HEALTH EXPENDITURE DATA

Child mortality is a very worrying demographic phenomenon especially in developing countries, which has attracted the attention of various researchers and policymakers. Today, combating this issue is considered an important objective, therefore many international organizations, such as the United Nations Children's Fund (UNICEF), the World Bank and the World Health Organization (WHO) have incorporated the objective of reducing child mortality into most of their future programs.

Healthcare financing, whether through private or public means, remains fundamental for the improvement of children's health status all over the world. So, one of the factors having a significant impact on the Global Child Mortality Rate is the Health Expenditures done by the nations across the world.

The role of health economics today is crucial because of growing international awareness of the close relationship between economic development and health. Furthermore, as health in childhood is one of the key predictors of health and productivity in later life, child mortality is an important indicator of socioeconomic development. In the regions, where health infrastructure is largely underdeveloped, increasing health expenditure will contribute to progress towards reducing the child mortality rates. Therefore, in order to achieve the same, the governments in those regions need to increase amounts allocated to health care service delivery.

Globally, the health spending is highly unequal. Hence, the main objective of this effort is to assess the impact of health expenditure on child mortality, as measured by under-five child mortality rates in countries with 50 highest child mortality rates for the period of 2000 to 2015. In addition, we also aim to predict the same using supervised machine learning models.

The data source for this study is the World Health Organization's Global Health Expenditure Database (GHED). This data in the GHED is collected by the WHO and is availed publicly by the World Bank Group according to the open data standards and licenses datasets under the Creative Commons Attribution 4.0 International license (CC-BY 4.0). They are labeled accordingly, and when they are accessed by users, users agree to comply with all of the terms of the respective licenses. GHED is the source of the health expenditure data republished by the World Bank and the WHO Global Health Observatory. The World Bank Group also quotes that in some cases it is not possible to make data available, either because the data are too sensitive, or have been lost or damaged. However, users may still benefit from the available metadata for these datasets.

DATASET DESCRIPTION

This data helps ensure health services are available and affordable when people need them. In particular, the data published here contribute to a better understanding of:

- How much do different countries spend on health?
- What are the financing arrangements to pay for health?

WHO works collaboratively with Member States and updates the database annually using available data such as health accounts studies and government expenditure records and where necessary, modifications and estimates are made to ensure the comprehensiveness and consistency of the data across countries and years.

This dataset comprises of Health Expenditure per Capita data for around 190 countries in the world for the time period of 2000 to 2018. This data is collected in local currencies for different nations and then is converted to US$ as per applicable and acceptable norms and conditions. The amount estimates of the current health expenditures include healthcare goods and services consumed during each year by both private and public organizations.

	Country Name	Country Code	Indicator Name	Indicator Code	2000	2001	2002	2003	2004	2005	...	2011	
0	Aruba	ABW	Current health expenditure per capita (current...	SH.XPD.CHEX.PC.CD	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	
1	Afghanistan	AFG	Current health expenditure per capita (current...	SH.XPD.CHEX.PC.CD	15.803164	15.803164	15.803164	17.035744	20.412764	23.890501	...	50.853474	5
2	Angola	AGO	Current health expenditure per capita (current...	SH.XPD.CHEX.PC.CD	12.998967	28.918121	29.049364	34.875187	49.810741	54.260777	...	122.107231	12

Screenshot of the Health Expenditure Dataset

	Location	Dim2	Indicator	Period	Dim1	Tooltip	IndicatorCode	FactValueForMeasure
0	Afghanistan	HIV/AIDS	Number of deaths	2017	0-27 days	0.17	MORT_100	0.17
1	Afghanistan	HIV/AIDS	Number of deaths	2017	1-59 months	16.83	MORT_100	16.83
2	Afghanistan	HIV/AIDS	Number of deaths	2017	0-4 years	17.00	MORT_100	17.00

Screenshot of the Child Mortality Dataset

As the time period information is on different axis in both the datasets i.e., on available in rows in the Health Expenditure dataset and available in a column in the Child Mortality dataset, the Health Expenditure dataset is melted & transformed using suitable python libraries in order to make both the datasets similar and compatible.

Under the hood, multiple data pre-processing techniques are applied for exploratory data analysis and then this dataset is fused with the Child Mortality data to find its one-to-one correlation with it by applying suitable machine learning models to predict the child mortality rate for the all the countries for future years, given the estimated values of the Health Expenditures per Capita for that time period.

	Location	Year	Deaths	HCE
0	Afghanistan	2000	126578.92	15.803164
1	Afghanistan	2001	131386.92	15.803164
2	Afghanistan	2002	124658.00	15.803164
3	Afghanistan	2003	117839.10	17.035744
4	Afghanistan	2004	117196.50	20.412764

Screenshot of the Dataset Formed After Fusing HI & CM Datasets

On applying the Decision Tree Regressor & AdaBoost Regressor ML models separately for different nations, we were able to predict the Child Mortality Rate for different countries. We got the following stats after applying the models on a dataset containing info about only 'Afghanistan':

```
y_pred = dt_regressor.predict(X_test)

# Results of Decision Tree Regressor
print("Mean absolute error =", round(sm.mean_absolute_error(y_test, y_pred), 2))
print("Mean squared error =", round(sm.mean_squared_error(y_test, y_pred), 2))
print("Median absolute error =", round(sm.median_absolute_error(y_test, y_pred), 2))
print("Explain variance score =", round(sm.explained_variance_score(y_test, y_pred), 2))
print("R2 score =", round(sm.r2_score(y_test, y_pred), 2))
```

```
Mean absolute error = 4497.7
Mean squared error = 23601575.42
Median absolute error = 4411.3
Explain variance score = 0.98
R2 score = 0.88
```

```
y_pred2 = ab_regressor.predict(X_test)

# Results of AdaBoost Regressor
print("Mean absolute error =", round(sm.mean_absolute_error(y_test, y_pred2), 2))
print("Mean squared error =", round(sm.mean_squared_error(y_test, y_pred2), 2))
print("Median absolute error =", round(sm.median_absolute_error(y_test, y_pred2), 2))
print("Explain variance score =", round(sm.explained_variance_score(y_test, y_pred2), 2))
print("R2 score =", round(sm.r2_score(y_test, y_pred2), 2))
```

```
Mean absolute error = 3952.7
Mean squared error = 21417215.42
Median absolute error = 4038.95
Explain variance score = 0.97
R2 score = 0.89
```

Screenshot of the Measures of Errors After the Application of ML Models