**INTRODUCTION**

**Overview**

The aim of this project is to predict the life expectancy (dependent variable) using 21 independent variables. The given data consists of 2938 samples. The data is processed by converting string values to integers and replacing NaN by mean of the corresponding feature. Further, train test splitting is done with 70% train and 30% test. This is followed by feature selection f\_regression, which outputs the 4 most important features used for prediction. At last, regression is done using random forest regressor and node red application is constructed which takes the feature inputs and gives the predicted value using regression model. The root mean square error of 2.0 was obtained between actual and predicted life expectancy.

**Purpose**

1. To find unknown life expectancy, given the feature set.
2. To create a node red application using the model constructed which outputs the predicted value.

**LITERATURE SURVEY**

**Existing problem**

The problem is to predict Life Expectancy using Machine Learning from independent features or variables: country, year, status, adult mortality, infant deaths, alcohol, percentage expenditure, hepatitis B, measles, BMI, under-five deaths, polio, total expenditure, diphtheria, HIV/AIDS, GDP, population, thinness 1-19 years, thinness 5-9 years, income composition of resources & schooling. Testing is done for prediction after training process.

**Proposed solution**

The following steps are considered in developing the best solution:

1. Data is processed by converting string values into integers and removing NaN containing samples or replacing NaN with zero or mean of their feature.
2. Data is split into train test with 70% train and 30% test.
3. Important features are selected using feature selection f\_regression.
4. Different regression models are used for prediction.
5. The performance of the prediction is calculated by using root mean square error.
6. That regression model having least root mean square error was used for node red application.

**THEORITICAL ANALYSIS**

**Block diagram**

Steps Followed:

* STEP 1: **Data Loading**: Data is loaded in Pandas Data Frame.
* STEP 2: **Data Preprocessing**: This step consists of two sub-steps namely
  + Encoding Strings
  + Handling NaN values
* STEP 3: **Train Test Split**: Data is split into the train and test.
* STEP 4: **Feature Selection**: Important features are selected using f\_regression.
* STEP 5: **Regression**: The Regression algorithm is used for the regression model.
* STEP 6: **Verification**: Test data is verified by calculating the root mean square error.

**Hardware / Software designing**

The hardware requirements for the project are as follows:

**CPU**: i5-9300H

**RAM**: 8 GB

**HDD**: 1 TB

The software requirements of the project are as follows:

**Environment**: IBM Cloud

**Software**: Watson Studio, Watson Machine Learning, Node Red

**Language**: Python

**Package**: watson\_machine\_learning\_client, types, botocore.client, ibm\_boto3, sklearn, pandas, numpy, math

**FLOWCHART**

**REGRESSION**

**FEATURE SELECTION**

**TRAIN TEST SPLIT**

**DATA**

**PREPROCESSING**

**EXPERIMENTAL INVESTIGATIONS**

In order to select the final model, the following experiments were carried out:

The code developed used the following methods:

1. Loading csv File:

Life Expectancy csv dataset was loaded in Watson Studio using packages botocore.client, ibm\_boto3 & types. The dataset consists of 2938 samples with 21 features.

1. Data Preprocessing:

Encoding of string values to integer takes place using unique function and the NaN values were replaced by the mean of the column using fillna function. The data contains 1649 rows with Nan and following columns contain Nan’s: adult mortality=10, alcohol=194, hepatitis B=553, BMI=34, polio=19, total expenditure=226, diphtheria=19, GDP=448, population=652, thinness 1-19 years=34, thinness 5-9 years=34, income composition of resources=167 & schooling=163, while target consists 10 NaN values.

1. Train Test Split:

Train Test Splitting is done randomly using train\_test\_split function. Splitting ratio was set to generate 70% train and 30% test data i.e. train data contains 2056 samples while test contains 882 samples.

1. Feature Selection:

Top four important features (Schooling, Adult Mortality, Income composition of resources & HIV/AIDS) are selected using f\_regression function.

1. Regression:

Regression model is constructed using the Random Forest Regressor.

In the experiment, the following pipelines were analyzed:

1. The removal of NaN containing rows using dropna function resulted in an rmse value of 146.40, using Gradient Descent algorithm.
2. The replacing of NaN values with 0 resulted in an rmse of 5.74, using Linear Regression model and KFold train test splitting.
3. The replacing of NaN values with mean of column resulted in the following:

* The rmse of 168.47, using Gradient Descent algorithm.
* The rmse of 4.35, using Linear Regression model and KFOLD train test splitting and using train\_test\_split, 4.24 was obtained.
* The rmse of 68.0, using Multi-Layer Perceptron Regression model in case of two hidden layers, each containing five neurons.
* The rmse of 9.59, using Support Vector Machine Regression model in case of rbf kernel with auto gamma and cost of twenty.
* The rmse of 2.44, using Decision Tree Regression model.
* The rmse of 2.0, using Random Forest Regression model by using f\_regression feature selection.

1. Train Test Splitting is done randomly using train\_test\_split function. Splitting ratio is 70% train and 30% test i.e. train data contains 2056 samples while test contains 882 samples.
2. Different Regression Models are constructed using:

* Linear Regression
* Multi-Layer Perceptron Regressor
* Support Vector Regressor
* Decision Tree Regressor
* Random Forest Regressor

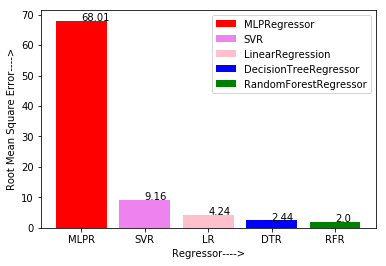
1. Root mean square error is used for verification between predicted and actual life expectancy.

**RESULTS & OBSERVATIONS**

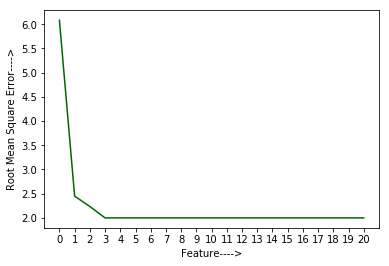
Following table shows the value of root mean square error obtained along with different regressors used in the project:

|  |  |
| --- | --- |
| REGRESSOR | RMSE |
| **LinearRegression** | **4.242640687119285** |
| **MLPRegressor** | **68.01470429252781** |
| **SVR** | **9.16515138991168** |
| **DecisionTreeRegressor** | **2.449489742783178** |
| **RandomForestRegressor** | **2.0** |

Following figure shows bar graph of root mean square error obtained along with different regressors used in the project:



Following line graph shows root mean square error with the number of features used in the project before feature selection:



Following is the node red application link of this project:

<https://node-red-foeir.eu-gb.mybluemix.net/ui/#!/0?socketid=L0SXuhAY-qPMQK-fAAAG>

**ADVANTAGES & DISADVANTAGES**

Advantages of life expectancy are as follows:

1. It can be used to find the causes of the death and relate the range of ages with life expectancy.
2. It gives more weight to deaths at younger ages.
3. It has been used nationally to monitor health inequalities.

Disadvantages of life expectancy are as follows:

1. It is a prediction, not a measurable fact. Instead of giving exact picture, it gives a predicted picture.
2. Using it as an indicator is suspect, unless you reveal the underlying data. It is better to use life duration statistics as an indicator.
3. At smaller geographies, it may be influenced by nursing homes in the area.

**APPLICATIONS**

Applications of regression are as follows:

1. Regression Analysis Basics:

This establishes the relationship between the independent variable and dependent variable. For example, if the sales of cars are to be related to GDP then the following equation is made:

car sales = (growth)\*c + b

where, c & b are to be determined using mathematical modelling or machine learning.

1. Multiple Regressions:

The number of potential independent variables is unlimited and the model is referred to as multiple regressions if it involves several independent variables. Regression models also can pinpoint more complex relationships between variables.

1. Predicting the Weather:

The linear regression is applied on the set of data and the coefficients are used to predict the rainfall based on the corresponding values of the parameters. The main advantage of this model is that this model estimates the rainfall based on the previous correlation between the different atmospheric parameters.

1. Optimization of Business Processes:

Another key use of regression models is the optimization of business processes. This requires building a model to understand the relationship between two or more variables.

Applications of life expectancy are as follows:

1. It gives the idea of many other perspectives like adult mortality of the country.
2. It reflects local conditions of nation. In less-developed countries, life expectancy at birth is relatively low as compared with more-developed countries because of high infant mortality rates (commonly due to [infectious disease](https://www.britannica.com/science/infectious-disease) or lack of access to a clean water supply).

**CONCLUSION**

Regression can be defined as a method that models a target value based on independent predictors. A regression can only have one dependent variable. The variable you are trying to estimate is referred to as dependent, while the variable you use in the model to predict the dependent variable is called independent. It finds the relationship between a dependent variable an independent variable, thus acting as a statistical tool.

The given problem in the project is based on regression. The given dataset Life Expectancy consists of 2938 samples with 21 independent variables and 1 dependent variable, all contained with NaN. Therefore, data preprocessing is required for cleaning which includes conversion of string values into integers and replacing NaN into mean of their attribute. Splitting of data takes place into train and test and for improving the result, feature selection is applied. Then, various regression models have been tested and which gives least root mean square error, that model was deployed into node red application. At last, a least rmse of 2.0 was obtained.

**FUTURE SCOPE**

Following are some future scope of the project, which enhances the result of life expectancy prediction:

1. The performance of all the regression models used increases by varying their parameters.
2. Ensemble method can be used, which uses the algorithm of taking average of results of various models.
3. The missing values i.e. NaN, can be calculated by using machine learning or deep learning.
4. The graphical GUI can also be implemented.

**BIBLIOGRAPHY**

Trappenberg, T. P. (2019). Machine learning with sklearn. *Fundamentals of Machine Learning*, 38–65. doi:10.1093/oso/9780198828044.003.0003

Hastie, Friedman, and Tibshirani, *The Elements of Statistical Learning*, 2001

Bishop, *Pattern Recognition and Machine Learning*, 2006

Scholkopf and Smola, Learning with Kernels, 2002

 Gareth James; Daniela Witten; Trevor Hastie; Robert Tibshirani (2013). [*An Introduction to Statistical Learning*](http://www-bcf.usc.edu/~gareth/ISL/). Springer. p. 204

Bermingham, Mairead L.; Pong-Wong, Ricardo; Spiliopoulou, Athina; Hayward, Caroline; Rudan, Igor; Campbell, Harry; Wright, Alan F.; Wilson, James F.; Agakov, Felix; Navarro, Pau; Haley, Chris S. (2015). ["Application of high-dimensional feature selection: evaluation for genomic prediction in man"](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4437376). [*Sci. Rep.*](https://en.wikipedia.org/wiki/Scientific_Reports) **5**: 10312

Guyon, Isabelle; Elisseeff, André (2003). ["An Introduction to Variable and Feature Selection"](http://jmlr.csail.mit.edu/papers/v3/guyon03a.html). [*JMLR*](https://en.wikipedia.org/wiki/Journal_of_Machine_Learning_Research). **3**

 David A. Freedman (27 April 2009). [*Statistical Models: Theory and Practice*](https://books.google.com/books?id=fW_9BV5Wpf8C&q=%22regression+analysis%22#v=snippet&q=%22regression%20analysis%22&f=false). Cambridge University Press. [*ISBN*](https://en.wikipedia.org/wiki/ISBN_(identifier)) [*978-1-139-47731-4*](https://en.wikipedia.org/wiki/Special:BookSources/978-1-139-47731-4)

*[Lemaréchal, C.](https://en.wikipedia.org/wiki/Claude_Lemar%C3%A9chal" \o "Claude Lemaréchal)* (2012). [*"Cauchy and the Gradient Method"*](https://www.math.uni-bielefeld.de/documenta/vol-ismp/40_lemarechal-claude.pdf)*(PDF)*. Doc Math Extra: 251–254

 Fabian Pedregosa; Gaël Varoquaux; Alexandre Gramfort; Vincent Michel; Bertrand Thirion; Olivier Grisel; Mathieu Blondel; Peter Prettenhofer; Ron Weiss; Vincent Dubourg; Jake Vanderplas; Alexandre Passos; David Cournapeau; Matthieu Perrot; Édouard Duchesnay (2011). [*"Scikit-learn: Machine Learning in Python"*](http://jmlr.org/papers/v12/pedregosa11a.html). Journal of Machine Learning Research. **12**: 2825–2830

["scikit-learn Version 0.23"](https://scikit-learn.org/stable/whats_new/v0.23.html)

**APPENDIX**

**Source code**

<https://github.com/SmartPracticeschool/llSPS-INT-1899-Predicting-Life-Expectancy-using-Machine-Learning/blob/master/Notebook%20Deployed.ipynb>