ANGELITE ARENDSE 600938
SIMEON LE ROUX 578047
IAN FASEEN 600148
KAGISO SEBATI 600836

BUSINESS INTELLIGENCE MILESTONE 4

BELGIUM CAMPUS ITVERSITY
10/10/2025

TABLE OF CONTENTS

1.1 Assessment of Results

Restate Business Objectives.

The main business goals of the project, as was outlined in the Business Understanding stage were to:

- Determine the major factors that cause certain health outcomes based on the data provided in the demographics and health surveys.
- Identify underserved or vulnerable populations according to geographical, social, or economic benchmarks.
- Direct strategic decision-making in resource allocation, program development, and implementation of policies.
- Create a strong data base to support the continuous monitoring and assessment of the public health programs.

The criteria of success were not only technical accuracy, but also impact, adoption, efficiency, and actionability e.g., the rate of immunisation, usage of dash boards by the stakeholders and enabling more efficient distribution of the resources.

Measure Against Business Objectives.

Two different models were created and compared Multiple Linear Regression (MLR) and Random Forest Regression (RF).

Multiple Linear Regression (MLR):

- Used as a transparent model of basis.
- Estimated extended linear relationships but had greater RMSE and less predictive power.
- Heteroscedasticity and nonlinearity were identified in residual plots which limited the robustness.

Random Forest Regression (RF):

- Performed better than MLR on 10-fold cross-validation in terms of lower RMSE, which satisfies the success criteria specified at the outset of the investigation of enhanced predictive accuracy.
- Self-proven resilience to multicollinearity and nonlinearity.
- According to the variable importance analysis, the most important predictors were the temporal persistence (valuez, valuelag1) followed by the service-related predictors (facility delivery, antenatal care) which had secondary yet significant explanatory value.

Assessment Against Goals:

- Accuracy: RF was higher in predictive accuracy compared to MLR but was less interpretable because of reliance on a small number of significant predictors.
- Robustness: RF passed the robustness test, generalizing effectively across folds and avoiding multicollinearity.
- Interpretability: MLR gave more understandable coefficients to consider policy implications, yet VIF of RF gave actionable feedback.
- Business Success Criteria: The models failed to generate the ideal explanatory power (Adjusted R2 < 1), but RF was more than 0.70 required in the case of reliability, and as such, they satisfied the technical criterion of business value.

Expansive Justification (Business Situation)

In the big picture, the outcomes should be very valuable to a business although not everything was achieved according to the technical goals:

- Policy Relevance: The observation that the temporal persistence (past values of health indicators) is a strong predictor of the present outcome in terms of skilled birth attendance supports the need to pursue long term, sustained interventions and not a single campaign.
- Resource Allocation: The fact that identifying the facility based delivery and skilled antenatal care as a secondary driver is highly precise would mean that investments towards the healthcare infrastructure and training are highly well justified.
- Actionability: Despite the partial recall of less dominant factors, the models nevertheless point to very evident intervention spots (e.g., facility access improvement, midwives support) that are government and NGO priorities.
- Efficiency Gains: With the most powerful predictors in mind, the stakeholders can have an even more efficient allocation of the resources into their possession, which can be seen as reaching the success criterion of the 15 percent increase in efficiency.
- Adoption Potential: It is possible to adopt the results into Power BI dashboards to monitor them continuously, which can be considered supporting the adoption criterion of 80% stakeholder use.

In conclusion, the interpretability of the Random Forest model is not as good as that of MLR, but the model has a better predictive power and strength that qualifies it as a useful tool of decision support. Even not ideal, the resulting insights are practical, policy-focused, and can initiate quantifiable changes in the path of health results of people.

1.2 APPROVED MODELS

Identification of Approved Models

After carefully comparing the performance of Multiple Linear Regression and Random Forest Regression, we've decided to move forward with the Random Forest model. Its results were consistently stronger, making it the preferred choice for future use and possible deployment.

Justification of Chosen Model

- After we compared the performance of Multiple Linear Regression and Random Forest, the Random Forest model stood out for its stronger predictive accuracy and overall robustness. While Multiple Linear Regression was easier to interpret, it ran into trouble with uneven error patterns and nonlinear relationships, making it harder to trust its predictions on new data. By comparison, the Random Forest model consistently produced lower RMSE scores and more stable results during cross-validation, making it a strong fit for our goal of building a model that's not just accurate, but also dependable in real-world use.
- Technically, Random Forest is a smart choice. It blends multiple decision trees using bootstrapped samples, which not only keeps overfitting in check but also helps it uncover the complex, nonlinear relationships that simpler models tend to miss. This was especially helpful when modelling how factors like facility-based delivery and antenatal care interact, two variables that don't follow a straight-line pattern but are crucial for predicting health outcomes.
- Random Forest not only delivers strong predictions; it also helps highlight what is driving those outcomes. Its feature importance rankings bring attention to the influence of time-based patterns and access to healthcare services, offering practical direction for where efforts can make the biggest difference. These insights make it easier for us to design interventions that improve equity and efficiency. And while Random Forest may not be as easy to interpret as Multiple Linear Regression, its ability to handle messy, real-world data and turn it into reliable, actionable guidance makes it a trustworthy tool for implementation.

Robustness and Scalability

- Random Forest holds up well when faced with new or unfamiliar data, thanks to the way it averages across multiple decision trees and handles multicollinearity without breaking a sweat. It's flexible enough to work across different regions and demographic groups, which makes it a strong fit for public health dashboards or automated reporting tools. Plus, it can easily take on new variables without needing a full rebuild, making it a practical choice for long-term health monitoring systems that need to grow and adapt over time.

Conclusion

While the MLR model was easier to interpret, Random Forest was ultimately the better fit. Its stronger accuracy, resilience, and alignment with our goal of delivering actionable, data-driven insights made it the clear choice. Deploying RF won't just improve the reliability of our predictions.

1.3 Review the Process

This section will review the processes taken during milestone 1 to 3, reflecting back onto the CRISP-DM methodology

Milestone/Phase 1

The goal of this phase was to address the correct business problem and clearly define success.

The business understanding was clearly defined the objective which was to leverage health data for actionable public health insights. It established quantifiable success criteria for example: a 15% increase in resource efficiency and the active use of dashboards by 80% of stakeholders. The determination was excellently executed; the project successfully transitioned from a technical idea to business objective, ensuring all downstream efforts target measurable organisational impact.

The initial data understanding identified the multi-score, multi-year DHS data covering Maternal/Child health, fertility and literacy. R was used to for the analysis and Power BI for visualisation. The data landscape and toolset were established early on, setting realistic boundaries for the scope, there for it was properly executed.

The data quality and EDA have identified key data issues and established an initial hypotheses which was that higher literacy is associated with improved child health outcomes. This analysis was properly executed and was crucial for identifying early trends and directing the focus of the subsequent data preparation steps.

Milestone/Phase 2

The goal was to transform raw data into a clean, non-redundant and statistically validate set of features for modelling. This phase demonstrated the strongest analytical rigor.

With the data selection the initial 13 datasets were reduced to 9 core datasets, based on four strict criteria: Relevance, data quality, feasibility and coverage. This was excellently executed. The focused approach ensured that resources were not wasted on data with low signal or higher redundancy making the model more robust.

The handling of data leakage and redundancy resulted in the exclusion of high-correlated, redundant fields, dropping “DenominatorUnweighted” in favour of “DenominatorWeighted” and “SurveyYearLabel” in favour of “SurveyYear”. This maintains model parsimony and efficiency, resulting that it was properly executed.

To prevent fatal errors “CILow” and “CIHigh” was a crucial exclusion from the Child Mortality and DHS quick stat datasets, which had a near-perfect correlation with the target “Value”. This was the single most important decision in Phase 2. Failing to exclude these fields would have introduced data leakage, resulting in a model that was 99% accurate in training but entirely useless for predicting new data.

Milestone/Phase 3

The modeling strategy adopted the Regression Model to predict the continues “Value” field. The strategy included using a complex model, Random Forest, as the primary technique and a simple model, Multiple Linear regression, as a neccessary benchmark. Using a benchmark model allowed for direct comparison of complexity versus interpretability, justifying the final choice. This is there for properly executed.

For model validation we used 10-fold cross-validation to ensure that the model results were not reliant on a single arbitrary data split, enhancing the generalisability of the findings. This confirmed the model’s robustness and ensures it can generalise well to new data.

The final assessment was concluded that the Random Forest model was confirmed as the superior predictive tool, offering higher accuracy and better robustness, therby passing the project’s predifined success criteria. This was succesfully completed due to the project concluding by a directly linking the technical outcome back to the initial business objectives, confirming the solution is fit for purpose.

The final verdict

The entire process, from phase 1 to 3, was executed correclty and with strong analytical decipline. The structure provided by the CRISP-DM methodology ensured that the project maintained focus, performed neccessary data cleaning, preventing critical errors like data leakage and ultimately delivered a validated model that met its initial success goals.

Technical and Methodology Review: Issues Missed or Overlooked

While the execution was rigorous, particulary in preventing data leakage during the feature selection, key methodological and documentation issues remain. Specifically, the project missed the oppertunity to perform a casual analysis using the perfectly correlated feature groups, failed to document the chosen imputation strategy for missing data, and did not address the high data granularity limitation introduced by the reliance on categorical identifiers as predictors. These issues introduced potential biases and limit the ultimate actionability and generalization of the final predictive model.

The project succesfully navigated the initial stages of the CRISP-DM cycle, culiminating in the selection of a robust Random Forest model for predicitng continues health indicators. The rigor applied in Phase 2, particularly the exclusion of confidence intervals, to mitigate data leakage, is commendable. However, the academic standard requires full transparency and mitigation of all identified analytical opportunities and limitations. The following sections detail critical issues that were either overlooked or insufficiently documented and necessitate a repeat or refinment of specific steps.

Detailed Review of Overlooked issues

Incomplete Data preparation: Undocumented Missing Value Imputation

The initial data quality assessment in Phase 1 noted the presence of missing values in certain varialbes, such as confidence intervals. The report, however, moves directly from feature selection to modeling without explicity documenting the final strategy employed to handle these missing data points.

- Issue: The chosen method was not reported. The choice if imputation technique significantly impacts model variance, bias and generalisability.
- Impact: Without this documentation, the model's performance cannot be fully verified, as simple imputation methods can artificially reduce data variance, leading to an underestimation of error and an over-optemistic view of the model's accuracy.
- Recommendation: The Data preparation phase must be repeated to include a documented section detailing the missing value imputation strategy applied to each dataset

Methodology Oversight: Casual Analysis Opportunity Missed

The correlation analyses in Phase 2 revealed instances of perfect or near perfect correlation between varialbes that were not target leakage indicators.

- Issue: The current approach was to exclude two of the three perfect correlated variables simply to avoid redundancy. While correct for a standard regression model, this decision failed to leverage this perfect correlation for a deeper understanding of the feature set.
- Impact: When variables are perfectly correlated, it often signifies that one is a direct inverse or categorial proxy of the other. The project missed the opportunity to explicitly state the causal or deinitional relationship between theseindicators and justify which single feature is the most interpretable representative of the latent construct.
- Recommendation: A brief Casual Analysis and Interpretation step should be added to Phase 2, explicitly defining the relationship among high correlated non-target features before exclusion.

Limitation in Predicitve Features: Over-reliance on Granular Identifiers

The final Random Forest model in Phase 3 relied heavily on a few dominant predictors, specifically "SurveyYear","CharacterisitcID" and "IndicatorOrder".

- Issue: "CharacterisitcID" and "IndicatorOrder" are essentially categorial identifiers or ordinal ranks used within the DHS data structure. Their predictive power confrims they contain information, but they do not provide actionable, real-world insight in isolation.
- Impact: The finding that "CharacteristicID is a strong predictor" is less actionalbe for a policy maker than "Maternal literacy levels are strong predictors." The model has essentially learned to segment the data by the survey's internal identifiers, which limits the interpretability and generalisability of the findings to a new or different datasets.
- Recommendation: A Feature Engineering step is needed. The team should attempt to map the highest-impact "CharacteristicId" and :IndicatorOrder" Values back to

their actual descriptive labels. The model should be re-run using these engineered, interpretable features to confirm if the predictive power is maintained. This refinment would enhance the actionalbility of Phase 3's results, linking the technical prediction back to the socioeconomic drivers identified in Phase 1.

Comprehensive Project Review

The project achieved its primary objective by producing a model that surpassed the Multiple Linear Regression benchmark in predictive accuracy, reduced RMSE. Crucial quality assurance was achieved by proactively preventing data leakage. However, the review identifies two critical methodological shortcomings: the non-documentation of the Missing Value Imputation strategy, and a failure to adequately interpret the highly predictive but non-actionable categorical features. Corrective action is required to address these deficiencies, ensuring the model's transparency and maximizing its ultimate value for policy makers.

Synthesis of Key Project Findings

The project followed the iterative CRISP-DM framework, successfully translating a broad public health challenge into a validated predictive model.

Phase 1: Business and Data Understanding

- Objective Confirmation: The core business objective ,to identify key drivers of health outcomes and inform resource allocation, was clearly established and remained the guiding principle throughout the modeling process.
- Initial Trends: Exploratory Data Analysis (EDA) suggested early hypotheses, such as the correlation between literacy rates and improved child health outcomes, which later guided feature selection.

Phase 2: Data Preparation and Feature Selection

- Feature Validation: Rigorous correlation analysis identified statistically significant linear relationships, justifying the inclusion of features like Precision ,a strong negative predictor, and IndicatorOrder.
- Success in Data Integrity: The exclusion of variables like DenominatorUnweighted and, critically, CILow and CIHigh (Confidence Intervals) prevented data leakage. This is the most significant quality assurance success, ensuring the final model's low RMSE is genuine and not an artifact of circular logic.

Phase 3: Modeling and Evaluation

- Model Performance: The Random Forest (RF) model was superior to the benchmark MLR, providing higher accuracy and better robustness across the 10-fold cross-validation procedure. This successful outcome met the predefined Accuracy and Robustness success criteria.

- Dominant Predictors: The RF model's high accuracy was driven primarily by SurveyYear, CharacteristicId, and IndicatorOrder.
- Weak Predictors: Postnatal checkups and other child treatment indicators were found to have little predictive value for the specific target, informing policy makers on where to de-prioritize focus.

Quality Assurance Issues Missed and Overlooked

While the technical execution was sound, two critical quality assurance issues were missed or overlooked, impacting transparency and actionability.

Issue 1: Undocumented Missing Value Imputation Strategy

- Finding: The initial quality assessment noted missing values across the nine selected datasets. However, the documentation for Phase 2 is silent on the specific methodology used to address these missing data points prior to modeling.
- Consequence: Failure to document the imputation strategy introduces an element of opacity and potential bias. For example, using a simple technique like mean imputation can artificially reduce the variance of the data, potentially leading to an over-optimistic Root Mean Square Error (RMSE) for the RF model.
- Quality Assurance Standard Violated: Transparency and reproducibility of the Data Preparation step.

Issue 2: Insufficient Interpretation of High-Impact Categorical Features

- Finding: The model's success is heavily reliant on the CharacteristicId and IndicatorOrder fields, which are internal survey identifiers. The report correctly identifies their dominance but concludes with a generalized variable importance ranking.
- Consequence: Identifying that CharacteristicId is important is not actionable for a policy maker. The policy utility is derived from knowing what that ID represents. This is an interpretability gap that limits the final stage of the CRISP-DM cycle Deployment.
- Quality Assurance Standard Violated: Actionability and Interpretability of the final model's core features.

Corrective Actions Required

To correct the oversights and ensure the project meets the highest academic and practical standards, the following steps must be repeated or implemented immediately:

- Document Imputation: This must be applied to Phase 2. The team must retroactively document the exact technique used to handle missing data and include a justification for that choice. This ensures the model's reported accuracy is verifiable.

- Map and Reengineer Categorial Features: This must be applied in Phase 2 and 3. The team must map the dominant “CharacteristicId” and “IndicatorOrder” values back to their original descriptive labels. The model should then be re-run with these descriptive features to confirm if the predictive power is maintained. This transforms a technical finding into an actionable policy lever.
- Causal Analysis of Correlated features: To be implemented in Phase 2. A brief explanatory analysis must be conducted on the perfectly correlated predictor groups. This step should confirm the inverse or proxy relationship between these features, fully justifying the exclusion of redundant variables beyond a mere statistical correlation measure.

Methodological Refinement

While the project achieved its technical objectives, future iterations must prioritize improved documentation, advanced feature engineering, and rigorous model interpretability. The proposed improvements include integrating a mandatory imputation justification ledger, transitioning from basic statistical feature selection to domain-driven feature engineering, and implementing advanced model explanation techniques to transform black-box predictions into transparent, actionable policy recommendations.

The current project successfully validated the Random Forest model's superiority over a Multiple Linear Regression benchmark, fulfilling the project's initial success criteria for accuracy and robustness. The project was commendable for its rigor in preventing data leakage. However, continuous improvement, a core tenet of the iterative CRISP-DM methodology, demands that we address limitations in documentation and the actionability of findings identified in the previous review. The following recommendations focus on strengthening the process in the Data Preparation and Modeling phases to maximize both the academic rigor and real-world utility of future public health data mining projects.

Proposed Improvements to the Project Process

Improvement in Phase 2: Data Preparation

The primary area for improvement in the data preparation phase lies in making the handling of data quality issues transparent and justifiable, addressing the current omission of the missing value imputation strategy.

Mandatory Imputation and Transformation Ledger

- Current Deficiency: The process failed to document the exact methodology used for handling missing values, leading to an interpretability gap regarding potential bias.
- Proposed Improvement: Future projects must integrate a formal Imputation and Transformation Ledger into the Data Preparation phase. This ledger will require:
    1. Variable: Identification of the variable requiring correction.
    2. Technique: Stating the exact method used.

3. Rationale: A brief explanation justifying the choice.
4. Transformation: Documenting all scaling or encoding applied (e.g., "Log transformation applied to 'Value' to stabilize variance and address skewness").

- Anticipated Benefit: This ensures full transparency, establishes traceability for results, and forces the team to choose statistically appropriate methods rather than arbitrary, default techniques, thereby minimizing artificial compression of the model's prediction error.

Proactive, Domain-Driven Feature Engineering

- Current Deficiency: Feature selection included highly predictive but difficult-to-interpret features like CharacteristicId and IndicatorOrder, which are merely survey identifiers.
- Proposed Improvement: The future process should mandate an expanded Feature Engineering sub-phase before the statistical correlation screen. The process should involve:
    1. Mapping and Interpretation: All categorical ID fields must be mapped back to their human-readable, policy-relevant labels.
    2. Creation of Composite Features: Develop new, hypothesis-driven features based on public health domain knowledge, such as a "Health Service Access Score" combining indicators like facility delivery percentage and antenatal care coverage.
    3. Modeling with Actionable Features: The final model would then be built and validated on these engineered, actionable features, ensuring that the model's high predictive power is derived from variables that policy makers can directly manipulate.
- Anticipated Benefit: Directly addresses the model actionability deficit. The output transitions from "Policy is driven by a strong CharacteristicId" to the more useful "Policy should target communities with low Maternal Education Level."

Improvement in Phase 3: Modeling and Evaluation (Model Interpretability)

While Random Forest provided high accuracy, its "black-box" nature limited direct interpretation, making policy communication difficult.

Integration of Advanced Model Explanation Techniques (SHAP/LIME)

- Current Deficiency: The model assessment relied on a simple Variable Importance ranking, which showed what features mattered, but not how they influenced the prediction.

- Proposed Improvement: Integrate a dedicated Model Explanation step using techniques such as SHapley Additive exPlanations (SHAP) or Local Interpretable Model-agnostic Explanations (LIME).
    - Global Interpretation (SHAP): Provides a globally consistent and theoretically sound quantification of how much each feature contributes to the final prediction, including the direction of the effect.
    - Local Interpretation (LIME): Explains individual predictions, allowing the team to drill down into why the model predicted a high mortality rate for a specific region or demographic group.
- Anticipated Benefit: Transforms the black-box RF model into a glass-box explanation tool. This is critical for the final deployment phase, as stakeholders require an explanation for a prediction before trusting the output.

Mandatory Stakeholder Review and Dashboard Prototyping

- Proposed Improvement: Introduce a mandatory Stakeholder Review Checkpoint at the end of the Evaluation phases. This involves building a low-fidelity Power BI dashboard prototype using the model's output and presenting it to a representative stakeholder group before final deployment.
- Anticipated Benefit: This checkpoint ensures that the model's output format, the terminology used in the visualizations, and the presented insights are immediately comprehensible and useful to the end-users, guaranteeing that the model delivers on the Phase 1 business objective of actionability.

Longer evidence base for smarter resource allocation and long-term planning in health interventions.

1.4 Determine the next steps

The Random Forest model has proven to be the most reliable and accurate forecasting tool, with the fulfilment from milestone 1 to 3. However, the review process identified areas that need improvement to enhance interpretability, transparency, as well as long term sustainability. The following steps outline the measures required to transition from a technical successful model to a fully deployable business solution.

1.4.1 Revisit and Document Data Preparation Decisions

Before deploying the model, the project team must go over the data preparation stage again to make sure that all data handling practises, particularly missing value imputation, are thoroughly documented and supported; by creating an Imputation and Transformation Ledger that documents the imputation process, the rationale for its selection, and any encoding or transformations that were used.

This will result in increased confidence in the model's stated performance indicators, transparency, and reproducibility.

1.4.2 Conduct Feature Re-engineering and Interpretability Enhancement

To improve the model's interpretability and policy relevance, it is necessary to re-map categorical identifiers to their descriptive, real-world meanings. This must be done by using feature mapping to associate survey identifiers with relevant descriptive variables. To confirm predictive stability, rerun the Random Forest model with the interpretable features.

The expected outcomes are improved interpretability and clear alignment between technological discoveries and practical public health strategies.

2 Implement Casual Analysis on Correlated Predictors

Clarifying the connections between the highly correlated variables and defending the inclusion or removal of particular features will be accomplished through a casual analysis stage. The team must determine the most significant representatives of each correlated feature group by interpreting perfect or nearly perfect correlations using statistical and domain-based reasoning.

Better model justification and a deeper comprehension of the underlying data relationships are expected.

1.4.3 Integrate Advanced Model Explanation Tools

Despite the Random Forest's excellent accuracy, incorporating contemporary explain ability strategies will improve stakeholder trust and offer more profound understanding of the factors influencing health outcomes. Team members must take action by using

LIME (Local Interpretable Model-agnostic Explanations) for case specific, local interpretations and SHAP for global feature contribution analysis.

Anticipated results include clear, comprehensible model output that may be used with assurance in academic reporting and decision reporting.

1.4.4 Prototype and Validate a Stakeholder Dashboard

Making sure the model's outputs are intelligible and useful to decision makers is crucial as the CRISP-DM cycle's last step places a strong emphasis on deployment and adoption. Project team must take action by creating a low-fidelity PowerBI dashboard prototype that displays interactive filters, important feature impacts, and model forecasts. This prototype can be showed to interested parties in health policy for evaluation and comments.

Anticipated results include involving stakeholders early on guarantees usability, promotes adoption, and confirms that the project achieves its business goals of providing insights based on data that can be put to use.

1.4.5 Plan for Continuous Model Improvement and Monitoring

In public health, predictive modelling is an iterative process. Retraining and recalibration will be necessary as new data becomes available in order to preserve the accuracy and relevance of the model., by creating a model maintenance plan that includes performance tracking metrics like RMSE drift and feature importance adjustments, frequent, retraining schedules and version control (GitHub).

Results include a long lasting, dynamic prediction system that can enable long term public health decision-making and adjust to new data trends.

1.4.6 Expand Future Research Directions

Future stages can investigate further machine learning methods and data integrations to broaden the predictive scope beyond the current improvements. The team can take action by assessing XGBoost or Gradient Boosting for performance comparison. Examine integrating with spatial or real-time datasets.