Final Project Report

Introduction

In this analysis, we aim to understand the impact of various factors such as GDP, adult mortality, alcohol consumption, and Hepatitis B vaccinations on the life expectancy of populations across different countries. This question is crucial for guiding resource allocation and policy formulation in the context of public health. By leveraging data analytics, specifically through regression models, we can provide actionable insights that policymakers and healthcare organizations can use to enhance health outcomes and improve quality of life globally. Given the constraints on resources and the varying health challenges across regions, this analysis is significant as it aids in prioritizing interventions and allocating funds more effectively.

Data Preprocessing

1. Dropping Null Rows for Life Expectancy and Adult Mortality

Life Expectancy and Adult Mortality: Both are crucial variables for any analysis involving health data. Since only 10 rows are missing these values, which is likely a small fraction of the entire dataset, removing these rows will have a minimal impact on the dataset's size and the statistical power of any analyses. Additionally, imputing these could introduce bias if not done carefully, especially since both variables are likely to be closely related to many other variables in the dataset.

2. Replacing Nulls with Mean for Alcohol, GDP, and Hepatitis B

Alcohol: Given that alcohol consumption data might be missing at random or not reported uniformly across all countries, replacing missing values with the mean can be a reasonable approach to maintain the dataset size and representativeness, particularly if alcohol consumption varies significantly from country to country.

GDP: The economic metric might be missing due to reporting differences or data availability issues. Using the mean GDP might not always reflect true economic conditions, especially for countries significantly poorer or richer than average. However, if detailed economic modeling isn't the primary focus of your analysis, mean imputation provides a straightforward method to retain these rows for other analyses involving health metrics.

Hepatitis B: A significant number of rows (553) are missing Hepatitis B data, suggesting challenges in data collection or reporting. Replacing these missing values with the mean ensures that the analysis can still proceed and leverages the available data, albeit with some potential loss of specificity in vaccination or infection rates.

Descriptive Analysis

To conduct a descriptive analysis of the data regarding the impact of various factors on life expectancy, we'll first summarize the main characteristics of the dataset. Based on the provided preprocessing steps, we have information on life expectancy, adult mortality, alcohol consumption, GDP, and Hepatitis B vaccinations across different countries.

1. Data Summary:

- Life Expectancy: The primary outcome variable reflecting the average lifespan in years.

- Adult Mortality: Reflects the probability of dying between the ages of 15 and 60 per 1000 population.

- Alcohol Consumption: Represents the average alcohol consumption per capita in liters of pure alcohol.

- GDP: Gross Domestic Product per capita, indicating the economic status of countries.

- Hepatitis B Vaccinations: Coverage or prevalence of Hepatitis B vaccinations.

2. Data Preprocessing:

- Handling Null Values:

- Null rows for life expectancy and adult mortality were dropped due to their critical importance, with only 10 rows affected.

- Null values for alcohol, GDP, and Hepatitis B were replaced with the mean to maintain dataset integrity.

3. Descriptive Statistics:

- Life Expectancy:

- Mean, median, minimum, and maximum life expectancy values.

- Distribution characteristics (skewness, kurtosis).

- Alcohol Consumption:

- Average alcohol consumption across countries.

- Distribution summary.

- GDP:

- Mean GDP per capita.

- Range of GDP values.

- Hepatitis B Vaccinations:

- Mean coverage or prevalence rate.

- Distribution details.

4. Correlation Analysis:

- Investigate correlations between life expectancy and other factors.

- Identify any significant correlations that could inform predictive modeling.

5. Visualization:

- Generate visual representations (scatter plots, histograms, box plots) to understand data distributions and relationships.

- Plot correlations matrices to visualize relationships between variables.

6. Insights and Anomalies:

- Identify any outliers or anomalies in the data.

- Discover patterns or trends that may indicate significant relationships between variables.

- Explore whether the imputation of missing values introduces bias or impacts the analysis.

7. Potential Predictive Analysis:

- Use regression models to predict life expectancy based on factors like GDP, adult mortality, alcohol consumption, and Hepatitis B vaccinations.

- Evaluate model performance and identify key predictors affecting life expectancy.

By conducting this descriptive analysis, we aim to gain insights into the relationships between various factors and life expectancy, which will guide subsequent predictive modeling for informing public health policies and resource allocation strategies.

Predictive Analysis using Snowflake Cortex ML Functions

Given the objective of understanding the impact of various factors on life expectancy across different countries and the provided data preprocessing steps, a suitable machine learning approach would be prediction.

The objective of the analysis is to understand the relationship between predictors (GDP, adult mortality, alcohol consumption, Hepatitis B vaccinations) and the target variable (life expectancy). Regression models are well-suited for predicting a continuous target variable based on one or more predictor variables.

The provided data preprocessing steps, such as dropping null rows for crucial variables like life expectancy and adult mortality, and imputing missing values for other variables, align well with the requirements of prediction modeling. These preprocessing steps ensure that the dataset is cleaned and ready for predictive modeling.

Regression models provide interpretable results, allowing policymakers and healthcare organizations to understand the impact of each predictor variable on life expectancy. This interpretability is crucial for guiding resource allocation and policy formulation, as it helps in identifying which factors have the most significant influence on improving health outcomes.

By predicting life expectancy based on various factors, policymakers can prioritize interventions and allocate resources more effectively. For example, if the model indicates that GDP has a significant positive effect on life expectancy, policymakers may focus on improving economic conditions in regions with lower GDP to enhance health outcomes.

Prediction models can provide actionable insights that can be used to develop targeted interventions and policies aimed at improving public health. For instance, if the model predicts that higher Hepatitis B vaccination rates are associated with increased life expectancy, policymakers can invest in vaccination programs to prevent Hepatitis B-related mortality and improve life expectancy.

Therefore, leveraging prediction modeling, particularly regression analysis, aligns well with the project objective of understanding the impact of various factors on life expectancy and providing actionable insights for policymakers and healthcare organizations.

Summary/Conclusion

This analysis explores the relationship between various factors, including GDP, adult mortality, alcohol consumption, and Hepatitis B vaccinations, and life expectancy across different countries. Understanding these relationships is essential for informing public health policies and resource allocation strategies. By employing regression models, we can derive actionable insights to improve global health outcomes.

Null rows for life expectancy and adult mortality were dropped due to their critical importance in health analyses. Imputing these values could introduce bias, considering their close relationship with other variables.

For alcohol consumption, GDP, and Hepatitis B, missing values were replaced with the mean. This approach maintains dataset size and representativeness while acknowledging the variability of these factors across countries. However, it's important to note potential limitations, especially in representing economic conditions accurately and loss of specificity in Hepatitis B vaccination rates.

Conclusion: This analysis provides a foundation for understanding the impact of various factors on life expectancy globally. By addressing missing data and employing robust statistical methods, we can generate insights crucial for prioritizing interventions and optimizing resource allocation in public health. Future work could involve exploring additional factors or refining modeling techniques to enhance the accuracy and relevance of findings.