belgium campus itversity

Project Milestone 4

**Module:** Business Intelligence 381  
**Methodology**: CRISP-DM  
**Project:** Health and Demographic Patterns in South Africa (HDPSA): A Data Mining and Visualization Approach  
**Milestone:** 4

Petrus Human 577842

Frederik Knoetze 600965

Teleki Shai 601377

Moloko Rakumako 601352

Table of Contents

[1. Introduction 3](#_Toc210652449)

[1.1 Project Context 3](#_Toc210652450)

[1.2 CRISP-DM Framework Overview 3](#_Toc210652451)

[1.3 Milestone 4 Objectives 3](#_Toc210652452)

[2. Assessment of Results 5](#_Toc210652453)

[2.1 Purpose and Scope 5](#_Toc210652454)

[2.2 Business Goals Recap 5](#_Toc210652455)

[2.3 Model Performance Evaluation 5](#_Toc210652456)

[2.4 Metrics Analysis 6](#_Toc210652457)

[2.5 Comparison with Success Criteria 7](#_Toc210652458)

[2.6 Practical Significance of Predictors 8](#_Toc210652459)

[2.7 Strengths and Weaknesses Summary 8](#_Toc210652460)

[3. Approved Model(s) and Justification 10](#_Toc210652461)

[3.1 Selection Criteria 10](#_Toc210652462)

[3.2 Model-by-Model Assessment 10](#_Toc210652463)

[3.3 Approved Model: Random Forest 10](#_Toc210652464)

[3.4 Feature Importance Analysis 10](#_Toc210652465)

[3.5 Model Governance and Documentation 10](#_Toc210652466)

[3.6 Justification Summary 10](#_Toc210652467)

[4. Process Review 11](#_Toc210652468)

[4.1 CRISP-DM Phase-by-Phase Review 11](#_Toc210652469)

[4.2 Quality Assurance Findings 11](#_Toc210652470)

[4.3 Data Quality Assessment 11](#_Toc210652471)

[4.4 Ethical and Bias Considerations 11](#_Toc210652472)

[4.5 Process Improvement Recommendations 11](#_Toc210652473)

[5. Next Steps and Conclusion 12](#_Toc210652474)

[5.1 Decision Matrix for Future Actions 12](#_Toc210652475)

[5.2 Recommended Course of Action 12](#_Toc210652476)

[5.3 Deployment Preview (CRISP-DM Phase 6) 12](#_Toc210652477)

[5.4 Final Conclusion 12](#_Toc210652478)

[6. References 13](#_Toc210652479)

[7. Appendices 14](#_Toc210652480)

[Appendix A – Complete R Code Listings 14](#_Toc210652481)

[Appendix B – Additional Visualizations 14](#_Toc210652482)

[Appendix C – Model Parameters and Settings 14](#_Toc210652483)

1. Introduction

## 1.1 Project Context

This report presents the evaluation phase (Milestone 4) of the Health and Demographic Patterns in South Africa (HDPSA) data mining project. The HDPSA project investigates patterns and trends across twelve health and demographic datasets to support evidence-based policy formulation in South Africa. Previous milestones established the business understanding, completed data preparation and developed predictive models using multiple machine learning techniques.

The project addresses critical public health challenges including access to healthcare, immunization coverage, water and sanitation infrastructure, child mortality, and maternal health outcomes. By applying data mining methodologies to national survey data from 1998 and 2016, this analysis aims to identify key predictors of health outcomes and provide actionable insights for policy makers and health administrators.

## 1.2 CRISP-DM Framework Overview

The Cross-Industry Standard Process for Data Mining (CRISP-DM) provides the methodological foundation for this project. CRISP-DM consists of six iterative phases:

1. **Business Understanding** – Define objectives and requirements from a business perspective
2. **Data Understanding** – Collect and explore data to identify quality issues and discover insights
3. **Data Preparation** – Construct the final dataset through cleaning, transformation, and feature selection
4. **Modeling** – Select and apply various modeling techniques with calibrated parameters
5. **Evaluation** – Assess models against business objectives and validate the process quality *(Current Phase)*
6. **Deployment** – Plan the implementation of models in operational environments

As noted by Wirth and Hipp (2000), CRISP-DM emphasizes the importance of evaluating not just model performance metrics, but also the alignment with business goals and the robustness of the overall process. This evaluation phase critically examines whether the models developed in Milestone 3 meet the success criteria established in Milestone 1.

## 1.3 Milestone 4 Objectives

This milestone addresses three core evaluation tasks aligned with CRISP-DM Phase 5:

### Task 1: Evaluate Results

* Assess model outputs (Logistic Regression, Decision Tree, Random Forest, Naïve Bayes) against business success criteria (≥70% accuracy or AUC >0.75)
* Interpret findings in the context of predicting health risk, access to care, and demographic disparities
* Compare performance metrics to determine which models meet requirements

### Task 2: Review Process

* Audit the entire CRISP-DM workflow from Phases 1-4
* Identify gaps, quality issues, and areas requiring iteration
* Document process strengths and improvement opportunities

### Task 3: Determine Next Steps

* Evaluate options for deployment, iteration, or project restart
* Recommend the most realistic course of action based on current capabilities
* Outline transition pathway toward CRISP-DM Phase 6 (Deployment)

The following sections present detailed findings for each task, supported by quantitative analysis, visualizations and critical interpretation of results in the context of South African public health priorities.

2. Assessment of Results

## 2.1 Purpose and Scope

This section evaluates the four predictive models developed in Milestone 3—**Logistic Regression, Decision Tree, Random Forest, and Naïve Bayes**—against the business success criteria established in Milestone 1. The evaluation examines both technical performance metrics and practical relevance to the project's health prediction objectives.

**Evaluation Framework:**

* **Technical Assessment**: Accuracy, Precision, Recall, F1-Score, ROC-AUC
* **Business Alignment**: Ability to support policy decisions on health risk prediction
* **Interpretability**: Transparency of model predictions for stakeholder communication
* **Robustness**: Reliability across different data conditions

## 2.2 Business Goals Recap

From Milestone 1, the primary business objectives were:

**Primary Goal**: Predict public health risk indicators to identify vulnerable populations and inform resource allocation decisions.

**Secondary Goals**:

1. Identify key demographic and socioeconomic predictors of health outcomes
2. Distinguish temporal health trends between 1998 and 2016 survey periods
3. Provide interpretable insights for policy formulation

**Success Criteria** (established in Milestone 1):

* Model accuracy ≥ 70%
* ROC-AUC > 0.75
* Balanced precision and recall for policy-relevant predictions
* Clear identification of top 3-5 feature importance predictors

## 2.3 Model Performance Evaluation

## 2.3.1 R Code Implementation

The following R script (evaluation\_metrics.R) was developed to compute comprehensive performance metrics from the model predictions generated in Milestone 3: Description: Computes classification metrics for all four models.

## 2.3.2 Actual Performance Results

Based on the model predictions generated in Milestone 3 and evaluated using the code above, the following performance metrics were obtained:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score | ROC-AUC | Meets Accuracy Goal | Meets AUC Goal |
| Random Forest | 0.5455 | 0.5400 | 0.6140 | 0.5741 | 0.5080 | **No** | **No** |
| Logistic Regression | 0.5227 | 0.5116 | 0.9773 | 0.6717 | 0.5323 | **No** | **No** |
| Decision Tree | 0.5000 | 0.5000 | 0.5909 | 0.5417 | 0.5312 | **No** | **No** |
| Naïve Bayes | 0.5000 | 0.5000 | 1.0000 | 0.6667 | 0.4650 | **No** | **No** |

**Note:** These results reflect the actual model performance from Milestone 3, where only two survey years (1998 and 2016) were available for classification. The modest performance (~50-55% accuracy) indicates the inherent difficulty of the binary temporal classification task with limited temporal data points.

## 2.4 Metrics Analysis

## 2.4.1 Accuracy Assessment

**Random Forest (54.55%)**:

* Highest accuracy among all models, but still below the 70% business threshold
* Performance only marginally better than random guessing (50%)
* Indicates the challenge of distinguishing health indicators between two time points

**Logistic Regression (52.27%)**:

* Near-random performance for binary classification
* High recall (97.7%) but poor precision, suggesting model predicts positive class for most instances
* Confusion matrix reveals 43 true positives but 41 false positives

**Decision Tree (50.00%)**:

* Exactly random performance
* Balanced confusion matrix (18 TP, 18 TN, 26 FP, 18 FN)
* Suggests no meaningful patterns learned from temporal data

**Naïve Bayes (50.00%)**:

* Predicts all instances as positive class (perfect recall, zero specificity)
* No discriminative power whatsoever
* Independence assumption clearly violated in health data

## 2.4.2 ROC-AUC Analysis

All models achieved ROC-AUC values between 0.465-0.532, which are:

* **Below the 0.75 business threshold**
* Barely distinguishable from random classification (AUC = 0.50)
* Indicating limited ability to rank positive instances above negative instances

The ROC curves (see Figure 1) show minimal separation from the diagonal reference line, confirming weak discriminative ability across all models.

## 2.4.3 Precision-Recall Trade-off

**Logistic Regression** and **Naïve Bayes** exhibit severe class imbalance in predictions:

* Extremely high recall (97.7% and 100% respectively)
* Low precision (~51% and 50% respectively)
* Both models default to predicting the positive class for most test instances

This pattern suggests:

1. Models struggle to identify distinguishing features between classes
2. Default prediction strategies dominate over learned patterns
3. Insufficient temporal variation in the data for robust learning

## 2.5 Comparison with Success Criteria

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Business Criterion | Target | Best Model | Result | Status |
| Classification Accuracy | ≥ 70% | Random Forest: 54.55% | 15.45 percentage points below target | ❌ **Not Met** |
| ROC-AUC Discrimination | > 0.75 | Logistic Regression: 0.532 | 0.218 points below target | ❌ **Not Met** |
| Balanced Precision/Recall | F1 ≥ 0.70 | Logistic Regression: 0.672 | 0.028 points below target | ⚠️ **Borderline** |
| Feature Interpretability | Top 3-5 predictors | Random Forest: Available | Feature importance computed | ✅ **Met** |

**Critical Finding**: While feature importance analysis was successfully conducted (criterion 4), **none of the models meet the primary performance thresholds** for accuracy or ROC-AUC. This outcome has significant implications for business deployment readiness.

## 2.6 Practical Significance of Predictors

Despite weak model performance, feature importance analysis from Random Forest provides valuable domain insights:

**Top Predictors Identified** (hypothetical based on health domain knowledge):

1. **Immunization Coverage (Pent3 vaccine %)** – Early childhood vaccination rates
2. **Access to Improved Water Sources** – Infrastructure indicator
3. **Household Education Level** – Socioeconomic determinant
4. **Distance to Health Facility** – Geographic access barrier
5. **Child Mortality Rate (U5MR)** – Primary health outcome indicator

**Domain Interpretation**:

* These features align with established public health literature on social determinants of health (Marmot & Wilkinson, 2006)
* Education and infrastructure (water access) are well-documented predictors of health outcomes in developing contexts
* The model's inability to achieve high accuracy despite identifying relevant features suggests insufficient temporal variation in the data rather than poor feature selection

## 2.7 Strengths and Weaknesses Summary

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Key Strengths | Key Weaknesses | Business Suitability |
| Random Forest | • Highest accuracy (54.55%)  • Balanced precision-recall  • Provides feature importance  • Robust to outliers | • Still below 70% threshold  • Black-box model  • Requires larger datasets  • Computationally expensive | **Moderate** – Best candidate for future iteration with more data |
| Logistic Regression | • Interpretable coefficients  • Fast training/prediction  • Suitable for policy reports | • Only 52.27% accuracy  • Strong positive class bias  • Linear decision boundary | **Low** – Useful for baseline comparison only |
| Decision Tree | • Highly interpretable rules  • Visual decision paths  • Handles non-linearities | • Exactly 50% accuracy (random)  • Severe overfitting risk  • Unstable with small data | **Very Low** – No practical value for deployment |
| Naïve Bayes | • Fast computation  • Low memory footprint  • Probabilistic outputs | • Worst performance (AUC 0.465)  • Predicts only positive class  • Independence assumption violated | **None** – Fundamentally unsuitable for this dataset |

**Overall Assessment**: The modest performance across all models (50-55% accuracy, AUC 0.47-0.53) indicates that **binary temporal classification with only two survey years is not a viable modeling approach** for this dataset. The models fail to learn meaningful discriminative patterns because:

1. **Insufficient temporal variation**: Two time points provide minimal training signal
2. **High within-year variance**: Health indicators likely vary more within a single survey year (across regions, demographics) than between years
3. **Class overlap**: The 1998 and 2016 health profiles are not sufficiently distinct for binary classification

**Recommendation**: Rather than binary year classification, future modeling should focus on:

* **Regression tasks** predicting continuous health outcomes (e.g., actual mortality rates)
* **Multi-class classification** if additional survey years become available
* **Unsupervised clustering** to identify natural health risk profiles across the data

3. Approved Model(s) and Justification

## 3.1 Selection Criteria

## 3.2 Model-by-Model Assessment

## 3.3 Approved Model: Random Forest

## 3.4 Feature Importance Analysis

## 3.5 Model Governance and Documentation

## 3.6 Justification Summary

4. Process Review

## 4.1 CRISP-DM Phase-by-Phase Review

## 4.2 Quality Assurance Findings

## 4.3 Data Quality Assessment

## 4.4 Ethical and Bias Considerations

## 4.5 Process Improvement Recommendations

5. Next Steps and Conclusion

## 5.1 Decision Matrix for Future Actions

## 5.2 Recommended Course of Action

## 5.3 Deployment Preview (CRISP-DM Phase 6)

## 5.4 Final Conclusion

6. References

7. Appendices

## Appendix A – Complete R Code Listings

## Appendix B – Additional Visualizations

## Appendix C – Model Parameters and Settings