Haggrén. *A Review: Remote Sensing Sensors*. 05 2018.

[10] Adriana Romero, Carlo Gatta, and Gustau Camps-Valls. Unsupervised deep feature extraction for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 54(3):1349–1362, 2016.

[11] Anil M. Cheriyadat. Unsupervised feature learning for aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1):439–451, 2014.

[12] Neal Jean, Sherrie Wang, Anshul Samar, George Azzari, David Lobell, and Stefano Ermon. Tile2vec: Unsupervised representation learning for spatially distributed data.

[13] Joshua Blumenstock. Don't forget people in the use of big data for development. *Nature*, 561:170–172, 09 2018.

[14] Esther Rolf, Jonathan Proctor, Tamma Carleton, Ian Bolliger, Vaishaal Shankar, Miyabi Ishihara, Benjamin Recht, and Solomon Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *CoRR*, abs/2010.08168, 2020.

[15] Quan Wang. Kernel principal component analysis and its applications in face recognition and active shape models. *CoRR*, abs/1207.3538, 2012.

[16] Karl Thurnhofer-Hemsi, Ezequiel López-Rubio, Miguel A. Molina-Cabello, and Kayvan Najarian. Radial basis function kernel optimization for support vector machine classifiers. *CoRR*, abs/2007.08233, 2020.