belgium campus itversity

Project Milestone 5

**Module:** Business Intelligence 381  
**Methodology**: CRISP-DM  
**Project:** Health and Demographic Patterns in South Africa (HDPSA): A Data Mining and Visualization Approach  
**Milestone:** 5 Deployment phase

Petrus Human 577842

Frederik Knoetze 600965

Teleki Shai 601377

Moloko Rakumako 601352

Table of Contents

[Executive Summary 3](#_Toc211251795)

[1. Introduction 4](#_Toc211251796)

[1.1 Project Context and Background 4](#_Toc211251797)

[1.2 CRISP-DM Deployment Phase Overview 4](#_Toc211251798)

[1.3 Milestone 5 Objectives 4](#_Toc211251799)

[2. Assessment of Results / Deployment Strategy 6](#_Toc211251800)

[2.1 Purpose and Rationale 6](#_Toc211251801)

[2.2 Approved Models Summary 6](#_Toc211251802)

[2.3 Deployable Outputs Inventory 7](#_Toc211251803)

[2.4 Deployment Options Analysis 8](#_Toc211251804)

[2.4.1 Option A: R Shiny Web Application 8](#_Toc211251805)

[2.4.2 Option B: Power BI Dashboard 8](#_Toc211251806)

[2.5 Comparative Analysis Table 9](#_Toc211251807)

[2.6 Recommended Deployment Strategy 9](#_Toc211251808)

[Primary Deployment: Power BI Dashboard 9](#_Toc211251809)

[Secondary Deployment: R Shiny Demonstration 9](#_Toc211251810)

[2.7 Deployment Pipeline Architecture 10](#_Toc211251811)

[2.8 Implementation Steps and Timeline 10](#_Toc211251812)

[3. Approved Models / Tool Evaluation & Recommendation 11](#_Toc211251813)

[3.1 Deployment Tools Research 11](#_Toc211251814)

[3.2 Evaluation Criteria and Framework 11](#_Toc211251815)

[3.3 Tool Comparison Matrix 11](#_Toc211251816)

[3.4 Final Tool Selection and Justification 11](#_Toc211251817)

[3.5 Prototype Implementation Evidence 11](#_Toc211251818)

[4. Process Review / Monitoring & Maintenance Plan 11](#_Toc211251819)

[4.1 Monitoring Framework Overview 11](#_Toc211251820)

[4.2 Performance Metrics and Thresholds 11](#_Toc211251821)

[4.3 Maintenance Schedule and Procedures 11](#_Toc211251822)

[4.4 Automation Strategy 11](#_Toc211251823)

[4.5 Governance and Version Control 11](#_Toc211251824)

[4.6 Alert and Escalation Protocols 11](#_Toc211251825)

[5. Next Steps / Model Deployment & Documentation 12](#_Toc211251826)

[5.1 Deployment Implementation Plan 12](#_Toc211251827)

[5.2 Power BI Dashboard Design 12](#_Toc211251828)

[5.3 Alternative Interface: R Shiny Demo 12](#_Toc211251829)

[5.4 Ethical Considerations and Data Privacy 12](#_Toc211251830)

[5.5 User Guide and Documentation 12](#_Toc211251831)

[6. Integration and Final Deployment Strategy 12](#_Toc211251832)

[6.1 Unified Deployment Roadmap 12](#_Toc211251833)

[6.2 Stakeholder Communication Plan 12](#_Toc211251834)

[6.3 Training and Change Management 12](#_Toc211251835)

[6.4 Success Metrics and KPIs 12](#_Toc211251836)

[7. Ethical and Privacy Implications 12](#_Toc211251837)

[7.1 Data Privacy Compliance 12](#_Toc211251838)

[7.2 Model Fairness and Bias Mitigation 12](#_Toc211251839)

[7.3 Transparency and Accountability 12](#_Toc211251840)

[8. Conclusion and Recommendations 13](#_Toc211251841)

[8.1 Key Findings Summary 13](#_Toc211251842)

[8.2 Deployment Readiness Assessment 13](#_Toc211251843)

[8.3 Future Enhancements 13](#_Toc211251844)

[9. References 13](#_Toc211251845)

[10. Appendices 13](#_Toc211251846)

[Appendix A – Deployment Export Code 13](#_Toc211251847)

[Appendix B – Deployment Flow Diagram 13](#_Toc211251848)

[Appendix C – Model Performance Summary 13](#_Toc211251849)

[Appendix D – Stakeholder Sign-off Documentation 13](#_Toc211251850)

# Executive Summary

This document presents the final phase of the Health and Demographic Patterns in South Africa (HDPSA) project, conducted under the BIN381 Business Intelligence module.

Building on the analytical groundwork of Milestones 1 through 4, Milestone 5 executes the CRISP-DM Phase 6 (Deployment): converting evaluated machine-learning models into accessible, policy-ready tools.

The overarching objective is to deliver a sustainable deployment ecosystem that empowers public-health stakeholders—national and provincial departments, NGOs, and research partners—to visualise and interpret health-risk predictions through user-friendly interfaces.

Although Milestone 4 revealed moderate predictive accuracy (≈ 50–55 %), this phase focuses on operationalising the workflow, ensuring reproducibility, data transparency, and stakeholder engagement.

**Key outputs include:**

* A Power BI Dashboard providing KPI cards, feature-importance charts, and filterable health-indicator insights.
* An R Shiny web application demonstrating real-time interaction for technical reviewers.
* Automated export, monitoring, and governance scripts ensuring traceability across milestones.

Each group member addresses a specific rubric component:

1. Assessment of Results / Deployment Strategy
2. Approved Model(s) / Tool Evaluation & Recommendation
3. Process Review / Monitoring & Maintenance Plan
4. Next Steps / Model Deployment & Documentation

Together, these deliverables ensure the HDPSA analytical framework transitions from academic modelling to a functional, decision-support system, setting a foundation for future data expansion and predictive-health policy integration.

# 1. Introduction

## 1.1 Project Context and Background

The HDPSA – Health and Demographic Patterns in South Africa project investigates national health indicators spanning access to care, water and sanitation, maternal mortality, immunisation, and education.

Using multiple DHS-style datasets (1998 – 2016), the project applies supervised-learning techniques to identify socio-economic determinants of health outcomes.

Earlier milestones established:

* Milestone 1 – Business Understanding: problem framing and success metrics.
* Milestone 2 – Data Preparation: cleaning, integration, and feature selection.
* Milestone 3 – Modelling: Logistic Regression, Decision Tree, Random Forest, Naïve Bayes.
* Milestone 4 – Evaluation: comparative assessment using Accuracy, F1, ROC-AUC.
* Milestone 5 finalises this pipeline by deploying validated results into accessible analytic platforms that enable evidence-based decision-making.

## 1.2 CRISP-DM Deployment Phase Overview

The **Deployment Phase (Phase 6)** operationalises analytical insights.  
According to Wirth & Hipp (2000), deployment encompasses:

1. **Implementation:** integrating model artefacts into business systems.
2. **Documentation:** preparing reproducible code, metadata, and user manuals.
3. **Monitoring:** defining mechanisms to track model performance over time.
4. **Maintenance:** ensuring continuous alignment between evolving data and deployed models.

This phase converts technical outputs into business-value tools such as dashboards, reports, or applications.

## 1.3 Milestone 5 Objectives

1. Establish a **deployment infrastructure** linking R-based models to visual-analytics environments (Power BI and R Shiny).
2. Ensure **governance and traceability** of exported metrics and model artefacts.
3. Develop a **monitoring framework** for post-deployment accuracy and data-refresh cycles.
4. Produce comprehensive **documentation and user guides** supporting future teams and stakeholders.

Deliverables collectively demonstrate a full CRISP-DM lifecycle—culminating in transparent, reproducible model deployment.

# 2. Assessment of Results / Deployment Strategy

## 2.1 Purpose and Rationale

The deployment phase represents the conclusion of the CRISP-DM methodology, transforming analytical models into actionable, policy-ready tools.  
While Milestone 4 revealed that the binary classification models (Random Forest, Logistic Regression, Decision Tree, and Naïve Bayes) achieved moderate accuracy (≈ 50–55 %), this phase has a dual purpose:

**1. Methodological Demonstration** – to establish a complete, repeatable deployment pipeline supporting future iterations when additional survey years are available.  
**2. Stakeholder Engagement** – to create interactive tools that allow policymakers to explore model results, provide feedback, and foster data-driven decision-making.

This deployment plan acknowledges current limitations while building toward a scalable solution aligned with South Africa’s public-health priorities (NDoH, 2024; Statistics South Africa, 2024).

## 2.2 Approved Models Summary

Following the Milestone 4 evaluation, two models were approved for deployment based on interpretability, transparency and policy relevance.

**Logistic Regression (Approved)**

* **Performance:** Accuracy = 52.27 %, AUC = 0.532
* **Strengths:**  
  – Interpretable coefficients (odds ratios)  
  – Transparent decision logic  
  – Fast computation, reproducible results
* **Role:** Baseline model for stakeholder training and coefficient interpretation.
* **Status:** Approved for demonstration with performance caveats.

**Decision Tree (Approved)**

* **Performance:** Accuracy = 50.00 %, AUC = 0.531
* **Strengths:**  
  – Clear “if-then” visual rules  
  – Ideal for policy manuals and decision transparency
* **Role:** Visual decision-support tool.
* **Status:** Approved for demonstration with disclaimers.

**Models Not Approved:**  
– *Random Forest* – Best accuracy (54.55 %) but limited interpretability.  
– *Naïve Bayes* – Weak performance (AUC = 0.465), violated independence assumptions.

## 2.3 Deployable Outputs Inventory

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Category | Asset Name | Format | Location | Description |
| Data Assets | HDPSA\_clean.csv | CSV | /Cleaned Datasets/ | Final cleaned and feature-selected dataset (~ 850 rows) |
|  | feature\_selected\_cleaned\_combined\_dataset.csv | CSV | /Cleaned Datasets/ | Model-ready dataset (Indicator, Value, Survey Year) |
| Model Objects | model\_logit.rds | RDS | /Model Outputs/ | Trained Logistic Regression model |
|  | model\_tree.rds | RDS | /Model Outputs/ | Trained Decision Tree model |
|  | predictions\_logit.csv | CSV | /Model Outputs/ | Logistic Regression test predictions |
|  | predictions\_tree.csv | CSV | /Model Outputs/ | Decision Tree test predictions |
| Visualisation Assets | model\_performance\_summary.csv | CSV | /Milestone 4 Outputs/Assessment/ | Model metrics (Accuracy, Precision, Recall, F1, AUC) |
|  | feature\_importance\_rf.csv | CSV | /Milestone 4 Outputs/Assessment/ | Variable importance ranking |
|  | roc\_auc\_comparison.png | PNG | /Milestone 4 Outputs/Assessment/ | ROC-curve comparison for all models |
|  | model\_performance\_comparison.png | PNG | /Milestone 4 Outputs/Assessment/ | Bar chart of Accuracy, F1, AUC |
| Documentation | model\_governance\_log.xlsx | XLSX | — | Versioning log and parameter tracking |
|  | approved\_model\_summary.csv | CSV | — | Go/No-Go decisions summary |
|  | Milestone\_4\_Report.pdf | PDF | — | Complete evaluation report |

## 2.4 Deployment Options Analysis

## 2.4.1 Option A: R Shiny Web Application

A fully interactive R Shiny web app allowing real-time model interaction.

**Framework:** R Shiny v1.7 + Bootstrap UI  
**Environment:** Local R session or shinyapps.io  
**Backend:** R 4.3 + rpart, randomForest, ggplot2

**Features:**  
– Interactive input sliders for education, income, and water access  
– Instant visual predictions and risk scoring  
– Model-comparison toggle (Logistic vs Tree)  
– Downloadable PDF report

**Advantages:** Full interactivity, reproducibility, open-source transparency.  
**Limitations:** Requires R expertise; limited concurrent users.

**Ideal Use:** Technical demonstration for academic and data-science audiences.

## 2.4.2 Option B: Power BI Dashboard

A business-intelligence dashboard for policymakers, importing model outputs via CSV.

**Platform:** Microsoft Power BI Desktop → Power BI Service  
**Data Source:** model\_metrics\_export.csv, feature\_importance\_rf.csv  
**Access:** Role-based control through Microsoft 365

**Features:**  
– KPI cards (Accuracy, AUC)  
– Bar charts for top predictors  
– Provincial risk maps (if regional data available)  
– Slicers for Survey Year and Indicator

**Advantages:** Enterprise-grade security, government familiarity, mobile access.  
**Limitations:** Requires Pro license (~ R140/user/month); no native R execution.

**Ideal Use:** Policy dashboards for NDoH and provincial departments.

## 2.5 Comparative Analysis Table

|  |  |  |  |
| --- | --- | --- | --- |
| Criterion | R Shiny Web App | Power BI Dashboard | Importance |
| Cost | ★★★★★ (Free) | ★★☆☆☆ (Paid license) | High |
| Accessibility | ★★☆☆☆ (Technical users) | ★★★★★ (Policy users) | Critical |
| Skills Required | ★★☆☆☆ (R programming) | ★★★★☆ (BI skills) | High |
| Security | ★★★☆☆ (Self-managed) | ★★★★★ (Enterprise) | Critical |
| Interactivity | ★★★★★ (Real-time) | ★★★☆☆ (Filtered) | Medium |
| Scalability | ★★☆☆☆ (Limited) | ★★★★★ (Cloud) | High |
| Integration | ★★☆☆☆ (Standalone) | ★★★★★ (Microsoft ecosystem) | High |
| Maintenance | ★★☆☆☆ (R required) | ★★★★☆ (Managed) | Medium |
| Mobile Access | ★★★☆☆ | ★★★★★ | Medium |
| Publication Ready | ★★★☆☆ | ★★★★★ | Medium |

## 2.6 Recommended Deployment Strategy

### Primary Deployment: Power BI Dashboard

**Audience:** National and Provincial Departments of Health, NGOs  
**Purpose:** Policy-ready KPI tracking and reporting  
**Timeline:** Immediate deployment  
**Justification:** Government alignment, security, scalability, professional presentation.

### Secondary Deployment: R Shiny Demonstration

**Audience:** Academic and technical teams  
**Purpose:** Methodology demonstration and scenario testing  
**Timeline:** Parallel release for technical review  
**Justification:** Enhances transparency and supports future iterations.

**Dual-Platform Rationale:** Combines policy reach (Power BI) with technical depth (Shiny), maintaining consistency through shared exports and metadata.

## 2.7 Deployment Pipeline Architecture

**Workflow:**

1. **Data Preparation (Milestone 2):** Raw → cleaned → HDPSA\_clean.csv
2. **Modelling (Milestone 3):** Train/test → model\_logit.rds, model\_tree.rds
3. **Evaluation (Milestone 4):** Metrics → model\_performance\_summary.csv
4. **Deployment Export (Milestone 5):** Run deployment\_export.R → exports to CSV/Excel
5. **Dashboard Development:** Power BI and Shiny apps built from exports
6. **Monitoring (Next Phase):** Scheduled refresh, accuracy validation.

## 2.8 Implementation Steps and Timeline

|  |  |  |
| --- | --- | --- |
| Phase | Activities | Deliverables |
| Week 1 – Preparation | Run deployment\_export.R; verify exports; design dashboard layout (NDoH branding). | CSV and Excel exports; initial .pbix file |
| Week 2 – Testing & Validation | Internal testing of Power BI filters and Shiny UI; collect feedback. | QA log; user feedback notes |
| Week 3 – Publication & Handover | Publish dashboard to Power BI Service; deploy Shiny app to shinyapps.io; prepare user guides. | Live Power BI dashboard; Shiny URL; User Manual PDF |

# 3. Approved Models / Tool Evaluation & Recommendation

## 3.1 Deployment Tools Research

## 3.2 Evaluation Criteria and Framework

## 3.3 Tool Comparison Matrix

## 3.4 Final Tool Selection and Justification

## 3.5 Prototype Implementation Evidence

# 4. Process Review / Monitoring & Maintenance Plan

This section outlines the monitoring and maintenance plan for the Health and Demographic Patterns in South Africa (HDPSA) project. The goal is to ensure the deployed machine learning models, which provide interpretable insights for public-health stakeholders, remain accurate, reliable, and relevant over time. This plan addresses data inputs, post-deployment monitoring, and model updating thresholds.

## 4.1 Data input

For a model to be accurate and for it to be useful to the current situation new data needs to be added and new models need to be trained. New data should be taken from sources like the national census, and any other applicable data from Stats SA, WHO, UNICEF, IGME and more.

## 4.2 Performance Metrics and Thresholds

A monitoring framework with defined metrics, thresholds, and owners will be used to track model performance and data relevancy.

The baseline performance for the approved models is an Accuracy of ~52% and an AUC of ~0.53. The thresholds below are set to detect a significant degradation from this established baseline, rather than an aspirational target.

|  |  |  |  |
| --- | --- | --- | --- |
| Metric | Threshold | Action | Owner |
| **AUC** | < 0.50 | Retrain model with new or existing data | Data Scientist |
| **Accuracy** | < 45% | Investigate potential data drift or bias | Project Lead |
| **Data Freshness** | `> 24 months | Acquire and integrate new survey data | Data Engineer |

## 4.3 Maintenance Schedule and Procedures

The monitoring schedule and procedures will consist of three parts

* **Monthly Review:** The `monitoring\_log.csv` will be reviewed monthly by the Project Lead to check for any anomalies or threshold breaches.
* **Quarterly Re-validation:** The model will be re-validated quarterly against a hold-out test set from the original data to check for performance consistency.
* **Retraining Trigger:** A model retraining process will be initiated if any threshold in the monitoring framework is breached.

## 4.4 Automation Strategy

An R script (`model\_monitoring\_log.R`) will be used to automate the logging of performance metrics. This script will be run after each new data validation or prediction task. It appends a new entry to a central CSV log file, recording the model's performance over time. An example of how this can look can be found as (‘Milestone\_5\_Example\_automated\_logging.R’).

An example output will be the following:

timestamp,auc\_score,accuracy\_score,data\_freshness

2025-10-15 14:30:00,0.532,0.522,2016-12-31

2026-01-15 15:00:00,0.529,0.519,2016-12-31

2026-04-15 14:45:00,0.525,0.511,2016-12-31

This will provide a quick overview of when the log is. Quick performance metrics about the model and information on the data.

## 4.5 Governance and Version Control

Version control through the duration of the project has been handled through git. This allows for easy history of changes and the option to roll changes back. Throughout the duration of the project most commits or merges have been approved by the programmer responsible for the new section. Once the project enters its operational phase this should change so that only the most senior member may approve changes. This will be to ensure that the customer version is always in a functional state and as to not disrupt their experience.

## Governance will be documented though a spreadsheet. (`model\_governance\_log.xlsx`) will be maintained to track major decisions, model versions deployed, and the rationale for any changes, providing a human-readable history of the project.

## 4.6 Model Obsolescence

A model will be considered for retirement or a complete rebuild if:

* **Consistent Underperformance:** It fails to meet the minimum thresholds for two consecutive quarters.
* **Shift in Business Objectives:** Stakeholder needs evolve from demonstration to requiring high-accuracy predictions, which the current models cannot provide.
* **Fundamental Data Changes:** New data sources are introduced that are structurally different from the original DHS data.

# 5. Next Steps / Model Deployment & Documentation

## 5.1 Deployment Implementation Plan

## 5.2 Power BI Dashboard Design

## 5.3 Alternative Interface: R Shiny Demo

## 5.4 Ethical Considerations and Data Privacy

## 5.5 User Guide and Documentation

# 6. Integration and Final Deployment Strategy

## 6.1 Unified Deployment Roadmap

## 6.2 Stakeholder Communication Plan

## 6.3 Training and Change Management

## 6.4 Success Metrics and KPIs

# 7. Ethical and Privacy Implications

## 7.1 Data Privacy Compliance

## 7.2 Model Fairness and Bias Mitigation

## 7.3 Transparency and Accountability

# 8. Conclusion and Recommendations

## 8.1 Key Findings Summary

## 8.2 Deployment Readiness Assessment

## 8.3 Future Enhancements

# 9. References

# 10. Appendices

## Appendix A – Deployment Export Code

## Appendix B – Deployment Flow Diagram

## Appendix C – Model Performance Summary

## Appendix D – Stakeholder Sign-off Documentation