A NEW WEBAPP FOR THE STUDY AND MODELING OF HUMAN PROGRESS

Luis Alberto Ahumada Abrigo, M. S.

Nima Zahadat, Ph.D.

The George Washington University

Spring 2021

ABSTRACT

Data from various academic institutions and international organizations shows enormous improvements in human well-being throughout the world in the past centuries. Unfortunately, there seems to be a gap between reality, characterized by improvements, and public perception, which tends to be negative about the current state of the world, skeptical about the world's future, or illiterate about the causes of progress. Projects like HumanProgress.org gather empirical data from reliable sources that look at worldwide long-term trends and host it in an accessible and interactive online platform designed for students, researchers, and the general public. This project goes one step further by developing a new interactive tool to visualize the relationship between variables, causation through machine learning models such as linear regression, and prediction through neural networks. The project introduces an online dashboard powered by Plotly and delivers two case studies analyzing the variable life expectancy.

TABLE OF CONTENTS

INTRODUCTION

One of the most analyzed questions in the social sciences is how countries develop. Why do some societies become successful, and why do others fail? What are the necessary conditions in a country to achieve development? From these questions, multiple other derived questions arise, for example, how do we measure which societies are more developed than others? What are the indicators that reflect the level of development of a country? Are these indicators correlated with the causes?

According to international organizations like the World Bank, among the most used indicators to measure development are per capita income, life expectancy, and the Human Development Index (HDI). Can researchers relate these indicators to independent variables that explain their variations? This work focuses on streamlining the response to these questions by providing a platform to model different variables and find measurable and quantitative explanations for changes in human progress indicators. The platform, powered by HumanProgress.org data, will allow researchers to venture possible answers to the questions above.

We used the data that HumanProgress.org, a non-profit organization, provides for free and open on its website to carry out this task.

HumanProgress.org is an initiative of the Cato Institute to introduce the general public to the vast majority of human progress indicators available today. The project consists of 2,100 datasets from more than 120 different sources, usually international organizations or academics with high reputations and recognition. The website allows users to display indicators by country, year, and type of chart (maps, line charts, bar charts, etc.)

Thanks to the internet and technology, multiple other platforms host historical data on countries and progress in general.

However, one of the main gaps in online platforms is the ability to establish relationships between variables. Available platforms such as HumanProgress or OurWorldInData only provide the ability to see the evolution of one variable at a time, but not the ability to correlate variables, view multiple indicators simultaneously, or perform models to generate predictions or measure causality.

The main objective of this work is to develop an online platform that allows researchers to visualize the relationship between different variables and carry out simple models that indicate the connection between variables.

This work will present a case study carried out with the developed platform and two models as a secondary objective. This case study seeks to understand which variables of human progress are most related to the improvement in life expectancy in a country.

We will divide the work into two models, a linear regression model and a deep neural network model, their key findings, and visualization.

LITERATURE REVIEW

Researchers have been trying for decades to answer why some countries turn out to be more successful than others. This challenge requires having a well-defined understanding of progress and development. Academics and social organizations have specialized in giving a clear definition of progress to evaluate the evolution of the historical performance of societies.

"Progress" is a measurable variable with empirical data. Although there are different indicators for each dimension of progress, they are often interconnected. For example, there is a relationship between low levels of infant mortality and high life expectancy levels. At the same time, high levels of sanitation and public sanitation tend to correlate with low levels of diseases such as cholera and other viruses.

That said, it is fair to say that there is a broader consensus among academics and experts on how to measure progress than on which are the necessary conditions for progress. In other words, it is easier to determine the variables that demonstrate the improvement of society than to explain the causes of how these variables develop over time.

We will divide this chapter into four sections.

First, we will present the main definitions of progress and development. Second, we will review the most accepted theories and explanations about the progress and development of societies throughout history. Third, we will look more specifically at life expectancy as an indicator of progress since we will use it to evaluate our models in the course of this paper.

Finally, we will summarize some of the main online initiatives available today to explore and visualize historical data on countries' progress.

Development and Progress

The United Nations Development Program (UNDP, 2018) generated an indicator called the Human Development Index (HDI) that is used to classify countries into four levels of human development. The indicator consists of life expectancy, education (literacy and enrollment levels), and economic measurements such as per capita income. Governments around the world use the HDI to evaluate their public policies. Also, international organizations and academic researchers use it to measure and compare the development of countries. The Human Development Index (HDI) focuses on three basic dimensions of human development: "the ability to lead a long and healthy life, measured by life expectancy at birth; the ability to acquire knowledge, measured by mean years of schooling and expected years of schooling; and the ability to achieve a decent standard of living, measured by gross national income per capita." (UNDP, 2018)

Similarly, the United Nations in 2000 proposed the Millennium Development Goals (MDGs). A set of 8 development goals to be achieved in 2015. Among them was to reduce poverty and extreme hunger, provide universal access to primary education, achieve gender equality, lower infant mortality, increase maternal health, combat diseases such as malaria and HIV/AIDS, and achieve environmental sustainability. As of today, progress towards the goals was uneven. Some countries achieved many goals, while others were not on track to realize any. (UN, 2015)

More recently, Steven Pinker, researcher and psychologist at Harvard, in his book "Enlightenment Now", proposes a definition of progress, which although it is broader, is also widely accepted due to its invocation of common sense. Pinker (2018, p.51) states:

> "What is progress? You might think that the question is so subjective and culturally relative as to be forever answerable. In fact, it's one of the easier questions to answer.
>
> Most people agree that life is better than death. Health is better than sickness. Sustenance is better than hunger. Abundance is better than poverty. Peace is better than war. Safety is better than danger. Freedom is better than tyranny. Equal rights are better than bigotry and discrimination. Literacy is better than illiteracy. Knowledge is better than ignorance. Intelligence is better than dull-wittedness. Happiness is better than misery. Opportunities to enjoy family, friends, culture, and nature are better than drudgery and monotony.
>
> All these things can be measured . If they have increased over time, that is progress. "

From this definition arise multiple categories where we can measure human progress. Each of these dimensions has numerous indicators that are used to track its evolution. The discussion about which are the most precise indicators for each dimension is even more extensive and is beyond this project's scope.

Theories on progress and development

After defining progress and development, it remains to explore the causes that generate progress and development. There are multiple theories about the development and failure of different countries. Below we present a summary of the most recognized hypothesis.

In 1776, Adam Smith (2002) argued that the origin of the prosperity of societies like England or the Netherlands has to do with the adoption of economic practices such as the division of labor, free enterprise, free competition, and free trade. In turn, he points out that mercantilist, absolutist, and centralist ideas such as monopolies, tariffs, protection of local production, and intervention by the

authorities in the economy, tend to predominate in countries that have not been enriched in the same way.

More recently, Douglass North (1991), a Nobel Laureate in economics, highlights the importance of a society's institutions. Both formal institutions such as constitution, laws, property rights, and more informal institutions such as customs, traditions, codes of conduct, are "humanly devised constraints that structure political, economic, and social interactions." Thus, for North, the Institutions would be a fundamental pillar to define the necessary conditions for a society to achieve order, security, and development.

Following North's tradition, Daron Acemoglu and James Robinson (2020) delve into the institutional political economy argument. They add that factors such as geography, climate, genetics, beliefs, culture, or religions are secondary to the importance of institutions. Their main thesis is that the institutions that govern a certain territory that will make it prosper. Most importantly, societies will succeed if they are respectful of private property, guarantee an effective separation of powers and enable the proper functioning of a free-market economy.

On the other hand, Jared Diamond (1997) gives more importance to biogeographic factors when explaining the dominance of certain civilizations over others on a historical level. He argues that important cultural differences that can mark the superiority of a society are the result of environmental differences that tend to be reinforced thanks to positive feedback loops throughout history.

In a slightly different line, great economists such as Ludwig Von Mises, Friedrich von Hayek, and John Maynard Keynes believed that the prevailing ideas among the citizens of a society and the development of intellectual movements can play a transformative role in societies. Polanyi (2012) states that these authors "understood that ideological beliefs can sustain an unsustainable social and economic order, or release forces of revolutionary change."

Following the same argument, Deirdre McCloskey (2011), suggests, through historical analysis from the 15th century, that the evolution of idiosyncrasy in a community with respect to certain ideas and social roles would explain the economic success of societies. In his book "Bourgeois Dignity: Why Economics Can't Explain the Modern World," McCloskey argues that the change in rhetoric about the valuation of social roles such as business, innovation, and entrepreneurship contributed as the main factor responsible for the economic success in Northwest Europe from the late 18th century.

This is a brief review of an extensive debate that is currently ongoing. Thanks to available technology and new data collection techniques, researchers increasingly obtain more and better data from the past,

study phenomena with greater precision, challenge visions, and find better explanations for these questions.

The use of life expectancy as an indispensable indicator of human progress

Life expectancy is a key metric for assessing population health. It is broader than infant mortality, which focuses solely on mortality at a young age. Life expectancy captures mortality along the entire life course. It tells us the average age of death in a population (Roser et al., 2019).

The average life expectancy at birth for people was about 30 years for most of human history. Children used to die before they reached their fifth birthday. In 1820, the global average life expectancy was still about 30 years. However, around that time, life expectancy in the West began rising at a rate of about three months per year. During the past 200 years, global life expectancy has more than doubled, now reaching more than 72 years according to the World Bank. That increase was largely a consequence of better nutrition and deployment of public health measures such as filtered water and sewers (Bailey and Tupy, 2020. p. 55).

The case study in this report will attempt to demonstrate the strength of life expectancy as an indicator of progress. Life expectancy at birth is not only relevant to measure the health of infants, but also it is a tool to measure the quality of life in a society as a whole.

Although the indicator strongly depends on the health of infants, it also reflects the reality of societies that affects adults. For example, countries with fatal violence, such as homicides and war, tend to have a lower life expectancy. At the same time, countries with better economic conditions reflected in indicators such as GDP per capita tend to score more in life expectancy due to their ability to provide medical services to the population to combat diseases, epidemics, vaccinate their population, etc.

Initiatives to understand progress data

Another relevant aspect to explore for this study are the initiatives dedicated to the collection, dissemination and visualization of international development data.

HumanProgress.org is an initiative of the Cato Institute to introduce the general public to the human progress indicators available today. It has more than 2,000 datasets from more than 120 different sources, international organizations, and academics with high reputation and recognition. The website provides the possibility of displaying indicators according to country and year and displaying maps and charts.

The editorial line of HumanProgress.org (2021) believes in the dissemination of a realistic understanding of the current situation in developed societies and the world at large. They emphasize

that almost all material conditions are much better compared to any period in the past. They add that the main reasons are embracing illustration values, such as reason, science, humanism, and freedom.

Our World in Data, an initiative of Max Roser, collects data from different sources and aims to organize the information by subject and present a brief history of each indicator and its evolution.

There are also organizations such as the World Bank, OECD, World Economic Forum that also collect progress indicators and generate platforms for their visualization thanks to their connections with governments worldwide.

V-Dem (Varieties of Democracy) is an independent research platform founded by Professor Staffan I. Lindberg in 2014. Its objective is to conceptualize and measure democracy around the world (V-Dem, 2021). For this, they have created indicators and categories to collect data and measure the quality of democracy around the world.

The Human Freedom Index (2020) is an annual index of economic and personal freedoms that seeks to measure the evolution of certain institutional indicators in a country and human and economic rights and liberties.

RESEARCH METHODOLOGY

Project Scope

Although all the organizations above have a website, none provide the ability to users to visually compare indicators to find relationships between variables or models intended to find causality. Instead, this is usually a task left for researchers in universities.

This project aims to improve the capacity of researchers, students, and interested people to visualize the relationships between variables and even deploy simple models to understand the behavior of multiple variables.

Methodology

The nature of this project is interactive and dynamic. Users must have the ability to view updated data seamlessly. Therefore, instead of requesting all the data from the organization, we developed a data scraping code that retrieves the data from the website.

The program, developed with Beautiful Soup, enters HumanProgress.org/datasets and accesses all the pages with a csv file. The program downloads the csv's files and saves them on the local machine.

We used Python 3.8 and common packages (pandas, Numpy), to preprocess the data. We performed data preprocessing techniques to clean the data and prepare it for machine learning algorithms and data visualization. We use linear regression interpolation to fill the missing values.

The paper focuses on a case study as an attempt to explain the "Life Expectancy" variable. We conducted an Exploratory Data Analysis (EDA) with the variables in this subset.

The case study consists of two models. The first model consists of a Multiple Linear Regression using the package statsmodel for Ordinary Least Squares (OLS). The second model is a Deep Neural Network (DNN) regression using Tensorflow. We will use Mean Squared Error (MSE) as the performance metric to evaluate the models.

We used Plotly to generate the visualizations, Dash to create the dashboard, and Heroku to deploy the dashboard as an interactive online app.

Finally, we do analyses and conclusions based on the results of all processes in the project.

DATA

We put together all the countries' indicators in a single file containing 9,900 observations and more than 2,100 indicators.

We made the following modifications to the original data:

- We reduced the number of countries to 162. This is a quantity regularly used by indices from different organizations that include countries whose data is reliable and consistent over time. Additionally, we renamed all to a World Bank standard to avoid inconsistencies.

- We reduced the number of indicators to 975 and excluded non-dynamic indicators (indicators that do not disaggregate the data by country).

- The time frame used is from 1960 to 2020. In addition, for those countries whose periods are not available in some indicators, we conducted a linear regression to fit the curve of each country. We then used an interpolation technique to fill the NAs.

The extracted data finally consists of a dataset with 9,883 rows, equivalent to each country analyzed for each year. And 977 columns that are equivalent to the selected indicators with enough data available to do the analysis.

VARIABLES

Our case study will be an attempt to explain the factors that influence the variable "Life Expectancy." Furthermore, we will seek to predict its variations using a set of independent variables. For this, we will use two models: multiple linear regression and Deep neural network regression.

The dependent variable is "Life expectancy at birth." The independent variables are mortality rate, vaccination, the Human Development Index, GDP per capita, GDP's annual growth rate, poverty, mean years of primary schooling, children labor, access to electricity, improved drinking water sources, life satisfaction, CO2 emissions, homicide rate.

Why life expectancy? As discussed in the literature review, life expectancy is a variable that, according to international organizations and the academy specialized in the analysis of human progress, tends to summarize several dimensions of human progress.

Intuitively, one might think that life expectancy refers only to living longer. But in reality, variables of all dimensions tend to affect this indicator. For example, countries where there is less violence and homicides tend to have higher life expectancy levels since there are many deaths from homicide. Though these regularly appear among men between 15 and 28 years old, these are registered in the statistics, and inevitably they will affect the average of a nation. Another example related to violence is wars. A country with civil wars, drug trafficking, or terrorism will also have levels of life expectancy affected.

Health indicators provide us with another example. By having a better response to serious diseases, more vaccines, more prevention, and more control over deaths associated with infant mortality, the life expectancy indicator will also increase.

Finally, another variable under analysis is the economy, how the best economic conditions in a country also tend to explain the best in life expectancy. A country where its citizens have a decent job and keep their employees in good condition with a salary that allows them to live generates greater mental and physical health that ultimately determines an improvement in life expectancy.

By having greater economic opportunities, citizens also choose healthier lifestyles. Check-in at the doctor more often, exercise, healthy free time, better diets, better education on maintaining health, etc.

We selected the independent variables based on two books, Ten Global Trends by Marian Tupy and Ronald Bailey and Enlightenment Now by Steven Pinker. Both books compare and justify the use of specific indicators to measure the desired variables. In addition, the selection of variables maintains a

fair representation of the different areas of progress in society, including health, education, economy, consumption, happiness, violence, and the environment.

Table 1 summarizes the selected variables, their unit of measure, and source.

Table 1: Variables

Indicator	Unit	Description	Source
Life expectancy at birth	years	Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.	World Bank
Mortality rate, children under 5	per 1,000 live births	Under-five mortality rate is the probability per 1,000 that a newborn baby will die before reaching age 5 if subject to current age-specific mortality rates.	World Bank
DTP3 Diphtheria, tetanus, pertussis vaccination	percent of children aged 0 to 12 months	DPT refers to a class of combination vaccines against three infectious diseases in humans: diphtheria, pertussis (whooping cough), and tetanus. This indicator reports data on those receiving the third dose of this vaccine.	Unicef
Human Development Index	scale 0-1	Human Development Index measures three basic dimensions of human development: life expectancy at birth; adult literacy rate (2/3 weight) and school enrollment ratio (1/3 weight); and GDP per capita (PPP in U.S. dollars).	U.N.
GDP per capita	2018 U.S. dollars	GDP per capita is gross domestic product divided by midyear population. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies, not included in the value of the products. This version contains GDP growth and levels adjusted for rapidly falling ICT prices. The data is in constant 2018 U.S. dollars.	The Conference Board
GDP, annual growth rate	percent	Annual percentage growth rate of GDP at market prices based on constant local currency. Aggregates are based on constant 2010 U.S. dollars. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources.	MDG

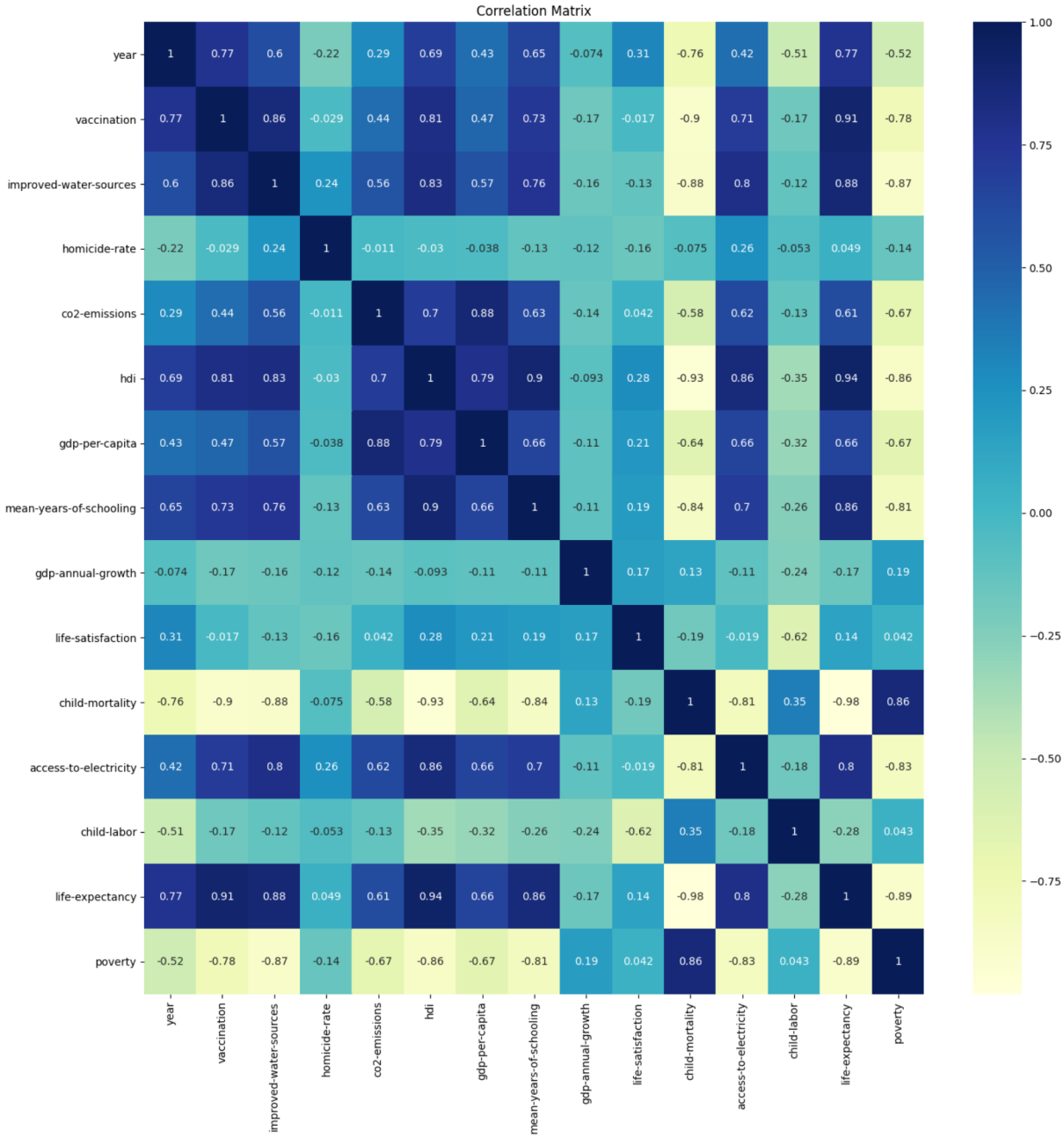
Poverty headcount ratio at $1.90 a day	percent of population, 2011 international dollars, PPP	Poverty headcount ratio at $1.90 a day is the percentage of the population living on less than $1.90 a day at 2011 international prices. As a result of revisions in PPP exchange rates, poverty rates for individual countries cannot be compared with poverty rates reported in earlier editions.	World Bank
Mean years of primary schooling	number of years	Estimates of average primary years of schooling for 15-64 year-olds from census data and interpolation procedures. Also includes projections.	Robert Barro, Jong-Wha Lee.
Economically active children	percent of children aged 7 to 14	Economically active children refers to children involved in economic activity for at least one hour in the reference week of the survey.	World Bank
Access to electricity	percent of population	Access to electricity is the percentage of the population with access to electricity. Electrification data are collected from industry, national surveys and international sources.	World Bank
Population using improved drinking water sources	percent	The proportion of the population using an improved drinking water source, including piped water; public tap/standpipe; borehole/tube well; protected dug well; protected spring; rainwater collection and bottled water.	MDG
Share of people who say they are very happy or quite happy	percent	This data is compiled from the survey question, "Taking all things together, would you say you are: happy/very happy, or unhappy/very unhappy?" To allow for greater simplicity, we have combined happy and very happy, and unhappy and very unhappy.	World Values Survey
CO2 emissions	per person metric tons	Carbon emissions per capita are measured as the total amount of carbon dioxide emitted by the country as a consequence of all relevant human (production and consumption) activities, divided by the population of the country.	World Bank
Homicide rate	per 100,000	Unlawful death purposefully inflicted on a person by another person. Data on intentional homicide should also include serious assault leading to death and death as a result of a terrorist attack.	UN Office on Drugs and Crime

EXPLORATORY DATA ANALYSIS

First, we create a correlation matrix (Figure 1) to check the relationship between the different variables. Life expectancy, our dependent variable, has a strong positive correlation with vaccination, improved water sources, the Human Development Index, mean years of schooling, access to electricity, and poverty, and a strong negative correlation with child mortality.

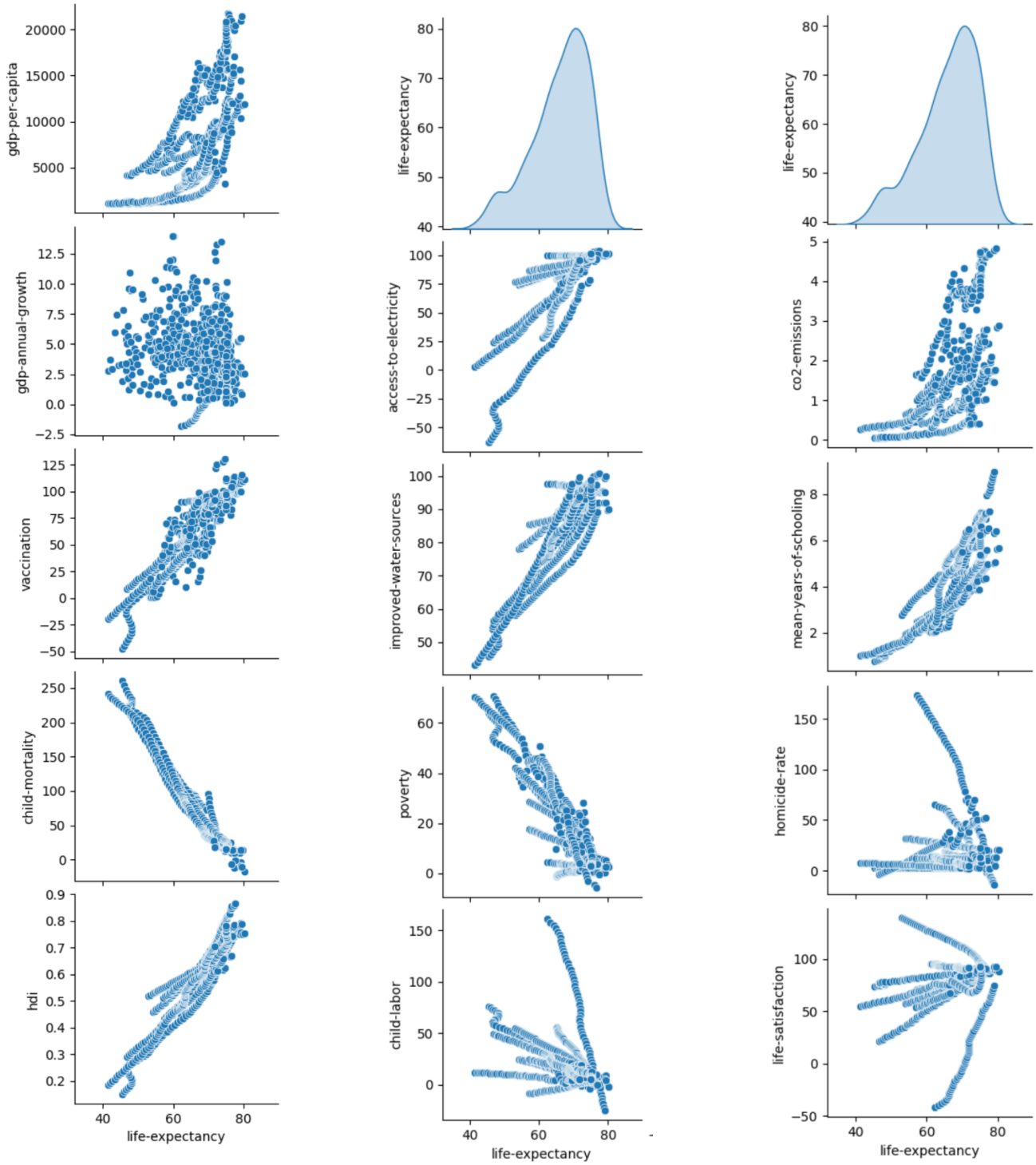
HDI is strongly related to life expectancy, education, and consumption variables because it is an index constructed from variables such as life expectancy at birth and child mortality. Furthermore, we see that human health variables are correlated with variables that improve living conditions, such as poverty and access to improved water sources.

Figure 1: Correlation Matrix



In Figure 2, we can see how the observations of the different variables behave versus life expectancy. As in the previous graph, the variables with a higher correlation coefficient tend to show an upward or downward trend versus life expectancy.

Figure 2: Scatter plot, independent variables vs. life expectancy



MODEL 1: MULTIPLE LINEAR REGRESSION

Feature selection

After a feature selection analysis, and with the use of performance indicators such as T-test, AIC, BIC, and Adjusted R2, we decided to remove variables that were not in the model, either because of their null relationship with life expectancy or because of their multicollinearity problems. A Single Value decomposition analysis also showed that certain variables have multicollinearity problems. For this reason, we removed the following variables through a backward stepwise regression: child labor, HDI, life satisfaction, CO2 emissions, and GDP annual growth.

Results

After running the model using the remaining variables, we ended up with a model with an Adjusted R-squared of 0.979. In other words, the model explains the variation in life expectancy by almost 98%.

Table 2: OLS Regression results

OLS Regression Results

Dep. Variable:	life-expectancy	R-squared:	0.979
Model:	OLS	Adj. R-squared:	0.979
Method:	Least Squares	F-statistic:	2774.
Date:	Wed, 05 May 2021	Prob (F-statistic):	0.00
Time:	03:05:45	Log-Likelihood:	-781.94
No. Observations:	488	AIC:	1582.
Df Residuals:	479	BIC:	1620.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	74.0869	1.131	65.515	0.000	71.865	76.309
vaccination	0.0602	0.005	12.705	0.000	0.051	0.070
improved-water-sources	-0.0628	0.012	-5.127	0.000	-0.087	-0.039
homicide-rate	0.0103	0.003	3.925	0.000	0.005	0.015
gdp-per-capita	9.728e-05	1.65e-05	5.879	0.000	6.48e-05	0.000
mean-years-of-schooling	0.4307	0.074	5.800	0.000	0.285	0.577
child-mortality	-0.0854	0.003	-26.534	0.000	-0.092	-0.079
access-to-electricity	-0.0162	0.003	-4.750	0.000	-0.023	-0.010
poverty	-0.0833	0.008	-10.955	0.000	-0.098	-0.068

Omnibus:	0.708	Durbin-Watson:	2.056
Prob(Omnibus):	0.702	Jarque-Bera (JB):	0.814
Skew:	0.058	Prob(JB):	0.666
Kurtosis:	2.836	Cond. No.	2.03e+05

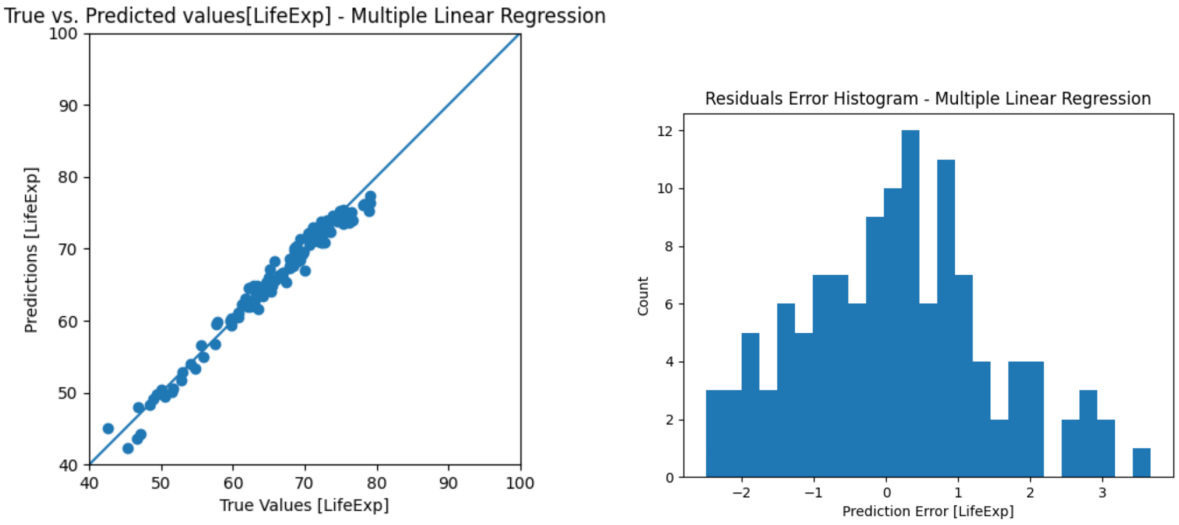
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.03e+05. This might indicate that there are strong multicollinearity or other numerical problems.

In Table 2, we can see how the p-value for all the independent variables is lower than our limit of 0.05 at a confidence level of 95%. Therefore, we can affirm that all the variables are statistically significant.

Also, our AIC and BIC are relatively low, and the model passes the T-test. The model indicates a good fit and allows us to reject the null hypothesis that there is no relationship between the variables studied.

The plot of predicted values is consistent with true values, and the residuals have a normal distribution close to zero. The MSE of this model is 1.45.

Figure 3: Multiple Linear Regression results



MODEL 2: DEEP NEURAL NETWORK REGRESSION

The second model is a Deep Neural Network sequential model for regression using Tensorflow. Before running the model, we perform a normalization process of the variables to standardize their units of measurement. Then we input all the variables into the deep neural network model. Table 3 shows how it is composed of the normalization layer, two hidden, nonlinear, Dense layers using the Relu nonlinearity, and a linear single-output layer. The model contains 4,801 trainable parameters.

Table 3: DNN Sequential model overview

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
normalization (Normalization (None, 8)                 17
_________________________________________________________________
dense (Dense)                (None, 64)                576
_________________________________________________________________
dense_1 (Dense)              (None, 64)                4160
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 65
=================================================================
Total params: 4,818
Trainable params: 4,801
Non-trainable params: 17
```

Results

The model can fit our data after approximately 30 epochs ending an error close to 0.

Figure 4: DNN loss history plot

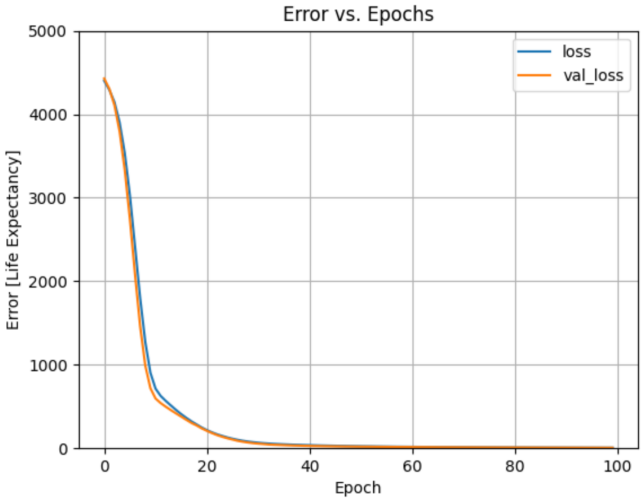
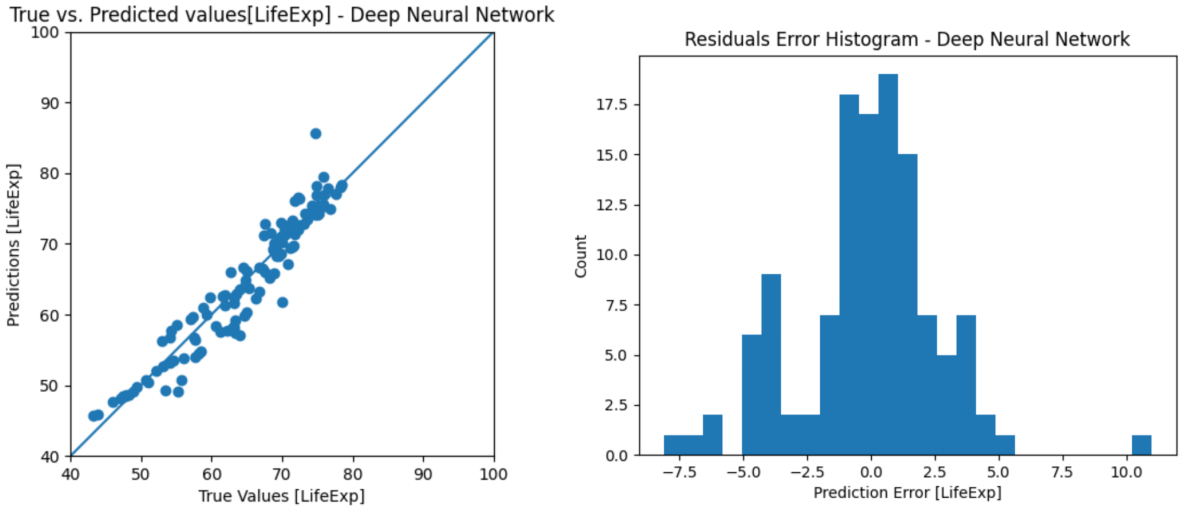


Figure 5: DNN results



While the sequential model has predicted values fairly close to true values, it did not get the precision of the Multiple Linear Regression model. This model has an MSE of 8.16, higher than the first model. This difference is also visible in the residual histogram, where the values are farthest from zero.

KEY FINDINGS

The multiple linear regression model seems to be more effective in this case than the Deep Neural Network model for regression. This occurs because the process in the first model involved a backward stepwise feature selection that removed variables that were not significant to analyze life expectancy. The neural network model, on the other hand, used all the variables. Although its performance was not bad, the result may have been slightly distorted by the influence of variables that do not impact life expectancy.

When studying variables of this type in the social sciences, it is advisable to use multiple linear regression models. Often the relationships between variables are not deterministic as in the physical sciences.

Neural networks will be most useful when carrying out analyzes where it is not strictly necessary or to understand causality or the "black box" but rather when it is important to have an accurate prediction.

Table 4 compares both models using Mean Squared Error as the evaluation metric.

Table 4: Model comparison

	Mean Squared Error [MSE]
Multiple Linear Regression	1.459183
Deep Neural Network	8.169984

DATA VISUALIZATION

We constructed a data visualization and modeling platform where users can replicate the case study. The platform uses Plotly Dash, and we used Heroku App to deploy it online. Users can access the application using the following URL: https://humanprogressorg.herokuapp.com/

The platform is divided into two parts. The first one is for data analysis and contains five charts. The user can select up to 6 variables and up to 4 countries and analyze the indicators over time. The first chart is a double-axis chart that allows users to see relationships between two variables in the same graph. The following four graphs are line charts that show the evolution of the variables selected for each country.

The second part allows users to perform models using Multiple Linear Regression from the statsmodel package and the Deep Neural Network regression from Tensorflow.

The user must select a dependent variable and four independent variables. The model will take approximately 15 seconds to process the model until it delivers the output.

The first graph of the output is a correlation matrix between the selected variables.

The following graphs are scatter plots between the dependent variable and each independent variable.

The next graphs are charts comparing the true values versus the predicted values of each model. In addition, we included a chart to see the distribution of the residuals for each model.

Finally, the platform outputs the summary table of coefficients of the OLS regression model of statsmodel.

CONCLUSION

This project contributed to users' and researchers' ability to perform simple analyses in an open and intuitive platform to understand the relationships between the variables of human progress.

For this, we developed an online application where users can freely and smoothly explore the variables, visualize them, and perform models.

The case studies showed that the linear regression model is useful to understand and explain correlations between variables. More specifically, some independent variables selected like vaccinations, income, child mortality, and improved water sources largely explain the changes in life expectancy levels of life expectancy globally.

In addition, it was a test for more advanced techniques such as DNN in social science data. Although the model was accurate, it is not ideal because it does not provide information on which variables affect the dependent variable more or less.

Project Limitations

Regarding the data of HumanProgress.org, one of the limitations is its timeframe. For many countries, there is no data available, especially in the last few decades. However, this seems to be a trend that is positively improving, since thanks to technology, it is becoming easier to collect data, organize and publish it openly.

We can refine the data visualization application by adding more data, more visualization capabilities, and more advanced and flexible models than are now available.

Recommendations for future research

Future research should attempt to explain the development of countries using the data available at HumanProgress.org.

Researchers can test hypotheses with new empirical data because many organizations recollect new data about the countries' past thanks to new methods for data collection.

Also, new models of quantitative data and qualitative data should be experimented with through techniques such as NLP and historical analyses.

REFERENCES

Acemoglu, Daron; Robinson, James. (2020). "Why Nations Fail - Why Nations Fail". whynationsfail.com.

Bailey, R. and Tupy, M. (2020). Ten Global Trends Every Smart Person Should Know: And Many Others You Will Find Interesting. Available at: www.tenglobaltrends.org

Diamond, Jared (1997). Guns, Germs, and Steel: The Fates of Human Societies. W.W. Norton & Company. ISBN 978-0-393-03891-0.

HumanProgress.org (2021). Who we are. Available at: https://www.humanprogress.org/about/

Human Freedom Index (2020). The Human Freedom Index. Available at: https://www.cato.org/human-freedom-index/2020

McCloskey, Deirdre (2011). Bourgeois Dignity: Why Economics Can't Explain the Modern World. University of Chicago Press. ISBN 978-0226556741

North, Douglass (1991). "Institutions". Journal of Economic Perspectives. 5 (1): 97–112. doi:10.1257/jep.5.1.97.

Pinker, Steven (2018). Enlightenment Now: The Case for Reason, Science, Humanism, and Progress.

Polanyi-Levitt, K. (2012). The Power of Ideas: Keynes, Hayek, and Polanyi. International Journal of Political Economy, 41(4), 5-15. Retrieved May 5, 2021, from http://www.jstor.org/stable/23408607

Roser, Max (2019). Life Expectancy. Available at: https://ourworldindata.org/life-expectancy

Smith, Adam (2002). The Wealth of Nations . Oxford, England: Bibliomania.com Ltd.

Tensorflow (2021). Sequential model. Available at: https://www.tensorflow.org/guide/keras/sequential_model

United Nations Development Programme (UNDP), (2018). Human Development Indices and Indicators: 2018 Statistical Update. UNDP. Available at: http://hdr.undp.org/sites/default/files/2018_human_development_statistical_update.pdf

United Nations (UN), (2015). Millennium Development Goals (MDGs). Available at: https://www.un.org/millenniumgoals/

Varieties of Democracy (2021). Varieties of Democracy, Homepage. Available at: https://www.v-dem.net/en/

Indicators

DTP3 Diphtheria, tetanus, pertussis vaccination: percent of children aged 0 to 12 months, 1980–2016: DPT refers to a class of combination vaccines against three infectious diseases in humans: diphtheria, pertussis (whooping cough), and tetanus. This indicator reports data on those receiving the third dose of this vaccine. Source: Unicef

Population using improved drinking water sources percent, 1990–2015. Source: MDG

Homicide rate per 100,000, 1990–2018. Source: UN Office on Drugs and Crime. https://www.humanprogress.org/dataset/homicide/

CO2 emissions, per person metric tons, 1960–2014. Source: World Bank. https://www.humanprogress.org/dataset/co2-emissions-per-person-2/data-table/

Human Development Index, scale 0-1, 1990–2015. Source: U.N. https://www.humanprogress.org/dataset/human-development-index/data-table/

GDP per capita, 2018 U.S. dollars, 1950–2019. Source: The Conference Board. https://www.humanprogress.org/dataset/gdp-per-capita/

GDP, annual growth rate, percent, 1961–2016. Source: World Bank. https://www.humanprogress.org/dataset/gdp-annual-growth-rate/data-table/

Share of people who say they are very happy or quite happy, percent, 1981–2020. Source: World Values Survey. https://www.humanprogress.org/dataset/share-of-people-who-say-they-are-very-happy-or-quite-happy/data-table/

Mortality rate, children under 5, per 1,000 live births, 1960–2019. Source: World Bank. https://www.humanprogress.org/dataset/mortality-rate-children-under-5/data-table/

Access to electricity, percent of population, 1990–2018. Source: World Bank. https://www.humanprogress.org/dataset/access-to-electricity/data-table/

Economically active children, percent of children aged 7 to 14, 1994–2016. Source: World Bank. https://www.humanprogress.org/dataset/economically-active-children/data-table/

Life expectancy at birth, years, 1960–2018. Source: World Bank.
https://www.humanprogress.org/dataset/life-expectancy-at-birth/data-table/

Poverty headcount ratio at $1.90 a day, percent of population, 2011 international dollars, PPP, 1979–2016. Source: World Bank.
https://www.humanprogress.org/dataset/poverty-headcount-ratio-at-1-90-a-day/data-table/

Mean years of primary schooling number, 1870–2040. Source: Robert Barro, Jong-Wha Lee.
https://www.humanprogress.org/dataset/mean-years-of-primary-schooling/data-table/

BIOGRAPHY

Luis Ahumada Abrigo is an MS student in data science at The George Washington University. Luis earned a BS in political science from Diego Portales University. Luis's research interests include Development Economics, Political Communication, Political Philosophy, International Development, International Trade, and Welfare Policy.

Dr. Nima Zahadat is a professor of data science, information systems security, and digital forensics. His research focus is on studying the Internet of Things, data mining, information visualization, mobile security, security policy management, and memory forensics. He has been teaching since 2001 and has developed and taught over 100 topics. Dr. Zahadat has also been a consultant with the federal government agencies, the US Air Force, Navy, Marines, and the Coast Guard. He enjoys teaching, biking, reading, and writing.