Individual Assignment 1
COSC2793
S3970066
Amay Viswanathan Iyer

Dataset Description and Problem Statement: The dataset provided contains information related to various factors that potentially influence life expectancy across different countries and years. The training dataset consists of 22 columns, including the target variable "TARGET_LifeExpectancy," which represents the life expectancy in years. The objective is to develop a predictive model that accurately estimates life expectancy based on the given set of features in the test dataset, which has the same structure as the training set but lacks the target variable column. To be able to achieve this, the following nine steps need to be followed through:

1. Preliminary Data Processing
2. Exploratory Data Analysis (EDA)
3. Data Cleaning
4. Outlier Handling
5. Data Processing
6. Model Development
7. Model Evaluation
8. Model Selection
9. Prediction

```
# Loading the datasets
df_train = pd.read_csv('train.csv')
df_test = pd.read_csv('test.csv')

# Displaying the basic information
print("Train dataset shape:", df_train.shape)
print("Test dataset shape:", df_test.shape)
print("\nFirst few rows of the Train dataset:")
print(df_train.head())
```

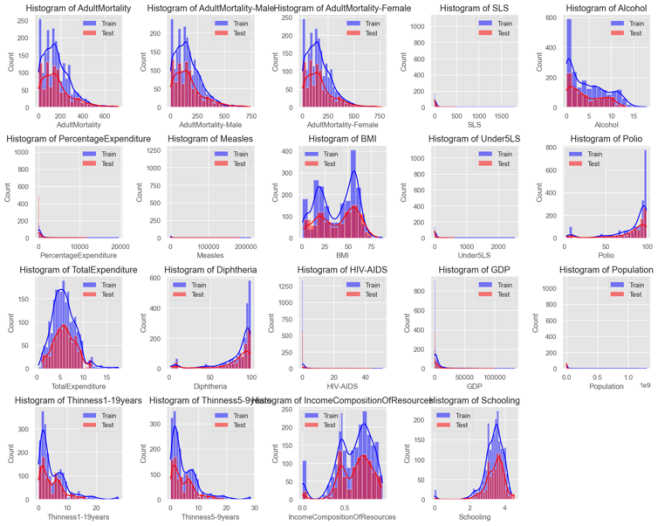
```
Train dataset shape: (2071, 24)
Test dataset shape: (867, 23)
```

```
Missing values in Train dataset:
ID                              0
TARGET_LifeExpectancy           0
Country                         0
Year                            0
Status                          0
AdultMortality                  0
AdultMortality-Male             0
AdultMortality-Female           0
SLS                             0
Alcohol                         0
PercentageExpenditure           0
Measles                         0
BMI                             0
UnderSLS                        0
Polio                           0
TotalExpenditure                0
Diphtheria                      0
HIV-AIDS                        0
GDP                             0
Population                      0
Thinness1-19years               0
Thinness5-9years                0
IncomeCompositionOfResources    0
Schooling                       0
dtype: int64
```

```
Missing values in Test dataset:
ID                              0
Country                         0
Year                            0
Status                          0
AdultMortality                  0
AdultMortality-Male             0
AdultMortality-Female           0
SLS                             0
Alcohol                         0
PercentageExpenditure           0
Measles                         0
BMI                             0
UnderSLS                        0
Polio                           0
TotalExpenditure                0
Diphtheria                      0
HIV-AIDS                        0
GDP                             0
Population                      0
Thinness1-19years               0
Thinness5-9years                0
IncomeCompositionOfResources    0
Schooling                       0
dtype: int64
```

Data Exploration & Initial Findings:

Train.csv	2071 rows, 24 columns	Composed of all the columns.
Test.csv	867 rows, 23 columns	Missing the LifeExpectancy Column that needs to be predicted



The attached image displays 19 histogram plots, each representing a continuous variable shared between the train and test datasets. The blue histograms represent the training data, while the red histograms represent the test data. The overlaid distributions in most of the plots suggest that most of the variables have similar spreads across both datasets. This observation implies that the train and test datasets are well-balanced and share comparable characteristics for these continuous features, which is a positive indicator for building a robust predictive model.

Data Cleaning by Value Substitution of predictable demographic metrics (using linear regression): In the histograms shown above, there are certain demographic metrics for which the value is impossible to be 0. (10) Especially economic expenditure, and income metrics that follow a steady, linear trend over the span of 16 years. (10) Given that none of the measurements are from over 100 years ago it is impossible for metrics like Schooling, Expenditure, and Income to be 0 (which skews the robustness of the model). (10) Metrics like these include:

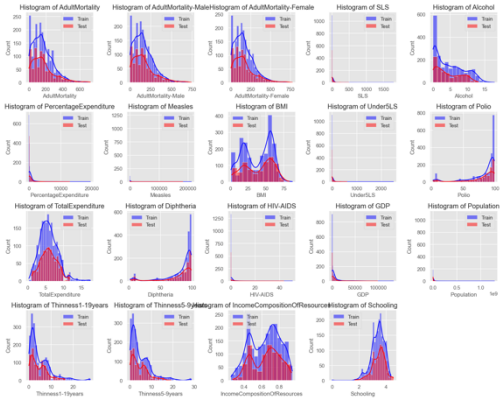
a. PercentageExpenditure
b. TotalExpenditure
c. GDP
d. IncomeCompositionOfResources
e. Schooling

To solve this issue, linear regression can be employed to predict the missing values based on the values that we know for each of the countries. (8) To achieve this, we follow these steps:

a. Iterate across each country with 16 rows (one row for each year)
b. Identify column values that are zero
c. If there are zeros present:
    a. Split data into non-zeroes (known) and zeroes (unknown)
    b. If less than two known values, replace the zero with global mean of column
d. If enough known values,
    a. Fitting linear regression model with known values where the year is the independent variable and the column value is the dependent variable (11)
    b. Predict unknown values with the corresponding year. (11)
    c. Replace the zeros with the predicted values

Individual Assignment 1
COSC2793
S3970066
Amay Viswanathan Iyer

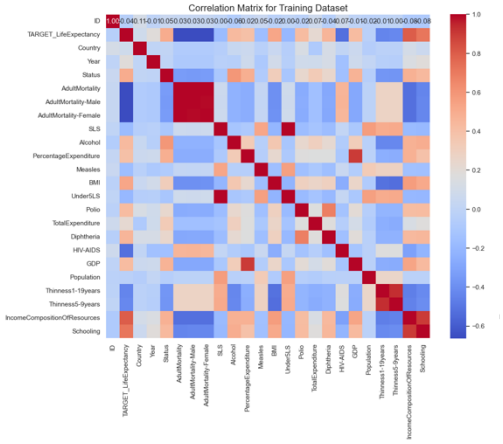
e. Repeating for training and test datasets.



The rationale underlying this is that these columns are expected to have a consistent yearly trend for each country, as supported by reputable sources such as the United Nations (UN) and human development index studies. (10) By leveraging the known values and the year information, the linear regression model can provide reasonable estimates for the missing values, allowing for a more accurate representation of the data. (9) This data cleaning and value substitution process aims to improve the quality of the dataset by replacing anomalous zero values with more plausible estimates based on the historical trends. By doing so, it helps to mitigate the impact of missing or erroneous data points on the overall statistics and subsequent analyses or modeling tasks. (9)

This method is not without its drawbacks. It is simply assumed that there exists a linear relationship between the year and the column values, which may not always be the case. (12) Therefore, it's crucial to validate the assumptions and assess the appropriateness of the linear regression model for each column and country combination. (13) Overall, the linear substitution portion demonstrates a data cleaning technique that utilizes linear regression to handle missing values in specific columns, considering the yearly trends and country-specific patterns, with the goal of improving the quality and reliability of the dataset for further analysis and modeling in the context of human development indicators. (9) The new spread is displayed in the accompanying histogram after the linear substitutions sadly, given that a large proportion of countries in the dataset aren't very economically developed (biased towards one 'Status'), some of the metrics are still skewed to one side. (10)

EDA Training Correlative Analysis:



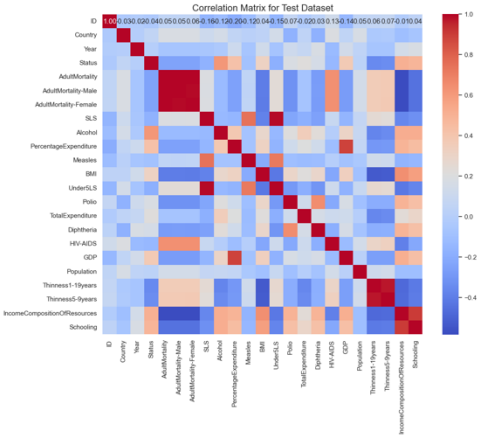
Strongly Positive Correlations:

- TARGET_LifeExpectancy has a strong positive correlation with IncomeCompositionOfResources (0.798) and Schooling (0.716). This suggests that higher income levels and better education are significantly associated with longer life expectancy, which is consistent with broader socioeconomic health determinants. (1)
- Diphtheria and Polio vaccination rates are positively correlated (0.688), indicating that countries with better vaccination coverage for one often have better coverage for the other. (2)

Strongly Negative Correlations:

- TARGET_LifeExpectancy is negatively correlated with AdultMortality, AdultMortality-Male, and AdultMortality-Female (around -0.66). (3) This indicates that higher mortality rates in adults are associated with lower life expectancy, which is a straightforward health indicator.
- HIV-AIDS prevalence negatively impacts TARGET_LifeExpectancy (-0.522), underscoring the disease's severe impact on public health. (4)

EDA Test Correlative Analysis:



Strong Positive Correlations:

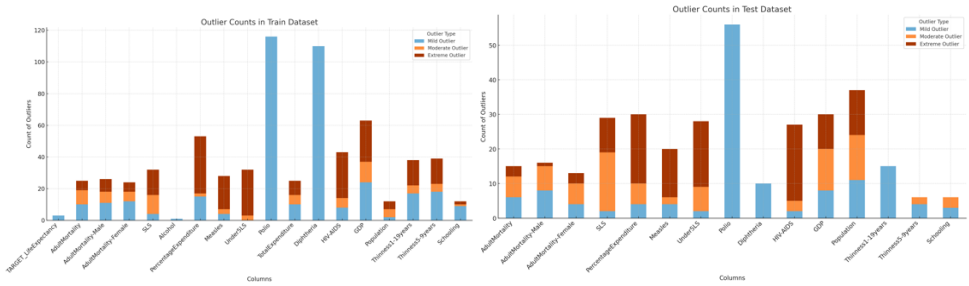
- Like the training set, IncomeCompositionOfResources and Schooling show a very high correlation with each other (0.904) and significant positive correlations with life expectancy proxies such as BMI (0.646 and 0.582 respectively), indicating that higher educational attainment and income levels are associated with better health outcomes.

Strong Negative Correlations:

- AdultMortality rates continue to show strong negative correlations with IncomeCompositionOfResources and Schooling (around -0.58), underscoring the impact of socioeconomic factors on mortality.

Individual Assignment 1
COSC2793
S3970066
Amay Viswanathan Iyer

Feature Scaling: Outlier handling by Winsorization:



The outliers were categorized into three distinct categories based on their z-scores: mild outliers (z-score between 3 and 3.5), moderate outliers (z-score between 3.5 and 4), and extreme outliers (z-score above 4). This categorization allowed for a more nuanced understanding of the severity of the outliers in the dataset. By setting specific thresholds for each category, I was able to identify and differentiate between outliers that deviated moderately from the central tendency (mild and moderate outliers) and those that were far more extreme in nature. This approach provided a clear framework for assessing the impact of outliers on the dataset and guided my decision-making process regarding outlier handling techniques. (5)

These are some of the inferences we can make from the outliers:

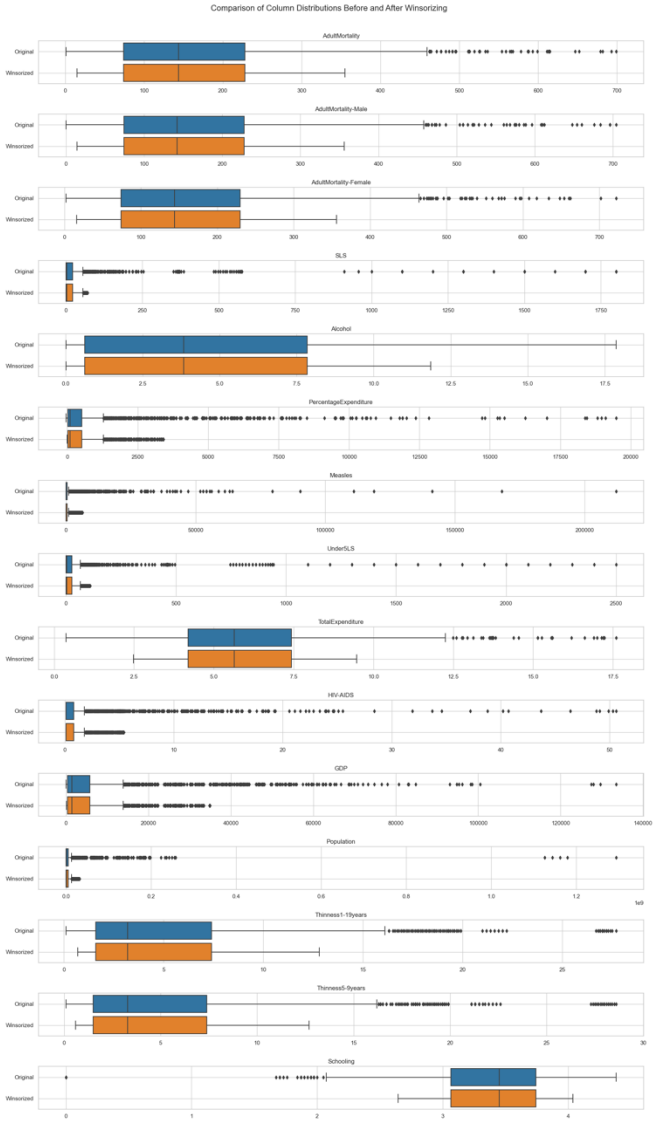
- Mild outliers in the TARGET_LifeExpectancy column suggest that some countries have exceptionally low life expectancies compared to most of the data.
- The strong negative correlation between life expectancy and adult mortality rates, along with the outliers in these variables, suggests that countries with exceptionally high mortality rates tend to have lower life expectancies.
- The negative correlations between immunization rates and disease cases, along with the outliers in these variables, demonstrate the effectiveness of vaccination in preventing outbreaks and the vulnerability of certain regions to disease spread.

My rationale to Winsorize only the Extreme Outliers:

- Initially, my intention was to winsorize extreme outliers and transform mild and moderate outliers to handle the large number of outliers present in the dataset. However, the process of transforming mild and moderate outliers proved to be too complicated and potentially distorted the data distribution (I have kept the jupyter notebook snippets towards the end of the notebook).
- Therefore, the decision was made to focus on winsorizing only the extreme outliers, defined as those with a z-score above 4. By targeting the most extreme values, the aim was to mitigate the impact of severe outliers on the overall statistics and subsequent analyses while preserving the general distribution of the data.

Impact of Winsorization (Winsorized Training data spread shown on right):

- Although the winsorization was applied to both the training and test set, it seems from the visuals that there was a much more significant effect on the training set after stricter limits were tried (when compared to effect on test set).
- The mean values of the winsorized columns shifted closer to the central tendency, reducing the effect of extreme outliers. The standard deviations decreased, indicating a reduction in the variability of the data after winsorization.
- The minimum and maximum values were capped at the specified percentiles, limiting the range of extreme values. The median values (50th percentile) remained relatively stable, suggesting that winsorization preserved the central tendency of the data.

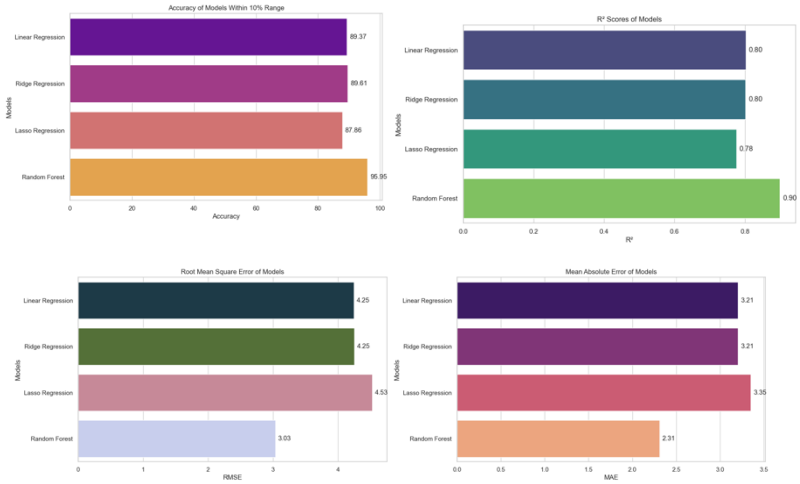


Individual Assignment 1
COSC2793
S3970066
Amay Viswanathan Iyer

Model Selection and Evaluation:

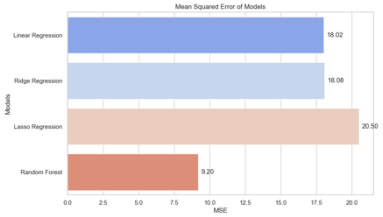
Initial Steps of Selection

a. I began by defining a dictionary of models including Linear Regression, Ridge Regression, Lasso Regression, and Random Forest Regression for which, I used cross-validation with 5 folds to evaluate each model's performance on the training data, calculating metrics such as RMSE, R^2, MAE, and a custom accuracy metric (percentage of predictions within 10% of the actual values).
b. I then visualized the results using bar plots to compare the performance of the models across different metrics and based on the initial evaluation, I narrowed down the focus to Ridge Regression and Lasso Regression for further analysis, as they showed promising results and are regularized versions of Linear Regression.



Validation Setup and Performance

a. I split the training data into features (X) and the target variable (y) for life expectancy and then used the train_test_split function from scikit-learn to split the training data into training and validation sets, with 20% of the data reserved for validation.
b. I then applied cross-validation with 5 folds to evaluate the models, ensuring a robust assessment of their performance them being; RMSE, MSE, MAE, R^2, and Accuracy to assess the model's performance.

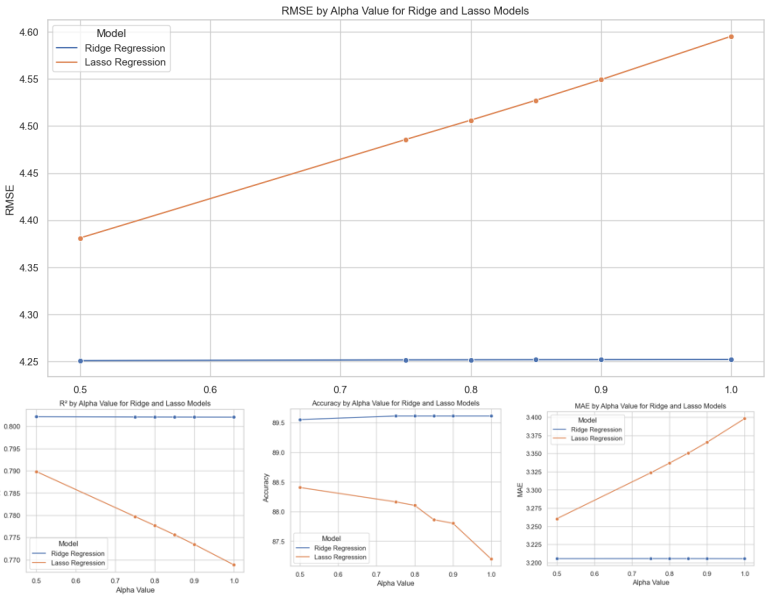


Regularization and Justification:

a. Ridge Regression and Lasso Regression are regularized linear regression techniques that add a penalty term to the ordinary least squares (OLS) objective function. Ridge Regression adds an L2 penalty term, which shrinks the coefficients towards zero, while Lasso Regression adds an L1 penalty term, which can shrink some coefficients to exactly zero. (6)
b. I justified the use of regularization to handle potential multicollinearity in the dataset, prevent overfitting, and improve the model's generalization performance. Regularization helps to control the model's complexity by constraining the magnitude of the coefficients, leading to simpler and more interpretable models. (6)

Hyperparameter-Tuning:

a. I identified the hyperparameter 'alpha', which controls the strength of regularization, with higher values resulting in stronger regularization, as the key parameter to tune for Ridge Regression and Lasso Regression. (7)
b. I defined a range of alpha values to test: [1.0, 0.9, 0.5, 0.75, 0.8, 0.85] and used cross-validation with 5 folds to evaluate the performance of Ridge Regression and Lasso Regression models for each alpha value, calculating metrics such as RMSE, R^2, MAE, and custom accuracy. (7)
c. I also visualized the results using line plots to analyze the impact of different alpha values on the model's performance. (7)
d. As displayed on the right, as the alpha approached 1.0 from 0.5, the accuracy % of the Ridge Regression model greatly improved and stabilized.
e. The RMSE, R^2, and MAE remained stable for the Ridge Regression model as the alpha approached 1.0.



Individual Assignment 1
COSC2793
S3970066
Amay Viswanathan Iyer

Prediction and Performance:

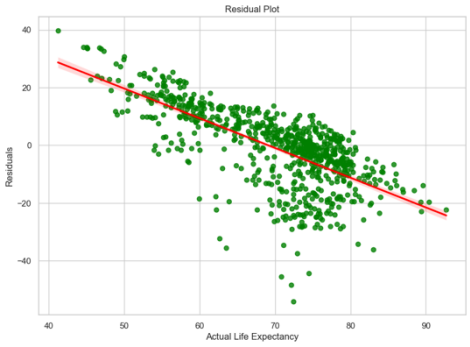
a. Once I chose Ridge as the ideal model, it was trained with its respective alpha value across the entire dataset to predict the Life Expectancy for the test dataset.

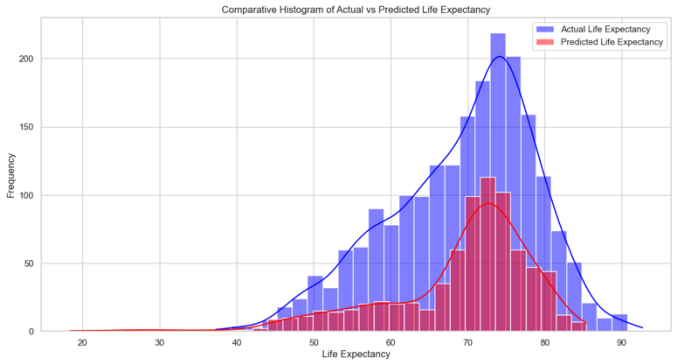
Evaluation Methods and Justification

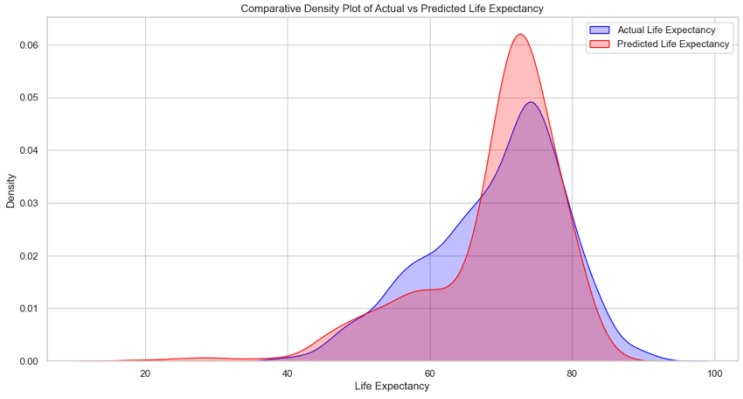
a. RMSE and MAE measure the average magnitude of errors present whereby the lower the value, the better.
b. R^2 values represent the proportion of variance explained by each of the features (which were scaled by winsorizing).
c. The custom accuracy metric computes the percentage of predictions with a 10% threshold of values, providing an intuitive measure of accuracy.
d. Ridge Regression demonstrated a stability in its performative robustness across fluctuating alpha values. The value of 1.0 enabled the right balance between complexity and performance.

Additional Linear Regression Justifications by addressing assumptions:

a. As per the initial assumptions of linear regression, the Ridge model also aligns with linearity, homoscedasticity, independence, and normality of residuals. (12) Below is the Residual plot exemplifying the normality of Residuals for our Ridge Regression Model when applied to predict Life Expectancy.
b. Finally, when the Predicted Life Expectancy spread is overlaid against the Training set's spread, it shows a fairly equivalent distribution of values by their respective frequency.
c. The peak density of the predicted values reaches much higher than the actual values indicating the limitation of Ridge Regression's ability to predictably follow a spread.







s3970066

ID	Predicted_LifeExpectancy
1	60.630698794724600
2	60.38454306213140
3	59.367084136470400
4	58.37196174304110
5	58.07647983015820

Individual Assignment 1
COSC2793
S3970066
Amay Viswanathan Iyer

References:

(1) Socioeconomic determinants of health and life expectancy: Source: Marmot, M. (2005). Social determinants of health inequalities. The Lancet, 365(9464), 1099-1104. https://doi.org/10.1016/S0140-6736(05)71146-6

(2) Vaccination coverage and disease prevention: Source: Andre, F. E., Booy, R., Bock, H. L., Clemens, J., Datta, S. K., John, T. J., ... & Schmitt, H. J. (2008). Vaccination greatly reduces disease, disability, death and inequity worldwide. Bulletin of the World Health Organization, 86, 140-146. https://doi.org/10.2471/BLT.07.040089

(3) Adult mortality rates and life expectancy: Source: Oeppen, J., & Vaupel, J. W. (2002). Broken limits to life expectancy. Science, 296(5570), 1029-1031. https://doi.org/10.1126/science.1069675

(4) HIV/AIDS impact on life expectancy: Source: Bor, J., Herbst, A. J., Newell, M. L., & Bärnighausen, T. (2013). Increases in adult life expectancy in rural South Africa: valuing the scale-up of HIV treatment. Science, 339(6122), 961-965. https://doi.org/10.1126/science.1230413

(5) Outlier detection and handling techniques: Source: Aguinis, H., Gottfredson, R. K., & Joo, H. (2013). Best-practice recommendations for defining, identifying, and handling outliers. Organizational Research Methods, 16(2), 270-301. https://doi.org/10.1177/1094428112470848

(6) Ridge regression and regularization: Source: Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67. https://doi.org/10.1080/00401706.1970.10488634

(7) Hyperparameter tuning and cross-validation: Source: Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95) (pp. 1137-1143). Morgan Kaufmann Publishers Inc.

(8) Linear regression for missing data imputation: Source: Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). Review: a gentle introduction to imputation of missing values. Journal of clinical epidemiology, 59(10), 1087-1091. https://doi.org/10.1016/j.jclinepi.2006.01.014

(9) Importance of data cleaning in machine learning: Source: García, S., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining. Springer. https://doi.org/10.1007/978-3-319-10247-4

(10) United Nations data on human development indicators: Source: United Nations Development Programme. (2020). Human Development Report 2020. The next frontier: Human development and the Anthropocene. United Nations Development Programme. http://hdr.undp.org/en/2020-report

(11) Handling missing data in longitudinal studies: Source: Engels, J. M., & Diehr, P. (2003). Imputation of missing longitudinal data: a comparison of methods. Journal of clinical epidemiology, 56(10), 968-976. https://doi.org/10.1016/s0895-4356(03)00170-7

(12) Assumptions of linear regression: Source: Osborne, J., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. Practical assessment, research, and evaluation, 8(1), 2. https://doi.org/10.7275/r222-hv23

(13) Assessing the appropriateness of linear regression models: Source: Chatterjee, S., & Hadi, A. S. (2015). Regression analysis by example (5th ed.). John Wiley & Sons. https://doi.org/10.1002/9781119202271