Nirmal Balachundhar, Biraj Parekh, Nibir Pathak

**Machine Learning for Public Health**

Our task is to determine the most important economic and healthcare traits of a country that can help the country improve its public health infrastructure. While certain traits such as the health of the economy and hospital spending appear obvious, a deeper investigation is required to both reveal new characteristics that may be just as important and to further optimize public health. This can be achieved by learning the factors that are pivotal to different public health systems in different countries. We desire to learn about features in public health systems which may have gone undetected in previous qualitative research which would help a country allocate its investments in different sectors of public health to maximize the wellbeing of their population. These different factors can range over investment in medical research, types of vaccinations programs, specific focus on sub-populations, the number of health personnel per capita, literacy rates, and so on.

Our dataset consists of various attributes that are related to public health for almost every country. These attributes include literacy rates, number of doctors and nurses per one-thousand people, health spending as a percentage of the GDP, total health spending per capita, prepaid private spending per capita, out-of-pocket spending per capita, developmental assistance for health spending per capita, neonatal mortality rates, under-5 mortality rates, the universal health coverage index, mortality from air pollution and a lack of sanitation. We attempted to balance our attributes with certain ‘good’ qualities with ‘bad’ qualities and neutral financial indicators. For example, our ‘good’ attributes include the total amount of physicians and nurses per every thousand and the universal health coverage index which includes essential health services, tracer indicators, and service coverage inequalities. Some of the ‘bad’ features would include the neonatal, under-5, air pollution, and sanitation rates. In addition to financial indicators, we found that a combination of these attributes represented a holistic view of any given country’s public health system so that we weren’t leaving out potentially important features. Our output values that we will be testing with utilizes the Health Access and Quality Index which summarizes healthcare access and quality for a given location. This is a known scientific standard that accounts for different factors and standardizes healthcare access measures across different countries. In addition, it is commonly used in many research publications. We decided to research and add some of the latter features to improve the robustness of our models. Two major problems regarding the data that seem inherent to our domain of study included smaller datasets and minor accounts of seemingly inaccurate data. Our total dataset consisted of about two hundred examples to reflect every country for the most recent year where data was taken. That didn’t produce incoherent results, but we also wish that we had access to some more recent and accurate data. In addition, it seemed that some of the data from our different sources collected data from the countries themselves. Some countries had certain feature values which obviously were not true, but we couldn’t account for that since they hadn’t been transparent about their actual values.

We used multivariate linear regression models and feed-forward neural networks to analyze our data. We thought that both of these would be most appropriate for our task since they are advantageous when attempting to answer mysterious regression problems. Our linear regression model predicted HAQ index values with an R2 score of 0.9219. This suggested that a linear correlation is present between our input variables and the HAQ index. We found that the UHC Index had the biggest corresponding weight. This suggests that countries should aim toward meeting the UHC goals that are set out by decreasing inequality of healthcare and maximizing coverage over essential health services such as doctors’ services, inpatient and outpatient hospital care, prescription drug coverage, pregnancy and childbirth. All of these components suggest that a country should attempt to expand their national health insurance to improve their public health system. In addition, prepaid private spending per capita had one of the highest weights but had a negative effect on the efficacy of a public health system because individuals are more likely to shift to private companies for insurance rather than adopt government spending which not only increases inequalities in health care but also hurts the general pool of individuals who are covered by government insurance since they are more likely to be at risk. Sanitation and under-5 mortality were the next largest predictors for an HAQ index and they both suggest that it might be effective to redistribute resources within the public health system with an increased awareness on sanitation in public areas. Our neural network consists of two hidden layers with many nodes in each layer so that they could accurately learn certain relationships between pieces of data.

We think that there is a lot of potential for future work in this area. First, public health systems would benefit from a tailored machine learning task that evaluated the effect of prepaid private insurance on the inequality between access to healthcare for the rich and the poor. We think that it may be important to evaluate this difference because it could shed light into the difference in the quality of healthcare that both groups receive based on the specific amount of prepaid private health care. Second, we think that public health systems could benefit from understanding how they can better optimize their systems under given budget constraints. Our project provides insight regarding which sub-sectors they should investigate, but how funding is split and how coverage differs expands within each of those sectors to more sectors. In order to truly create an optimal system, one must further study the relationships between investments and allocation of resources in each of those categories.

All the members in our group decided that we wanted to best understand the entire project from a holistic standpoint, so we all contributed to the project an equal amount. This included initial research, data mining, model research and evaluation, and the actual project itself.