|  |  |  |
| --- | --- | --- |
| Internship Title | : | Predicting Life Expectancy using Machine Learning – SB25411 |
| Project Title | : | Predicting Life Expectancy using Machine Learning |
|  |  |  |
|  |  |  |

* **Introductions**

Life expectancy is the key metric for assessing population health. Broader than the narrow metric of the [infant and child mortality](https://ourworldindata.org/child-mortality), which focus solely at mortality at a young age, life expectancy captures the mortality along the entire life course. It tells us the average age of death in a population.

In this project, the model tries to predict the life expectancy of a country by considering various features.

* **Problem Definition and Algorithm**
  + Task Definition
    - Life expectancy is a statistical measure of the average time a human being is expected to live, Life expectancy depends on various factors: Regional variations, Economic Circumstances, Sex Differences, Mental Illnesses, Physical Illnesses, Education, Year of their birth and other demographic factors. This problem statement provides a way to predict average life expectancy of people living in a country when various factors such as year, GDP, education, alcohol intake of people in the country, expenditure on healthcare system and some specific disease related deaths that happened in the country are given.
  + Algorithm Definition
    - Ridge regression was used since it is a technique for analyzing multiple **regression** data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value.
  + **Experimental Analysis**
    - Methodology

The model evaluation started by reading the dataset. Then the datatypes of each column were studied in order to identify if there is any categorical data. Status and Country columns from the dataset were identified containing categorical data. The columns were label encoded and a group by was performed on the dataset by using Country column. After the initial cleansing of the dataset Exploratory Data analysis was performed on the dataset. Scatter plots were used to find the relation between a particular column and the column to be predicted which was **Life Expectancy.**

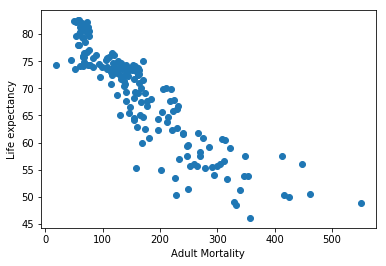


Fig 1

Fig 1 shows the scatter plot between Life Expectancy and Adult Mortality which shows a negative correlation. Such plots were drawn for other columns as well and necessary information was extracted by looking at the graphs.

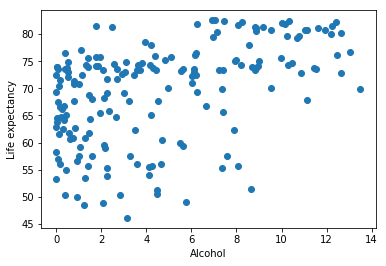


Fig 2

Fig 2 shows a scatter plot between Alcohol and Life Expectancy from which we could deduce that there is no correlation between Alcohol and Life Expectancy.

Heat map was also plotted inorder to find the correlation.

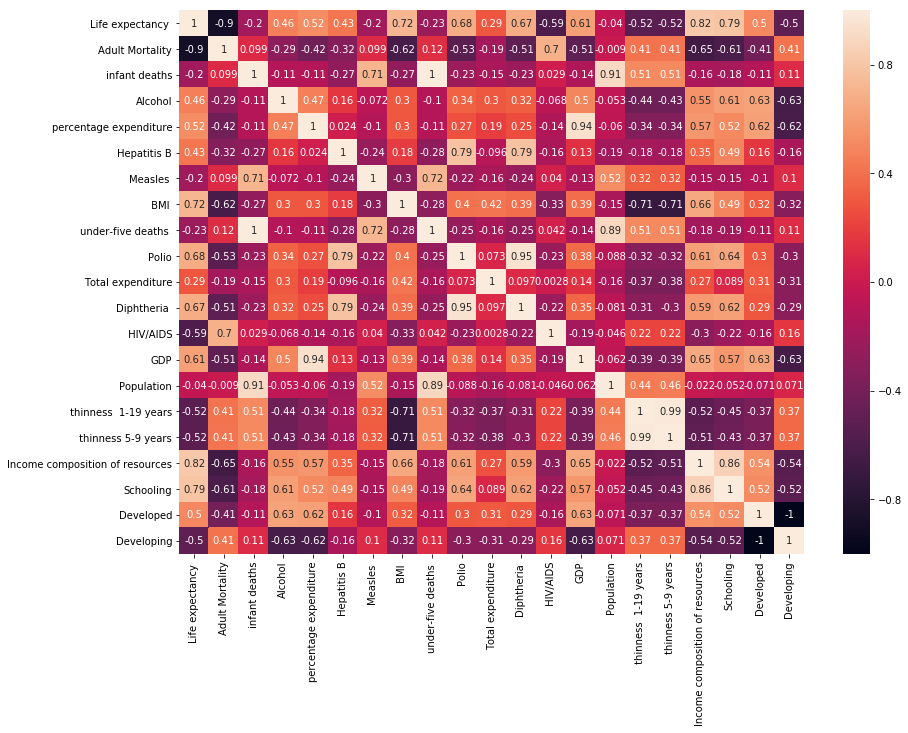
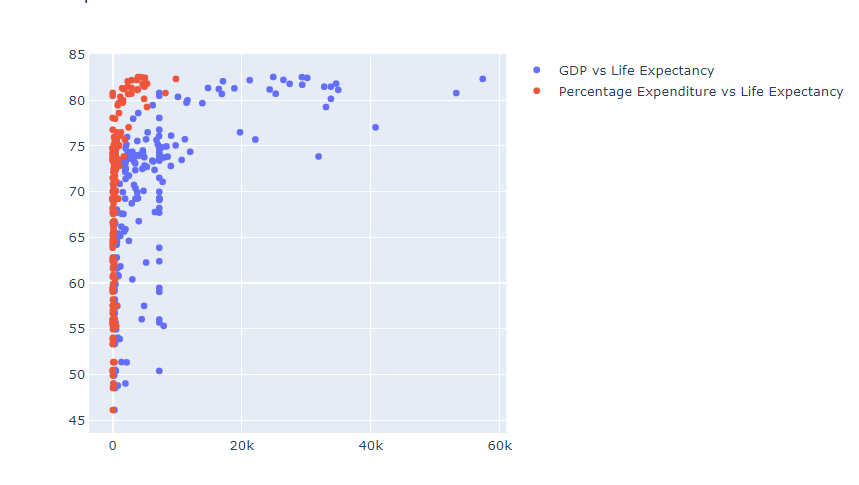
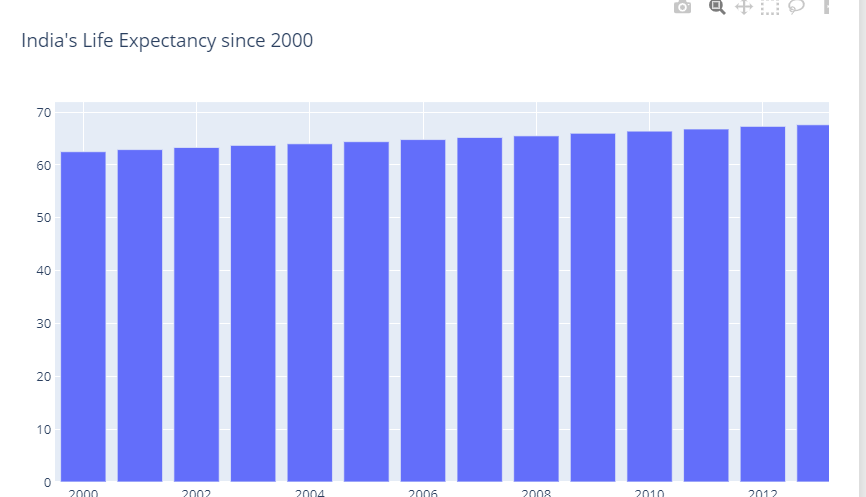


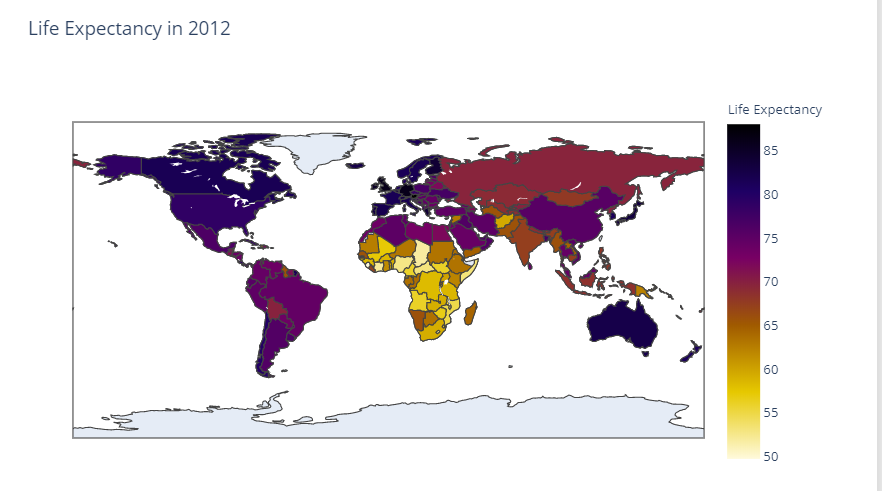
Fig 4

The legend tells that the warmer colors show higher and positive correlation, while the colder low or negative. From figure 4, there is a very high correlation between thinness of 5-9 year-old and that of 1-19 year-old. Also between population and infant deaths, under 5 deaths, another is between schooling and income composition of resources. On the other hand Life expectancy and Adult Mortality are very highly negatively correlated.

The dataset was then preprocessed. Null values were also replaced by the mean and the dataset was divided into training set and test set with 70:30 ratio. Then Ridge regression was applied. Best parameters and R square score was found for training as well as the test data.

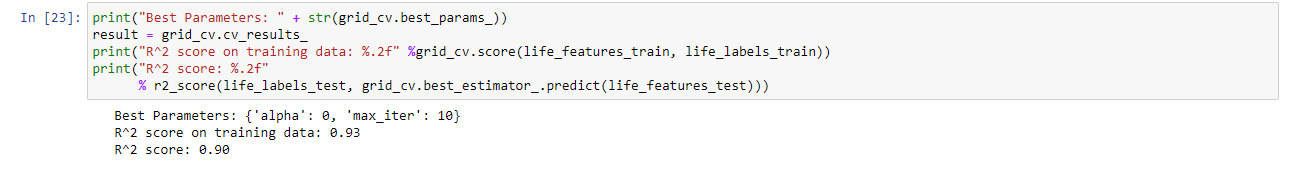






* + - Results

The following results were obtained:



The Ridge Regression uses L2 regularization to minimize the weights of the coefficients, this is controlled with the hyperparameter alpha. Increasing the value of alpha decreases the weights of the coefficients.  
A grid search with cross-validation on the grid regression with alpha varying between 0 and 10 and having 3 max iterations of 10, 100 and 1000 was performed. Finally the best parameters here are alpha = 1, and max iterations = 10. The R square on the training data is 94% compared to 92 % on the standard linear model. The R square on the test data was found to be 90%

* **Future Work**

We could possibly collect more data by expanding the scope to cities instead of countries, and to explore other features (factors) affecting life expectancy. Also, we could split the data to male and female categories for such life expectancy regression analysis.

* **Conclusion**

The best parameters were found to be alpha : 0 and max iterations to be performed : 10.

R Square score on training data was found to be 0.93.

R Square score on test data was found to be 0.90.

* **References**
* Dataset - kaggle.com/kumarajarshi/life-expectancy-who
* <https://hackernoon.com/regression-analysis-on-life-expectancy-4wp34rf>
* <https://www.datasciencesociety.net/using-machine-learning-to-explain-and-predict-the-life-expectancy-of-different-countries/>
* <https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db>