# Data Engineering Master Project: Life Expectancy Dataset

Citizen Data Science Program – Level 2

**Guidelines:**

1. Answer each question with a clear and concise explanation of the Operator you used. Explain why you chose that particular operator and how it helps.
2. Outline the steps you followed to arrive at the solution.
3. Provide a screenshot of your process that demonstrated the relevant portion of your answer.
4. Name the document as the {project name - your name} before submitting it.
5. Ensure that you include both your answers word document and RapidMiner process as attachments in a reply to the person who sent you the assessment.
6. Write your name and email at the bottom of this page

## Project Description

This project consists of 13 questions related to your process in RapidMiner of the Life Expectancy dataset. The questions are designed to assess your understanding of the key concepts and your ability to apply them in practical scenarios.

You will work on the Life Expectancy dataset using RapidMiner to predict life expectancy values for people. The dataset provided by the World Health Organization (WHO) encompasses a comprehensive collection of socio-economic and health-related indicators alongside corresponding life expectancy data for multiple countries spanning several years.

The dataset comprises various attributes, including but not limited to:

- Country: Name of the country under observation.
- Year: The year of data collection.
- Status: Indicates whether the country is classified as developed or developing.
- Life Expectancy: The average life expectancy at birth for a given country-year combination.
- Economic Indicators: Gross Domestic Product (GDP), GDP per capita, income composition of resources, and expenditure on healthcare.
- Health Indicators: Immunization coverage for various diseases, mortality rates (e.g., infant mortality rate, under-five mortality rate), percentage of the population with access to improved sanitation facilities and clean drinking water.

**Name: Abdullah Mohammad Al Talaq**

**Email: AlTalaqA@sabic.com**

## Section 1: Data Loading and Exploration

1. Import the Life Expectancy dataset into RapidMiner and generate summary statistics for numerical variables (mean, standard deviation, min, max, etc.) along with the distribution of each variable within the dataset.

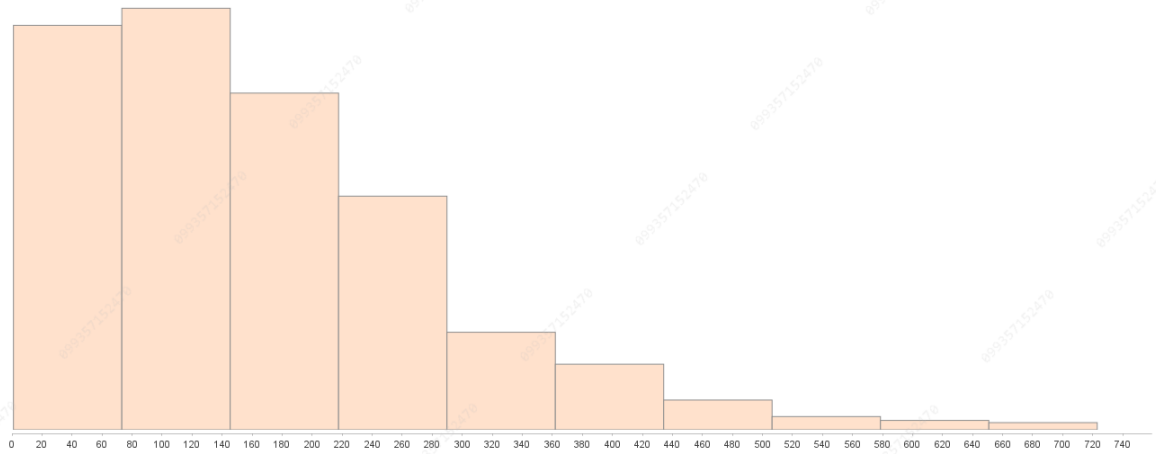
### Adult Mortality

**Summary**

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 14.47%
Stability: 1.16%
Valid: 84.38%

**Distribution**




**Statistics**

| Name | Value |
|---|---|
| Minimum | 1 |
| Maximum | 723 |
| Average | 164.817 |
| Standard Deviation | 124.434 |

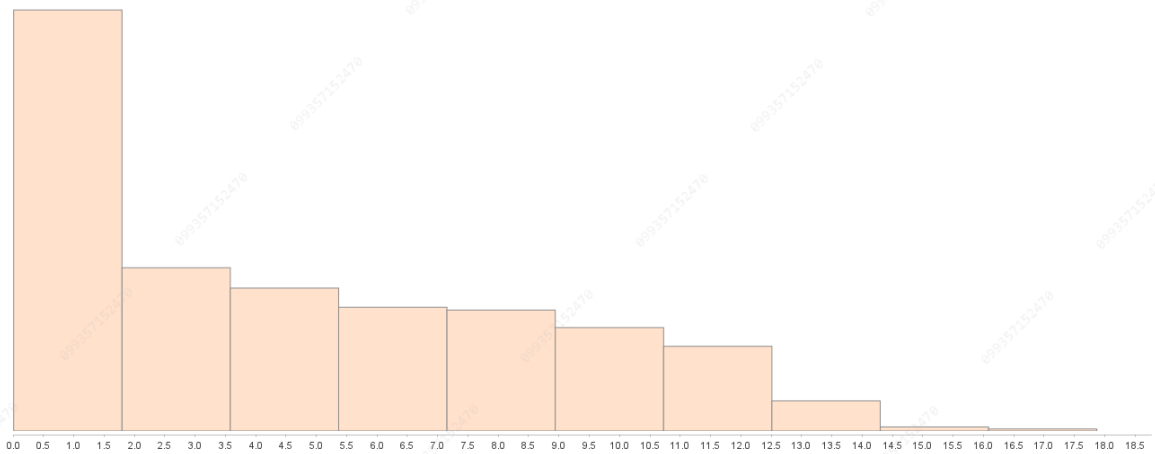
### Alcohol

**Summary**

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.41%
Stability: 10.28%
Valid: 89.31%

**Distribution**




**Statistics**

| Name | Value |
|---|---|
| Minimum | 0.010 |
| Maximum | 17.870 |
| Average | 4.607 |
| Standard Deviation | 4.045 |

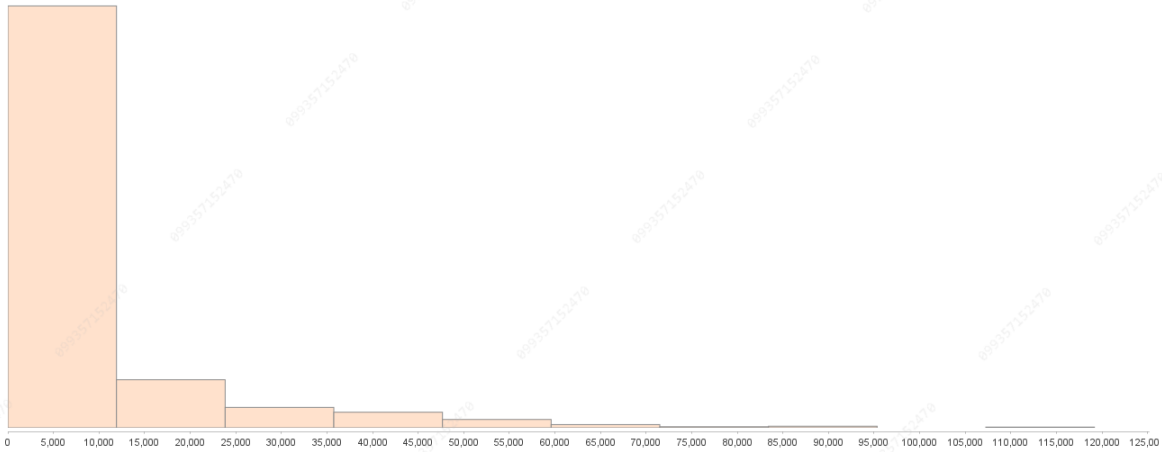
## GDP

### Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.00%
Stability: 1.67%
Valid: 98.33%

### Distribution




### Statistics

| Name | Value |
|---|---|
| Minimum | 1.681 |
| Maximum | 119172.742 |
| Average | 7475.594 |
| Standard Deviation | 13728.462 |

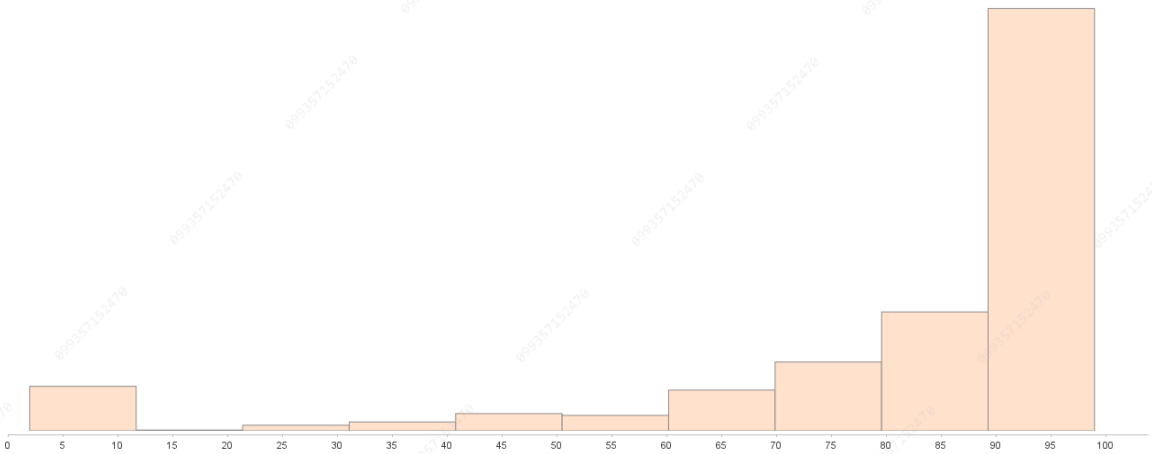
## Diphtheria

### Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 2.76%
Stability: 11.91%
Valid: 85.33%

### Distribution




### Statistics

| Name | Value |
|---|---|
| Minimum | 2 |
| Maximum | 99 |
| Average | 82.075 |
| Standard Deviation | 23.917 |

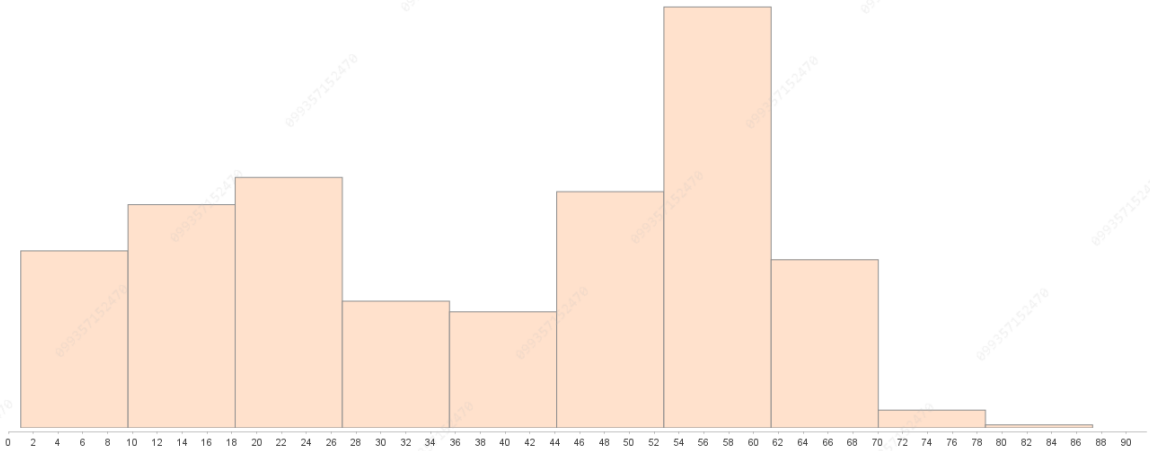
## BMI

**Summary**

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 2.04%
Stability: 0.75%
Valid: 97.21%

**Distribution**




**Statistics**

| Name | Value |
|---|---|
| Minimum | 1 |
| Maximum | 87.300 |
| Average | 38.020 |
| Standard Deviation | 20.175 |

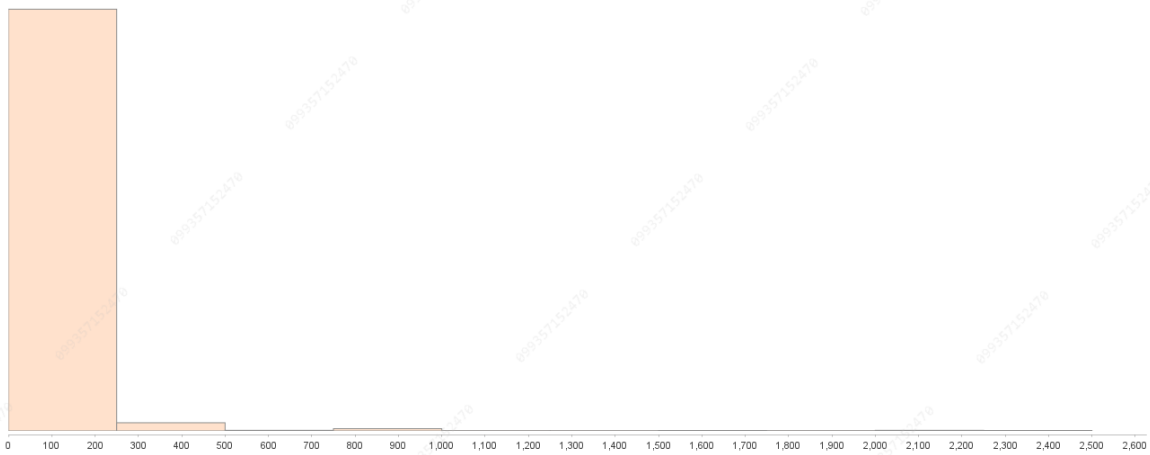
## under-five deaths

**Summary**

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 8.58%
Stability: 26.72%
Valid: 64.70%

**Distribution**




**Statistics**

| Name | Value |
|---|---|
| Minimum | 0 |
| Maximum | 2500 |
| Average | 42.036 |
| Standard Deviation | 160.446 |

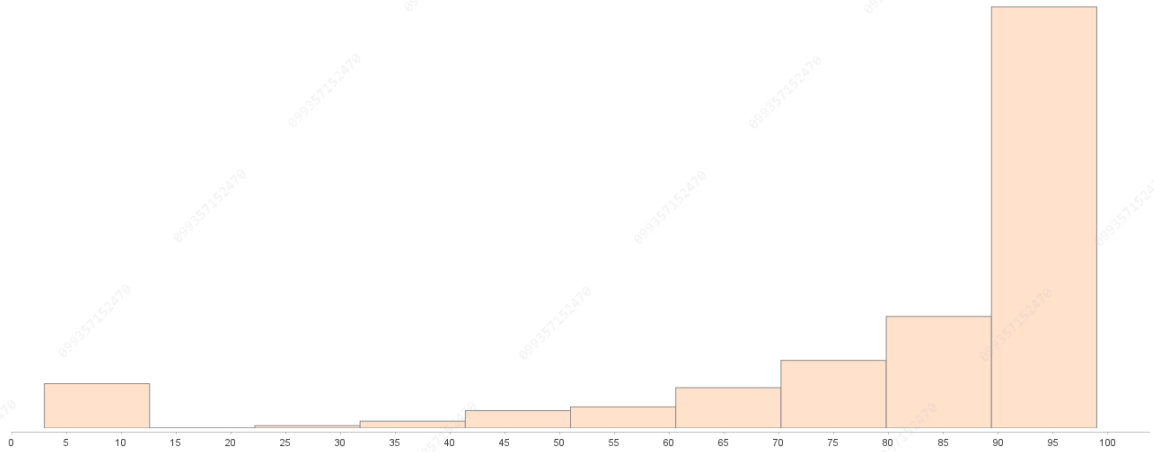
## Polio

### Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 2.48%
Stability: 12.80%
Valid: 84.72%

### Distribution




### Statistics

| Name | Value |
|---|---|
| Minimum | 3 |
| Maximum | 99 |
| Average | 82.308 |
| Standard Deviation | 23.637 |

## Measles

### Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 32.61%
Stability: 33.46%
Valid: 33.93%

### Distribution


### Statistics

| Name | Value |
|---|---|
| Minimum | 0 |
| Maximum | 212183 |
| Average | 2419.592 |
| Standard Deviation | 11467.272 |

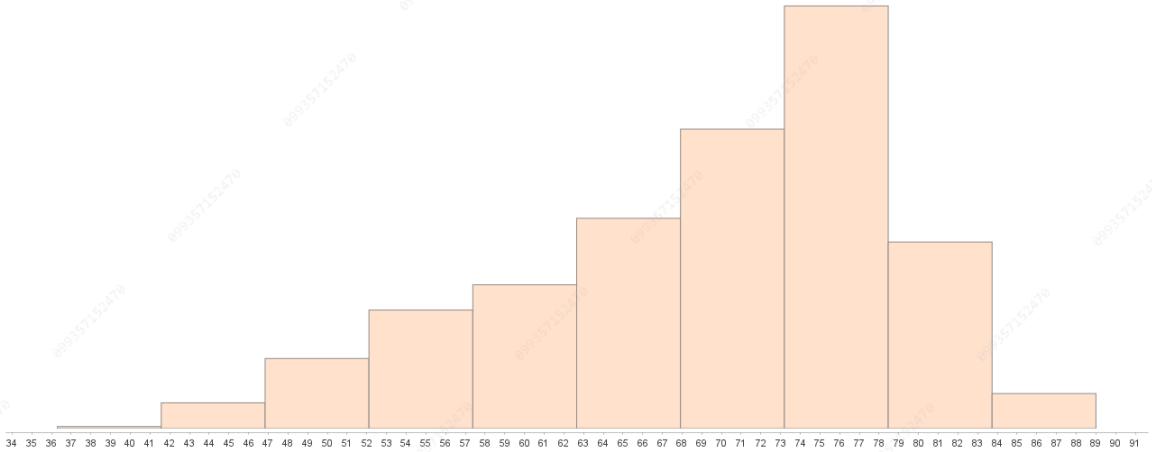
## Life expectancy

**Summary**

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 1.57%
Stability: 1.53%
Valid: 96.90%

**Distribution**




**Statistics**

| Name | Value |
|---|---|
| Minimum | 36.300 |
| Maximum | 89 |
| Average | 69.196 |
| Standard Deviation | 9.537 |

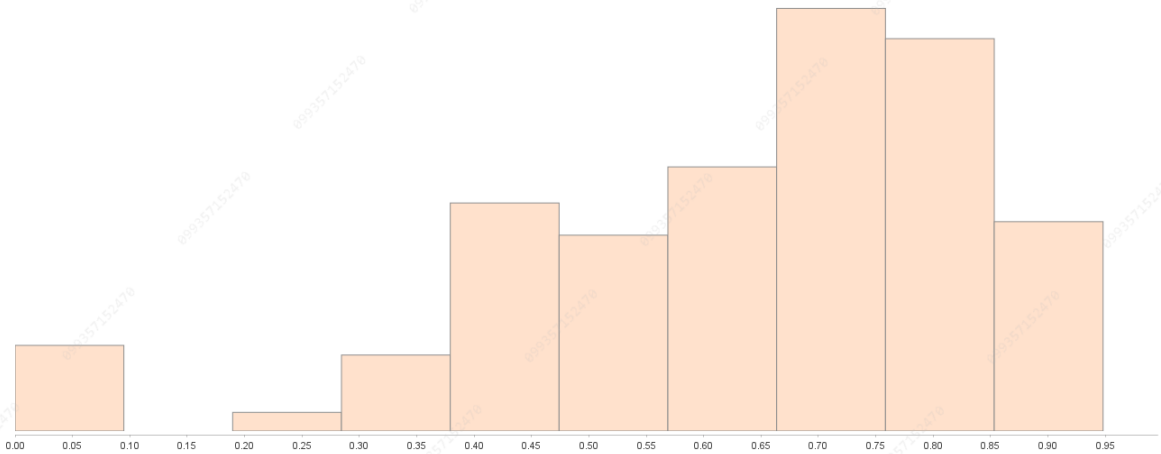
## Income composition of resources

**Summary**

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.03%
Stability: 4.53%
Valid: 95.44%

**Distribution**




**Statistics**

| Name | Value |
|---|---|
| Minimum | 0 |
| Maximum | 0.948 |
| Average | 0.631 |
| Standard Deviation | 0.211 |

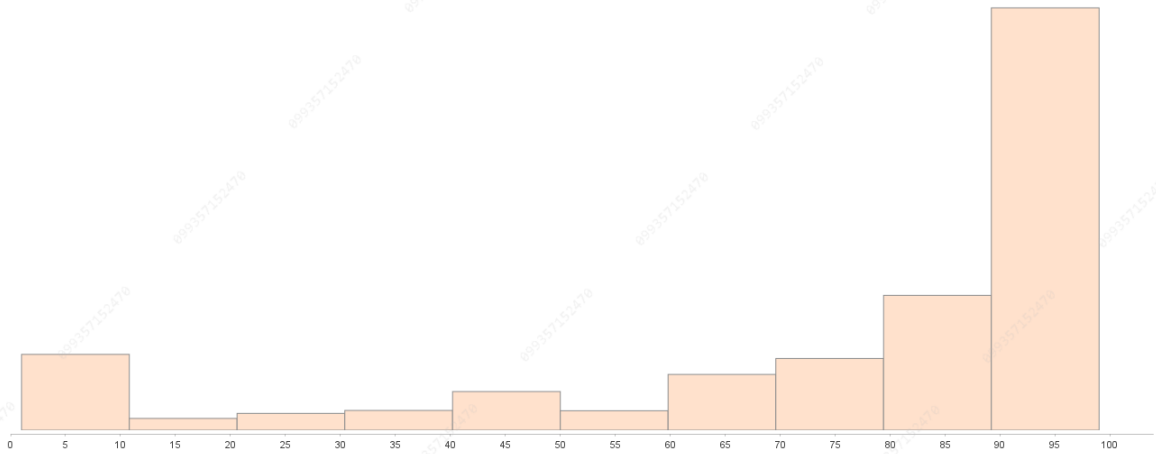
## Hepatitis B

Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 2.96%
Stability: 8.71%
Valid: 88.33%

Distribution




Statistics

| Name | Value |
|---|---|
| Minimum | 1 |
| Maximum | 99 |
| Average | 75.684 |
| Standard Deviation | 28.852 |

## HIV/AIDS

Summary

Number

Missing: 0.00%
Infinite: 0.00%
ID-ness: 0.54%
Stability: 60.62%
Valid: 38.84%

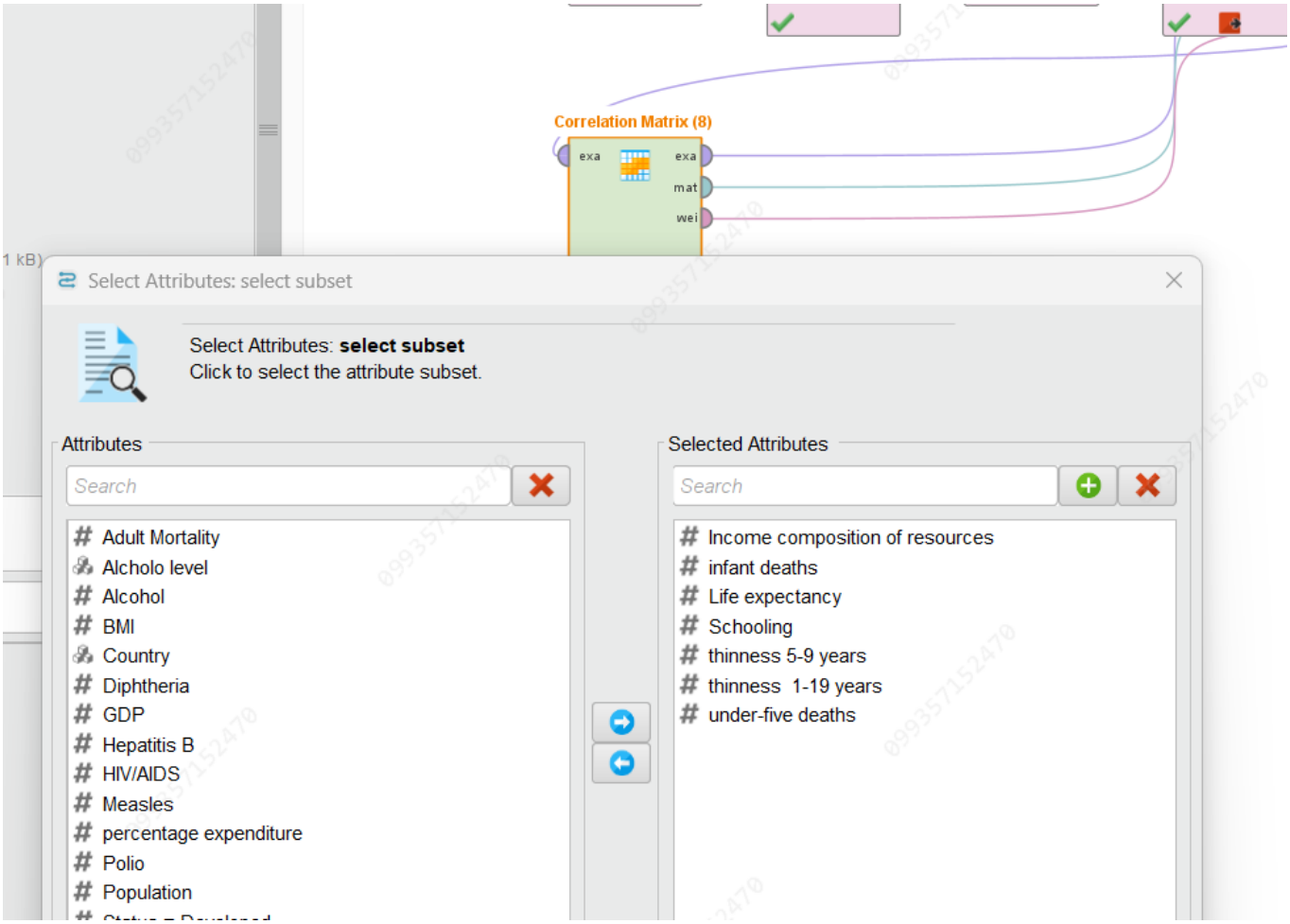
Distribution


Statistics

| Name | Value |
|---|---|
| Minimum | 0.100 |
| Maximum | 50.600 |
| Average | 1.742 |
| Standard Deviation | 5.078 |

2. Explore different relations between the attributes within the dataset that might help in predicting the Life Expectancy and specify the type of the relation.

| First Attribute | Second Attribute | Correlation |
|---|---|---|
| infant deaths | under-five deaths | 0.997 |
| thinness 1-19 years | thinness 5-9 years | 0.946 |
| Income composition of resources | Schooling | 0.850 |
| Life expectancy | Schooling | 0.795 |
| Life expectancy | Income composition of resources | 0.706 |

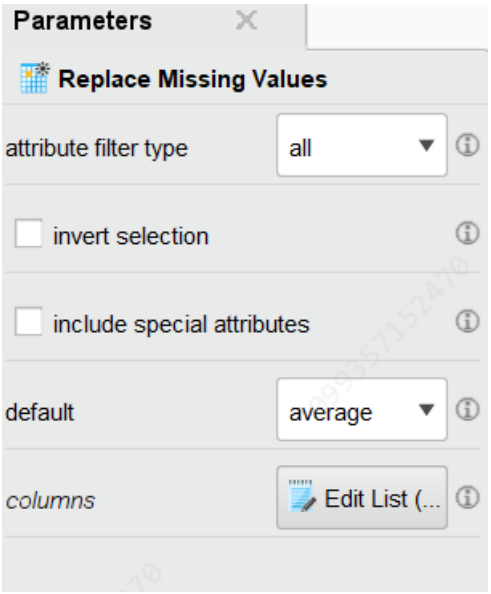



used correlation matrix.

## Section 2: Data Preprocessing

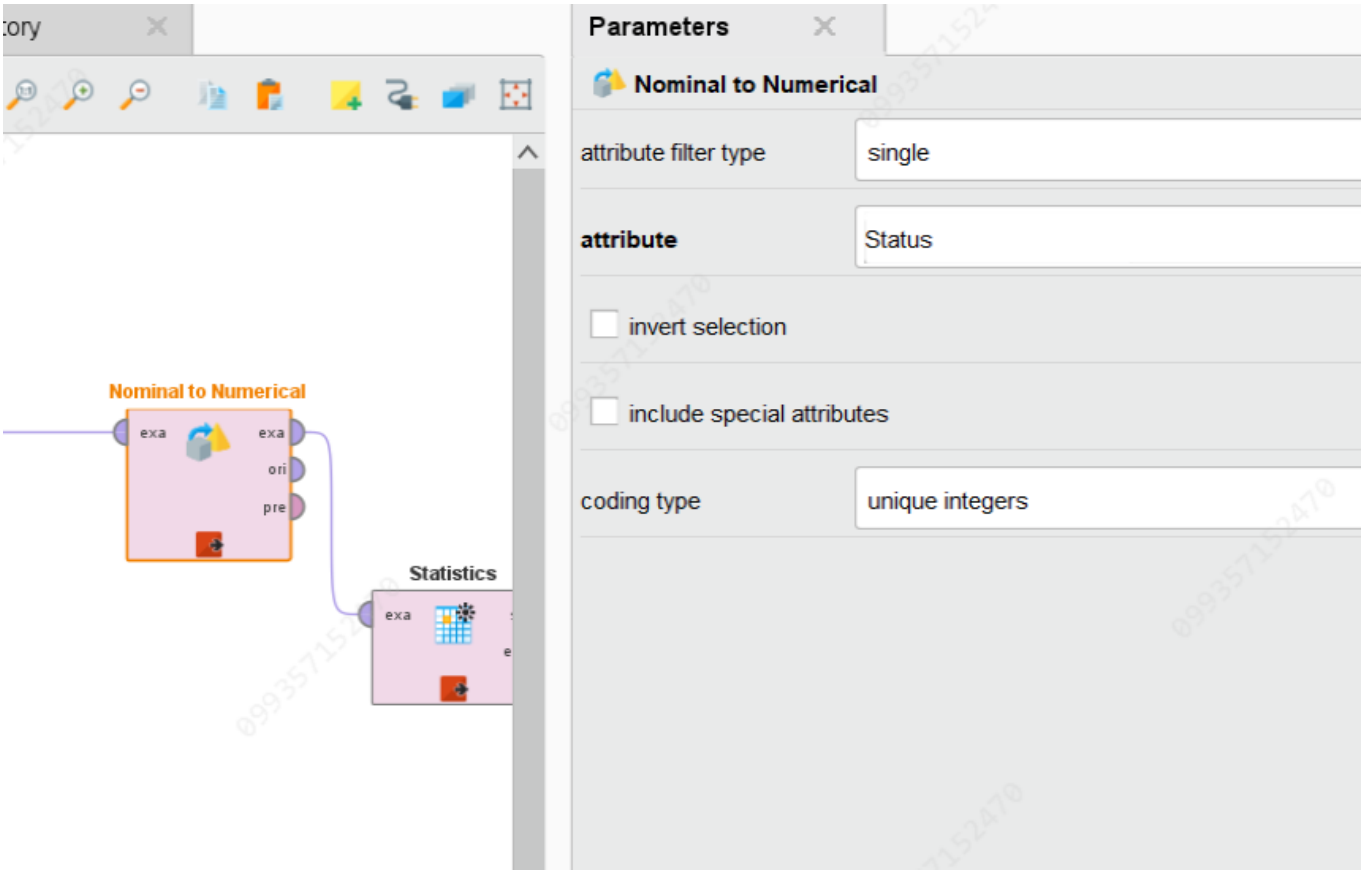
1. Identify if there is any missing value in the dataset and handle them.

| Name | Type | Missing |
|---|---|---|
| Status | Polynominal | 0 |
| Life expectancy | Real | 10 |
| Adult Mortality | Integer | 10 |
| infant deaths | Integer | 0 |
| Alcohol | Real | 194 |
| percentage expenditure | Real | 0 |
| Hepatitis B | Integer | 553 |
| Measles | Integer | 0 |
| BMI | Real | 34 |
| under-five deaths | Integer | 0 |
| Polio | Integer | 19 |
| Total expenditure | Real | 226 |
| Diphtheria | Integer | 19 |
| HIV/AIDS | Real | 0 |
| GDP | Real | 448 |
| Population | Integer | 652 |
| thinness 1-19 years | Real | 34 |
| thinness 5-9 years | Real | 34 |
| Income composition of resourc... | Real | 167 |




| Name | Type | Missing |
|---|---|---|
| Country | Nominal | 0 |
| Year | Real | 0 |
| Status | Nominal | 0 |
| Life expectancy | Real | 0 |
| Adult Mortality | Real | 0 |
| infant deaths | Real | 0 |
| Alcohol | Real | 0 |
| percentage expenditure | Real | 0 |
| Hepatitis B | Real | 0 |
| Measles | Real | 0 |
| BMI | Real | 0 |
| under-five deaths | Real | 0 |
| Polio | Real | 0 |
| Total expenditure | Real | 0 |
| Diphtheria | Real | 0 |

2. Encode binary categorical variables in the dataset to have numerical values.

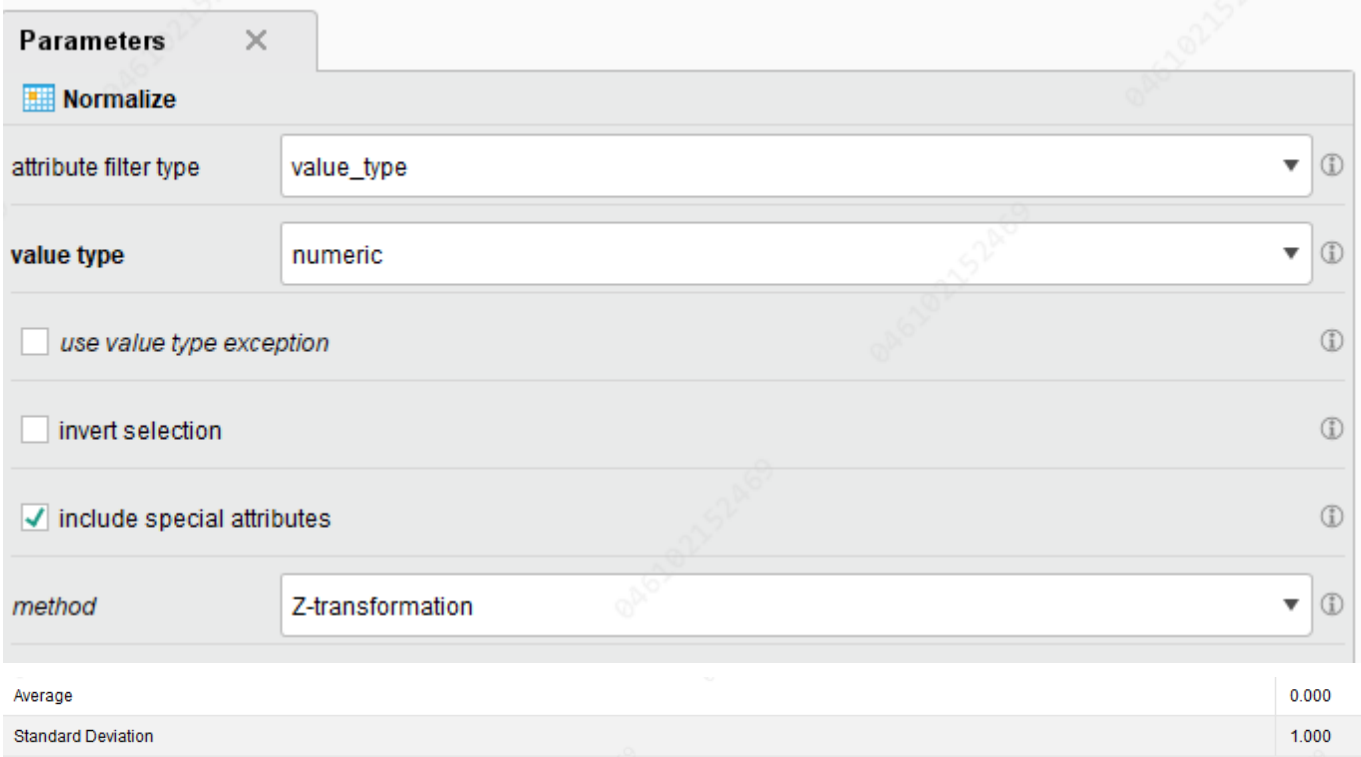



| Status ↑ |
|---|
| 0 |
| 0 |
| 0 |
| 0 |
| 0 |
| 0 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |
| 1 |

0: Developing

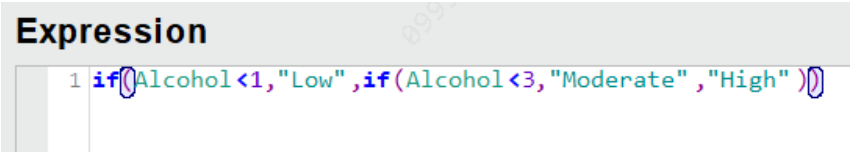
1: Developed

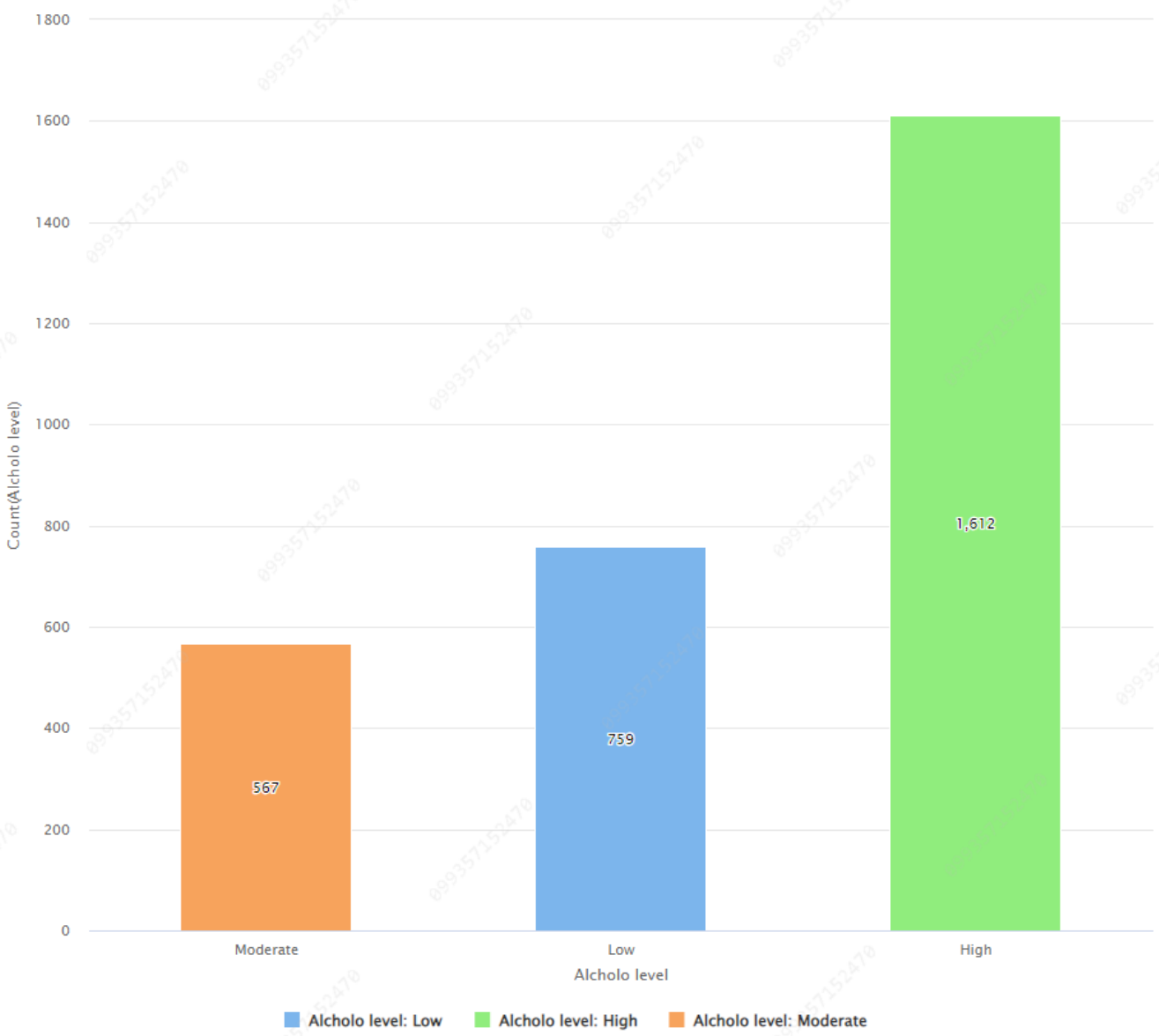
3. Normalize numerical variables to have a mean of 0 and a standard deviation of 1.




## Section 3: Feature Engineering

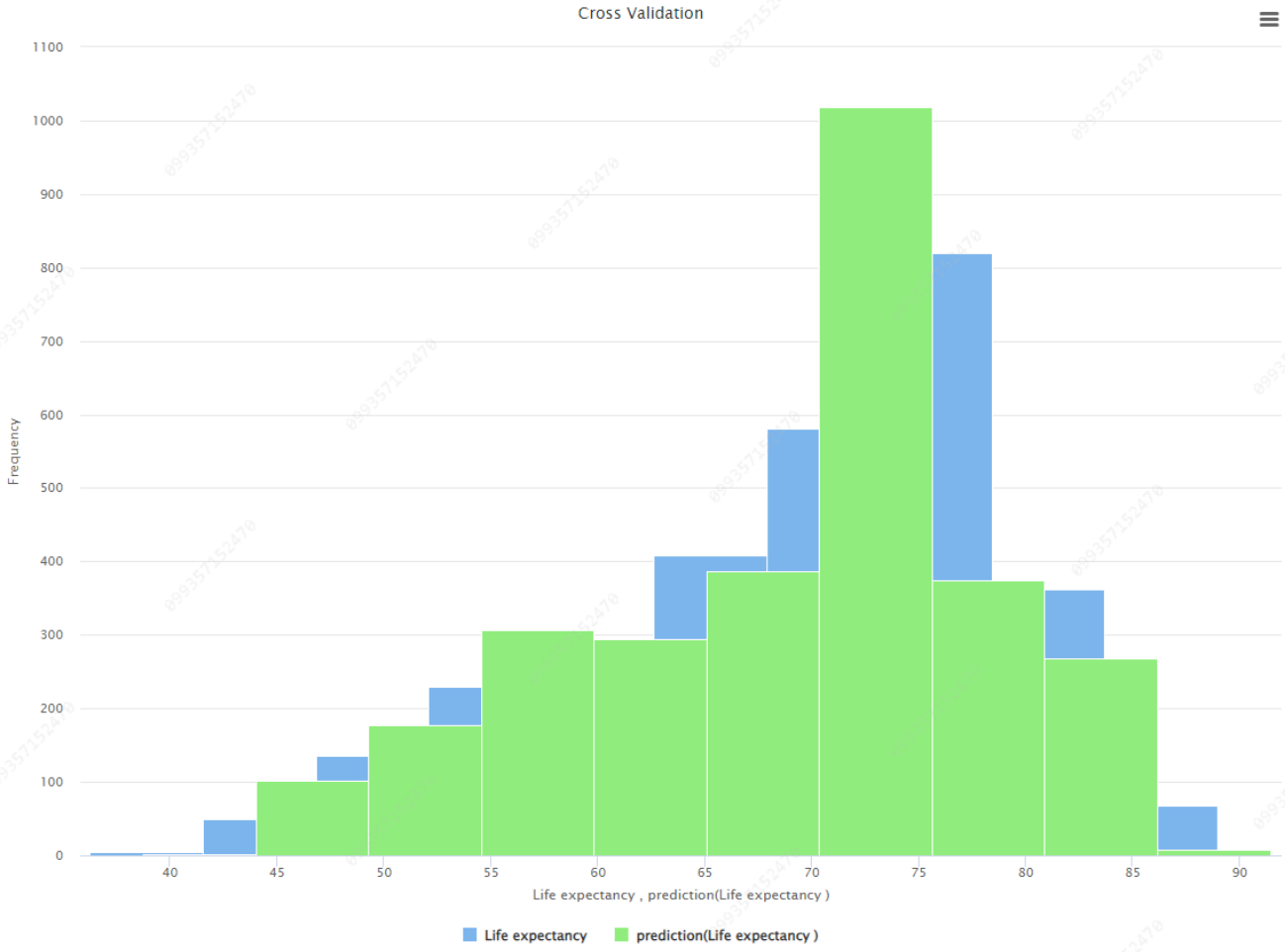
1. Create a new attribute from existing ones called “Alcohol Level”, divide the scores into three levels (Low, Moderate, High).

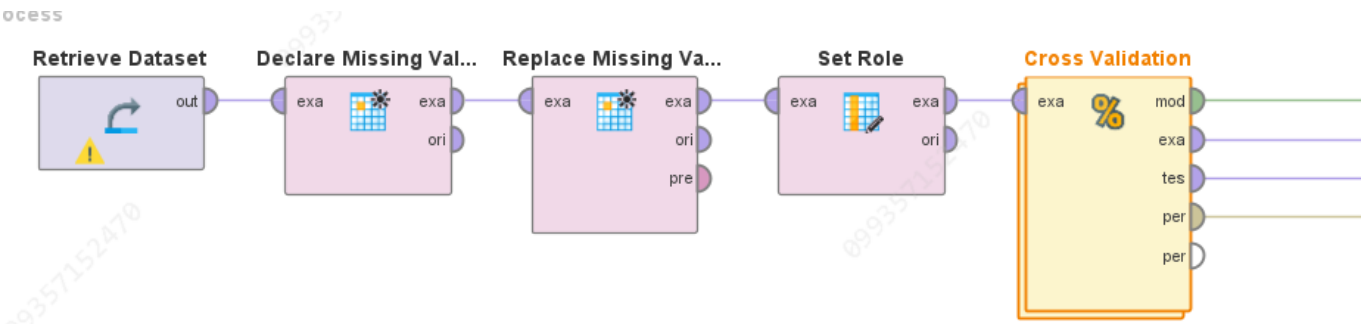


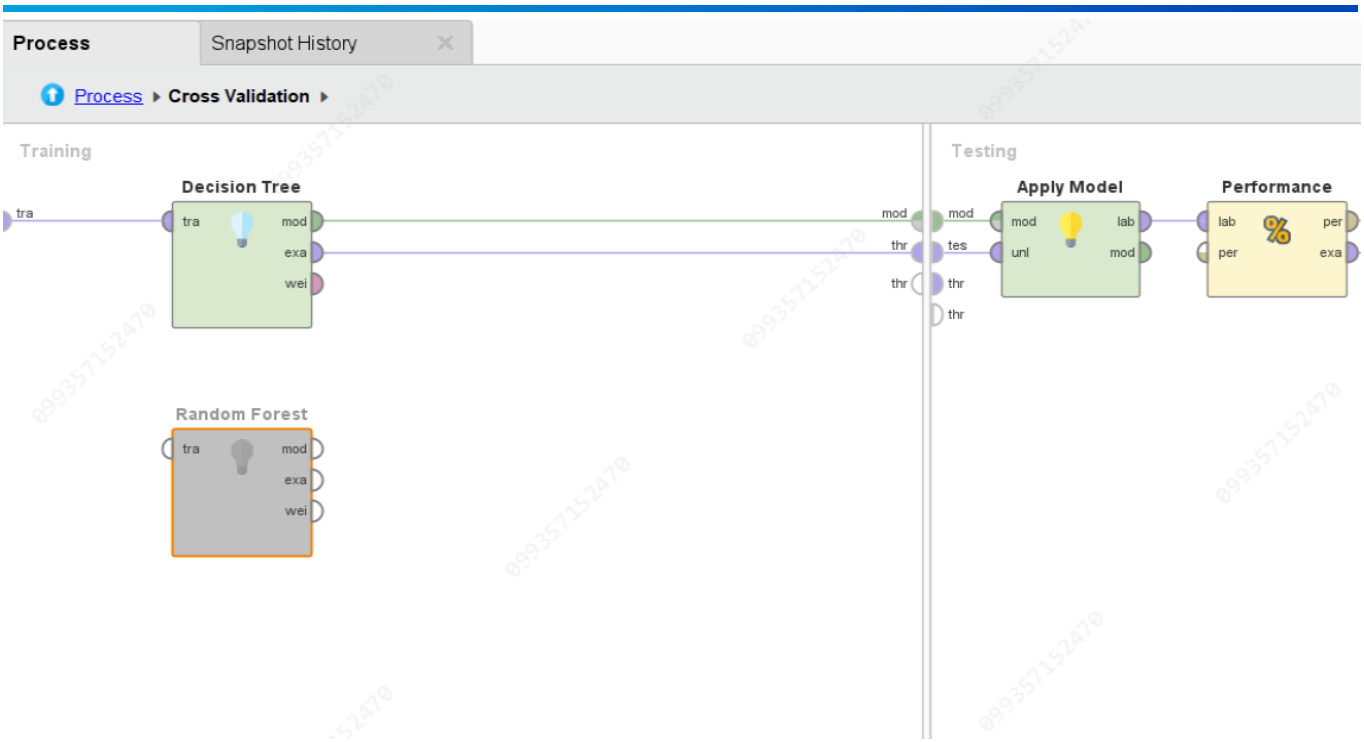



I Used these ranges for classing alcohol level by reading some of an article on this website: [Blood Alcohol Level Chart and Easy Guide]

2. Apply cross-validation to the data by using 10 folds with shuffled sampling to assess the generalization of the model.

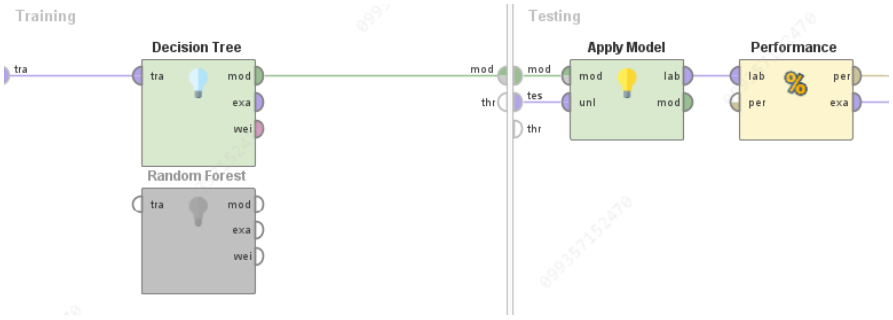





## Section 4: Model Building

1. Train and build a regression model (e.g., logistic regression, random forest) and comment on performance of the model.




The model predictions were close to actual results.

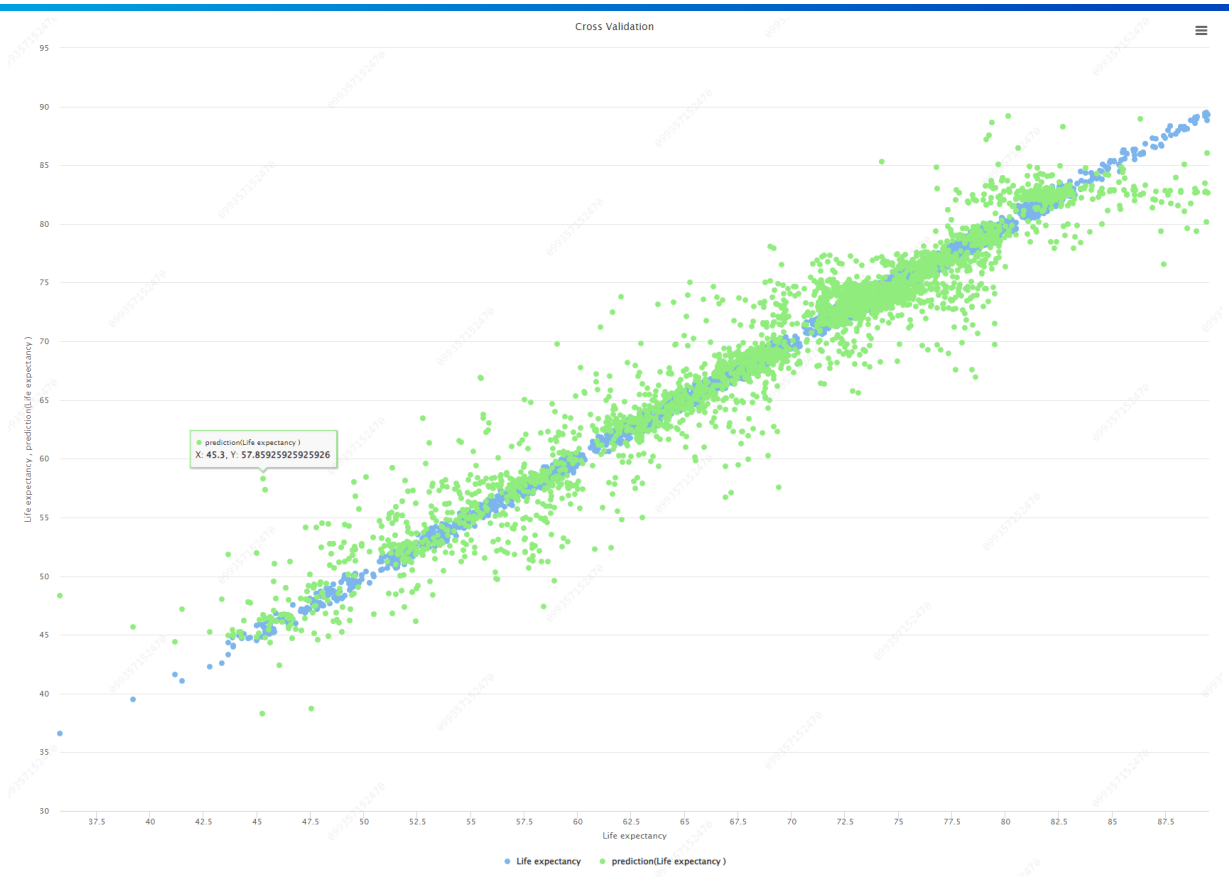



*Figure: DT Performance*

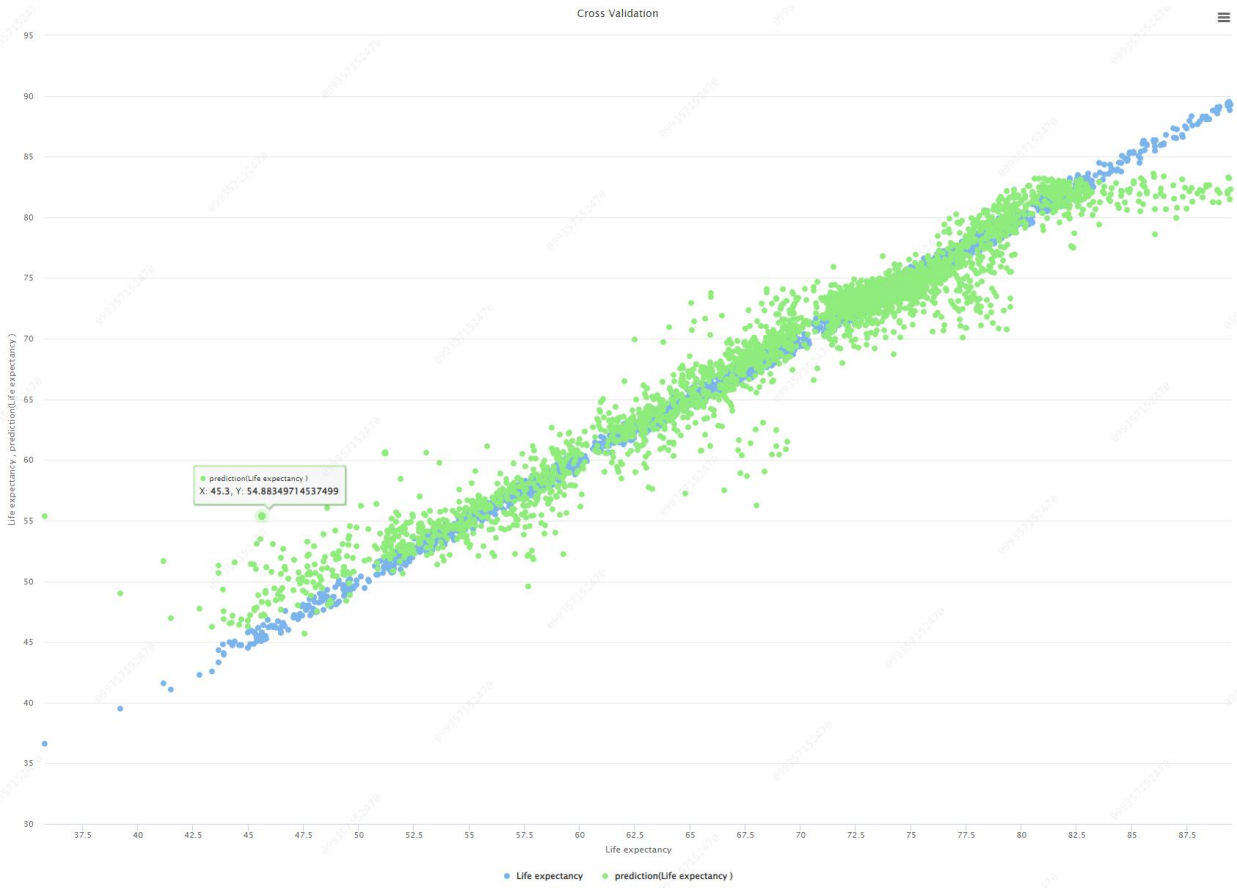
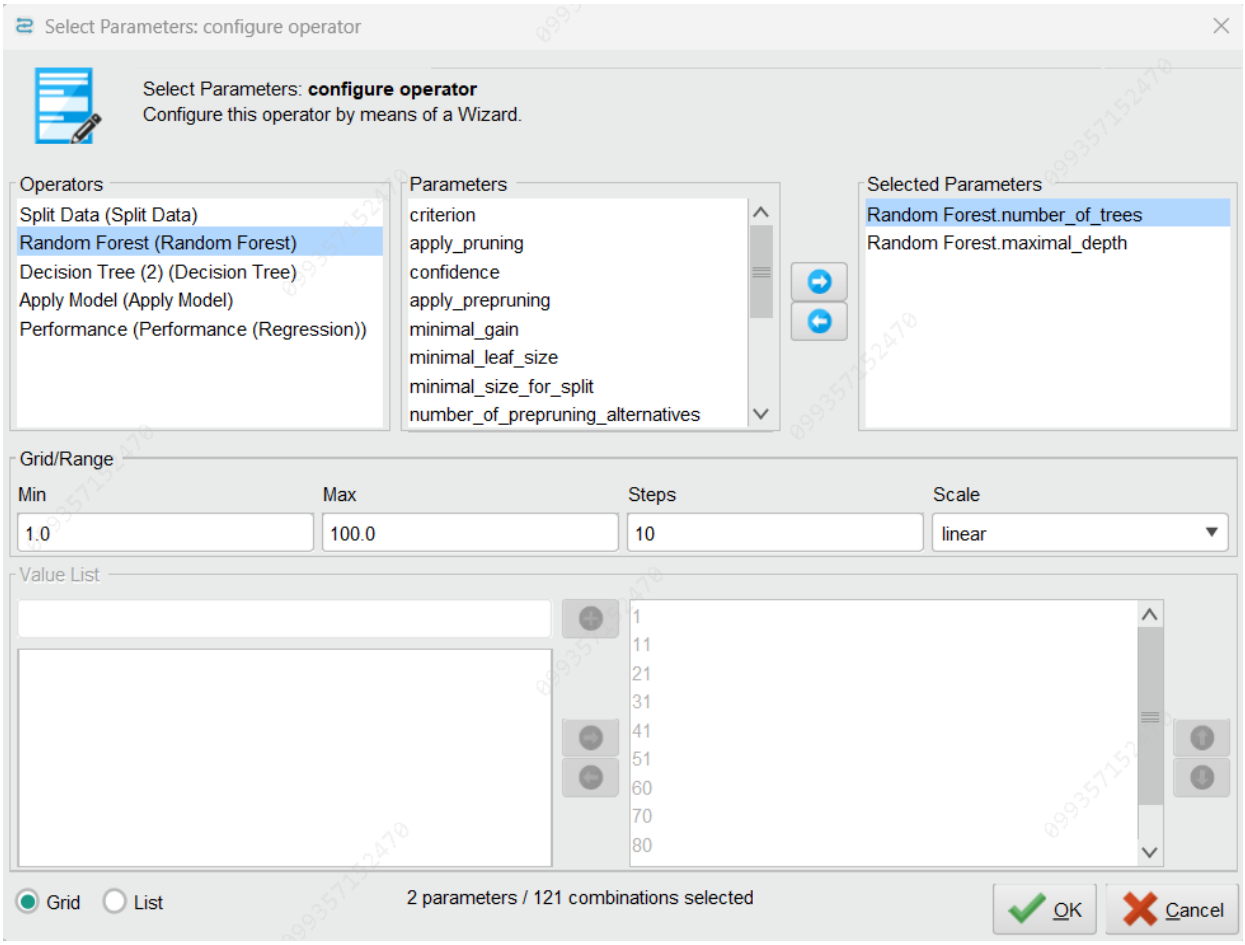



*Figure: RFPerformance*

2. Use optimization techniques to go over different model configurations (e.g., hyperparameters) for comparison. Mention the parameters that you include for optimization.



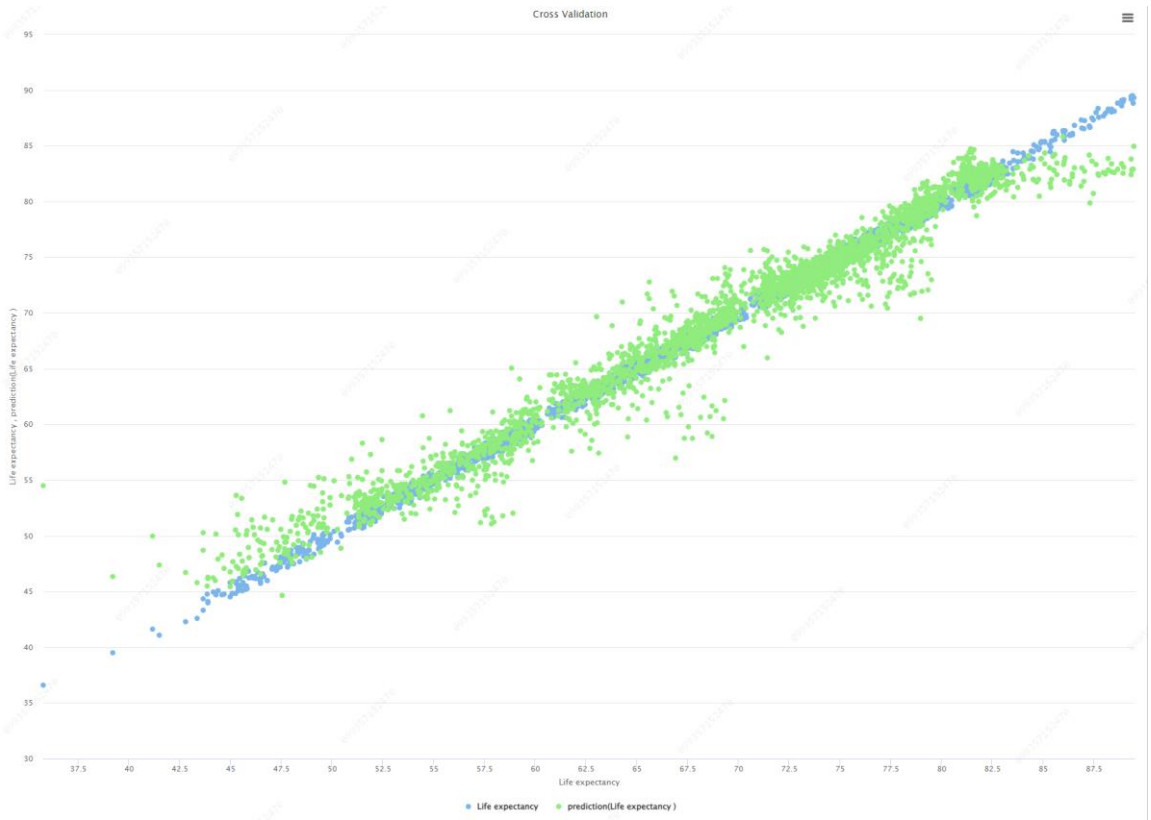




*Figure 1 : RF after using optimization technique.*

## Section 5: Sampling and Handling Errors:

1. Explain the functionality of "Handle Exception" operator and how it can be used in this pipeline to handle errors.
    a. "Handle exception" contains two blocks "Try and Catch." The try block lets you test a block of code for errors. The except block lets you handle the error to prevent run time error.
    Can be used in this pipeline by

2. What are sampling and weighting methods? Explain how they can be used to help in improving model performance. (In General)
    a. Sampling: Techniques used to select a subset of data from large dataset. Sampling can address issues like imbalance data or reduce time training by select short size of the data.
    b. Weighting: assigning different importance to data point during training. For example, higher weights can be given to underrepresented classes in a classification to reduce bias.

## Section 6: Advanced Topics (Optional):

1. Use Python scripts to perform data analysis and do summary statistics of the Life Expectancy data (It may require python libraries to be installed).

**Feel free to explore and experiment with the dataset and RapidMiner's capabilities beyond the questions mentioned above.**