ITCS-6162 Knowledge Discovery in Databases

Y-NOTS!

Team Members

- Sushma Keddallu Suriyaprakash
- Srinath Muralinathan
- Muthu Priya Shanmugakani Velsamy
- Nivedita Veeramanigandan
- Balasundaram Avudai Nayagam
- Arjun Manevannan

Picturizing the Life-span by Analysis of Economy and Public Health data

Project and Team Introduction

Life expectancy is a reflection on the quality of life in a country, since individuals can hope to live longer lives. It is an estimate of an individual's lifespan derived from averaging the age all individuals who die in a particular year. Our project deals with determining the life expectancy of a region based on analysis of immunization factors, mortality factors, economic factors, social factors and other health related factors. We take into account these factors across ethnicity and geographical regions. Since the observations this dataset are based on different countries, it will be easier for a country to determine the predicting factor which is contributing to lower value of life expectancy. On performing analysis on the data, we produce life expectancy as our model's outcome. This will help in suggesting a country which area should be given importance in order to efficiently improve the life expectancy of its population.

Data and Source description

The dataset is acquired from Kaggle. The dataset related to life expectancy focuses on, the health factors of 193 countries and the details have been collected from WHO. The data consists of 22 Columns and 2938 rows. Our model's predicting value will be Life expectancy.

Application of the CRISP-DM Process

- **Business Knowledge**
  - The healthcare industry is one of the largest industries in the world, and it has a direct effect on the quality of life of people in each country. This project will focus on healthcare domain. The data source is acquired from World Health Organization's (WHO) Global Health Observatory (GHO) data repository and United Nations website. The dataset is related to the life-

expectancy of humans, based on various dimensions such as, health, immunization, economic and social factors.

- **Data Understanding and EDA**
  - We are in the process of performing initial study on data to discover if there are patterns, to examine anomalies, to test hypothesis on the data set, and to check the assumptions with the help of summary statistics and visual representations.
- **Data Preparation**
  - Data preparation for modelling is underway. We are still in the process of deciding the data transformation methods.
- **Machine Learning**

The dataset will be split into two data 80:20 ratio where 80% of data are used for training the model and 20% of data are used for testing the model. The model will be trained through various regression techniques like:

- Linear Regression
- Random Forest
- Lasso Regression
- Elastic Net regression

since our data contains more of numerical data values. While validating our machine learning model by its learning and behavior with the new data, we are using 2 essential techniques:

- Overfitting of data
- Underfitting of data

which are majorly responsible for evaluating the performances of the machine learning algorithms. Underfitting data can be avoided by using more data and also reducing the features by feature selection. We can also use Unification technique and dimensionality reduction techniques for feature selection. Overfitting data can be avoided through one of the common methodologies called cross-validation. Once we avoid both overfitting and underfitting of data, we will achieve the good fit on the data. A good fit data will have less error rate while training and testing the data.

The attributes are examined through their degree of correlation or multicollinearity with other attributes. The correlation among the attributes are visualized through the heatmap. While training and testing the model, the difference between the predicted and the observed values can be calculated through:

- Root Mean Square Error (RMSE)

where RMSE provides the absolute measure of fit. This will be useful in predicting the accuracy and efficiency of the trained model.

The data is evaluated using analytical and logical reasoning. Each attribute in the data are considered for the evaluation. The education column plays an important role in the population growth. The more knowledge an individual has, the more he or she can make informed life decisions, and improve his or her quality of life.

The attributes like Polio, Measles, Hepatitis-B, Diphtheria can be used to analyze the immunization level of the people. The dense and sparsely populated countries have different cultures, behavior, social and economic status. Based on all these attributes we will predict the life expectancy of the people in various countries.

- **Evaluation**
    - The data is evaluated using analytical and logical reasoning. Each attribute in the data are considered for the evaluation. The education column plays an important role in the population growth. The more knowledge an individual has, the more he or she can make informed life decisions, and improve his or her quality of life. The attributes like Polio, Measles, Hepatitis-B, Diphtheria can be used to analyze the immunization level of the people. The dense and sparsely populated countries have different cultures, behavior, social and economic status. Based on all these attributes we will predict the life expectancy of the people in various countries.
- **Conclusion**
    - Our model will help in suggesting various countries in order to efficiently improve the life expectancy of its population through immunization or by increasing their investment on healthcare field to improve the available medications.