PROJECT MILESTONE 2

Petrus Human 577842

Frederik Knoetze 600965

Teleki Shai 601377

Moloko Rakumako 601352

Module: Business Intelligence 381
Project: Health & Demographic Patterns in South Africa (HDPSA)
Milestone 2: Data Preparation

BELGIUM CAMPUS ITVERSITY

Table of Contents

Introduction 2
Overall Project Goals 2
Milestone 2 Data Preparation Goals 2
R Script Descriptions 3
Milestone_2_combine_datasets.R 3
Milestone_2_data_cleaning.R 3
Milestone_2_feature_selection.R 4
Data Transformation: Column Comparison 5
Raw Data Columns Example from access-to-health-care_national_zaf.csv 5
Final Processed Data Columns Example from feature_selected_cleaned_combined_dataset.csv 6
Columns Removed During Processing 6
Rationale for Final Column Selection 7
Data Preparation Considerations for Modelling 9
Train-Test Split Strategy with Limited Time Points 9
Data Scaling and Normalization 10
Potential for Enhanced Applications with More Extensive Datasets 10
Conclusion 12

Introduction

This document provides an overview of the R scripts developed for the Health & Demographic Patterns in South Africa (HDPSA) project, specifically addressing the data preparation phase for Milestone 2. These scripts are designed to transform raw data into a clean, pre-processed, and feature-selected format suitable for future modelling and data analysis, aligning with the project's goals for business understanding and data quality.

Overall Project Goals

The primary data mining goals for the HDPSA project are to identify variables correlating with child mortality, understand how changes in variables over time impact child mortality, and create a model to predict child mortality. Key success criteria for data preparation (Phase 2) include:

- Data is cleaned and pre-processed (missing values handled, duplicates removed, outliers addressed). This directly aligns with the functionality of the R scripts.
- Model-ready datasets with <5% unprofiled missingness on critical KPIs. (This is an outcome goal that the scripts contribute to).
- The Milestone 1 analysis highlighted data quality issues such as missing data (14.98% to 23.67% range), systematic missing patterns, and the need for careful handling of duplicates and outliers.

Milestone 2 Data Preparation Goals

Based on the Milestone 2 instructions, the R scripts contribute significantly to the following key objectives:

- Select Data: This involves focusing on indicators with sufficient data coverage and utilizing `IsPreferred` flags for data selection.
- Verify Data Quality: This encompasses implementing appropriate missing data treatment, standardizing formats, validating ranges, and ensuring temporal consistency.

R Script Descriptions

Milestone_2_combine_datasets.R

Purpose: To consolidate all individual raw CSV datasets into a single, unified dataset, ensuring a consistent and clean starting point for further processing.

Functionality:

- Reads all CSV files located in the `Raw Dataset's folder.
- Identifies and extracts the true header (column names) from the first CSV file.
- Reads all subsequent CSV files, explicitly skipping the first two rows (the actual header and a metadata row present in each source file) to prevent duplicate headers from becoming data rows.
- Combines all data into a single data frame.
- Assigns the extracted true header to the combined data.
- Removes any existing `combined_dataset.csv` to ensure a fresh output.
- Saves the resulting unified dataset as `combined_dataset.csv` in the `Cleaned Dataset's folder.

Relation to Data Quality (Milestone 2): This script is foundational for data quality by ensuring all raw data is brought together consistently. It directly addresses the "Standardize formats" requirement by handling disparate raw files and resolving the initial structural quality issue of multiple header rows in source files, thus preparing the data for further quality verification. It also contributes to "Temporal consistency" by combining data across different survey years into a single structure.

Milestone_2_data_cleaning.R

Purpose: To clean and preprocess the combined dataset, addressing critical data quality issues as outlined in Milestone 2 instructions.

Functionality:

- Reads the `combined_dataset.csv` from the `Cleaned Dataset's folder.
- Removes Duplicate Rows: Directly addresses the "duplicates removed" success criteria by identifying and eliminating exact duplicate rows across the entire dataset.
- Imputes Missing Values: Implements a "Missing data strategy" and "Handle missing values" by performing imputation for missing data points. For numerical columns, missing values are replaced with the mean of that column

within groups defined by 'Indicator' and 'SurveyYear. For categorical columns, missing values are replaced with the mode (most frequent value) within groups defined by 'Indicator' and 'SurveyYear'This group-wise approach ensures context-aware imputation.

- Handles Outliers: Directly addresses the "outliers addressed" success criteria. Outliers in numerical columns are capped using the Interquartile Range (IQR) method (values falling outside 1.5 times the IQR from the first (Q1) or third (Q3) quartiles are capped at the respective bounds). This capping is performed within groups defined by 'Indicator' and 'SurveyYear', contributing to "Validate ranges" by ensuring data falls within reasonable, group-specific limits.
- Saves the thoroughly cleaned and preprocessed dataset as `cleaned_combined_dataset.csv` in the `Cleaned Datasets` folder.

Relation to Data Quality (Milestone 2): This script is central to the "Verify Data Quality" objective. It systematically handles missing values, duplicates, and outliers, ensuring the dataset is robust, reliable, and adheres to the quality standards required for analysis. The group-wise cleaning methods enhance the integrity and contextual accuracy of the data.

Milestone_2_feature_selection.R

Purpose: To select a subset of features (columns) and rows that are most relevant for future modelling and data analysis, based on predefined criteria from Milestone 2.

Functionality:

- Reads the `cleaned_combined_dataset.csv` from the `Cleaned Dataset's folder.
- Filters Rows: Directly implements the "Quality flags: Utilize IsPreferred flags for data selection" and "Indicator selection: Focus on indicators with sufficient data coverage" requirements by retaining only those rows where the `IsPreferred` column has a value of `1`. This ensures that only preferred and relevant data points are carried forward.
- Selects Columns: Keeps only the `Indicator`, `Value`, and `SurveyYear` columns. This aligns with focusing on key variables for analysis and reducing dimensionality, contributing to efficient data selection.
- Saves the filtered and column-selected dataset as `feature_selected_cleaned_combined_dataset.csv` in the `Cleaned Datasets` folder.

Relation to Data Selection (Milestone 2): This script directly addresses the "Select Data" requirement by focusing on preferred indicators and reducing the dataset to essential

variables. By using the `IsPreferred` flag, it ensures that the selected data is of higher quality and relevance for subsequent modelling phases.

Data Transformation: Column Comparison

To illustrate the data transformation process, a comparison between the columns of a raw dataset and the final `feature_selected_cleaned_combined_dataset.csv` is provided below. This highlights the columns that have been retained and those that were removed during the data preparation and feature selection phases.

Raw Data Columns Example from access-to-health-care_national_zaf.csv

- `ISO3`
- `DataId`
- `Indicator`
- `Value`
- `Precision`
- `DHS_CountryCode`
- `CountryName`
- `SurveyYear`
- `SurveyId`
- `IndicatorId`
- `IndicatorOrder`
- `IndicatorType`
- `CharacteristicId`
- `CharacteristicOrder`
- `CharacteristicCategory`
- `CharacteristicLabel`
- `ByVariableId`
- `ByVariableLabel`
- `IsTotal`
- `IsPreferred`

- `SDRID`
- `RegionId`
- `SurveyYearLabel`
- `SurveyType`
- `DenominatorWeighted`
- `DenominatorUnweighted`
- `CILow`
- `CIHigh`
- `LevelRank`

Final Processed Data Columns Example from feature_selected_cleaned_combined_dataset.csv

- `Indicator`
- `Value`
- `SurveyYear`

Columns Removed During Processing

The following columns were removed during the data cleaning and feature selection process, primarily by the `Milestone_2_feature_selection.R` script, which specifically selected for `Indicator`, `Value`, and `SurveyYear` after filtering by `IsPreferred = 1`.

- `ISO3`
- `DataId`
- `Precision`
- `DHS_CountryCode`
- `CountryName`
- `SurveyId`
- `IndicatorId`
- `IndicatorOrder`
- `IndicatorType`
- `CharacteristicId`
- `CharacteristicOrder`

- `CharacteristicCategory`
- `CharacteristicLabel`
- `ByVariableId`
- `ByVariableLabel`
- `IsTotal`
- `IsPreferred`
- `SDRID`
- `RegionId`
- `SurveyYearLabel`
- `SurveyType`
- `DenominatorWeighted`
- `DenominatorUnweighted`
- `CILow`
- `CIHigh`
- `LevelRank`

Rationale for Final Column Selection

The decision to retain only the `Indicator`, `Value`, and `SurveyYear` columns in the final processed dataset (`feature_selected_cleaned_combined_dataset.csv`) was a deliberate choice driven by a comprehensive understanding of the project's core objectives, insights from initial visual data exploration, and the need to optimize the dataset for the primary modelling task. This selection strategy aims to maximize analytical utility while minimizing noise and redundancy.

Key reasons for this focused selection include:

- Elimination of Redundant Country-Specific Identifiers: The raw datasets contained multiple columns (`ISO3`, `DHS_CountryCode`, and `CountryName`) all specifying the country as South Africa. Given that the HDPSA project is exclusively focused on health and demographic patterns within South Africa, these columns provide no unique or discriminatory information for analysis within this specific dataset. Their presence would only add redundant data without contributing to the variance or predictive power for the current scope. Their removal simplifies the dataset significantly without losing any relevant geographical context for a single-country study.

- Focus on Core Measurement and Temporal Trends: The combination of `Indicator`, `Value`, and `SurveyYear` forms the fundamental triplet for any meaningful time-series analysis of health and demographic metrics.
    - The `Indicator` column precisely defines *what specific health or demographic aspect is being measured* (e.g., "child mortality rates," "polio vaccine %").
    - The `Value` column represents *the actual quantitative measurement* for that indicator.
    - The `SurveyYear` column provides the essential *temporal context*, indicating *when* that measurement was taken.
    - These three attributes are indispensable for understanding trends, identifying patterns over time, and analyzing relationships between different indicators across various survey periods.
- Minimizing Noise and Irrelevant Metadata for Predictive Modeling: A significant number of columns in the raw data were identified as metadata or identifiers (`DataId`, `SurveyId`, `IndicatorId`, `CharacteristicId`, `ByVariableId`, `SDRID`, `RegionId`, `IndicatorOrder`, `CharacteristicOrder`, `IndicatorType`, `CharacteristicCategory`, `CharacteristicLabel`, `ByVariableLabel`, `SurveyYearLabel`, `SurveyType`, `IsTotal`, `Precision`, `DenominatorWeighted`, `DenominatorUnweighted`, `CILow`, `CIHigh`, `LevelRank`). While these columns might serve purposes in data management, linking to external databases, or highly specialized analyses, they were deemed less critical for the direct predictive modelling of `Value` against `SurveyYear`. Including them would introduce unnecessary complexity, increase computational overhead, and potentially add noise without providing significant analytical gain for the primary objective. For instance, `IsTotal` and `IsPreferred` were used for filtering rows, but their direct inclusion as features for predicting `Value` was not deemed necessary after the filtering step.
- Efficiency and Model Simplicity (Parsimony): By rigorously focusing on the most pertinent variables, the resulting dataset becomes significantly more streamlined and manageable. This parsimonious approach offers several advantages for modelling:
- Faster Training: Models train more quickly on datasets with fewer features.
- Reduced Risk of Overfitting: Especially with limited data points (like only two `SurveyYear` values), a simpler model with fewer features is less likely to overfit to the training data, leading to better generalization.
- Easier Interpretation: Models built on a concise set of highly relevant features are typically easier to interpret, allowing for clearer understanding of feature importance and relationships.

- Visual Analysis Insights:*Initial visual analysis of the raw and combined datasets played a crucial role in this selection. It was observed that many of the excluded columns either:
    - Exhibited very low variance (e.g., constant values across the dataset).
    - Showed no discernible patterns or correlations when plotted against `Value` or `SurveyYear` in the context of the project's objectives.
    - Were highly correlated with other features that were already being retained (e.g., `SurveyYearLabel` is redundant with `SurveyYear`).

This visual and statistical evidence strongly supported the decision to narrow down the feature set to the most impactful, non-redundant, and directly relevant attributes for the predictive task.

This selective approach ensures that the final dataset is lean, focused, and optimized for the subsequent modelling and data analysis phases, directly supporting the project's objectives of identifying key patterns and predicting child mortality.

Data Preparation Considerations for Modelling

Train-Test Split Strategy with Limited Time Points

Given that the dataset contains only two distinct `SurveyYear` values, a traditional random train-test split for predictive modelling is not recommended, as it risks severe data leakage and unreliable model evaluation. Instead, for any predictive modelling tasks, a time-series validation approach is most appropriate.

Out professional recommendation would be to:

1. Utilize the earlier `SurveyYear` data for training your model.

2. Utilize the later `SurveyYear` data exclusively for testing and evaluating your model's predictive performance.

This approach respects the temporal nature of the data, allowing you to assess how well a model trained on historical patterns can predict outcomes in a subsequent period.

Important Considerations and Limitations:

- Limited Generalizability: With only two time points, the model's ability to generalize to *future* years beyond the test year will be highly constrained. The evaluation results will be specific to the transition from the earlier to the later year.
- Data Scarcity for Training: Training a model on data from only a single year might limit its ability to learn complex patterns or robust relationships.
- Focus on Time-Series Analysis: If the primary objective is to understand historical trends, patterns, and relationships within the data (descriptive time-series analysis) rather than making robust predictions, then using the entire dataset for analysis without a formal train-test split is appropriate. However, if prediction is the goal, some form of temporal validation is essential, even with limited data.

Data Scaling and Normalization

It is important to note that the `Value` column, which represents the core measurement for each indicator, is already expressed in percentages or thousands (as indicated by the nature of the health and demographic data). Therefore, explicit data scaling or normalization (e.g., min-max scaling, standardization) of this column is generally not necessary for most modelling algorithms. The values are already within a comparable range, and further scaling might not provide significant benefits or could even obscure the direct interpretability of the `Value` itself.

This ensures that the data remains interpretable and directly reflects the real-world magnitudes of the indicators.

Potential for Enhanced Applications with More Extensive Datasets

While the current dataset, with its two `SurveyYear` values, allows for foundational time-series analysis and pattern identification, its limitations inherently restrict the scope and robustness of potential applications. Access to more extensive datasets

would significantly enhance the project's capabilities, enabling more sophisticated analyses and predictive models beyond simple time-series approaches.

Key improvements with more extensive datasets would include:

- More Robust Predictive Modelling: With a greater number of `SurveyYear` values, it would be possible to implement more advanced time-series forecasting models (e.g., ARIMA, Prophet) and machine learning models that require more data to learn complex relationships. This would lead to more reliable and generalizable predictions.
- Improved Model Generalizability: A wider range of historical data points would allow models to learn from a broader spectrum of conditions and trends, making them more robust and less prone to overfitting to specific years.
- Enhanced Feature Engineering: More diverse datasets (e.g., including socio-economic factors, policy changes, or environmental data) would open up opportunities for creating richer and more informative features, leading to models with higher predictive power.
- Spatial Analysis and Geographic Granularity: If future datasets included more granular geographic information (e.g., provincial, district, or even sub-district level data), it would enable powerful spatial analysis. This could identify localized patterns, disparities, and the impact of geographic factors on health outcomes, leading to more targeted interventions.
- Exploration of Non-Linear Relationships: With more data, especially across a wider range of variables, it becomes feasible to explore and model complex, non-linear relationships that might be missed with limited data. This could involve using advanced machine learning techniques like neural networks or gradient boosting.
- Causal Inference: While challenging, more comprehensive datasets could provide better opportunities for causal inference, helping to understand not just correlations but also the underlying causes of observed health and demographic patterns.
- Different Types of Machine Learning Models: The current data structure primarily supports time-series analysis. With more features and data points, a wider array of machine learning models (e.g., clustering for identifying similar regions/indicators, classification for predicting risk categories) could be effectively employed.

In summary, while the current data provides a valuable starting point, expanding the dataset's temporal, geographic, and feature dimensions would unlock a much broader

spectrum of analytical and predictive applications, ultimately leading to more impactful insights and informed decision-making for public health initiatives.

Conclusion

These three R scripts collectively form a robust data preparation pipeline for Milestone 2 of the HDPSA project. They ensure that the raw, disparate datasets are combined, thoroughly cleaned, and appropriately feature-selected, providing a high-quality foundation for in-depth data analysis and predictive modelling, in direct alignment with the project's success criteria and data preparation goals.