COMP 560 Project: Analyzing the Relationship Between Environmental and Clinical Risk Factors and National GDP

Siddhant Saxena[1] **Joanna Jipson**[1] **Justin Li**[1] **Nhu Le**[1]

Abstract

This paper examines the relationship between national environmental risk factors, clinical health statistics, and economic performance, measured through country-level GDP. Using a consolidated dataset containing environmental, clinical, demographic, and economic indicators, we analyze which variables most strongly correlate with economic outcomes. Our methodology explores a set of supervised learning techniques, including linear regression, logistic regression, ridge regression, and stochastic gradient descent based models, along with basic hyperparameter tuning procedures. After generating model outputs and examining feature influence, we route the results to the Anthropic API to produce structured, high-level interpretations of the observed patterns and correlations. This combination of statistical modeling and large language model assisted analysis provides a flexible framework for understanding how health and environmental vulnerabilities may be associated with macroeconomic stability, offering a foundation for future, more domain-specific modeling.

1. Introduction

Accurately understanding the factors that influence a country's economic performance is a central question in global development, public policy, and international economics. Traditional models focus on capital, labor, and institutions, but recent work suggests that environmental risk factors and population level clinical health statistics may also shape long term outcomes. Environmental pressures such as pollution, natural disasters, and climate related risks can affect productivity and infrastructure, while clinical indicators like disease prevalence, mortality, and healthcare access influence labor force participation and overall economic capacity.

In this project, we use a consolidated dataset of environmental, clinical, demographic, and economic indicators across countries to study how these nontraditional risk factors relate to Gross Domestic Product (GDP). Our project goals are to build a modeling system that accurately predicts national GDP from these variables and to generate meaningful interpretations of the models using a large language model. To do this, we preprocess the data, apply regression based and stochastic gradient descent based methods, and use the Anthropic API to produce structured explanations of model outputs. This work is exploratory rather than causal, but it provides an initial framework for understanding how health and environmental vulnerabilities may be associated with economic performance.

2. Preparing datasets

For our analysis, our data was taken from 3 separate datasets. The first dataset was taken from Kaggle, which was originally sourced from the World Health Organization (WHO). This data describes deaths related to specific risk factors for each country for the years 1990-2017. For our project we chose to limit the data from years 1990-2009. Although the original dataset contained 31 variables, we selected 15 that were most relevant to our analysis. Our second dataset contains the annual population statistics for countries around the world. The sources for this data are History Database of the Global Environment (HYDE), Gapminder, and the United Nations World Population Prospects (UN WPP). We chose to use years 1990-2009 for this data. The third dataset records the GDP for different countries in US dollars from years 1988-2022. This dataset comes from World Integrated Trade Solutions (WITS). We chose to use GDP from the years 1990-2009. To answer our question, we chose to merge these 3 datasets. The datasets had different country names, which made it difficult to merge on the Country column. To fix this, we standardized the country names across the datasets. From there our merged dataset had 35 variables. To cut down

[1] Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. Correspondence to: Siddhant Saxena <sisaxena@unc.edu>, Justin Li <jql@ad.unc.edu>, Joanna Jipson <jjips@ad.unc.edu>, Nhu Le <nhule@unc.edu>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

on rows, we chose to filter the data by removing cases for territories, regions, and small territories that are not considered countries. We also chose to exclude countries that had very small populations because oftentimes their populations were incomplete or unrepresentative. Our final dataset had a total of 35 variables and 1981 rows.

3. Method

We chose to implement 3 models to estimate country-level GDP from environmental and clinical risk factors. The first model made is a linear regression which was optimized using stochastic gradient descent (SGD). This model assumes that there is a linear relationship between the predictors and GDP. We chose to use SGD because it is traditionally more effective on larger datasets, and allows for quicker progress early in training. The second model was a logistic regression model trained using logistic loss optimized by gradient descent. This approach iteratively updates the model parameters to minimize the logistic loss function, enabling the classifier to determine whether a country-year should be labeled as having high GDP or low GDP. The final model was a ridge regression estimator, which incorporates L2 regularization to stabilize coefficient estimates. The data was split into training and test sets to evaluate the generalization of the model. Each model was fit on the training set and then assessed on the test set using appropriate evaluation metrics. For our regression model we used mean-squared error (MSE) and R^2. To measure the performance of our classification model, we used accuracy score, and F1 scores. We also computed feature importances or coefficient magnitudes to assess which variables contributed most strongly to predictions. To generate concise and accurate results, we included prompt engineering using the Anthropic Claude API. The prompt was structured to summarize outputs in a neutral, and academically appropriate style. This was done by assigning a clear role to the API, which added constraints regarding its tone, discouraging informal language. Additionally, by grounding the model firmly in a summarization task, the prompt prevents it from drifting into irrelevant discussion, keeping it aligned with the models goals. We submitted this prompt, along with the model output JSON to the Claude API.

4. Results

Using linear regression, ridge regression, and stochastic gradient descent based models, we first treated GDP prediction as a regression task. We evaluated performance using mean squared error (MSE), which measures the average squared difference between predicted and true GDP, and R^2, which measures the proportion of variance explained by the model. The linear regression model performed best, with an MSE of approximately 2.14×10^{23} and an R^2 of about 0.47. Ridge regression and the SGD based model performed worse, with R^2 scores of about 0.37 and 0.24 respectively. Overall, these results indicate moderate predictive accuracy and suggest that a simple linear model captures a meaningful portion of the variation in GDP.

We also framed the problem as a binary classification task, predicting whether a country falls into a high or low GDP group using logistic regression with both liblinear and gradient descent based solvers. We evaluated these models using accuracy, the proportion of correctly classified countries, and the F1 score, the harmonic mean of precision and recall. The liblinear model achieved an accuracy of about 0.85 and an F1 score of about 0.86, while the gradient descent variant achieved an accuracy of about 0.84 and an F1 score of about 0.82. In general, prediction accuracy was stronger for higher GDP countries and weaker for countries with very low GDP, suggesting that the relationship between the features and GDP is more stable in higher income settings.

Finally, models using the combined set of environmental and clinical features outperformed those using either group alone, indicating that both domains contribute useful information. To interpret these patterns, we routed model outputs and summary statistics to the Anthropic API, which generated structured, high level explanations of which features appeared most influential and how they tended to differ between high and low GDP profiles. These LLM based interpretations were helpful for turning raw metrics into more intuitive narratives about how environmental and clinical risks relate to economic performance.

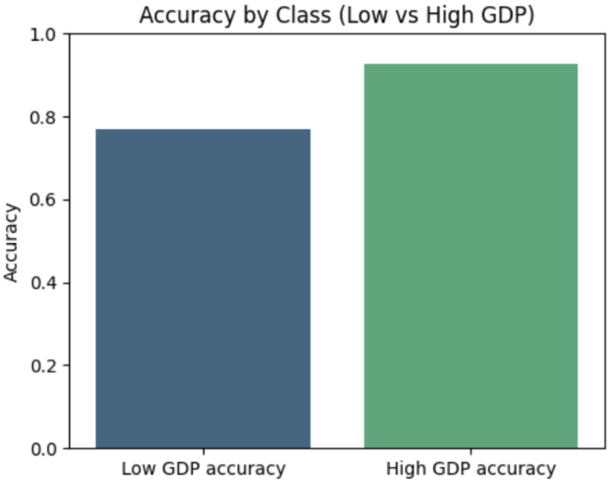

Figure 1.

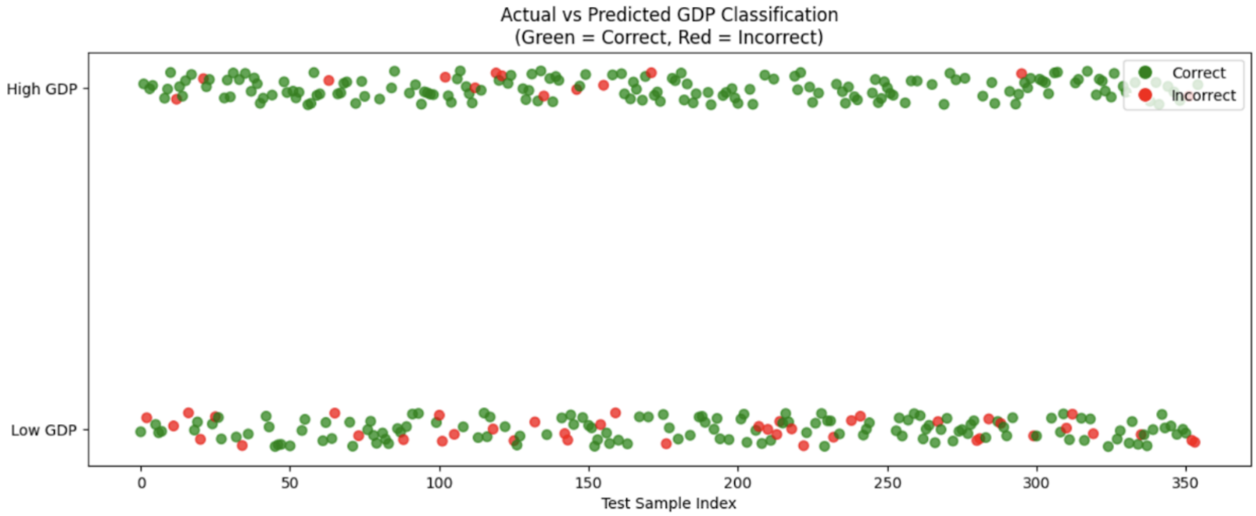

Figure 2.

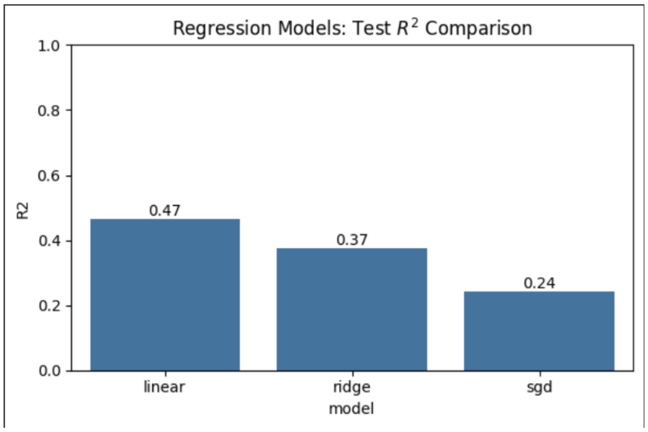

Figure 3.

5. Discussion

Throughout this project, several insights emerged regarding both the modeling process and the quality of the underlying data. Early experiments revealed significant challenges with missing values. When training the initial models, we encountered repeated NaN outputs caused by incomplete or improperly formatted entries in the dataset. This required additional preprocessing and cleaning before any meaningful analysis could be performed. Once these issues were resolved, the models were able to train consistently, but this obstacle highlighted the importance of rigorous data validation in cross-country analyses that combine multiple sources.

Another challenge involved the computational environment. The notebook kernel frequently failed to update or reflect changes during iterative testing, which slowed down experimentation and required multiple restarts. While not directly related to the modeling itself, this issue affected workflow efficiency and underscored the need for a more stable environment or a structured pipeline for repeated experiments.

Modeling choices also presented difficulties. For example, stochastic gradient descent with a linear objective initially produced unrealistic predictions, often outputting unusually large or unstable values. This was ultimately traced to an insufficient number of iterations during optimization. Increasing the iteration count and adjusting learning rates stabilized the results and produced more interpretable outputs. We also experimented with ridge regression as a regularized version of linear regression, which helped prevent overly large coefficients and produced more stable predictions. These experiences illustrate how sensitive gradient-based methods can be to hyperparameter choices and how easily they can diverge when working with heterogeneous global datasets.

Despite these challenges, several approaches performed reasonably well. Linear regression produced interpretable baseline relationships, and logistic regression helped validate how certain features behaved under different transformations or binary framing of the data. The Anthropic API added an additional layer of interpretability by generating structured explanations of feature behavior, helping us contextualize correlations identified by the models.

There are several limitations to note. The analysis is correlational and cannot establish causal relationships. The dataset, while broad, does not include important economic, institutional, or political variables that may influence GDP.

Additionally, the cross-sectional structure prevents us from examining how changes in environmental or clinical indicators over time affect economic performance. Future work could extend this study by incorporating time-series data, expanding the feature set to include additional socioeconomic controls, or applying causal inference techniques to better understand underlying mechanisms.

Overall, this project demonstrates that environmental and clinical factors show meaningful statistical relationships with national GDP, but also highlights the importance of careful data handling, thoughtful model selection, and transparent interpretation when analyzing complex global datasets.

6. Conclusion

In this project, we examined how environmental risk factors and clinical health statistics relate to national economic performance, measured by GDP. Using a multi-domain dataset with country-level indicators, we trained a range of supervised learning models and evaluated their ability to capture patterns between risk profiles and economic outcomes. Across several model families and feature representations, we consistently found that both environmental and clinical variables showed meaningful correlations with GDP, suggesting that health and environmental conditions are informative signals of a country's broader economic status. To help interpret these patterns, we routed model outputs to the Anthropic API, using a large language model to generate structured, high-level explanations of which factors appeared most influential and how they might interact.

Although our analysis is correlational and does not establish cause and effect, the results support the broader idea that environmental and clinical conditions are closely intertwined with economic outcomes. By integrating these dimensions into quantitative modeling and leveraging large language model assisted interpretation, this project provides an initial framework for understanding how health and environmental vulnerabilities may be associated with a country's long-term development trajectory and economic stability.

References

Our World in Data. Sources of our population dataset. https://ourworldindata.org/grapher/sources-population-dataset, 2024. Accessed: 2025-11-18.

Varpit94. Worldwide deaths by country/risk factors. https://www.kaggle.com/datasets/varpit94/worldwide-deaths-by-risk-factors, n.d. Accessed: 2025-11-18.

World Bank. Gdp by country in current us$ (indicator: Ny.gdp.mktp.cd), 1988–2022. https://wits.worldbank.org/CountryProfile/en/country/by-country/startyear/ltst/endyear/ltst/indicator/NY-GDP-MKTP-CD, n.d. Accessed: 2025-11-18.