Analysis of Life Expectancy as a function of Health and Environmental Conditions

Aanchal Narendran
Computer Science and Engineering
PES University
Bangalore, India
aanchalnarendran@gmail.com

Arushi Kumar
Computer Science and Engineering
PES University
Bangalore, India
arushi32001@gmail.com

Ayush Godbole
Computer Science and Engineering
PES University
Bangalore, India
godbole.ayush@gmail.com

Ayush Kapasi
Computer Science and Engineering
PES University
Bangalore, India
kapasiayush@gmail.com

***Abstract*—The project aims to aid in understanding the impact of various medical statistics and environmental conditions on the Life Expectancy of the citizens across a multitude of countries. The primary aim is prediction of Life Expectancy with Regression models based on features obtained via statistical methods. The data used in this project has been collected from the World Health Organisation and includes various attributes such as Health Expenditure, Air Pollution, Unsafe Sanitation, GDP and so on. The project additionally aims to explore life expectancy as a function of Time. We aim to develop a clustering strategy that breaks away from this binary classification. We benchmark these models against a gradient boosting model and its metrics**

***Index Terms*—Life expectancy, Linear Regression, Clustering, Feature Selection, Boosting**

I. Introduction

Life Expectancy is a statistical measure used to estimate how long an organism born in a country under specific demographics is likely to live. In this project, the aim is to predict the life expectancy of Humans across various countries around the globe based on a multitude of lifestyle, health and environmental factors. The World Health Organisation often estimates Life Expectancy based on a plethora of characteristics. Over the past few decades, leaps in the medical domain and increased expenditure in medical infrastructure has lead to an increase in the estimate of life expectancy.

This paper aims to predict the life expectancy of various countries around the globe without restricting its scope to prominent and famous countries. The objective is to avoid bias and to build a more generalised model that can predict life expectancy for a multitude of nations.

The goal of this project is to use regression analysis to predict the life expectancy of a certain country. The regression model is built on a few key factors, determined via Exploratory Data Analysis and Feature Selection to constrain the impact of human bias. This project also aims to explore life expectancy as a function of time using various Time Series Analysis , contingent on the exploratory data analysis. Additionally, the project aims to enhance predictions with clustering.

II. Literature Review

A. Impact of pollution on Life Expectancy

The objective of this paper [1] was to gauge the impact of various air pollutants on the life expectancy of the citizens of Romania using Machine Learning algorithms. The author focuses on understanding the longstanding impacts of these metrics from various social and economical traits.

The author focuses on using correlation and linear regression to help establish the significance of the papers' finds. The paper mentions 3 main hypotheses considered by the author which focus on correlating and flagging the influence of air pollutants on the Life Expectancy. The dataset chosen is collated from the National Institute of Statistics.

The results of the authors' hypotheses are verified using a hypothesis test for hypotheses 1 and 2, which talk about the correlation of air pollutants on the life expectancy at 65 and at birth, respectively. The author shows that both of these tests are partially verified and there exists a negative correlation between the two attributes. The author uses a linear regression model to analyse life expectancy.

The author limits themselves to visualising life expectancy as a function of air pollutants. Decreasing quantities of pollutants may not directly impact the increased life expectancy due to spurious correlation. In this case, the hidden variable is urbanisation. Between 2000 to 2015, Romania's seen an increase in literacy and higher use of modern products in the urban cities to ease daily life. Moreover, NGOs in Romania play a key role in boosting the countries' life expectancy. In our project, we aim to address this by merging datasets that accounts for both, environmental conditions and key statistics such as health expenditure and GDP.

B. Life Expectancy as a function of Health status in EMR

The objective of this paper [2] is to estimate the health production function in the East Mediterranean region. Performed using econometric methods, it identifying a relationship between life expectancy and socio-economic factors from 1995 to 2007. The authors chose to use the Grossman theoretical model to build their model.

The Grossman model assumes that a human beings health is like capital goods. Over time the amount of goods decreases. However, it could be increased, to an extent, by investing in Medical care. The model assumes that at some point, the human reaches a minimum threshold. Below a threshold, the human cannot survive. Medical care is not limited to Hospital Expenditures, but also includes, Nutritional Intake, Income, Education and various other lifestyle traits. The model includes Social, Economical and Environmental attributes. The attributes considered were Income, Health Expenditure, Food Availability, to name a few.

The author showed via the regression coefficients that income had an impact on life expectancy. The author could conclusively show that the food production rate has an impact on life expectancy. However, the impact of certain attributes, like Health Expenditure, was inconclusive for different genders.

The author admits that multicollinearity may have impacted variables such as income and employment ratio. Taking this into account, the current approach used GDP. The model does not factor in the knowledge that certain countries in this region are extremely well-developed and wealthy while a few are war-torn. This cause a model to have conservative or liberal estimates. The current projects aim to address this by using clustering and building optimised models for countries that are alike in terms of the attributes considered even though they are geographically apart.

C. Life expectancy as function of Health Expenditure cross-country

This paper [3] aimed to analyse the dynamic nature of the Healthcare domain owing to a constant increase in Healthcare expenditure. This paper views the Healthcare field as a function, whose inputs are infrastructure and expenditure, with its output being a change in life expectancy of the citizens. The paper goes one step above highlights the difference in a models significance based on the country.

The author applies panel data analysis grouped by the geographical location and income level of the country. The dataset used covers data from 1995 - 2010 for different countries. The author uses a Fixed Effects Model since the data is longitudinal, time-series data, i.e there exists time series data divided by countries within a fixed period.

The author shows that developed countries tend to have a higher life expectancy than developing countries. However, there are cases where a geographical region is developed, yet there is a variation in the life expectancy (The European Region). The author classifies the countries into 4 sub-categories on the basis of income. Based on this classification, the author observes high values of R-Squared, which support the models' analysis.

The author supports the need to recreate this project on bigger datasets that consider other factors that directly impact the health status and consequentially, the life expectancy of the country. Despite Health Expenditure being a good metric to predict, there exist various Social, Economical and Environmental aspects that dictate the life expectancy of the citizens of a country. The current project addresses this issue by analysing a merged dataset that includes attributes from all the domains.

D. Analysis of Life Expectancy between developed and developing countries

The goal of this paper [4] is to use various regression models to establish a relation between various Socioeconomic factors and Life Expectancy. The paper focuses on classifying the countries into two non-overlapping divisions,i.e Developed and Developing countries. The comparative analysis of these models and the features seeks to show their effect on predicting Life Expectancy. The paper aims to predict Life Expectancy for developed and developing countries.

The author performed Exploratory Data Analysis by handling the missing values and performing correlation analysis by using heat-maps. The author chose to implement three different regressors, i.e linear regressor, decision tree regressor and random forest regressor.

The Random Forest Regressor was the best model with training and test R-squared values of 0.99 and 0.95. The author uses visualisation analysis to show and conclude the initial hypothesis, i.e the life expectancy of developed countries is better than the developing countries.

Although the author analysis the models, there is no explicit information about the performance of the models against each other. Apart from accuracy and error, the Author limits finding other metrics such as F-Score, Time taken and storage. The current project aims to address this by considering a different set of bench-marking metrics. The author does not consider the impact of various environmental conditions in the prediction of life expectancy. The current problem aims to address this by merging datasets with relevant data to aid in prediction.

The above resources aided us in isolating and formulation our problem statement which will be addressed in detail in the upcoming sections. Prior to this, a deep dive into the current dataset would assist in setting the stage for easier and better comprehension of the problem statement.

III. Proposed Solution

The problem statement involves assessing the contribution of various socio-economic, health and environmental factors on the Life Expectancy values of any country around the globe. Previously, Life Expectancy was computed cross-country [3] or based on developed or developing nations [4]. However, the analysis and models presented in the solution show that Life Expectancy can be modelled just as well independent of Country and Year.

The Life Expectancy of a nation at any given point in time could be modelled purely based on its socio-economic factors, health conditions and environmental status. These insights help the Governments of Countries to assess methods that could be implemented to curb a factor that negatively impacts Life

Expectancy. Simultaneously, it provides insights into factors that could help improve the Life Expectancy of the Countries citizens.

The initial step of the approach was to Pre-process the dataset and merge the two datasets being harnessed to develop the models. As a part of preprocessing, the data was cleaned and normalised. Feature extraction was also performed using Principal Component Analysis to highlight the best of the batch. About 5 different models were used to model the data to assess which model has the most optimal performance. The models used are as follows: Linear Regression, Decision Tree Regression, Random Forest Regression, Extreme Gradient Boosting and Clustering using K-Means. The decision to use the first three models is based on prior work done in [4]

A. *Dataset*

The dataset for this project was created by integrating data from two different datasets publicly available on Kaggle [5] [6]. Both of these datasets contain information about various Health, Social, Economical, Environmental attributes and the deaths caused due to environmental conditions. The merged dataset contains data about 164 countries between the 2000 - 2015 period.

Due to a large number of attributes, a few of them were handpicked based on literature review and domain knowledge for exploratory data analysis. Some of these attributes are GDP, Health Expenditure, Income, Schooling, Unsafe Water Sources, Unsafe Sanitation, Air Pollution levels, Diet and the Status of the country.

It is worth noting that there was a significant lack of available data in a few cases. Most of these missing values were in the early 2000s, where data collection was not as important a task. Moreover, a considerable amount of developing nations did not possess the means to collect and store data. The project makes use of various imputation methods to assess and handle any missing records.

B. *Data Preprocessing*

The main goal of this project is to assess and predict Life Expectancy based on certain socioeconomic, Health, and Environmental conditions. Due to unavailability of a single source, two datasets were integrated and cleaned to build a single data source for this project. Null values in the dataset were dropped in cases where the significance of the rows was minimal. Columns whose data was either redundant or statistically insignificant using T-tests were dropped.

The dataset was visualised using a variety of visualisation techniques 3. Key features, such as Expenditure on Health resources, Adult Mortality, Schooling and so on, have a normal distribution.

In the above figure 1, we see that schooling and income have a strong, positive correlation with life expectancy. Similarly, we notice that adult mortality has a strong, negative correlation with life expectancy. However, some factors such as Pollution and GDP have a lower correlation coefficient. Pearson's correlation coefficient has a few limitation with respect to the category of equations it can effectively model.

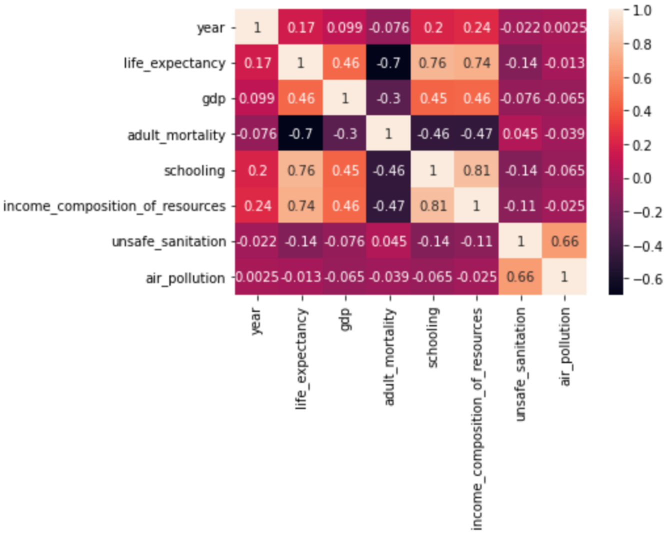

Fig. 1. Correlation Heat-map of certain variables

Since the status of the country was found to be statistically significant to the Life Expectancy, it was vectored into two columns i.e Developed and Developing with binary values. The features of the model were Winsorized. Winsorization helps in dealing with outliers. Due to the varied difference between the Life Expectancy of developed nations and developing nations, winsorization helps in making the data more uniform and ensures that legitimate values are not treated as outliers. The percentile for each feature is found using box plots.

As a part of preprocessing, the dataset was analysed for its competency to be modelled as Time Series data. The data was plotted and decomposed into 4 components for China. The decomposed components showed that the Life Expectancy values followed a linear, increasing trend with time. However, the data showed little-to-no seasonality and cyclic components. One plausible reasoning behind this is that the data is collected yearly and there may be no seasons to divide the data into. Another potential reason could be that the data covered only 15 years per country and that might have been far too little data to effectively visualise seasonality. The Durbin-Watson test statistic showed auto-correlation in the data a significant extent. With a value of 0.0 and a test significance level of 1%, the value falls well between the 0 to dl value as specified in the paper by Durbin and Watson [7]. For these reasons, the analysis concluded that modelling the above data with Time Series may provide an effective model.

The dataset at this point has about 46 columns. This introduces the curse of dimensionality and multiple columns which might be collinear. Principal component Analysis is used which reduces the feature space from 46 to 24 attributes. Some of the key features being considered are GDP, Population, Health Expenditure and so on.

IV. Fitting the Models

A Multiple Linear Regression model was built to understand the influence of the 24 principal components extracted on the Life Expectancy of the country. The model effectively manages to solidify the results obtained from Correlation Analysis. Due to minimal auto-correlation, we can safely conclude the results are due to the influence of other components. We see a cross-validation score measure of 0.88 and a root mean squared error measure of 3.40. The Linear Regression model doesn't perform well due to the presence of skewed data.

Decision tree regression builds decision trees to predict the values of Life Expectancy. Exploratory data analysis shows that the features of the dataset do contribute to life expectancy and that's crucial in justifying the usage of decision tree regression in this scenario. The model shows a cross-validation score of 0.85 and a root mean squared error of 3.69. The decision tree regression model seems to perform relatively better than the multiple linear regression model. It was observed that the Decision tree performed extremely well on a subset of the training data, but did not perform well on the test data despite varying the depth.

Random forest regression was done to replicate the results in [4] to validate the assumptions and results isolated in this approach to modelling Life Expectancy. Random forest regression models are much more robust than the aforementioned regression models, to outliers and skewed data. Owing to this, there is a considerable improvement between the metrics of the Random forest regression model and its predecessors. The random forest regression model manages to get a cross-validation score of 0.932 and a root mean squared error of 2.65. Since Random Forest regressors cannot view all the data at once, it seems to have overcome the problem of over-fitting.

Extreme Gradient Boosting is an ensemble method that uses a group of models to obtain optimal solutions and predictions. In this approach, its Regressor functionality was used to predict Life Expectancy. A probable reason for its strength here is that it can efficiently model non-linear relationships. Previously, we notices that the correlation heat-map showed low values for certain features. The Gradient Boosting model gives a cross-validation score of 0.92 and a root means squared error of 2.44.

A train-test split of 80% and 20% is used for all of the regression models. The models use k-fold cross-validation to establish their finds. The metrics used for analysis are Cross-Validation Score and Root Mean Squared Error.

The Time series model relies on modelling the given dataset as a function of time based on the auto-correlation of its values in the target variable with values from a certain amount of lags. Here, due to the existence of a pure trend line, Holt's exponential Smoothing model is used to forecast the data. Based on the visualisation of forecasted values and the root mean squared error of 0.204, it is conclusively proved that the data could be modelled as a Time Series model. However, this model can be run for only one country at a time owing to the fact that countries share the same datetime stamp and Time Series relies on a unique datetime stamp for each example.

In previous work, many authors choose to work with Countries in two classes i.e developed and developing. In this approach, clustering algorithms such as k-means is used to show that the data is better represented and clustered when the number of clusters is greater than 2. Hierarchical clustering and dendrograms along with graphs are used to visualise the same. The graphs presented as a part of clustering do not depict the entire picture since we can only model graphs in 2 Dimensions.

V. Results and Inferences

A. Expected Results

Throughout this discussion, certain assumptions have been made regarding the relationship between features and the target variables based on correlation analysis and T-test statistics for significance. Due to certain strong and linear correlations depicted by the Pearson correlation coefficient, it is assumed that the Linear Regression model would perform best.

During Literature Review, it is observed that various papers seek to classify the countries into two main clusters. The countries, in general, are usually classified into developed and developing. It is assumed that both categories of nations are linearly separable and do not interact with each other.

Since the dataset contains data segregated based on years, it is assumed that this data can be modelled using time series models. Time series models also rely on the auto-correlation of data to an extent and it is assumed that this data is auto-correlated.

B. Observed Results

Based on the results from Analysis of the data, it is conclusively demonstrated that the dataset can be modelled using the Holt Exponential Smoothing model. The model has two coefficients, one of which aids in effectively modelling the trend line. However, this model poses a limitation in terms of generalisation. The model can effectively predict for one country at a time owing to issues in datestamp. The model effectively forecasts data with the parameters and metrics in I.

TABLE I
Parameters and Metrics of Holt Exponential Smoothing

Level Coefficient	Trend Coefficient	RMSE
0.8	0.2	0.205

The models are built on the assumption that the data can be linearly modelled based on the correlation Heat Map and the reduction of dimensions. However, it is observed that the Linear regression model performs the least favourably against other tested models. This could be because the data is partly skewed and there is a large variance between various features. This variance is because different countries have different expenditures, cultures and socio-economic backgrounds. Most of these latent variables cannot be explicitly modelled by regression functions. The table II below depicts the values for various metrics the Regression models were tested against.

TABLE II
METRICS OF REGRESSION MODELS

Model	CrossValScore	RMSE
Linear Regression	0.88	3.40
Decision Tree Regression	0.85	3.69
Random Forest Regression	0.93	2.65
Extreme Gradient Boosting	0.92	2.44

The clustering model shows, to an extent, that the dataset can be split into more than two clusters. The dendrogram extracted for the given dataset shows that it can be effectively clustered better based on its principle components into more than two clusters. In the below figure 2, we see the dendrogram for the clustering of data which shows a 8 cluster configuration. However, the clusters cannot be effectively visualised due to the high dimensionality of data and restriction to 2-D visualisation.

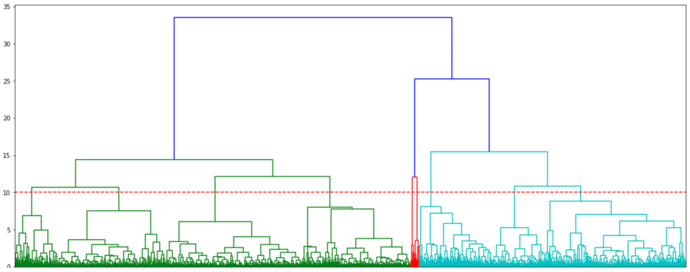

Fig. 2. Dendrogram for Clustering

VI. CONCLUSION

This project aimed to highlight the impact of various factors in the prediction of Life Expectancy. Previous work uses one out of 3 categories of factors which leaves room for improvement. Moreover, the models built show, that Life Expectancy, to an extent, may not necessarily need Country information to be predicted. It can be predicted as a function of the countries socio-economic conditions, health factors and environmental conditions.

The project also sought to shine light on evaluating Life Expectancy as a function of time and does this effectively as demonstrated in the results section. This is partially due to the fact that as time progresses nations tend to grow in terms of quality of life and that inadvertently increases the Life Expectancy.

As observed in the results section, the models seem to attain a high cross-validation score, which further supports the initial problem statement. The project explores the clustering of countries with regards to life expectancy prediction into more clusters than the typical binary classification. However, high dimensionality impedes visualising this clustering. In conclusion, this project has highlighted the possibilities of a generalised model for Life Expectancy and delves into breaking away from the binary classification of countries. The project demonstrates the visualisation of Life Expectancy as a function of time.

VII. ACKNOWLEDGMENTS

We are grateful to Dr. Gowri Srinivasa and the entire Data Analytics team, for motivating us and providing us with this opportunity to get hands-on experience in the domain. This project has proven to be a great experience in learning how data analytics is used to impact decisions that affect humans. We also learnt a great deal about how minuscule economic or lifestyle changes affect the life expectancy of a nation. We would like to thank the Computer Science and Engineering department at PES University, for always inspiring us to conduct frequent research and inculcating a problem-solving discipline in us.

REFERENCES

[1] G. M. Toma, ”Public health management: life expectancy and air pollution”, in Conf. International Conference on Business Excellence,2017
[2] M. Bayati, R. Akbarian, Z. Kavosi, ”Determinants of Life Expectancy in Eastern Mediterranean Region:A Health Production Function”, International Journal of Health Policy and Management,pp. 57 - 61, 2013
[3] E. Jaba, C. Balan, I. Robu, ”The relationship between life expectancy at birth and health expenditures estimated by a cross-country and time-series analysis”, Procedia Economics and Finance, vol. 15, pp. 108 - 114, 2014
[4] S. S. Meshram, ”Comparative Analysis of Life Expectancy between Developed and Developing Countries using Machine Learning ”, in conf. IEEE Bombay Section Signature Conference, 2020
[5] K. Rajarshi, ”Life Expectancy (WHO)”, 2018
[6] A. Verma, ”Worldwide deaths by country/risk factors”,2021
[7] J. Durbin, G. S. Watson, ”Testing for Serial Correlation in Least Squares Regression. II”,1951

VIII. APPENDIX

A. *Contributions*

1) Aanchal Narendran: Literature Survey, Exploratory Data Analysis for Time Series, tuning hyperparameters for Gradient Boosting, Cross validation for gradient boosting, Visualisations, Holt Exponential Smoothing
2) Arushi Kumar: Exploratory Data Analysis, Linear Regression, hyperparameter tuning for Decision Tree Regression and Random Forest Regression, Cross Validation for the same
3) Ayush Godbole: Exploratory Data Analysis, Pre-Processing for Clustering, tuning hyperparameters for K-Means Clustering, Clustering Visualisations
4) Ayush Kapasi: Pre-Processing for Clustering, tuning hyperparameters for K-Means Clustering, Clustering Visualisations

B. *Additional Visualisation*

- In 3, we see that the Life Expectancy follows a normal distribution.
- In 4, we see that the best performing regression models are almost close to the true value of prediction, if the predictions are incorrect

- In 5, we see the impact that the winsorization process has on the outliers in the dataset
- In 7, the green line depicts the forecasted values from the model, while the black line indicates the training values and the blue line indicates the test values.
- In **??**, we see that the K-Means clustering algorithm has efficiently divided the dataset into 8 different clusters

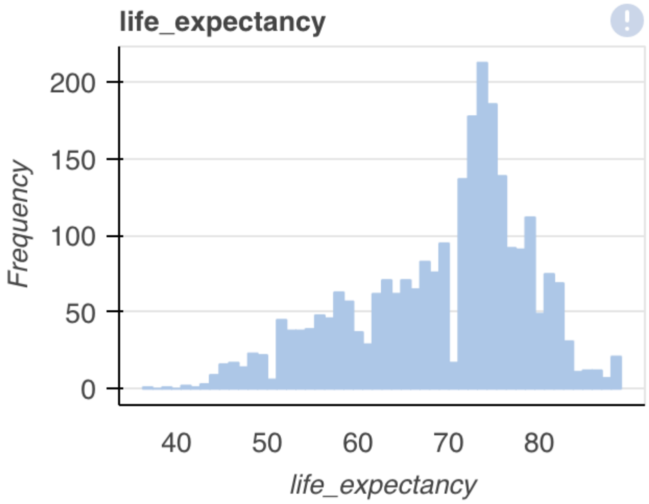

Fig. 3. Histogram of Life Expectancy depicting Normal distribution

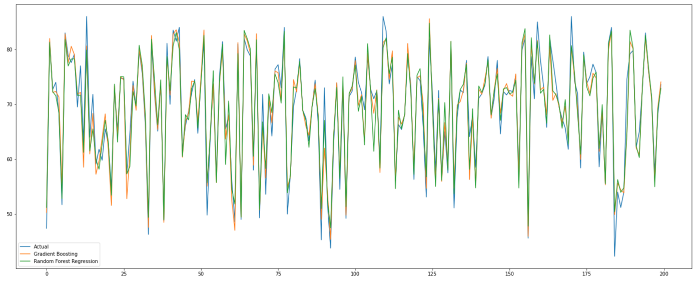

Fig. 4. Comparison of predictions from Gradient Boosting and Random Forest Regression against the true value

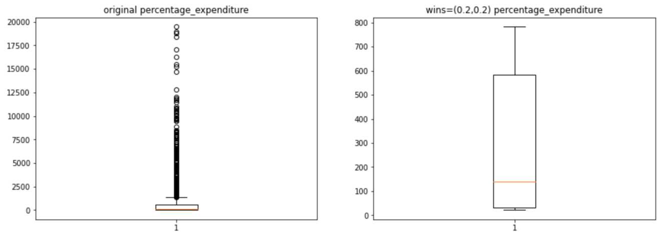

Fig. 5. Effect of Winsorization on data

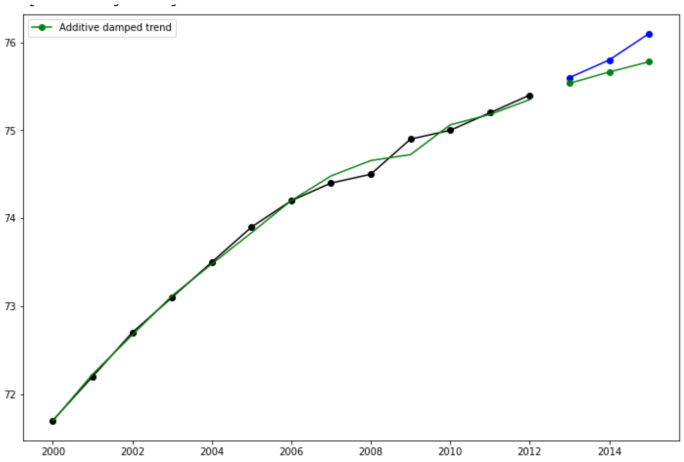

Fig. 6. Forecasted vs Actual values for Holt Exponential Smoothing

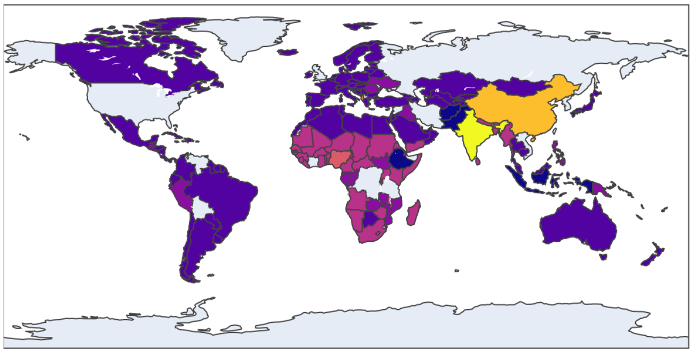

Fig. 7. Chloropleth demonstrating the clustering of countries based on K-Means