	Assessment:	Project
	Subject:	Business Intelligence 381
	Total:	100

Health and Demographic Patterns in South Africa (HDPSA): A Data Mining and Visualization Approach

Introduction

Access to quality healthcare, clean water, safe sanitation, and proper nutrition remains a major challenge in many parts of South Africa. These factors are especially important for mothers, children, and young people. Despite efforts made by government and non-profit organisations, many families still face problems such as poor access to clinics, low immunization rates, and unsafe living conditions. This project will focus on understanding the everyday challenges people face in their communities by exploring real-world data collected from different regions in South Africa. The datasets include information about water sources, types of toilets used, how easily people can reach health facilities, child nutrition, and how families have responded to health campaigns like those during the COVID-19 outbreak.

By examining the publicly available datasets, your team will discover patterns and trends that show how certain conditions are linked to others; for example, how distance to water affects child health, or how household living conditions may influence access to medical treatment. Using the findings, you will design easy-to-understand dashboards and summaries using **Power BI** for visuals and **R** for data exploration and analysis.


Outline

To carry out this investigation, your team will follow the Cross-Industry Process for Data Mining (CRISP-DM) methodology. This approach breaks the project into clear steps, starting from understanding the problem, preparing and analysing the data, and ending with presenting the findings. You will use R for analysis and Power BI to create easy-to-understand visuals and dashboards.

You should follow the CRISP-DM stages shown on Figure 1 below to plan and implement the data mining and business intelligence project. CRISP-DM was initially created as an open standard for data mining processes across industries but has since become the most common methodology for data mining, analytics, and data science projects. Therefore, understanding the CRISP-DM framework is very important before beginning this project.

Datasets to Use:

- Access to Health Care
- Child Mortality
- HIV Behavior
- Immunization
- Toilet Facilities
- Water
- COVID-19 Prevention
- Maternal Mortality
- Literacy
- ARI Symptoms
- Anthropometry

	Assessment:	Project
	Subject:	Business Intelligence 381
	Total:	100

- IYCF (Infant & Young Child Feeding)

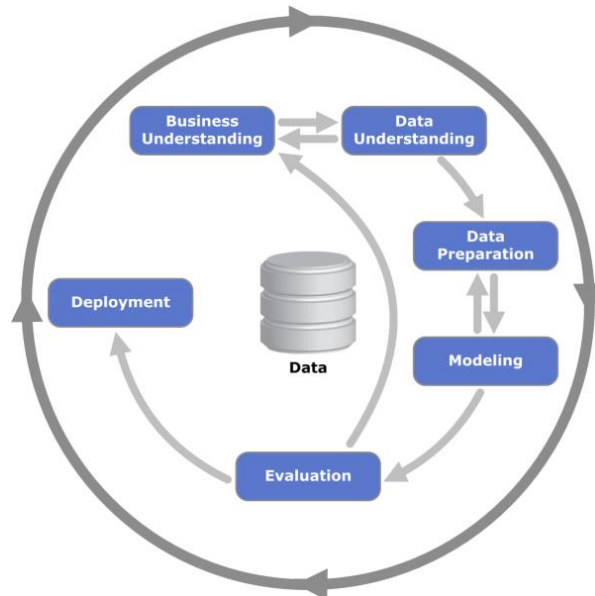


Figure 1: CRISP-DM

Project Steps

The project outline underscores the use of CRISP-DM as a structured approach to guide you through the data science project, emphasizing its importance in the context of real-world data analysis and modelling. You may need to gather extra information on related projects and scenarios to establish a strong business case for the project.

The entire project consists of the following steps which are then broken into milestones that should be submitted on given due dates.

1. Business Understanding:


- Clearly define the business problem and objectives.
- Identify stakeholders and their requirements.
- Determine the success criteria for the HDPSA project.
- Recognise the importance of CRISP-DM as the methodology to guide the project.

2. Data Understanding:

- Explore the provided dataset, including variables like mortality rates, distance to health facilities, facility types, immunization rates, water sources, vaccination rates, etc.
- Document data quality problems, missing values, duplicates, and outliers.
- Perform preliminary data visualizations and analysis to gain initial insights.
- Create preliminary dashboards to visualize data distributions, correlations, and initial insights to identify patterns and relationships in the data before moving on to the modelling phase.

3. Data Preparation:


- Clean and preprocess the data to address quality issues (e.g., impute missing values, remove duplicates, handle outliers).
- Transform and encode categorical variables as needed.

	Assessment:	Project
	Subject:	Business Intelligence 381
	Total:	100

- Discretize or scale numerical variables if necessary.
 - Split the data into training and testing sets for model evaluation.
- 4. Modelling:**
- Select suitable algorithm(s) (e.g., logistic regression, decision tree, random forest) for classification, clustering, association rules or time-series analysis.
 - Explore and identify significant variables to be used based on your analysis.
 - Justify your choice of predictor variables by employing methods to evaluate the interestingness, importance, and relevance of these variables.
 - Train and fine-tune the data mining model using the training data.
 - Visually represent the model's performance metrics, confusion matrices, and other evaluation results to facilitate interpretation.
 - Document the model selection process and parameter tuning, including the rationale behind predictor variable selection.
- 5. Model Evaluation:**
- Evaluate the classification model's performance using appropriate metrics (e.g., accuracy, precision, recall, F1-score, Silhouette Scores, Dunn Index).
 - Compare the models' performance against the baseline.
 - Visualize and interpret the results.
- 6. Deployment:**
- Discuss deployment options for the HDPSA project.
 - Develop and document a plan for deploying the model in a real-world scenario.
 - Address considerations such as data input, monitoring, and updating the model.
 - Consider ethical and privacy implications of deploying the model, as guided by CRISP-DM.
- 7. Final Report:**
- Summarise the entire project, including the business problem, data exploration, data preparation, modelling, evaluation, and deployment plan highlighting the use of CRISP-DM as the methodology guiding the project.
 - Provide insights and recommendations based on the model's findings.
 - Include visualizations, code snippets, and explanations throughout the report.
 - Submit the project report along with any code and visualizations used for analysis and modelling.


Deliverables:

- Project Report (in PDF or document format) detailing all project steps.
- Code (R and R Markdown) used for data analysis and modelling.
- Power BI project file with data models, visualizations and interactive dashboards.
- Short Presentation summarising the key findings and recommendations.

	Assessment:	Project
	Subject:	Business Intelligence 381
	Total:	100

Detailed Breakdown (Based on CRISP-DM)

Milestone	Phase	Suggestions	Tools	Deliverables
1	Business Understanding	Define the main goals: e.g., identify health risk zones, analyze healthcare access, and link hygiene to disease. Define key questions like: "What predicts child mortality?" or "Is HIV behavior linked to literacy?"	R Markdown / Word	Problem definition document List of project questions and KPIs Stakeholder objectives summary
1	Data Understanding	Explore all 12 datasets in R. Check dimensions, structures, distributions, missing values, outliers. Create data dictionaries.	R	R Markdown script for EDA Summary tables of key variables Visuals (bar charts, histograms, boxplots) Data dictionary (Excel/PDF)
2	Data Preparation	Clean and merge datasets where relevant (e.g., literacy + mortality, water + ARI). Handle missing values, create new variables (e.g., "Risk_Level"), normalize data if needed.	R	Cleaned and merged datasets in R Codebook of transformations RDS or CSV files of final datasets
3	Modeling (Clustering)	Use K-Means and Hierarchical Clustering to group provinces or communities by health risks (e.g., toilet + water + mortality). Determine optimal clusters (elbow method).	R	Clustering script (with visuals) Cluster assignment table Interpretation of clusters
3	Modeling (Classification & Association Rules)	Use Decision Trees / Random Forests to predict child mortality. Apply Apriori to find rules (e.g., "If low literacy and no toilet → High ARI").	R	Classification model performance report Confusion matrix / accuracy Association rules (lift/support/confidence)
4	Evaluation	Interpret results in context. Discuss accuracy, patterns, insights. Reflect on the significance and limitations. Compare regions or population segments.	R & Word	Evaluation report Insights summary Recommendations based on findings
5	Deployment (Dashboard & Reporting)	Create Power BI dashboards and Shiny App showing key trends: access to care, mortality clusters, regional risks, hygiene coverage. Add slicers for province, age group, gender.	Power BI, Shiny App	Interactive Power BI dashboard and Shiny Apps
6	Final Report	Final Project Report and Presentation		Final project report (PDF) Group presentation (PPT)

	Assessment:	Project
	Subject:	Business Intelligence 381
	Total:	100

Grading Criteria:

Criteria	Weight
Understanding of the business problem	10%
Data exploration and quality assessment	10%
Data preprocessing and transformation	15%
Model selection and training	15%
Model evaluation and comparison	10%
Dashboards and Visualizations	20%
Deployment plan and considerations	10%
Documentation and presentation	10%

Additional Information

- This is a group project, but the group may not exceed four people.
- All work must be original. Copying another group's work will not be tolerated.
- Includes names of all group members on the Cover Page.
- Submit your project milestones electronically on Moodle (BC Connect) before the due dates.
- All writing must be correctly cited and referenced.
- **Plagiarism is a serious offence.** Belgium Campus uses software that can scan for plagiarism and a student caught doing this will get 0 for this assignment.
- No mark will be awarded if the assignment is not uploaded via BC Connect.
- Late assignments will not be accepted; missing the deadline is an automatic 0.