Course Code : AIT302

Course Name : Statistical Learning

Lecturer : Dr Shamini Raja Kumaran

Academic Session : 2022/09

Assessment Title : Final Project

Submission Due Date : 27th December

Prepared by :

Student ID	Student Name
AIT2009357	Bi Xiaoyang
AIT2009362	He Enhao
AIT2009376	Wang Qipeng

Date Received :

Feedback from Lecturer:

Mark:

Own Work Declaration

I/We hereby understand my/our work would be checked for plagiarism or other misconduct, and the softcopy would be saved for future comparison(s).

I/We hereby confirm that all the references or sources of citations have been correctly listed or presented and I/we clearly understand the serious consequence caused by any intentional or unintentional misconduct.

This work is not made on any work of other students (past or present), and it has not been submitted to any other courses or institutions before.

Signature: 毕骁杨 何思豪 Wang Qipeng

Date: 2022/12/27

Analysis of
Worldwide Life Expectancy
in the period between 2000 and 2015
Based on Statistical **Regression**
Method

Prepared by

Bi Xiaoyang

He Enhao

Wang Qipeng

Table of Contents

Abstract

This project focuses on exploring the relationship between life expectancy and the factors that may influence it across all countries in the world between 2000 and 2015. The core question here is “Which factors affect life expectancy most?”. We used the dataset provided by World Health Organization (WHO), which can indicate its authenticity, and we used both dropping and replacement methods to eliminate the null values in it. After pre-processing, two methods of feature selection were applied: Mutual Information and Boruta, which led to conclusions about which factors have the greatest impact on life expectancy. Regression models were then used on the well-processed dataset, and we analyzed their results to answer the questions in context. In the above processing and analysis, criteria (e.g. R2 score) that measure the dataset and the performance of models have been observed and visualizations have been implemented. It is believed that the project will help to formulate policies that a country should implement accordingly in order to effectively increase the life expectancy of its people.

Keywords: Regression, Life Expectancy, Boruta, Feature Selection

1. Introduction

Longevity has been one of the major directions of scientific and technological development of human beings, and with the advancement of medical science and immunology, the global average life expectancy has increased from 47 years in 1950 to 73.2 years today. Such remarkable improvements have led to a considerable amount of research into the factors that affect life expectancy. However, we found that most of the early relevant studies only considered demographic variables, income composition, and mortality, and did not take into account the effects of immunization and the Human Development Index (HDI). Also, some of the past studies focused on multivariate linear regressions of the **same** year for different countries. To fill this gap, we address all the factors described previously by developing regression models based on mixed effects models and multiple linear regressions while considering data for all countries **from 2000 to 2015**.

Correspondingly, we selected **Life Expectancy (WHO) dataset** for our project.

In response to our desire to complement the previous study and based on our understanding of the dataset, we proposed the following questions and attempted to answer them with the analysis of the dataset.

- Do various predicting factors which has been chosen initially really affect the Life expectancy? What are the predicting variables actually affecting the life expectancy?
- Should a developing country(i.e Malaysia, China) increase its healthcare expenditure in order to improve its average lifespan?
- How does Infant and Adult mortality rates affect life expectancy?
- What is the impact of schooling on the lifespan of humans?
- Does Life Expectancy have positive or negative relationship with drinking alcohol?

To obtain answers to these questions, two data pre-processing methods were employed to obtain the most suitable dataset for regression application. We also adopted two feature selection methods, Mutual Information and Boruta Algorithm, to find the features that have the greatest impact on life expectancy from different perspectives. Finally, different regressions were implemented to the dataset to confirm our assumption. The results were analyzed along with visualization figures.

The main content of report is organized as follow.

In **section 2**, we focused on the survey of the literature on regression analysis of life expectancy and feature selection, which provides a clear understanding of the previous work about analysing the life expectancy and what we are supposed to do with the dataset.

In **section 3**, we elaborated information about our dataset, as well as data pre-processing, feature selection, regression models and visualization and how to implement them in code.

In **section 4**, we represented the result by charts(visualization) and text to analyze the results in detail and made certain inferences about their causes.

In **section 5,6**, we did the extra work and gave our suggestion about how to enhance the life expectancy and made a conclusion of the whole project.

1.1 Contribution of each team member

In this section, we provide the work segregated among each team members.

Bi Xiaoyang	Writing code for building models, do experiments, comparisons(MY and CN),Time series predicting and visualization, writing the literature review, methodology part of report	**33.3%**
Wang Qipeng	Writing report(introduction and analysis part) and write the code of data pre-processing part[filling missing value(second method)]	**33.3%**
He Enhao	Writing code of Feature selection, data pre-processing part[deleting missing value(First method)], writing report of methodology, experiment	**33.3%**

	and analysis, conclusion part	

2. Literature review

In this section, we focuses on earlier work analysing life expectancy with social development, which is a multidimensional subject with multiple dimensions. It encompasses the economics, politics, culture, and so on in a broad sense, but merely the development of education, healthcare, the environment, and so on in a narrow meaning. In a broad sense, social development is related to, but distinct from, economic development. And also, our project focus on feature selection. So we surveyed the literature on analysis of life expectancy and feature selection, which allowed us to learn about the forefront research in these fields and to apply them in our subsequent project.

2.1. Regression analysis of life expectancy

In the project, we mainly use regression models, including linear regression models[1], ridge regression models[2], random forest regression models[3], and multilayer perceptron regression models[4]. Many researchers have explored the impact of social development on life expectancy[6][7][8]. Aalen et.al [5]simulation as well as by data from a clinical trial of treatment of carcinoma of the oropharynx. However, social development is a multifaceted term with significant regional and temporal differences. Furthermore, different factors of social development may have varying effects on life expectancy in different places and time periods. Healthcare input and development are intimately related to population health. Prior to the Chinese economic reform in 1978, a massive development of basic healthcare contributed significantly to the enhancement of life expectancy[8]. After 1978, some scholars believed that there were some discrepancies in the spatial distribution of life expectancy in China due to the unequal spatial distribution of healthcare resources and services [9]. Ming and Dong[10] reported that for life expectancy promotion, it was not the increase in healthcare provisions but the increase in the usage efficiency of the healthcare services that mattered. However, in a trans-national study in Europe, Heuvel et al.[11] argued that health-care investments had little effects on the promotion of life expectancy, while investments in social protection could greatly improve life expectancy, which also suggested the indirect importance of social protection and harmony.

The development of education is a vital dimension of social development that may also greatly influence life expectancy. Most previous studies indicated that education development could contribute greatly to life expectancy, but the marginal enhancing effect was usually smaller than that of economic development[12][13]. Some people argued that the marginal enhancing effect of education is the most important factor for the promotion of life expectancy[14].Other evidence suggestedthat there was a non-significant effect of education develop-ment on life expectancy and other health indicators[15].

The role of environmental development in life expectancy promotion has gradually caught researchers' attention in recent years. Based on the comparison of life expectancy values between some developed nations and some postsocialist countries, Feacham[16] noted that the environmental disadvantages produced many detrimental impacts on life expectancy among the residents of postsocialist countries. In addition, in astudy conducted in 156 nations, Gulis[14] also reported that a better aquatic environment, more than improvements in economy and education development, could most significantly increase life expectancy. What is more, an investigation conducted in China suggested that owing to its severe air pollution status, residents in northern China would, on average, have a life expectancy that was 5 years shorter than residents in southern China[17].

2.2 Feature selection

Feature selection is divided into three main categories, Filter Methods[18], Wrapper Methods[19] and Embedded Methods[20].Filter Methods compute some measure from the general features of the training data as a processing step prior to modeling, scoring the features or a subset of features.Filter Methods can be further classified according to their use of Filter Methods can further classify features or feature subsets based on the filter measures they use, i.e., information, distance, dependency, consistency, similarity, and statistical measures, such as information gain [21], chi-square [22], Mutual Information[23] and Relief [24].

Wrapper Methods uses any independent modeling algorithm to train a predictive model using a subset of candidate features. The feature set is usually scored using the test performance on a fixed set. Alternatively, in random forests, feature subsets can be selected based on the estimated feature importance scores. The Boruta[25] is one of the method of it.

Embedded Methods use feature selection as part of the execution of the modeling algorithm. These methods tend to be more computationally efficient than Wrapper Methods because they integrate both modeling and feature selection. As with Wrapper Methods, the features selected by Embedded Methods depend on the learning algorithm.Examples of Embedded Methods are Lasso [26], Elastic Net [27], and various decision tree based algorithms such as CART [28], C4.5 [29], and XGBoost [30].

3. Methodology

3.1. Dataset

In the project, we obtained the dataset from Kaggle, which covers the data of the factors affecting the life expectancy of 193 countries from 2000 to 2015 provided by the World Health Organization (WHO). The final merged file (final dataset) consists of 22 columns and 2938 rows which means 19 predicting variables. **All predicting variables were then divided into several broad categories: immunization related factors, mortality factors, economy factors, and social factors.**

The dataset has some outliers and NaN present, which we believe may be based on realistic factors (such as possible wars or the absence of country-related statistics). We will illustrate the processing of them in detail in the data preprocessing section.

Overall, this is a dataset with a reliable source (official WHO contribution), but also a challenging one that can meet a variety of analytical directions. We believe that the analysis performed on this dataset can identify factors that influence life expectancy the most, which will help to suggest policies that a country should implement accordingly in order to effectively increase the life expectancy of its people.

And we made the following table to present each independent variable’s meaning.

Country	Country
Year	Year
Status	Developed or Developing status
Life_Expectancy	Life Expectancy in age
Adult_Mortality	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
Infant_Deaths	Number of Infant Deaths per 1000 population
Alcohol	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
Percentage_Expenditure	Expenditure on health as a percentage of Gross Domestic Product per capita(%)
HepatitisB	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
Measles	Measles - number of reported cases per 1000 population
BMI	Average Body Mass Index of entire population
Under_Five_Deaths	Number of under-five deaths per 1000 population
Polio	Polio (Pol3) immunization coverage among 1-year-olds (%)
Total_Expenditure	General government expenditure on health as a percentage of total government expenditure (%)
Diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
HIV/AIDS	Deaths per 1 000 live births HIV/AIDS (0-4 years)
GDP	Gross Domestic Product per capita (in USD)
Population	Population of the country
Thinness_10-19_Years	Prevalence of thinness among children and adolescents for Age 10 to 19 (%)
Thinness_5-9_Years	Prevalence of thinness among children for Age 5 to 9(%)
Income_Composition_Of_Resources	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)

Schooling	Number of years of Schooling(years)

Table 1. *Explaination of each variable in this dataset*

3.2. Data Preprocessing

In practical applications of statistical learning, datasets are often incomplete (presence of null values), inconsistent (data type not numeric), and susceptible to noise (occurrence of abnormal values). Unprocessed, low-quality data will lead to low-quality mining results. Just like chefs nowadays want to make delicious steamed fish, if they don't process the fish such as scaling, they will definitely not make the delicious taste in our mouths. Data pre-processing is an effective tool to solve the above problems.

The dataset has 22 columns and nearly 3000 entries and with the inequality in the number of observations from each feature, it is obvious that null values exist. We can also know that there are 20 Quantitative and 2 Qualitative features. From the information shown we know that the parts we need to process are the following.

1. Null values
2. Object variables

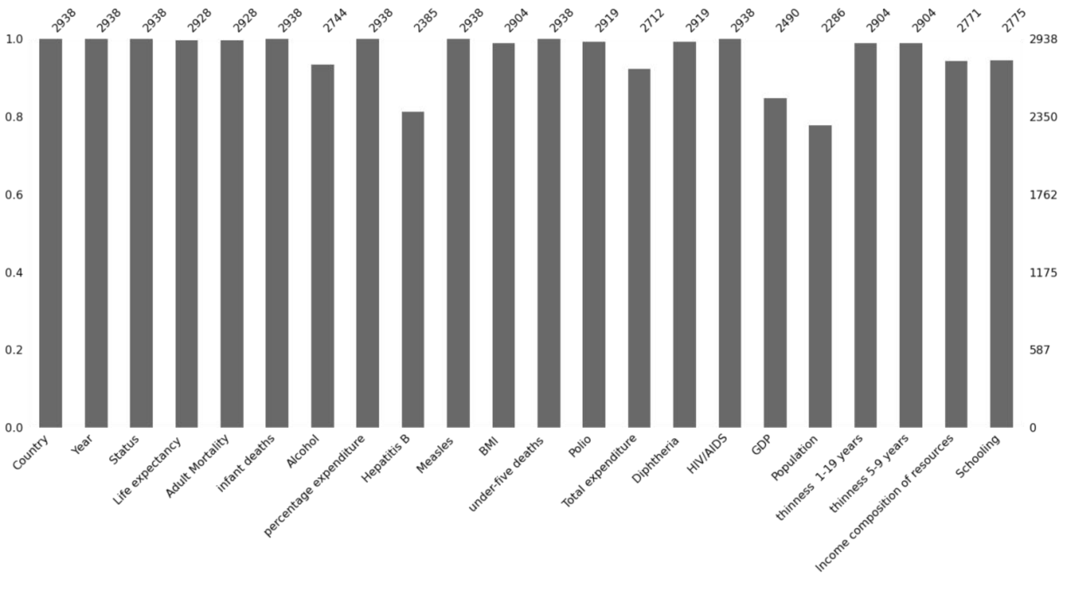
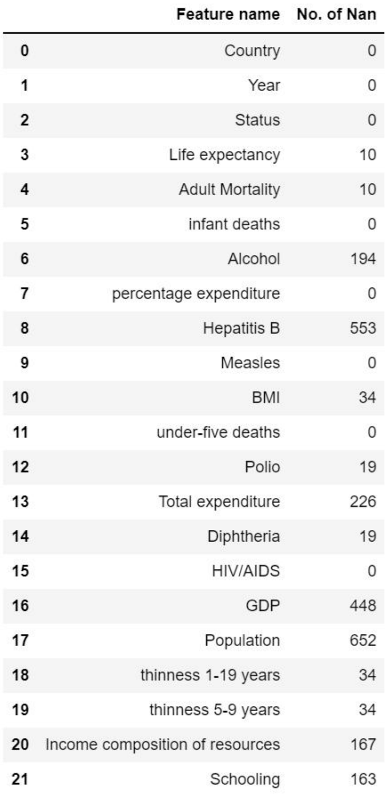


Fig 1. *number of normal values in each feature*



	Feature name	No. of Nan
0	Country	0
1	Year	0
2	Status	0
3	Life expectancy	10
4	Adult Mortality	10
5	infant deaths	0
6	Alcohol	194
7	percentage expenditure	0
8	Hepatitis B	553
9	Measles	0
10	BMI	34
11	under-five deaths	0
12	Polio	19
13	Total expenditure	226
14	Diphtheria	19
15	HIV/AIDS	0
16	GDP	448
17	Population	652
18	thinness 1-19 years	34
19	thinness 5-9 years	34
20	Income composition of resources	167
21	Schooling	163

Fig 2. *number of null values in each feature*

Given the prevalence of null values, our initial approach was to drop these values.

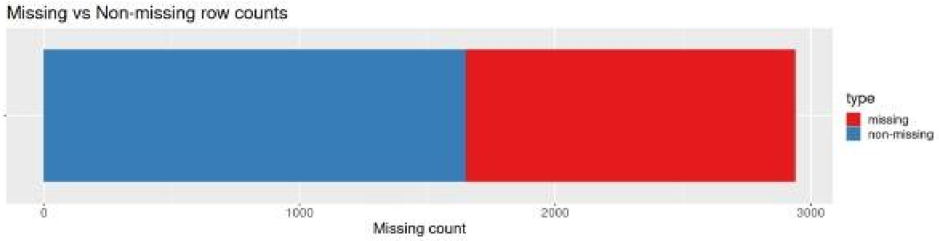


Fig 3. *number of rows with missing values vs without ones*

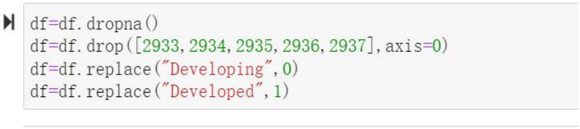
However, after accounting for the number of rows containing null values, we found that up to 44% of the data would be lost in the drop operation. We tried to propose another pre-processing method based on **constructing a linear regression model to fill the feature containing null value with another feature that has the strongest correlation with it.**

We will implement two types of pre-processing methods and observe the R2 score they get after simple linear regression to determine which method yields a dataset **more suitable for regression**.

Since the dropping method is to drop all rows containing null values, which is both common and easy to understand, we focus more on our proposed replacement method in the methodology section. The following is the principle and code implementation of the two methods.

a. Dropping method

As we have more than 40% of missing data, in order to be more consistent with the real-world processing of the dataset and to preserve the data distribution of the original dataset to the greatest extent possible, we decided to remove all nulls without filling them.



```
df=df.dropna()
df=df.drop([2933,2934,2935,2936,2937],axis=0)
df=df.replace("Developing",0)
df=df.replace("Developed",1)
```

Fig 4. *Code of Dropping Method*

b. Filling method

We have already outlined the idea of this method above, here is a detailed analysis and treatment of each feature of the dataset.

For the features such as 'Life expectancy' and 'Adult Mortality' that only have few null values, we fill with the average value of their corresponding columns.

Features filled with average:

a) Life expectancy

b) Adult Mortality

c) Polio

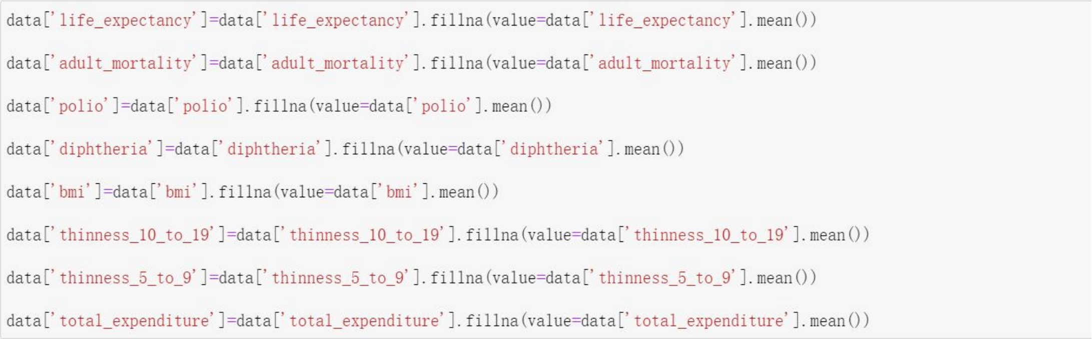
d) Diphtheria

e) BMI

f) thinness 10-19 years

g) thinness 5-9 years

h) total expenditure



```
data['life_expectancy']=data['life_expectancy'].fillna(value=data['life_expectancy'].mean())

data['adult_mortality']=data['adult_mortality'].fillna(value=data['adult_mortality'].mean())

data['polio']=data['polio'].fillna(value=data['polio'].mean())

data['diphtheria']=data['diphtheria'].fillna(value=data['diphtheria'].mean())

data['bmi']=data['bmi'].fillna(value=data['bmi'].mean())

data['thinness_10_to_19']=data['thinness_10_to_19'].fillna(value=data['thinness_10_to_19'].mean())

data['thinness_5_to_9']=data['thinness_5_to_9'].fillna(value=data['thinness_5_to_9'].mean())

data['total_expenditure']=data['total_expenditure'].fillna(value=data['total_expenditure'].mean())
```

Fig 5. *Code of filling these features by average*

As for other features with null values, we found that the dataset correlates with life expectancy by statistical data on various factors such as population, economy, infection of specific diseases, etc., so we believe that filling null values by taking crude averages does not correspond to the reality of the wide variation among country conditions. Another consideration is that there are more null values for other features and it is unlikely that these values will remain continuous, so filling them with the same values (average) doesn't make sense. The solution to above is that we try to find another column that correlates to them the most, and use them to represent these lost data.

We calculated the correlation matrix of the dataset and visualized it with a heatmap. The deeper the colour in the heatmap is, the stronger the correlation is, and we can find the most correlated features more intuitively. Meanwhile, we also use the value of the correlation matrix to double check to ensure accuracy in order not to make mistakes due to the difficulty in distinguishing colour nuances.

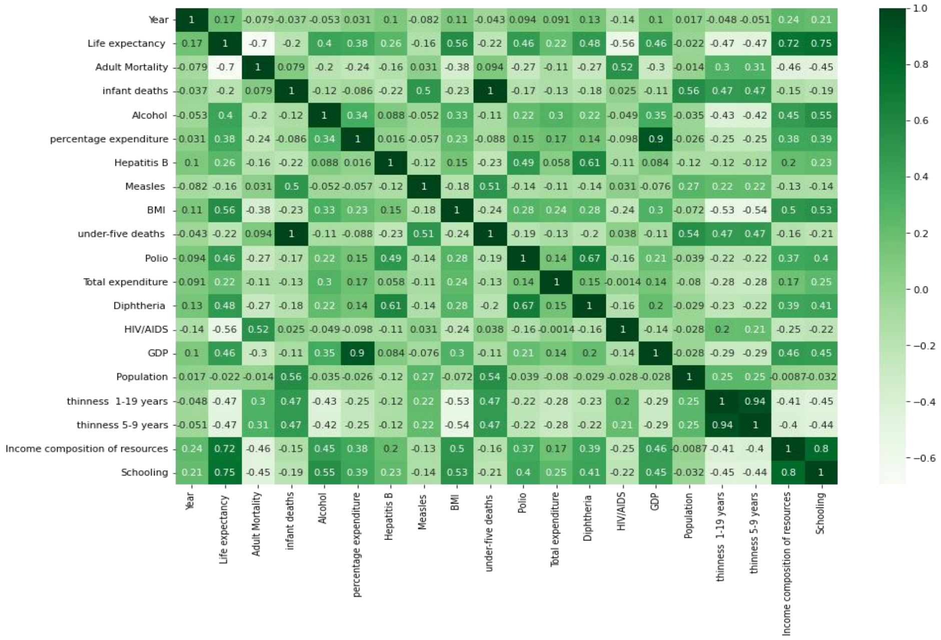


Fig 6. *Correlation Matrix of the Original Dataset*

We first select another column in the heatmap with the highest feature correlation with the one we want to fill, and if this value is promisingly high, we then replace the null value by constructing their correlation. We use the two sets of data with null values removed to train a linear regression model, and use the resulting model to calculate the missing values.

For example, from the above correlation matrix 'Alcohol' feature nicely correlates with the 'Schooling' feature. We plot a Scatterplot between them and observe the trend.

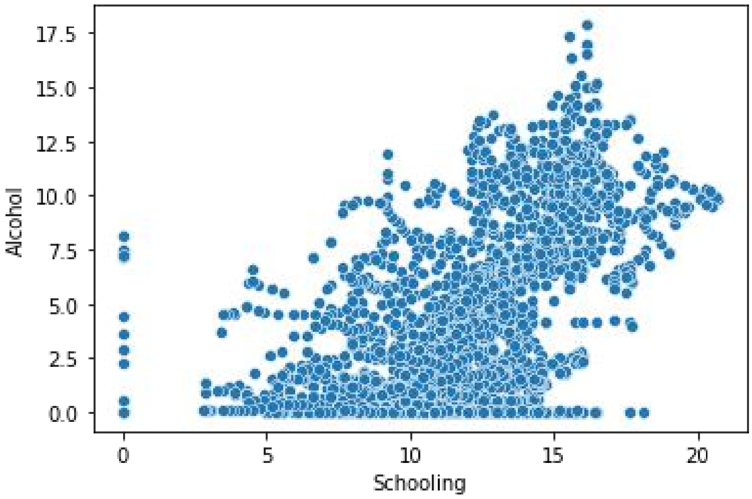


Fig 7. *relation between column 'Schooling' and 'Alcohol'*

We then drop the null values that exist and train to construct a linear regression of the two with the remaining data. The feature containing the null value is calculated by linear regression of the corresponding quantity in the other feature value, and thus be filled.

And we repeat this for imputing NaNs for all other features as well.

Here is the list of features that we imputed with other features:

- Impute 'BMI' feature with 'Life expectancy' feature.
- Impute 'Total expenditure' with 'Alcohol' feature.
- Impute 'GDP' feature with 'percentage expenditure ' feature.
- Impute 'Population' feature with 'Infant death' feature.
- Impute 'Thin 1-19' feature with 'BMI' feature.
- Impute 'Thin 5-9' feature with 'BMI' feature.
- Impute 'Schooling' feature and 'Income Composition of resources' feature with 'Life expectancy' feature.

And finally, we check again for finishing imputing Mean Value:

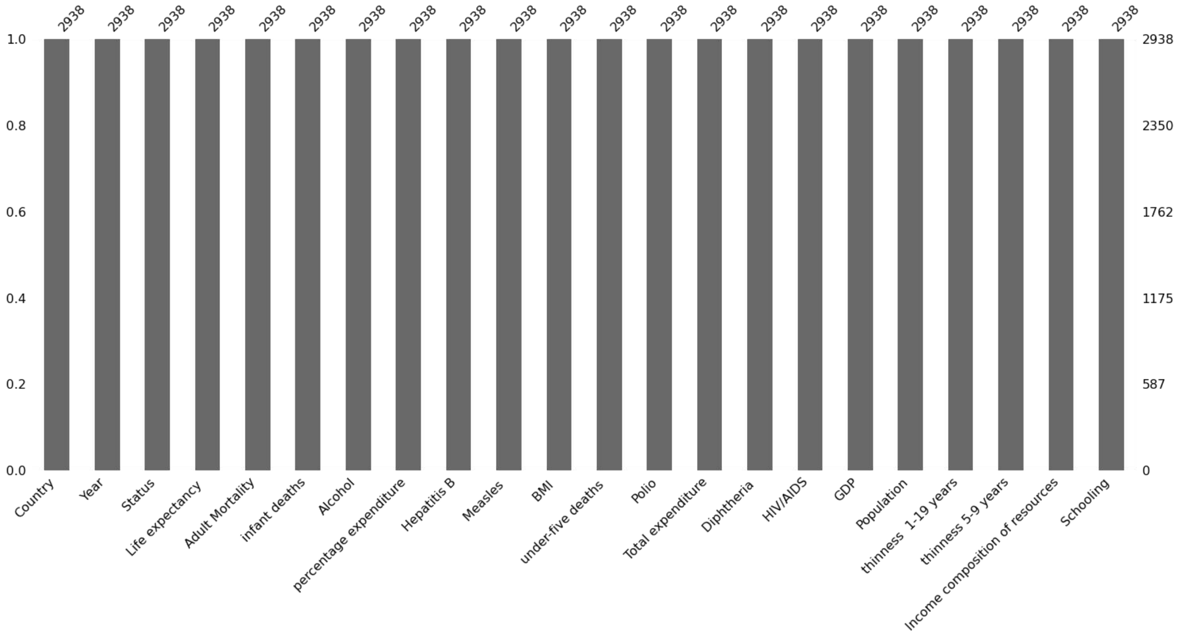


Fig 8. *number of features after pre-processing*

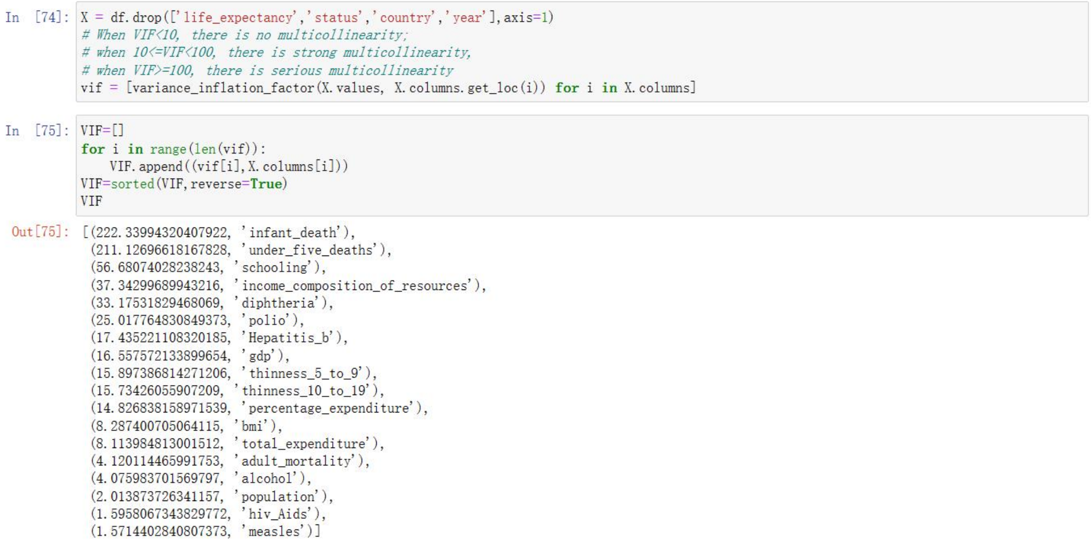
And we see all the missing values have been imputed well.

3.3. Feature Selection

Feature selection is the process by which a subset of relevant features, or variables, are selected from the original dataset. The subset both reduces the dimensionality and retains the most impactful features of the original dataset. Feature selection is the fundamental step in machine learning pipelines. When disposing of a bunch of features, an ideal situation is that the most relevant ones are selected and others are discarded. In the project, we applied three methods of feature selection: **Multicollinear feature deletion, Mutual Information and Boruta Algorithm.**

3.3.1. Multicollinear feature deletion

Multicollinearity is an occurrence that two or more predicting variables have high linear intercorrelation among them. It may lead to bad performance of regression models, as well as introduce feature redundancy in dataset. We supposed high multicollinearity is likely to exist in our dataset because we observed that some variables have high correlation with each other from correlation matrix. Therefore, we utilized Variance Inflation Factor (VIF) to measure the multicollinearity of each predicting variable with all other variables. Figure 9 shows the code implementation while figure 10 is its visualization.



```
In  [74]: X = df.drop(['life_expectancy','status','country','year'],axis=1)
          # When VIF<10, there is no multicollinearity;
          # when 10<=VIF<100, there is strong multicollinearity,
          # when VIF>=100, there is serious multicollinearity
          vif = [variance_inflation_factor(X.values, X.columns.get_loc(i)) for i in X.columns]

In  [75]: VIF=[]
          for i in range(len(vif)):
              VIF.append((vif[i],X.columns[i]))
          VIF=sorted(VIF,reverse=True)
          VIF

Out[75]: [(222.33994320407922, 'infant_death'),
           (211.12696618167828, 'under_five_deaths'),
           (56.68074028238243, 'schooling'),
           (37.34299689943216, 'income_composition_of_resources'),
           (33.17531829468069, 'diphtheria'),
           (25.017764830849373, 'polio'),
           (17.435221108320185, 'Hepatitis_b'),
           (16.557572133899654, 'gdp'),
           (15.897386814271206, 'thinness_5_to_9'),
           (15.73426055907209, 'thinness_10_to_19'),
           (14.826838158971539, 'percentage_expenditure'),
           (8.287400705064115, 'bmi'),
           (8.113984813001512, 'total_expenditure'),
           (4.120114465991753, 'adult_mortality'),
           (4.075983701569797, 'alcohol'),
           (2.013873726341157, 'population'),
           (1.5958067343829772, 'hiv_Aids'),
           (1.5714402840807373, 'measles')]
```

Fig 9. *Code implementation of VIF*

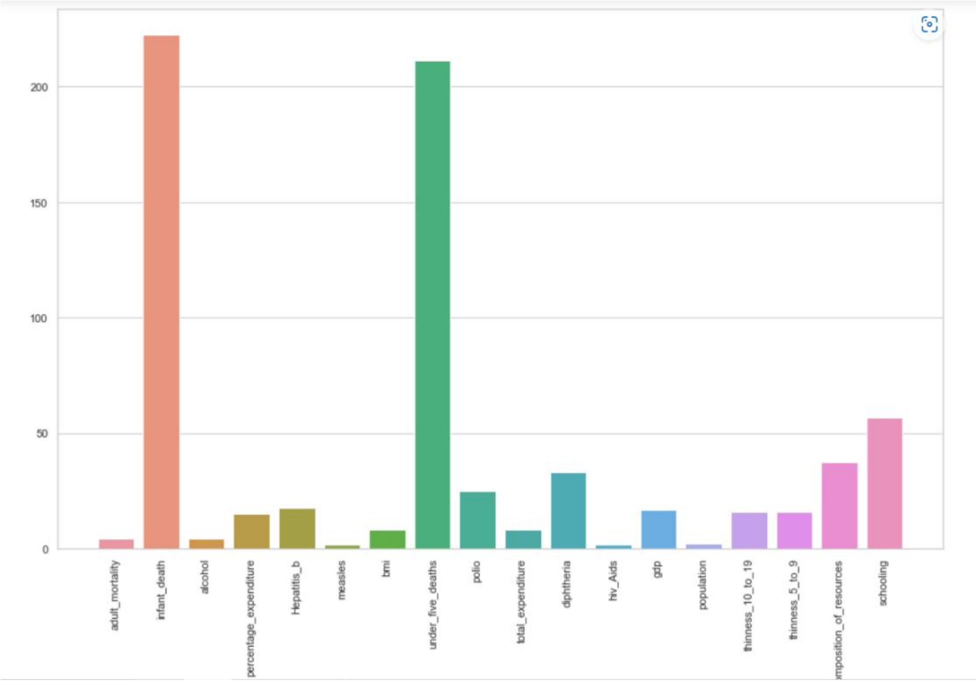


Fig 10. *visualization of VIF*

We can see that both “under_five_deaths” and “infant_death” have pretty high multicollinearity. It is easily understood because they almost measure the same thing (infants are of course under five). It can be seen from the correlation matrix as well that their correlation is 1, which indicates that they can almost be treated as the same variable. Therefore, we decided to delete one of these two from the feature set to reduce multicollinearity. We deleted “infant_death” and got a new feature set:

['status', 'adult_mortality', 'alcohol', 'percentage_expenditure','Hepatitis_b', 'measles', 'bmi', 'under_five_deaths', 'polio','total_expenditure', 'diphtheria', 'hiv_Aids', 'gdp', 'population','thinness_10_to_19','thinness_5_to_9','income_composition_of_resources', 'schooling']

3.3.2. Mutual Information

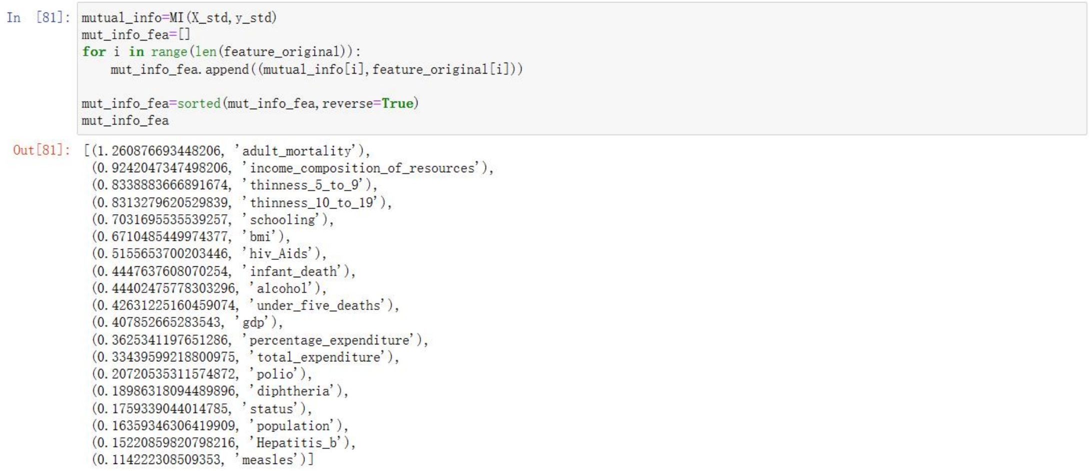
Mutual information is a quantity that measures the dependency between two random variables, that is to say, the information gained of one variable when observing another variable. Its formal definition in the case of continous random variables is:

$$\mathrm{I}(X;Y)=\int_{y}\int_{x}P_{(X,Y)}(x,y)\log(\frac{P_{(X,Y)}(x,y)}{P_{X}(x)P_{Y}(y)})$$

In the perspective of entropy, it can be expressed as $H(X) - H(X|Y)$, where $H(X)$ is marginal entropy of X and $H(X|Y)$ means the conditional entropy of X given Y. Because of the mutuality, we can get the following equation:

$$\begin{aligned}
&I(X;Y) = H(X) - H(X \mid Y) \\
&\ldots\ldots\ldots.. = H(Y) - \mathrm{H}(Y \mid X) \\
&\ldots\ldots\ldots.. = H(X) + H(Y) - H(X,Y) \\
&\ldots\ldots\ldots.. = H(X) + H(Y) - H(X,Y) \\
&\ldots\ldots\ldots.. = H(X) + H(Y) - H(X,Y) \\
&\ldots\ldots\ldots.. = H(X;Y) - H(X \mid Y) - H(Y \mid X)
\end{aligned}$$

When applying to our dataset, we computed mutual information between each predicting variable and the response variable, based on which we ranked all the predicting variables. Higher mutual information means more dependency with the response variable. The code implementation and result are shown below.



```
In  [81]: mutual_info=MI(X_std,y_std)
          mut_info_fea=[]
          for i in range(len(feature_original)):
              mut_info_fea.append((mutual_info[i],feature_original[i]))

          mut_info_fea=sorted(mut_info_fea,reverse=True)
          mut_info_fea

Out[81]: [(1.260876693448206, 'adult_mortality'),
          (0.9242047347498206, 'income_composition_of_resources'),
          (0.8338883666891674, 'thinness_5_to_9'),
          (0.8313279620529839, 'thinness_10_to_19'),
          (0.7031695535539257, 'schooling'),
          (0.6710485449974377, 'bmi'),
          (0.5155653700203446, 'hiv_Aids'),
          (0.4447637608070254, 'infant_death'),
          (0.44402475778303296, 'alcohol'),
          (0.42631225160459074, 'under_five_deaths'),
          (0.407852665283543, 'gdp'),
          (0.3625341197651286, 'percentage_expenditure'),
          (0.33439599218800975, 'total_expenditure'),
          (0.20720535311574872, 'polio'),
          (0.18986318094489896, 'diphtheria'),
          (0.1759339044014785, 'status'),
          (0.16359346306419909, 'population'),
          (0.15220859820798216, 'Hepatitis_b'),
          (0.114222308509353, 'measles')]
```

Fig 11. *Code and result of mutual information*

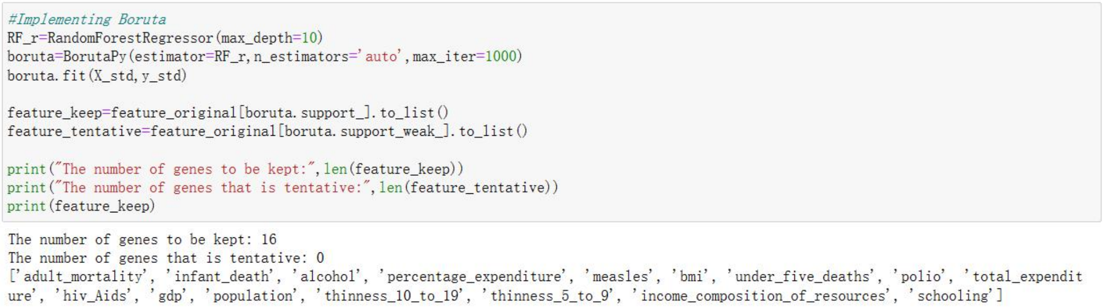
Then we selected the top ten features without “infant_death” to get the feature set:

['adult_mortality','income_composition_of_resources','thinness_5_to_9','thinness_10_to_19','schooling','bmi','hiv_Aids','alcohol','under_five_deaths','gdp'].

3.3.3. Boruta Algorithm

Boruta algorithm is a wrapper method for feature selection which uses random forest as the learning model to calculate importance of each feature. It introduces shadow features, which are a randomly shuffled version of original features, to be the criterion. A random forest regression model is fitted to both original features and shadow features, followed by calculation of importance of every feature. Subsequently, the highest feature importance in shadow features is treated as a threshold, which means every original feature will be considered to be useful if its feature importance is higher than the threshold. Otherwise, it will be regarded as useless. This process will be repeated multiple times. If a feature is useful in most time (there is also a threshold here), then it will be selected in the new feature set. And features that can't exceed the threshold will be discarded.

We implemented Boruta algorithm by its library. The code implementation and results are as follows.



```
#Implementing Boruta
RF_r=RandomForestRegressor(max_depth=10)
boruta=BorutaPy(estimator=RF_r,n_estimators='auto',max_iter=1000)
boruta.fit(X_std,y_std)

feature_keep=feature_original[boruta.support_].to_list()
feature_tentative=feature_original[boruta.support_weak_].to_list()

print("The number of genes to be kept:",len(feature_keep))
print("The number of genes that is tentative:",len(feature_tentative))
print(feature_keep)
```

```
The number of genes to be kept: 16
The number of genes that is tentative: 0
['adult_mortality', 'infant_death', 'alcohol', 'percentage_expenditure', 'measles', 'bmi', 'under_five_deaths', 'polio', 'total_expendit
ure', 'hiv_Aids', 'gdp', 'population', 'thinness_10_to_19', 'thinness_5_to_9', 'income_composition_of_resources', 'schooling']
```

Fig 12. *Code and result of Boruta algorithm*

Since the "infant_death" should be discarded, our new feature set is:

['adult_mortality', 'alcohol', 'percentage_expenditure', 'measles', 'bmi','under_five_deaths','polio', 'total_expenditure', 'hiv_Aids', 'gdp', 'population', 'thinness_10_to_19','thinness_5_to_9','income_composition_of_resources', 'schooling']

3.4. Regression models

In this section, we put our dataset, which has been finely processed above, into different regression models. Well-trained regression models can both precisely illustrate the relationship between features and dependent variables and can also be used for practical usages such as prediction, which is vitally important for our application-oriented project.

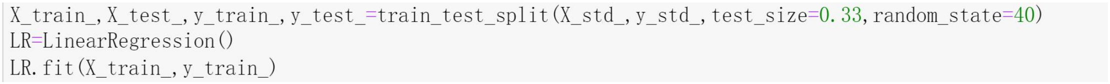
Therefore, choosing the best-performing training model is our goal in this part of the experiment.

According to Unique Variable Principle, we use a 2:1 ratio to split the training set and testing set when applying all the four models.

The following are the four regression models we applied in the project.

a) Linear regression

Linear regression fits a linear model with coefficients w = (w1, …, wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation. It is the earliest regression model and can be implemented as follow.

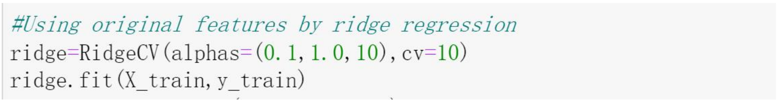


```
X_train_,X_test_,y_train_,y_test_=train_test_split(X_std_,y_std_,test_size=0.33,random_state=40)
LR=LinearRegression()
LR.fit(X_train_,y_train_)
```

Fig 13. *Code implementation of Linear Regression*

b) Ridge regression

Ridge regression (Tikhonov regularization) is a kind of biased estimation regression method dedicated to the analysis of covariance data, which is essentially a modified least squares estimation method. By giving up the unbiased nature of least squares, the regression coefficients are more realistic and reliable at the cost of losing some information and reducing accuracy. Since we have already noticed the co-linearity of the dataset in our previous work, we believe that ridge regression is also well suited to deal with this dataset.

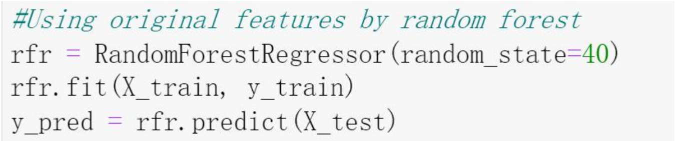


```
#Using original features by ridge regression
ridge=RidgeCV(alphas=(0.1,1.0,10),cv=10)
ridge.fit(X_train,y_train)
```

Fig 14. *Code implementation of Ridge Regression*

c) Random Forest Regression

Random Forest Regression (RFR) is an important branch of Random Forest. The Random Forest Regression model builds multiple unrelated decision trees by randomly selecting samples and features to obtain prediction results in a parallel manner. Each decision tree can produce a prediction result from the extracted samples and features, and the regression prediction result of the whole forest is obtained by combining the results of all trees and taking the average. We learned that the random forest regression algorithm works best in scenarios that require relatively low data dimensionality (tens of dimensions), while requiring high accuracy. The dataset applied in this project exactly fits this scenario, so we have high expectations for random forest regression, and hopefully, this AI based model can outperform the others.

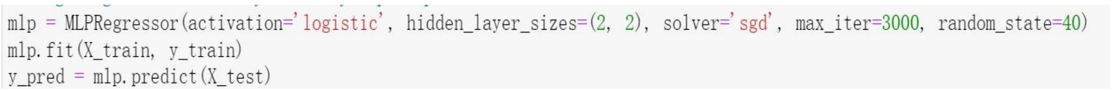


```
#Using original features by random forest
rfr = RandomForestRegressor(random_state=40)
rfr.fit(X_train, y_train)
y_pred = rfr.predict(X_test)
```

Fig 15. *Code implementation of Random Forest Regression*

d) Multilayer Perceptron

Multi-layer perceptron is another AI model commonly used for regression problems. It adds fully connected hidden layers between the output and input layers, and transforms the output of the hidden layer by an activation function, and can also update the parameters using a back propagation (BP) algorithm. In this project, we set the hidden layer to 2*2 and use logistic as the activation function.



```
mlp = MLPRegressor(activation='logistic', hidden_layer_sizes=(2, 2), solver='sgd', max_iter=3000, random_state=40)
mlp.fit(X_train, y_train)
y_pred = mlp.predict(X_test)
```

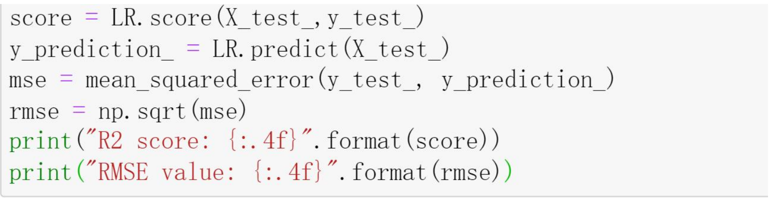
Fig 16. *Code implementation of Multilayer Perceptron*

3.5. Visualization

Although the data itself and some of the calculated values can indicate a lot of needed information, we still want to use charts and tables to provide more visual representation. We use matplotlib package, seaborn package, plotly package and other visualization methods.

3.6. Qualitative Performance Measures

We used two metrics, **RMSE and R2 score, to measure the performance of the model.** In the code of calculating these two values, we called the functions from the sklearn.linearmodel library, and we listed the code implementation here to avoid redundancy.



```
score = LR.score(X_test_,y_test_)
y_prediction_ = LR.predict(X_test_)
mse = mean_squared_error(y_test_, y_prediction_)
rmse = np.sqrt(mse)
print("R2 score: {:.4f}".format(score))
print("RMSE value: {:.4f}".format(rmse))
```

***Fig 17.** Code of calculating R2 score and RMSE*

3.6.1. Root mean square error

Root mean square error is a polynomial counting rule. It represents the error-index, which is used to compare diferent forecasting models. This is the square root of the square diferences measured between prediction and actual observation. The model having low MSE values is better. If the RMSE value is 0.00, it denotes no error value occurs in the prediction model. where n is the number of samples, P is the predicted value and A is actual value.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - A_i)^2}$$

3.6.2. R2 score

The R2 score, or coefficient of determination, reflects the proportion of the total variation in the dependent variable that can be explained by the independent variable through the regression relationship:

$$\mathrm{R}^2 = 1 - \frac{\sum_i (y_i - y_i)^2 / n}{\sum_i (y_i - \hat{y})^2 / n} = 1 - \frac{RMSE}{Var}$$

For R2 Score, it can be interpreted as using the mean as the error benchmark to see if the prediction error is greater or less than the mean benchmark error.

- R2_score = 1, the predicted and true values in the sample are exactly equal, without any error, indicating that the better the independent variable explains the dependent variable in the regression analysis.

- R2_score = 0. At this point the numerator equals the denominator and each predicted value in the sample is equal to the mean value.

4. Experiment and Analysis

4.1. Visualization Results

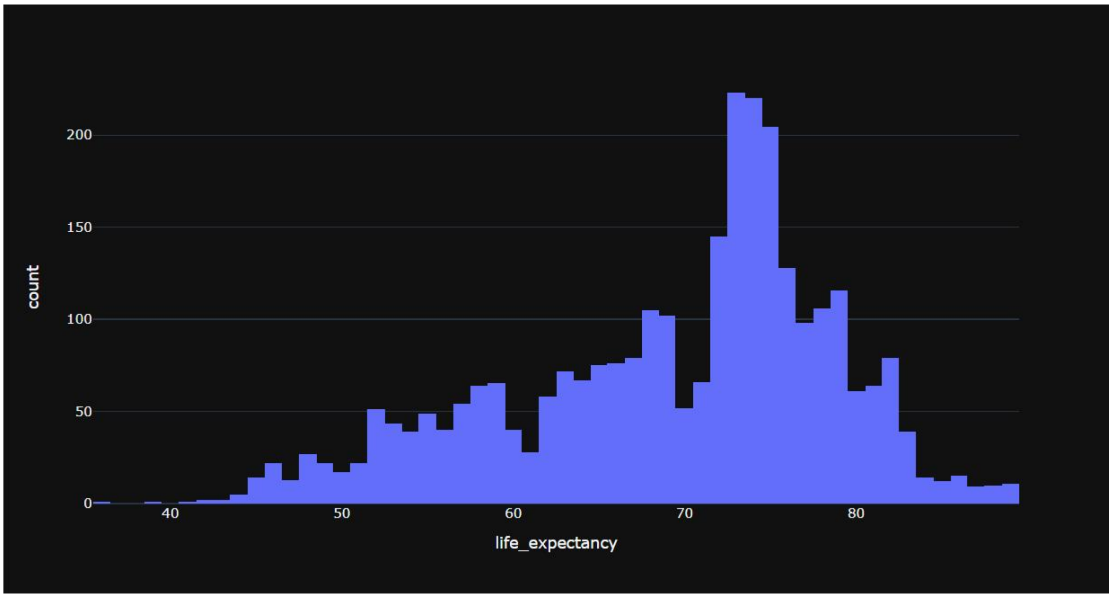
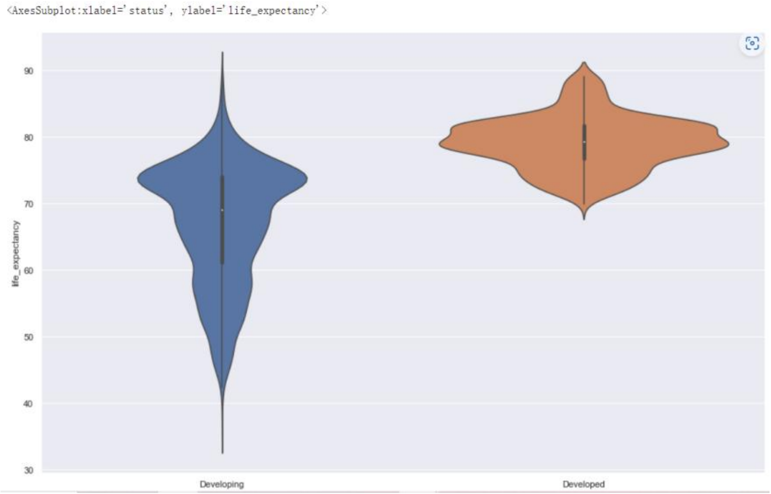


Fig 18. *Distribution of Life Expectancy according to the age*

The diagram above shows life expectancy is high between the age of 70 to 75 all over the world



```
<AxesSubplot:xlabel='status', ylabel='life_expectancy'>
```

Fig 19. *Counting the developed and developing 's life expectancy from 2000-2015*

The figure above shows developing countries have a low life expectancy and developed countries have high life expectancy all over the world. That can relate to multiple factors for example healthcare and income.

4.2. Experiments and analysis

In this section, we designed three different experiments to reach conclusions step by step. First, we compared the performance of two methods to dispose of null values by simple linear regression. Second, we applied four regression models to the original feature set to determine the best model in our dataset. Third, we fitted that best model to other three feature sets obtained by three feature selection methods to confirm our assumption and answer those questions.

4.2.1. Comparing the performance of two NaN-processing methods by Linear Regression

	Dropping method	**Filling method**
R2 score	0.8304	0.8049
RMSE value	0.4202	0.4298

As the experimental results show, the dropping method is better in terms of both RMSE and R2 score. This agrees with our expectation because we have already perceived some disadvantages in the filling method. First of all, we discovered that there exist some abnormal zero values in dataset, for example in percentage expenditure, it is impossible for a country to expend no money on health. Nevertheless, we can’t simply transform all the zero values into NaN since there are also some normal zero values. When we dropped NaN, we found that it was lucky of us that most of the abnormal zero values were dropped as well, and thus we didn’t need to care about them any more. By contrast, filling method doesn’t have such luck. Because of the complexity of zero values, we failed to deal with them in filling method, causing more noise to remain in the dataset.

Moreover, filling method is not that reasonable at some time. For example, we filled with mean in the column life expectancy. It can be imagined that there is a undeveloped country with low life expectancy around 40. And since there are some null values in this country, we filled them with the whole average 69, which is much larger than what it should be. Therefore, filling method may introduce more outliers into our dataset.

On account of these drawbacks of filling method, there is no wonder that it can't beat simple dropping method. Hence, we utilized the dataset preprocessed by dropping method to conduct following experiments.

4.2.2. Comparing the performance of applying different regression models

	Linear Regression	**Ridge Regression**	**Random Forest**	**Multilayer Perceptron**
R2 score	0.8304	0.8304	0.9590	0.8008
RMSE value	0.4202	0.4202	0.2066	0.4553

Based on the results we found that **random forest regression** had the best performance in terms of both R2 score and RMSE. Thus, we selected random forest to continue our experiments. Besides, experimental results show that ridge regression has same performance with linear regression. It may indicate that overfitting problem is trivial here, so regularization in ridge regression doesn't enhance the performance.

4.2.3. Using Random Forest to make prediction on other three feature sets

	Original feature set	**Deleting high multicollinear features**	**Selected by Mutual Information**	**Selected by Boruta**
R2 score	0.9590	0.9587	0.9609	0.9597

RMSE value	0.2066	0.2072	0.2018	0.2047

The table represents R2 score and RMSE of applying random forest regression to different feature sets. It can be seen that all the feature sets selected by three methods have similar performance with original one. Therefore, **we can use fewer features to make predictions without reducing accuracy.** That means, features discarded by these selection methods are indeed useless in prediction, that is to say, they have little impact on life expectancy. The three features discarded by Boruta algorithm are "Diphtheria", "Hepatitis_b" and "Status", which also has little mutual information with life expectancy, **0.19**, **0.15** and **0.18** respectively. We conclude that immunization coverage of diphtheria and hepatitis_b have little effect on life expectancy. "Status" is left out because we suppose there may be other reasons why it is discarded, like it is the only discrete feature in continuous features.

Moreover, we conclude that **adult mortality** and **income composition of resources** have highest influence on life expectancy since they have both the highest mutual information and feature importance in boruta. That's our answer to the core question proposed previously. Since adult mortality is too general and can be affected by too much factors, we suggest that governments should be devoted to improving income structures of residents in order to extend expected life. Below is the relationship between income composition and life expectancy. We can clearly see that they have high intercorrelation.

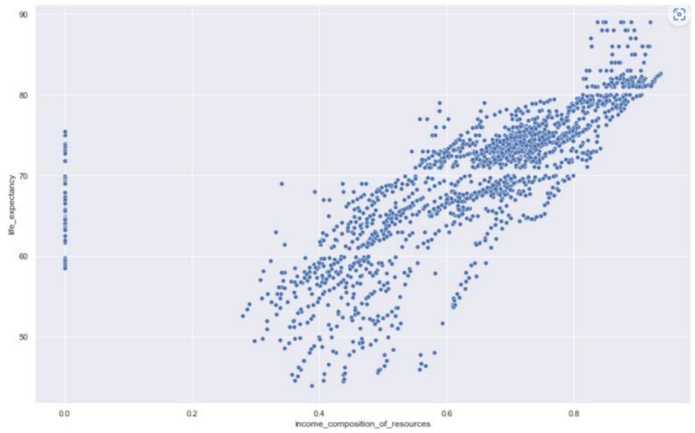


Fig 20. Scatter plot of income composition and life expectancy

5. Extra and future Work

Also, we implement the comparisons of two developing countries **Malaysia(MY) and China(CN)'**s life expectancy, observing some affected factors and do Time series predicting about MY.

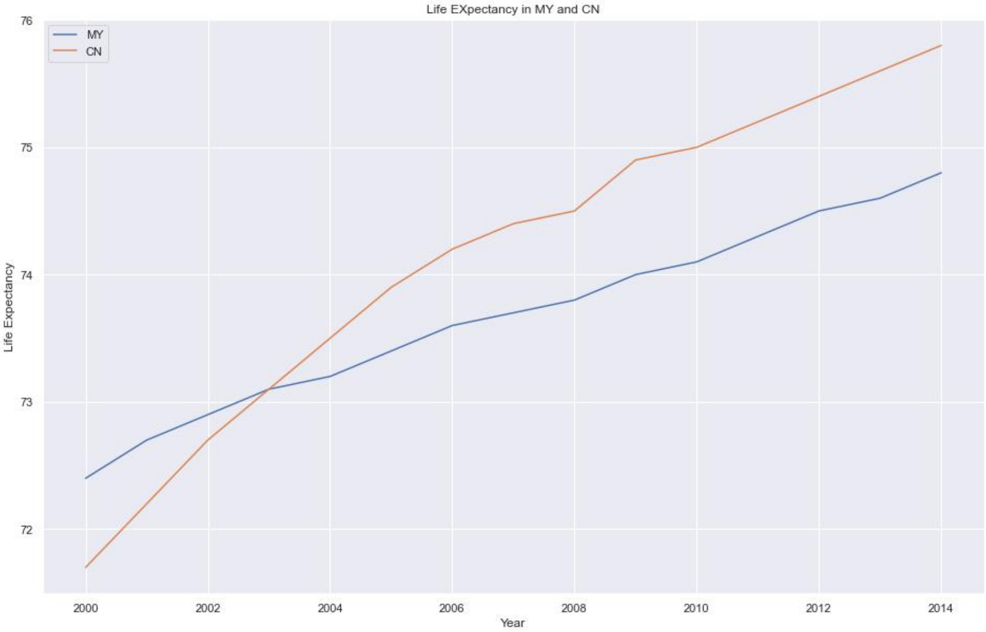


Fig 21. The comparison of MY and CN's life expectancy

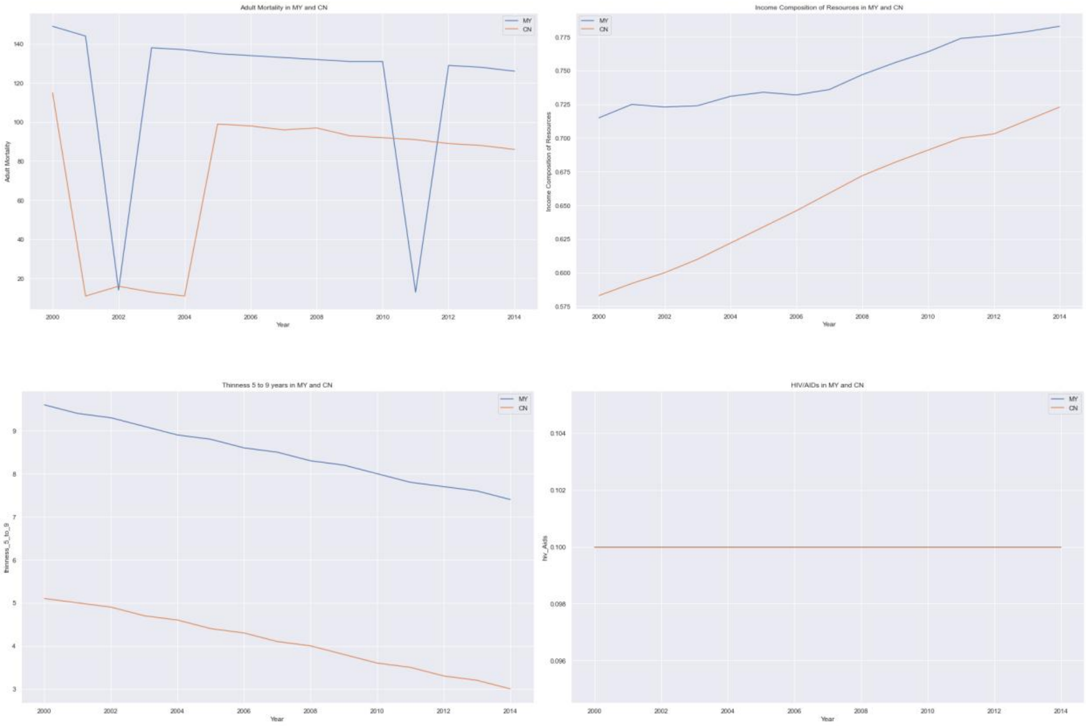
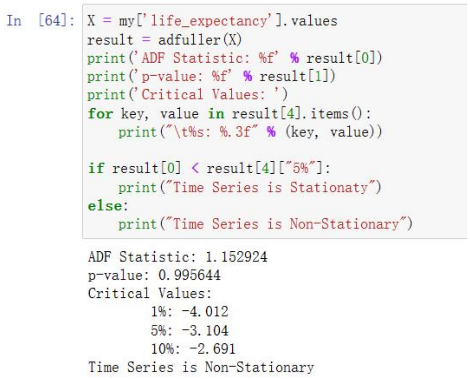


Fig 22. *The trend of comparison the MY and CN in* ***Adult mortality (top left)****,* ***income composition (top right)****,* ***child malnutrition rate (bottom left)****,* ***HIV/AIDS mortality (bottom right)***

For **Time series analysis** part, due to its data is Non-stationary(after **ADF test** and see its **ACF/PACF plot**), so we also implement Linear Regression on AutoRegression. We can see that our predicted value is always **higher** than actual value of life expectancy in MY, we guess it may occurs some **"Black Swan" event** in MY between 2000-2015 years.



```
In  [64]: X = my['life_expectancy'].values
          result = adfuller(X)
          print('ADF Statistic: %f' % result[0])
          print('p-value: %f' % result[1])
          print('Critical Values: ')
          for key, value in result[4].items():
              print("\t%s: %.3f" % (key, value))

          if result[0] < result[4]["5%"]:
              print("Time Series is Stationaty")
          else:
              print("Time Series is Non-Stationary")

          ADF Statistic: 1.152924
          p-value: 0.995644
          Critical Values:
                  1%: -4.012
                  5%: -3.104
                  10%: -2.691
          Time Series is Non-Stationary
```

Fig 23. *The* ***ADF Test*** *for our data to put in Time series Analysis*

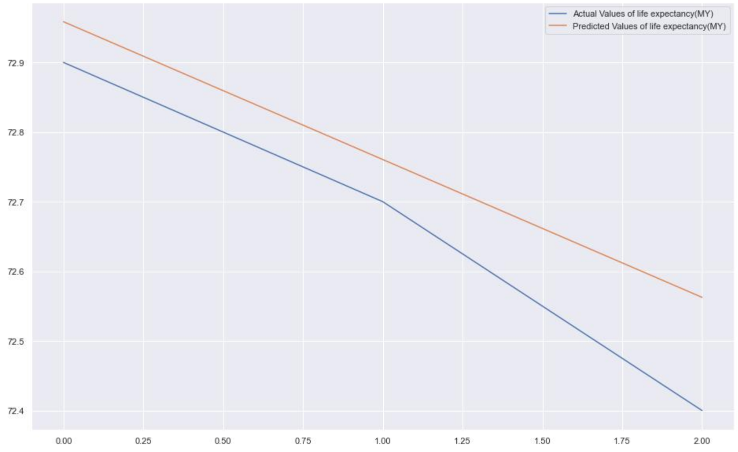


Fig 24. *The Predicted value vs. Actual value of life expectancy in Time series Analysis*

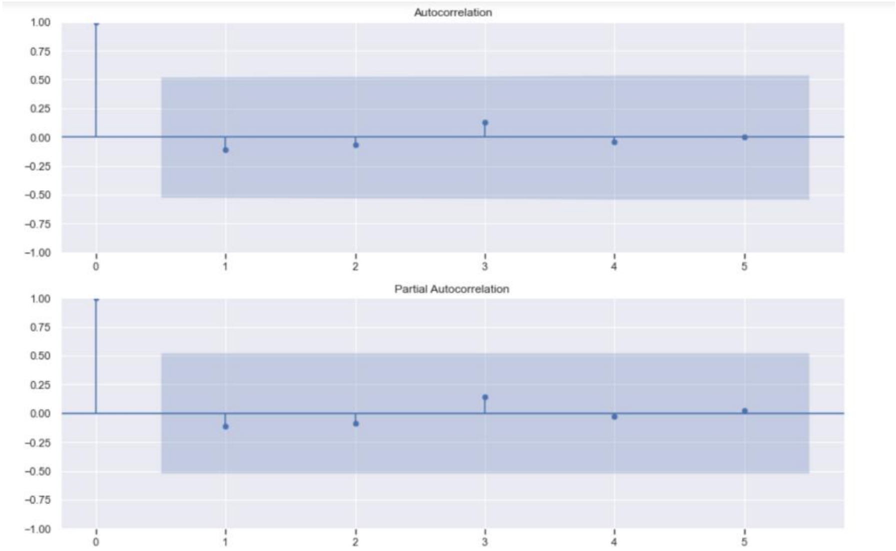


Fig 25. *ACF/PACF plot in Time series analysis*

6. Conclusion

In this project, we mainly focused on exploring the relationship between predicting variables and life expectancy. We proposed multiple questions in the introduction and the core one is what are the predicting variables most affecting life expectancy? Then we utilized three feature selection methods to obtain three different feature sets, which can be regarded as the assumed best features. Finally, after the experiments used to compare different preprocessing methods and regression models, we conducted experiments to confirm our assumptions and answered the core question and did the extra work about comparisons life expectancy between MY and CN and simple Time series analysis. Due to time constraints, our project can only be limited to this, but in the future, we will continue to explore these issues, of course not limited to these issues, and become explorers in the field of statistics

Reference

1. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. John Wiley & Sons.

2. Marquardt, D. W., & Snee, R. D. (1975). Ridge regression in practice. *The American Statistician*, *29*(1), 3-20.Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. *Korean journal of anesthesiology*, *70*(4), 407-411.

3. Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

4. Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, *32*(14-15), 2627-2636.

5. Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in medicine*, *8*(8), 907-925.

6. Ebenstein, A., Fan, M., Greenstone, M., He, G., Yin, P., & Zhou, M. (2015). Growth, pollution, and life expectancy: China from 1991-2012. *American Economic Review*, *105*(5), 226-31.

7. Ma S. (1989). Analysis on the factors of life expectancy. Popul Res ;13:14-8.

8. Sen A. (1999). Development as freedom. New York: Oxford University Press;

9. Yang F. (2015). Equalization of medical and health services and life expectancy. China Popul Newsl August 31st:1-2.

10. Ming Y, Dong Z.(2010). Life expectancy of China's population analysis of the impact of factors. Theor Res :47-50.

11. Heuvel vd, A W, Marinela O. How important are health care expenditures for life expectancy? A comparative, European analysis.J Am Med Dir Assoc 2017;18:276.-9-.-12.

12. Chen C, Zhou T, Chen G.(1997) Analysis on influencing factors of life expectancy.J Math Med ;10:71-2.

13. Ling CH(2017), Ahmed K, Muhamad R, Shahbaz M, Loganathan N.Testing the social cost of rapid economic development in Malaysia: the effect of trade on life expectancy. Soc Indicat Res:130:1005-23.

14. Gulis G.(2000) Life expectancy as an indicator of environmental health. Eur J Epidemiol 16:161-5.

15. Qi L(2008). Interrelationship between growth, environment and population health: an empirical analysis based on China's provincial data. China Popul Resour Environ :18:169-73

16. Feachem R. Health decline in Eastern Europe. Nature1994;367:313 4.

17. Chen Y, Ebensteinb A, Greenstone M, Li H(2013) Evidence on the impact of sustained exposure to air pollution on lifeexpectancy from China's Huai River policy. Proc Natl Acad Sci USA:110:12936 41.

18. Kohavi, R., & Sommerfield, D. (1995, August). Feature Subset Selection Using the Wrapper Method: Overfitting and Dynamic Search Space Topology. In *KDD* (pp. 192-197).

19. Lal, T. N., Chapelle, O., Weston, J., & Elisseeff, A. (2006). Embedded methods. In *Feature extraction* (pp. 137-165). Springer, Berlin, Heidelberg.

20. Duch, W. (2006). Filter methods. In *Feature Extraction* (pp. 89-117). Springer, Berlin, Heidelberg.

21. Kent, J. T. (1983). Information gain and a general measure of correlation. *Biometrika*, *70*(1), 163-173.

22. Wilson, E. B., & Hilferty, M. M. (1931). The distribution of chi-square. *Proceedings of the National Academy of Sciences*, *17*(12), 684-688.

23. Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical review E*, *69*(6), 066138.

24. Kononenko, I. (1994, April). Estimating attributes: Analysis and extensions of RELIEF. In *European conference on machine learning* (pp. 171-182). Springer, Berlin, Heidelberg.

25. Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of statistical software*, *36*, 1-13.

26. Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418-1429.

27. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301-320.

28. Lewis, R. J. (2000, May). An introduction to classification and regression tree (CART) analysis. In *Annual meeting of the society for academic emergency medicine in San Francisco, California* (Vol. 14). California: Department of Emergency Medicine Harbor-UCLA Medical Center Torrance.

29. Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

30. Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., & Chen, K. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, *1*(4), 1-4.

MARKING RUBRICS

		Score and Descriptors				
No.		**Poor**	**Average**	**Excellent**	**Weight (%)**	**Mark**
Component 1: Project Development						
		0-2	**3**	**4-5**	**5**	
1	Use of Data Set	No application pre-processing data.Only use the raw data.	Apply partical complete of data pre-processing.	Apply complete data pre-processing methods on data sets.		
		0-2	**3**	**4-5**	**5**	
1	Code Quality	Design codes that are poorly structured. Not fully functional. Not documented.	Codes are sufficiently documented and mostly functional. Satisfactorily structured.	Codes are fully functional and well structured and documented.		
		0-2	**3**	**4-5**	**5**	
2	Method Functionalities	Incomplete development of the method.	Complete development of method.	Completed\ enhanced the developed methods.		
		0-2	**3-5**	**6-10**	**10**	
3	Performance Evaluation	Lack evaluation of performance and analysis.	Partially complete performance evaluation.	Provide all analysis and performance evaluation.		
				Subtotal	**25**	
Component 2: Project Report + Project Demonstration						
		0 - 6	**7 - 10**	**11-15**	**15**	
1	Project Report	Poor writing quality Poor or no formatting / presentation Lack of discussion for statistical method's results.	Satisfactory writing quality, grammar and flow. Substantial content on the statistical methods.	Good writing quality, grammar and flow Well formatted and good presentation Demonstrate excellent experimental analysis of statistical methods.		
		0-2	**3**	**4-5**	**5**	
1	Presentation	Poor quality slides Poor time management Speech that is unclear.	Satisfactory quality slides Speech that is satisfactory and	High quality slides Good time management Speech that is clear		

			understandable.	and impactful.		
		0-2	**3**	**4-5**	**5**	
2	Demonstration	Poor demonstration that is unclear of implementation methods.	Satisfactory demonstration of implementation of methods.	Excellent demonstration is fully functioning and logical.		
				Subtotal	**25**	
				Grand Total	**50**	

Note to students: Please attach this appendix together with the submission of coursework