Suhil Jugroop Makhosandile Nkosi Nkosinathi Remoratile Sibusiso Mogorosi Michael Anyimadu

Uncovering Health Inequities: A BI Exploration of Literacy, ARI Symptoms, and Maternal Mortality

BIN 371 MIlestone 1

Table of Contents

[Background 2](#_Toc205478879)

[Business Objectives 2](#_Toc205478880)

[Success Criteria 2](#_Toc205478881)

[Inventory of Resources 3](#_Toc205478882)

[Human Resources 3](#_Toc205478883)

[Data Resources 3](#_Toc205478884)

[CSV Files Used: 3](#_Toc205478885)

[Source: 3](#_Toc205478886)

[Tools: 3](#_Toc205478887)

[Risks, Assumptions, and Constraints 4](#_Toc205478888)

[Risks 4](#_Toc205478889)

[Assumptions 4](#_Toc205478890)

[Constraints 4](#_Toc205478891)

[Data Mining Goals & Success Criteria 5](#_Toc205478892)

[Goals 5](#_Toc205478893)

[Success Criteria 5](#_Toc205478894)

[Data Description 6](#_Toc205478895)

# Background

South Africa faces significant challenges in maternal and child health, with disparities driven by socioeconomic factors, limited healthcare access, and varying literacy levels. Despite efforts by government and non-profit organizations, preventable health issues persist, particularly in vulnerable communities. This project leverages nationally representative datasets to uncover critical patterns linking literacy, maternal mortality, child malnutrition (anthropometry), and acute respiratory infections (ARI). By focusing exclusively on these four domains: literacy, maternal mortality, ARI symptoms, and anthropometry—we aim to reveal actionable insights for targeted health interventions.

# Business Objectives

1. **Identify Predictors of Health Risks**: Determine key factors contributing to poor maternal and child health outcomes, such as stunting, wasting, and pregnancy-related mortality.
2. **Analyse Literacy-Health Correlations**: Investigate how literacy levels (e.g., women’s education and reading proficiency) correlate with child health indicators (e.g., ARI treatment rates) and maternal survival.
3. **Cluster High-Risk Populations**: Segment communities or demographics using anthropometric data (e.g., BMI, child growth metrics) and ARI prevalence to identify groups needing urgent interventions.
4. **Inform Policy and Interventions**: Translate findings into evidence-based recommendations for policymakers and health organizations to prioritize resource allocation and program design.

# Success Criteria

1. **Actionable Insights**: Deliver clear, statistically supported patterns (e.g., "Low maternal literacy correlates with reduced ARI treatment-seeking behaviour").
2. **Stakeholder Alignment**: Ensure recommendations address priorities of key stakeholders (e.g., South Africa’s Department of Health).
3. **Risk Group Identification**: Successfully classify ≥3 distinct high-risk population segments using clustering models.
4. **Policy Impact**: Propose ≥5 targeted interventions (e.g., literacy-linked health campaigns) validated by data-driven evidence.
5. **Reproducibility**: Provide fully documented code and visualizations to enable stakeholders to replicate analyses for future planning.

# Inventory of Resources

## Human Resources

* **Group Members**: 2–4 students with skills in R, Power BI, and business analysis.
* **Lecturer/Supervisor**: Provides guidance and assessment.
* **Stakeholders**: Simulated representation of NGOs, South African Department of Health, and health data analysts.

## Data Resources

## CSV Files Used:

* + maternal-mortality\_national\_zaf.csv
  + literacy\_national\_zaf.csv
  + anthropometry\_national\_zaf.csv
  + symptoms-of-acute-respiratory-infection-ari\_national\_zaf.csv

## Source:

Publicly shared national health datasets in structured tabular format.

## Tools:

* + **R / RStudio**: For statistical analysis and data cleaning.
  + **Power BI**: For dashboards and visual reporting.

**Technical Resources:**

* Local machines with internet access, R installed, and Power BI Desktop.
* RMarkdown environments for script-based work and documentation.

Risks, Assumptions, and Constraints

## Risks

* **Data Quality Issues**: Missing values (esp. in Anthropometry), inconsistent formats across datasets.
* **Time Constraints**: Limited time for deep model optimization across all CRISP-DM stages.
* **Skill Gaps**: Risk of misunderstanding R syntax or Power BI integration for less-experienced team members.

## Assumptions

* The data provided is representative of national trends.
* Indicators like literacy, ARI, and maternal mortality are linked in meaningful and detectable ways.
* All tools and resources will remain accessible throughout the semester.

## Constraints

* **File Size and Scope**: Limited to only four datasets to stay within project scope.
* **Software Dependency**: Must rely on R and Power BI, even where other tools might offer better fit (e.g., Python or Tableau).
* **Academic Boundaries**: Work must be original, citation-compliant, and presented in line with campus standards.

# Data Mining Goals & Success Criteria

## Goals

1. Discover key factors that predict poor maternal and child health outcomes, focusing on literacy, respiratory symptoms (ARI), and anthropometric indicators.
2. Explore correlations between maternal status, education levels, and child anthropometry.
3. Segment population groups based on health risk levels using clustering.
4. Generate actionable rules using classification and association techniques.

## Success Criteria

* Identify significant predictors of maternal and child health with measurable metrics
* Achieve >= 70% in classification models
* Generate interpretable clusters showing disparities if any
* Present results in Power BI with slicers and visuals to support executives in key decisions

# Data Exploration

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Dataset | Records | Columns | Key Variables | Duplicated Records | Empty Values | NA Values |
| Maternal Mortality | 22 | 29 | Indicator, Value, Precision, Period, Subgroup | 0 | 43 | 168 |
| ARI Symptoms | 27 | 29 | Indicator, Value, Precision, Subgroup, Sex | 0 | 26 | 96 |
| Literacy | 21 | 29 | Indicator, Value, Sex, Age, Subgroup | 0 | 27 | 124 |
| Anthropometry | 38 | 29 | Indicator, Value, Sex, Age, Region | 0 | 6 | 132 |

## Issues found

Columns names were repeated twice in different separation methods; several records were left with NA values or were duplicated.

## R Code used

All four csv’s were read into individual data frames then each data frame was placed in a list for easy looping.

df\_list <- list(df1, df2, df3, df4)

# Function to find duplicates

check\_duplicates <- function(data, show = TRUE) {

duplicate\_rows <- data[duplicated(data), ]

if (show) {

if (nrow(duplicate\_rows) > 0) {

cat("Duplicates found:\n")

print(duplicate\_rows)

} else {

cat("No duplicate rows found.\n")

}

}

return(invisible(duplicate\_rows))}

#Function to check empty values

check\_empty\_values <- function(data) {

na\_counts <- sapply(data, function(x) sum(is.na(x)))

empty\_counts <- sapply(data, function(x) sum(x == "", na.rm = TRUE))

string\_na\_counts <- sapply(data, function(x) sum(x == "NA", na.rm = TRUE))

result <- data.frame(

Column = names(data),

NA\_Count = na\_counts,

Empty\_String\_Count = empty\_counts,

"Text\_'NA'\_Count" = string\_na\_counts

)

result <- result[rowSums(result[,-1]) > 0, ]

if (nrow(result) == 0) {

cat("No missing or empty values found.\n")

} else {

cat("Columns with missing/empty values:\n")

print(result)

}

return(invisible(result))

}

#Dataframe Names

print("Dataframe Column Names")

for(i in 1:length(df\_list)){

cat("Dataframe", i, "\n")

print(names(df\_list[[i]]))

cat("\n")

}

#Duplicates Check

print("Dataframe Duplication Tests")

for(i in 1:length(df\_list)){

cat("Dataframe", i, "\n")

check\_duplicates(df\_list[[i]])

cat("\n")

}

#Empty Values Check

print("Dataframe Empty Valus Check")

for(i in 1:length(df\_list)){

cat("Dataframe", i, "\n")

check\_empty\_values(df\_list[[i]])

cat("\n")}