****

**Best Machine Learning Model for WHO’s Life Expectancy**

**Saba Soltanmohammadi, B1923164**

***Abstract*** *- Life expectancy of individuals is a vital metric that is determined by various factors. This report aims to explore the factors that affect life expectancy and build a predictive model to estimate life expectancy based on those factors. The dataset used in this report is collected from:* [*https://www.kaggle.com/datasets/kumarajarshi/life-expectancy -who*](https://www.kaggle.com/datasets/kumarajarshi/life-expectancy%20-who) *which is mainly focused on the World Health Organization data and it is “Regression" type data set, and includes variables related to healthcare, economic, and social factors for different countries in fifteen years from 2000 to 2015. The exploratory data analysis revealed interesting patterns and relationships between the variables. For example, the correlation between life expectancy and schooling was found to be positive, indicating that countries with higher education rates tend to have higher life expectancies. The report also conducted feature selection using various techniques such as mutual information and f-regression, to identify the most important variables that affect life expectancy. Using the selected features, multiple regression models were developed to predict life expectancy. The models used in this report include linear regression, decision tree regression, random forest regression, K-nearest neighbors’ regression, polynomial, and ensemble. The performance of the models was evaluated using metrics such as MAPE (mean absolute percentage error) and R-squared.*

***The results of the analysis indicated that economic, healthcare, and social factors have a significant impact on life expectancy. The best-performing model was the random forest regression model, which achieved a MAPE 5.66 and an R-squared score of 0.95 on the test dataset.***

1. **Introduction**

The World Health Organization defines life expectancy as the average number of years a person can anticipate living in "full health." Over the world, increasing life expectancy has been one of the main objectives of public health initiatives.

A person has a better probability of surviving longer as time goes on. The World Health Organization (WHO) looked at factors that affect people's longevity between 2000 and 2015 in 193 developed and developing nations. So, improved judgements about the health of society can result from a greater understanding of lifespan issues. In recent years, there has been an increase in interest in modelling life expectancy and identifying the most significant factors using methods of machine learning. Many research have examined the connections between life expectancy and various social and health-related aspects using accessible datasets. This research have uncovered a wide range of potential life expectancy determinants, such as income, education, and healthcare access.

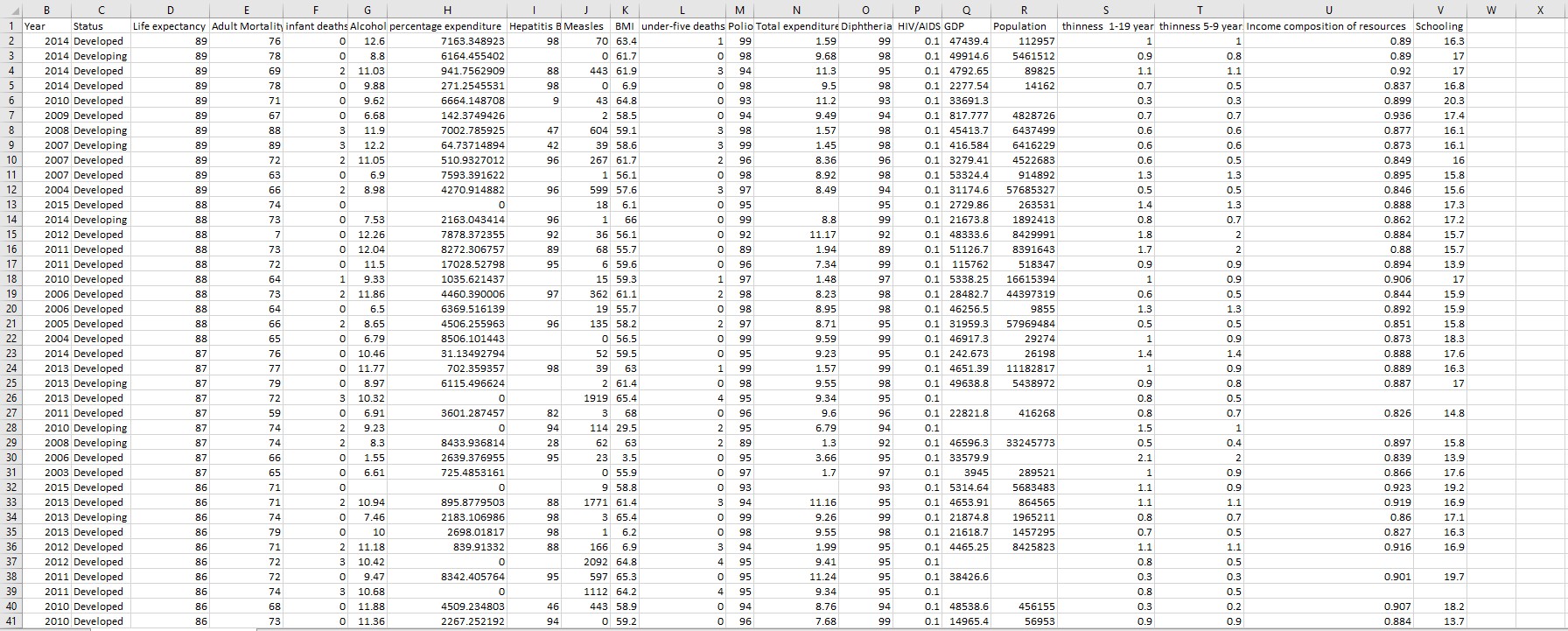
Machine learning techniques offer new opportunities for exploring these components in greater depth and making novel predictions.  Considering that one, the current work aims to use methods from machine learning to build a model that forecasts life expectancy based on a range of social and health-related data. This model can be used to identify the most important predictors of life expectancy and to direct public health policies and initiatives with the aim of improving population health.

## Data exploration and features selection

The Worldwide Health Organization gathers life expectancy data through an annual report from agreed countries and makes it available on its website. This dataset was obtained from: [https://www.kaggle.](http://www.kaggle.com/datasets/kumarajar)c[om/datasets/kumar](http://www.kaggle.com/datasets/kumarajar)ajar expectancy-who. There are 2939 rows and 22 columns in the table (life expectancy data).

The following table shows a description of each column:

|  |  |
| --- | --- |
| Country | specify the list of countries |
| Year | Years from 2000 to 2015 |
| Status | developing and developed countries |
| Life expectancy | life expectancy in age |
| Adult Mortality | Mortality range |
| infant deaths | infant death count |
| Alcohol | Alcohol consumption |
| percentage expenditure | expenditure on health |
| Hepatitis B | Type of virus immunisation |
| Measles | Type of virus immunisation |
| BMI | obesity indicator |
| under-five deaths | count of under 5 deaths |
| Polio | Type of virus immunisation |
| Total expenditure | General Government expenditure on health |
| Diphtheria | Type of virus immunisation |
| HIV/AIDS | Type of virus immunisation |
| GDP | Gross Domestic Product Per Capita |
| Population | the population of the country |
| thinness 1-19 years | thinness among children and adolescents between 1 and 19 |
| thinness 5-9 years | thinness among children between 5 and 9 |
| Income composition of resources | Society Income class |
| schooling | Years of education |

**Figure 1**

## Load the data

This article explains how to use Python to load data into a Pandas DataFrame and give index labels. The first column of a CSV file can be utilised as the index column in the final DataFrame thanks to index labelling in CSV files. When the labels for the rows in the DataFrame are taken from the values in the index column, this is useful. When it comes to life expectancy data, for instance, assigning the index to the first column (country names) enables access to life expectancy data for particular nations as well as merging or connecting the data with other datasets that can utilise the country name as a common identifier.



## Check for missing value

For dealing with missing values different methods have been taken into consideration: (Nayak, 2022)

* Forward fill: During data imputation, missing values in time series data are filled using the following value in the sequence. When missing values are discovered at the beginning of a dataset and we need to fill the missing value with next available value, mostly this technique is used. When the pattern of missing data missing values are dispersed randomly over the dataset, forward fill can be used to fill in the gaps left by missing data.
* Backfill: is commonly used when there are missing values in a time series data set. This method is commonly used when missing data is assumed to be the same as the previous value, such as when data is recorded in consecutive periods. Back filling, on the other hand, can introduce biases in the data if the pattern changes suddenly.
* Mean: Python’s mean () function calculates the arithmetic mean or average of values in a list or array. The function returns the sum of all values in the list or array divided by the number of values. Alternatively, the function can be called as a method of a NumPy array using the NumPy library.
* Drop missing values: remove rows which have null values with drop() function.
* Dealing with zero values

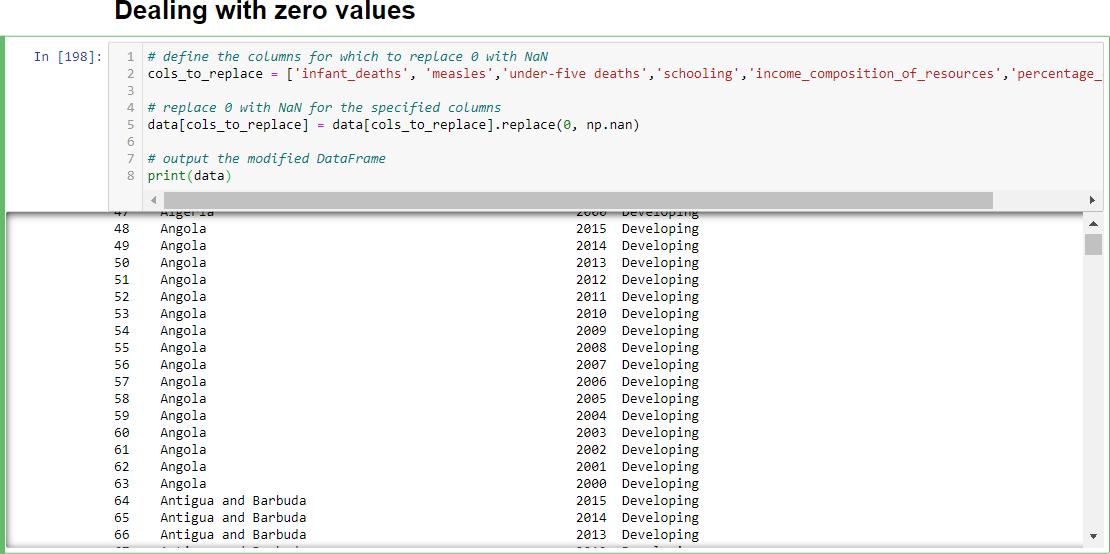
By using isnull() function, it is possible to find out the number of null values. Below command has been used to find the count of missing values for each column:

data.isnull() returns a DataFrame with the same shape as the original data, where each element is True if the corresponding element in the original data is null (missing), and False otherwise.

sum() method returns the sum of all the True values in each column. So, data.isnull().sum() returns the number of null values in each column. data.shape[0] returns the number of rows in the original data. Therefore, data.isnull().sum()/data.shape[0]∗100 returns the percentage of null values in each column.

This command represents that below columns from original file have null values in percentage:

|  |  |
| --- | --- |
| Life expectancy | 0.34 |
| Adult Mortality | 0.34 |
| Alcohol | 6.6 |
| Hepatitis B | 18.82 |
| BMI | 1.16 |
| Polio | 0.65 |
| Total expenditure | 7.69 |
| Diphtheria | 0.65 |
| GDP | 15.25 |
| Population | 22.19 |
| thinness 1-19 years | 1.16 |
| thinness 5-9 years | 1.16 |
| Income composition of resources | 5.68 |
| Schooling | 5.55 |

After careful consideration with data.describe().T which is used to generate a statistical summary of numerical columns, it is found out that the minimum values for some columns are 0 which is may not logical and need to be changed to NaN values.

**

After changing to NaN values, NaN values for all columns have been counted again:

*Graphical user interface

Description automatically generated with medium confidence*

|  |  |  |
| --- | --- | --- |
| **column** | **null percentage** | **how to deal** |
| Life expectancy | 0.34 | drop null rows |
| Adult Mortality | 0.34 | drop null rows |
| infant deaths | 28.86 | mean based on status of the country |
| Alcohol | 6.6 | back fill |
| percentage expenditure | 20.8 | mean based on status of the country |
| Hepatitis B | 18.82 | mean based on status of the country |
| measles | 33.46 | mean based on status of the country |
| BMI | 1.16 | back fill |
| under five deaths | 26.72 | mean based on status of the country |
| Polio | 0.65 | forward fill |
| Total expenditure | 7.69 | mean based on status of the country |
| Diphtheria | 0.65 | mean based on status of the country |
| GDP | 15.25 | mean based on status of the country |
| Population | 22.19 | drop column |
| thinness 1-19 years | 1.16 | mean based on status of the country |
| thinness 5-9 years | 1.16 | mean based on status of the country |
| Income composition of resources | 10.11 | mean based on status of the country |
| Schooling | 6.5 | mean based on status of the country |
|  |  |  |

Table 1: Table percentage of missing values of each column with the way they addressed.

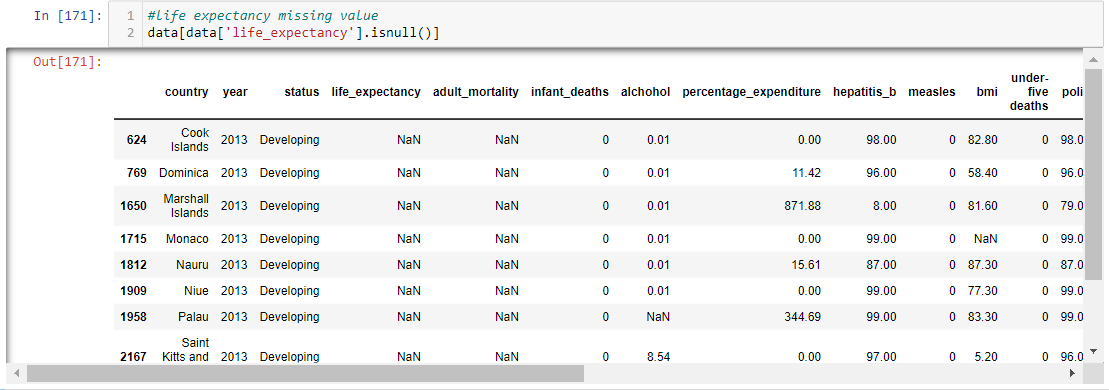
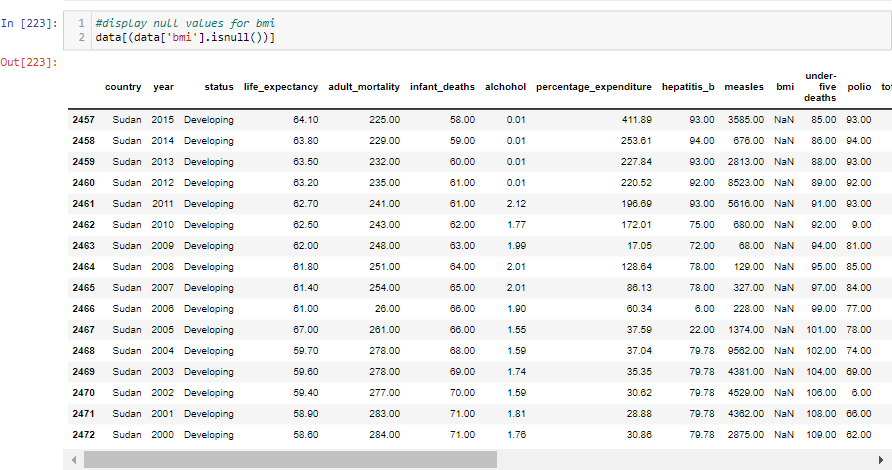
* Life expectancy has 10 NaN rows in different countries, but same year (2013) and same status (Developing). Because the rate of Null value is low in life expectancy, it is decided to drop rows which had null values.
* Adult mortality has the same condition and NaN values are in same row with life expectancy’s NaN value. So, after dropping the NaN values for life expectancy, adult mortality’s NaN values were dropped as well.
* The same action has been done for BMI. After finding out the count of NaN rows, it shows that BMI has missing values just for one country which is Sudan. So, dropping NaN rows has been decided for dealing with BMI missing values.
* After careful investigation in data, South Sudan has been dropped from data set as well because it has so many missing values in different features which dealing with different methods will cause biased data for South Sudan.
* Population has been dropped as well because it has no relation with solving the overall main problem.

Figure 2: life expectancy null values



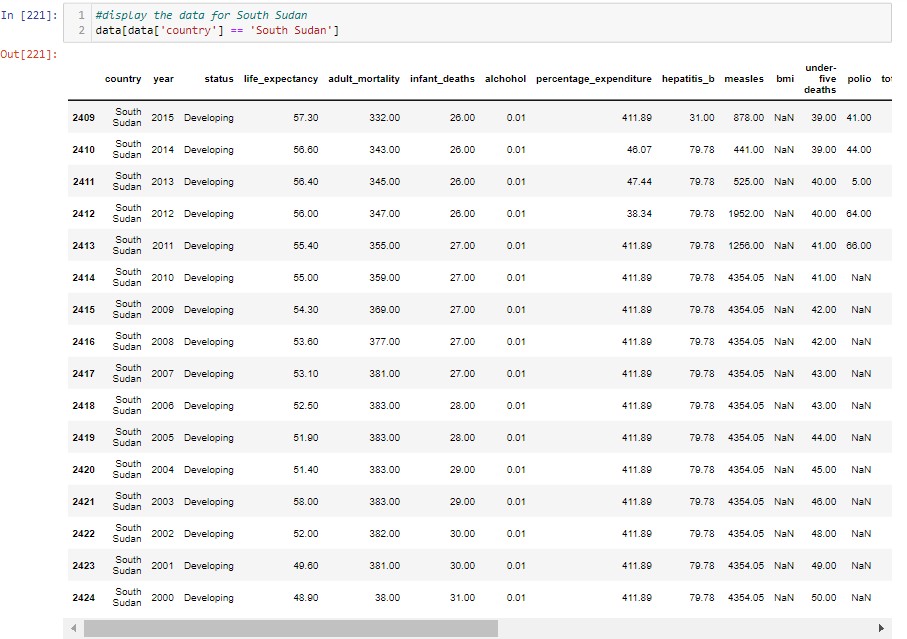
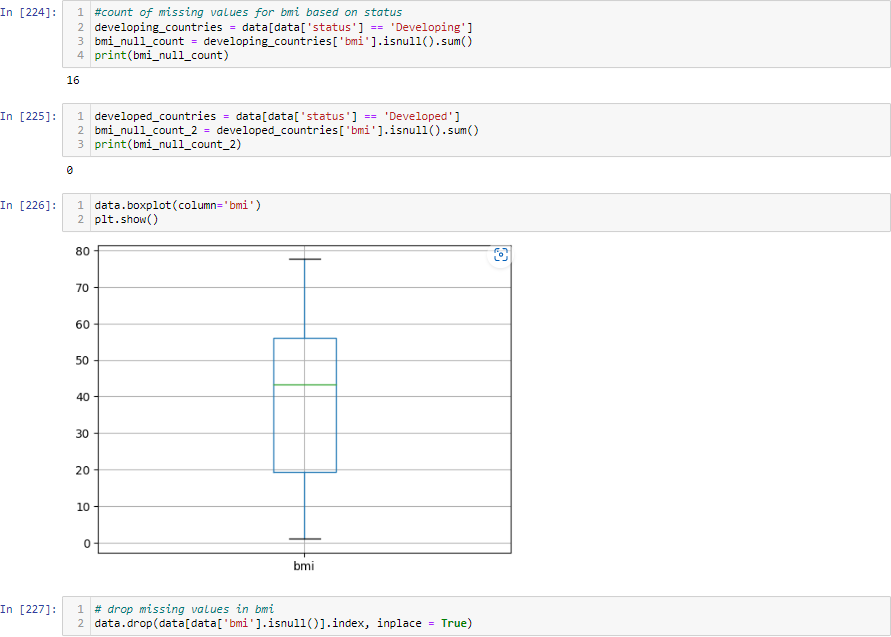


Figure 3: bmi null values





*Table

Description automatically generated*

Figure 4: population null values

* Missing values for alcohol are mostly different developing countries in 2015 and it was filled with backward filling:

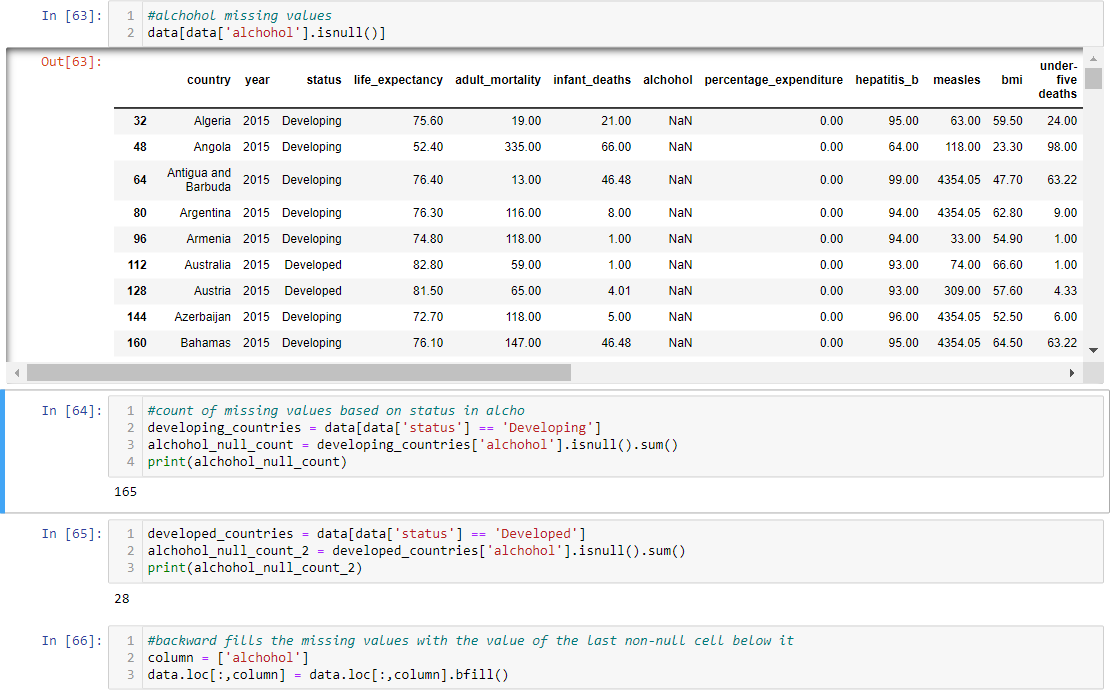


Figure 5: alcohol null values

* Missing infant deaths values are combination of developing and developed countries in different years and it was filled with mean value of each type of country:

*Graphical user interface

Description automatically generated with low confidence*

Figure 6: infant deaths null values

Graphical user interface, text, application, email

Description automatically generated

* Missing values for percentage expenditure, same as infant deaths, are combination of developing and developed countries in different years and it was filled with mean value of each type of country. This is the same for hepatitis B, Measles, total expenditure, GDP, percentage expenditure, income composition of resources, schooling, and under five deaths:

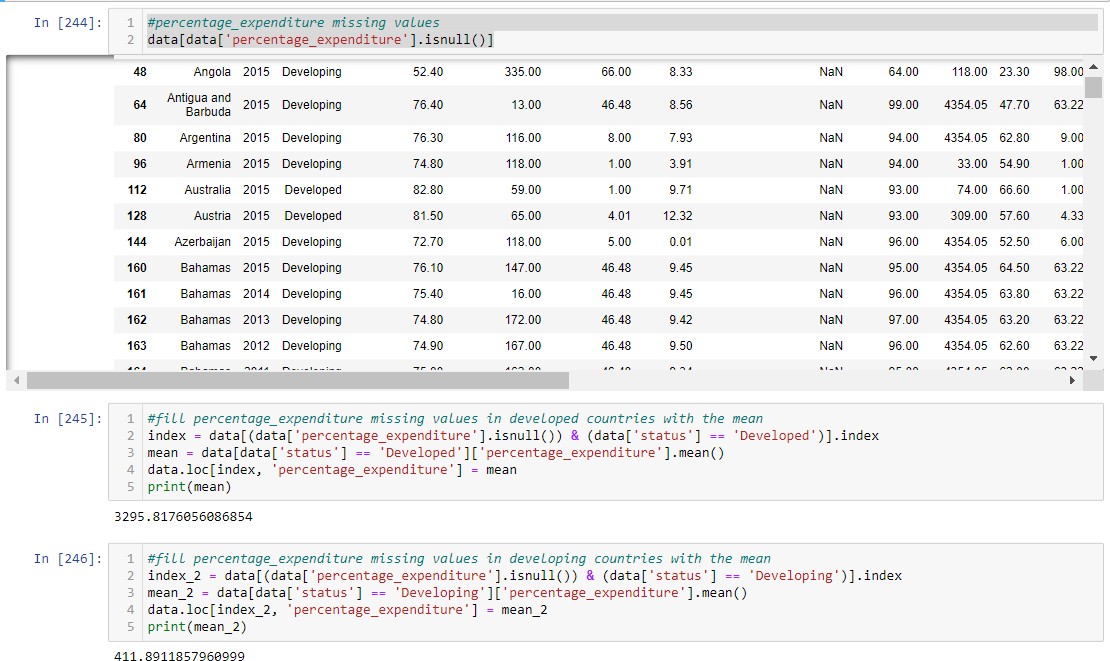


Figure 7: percentage expenditure null values

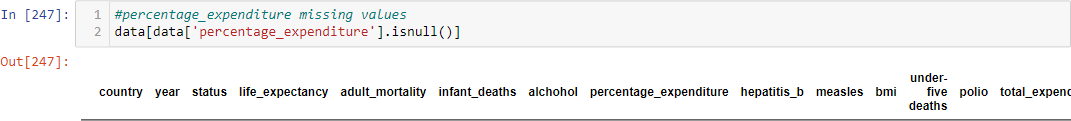
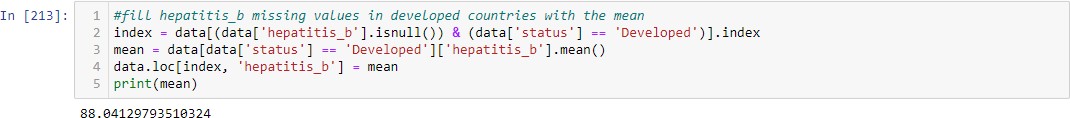
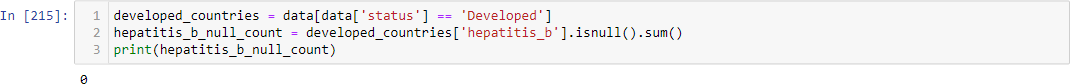




Figure 8: hepatitis B null values





Graphical user interface

Description automatically generated with medium confidence

Figure 9: measles null values

Graphical user interface, text, application, email

Description automatically generated

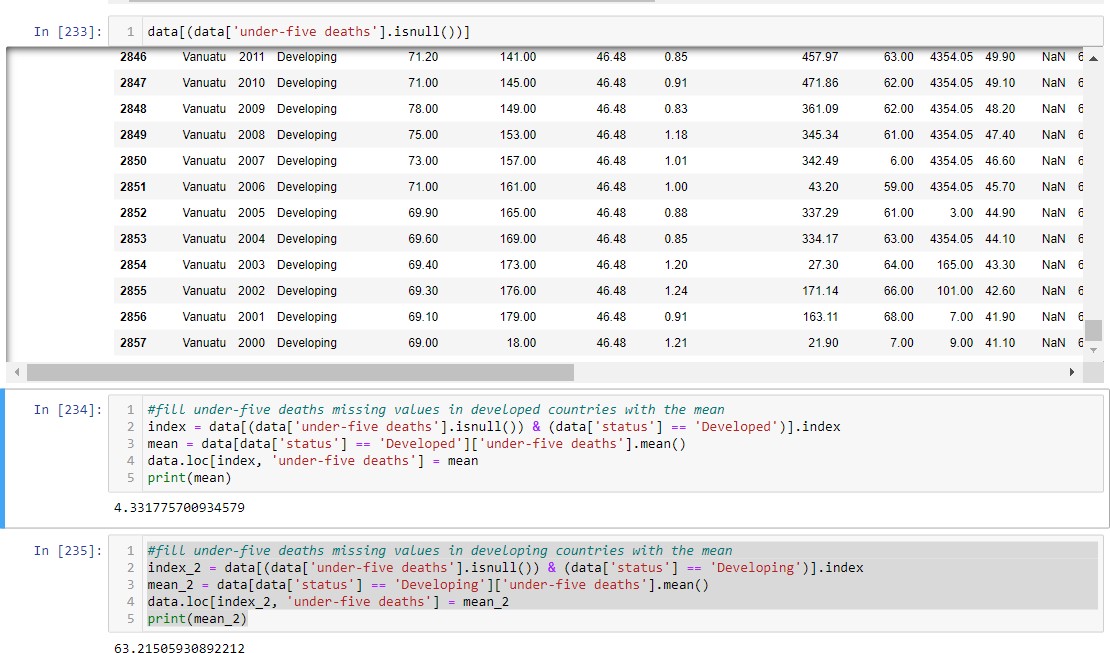


Figure 10: under-five deaths null values

Graphical user interface, application

Description automatically generated

Figure 11: total expenditure null values

Graphical user interface, text, application, email

Description automatically generated

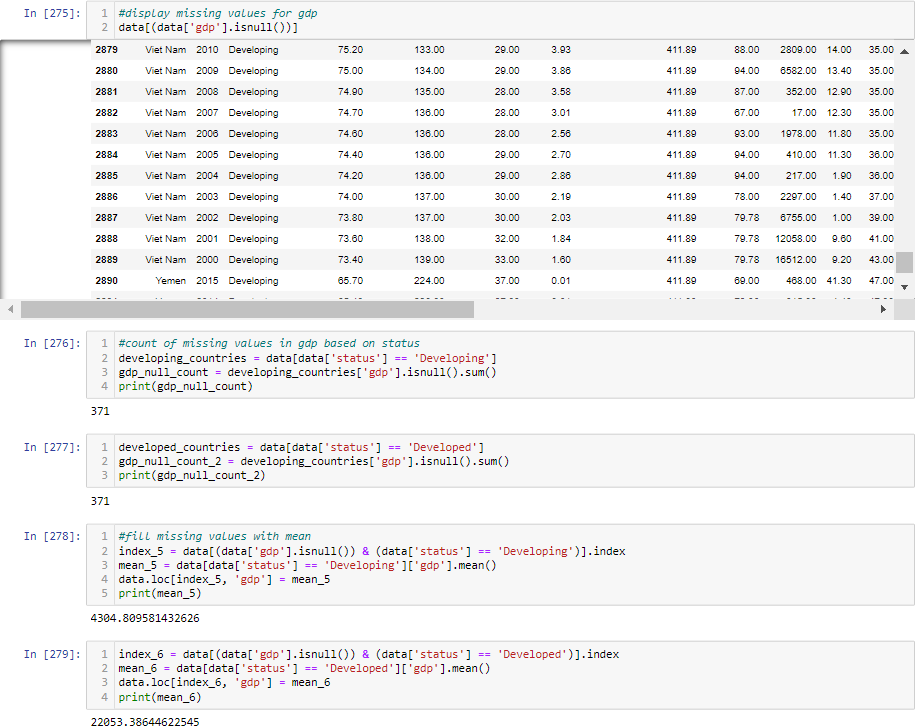
**

Figure 12: GDP null values

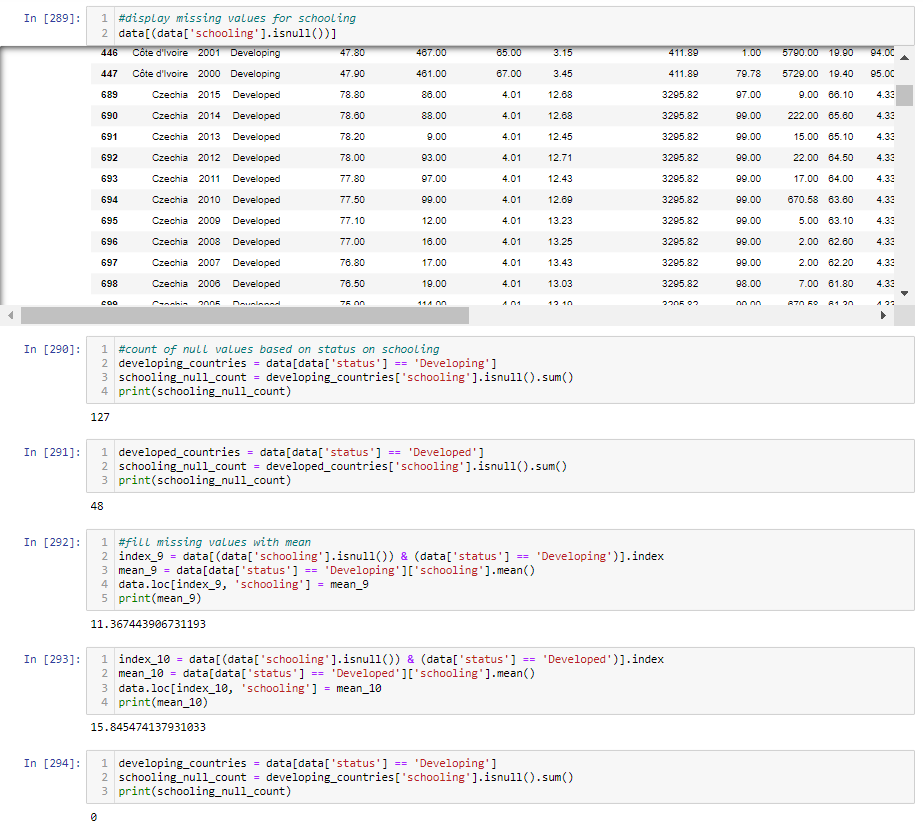


Figure 13: schooling null values

* Missing values for Polio and diphtheria are filled with forward fill because both have missing values from the beginning of time series (2000) and are related to two developing countries with 8 rows missing value.



Figure 14: Polio null values

*Graphical user interface, application

Description automatically generated*

Figure 15: Diphtheria null value

## Explore the variables and correlation analysis

For exploring variables, techniques include performing hypothesis testing to determine if there is a significant difference between groups or correlations between variables can be taken into consideration. In this report, exploring the variables has been done by visualizations such as histograms, box plots, scatter plots, and correlation matrices (heatmap) to understand the distribution of each variable and any potential relationships or patterns among them. In heatmap it is represented that life expectancy has strong correlation to income composition of resources, schooling, and adult mortality.

*Chart

Description automatically generated*

Figure 16: heatmap for correlation

To classify life expectancy based on status, general scatter plot has been decided to be plotted for all features.

Graphical user interface, application, timeline

Description automatically generated

Figure 17: scatter plot for life expectancy based on statue.

Below plots show positive correlation between schooling and income composition of resources and life expectancy.

Chart

Description automatically generatedChart, scatter chart

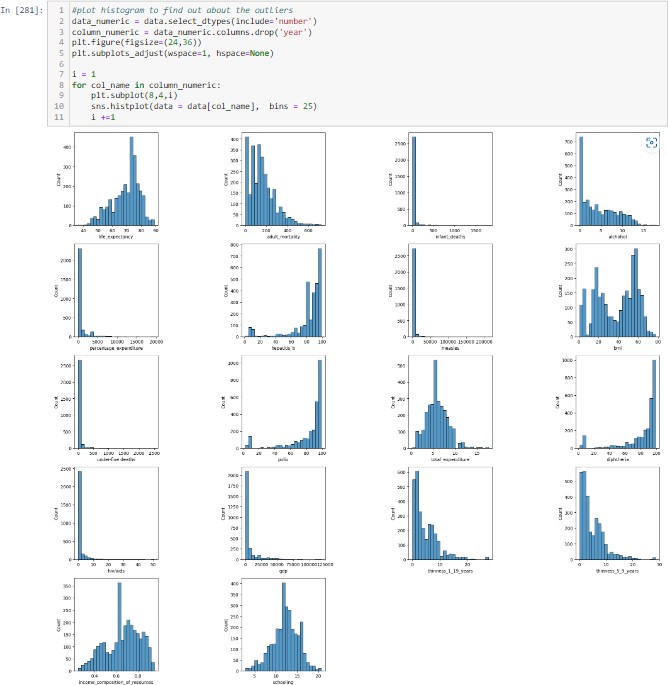
Description automatically generated

Figure 18: histogram for variance and normal distribution

*Chart, box and whisker chart

Description automatically generated*

*Chart, box and whisker chart

Description automatically generated*

Figure 19: box plot for visualizing distribution

As a part of variable exploration and preprocessing the data before training, the method for finding the scale of outliers and dealing with them has been taken into consideration. [[1]](#_bookmark0) Outliers are data points that are significantly different from other data points in a dataset. They can occur due to errors in data collection or measurement, or they may represent genuine extreme values. Outliers can have a significant impact on data analysis, so it is critical to deal with them appropriately. Before dealing with outliers, it is necessary to identify them, which has been accomplished by classifying z-scores less than -3 and greater than 3 as outliers. A value with a z-score of -3 or less is more than three standard deviations below the mean, while a value with a z-score of 3 or more is more than three standard deviations above the mean. Yeo-Johnson transformations are a method used in data preprocessing to deal with outliers. [2]

*A picture containing text

Description automatically generated*

Figure 20: check for outliers

*Graphical user interface, text, application

Description automatically generated*

## Identify the target variable

Identifying the target variable is an important step in any predictive modeling problem. In the case of life expectancy, the target variable is life expectancy itself. In the dataset we are working with, the life expectancy is given in the column named ’Life expectancy’. This column has been used as the target variable in our modeling process.

It is important to note that the target variable should be carefully selected based on the problem we are trying to solve. In this case, the problem is to predict life expectancy based on various factors such as income, education, and healthcare. Hence, life expectancy is an appropriate target variable.

## Feature Selection

According to a study by Gajawda in 2019, feature selection is the process of choosing the most important features from a given dataset. This can help improve the performance and efficiency of a machine learning model. To determine the significance of each feature in predicting life expectancy, the ANOVA test was used on all the features in the dataset. The results showed that features were significant, with "status" and "country" having the largest impact and "alcohol" having the smallest impact. ANOVA is a statistical method used to compare the means of three or more groups of data, and it is a popular method for feature selection because it can handle both continuous and categorical variables and can detect complex interactions between features. The study found that country, status, adult mortality, polio, diphtheria, hiv/aids, and schooling were the most relevant features in predicting life expectancy. In life expectancy data set, ANOVA feature selection has been used to check which features are more relevant to the target variable. After investigation on the result, it may conclude that country, status, adult mortality, polio, diphtheria, hiv/aids, schooling are more relevant to the life expectancy.

Table

Description automatically generated

Statistical methods such as the F-test and mutual information can also be used to select features from a dataset. The F-test determines the degree of linear dependence between two variables and can be used to assess the relationship between each feature and the target variable. A low F-test p-value indicates that a feature is more relevant and should be included in the model. Mutual information quantifies the amount of information provided by each feature about the target variable. A feature with a high mutual information value is more relevant and should be chosen for the model. F-test and mutual information are both commonly used in feature selection and can be combined to improve the performance of a machine learning model.

It is important to calculate these scores after splitting the data into training and testing sets because the feature selection process should only be done on the training data to prevent information leakage. Information leakage occurs when information from the testing set is used to select features or train a model. This can result in overfitting, where the model performs well on the testing data but poorly on new, unseen data. By calculating F-test and mutual information score on the training data only, we ensure that the feature selection process is based only on information available during training and not influenced by testing data. Also, splitting a dataset into training and testing sets is a common practice in machine learning to evaluate the performance of a predictive model. The training set is used to fit the model to the data, while the testing set is used to evaluate the model’s ability to generalize new, unseen data. This approach helps to ensure that the model has not simply memorized the training data but can make accurate predictions on new, independent data.

Graphical user interface, text, application, email

Description automatically generatedIn our model, F-score and mutual information have been used to select top 12 features and fit it to our train data set. Once the data is split, various predictive models can be applied to the training set, such as linear regression, decision trees, random forest, SVM and ensemble which are used in this prediction. The model’s performance on the training set can then be evaluated to optimize model’s performance.

Chart

Description automatically generated

Chart

Description automatically generated Figure 21: features election with mutual information

Figure 22: feature selection with F-score

*Diagram, text

Description automatically generated*

*Chart, scatter chart

Description automatically generated*

## Experiments and Results

As (Understand Advanced Metrics, n.d.) mentioned, in machine learning, an experiment involves testing the model on a dataset that is separate from the one used to train it and assessing how well it generalizes to new data. The goal of an experiment is to understand the behaviour of the model and its ability to make accurate predictions on unseen data.

Experiments involve multiple iterations where various configurations of the model are changed, and the performance is evaluated to find the best-performing model. The results of an experiment can be used to improve the model, select the best model for deployment or compare different models to determine which one is the most effective.

In machine learning, it is important to evaluate a model's performance in order to determine how well it can predict outcomes. MAPE (mean absolute percentage error) and R-squared (R2) score are two commonly used metrics for evaluating regression models. The better the predictions, the closer the MAPE value is to zero. MAPE can compare forecast error between different data set sizes and time scales because it specifies the size of the error as a percentage rather than actual values. R2 score closer to 1 indicates that the model can explain a larger portion of the variance in the target variable, while a score closer to 0 indicates that the model is not a good fit for the data. By calculating both MPAE and R2 score, we can get a good understanding of how well our model is performing and make any necessary adjustments to improve its performance and compare models together.

## Linear Regression Model

Linear regression is a popular predictive model that can be used to investigate the relationship between a dependent variable (life expectancy in this case) and one or more independent variables (features). Linear regression can be used in the context of the life expectancy dataset to investigate how various features (such as income, education, and healthcare indicators) are related to life expectancy. [4]

Graphical user interface, text, application, email

Description automatically generated

## Decision Tree Model

Life expectancy can be influenced by a variety of factors such as age, gender, lifestyle habits, and environmental factors, among others. Decision trees are relatively easy to interpret and can help identify which features are most important in predicting life expectancy. [[4]](#_bookmark3)

Graphical user interface, text, application, email

Description automatically generated

## KNN Model

KNN (K-Nearest Neighbors) is a simple and effective algorithm for regression problems like life expectancy prediction. The basic idea of KNN is to find the K closest data points to a new data point in the training set, and then use their target values to predict the target value of the new data point. In the case of life expectancy prediction, it is reasonable to assume that similar countries (based on their features) may have similar life expectancies. Therefore, KNN can be a good fit for this problem since it makes predictions based on the similarity of data points. (Ohri, 2021)

Graphical user interface, text, application, email

Description automatically generated

## Random Forest Model

Random Forest is a popular machine learning algorithm for regression problems such as life expectancy prediction. The benefit of using Random Forest to predict life expectancy is that it can handle non-linear relationships between predictors and the target variable, as well as missing values and outliers in the data. Furthermore, Random Forest can identify important features and their interactions, which can provide insights into factors that influence life expectancy.

Graphical user interface, text, application, email

Description automatically generated

## SVM Model

Graphical user interface, text, application, email

Description automatically generatedSVM (Support Vector Machines) can be used for regression problems as well as classification problems. In the case of the life expectancy data set, SVM regression can be used to predict life expectancy based on certain features or attributes such as income, education level, health status, etc. SVM works by finding the hyperplane in a high-dimensional space that maximizes the distance between the closest points of different classes or in this case, different values of life expectancy.[4]

## Ensemble Model

Ensemble methods frequently outperform individual models by lowering the risk of overfitting and improving model stability.

To make a final prediction, the ensemble model can take the average or weighted average predictions made by individual models. This can help improve the model's accuracy and reduce the impact of outliers in the data set.

*Graphical user interface, text, application, email

Description automatically generated*

## Polynomial Model

In the context of the life expectancy dataset, there may be some non-linear relationships between the predictor and response variables that are not captured by a linear regression model. The polynomial regression model can capture these non-linear relationships and provide a better fit to the data. [[4]](#_bookmark3)

Graphical user interface, text, application, email

Description automatically generated

## Conclusion

We discovered that the Random Forest model outperforms other models in terms of both MAPE and R-squared score after running different regression models on our dataset. The Random Forest model has the lowest MAPE of 5.67% and the highest R-squared score of 0.95 out of all the models tested. According to these results, the Random Forest model is the most accurate and reliable model for our dataset, making it the best model for predicting future outcomes.

Comparison between models in R2 score and MAPE has been plotted in bar charts.

## Discussion and Future Work

The Random Forest model outperforms other models in predicting the target variable, according to the results of our regression model with feature selection. We can improve the performance of the regression model and reduce noise introduced by irrelevant or redundant features by selecting the most informative features.

In the future, we can investigate more advanced feature selection techniques such as Recursive Feature Elimination (RFE) and L1 regularisation to see if we can improve model performance even further. [5]

Graphical user interface

Description automatically generated with medium confidence

## References

1. Rousseeuw. “Transforming variables to central normality”. In: (Mar. 2021). doi: [10.1007/s10994-02](https://doi.org/10.1007/s10994-021-05960-5) [1-05960-5](https://doi.org/10.1007/s10994-021-05960-5).
2. Riani Marco, Atkinson Anthony, and Corbellini Aldo. “Automatic robust Box-Cox and extended Yeo- Johnson transformations in regression”. In: (June 2022). doi: [10.1007/s10260-022-00640-7$\#$aut](https://doi.org/10.1007/s10260-022-00640-7%24\#%24auth-Marco-Riani) [h-Marco-Riani](https://doi.org/10.1007/s10260-022-00640-7%24\#%24auth-Marco-Riani).
3. Asaithambi Sudharsan. “why-how-andwhen-to-apply-feature-selection-e9c69adfabf2 > sampath kumar gajawada”. In: (June 2018).
4. Ohri Ajay. In: (June 2022).
5. Prabhakaran Selva. In: (June 2018).

Rushikesh Pupale, June 16, 2018 *<*[*https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989*](https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989)*>*

[Gurucharan M K, Jul 14,2020](Gurucharan M K, Jul 14,2020 <https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda>) *[<https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda>](Gurucharan M K, Jul 14,2020 <https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda>)*

Understand advanced metrics, *<*[*https://help.anaplan.com/685ff9b2-6370-46ba-af10-679405937113-Understand- advanced-metrics*](https://help.anaplan.com/685ff9b2-6370-46ba-af10-679405937113-Understand-%20%20advanced-metrics)*>*

Kaijub, August 2022,[*https://www.kaggle.com/code/kaiju8/life-expectancy-who-ml-project#Plotting*](https://www.kaggle.com/code/kaiju8/life-expectancy-who-ml-project#Plotting)

[Stephen Allwright](https://stephenallwright.com/author/stephen/), 10 Aug 2022,[*https://stephenallwright.com/good-mse-value/*](https://stephenallwright.com/good-mse-value/)

[Walker Rowe](https://www.bmc.com/blogs/author/walker_rowe/), July 5, 2018*,* <[*https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/#:~:text=There%20is%20no%20correct%20value,means%20the%20model%20is%20perfect.*](https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/#:~:text=There%20is%20no%20correct%20value,means%20the%20model%20is%20perfect.)*>*

[Rob Hyndman](https://stats.stackexchange.com/users/159/rob-hyndman), June 17, 2011, <[*https://stats.stackexchange.com/questions/11636/the-difference-between-mse-and-mape#:~:text=MSE%20is%20scale%2Ddependent%2C%20MAPE,when%20percentages%20make%20no%20sense*](https://stats.stackexchange.com/questions/11636/the-difference-between-mse-and-mape#:~:text=MSE%20is%20scale%2Ddependent%2C%20MAPE,when%20percentages%20make%20no%20sense)*.>*

Ajay Ohri, Jun 16, 2022,

*<*[*https://u-next.com/blogs/machine-learning/popular-regression-algorithms-ml/*](https://u-next.com/blogs/machine-learning/popular-regression-algorithms-ml/)*>*

[Dominik Polzer, Jun 21, 2021](https://dmnkplzr.medium.com/?source=post_page-----fc3c8890f9e3--------------------------------)

<[*https://towardsdatascience.com/7-of-the-most-commonly-used-regression-algorithms-and-how-to-choose-the-right-one-fc3c8890f9e3*](https://towardsdatascience.com/7-of-the-most-commonly-used-regression-algorithms-and-how-to-choose-the-right-one-fc3c8890f9e3)*>*

Sudharsan Asaithambi,Jun 31,2018,

*<*[*https://towardsdatascience.com/why-how-and-when-to-apply-feature-selection-e9c69adfabf2*](https://towardsdatascience.com/why-how-and-when-to-apply-feature-selection-e9c69adfabf2)*>*

Sampath Kumar Gajawada, Oct 19, 2019 ,Cornell Statistical Consulting Unit,October, 2020,

*<https://cscu.cornell.edu/wp-content/uploads/68\_regfit.pdf>*

[Eddie\_4072](https://www.analyticsvidhya.com/blog/author/eddie_4072/), February 15th, 2023,

*<*[*https://www.analyticsvidhya.com/blog/2021/05/dealing-with-missing-values-in-python-a-complete-guide/*](https://www.analyticsvidhya.com/blog/2021/05/dealing-with-missing-values-in-python-a-complete-guide/)*>*

Lopamundra Nayak, Feb 16, 2022,

*<*[*https://medium.com/nerd-for-tech/dealing-with-missing-data-using-python-3fd785b77a05*](https://medium.com/nerd-for-tech/dealing-with-missing-data-using-python-3fd785b77a05)*>*

[Marco Riani](https://link.springer.com/article/10.1007/s10260-022-00640-7#auth-Marco-Riani), [Anthony C. Atkinson](https://link.springer.com/article/10.1007/s10260-022-00640-7#auth-Anthony_C_-Atkinson), [Aldo Corbellini](https://link.springer.com/article/10.1007/s10260-022-00640-7#auth-Aldo-Corbellini), Automatic robust Box–Cox and extended Yeo–Johnson transformations in regression, [8 June, 2022](https://link.springer.com/article/10.1007/s10260-022-00640-7#article-info)

*<https://link.springer.com/article/10.1007/s10260-022-00640-7>*

Jakob Raymaekers & Peter J. Rousseeuw, Transforming variables to central normality, March 2021,[*https://link.springer.com/article/10.1007/s10994-021-05960-5*](https://link.springer.com/article/10.1007/s10994-021-05960-5)

[Selva Prabhakaran](Selva%20Prabhakaran), June 7, 2018*, <https://www.machinelearningplus.com/machine-learning/feature-selection*