Data Visualiaztion using pySpark

*

1st Dev Wankhede, 2nd Tapan Mahata, 3rd Diganta Diasi
Indian Institute of Technology Guwahati
Guwahati, ASSAM

I. Introduction

Introduce the objective of the report, which is to analyze various indicators related to child nutrition and health using data from the "Health Report" dataset. The report focuses on exploring relationships between different health metrics and applying machine learning techniques for analysis. PySpark enables efficient handling of large child nutrition datasets, utilizing distributed computing for parallel processing. Integrated with MLlib, it supports advanced analytics like regression and clustering, providing valuable insights into child health. By streamlining data preprocessing, modeling, and visualization, PySpark empowers researchers to make evidence-based decisions for improving child nutrition outcomes.

II. Description of Dataset

The dataset consists of standardized data related to various indicators of child and adolescent health and nutrition across different states and regions of India. The data is organized by country, state, year, year code, residence type, and various health and nutrition metrics. The dataset provides valuable insights into the nutritional status and health conditions of children and adolescents in different regions of India. Analyzing this data can help identify areas that require targeted interventions and resources to address malnutrition and improve overall health outcomes among children and adolescents.

III. Handling Missing Data

Before performing analysis and visualization, missing data was identified and handled appropriately. The following steps were taken: Identifying Missing Values: The number of missing values in each column was calculated using Spark SQL functions. Handling Missing Values: Missing values were handled differently for numerical and categorical columns. For numerical columns, missing values were imputed with the mean value using the Imputer class from the pyspark.ml.feature module. For categorical columns, missing values were replaced with the most frequent value.

IV. Technology Stack Overview

Frontend: React React offers a component-based approach for building dynamic user interfaces, ensuring scalability and code maintainability.

Backend: Flask and PySpark

Flask: A lightweight Python web framework for developing RESTful APIs, providing flexibility and ease of integration. PySpark: The Python API for Apache Spark enables scalable data processing, feature engineering, and machine learning tasks with efficiency. Database: MongoDB MongoDB, a NoSQL database, stores JSON-like documents and offers high scalability and flexibility, ideal for managing diverse datasets.

Project Summary: Combining React for the frontend, Flask and PySpark for the backend, and MongoDB for the database, our project delivers a robust solution for exploring child and adolescent health and nutrition data in India.

V. Data visualize using various plot and Algorithm

A. Histogram

The histogram function provided here offers a visual representation of the distribution of various child malnutrition indicators. Utilizing a color palette reminiscent of a clear sky and bordered by a sleek black outline, these histograms draw attention to the frequency of occurrences within each bin. With 20 bins strategically arranged, the histograms encapsulate the prevalence of malnutrition, painting a vivid picture of its impact on child health. Each histogram is tailored to a specific malnutrition indicator, enabling a focused examination of the distribution patterns. Through this visual exploration, we gain valuable insights into the extent and severity of malnutrition among children under the age of 5 years.

B. Scatter Plot

The scatter plots depicted here offer a visually appealing representation of the relationship between pairs of child malnutrition indicators. Each scatter plot provides a clear visualization of the correlation, if any, between two specific indicators. Through the strategic use of color and transparency, these plots effectively convey the density and distribution of data points across the range of values for each indicator. By plotting these indicators against each other, potential associations or trends between different malnutrition categories can be discerned, aiding in the identification of potential areas for further investigation or intervention. With their intuitive design and comprehensive depiction of data relationships, these scatter plots serve as valuable tools for exploring and analyzing patterns in child malnutrition prevalence.

C. Box Plot

The box plot visualizations showcased here offer a succinct overview of the distribution and variability in various child malnutrition indicators. Each box plot is carefully crafted to encapsulate key statistical measures such as the median, quartiles, and outliers. With a focus on clarity and precision, these plots provide valuable insights into the distributional characteristics of each indicator. Through their concise yet informative presentation, they enable a quick and intuitive comparison across different malnutrition categories. By leveraging the power of visualization, these box plots serve as effective tools for identifying potential trends, patterns, and disparities in child malnutrition prevalence.

D. Pie Chart

The pie charts showcased above offer a concise yet insightful visualization of the distribution of child malnutrition indicators across different categories. By grouping the data based on specified criteria such as residence type or state, these charts provide a clear breakdown of the proportion of malnourished children within each category. The use of vibrant colors and percentage labels enhances the readability and interpretability of the charts, allowing viewers to quickly grasp the relative prevalence of malnutrition indicators within each group. Through these visually appealing pie charts, key patterns and disparities in malnutrition prevalence among different demographic segments can be easily identified, facilitating targeted interventions and policy decisions aimed at addressing the most vulnerable populations

E. Correlation Matrix

The correlation matrix heatmap above provides a succinct overview of the relationships between different variables in the dataset. Each cell in the heatmap represents the correlation coefficient between two variables, with values ranging from -1 to 1. A coefficient closer to 1 indicates a strong positive correlation, while a coefficient closer to -1 suggests a strong negative correlation. Values near 0 indicate little to no linear relationship between the variables. The use of color gradients and annotation makes it easy to identify patterns and associations within the dataset at a glance. This visualization aids in understanding the interdependencies among variables, guiding further analysis and decision-making processes.

F. Linear Regression Analysis

The linear regression analysis conducted in the above code explores the relationship between various indicators related to child malnutrition under the age of 5 years. Utilizing the selected columns, the study investigates how factors such as stunted children, wasted children, and underweight children influence severe malnutrition indicators.

The process begins by assembling the features and selecting the target column for each regression analysis. Training and test datasets are then split for model evaluation. Linear regression models are fitted to the training data, and predictions are made for the test data. The scatter plots visualize the relationship between the selected indicator and the corresponding severe malnutrition indicator. The plots showcase how changes in stunted, wasted, or underweight children percentages impact severe malnutrition levels, providing insights into potential intervention strategies.

G. K-means Clustering Analysis

This analysis focuses on clustering children under 5 years based on their percentage of stunted and severely stunted conditions. After assembling and scaling the feature vectors, K-means clustering is applied to identify distinct clusters within the dataset. The resulting clusters are visualized using a scatter plot, with each cluster representing a different combination of stunted and severely stunted children. This visualization aids in understanding the distribution of stunted and severely stunted children across different groups, providing insights into the prevalence and severity of malnutrition among children under 5 years

H. GMM Clustering

The Gaussian Mixture Model (GMM) clustering algorithm offers valuable insights into the dataset's underlying structure. By selecting pertinent columns like 'Severely thin adolescents age group of 10 to 14 years less than -3 SD (%)' and 'Severely thin adolescents age group of 15 to 19 years less than -3 SD (%)', we aim to discern patterns and groups related to adolescent thinness.

After assembling and scaling the features, the GMM algorithm partitions the data into clusters, with 3 clusters predefined in this case. The resulting clusters are depicted in a scatter plot, where each point represents a data point colored by its assigned cluster. Additionally, centroids of each cluster are marked for reference.

The plot offers a concise overview of the clustering outcome, illustrating how data points are grouped into distinct clusters based on their characteristics concerning adolescent thinness. This analysis aids in comprehending the distribution and composition of different subgroups within the dataset, facilitating further exploration and interpretation of the data's underlying structure.

I. Line Plot Visualization

The visualizations above depict the trend of stunted children under the age of 5 years and severely stunted children over the years. Each line plot illustrates the respective percentage of children affected by stunting, with the x-axis representing the years and the y-axis indicating the percentage. The plots provide a clear depiction of how these percentages have evolved over time, aiding in understanding the temporal patterns of child malnutrition. The upward or downward trends observed in the plots can inform policymakers and stakeholders about the effectiveness of interventions and the need for further action in addressing childhood malnutrition.

article hyperref

Please visit our GitHub repository for the project details: https://github.com/tapan0p/Data-Visualisation-in-Pyspark-NITIAYOG-Health-dataset