Project Executive Summary

Christian Buonfiglio, Addison Gangwish, Aaron Janaszak, Lucas Mueller

---

Introduction

The project began with a rather ambiguous direction. As the four of us were slated to start our consulting with Medtronic, our general topic was 'healthcare'. An afternoon spent scouring the internet for usable, reliable, and interesting data guided us towards the final scope of our project: looking at global health metrics (such as accessibility to drinking water, percentage of underweight adults, and eventually many others), how are those correlated to the adult mortality rate? Moreover, if there is a correlation between these statistics, can we use machine learning capabilities to predict mortality rate from these disparate metrics? Can we forecast the metrics themselves in time? Our exploratory questions took the following forms:

- Trends
    - What are the general trends of these statistics over the years 2000 – 2020?
- Correlations
    - Are there any correlations between individual metrics that may or may not be intuitive, and can we explain why they might be connected?
- Comparisons
    - Can we use common features between countries, like population, to make comparisons between their statistics and draw conclusions about the effects those statistics have on the mortality rate?
- Predictions
    - Can we predict the mortality rate given the current data from a country?
    - Can we predict how the current data for a country will change?

Of course, these four categories were the most generic form of the questions, while our actual list more specifically detailed each nuance, and it was these questions that helped focus our research.

Research Presentation

Our initial round of data gathering started at the World Health Organization. Knowing that the scope of our project would be global, this resource seemed to be the best. Using the API features of this organization's data, we were able to extract

comprehensive data regarding percentage of a country's population having access to drinking water, percentage of underweight adults, and mortality rates. This data was given by country by year, i.e., "AFG 2015", "AFG 2016" and so on.

The countries in this data were listed with their ISO-3166 designation, a 3-letter code set by the International Organization for Standardization (ISO) to represent countries in places where the entire country's name written out would be impractical. While some of these designations are rather intuitive, it was helpful to have access to the name. To this end, we found a website (www.statisticstimes.com) that we were able to web-scrape and thus create a table that could relate an ISO-3166 code to that country's formal name. More importantly, this also gave us the country's geographic region, for instance, West Africa or Central America. This was useful for predicting missing values in some countries.

At this point, we really started shaping out the structure of the project. To ensure we would be able to run individual sets of data independently of each other, we created a unique producer/consumer for each dataset in Azure's DataBricks environment. Our consumers wrote their respective data to a .csv file in the datalake. A separate set of code was written to take those .csv files and write them to a SQL database, imputing missing values by the mean for their geographic region along the way. This process was automated using Azure's Data Factory.

When we started applying machine learning models to our data, we realized the results were erroneously dependent on a country's name instead of prioritizing the numerical data we were feeding it. After addressing that issue, we noticed that our models were not producing results that met our expectations of correlation coefficient 0.85 or higher.

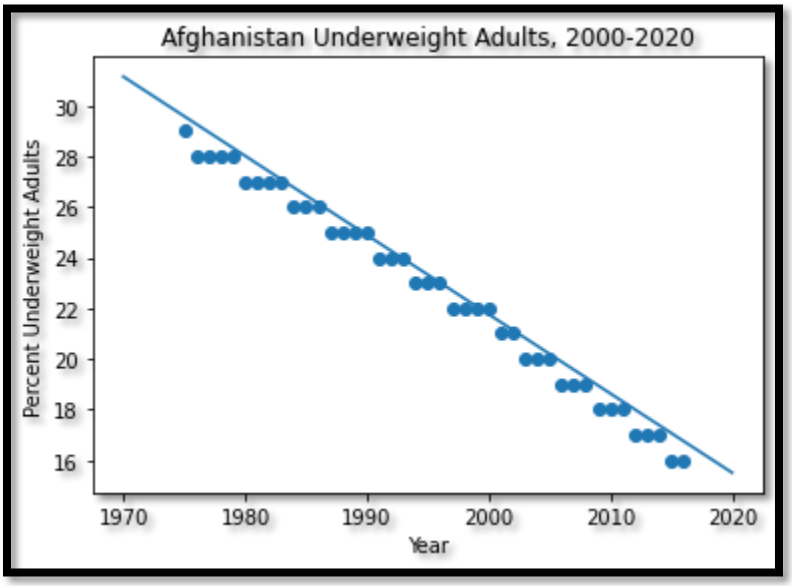
To improve our results, we decided to find more data and give the machine learning model more information. We added data for the prevalence of malaria, per capita population of doctors, and tuberculosis. Again, we created consumers and producers for each data set, ran them through our data factory, and applied the models to them. The number of doctors unexpectedly reduced the effectiveness of our model, so we ended up removing that set of data.

Finally, we decided we wanted to find possible trends in each of the indicator variables, such as malaria prevalence or water access, over time. We initially attempted to model this with a single linear model independent of country, however we found it was much more effective to perform a simple linear regression for each country's data, with an independent variable of year and dependent variable of the indicator in

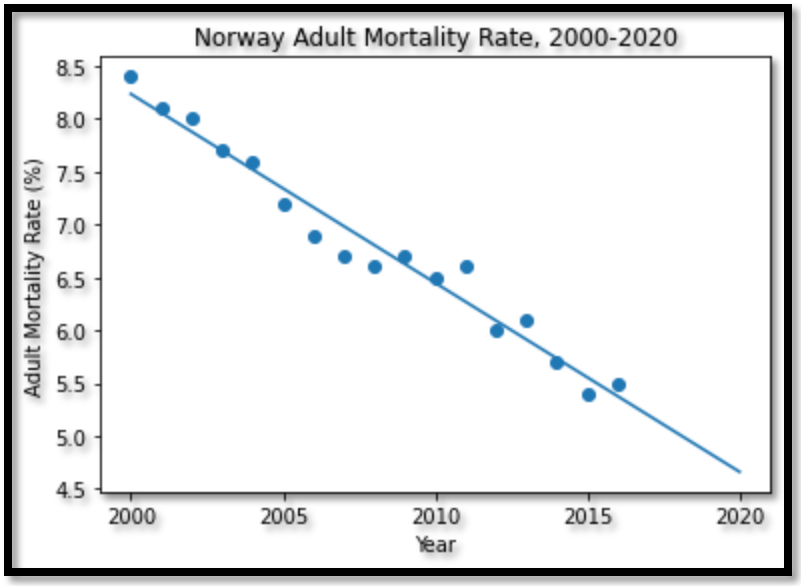
question. This gave us highly accurate (R-squared of 0.85 or higher) time forecasts for about 90% of our countries for each indicator.

Results

We found that in general, *all* indicator values had a consistent trend towards greater development. As an example, we take Afghanistan, which in 1975 had a 29% rate of underweight adults, which has been cut to 16% in 20. Similarly, 31.6% of adults didn't live to age 60 in Afghanistan in 2000, while in 2016 the numbers had dropped to 24.5%.



Nor is this unique to Afghanistan, we can see this trend even in "developed" countries, such as Norway. In 2000, 8.4% of adults didn't survive to age 60. But in 2016, 5.5% of Norwegian adults failed to live that long. In fact, 178 out of 183 countries in our survey had a reduction in mortality rate.

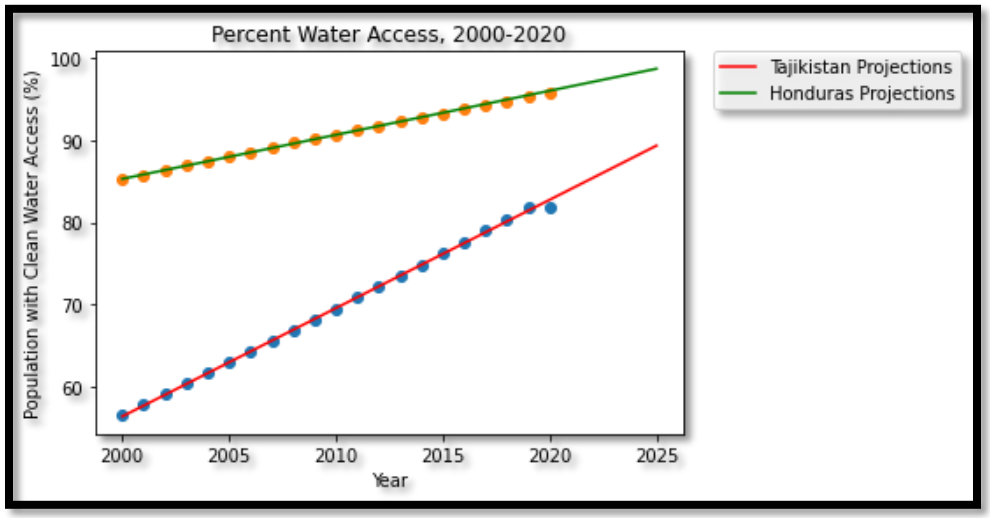


Likewise, in 9 out of 10 countries water access increased or remained level, and in 9 out of 10 countries with malaria endemic to the nation, the incidence rate of malaria decreased. These are extremely positive signs for human development across the world.

We also have found that certain regions (unsurprisingly) are much more likely to have access to safe drinking water, food, and are less likely to be infected with certain diseases. For instance, the regions with the greatest mortality rate include Southern (45%), Western (33%), Eastern (33%), and Middle Africa (30%), which each have an adult mortality rate of at least double Northern Africa (mortality rate 14%, comparable to that of East Asia at 12%). Likewise, Northern Europe, Southern Europe, and Northern America (including the United States and Canada but not Mexico) have a mortality rate all around 10%.

Our machine learning model, using *Pandas* and *scikit-learn*, had an R-squared value of 0.97 when comparing the training data to the test data. This model uses the extremely randomized trees method. We performed modeling both via the python *sklearn* module and in pyspark. Our pyspark machine learning models consisted of basic multilinear regression (r-squared 0.73), generalized linear regression (r-squared 0.73), basic decision tree (r-squared 0.87), and GBT regressor (r-squared 0.89). For *sklearn*, we have performed k-nearest-neighbors (r-squared 0.88), decision tree regression (r-squared 0.93), MLP regressor (r-squared 0.63), gaussian process regressor (r-squared –0.64), random forest (r-squared 0.95), gradient boosting regressor (r-squared 0.91), and radius-nearest-neighbors (r-squared NaN) tests, as well as the extra trees forest that we ultimately chose with its r-squared value of 0.97.

We also constructed a time series, performing a basic linear fit on each country in a given dataset (such as water availability or malaria rate), and then outputting predictions for the slope and intercept of the line of best fit. We obtained an r-squared value of 0.85 or higher for 100% of countries with the underweight adults, 92% of the countries for water access, 72% of the countries for mortality, 65% of the countries for malaria incidence, and 35% of the countries for Tb incidence. From this we can see that the diseases remain the hardest to predict, as should be intuitively obvious since disease breakouts unlike infrastructure are not nearly so linear.



Conclusions and Recommendations

Using our trained model in Power BI revealed some interesting exceptions. When we took our predicted values for the indicators in future years and fed them into the model, the mortality rate reacted unexpectedly in some cases. We had some theories as to why this is happening. We see the most unlikely model predictions in countries where the indicator values are at an extreme (either 0% or 100%). The model could be ignoring values if they are at either extreme, creating a false over-dependence on the remaining indicators. Japan is a good example. With its malaria and TB indicators at (or very very close) to zero percent and its access to water very close to 100 percent, the model could be fixating on Japan's percent of underweight adults (5.9%). If the model is ignoring the other indicators and fixating on this statistic, it might explain the aberrant behavior.

We found that in the vast majority of cases, countries are substantially improving on all the indicator variables we used in this project, and accordingly, mortality rates among

adults have decreased drastically. We also found projections that indicate further reductions in these variables and in mortality rates.

Our model's correlation coefficient of 0.97 is exceptionally good, meaning that there is not a huge margin for improvement. However, we expect that there are several other key indicators for a country's mortality rate, which could improve our model to an even higher coefficient. One of the indicators we had to forgo was the number of doctors per 1,000 people in the country, which simply didn't have enough data from the WHO, which was missing over half the years in our sample set.

If we were to add more indicators, we considered using HIV data (the 4th leading cause of death in Africa, which was the source of most of our outlier data points) but did not have time to incorporate it into the model. Finally, the largest killer worldwide, ischemic heart disease, failed to make it into our model since the WHO does not have any incidence values for it, only mortality figures, which of course would be correlated with mortality as a whole.

In general, one of the biggest causes of outliers in our dataset was African nations with very high mortality rates. This means that we haven't quite captured all the causes of death in those nations, since our model is still predicting lower figures than are truly present. Looking at HIV and diarrheal diseases, both leading causes of death there, would be key to corralling these outliers into our model.

Finally, our model did not use most of the WHO's breakout groups besides year. Future studies might benefit from taking a narrower look at sex or age groups with, for instance, underweight adults or malaria incidence, and generate dummy variables to improve the model further