# Section 3: Process Review

**Author:** Person 3 (Data Quality & Process Audit Specialist)  
**Date:** February 2025

## 3.1 Executive Summary

This section provides a comprehensive audit of the CRISP-DM process execution across Phases 1-4, evaluating data quality, process adherence, and identifying areas for improvement. The analysis reveals that while the process was methodically executed, significant limitations in data scope and model performance highlight the need for substantial improvements before production deployment.

## 3.2 Data Quality Assessment

### 3.2.1 Dataset Overview

The final cleaned dataset contains 294 records across 3 variables (Indicator, Value, SurveyYear), representing a total of 147 health indicators measured across two survey years (1998 and 2016). The dataset size of 0.02 MB indicates a relatively small-scale analysis.

### 3.2.2 Data Completeness Analysis

**Missing Data Assessment:** - **Missing Values:** 0% across all variables - **Data Completeness:** 100% - no missing values detected - **Data Integrity:** All records contain complete information

**Outlier Detection:** - **Value Variable:** 33 outliers detected (11.22% of records) - **SurveyYear Variable:** 0 outliers (as expected for categorical data) - **Outlier Impact:** Moderate - outliers in health indicator values may represent genuine extreme cases rather than data errors

### 3.2.3 Data Distribution Analysis

**Target Variable Distribution:** - **Class Balance:** Perfect 1:1 ratio (147 records each for 1998 and 2016) - **Temporal Coverage:** Limited to two time points, 18 years apart - **Representativeness:** Balanced sampling across survey years

**Indicator Distribution:** - **Total Indicators:** 147 unique health indicators - **Coverage:** Comprehensive across maternal health, child health, immunization, water/sanitation, and nutrition domains - **Consistency:** Each indicator appears exactly twice (once per survey year)

## 3.3 CRISP-DM Process Review

### 3.3.1 Process Execution Assessment

| Phase | Goal | Executed | Issues Found | Improvement Needed |
| --- | --- | --- | --- | --- |
| **1. Business Understanding** | Define business objectives | ✓ Completed | Limited to 2 survey years | Include more survey years |
|  | Identify success criteria | ✓ Completed | Thresholds may be too ambitious | Reassess success criteria |
|  | Stakeholder analysis | ✓ Completed | Limited stakeholder input | Engage more stakeholders |
| **2. Data Understanding** | Initial data collection | ✓ Completed | Small sample size (294 records) | Collect additional data sources |
|  | Data quality assessment | ✓ Completed | Missing values in some indicators | Implement better imputation |
|  | Data exploration | ✓ Completed | Limited temporal coverage | Expand temporal analysis |
| **3. Data Preparation** | Data cleaning | ✓ Completed | Some data loss during cleaning | Preserve more data |
|  | Feature engineering | ✓ Completed | Basic feature engineering only | Advanced feature engineering |
|  | Data integration | ✓ Completed | Potential information loss | Better integration strategy |
| **4. Modeling** | Model selection | ✓ Completed | All models performed poorly | Try ensemble methods |
|  | Model training | ✓ Completed | Limited hyperparameter tuning | Implement grid search |
|  | Model evaluation | ✓ Completed | No cross-validation used | Add proper validation |

### 3.3.2 Quality Assurance Checklist

| Category | Check Item | Status | Notes |
| --- | --- | --- | --- |
| **Data Quality** | Missing values identified and handled | ✓ PASS | Missing values handled via omission |
|  | Outliers detected and addressed | ✓ PASS | Outliers identified but not removed |
|  | Data types correctly assigned | ✓ PASS | All data types appropriate |
|  | Consistent data formats | ✓ PASS | Consistent formatting applied |
|  | No duplicate records | ✓ PASS | Duplicates removed successfully |
| **Process Quality** | Business objectives clearly defined | ⚠ PARTIAL | Limited to 2-year analysis |
|  | Success criteria measurable | ⚠ PARTIAL | 70% accuracy threshold ambitious |
|  | Data collection process documented | ✓ PASS | Process well documented |
|  | Cleaning steps reproducible | ✓ PASS | Steps are reproducible |
| **Model Quality** | Models properly trained | ⚠ PARTIAL | All models underperformed |
|  | Performance metrics calculated | ✓ PASS | Comprehensive metrics calculated |
|  | Results interpreted correctly | ⚠ PARTIAL | Results show poor performance |
| **Documentation** | Code well-commented | ✓ PASS | Code is well documented |
|  | Process documented | ✓ PASS | Process clearly documented |
|  | Results reproducible | ✓ PASS | Fully reproducible |

## 3.4 Ethical Considerations and Bias Analysis

### 3.4.1 Identified Biases

| Bias Type | Description | Impact | Mitigation |
| --- | --- | --- | --- |
| **Temporal Bias** | Only 2 survey years (1998, 2016) - limited temporal coverage | High - Cannot detect trends or patterns over time | Include more survey years if available |
| **Geographic Bias** | National-level data only - no regional variation analysis | Medium - Misses regional health disparities | Analyze provincial/regional data if available |
| **Indicator Selection Bias** | Limited to 12 health indicators - may miss important factors | High - May exclude critical health determinants | Expand indicator selection based on literature |
| **Sample Size Bias** | Small sample size (294 records) - limited statistical power | High - Results may not be statistically significant | Collect additional data or use synthetic data |
| **Measurement Bias** | Different survey methodologies between years | Medium - Affects comparability between years | Standardize measurement approaches |
| **Representation Bias** | May not represent all population segments equally | Medium - May not capture all demographic groups | Ensure representative sampling across demographics |

### 3.4.2 Fairness and Equity Considerations

The analysis reveals several ethical concerns that must be addressed: - **Limited Demographic Representation:** The dataset may not adequately represent all population segments - **Temporal Limitations:** 18-year gap between surveys may not capture recent health trends - **Geographic Aggregation:** National-level analysis may mask important regional disparities

## 3.5 Process Improvement Recommendations

### 3.5.1 Priority Improvements

| Phase | Current State | Recommended Improvements | Priority | Effort Required |
| --- | --- | --- | --- | --- |
| **Business Understanding** | Limited to 2-year analysis scope | Expand to multi-year analysis, engage more stakeholders, define realistic success criteria | High | Medium |
| **Data Understanding** | Small dataset with limited indicators | Collect additional data sources, include regional data, expand indicator selection | High | High |
| **Data Preparation** | Basic cleaning and feature engineering | Implement advanced feature engineering, better missing data handling, data validation | Medium | Medium |
| **Modeling** | Simple models with poor performance | Try ensemble methods, hyperparameter tuning, cross-validation, advanced algorithms | High | Medium |
| **Overall Process** | Manual process with limited automation | Implement automated data pipelines, version control, continuous monitoring | Medium | High |

### 3.5.2 Specific Technical Improvements

**Data Collection Enhancements:** - Include additional survey years (2003, 2008, 2011) for better temporal coverage - Collect provincial/regional level data for geographic analysis - Expand indicator selection based on literature review - Implement data validation at source

**Modeling Improvements:** - Implement proper cross-validation strategies - Apply hyperparameter tuning using grid search - Explore ensemble methods (Voting, Bagging, Boosting) - Consider advanced algorithms (XGBoost, Neural Networks)

**Process Automation:** - Develop automated data pipelines - Implement version control for models and data - Create continuous monitoring dashboards - Establish model retraining schedules

## 3.6 Key Findings and Conclusions

### 3.6.1 Process Strengths

1. **Methodical Execution:** All CRISP-DM phases were systematically completed
2. **Data Quality:** High data completeness with no missing values
3. **Documentation:** Comprehensive code documentation and process tracking
4. **Reproducibility:** Fully reproducible analysis with clear audit trail

### 3.6.2 Critical Limitations

1. **Data Scope:** Limited temporal coverage (2 years, 18 years apart)
2. **Sample Size:** Small dataset (294 records) limits statistical power
3. **Model Performance:** All models failed to meet business success criteria
4. **Geographic Coverage:** National-level analysis misses regional variations

### 3.6.3 Recommendations for Future Iterations

1. **Immediate Actions:**
   * Reassess business success criteria to be more realistic
   * Collect additional survey years if available
   * Implement proper cross-validation
2. **Medium-term Improvements:**
   * Expand data collection to include regional data
   * Implement advanced feature engineering
   * Try ensemble modeling approaches
3. **Long-term Strategy:**
   * Develop automated data pipelines
   * Establish continuous monitoring systems
   * Create stakeholder engagement framework

## 3.7 Conclusion

The CRISP-DM process was executed with high methodological rigor, but the project’s success was fundamentally limited by data constraints. The small sample size, limited temporal coverage, and ambitious success criteria resulted in models that cannot meet business objectives. However, the process itself was well-documented and reproducible, providing a solid foundation for future iterations.

**Key Takeaway:** While the technical execution was sound, the project requires significant data enhancement and process refinement before it can deliver actionable insights for South African health policy decisions. The recommended improvements focus on expanding data scope, refining success criteria, and implementing more sophisticated modeling approaches.

**Next Steps:** This process review should inform the team’s decision on whether to proceed with deployment, further iteration, or project restart, as outlined in Section 4 of the final report.