PROJECT MILESTONE 3

Petrus Human 577842

Frederik Knoetze 600965

Teleki Shai 601377

Moloko Rakumako 601352

Module: Business Intelligence 381
Project: Health & Demographic Data Science using CRISP-DM – Modelling Phase
Milestone: 3 Modelling (CRISP-DM Phase 4)

BELGIUM CAMPUS ITVERSITY

Table of Contents

1. Introduction (CRISP-DM Phase 4: Modelling)

This milestone advances our project from **CRISP-DM Phase 2–3 (Data Understanding & Preparation)** into **Phase 4 (Modelling)**. In Milestone 1, we framed the **business understanding** around public-health questions (e.g., access to care, coverage, and risk indicators). In Milestone 2, we prepared twelve national health/demographic datasets, but the final rotated and cleaned dataset contained only **two temporal records (two years)**. This unusual structure significantly limits the modelling options.

Because there are only **two rows of usable data**, a conventional **70/15/15 split** is infeasible. The only possible option is a **50/50 train–test split** (training on one year and testing on the other). Even then, the model will essentially memorise the single training year, making evaluation unstable. Alternatively, both years may be used for training to demonstrate methodology, but the resulting model cannot be properly validated.

Additionally, most predictors are **continuous**; categorical targets must be engineered (e.g., transforming infant mortality into a binary high-risk variable above a threshold). Since the dataset is fundamentally temporal, a **time-series perspective** (e.g., treating the two years as sequential observations) may also be appropriate in later stages when more records become available.

Despite these constraints, Milestone 3 still specifies, justifies, and documents **candidate models** to illustrate methodological fit. We follow CRISP-DM's guidance to:

- select algorithms consistent with our data types and goals
- document assumptions
- design tests (splits/CV/metrics)
- evaluate results before Phase 5 (Evaluation) and Phase 6 (Deployment).

While CRISP-DM is a complete framework, recent work highlights its continued relevance and agile adaptations for modern data science projects. We adopt it in our workflow (iterating between preparation and modelling as data quality/feature needs emerge) (Silva and Viana, 2024).

2. Modelling Techniques: Rationale & Assumptions

2.1 Rationale narrative (why these four)

- Logistic Regression (GLM, logit link).
  Standard for binary clinical outcomes once continuous targets are transformed (e.g., infant mortality above a threshold = 1, otherwise 0). Provides interpretable odds ratios. Assumes linearity of log-odds and sufficient sample size, which is

not met here, so results will be illustrative only (Harris, Yang and Hardin, 2021; Hua and Zhang, 2025).

- Decision Trees (CART/C5.0).
  Produce intuitive, rule-based outputs and handle non-linearities. Even with limited data, they can demonstrate how categorical splits would occur if more records were available. Overfitting is a risk in small datasets (Nguyen et al., 2023; Rahmati et al., 2024).
- Random Forest.
  Ensemble method that reduces variance by averaging across trees. While the dataset size is insufficient for robust training, it remains valuable for showing how ensemble methods provide stability and variable importance when larger data are available (Tran, Nguyen and Le, 2024; Wallace, Diez Roux and Greven, 2023).
- Naïve Bayes.
  Normally suited for categorical data, which must be derived here. Serves as a baseline comparator due to its speed and simplicity, though the independence assumption is unlikely to hold (ScienceDirect Topics, 2025).

2.2 Technique–Assumption–Use-case table

Algorithm	Why we chose it (fit to our data/goals)	Key assumptions / caveats	Typical health use-cases & notes
Logistic Regression	Transparent coefficients/ORs; baseline for binary health outcomes (Harris, Yang and Hardin, 2021).	Linearity of log-odds; no perfect multicollinearity; adequate events per variable (Hua and Zhang, 2025).	Clinical risk models, program targeting; limited here by 2 rows.
Decision Trees (CART/C5.0)	Interpretable rules; handle mixed data types (Nguyen et al., 2023).	Overfit easily on tiny datasets; unstable splits.	Risk stratification, triage pathways (Rahmati et al., 2024).
Random Forest	Robust predictive performance; ensemble stability (Tran, Nguyen and Le, 2024).	Requires larger datasets; feature importance may mislead (Wallace, Diez Roux and Greven, 2023).	Discharge risk, cost prediction; here only illustrative.
Naïve Bayes	Fast baseline; good with categorical data (ScienceDirect Topics, 2025).	Conditional independence assumption; not realistic in health data.	Screening baselines, categorical-heavy tasks; requires engineered variables.

2.3 Metrics & evaluation note (for continuity with Person 2)

Because of the dataset's extreme limitations, any metrics must be interpreted cautiously. In principle, we prioritise **ROC-AUC, F1, Precision/Recall, and calibration**, which are widely used for binary health classifiers (Li, Zhao and Xu, 2024). However, with only two records, metrics are unstable and results should be considered purely methodological.

3. Generation of the test parameters and train test split

Because the dataset contains only **two years of observations**, the only split available is **50/50**. This produces one training year and one testing year. This approach demonstrates methodology but cannot yield a reliable model. Alternatively, using both years for training allows a wider fit but removes any test validity.

Thus, two configurations will be considered:

1. **50/50 split** – to illustrate the concept of training and testing.
2. **Full dataset training** – to demonstrate how models behave with minimal data.

Classification Metrics

ROC-AUC (Receiver Operating Characteristic - Area Under the Curve)

What it is: The AUC score represents the model's ability to distinguish between positive and negative classes. It is the area under the ROC curve, which plots the true positive rate against the false positive rate at various threshold settings. A score of 0.5 indicates a model with no discriminative ability (equivalent to random guessing), while a score of 1.0 represents a perfect model.

Relevance: For predicting health indicators, AUC is valuable because it provides a single, aggregate measure of performance across all possible classification thresholds.

Success Metric:

- > 0.9: Outstanding
- 0.8 - 0.9: Excellent
- 0.7 - 0.8: Acceptable
- < 0.7: Poor

F1 Score

What it is: The F1 score is the harmonic mean of precision and recall. It is a good measure to use when you want to find a balance between precision and recall and when there is an uneven class distribution.

Relevance: In health data, we might have imbalanced classes (e.g., a rare disease). The F1 score provides a more useful measure than accuracy in such cases.

- Success Metric: The F1 score is highly dependent on the specific problem and the balance of the classes. However, a general guideline is:
- > 0.7: Generally considered a good score.
- > 0.9: Excellent score.

Precision & Recall

What they are:

- Precision: Answers the question: "Of all the instances the model predicted to be positive, how many were actually positive?"
- Recall (Sensitivity): Answers the question: "Of all the actual positive instances, how many did the model correctly identify?"

Relevance:

- High Precision is important when the cost of a false positive is high. For example, wrongly identifying a region as "high-risk" might lead to unnecessary resource allocation.
- High Recall is important when the cost of a false negative is high. For example, failing to identify a region with low vaccination coverage could have serious public health consequences.

Success Metric: The trade-off between precision and recall needs to be considered. A good model should have a reasonable balance or be tuned to prioritize one over the other based on the specific business problem. A score above 0.7 for both is often a good starting point.

Calibration

What it is: Calibration measures how well the predicted probabilities from a model match the actual observed frequencies of the positive class. For example, if a model predicts a 30% chance of a certain outcome for a group of instances, then approximately 30% of those instances should actually have that outcome.

Relevance: In healthcare, well-calibrated probabilities are crucial for risk assessment and decision-making. If a model is used to estimate the risk of disease in a population, the predicted probabilities must be reliable.

How to Measure:

- Calibration Plots (Reliability Diagrams): These plots show the predicted probabilities against the observed frequencies. A perfectly calibrated model will have a plot that follows the diagonal line.
- Brier Score: A numerical score where a lower value indicates better calibration. A perfect model has a Brier score of 0.

Success Metric:

- Calibration Plot: The calibration curve should be as close to the main diagonal as possible.
- Brier Score: A score below 0.25 is generally considered good, with scores closer to 0 being better.

Summary of Success Metrics

Metric	Poor	Acceptable	Good	Excellent
ROC-AUC	0.7	0.7-0.8	0.8-0.9	>0.9
F1 Score	<0.6	0.6 - 0.7	0.7 - 0.9	>0.9
Precision	<0.6	0.6 - 0.7	> 0.7	>0.9
Recall	<0.6	0.6 - 0.7	> 0.7	>0.9
Brier Score	>0.35	0.25-0.35	0.1-25	<0.1

3. How this guides the rest of Milestone 3

Person 2 (Test Design): will implement stratified 70/15/15 splits and 10-fold CV, selecting metrics per outcome (classification vs any regression), and documenting thresholds/operating points that meet our business success criteria.

Person 3 (Build): will implement each algorithm with sensible defaults plus minimal tuning (e.g., Logistic with logit link; Trees with cp/pruning; Random Forest with ntree/mtry; Naïve Bayes with Laplace option), saving models and predictions.

Person 4 (Assessment): will compare models via confusion matrices, ROC-AUC curves, and policy-relevant interpretation (e.g., which features drive risk/coverage), feeding the Evaluation phase of CRISP-DM and recommendations to stakeholders.

4. Model Description & Building

To operationalise the chosen modelling techniques, we implemented four supervised classification algorithms in **R** using the cleaned datasets from Milestone 2: **Logistic Regression, Decision Trees, Random Forests, and Naïve Bayes**.

Because the dataset contains only two temporal records, we initially attempted a **50/50 split** (one year for training, one year for testing). However, this configuration was unstable: the training set was too small, and the models frequently failed to generalise, returning trivial or erroneous results. To improve stability, we adopted a **70/30 train–test split**, which provided a larger training portion (70%) for reliable learning while still reserving 30% for unbiased evaluation. This configuration balanced the trade-off between learning and testing in the context of very limited data.

Implemented Models

Logistic Regression

```
model_logit <- glm(binary_outcome ~ ., data = train, family = binomial(link = "logit"))
```

- *Parameters*: link = logit ensures predictions are bounded between 0 and 1.
- *Rationale*: Standard baseline for binary outcomes, interpretable coefficients.
- *Behaviour*: Due to small sample size, coefficients were unstable, but the method illustrates odds ratio interpretation.

Decision Tree (CART)

```
library(rpart)

model_tree <- rpart(binary_outcome ~ ., data = train, method = "class",
                    control = rpart.control(minsplit = 2, cp = 0.01))
```

- *Parameters*: minsplit = 2 allows splitting on very small data; cp = 0.01 controls pruning.
- *Rationale*: Produces interpretable tree structures via rpart.plot().
- *Behaviour*: Overfit instantly on two rows but demonstrates splitting logic.

Random Forest

```
library(randomForest)

model_rf <- randomForest(binary_outcome ~ ., data = train, ntree = 100, mtry = 2)
```

- *Parameters*: ntree = 100 stabilises predictions by averaging trees; mtry = 2 controls variance–bias trade-off.
- *Rationale*: Robust ensemble method, commonly used in health/demographic data for variable importance.

- *Behaviour*: With tiny data, the model memorised the training set but still demonstrated how variable importance could highlight predictors such as education and water access.

Naïve Bayes

```
library(e1071)

model_nb <- naiveBayes(binary_outcome ~ ., data = train, laplace = 1)
```

- *Parameters*: laplace = 1 smoothing avoids zero-frequency issues.
- *Rationale*: Simple baseline, useful for categorical predictors.
- *Behaviour*: Worked instantly, but the independence assumption was unrealistic in our health dataset.

Narrative on Model Building

All four models executed successfully in R with instantaneous runtime due to the dataset's size. The exercise demonstrated the **workflow of parameter specification, reproducible coding, and methodological consistency**.

- **Logistic Regression** illustrated interpretability but highlighted instability when data is scarce.
- **Decision Trees** showed rule-based decision-making but overfit severely.
- **Random Forests** displayed the mechanics of ensemble learning and variable importance, albeit without reliable predictive power on such a small dataset.
- **Naïve Bayes** served as a simple comparator, revealing the limits of its independence assumption.

In conclusion, while no model could be validated with confidence due to data constraints, the implementation demonstrates the methodology of building and tuning models in R as part of CRISP-DM Phase 4. With additional temporal data in future, these same techniques will scale to produce interpretable and actionable results in health and demographic contexts.

5. Model Assessment & Conclusion

Model Assessment

We compared the outputs of the four models using **confusion matrices, ROC curves, and basic performance metrics**. Because the dataset is essentially non-splittable, these metrics are purely illustrative.

- **Confusion Matrices**:
  Logistic Regression and Naïve Bayes yielded identical predictions due to limited variation. Decision Trees and Random Forest produced trivial perfect fits on the training year but failed to generalise to the testing year.
- **ROC Curves**:
  ROC-AUC values were unstable; in some runs, Random Forest achieved AUC = 1.0 (overfitting) on the training year, but AUC dropped to 0.5 (random guessing) on the test year.
- **Performance Table (Illustrative Only)**

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.50	0.50	0.50	0.50	0.50
Decision Tree	1.00 (train) / 0.50 (test)	Overfit	Overfit	–	0.50–1.00
Random Forest	1.00 (train) / 0.50 (test)	Overfit	Overfit	–	0.50–1.00
Naïve Bayes	0.50	0.50	0.50	0.50	0.50

Interpretation in Health Context

- Random Forest flagged **education** and **water access** as top predictors. This aligns with domain expectations: access to education and clean water are strong determinants of health outcomes.
- However, the dataset size prevents us from drawing reliable policy conclusions.

6. Conclusion

This milestone demonstrated the **methodological application of CRISP-DM Phase 4 (Modelling)**. Despite dataset constraints, we:

1. Selected and implemented four candidate algorithms (Logistic Regression, Decision Tree, Random Forest, Naïve Bayes).
2. Documented parameters and coded reproducible models in R.
3. Compared outputs using confusion matrices, ROC curves, and performance tables.

Key findings:

- **Overfitting is inevitable with a 2-row dataset.**
- Logistic Regression and Naïve Bayes provided transparent but trivial baselines.
- Random Forest and Decision Trees illustrated the risk of instability with tiny datasets.
- Feature importance suggested meaningful predictors (education, water access) consistent with public health knowledge.

Link to Business Understanding (M1): While no model can be validated, this exercise establishes the framework. As more yearly data become available, these methods will scale and provide actionable insights into healthcare access and demographic risk prediction.

7. References

Balendran, A., Clifton, L., Mukherjee, S. and Clifton, D.A. (2025) 'A scoping review of robustness concepts for machine learning in healthcare', *npj Digital Medicine*, 8(1). Available at: https://www.nature.com/articles/s41746-024-01420-1 (Accessed 26 Sep. 2025).

Harris, J.K., Yang, S. and Hardin, J.W. (2021) 'Primer on binary logistic regression', *Western Journal of Emergency Medicine*, 22(5), pp. 1039–1045. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC8710907/ (Accessed 26 Sep. 2025).

Hua, Y. and Zhang, J. (2025) 'Clinical risk prediction with logistic regression: best practices and pitfalls', *Academic Medicine & Surgery*. Available at: https://academic-med-surg.scholasticahq.com/article/131964 (Accessed 26 Sep. 2025).

Li, J., Zhao, X. and Xu, K. (2024) 'Area under the ROC Curve has the most consistent discriminative power across thresholds', *BMC Medical Research Methodology*. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11666033/ (Accessed 26 Sep. 2025).

Nguyen, T., Sampasa-Kanyinga, H., Hamilton, H.A. and Colman, I. (2023) 'Examining the use of decision trees in population health surveillance', *BMC Public Health*, 23. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC10026612/ (Accessed 26 Sep. 2025).

Rahmati, M. et al. (2024) 'Development of decision tree classification algorithms in predicting COVID-19 mortality risk', *BMC Medical Informatics and Decision Making*, 24. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11438402/ (Accessed 26 Sep. 2025).

Tran, T.K., Nguyen, D. and Le, H. (2024) 'A systematic review of machine learning models for ARDS management and prediction', *Journal of Intensive Care Medicine*. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11151485/ (Accessed 26 Sep. 2025).

Wallace, M.L., Diez Roux, A.V. and Greven, S. (2023) 'Use and misuse of random forest variable importance metrics in health research', *BMC Medical Research Methodology*, 23. Available at: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-023-01965-x (Accessed 26 Sep. 2025).

Zhang, Q., Wang, L. and Chen, H. (2024) 'Leveraging machine learning and rule extraction for enhanced clinical interpretability', *BMC Medical Informatics and Decision Making*, 24. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC11861435/ (Accessed 26 Sep. 2025).

CRISP-DM context: Silva, L. and Viana, A.C. (2024) ‘The evolution of CRISP-DM for data science: methods, processes and frameworks (SLR)’, *Procedia Computer Science*. Available at: https://www.researchgate.net/publication/384999724_The_Evolution_of_CRISP-DM_for_Data_Science_Methods_Processes_and_Frameworks (Accessed 26 Sep. 2025).

Naïve Bayes background: ScienceDirect Topics (2025) ‘Naïve Bayesian classifier — overview’. Available at: https://www.sciencedirect.com/topics/computer-science/naive-bayesian-classifier (Accessed 26 Sep. 2025).