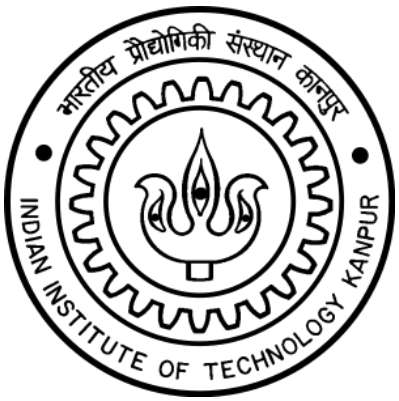
Department of Computer Science and Engineering
Indian institute of technology kanpur



CS685 Project Report

Relationship between Health and Education

GROUP: 13

Submitted by

Roll No	Name and Department
150305	Jaya Aditya C (MSE)
150323	Kashish Mittal (EE)
150658	Shankey Sahu (MSE)
150693	Shrey Agrawal (MTH)
150801	Vinay Kumar Agarwal (ME)

Under the guidance of
Mr. Arnab Bhattacharya

Contents

1 Introduction

In this project, We aim to examine, classify and summarize the data corresponding to the state of World's Children using fundamentals of data mining. We have focused towards identifying relations and trends among different variables along with identifying key features that determine important macro-variables and to establish a correlation between the health and education factors of a nation. Our aim was to use most of the methods we learned in our course CS685.

2 Dataset

We have analyzed datasets pertaining to health-care and education of children of all the countries from the United Nations database. It consists of both macro-variables and micro-variables related to health care and education. The data for health was spread across 69 files with 122 attributes. Similarly for education we had 23 files with 49 features in them. Following are the major features that the files incorporates:

1. Male-female Attendance in secondary education
2. Enrollment ratio of Male vs Female
3. Female-overall Literacy rate
4. Urban to Rural attendance ratio
5. Expenditure allocated towards health
6. Life expectancy
7. Ratio of male to female life expectancy
8. Fertility rate
9. Infant Mortality rate
10. Proportion Access to improved sanitation
11. Number of under five deaths
12. Proportion immunized against diseases
13. HIV AIDS- number of people and spread of awareness
14. Maternal mortality rate
15. Access to clean water source
16. Obesity and malnutrition
17. Vaccines sponsored by governments
18. Birth registration

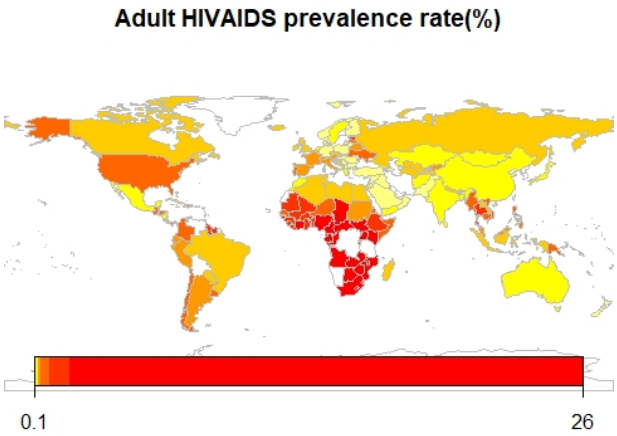


Figure 1: Adult HIV AIDS prevalence Rate

Figure 1 World heat map indicates the distribution rate for the Adult HIV Aids Prevalence. Indicating India on the very low side.

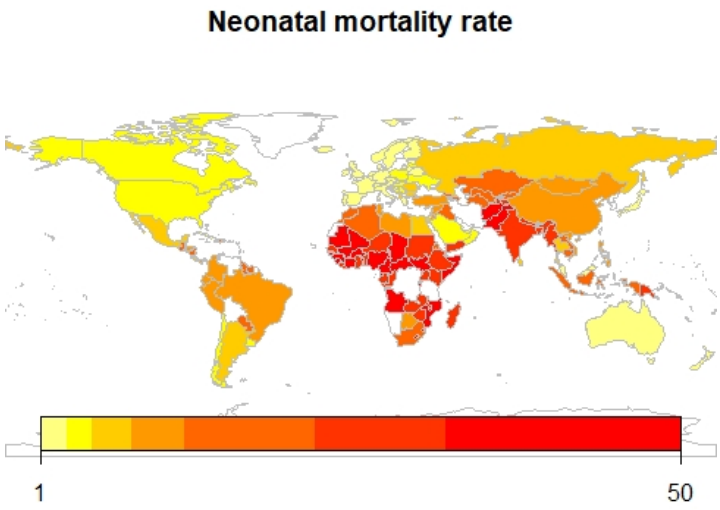


Figure 2: Neonatal Mortality Rate

Figure 2 heat map for neonatal Mortality rate indicates the high rate in India and parts of Africa. While North America having the very low rate

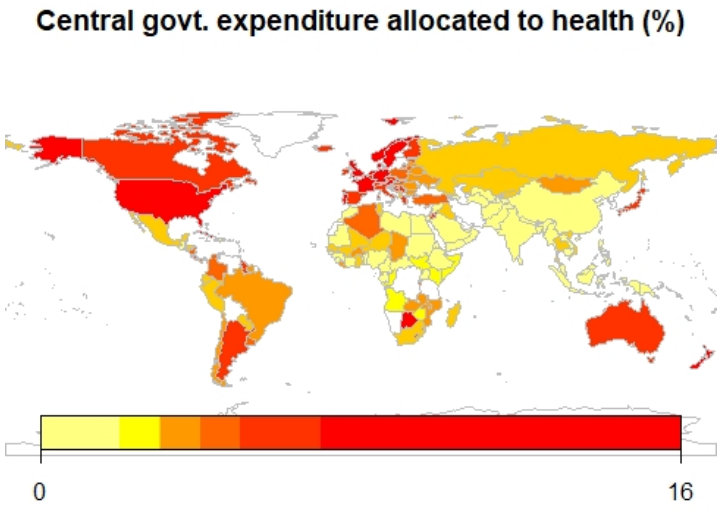


Figure 3: Central government expenditure allocated to health.

Figure 3 heat map shows the expenses allocated by central government of the respective country on their health, India being on the lower side while parts of America are at the higher side where government are more focused on health

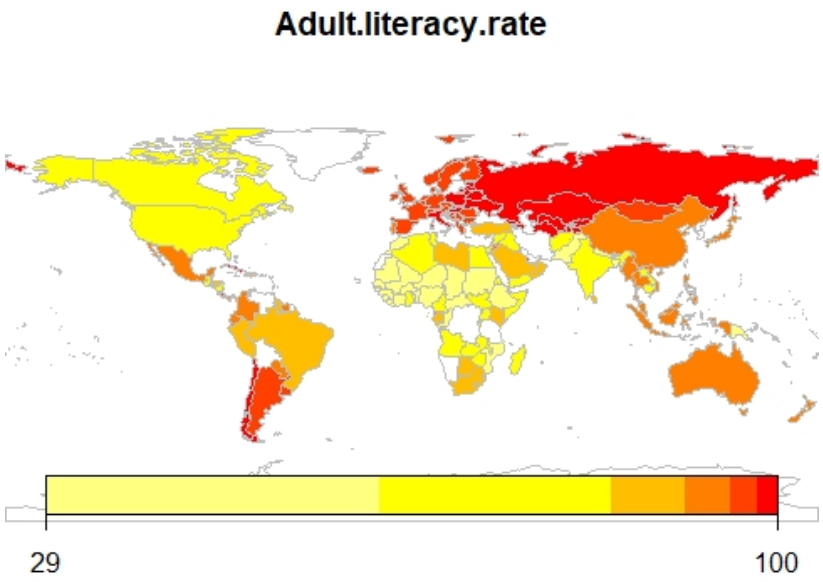


Figure 4: Adult literacy rate

.

This world heat map on adult literacy rate ranging till 100, highest can be observed in the Russia region while India being the average one.

Table 1: **Data Statistics**

Attribute	Europe	Oceania	Africa	Asia	Americas	World
Ratio of life expectancy at birth of females to males	108.73	105.84	104.29	106.32	107.74	106.0
Total fertility rate (TFR) 2011.0	1.57	3.43	4.38	2.53	2.36	2.4
Neonatal mortality rate	3.38	11.18	30.07	14.18	10.72	22.0
Annual number of under-five deaths	1.23	0.81	124.12	120.98	83.48	6914.0
Institutional deliveries	99.20	86.86	58.45	77.81	84.39	61.0
Antenatal care coverage for at least four visits	80.44	67.04	51.72	68.49	82.20	50.0
Life expectancy at birth 2011.0	77.44	71.18	57.38	71.42	73.70	69.0
Proportion of 1 year-old children fully immunised against Hepatitis B	91.27	89.81	81.18	88.14	90.98	75.0
People living with HIVAIDS Total	44.40	19.60	924.93	207.30	477.93	34000.0
Central government expenditure allocated to health'	6.11	7.52	2.94	2.66	4.18	6.0
Crude death rate (CDR) 2011.0	10.29	6.03	11.13	6.34	6.34	8.0
Crude birth rate (CBR) 2011.0	10.95	24.21	33.97	20.26	19.18	19.0
Annual number of births (Births)	203.93	40.84	1298.61	2812.44	1602.79	135056.0
Infant mortality rate (IMR) 2011.0	4.98	19.81	60.21	24.02	17.86	37.0
Prevalence of overweight (moderate and severe) - WHO	12.49	4.22	7.36	8.53	6.97	7.0
Proportion of population with sustainable access to an improved water source Total	99.35	90.09	70.93	88.81	91.42	89.0
Antenatal care coverage (ANC)	98.66	90.93	82.78	88.64	94.93	81.0
Reduction in under-five mortality rate 2000.0	38.65	25.12	28.21	37.4	29.86	29.0
Proportion of infants with low birth weight	6.10	11.5	13.35	11.21	9.50	15.0
Lifetime risk of maternal death	9122.64	1003.43	135.10	2005.02	866.53	180.0

2.1 Data Pre-Processing

The data had a lot of inconsistencies and we dealt with them in the following manner:

- Some files had three or four sub-features for each country so we had to flatten them.
- In some files the data included aggregated values for a defined group of countries, like South-East Africa, Pacific, Caribbean. So we created new entries in these files by swapping these regions with the countries in them.
- Some countries were missing from some files. So a Null entry for them was introduced to make them consistent. Finally all the files had 193 countries in them.
- Merged the files in two files, one for health the other for education.
- Handled the missing values of all countries by swapping them with the mean of the data for all the countries corresponding to its region. Like the data for Jamaica is missing, so we substituted it with the mean of data for all countries in Caribbean. In case of missing values in the region, missing values of the continent were used.
- Features with missing value threshold greater than 0.5 were dropped
- Applied max-min Normalization on them.

$$Normalized(x_i) = \frac{x_i - Min_i}{Max_i - Min_i}$$

2.2 Approach

- Our aim was to find the major factors contributing to health and education(feature selection) in the world. This was achieved by sorting the features on the basis of variance.
- After selecting the major features we went on to train Regression Models to predict the literacy rate and Infant Mortality Rate.
- Using the features from health and education trained a decision tree to predict if a country is developed or underdeveloped.
- See how features of health and education are correlated to each other. We used cosine similarity for that.

$$cs(\vec{v1}, \vec{v2}) = \frac{\vec{v1}.\vec{v2}}{\left\|\vec{v1}\right\|\left\|\vec{v2}\right\|}$$

- We Plot heat maps, scatter plots and drew inferences from them to correlate features with each other.

3 Analysis

3.1 Co-Relations

There were too many features in our data, and these many could have an issue later in the modelling if there was any correlation between the features. Following are some of the top correlation of features of health and education out of 5778 combinations possible.

Table 2: **Health**

Attribute 1	**Attribute 2**	**Corr.**
Total fertility rate (TFR) 2011.0	Under-five mortality rate (U5MR) Deaths per 1000 live births 2000.0 Total	0.85
Total fertility rate (TFR) 2011.0	Crude birth rate (CBR) 2011.0	0.98
Neonatal mortality rate	Proportion of population with access to improved sanitation Percent Total	-0.85
Neonatal mortality rate	Life expectancy at birth 2011.0	-0.9
Neonatal mortality rate	Under-five mortality rate (U5MR) Deaths per 1000 live births 2000.0 Total	0.94
Neonatal mortality rate	Under-five mortality rate (U5MR) Deaths per 1000 live births 2011.0 Female	0.96
Neonatal mortality rate	Under-five mortality rate (U5MR) Deaths per 1000 live births 2011.0 Male	0.96
Neonatal mortality rate	Under-five mortality rate (U5MR) Deaths per 1000 live births 2011.0 Total	0.96
Neonatal mortality rate	Under-five mortality rate (U5MR) Rank 2011.0 Total	-0.94
Neonatal mortality rate	Maternal mortality ratio (MMR)	0.86
Neonatal mortality rate	Infant mortality rate (IMR) 2011.0	0.98
Average annual rate of reduction of under-five mortality rate 2000-2011	Reduction in under-five mortality rate 2000.0	0.98
Proportion of population with access to improved sanitation Percent Total	Under-five mortality rate (U5MR) Deaths per 1000 live births 2000.0 Total	-0.85
Proportion of women 20-24 year-olds who gave birth before age 18	Adolescent birth rate	0.86
Population size Thousand Total	Annual number of births (Births)	0.94
Population size Thousand Total < 18 yr	Annual number of births (Births)	0.99
Population size Thousand Total < 5 yr	Annual number of under-five deaths	0.86
Population size Thousand Total < 5 yr	Annual number of births (Births)	1.0
Annual number of under-five deaths	Annual number of births (Births)	0.87
Prevalence of underweight (severe) - WHO	Prevalence of underweight (moderate and severe) - WHO Percent Total < 5 yr	0.93
Life expectancy at birth 2011.0	Under-five mortality rate (U5MR) Deaths per 1000 live births 2000.0 Total	-0.92
Life expectancy at birth 2011.0	Under-five mortality rate (U5MR) Deaths per 1000 live births 2011.0 Female	-0.89
Life expectancy at birth 2011.0	Under-five mortality rate (U5MR) Deaths per 1000 live births 2011.0 Male	-0.89
Life expectancy at birth 2011.0	Under-five mortality rate (U5MR) Deaths per 1000 live births 2011.0 Total	-0.89
Life expectancy at birth 2011.0	Under-five mortality rate (U5MR) Rank 2011.0 Total	0.89
Life expectancy at birth 2011.0	Maternal mortality ratio (MMR)	-0.86
Life expectancy at birth 2011.0	Crude birth rate (CBR) 2011.0	-0.86
Life expectancy at birth 2011.0	Infant mortality rate (IMR) 2011.0	-0.9
Proportion of 1 year-old children fully immunised against polio (OPV)	Proportion of 1 year-old children immunised against DPT1	0.89
Proportion of 1 year-old children fully immunised against polio (OPV)	Proportion of 1 year-old children immunised against measles	0.88
Proportion of 1 year-old children fully immunised against polio (OPV)	Proportion of 1 year-old children fully immunised against DPT	0.96
Under-five mortality rate (U5MR) Deaths per 1000 live births 2000.0 Total	Maternal mortality ratio (MMR)	0.88
Under-five mortality rate (U5MR) Deaths per 1000 live births 2000.0 Total	Crude birth rate (CBR) 2011.0	0.88
Under-five mortality rate (U5MR) Deaths per 1000 live births 2000.0 Total	Infant mortality rate (IMR) 2011.0	0.95
Under-five mortality rate (U5MR) Deaths per 1000 live births 2011.0 Female	Maternal mortality ratio (MMR)	0.89
Under-five mortality rate (U5MR) Deaths per 1000 live births 2011.0 Female	Infant mortality rate (IMR) 2011.0	0.99
Under-five mortality rate (U5MR) Deaths per 1000 live births 2011.0 Male	Maternal mortality ratio (MMR)	0.89
Under-five mortality rate (U5MR) Deaths per 1000 live births 2011.0 Male	Infant mortality rate (IMR) 2011.0	0.99
Under-five mortality rate (U5MR) Deaths per 1000 live births 2011.0 Total	Maternal mortality ratio (MMR)	0.89
Under-five mortality rate (U5MR) Deaths per 1000 live births 2011.0 Total	Infant mortality rate (IMR) 2011.0	0.99
Under-five mortality rate (U5MR) Rank 2011.0 Total	Crude birth rate (CBR) 2011.0	-0.86
Under-five mortality rate (U5MR) Rank 2011.0 Total	Infant mortality rate (IMR) 2011.0	-0.91
Maternal mortality ratio (MMR)	Infant mortality rate (IMR) 2011.0	0.87
Proportion of 1 year-old children immunised against DPT1	Proportion of 1 year-old children fully immunised against DPT	0.94
Proportion of 1 year-old children immunised against measles	Proportion of 1 year-old children fully immunised against DPT	0.89
Prevalence of underweight (moderate and severe) - WHO Percent Total < 5 yr	Prevalence of stunting (moderate and severe) - WHO	0.87

Table 3: **Education**

Attribute 1	Attribute 2	Corr.
Female 15-24 yr Literacy rate of 15-24 year-olds	Male 15-24 yr Literacy rate of 15-24 year-olds	0.96
Female 15-24 yr Literacy rate of 15-24 year-olds	Adult literacy rate	0.96
Female 15-24 yr Literacy rate of 15-24 year-olds	Ratio of adult literacy rate of females to males	0.91
Male 15-24 yr Literacy rate of 15-24 year-olds	Adult literacy rate	0.93
Urban Net attendance ratio in primary education (NAR)	Rural Net attendance ratio in primary education (NAR)	0.91
Female Net attendance ratio in primary education (NAR)	Male Net attendance ratio in primary education (NAR)	0.98
Poorest 20% Net attendance ratio in primary education (NAR)	Richest 20% Net attendance ratio in primary education (NAR)	0.87
Ratio of urban to rural Net attendance ratio in primary education (NAR)	Rural Net attendance ratio in primary education (NAR)	-0.86
Male Gross enrolment ratio in pre-primary education	Female Gross enrolment ratio in pre-primary education	1.0
Male Gross enrolment ratio in primary education	Female Gross enrolment ratio in primary education	0.91
Adult literacy rate	Ratio of adult literacy rate of females to males	0.92
Gross enrolment ratio in lower secondary education	Female Net enrolment ratio in secondary education	0.86
Poorest 20% Adult support for learning	Male Adult support for learning	0.96
Poorest 20% Adult support for learning	Total Adult support for learning	0.96
Poorest 20% Adult support for learning	Female Adult support for learning	0.96
Male Adult support for learning	Total Adult support for learning	1.0
Male Adult support for learning	Richest 20% Adult support for learning	0.93
Male Adult support for learning	Female Adult support for learning	0.99
Total Adult support for learning	Richest 20% Adult support for learning	0.94
Total Adult support for learning	Female Adult support for learning	1.0
Richest 20% Adult support for learning	Female Adult support for learning	0.94
Male Attendance in early childhood education	Female Attendance in early childhood education	1.0
Male Net enrolment ratio in secondary education	Female Net enrolment ratio in secondary education	0.96
Female Net enrolment ratio in secondary education	Gross enrolment ratio in upper secondary education	0.86
Male Net enrolment ratio in primary education (NER)	Female Net enrolment ratio in primary education (NER)	0.96

To complete our Exploratory data analysis, we started looking looking at features among the health and education datasets with co-relations > 0.8 and shortlisted a few co-relations between features which were non-trivial. All these correlations can be interpreted and certain insights can be derived for countries which have compromised health and educational facilities.

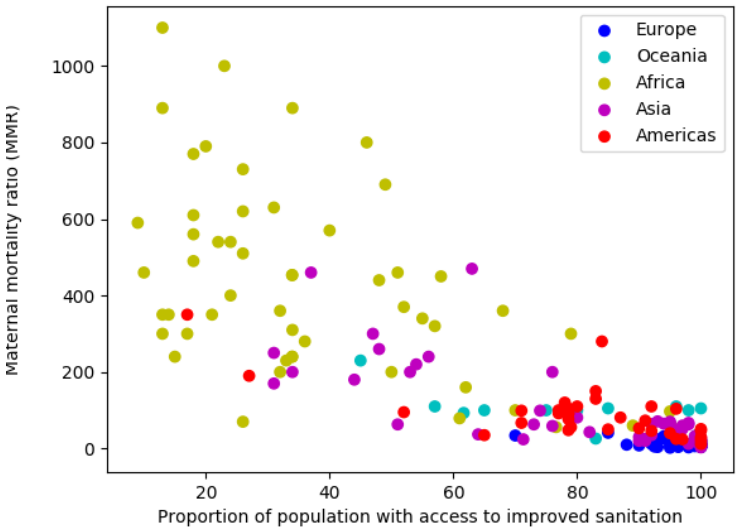


Figure 5: Proportion of population with access to improved sanitation vs Maternal mortality ratio (MMR)

Fig. 5 highlights an evident relationship between Maternal Mortality rate and Proportion of population with access to improved sanitation. Multiple studies have highlighted the importance of hygiene and improved sanitation for woman's health. Especially for pregnant women, unhygienic sanitation can be a root cause of many diseases and pose serious threats to their life. Also it can be inferred from the plot that countries belonging to Africa have the highest mortality rate and lowest proportion of population having access to improved sanitation. Countries

from Americas and Europe perform substantially well in both these metrics.

As shown in Fig. 6, there is a strong correlation between the Under five mortality rate and Ma-

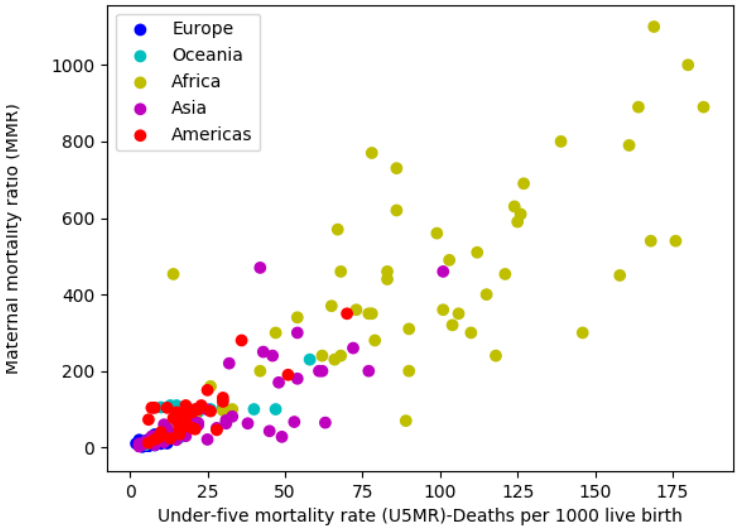


Figure 6: Under-five mortality rate (U5MR) Deaths per 1000 vs Maternal mortality ratio (MMR)

ternal mortality ratio. We believe that this highlights the lack of adequate health-care facilities in that country. Again here, the countries belonging to Africa are adversely placed and there is an urgent need to improve the facilities of health-care in these countries.

Underweight refers to having substantially low weight for given age of child and Stunting refers

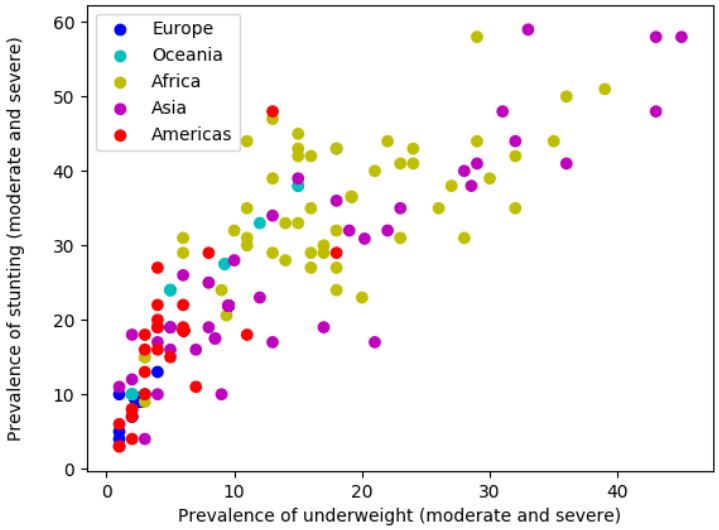


Figure 7: Prevalence of underweight (moderate and severe) vs Prevalence of stunting (moderate and severe)

to having substantially low height for the given age of child. The strong correlation shown in Fig. 7 between the these two features is somewhat intuitive as they both indicate nutritional deprivation and lack of hygiene and sanitation. Stunting may be caused due to birth defects. Food shortage and poverty levels could be major reasons behind alarmingly high rates in some Asian and African countries. Effective governmental schemes need to be implemented to counter

these problems.

A strong negative correlation between female literacy rate and Under five mortality rate can be

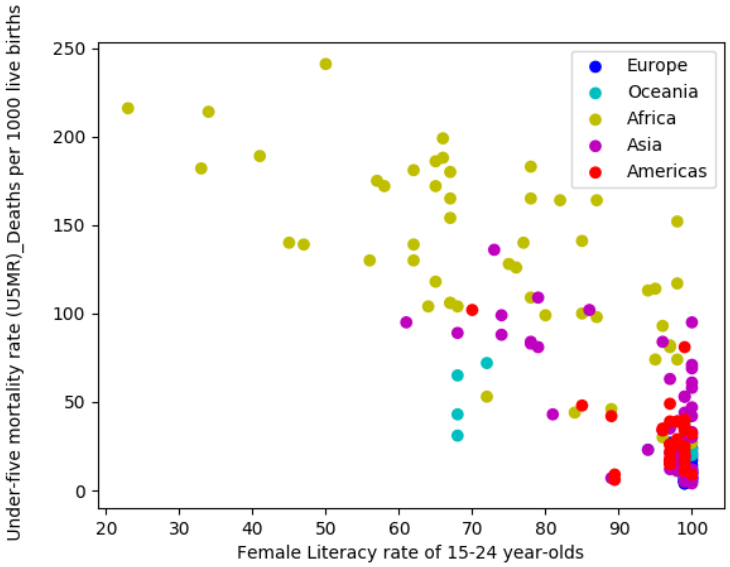


Figure 8: Literacy rate of 15-24 year-old vs Under-five mortality rate per 1000 live births

seen in Fig. 8. Literate individuals are more aware about the health hazards, birth complications and nutritional requirements of the new-borns. They inherently make required efforts to ensure hygiene, nutrition and prioritize the new-born's health. Regions corresponding to Americas and Europe boast of high literacy rates and low mortality rates at the same time.

Under-weight is a condition which is often a consequence of malnutrition and inadequate heath-

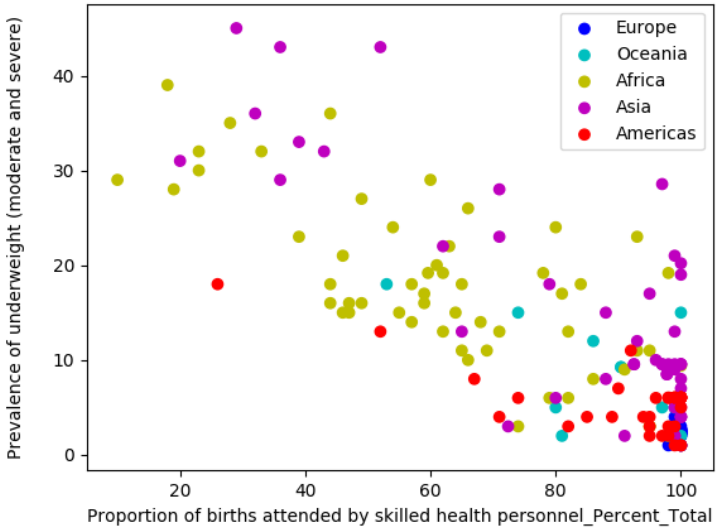


Figure 9: Proportion of births attended by skilled health personnel vs Prevalence of underweight

care and hygiene. Deliveries carried out in institutions or in presence of skilled personnel are more carefully performed and generally proper instructions are given to the households regarding nutrition and hygiene of the child. Thus guidance from health professionals can help prevent certain adverse health conditions among children. The Fig. 9 shows the strong inverse relation between Prevalence of Underweight and Proportion of births attended by skilled health professionals.

As shown in Fig. 10, countries belonging to Asia and Europe have witnessed substantial reduc-

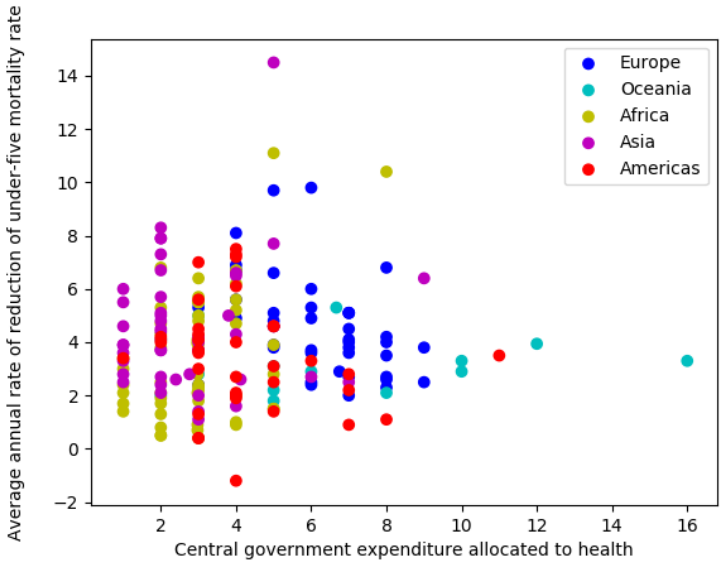


Figure 10: Central government expenditure allocated to health vs Average annual rate of reduction of under-five mortality rate

tion in infant mortality rate. The reduction in Europe can be attributed to increased expenditure by central government towards health-care. However in the case of American countries even after inflow of funds from the central government, reduction has not been aspiring. Fig. 11 shows the

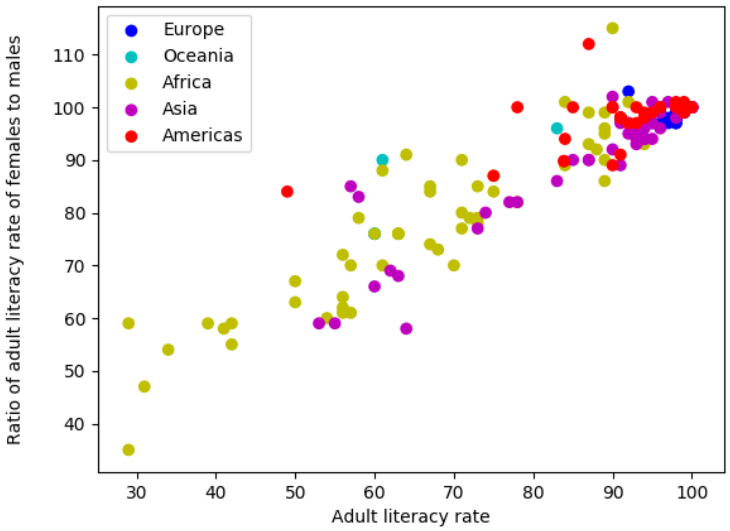


Figure 11: Ratio of adult literacy rate of females to males vs Adult literacy rate

relationship between Ratio of adult literacy rate of female to males and Overall adult literacy rate. As it can be seen, there is a strong positive correlation between the two. This shows that more educated societies have far more educated women than less educated societies. In other words ratio of of women being denied education is more in less educated countries.

3.2 Classification

To test the reliability and completeness of the date, We trained a Decision tree model that classified the countries as developed or developed using the values of attributes from health and education datasets. The World Bank, International Monetary Fund (IMF) and other international agencies consider developed and developing countries in terms of Gross National Income per capita. We decided to classify them on the basis of health and education indexes only and see how accurately it performed.

We divided the data set in two part, training and validation set in the ratio of 70 to 30. We got an accuracy of 94%. Following are the results:

Country	**Actual**	**Predicted**
Botswana	Developing	Developing
South Sudan	Developing	Developing
Democratic People's Republic of Korea	Developing	Developing
Indonesia	Developing	Developing
Brunei Darussalam	Developing	Developing
Kyrgyzstan	Developing	Developing
Costa Rica	Developing	Developing
Zambia	Developing	Developing
New Zealand	Developed	Developing
Libya	Developing	Developing
Suriname	Developing	Developing
Kenya	Developing	Developing
Panama	Developing	Developing
Montenegro	Developed	Developed
Ukraine	Developed	Developed
Ghana	Developing	Developing
Estonia	Developed	Developed
Slovenia	Developed	Developed
Finland	Developed	Developed
Turkmenistan	Developing	Developing
Mauritania	Developing	Developing
Iraq	Developing	Developing
Jordan	Developing	Developing
Poland	Developed	Developed
Honduras	Developing	Developing
Liechtenstein	Developed	Developed
Algeria	Developing	Developing
Tonga	Developing	Developing
Saudi Arabia	Developing	Developing
Swaziland	Developing	Developing
Myanmar	Developing	Developing
Morocco	Developing	Developing
Gambia	Developing	Developing
Cape Verde	Developing	Developing
Senegal	Developing	Developing
Central African Republic	Developing	Developing
Bangladesh	Developing	Developing

Mauritius	Developing	Developing
Serbia	Developed	Developed
Slovakia	Developed	Developed
Israel	Developed	Developing
Benin	Developing	Developing
Monaco	Developed	Developed
United Kingdom	Developed	Developed
San Marino	Developed	Developed
The former Yugoslav Republic of Macedonia	Developed	Developing
Japan	Developed	Developed
Colombia	Developing	Developing

The variables that had heavy importance in the decision tree included Female net attendance ratio, under 5 mortality rate and prevalence of obesity.The decision nodes for the classifier are as follows:

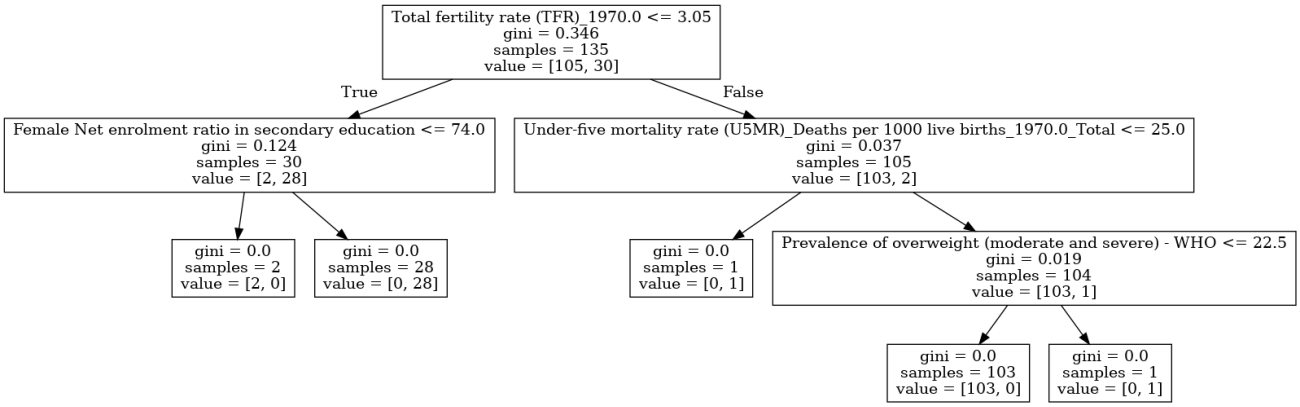


Figure 12: Decision Nodes of the classifier

3.3 Regression

Next We trained multiple Linear Regression models with attributes from both the Education datasets as independent features and some prominent health attributes as the dependent variables. This was done to see if Health Macro-Variables show dependence on other variables. Ten independent variables with the highest correlation with the dependent variable were used to train the model. Most of variables showed spectacular fits and educational features could be employed to explain the variance of the health macro-variables. Following are the models and their R^2 values.

Feature	$R^2 Value$
Neonatal mortality rate	0.864
Proportion of population with access to sanitation	0.837
Institutional deliveries	0.756
Life expectancy at birth	0.856
Adolescent birth rate	0.789
Under-five mortality rate	0.887
Maternal mortality ratio	0.803
Crude birth rate	0.835
Antenatal care coverage (ANC)	0.776

We can see that R^2 value is fairly high for all the models. This supports the claim is education is fairly important aspect of how a society fairs at child and maternal health. If we take a look at the correlation between health features and literacy rate, we can see that it is significant for most of them. This supports our claim made on the basis of regression modelling.

3.4 Clustering

Next we wanted to see similarities among different countries with respect to different attributes pertaining to health-care and education. Hence we applied clustering to classify all the countries into groups based on the education and health and then plotted them on a heat map. Following are some of the inferences:

- We can clearly see from the heat map the countries in the same geographical are clubbed together.
- The whole Latin America and Mexico are in the same cluster, which should be if we consider the socio-economic factors in that region and the history of this region.
- The United States is in the same cluster as that of most of Europe, which is what should be expected as most of countries in this region are developed as per World Bank.
- The Sub-tropical region of Africa and Asia are in the same cluster.

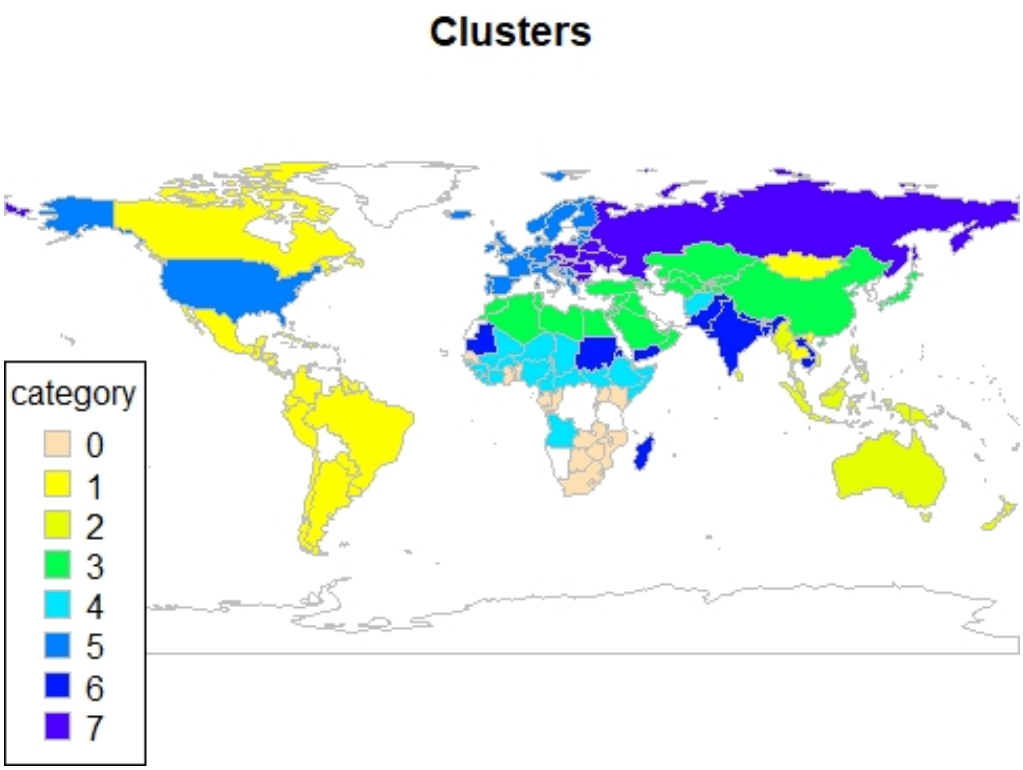


Figure 13: Heatmap of countries on the basis of clustering

Table 6: **Education vs Health**

Attribute 1	**Attribute 2**	**Corr.**
Proportion of births attended by skilled health personnel	Adult literacy rate	0.986
Proportion of 1 year-old children immunised against TB (BCG)	Proportion of pupils starting grade 1 who reach grade 5	0.984
Proportion of 1 year-old children fully immunised against DPT	Male 15-24 yr Literacy rate of 15-24 year-olds	0.984
Antenatal care coverage (ANC)	Adult literacy rate	0.984
Proportion of new borns protected against tetanus	Adult Literacy rate of 15-24 year-olds	0.983
Proportion of 1 year-old children fully immunised against Hepatitis B	Male 15-24 yr Literacy rate of 15-24 year-olds	0.982
of 1 year-old children fully immunised against polio (OPV)	Adult Literacy rate of 15-24 year-olds	0.982
Proportion of births attended by skilled health personnel	Richest 20% Adult support for learning	0.982
Proportion of 1 year-old children immunized against measles	Adult literacy rate	0.981
Antenatal care coverage (ANC)	Gross enrolment ratio of girls to boys in primary education	0.98

3.5 Prediction of Male and Female Literacy Rate

We aimed to predict the Literacy rate of the Male population and the Literacy rate of the Female population using the education attributes. We trained a multi layer perceptron on the top featured related to male population and female population respectively. Following are the results that we got:

Feature	$R^2 Value$
Male Literacy Rate rate	0.67
Femalr Literacy Rate	0.67

3.6 Comparison of Male and Female mortality rate

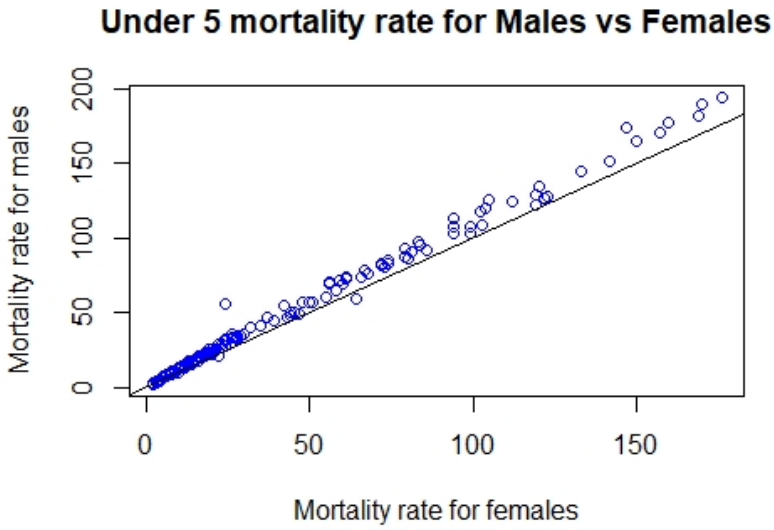


Figure 14: Male vs Female mortality rate

Fig. 14 shows an interesting observation, the under five mortality rate for males is consistently higher than that for females. This counter-intuitive fact is true for 191 out of 193 countries. As the observation is being made on under 5 mortality rate, social or cultural factors cannot be attributed as the cause. Hence, this strongly suggests that females have a biological advantage over males. Indeed this has been scientifically verified, the second X chromosome can compensate if one of them undergoes mutation. Women also have a better resistance to biological aging while estrogen is credited to facilitating elimination of bad cholesterol thus preventing heart diseases.

3.7 Relationship between Adult literacy rate and Prevalence of HIV AIDS

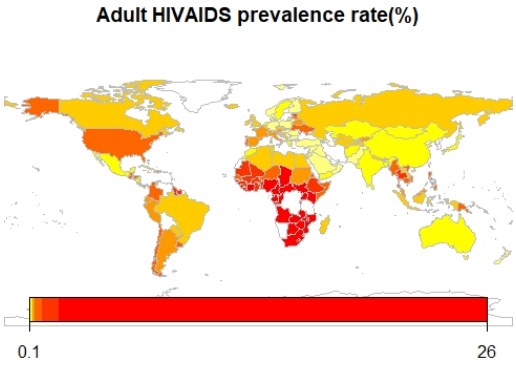


Figure 15: Adult HIV AIDS prevalance

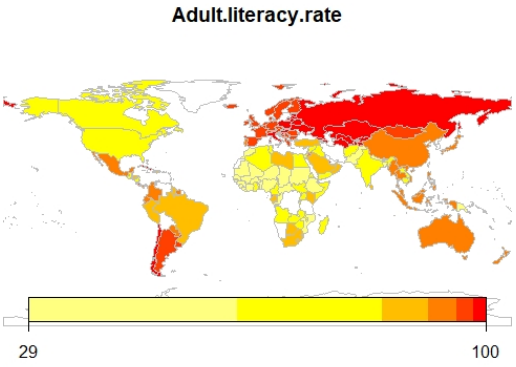


Figure 16: Adult Literacy Rate

On observing Fig. 15 and Fig. 16 , we saw a stark contrast between the two plots. The prevalence HIV AIDS is generally very high in the countries having low literacy rate, for example most of Africa. Upon further research, it was found that low literacy rate is indeed credited for high HIV AIDS cases and many organizations including UNESCO are targeting literacy to reduce the spread of HIV AIDS. As illiterate women and men have no access to written information, they remain unaware of many national and international issues affecting them that are increasingly being communicated through printed materials. The less they know, the more vulnerable they become, Because they do not have the skills to read and are therefore unable to know what HIV is and how it is spread, they are unable to protect themselves. In some cases, they may also be given incorrect information verbally but are in no position to verify this with reliable printed information.

3.8 Maternal mortality rate and Institutional Deliveries

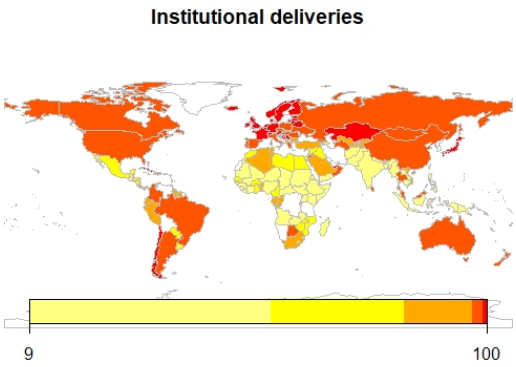


Figure 17: Institutional Deliveries

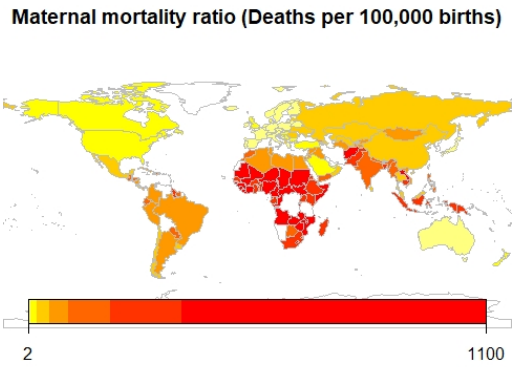


Figure 18: Maternal Mortality Ratio

Skilled birth attendance and institutional delivery have been advocated for reducing maternal mortality. The above plots are yet another proof that giving birth in a medical institution in the supervision of trained professionals promotes child survival and reduces the risk to mother's health.
There have been several governments who have taken this matter seriously and went ahead to implement incentive schemes that promoted institutional deliveries which have successfully fulfilled their purpose. For example Janani Suraksha Yojana (JSY) by Indian Government, which

increased institutional deliveries by 42.6 percent and sharply reduced the maternal mortality rate.

3.9 Birth registration and Vaccinations

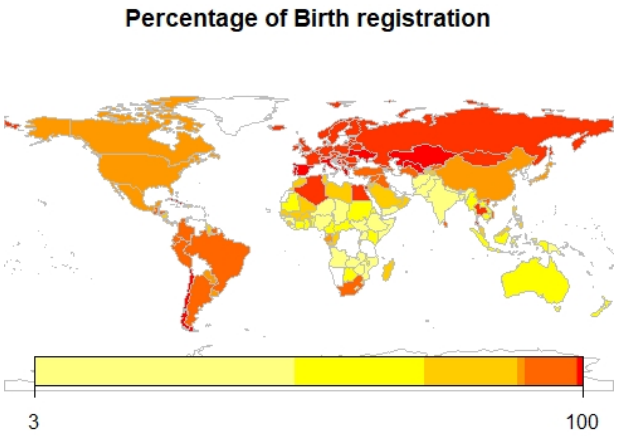


Figure 19: Percentage of Birth registration

Figure 19 heat map indicates the distribution rate for Percentage of registration of birth as they are born, which indirectly can be related as the percentage of new borns being vaccinated since assuming they are being registered. As can be observed from the map Russians are very active with respective to registrations, while India and africa being on the lower side

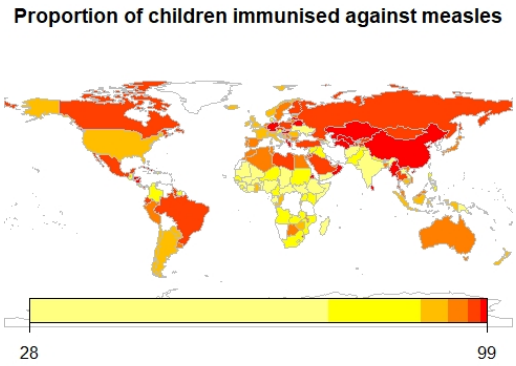


Figure 20: Proportion of children immunized against measles

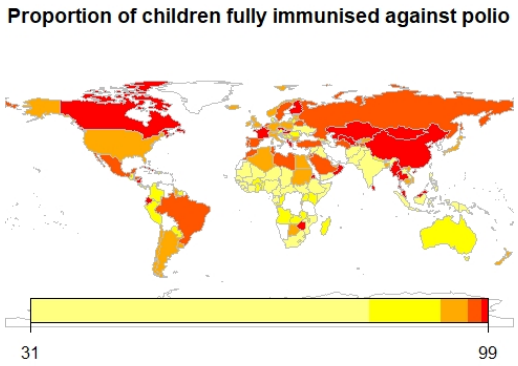


Figure 21: Proportion of children fully immunised against polio

As it can be seen in the figures 19, 20 and 21, Percentage of Children immunized against Polio and Measles follow a similar trend to that of percentage of Births registered across the world. It is often emphasized and is also evident from the above plots that extent of Birth registration plays an integral role in the success of health-care policies and vaccination schemes. In many developing countries today, the births of a substantial share of children go unregistered. Unregistered children are often unable to gain access to health care services or pay more for those services than a registered child. But a birth certificate means a child can access medical treatment and the vaccinations he or she needs to stay healthy.

3.10 Effectiveness of vaccines for Child Mortality rate

In order to study the effectiveness of studying various vaccines in affecting the Child Mortality, we chose to train a Decision tree between Under 5 Mortality Rate and Propotions zof Children under 1 Years immunised against various diseases. We later determined the feature importances of the various vaccines, to determine efficacy of vaccines in reducing mortality. This process was initialy done to regionwise and applied to all countries. The rationale behind applying this procedure is because some diseases in might be more prevalent in some regions than others. The results of this analysis are follows:

Attribute	DPT1	Measles	DPT	HiB	TB (BCG)	Hepatitis B	Polio (OPV)
Europe	0.07	0.02	0.08	0.28	0.05	0.14	0.36
Oceania	0.23	0.01	0.14	0.10	0.13	0.05	0.34
Africa	0.49	0.17	0.04	0.05	0.06	0.06	0.13
Asia	0.14	0.04	0.15	0.12	0.07	0.18	0.29
Americas	0.22	0.12	0.15	0.20	0.15	0.03	0.13
World	0.29	0.34	0.03	0.07	0.08	0.06	0.13

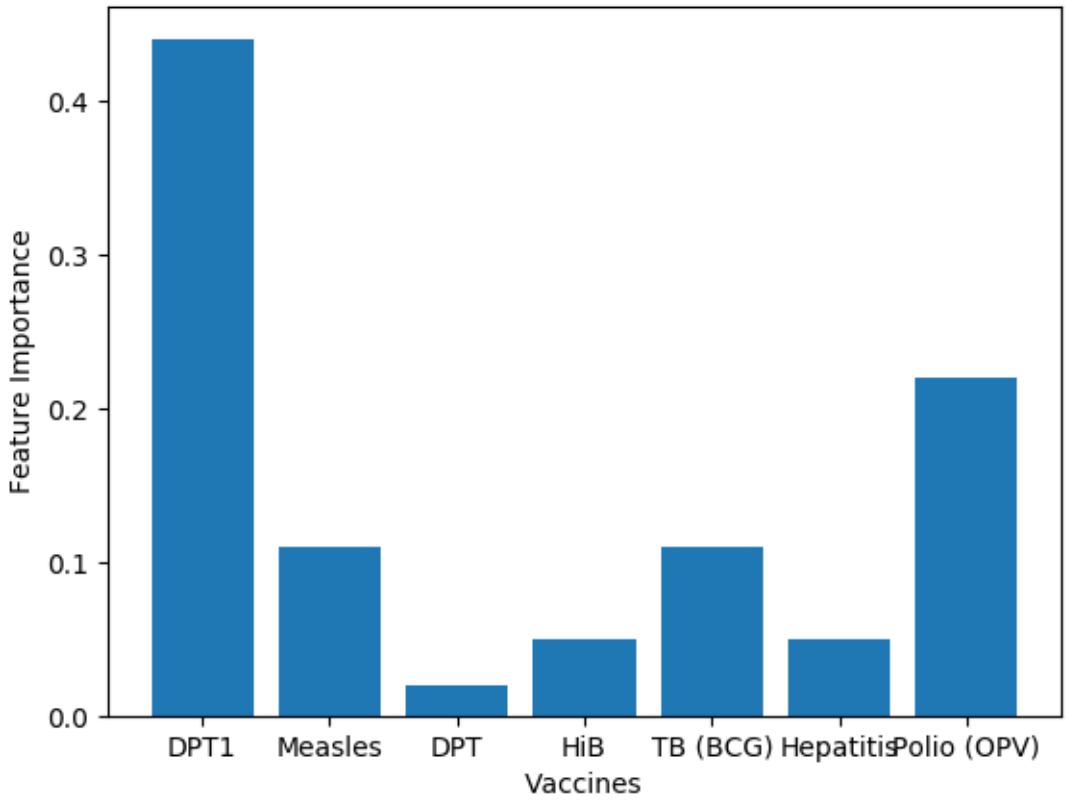


Figure 22: Bar plot of importances of all vaccines

4 Limitations

As the collected data was very inconsistent and very useful data were missing for different countries, also different files have data for different countries which was hard to get the common intersection to get valuable insights. There were too many variables to work on out of which many were correlated and hard to ignore. While merging the data many different files were having different spellings of the same country. Other than that working on such a big amount of data in the limited time period was hard for our group of 5 members to collect more insights which were possible from the data set to collect.

5 Conclusions

In this project, we have attempted to make a comprehensive analysis of various health and educational features and predicted a few macro indicators based on these values. Co-relations between various non trivial features were studied and visualizations were made to draw inferences from the co-relations. Co-relations were later utilised for feature selection to predict macro indices such as development classification, Mortality rates, etc. We were able to obtain good accuracy scores for all models that we trained. By Clustering on the countries on all features, we were able to get clusters of countries that have a similar development index.

In conclusion, analysis of health and education have provided us fascinating insights into this avenue. These seem to the trivially corrleated and it has been estabilished that education plays an important role in how the newborns of a country are at their health.

6 Future Scope

In our project we mainly focused on the age group 0-5. This study can be extended to other age groups as well. We have shown the relationship between AIDS and education. It can be similarly done with other diseases.

We took data set only from the 21^{st} century. United Nations archives contain data from many more years. These data can be used to find trends in the data. It has been established that education and health are correlated this knowledge can be used to a much greater extent.

7 Acknowledgement

The completion of this project required guidance throughout, without which the idea would have remained far from success. We are privileged to have received all of it from our instructor Dr. Arnab Bhattacharya. His lectures have helped immensely in gaining the required background knowledge. He was always willing to entertain our doubts and help us in ways he can. We owe our deepest gratitude to him and would like to thank him for all the support.

References

[1] Bhattacharya, A. (2018) CS685: Data Mining, class notes [PDF slides]. Retrieved from `https://web.cse.iitk.ac.in/users/ cs685/lectures.php`

[2] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[3] Jones E, Oliphant E, Peterson P, et al. SciPy: Open Source Scientific Tools for Python, 2001-, `http://www.scipy.org/` [Online; accessed 2018-11-16].

[4] United Nations Database, 2017 `http://data.un.org`

[5] Travis E, Oliphant. A guide to NumPy, USA: Trelgol Publishing, (2006).

[6] Wes McKinney. Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, 51-56 (2010) (publisher link)