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 (cleaning, harmonising codes, creating features and splitting data), delivering reproducible R scripts and a data dictionary.

Now, Milestone 3 specifies, justifies and implements **candidate models** appropriate for our (mostly tabular) health survey data, with attention to **interpretability**, **robustness**, and **evaluation design**. 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 in our workflow (iterating between preparation and modelling as data quality/feature needs emerge) (Silva and Viana, 2024).

Given our targets are primarily **classification** (e.g., coverage yes/no, high-risk vs not), and our predictors mix **categorical** and **numeric** variables with potential non-linearities and interactions, we shortlist four algorithms widely used in health analytics: **Logistic Regression**, **Decision Trees (CART/C5.0)**, **Random Forest**, and **Naïve Bayes**. These offer a spectrum from **high interpretability** (logistic, trees) to **strong predictive robustness** (random forest), plus a **simple baseline** (naïve Bayes). That 4 algorithms is commonly recommended for tabular health data benchmarking. Evidence from recent healthcare studies supports these choices and clarifies their trade-offs in accuracy, interpretability and robustness (Balendran et al., 2025; Harris, Yang and Hardin, 2021; Nguyen et al., 2023; Rahmati et al., 2024; Wallace, Diez Roux and Greven, 2023).

2. Modelling Techniques: Rationale & Assumptions

2.1 Rationale narrative (why these four)

- **Logistic Regression (GLM, logit link).**
  Standard for binary clinical outcomes and risk prediction. Provides odds ratios and clear coefficient interpretation. Assumptions are explicit (linearity of log-odds, no perfect multicollinearity, independent observations) and well-documented in

recent clinical reviews. This transparency is valuable for public-health stakeholders (Harris, Yang and Hardin, 2021; Hua and Zhang, 2025).

- **Decision Trees (CART/C5.0).**
  Highly interpretable (human-readable rules/paths), handle mixed types and capture non-linearities & interactions without manual feature engineering. Recent health studies show effective use for population surveillance and COVID-19 mortality risk segmentation (Nguyen et al., 2023; Rahmati et al., 2024).
- **Random Forest.**
  An ensemble of trees (bagging + random feature subspace) that typically improves generalisation and robustness on tabular health data. Resilient to outliers and noise, with strong performance reported in current healthcare applications (e.g., discharge risk, costs). Ongoing work in digital medicine stresses robustness considerations in clinical ML (Tran, Nguyen and Le, 2024; Balendran et al., 2025; Wallace, Diez Roux and Greven, 2023).
- **Naïve Bayes.**
  A lightweight baseline with a simple probabilistic assumption (conditional independence). Despite its naive assumption, it is often surprisingly competitive on high-dimensional tabular data, fast to train, and useful as a calibration point for more complex models (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. Strong baseline for binary outcomes. Easy to communicate to policy teams (Harris, Yang and Hardin, 2021).	Linearity of log-odds; no perfect multicollinearity; independent observations; adequate events per variable. Model misspecification reduces calibration (Hua and Zhang, 2025).	Clinical risk models, coverage yes/no, program targeting; use interactions/splines if needed; check VIFs & calibration (Hua and Zhang, 2025).
Decision Trees (CART/C5.0)	Interpretable rules, handle mixed types, missingness heuristics, capture non-linearities/interactions (Nguyen et al., 2023).	Greedy splits can overfit, unstable to small data changes; control with pruning/minsplit/CP.	Population segmentation (e.g., youth mental health), triage pathways, mortality risk stratification incl. COVID-19 (Rahmati et al., 2024).
Random Forest	Robust predictive performance on tabular health data. Reduces variance via bagging; handles many variables,	Less interpretable than single tree, variable importance must be interpreted carefully, tune `mtry`, `ntree`, class	Discharge/LOS risk, cost prediction, multi-factor risk indices, good default when accuracy is a

	variable importance for signal-finding (Tran, Nguyen and Le, 2024).	balance. Robustness must still be verified across shifts (Wallace, Diez Roux and Greven, 2023).	priority (Balendran et al., 2025).
Naïve Bayes	Fast baseline, performs well when independence holds approximately, good with many categorical features, serves as sanity check (ScienceDirect Topics, 2025).	Conditional independence assumption, often over-confident probabilities and may underperform when features interact strongly.	Screening/triage baselines, text/categorical-heavy data; compare against logistic/tree/forest to judge added value.

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

Because our targets are mostly binary, we will prioritise **ROC-AUC, F1, Precision/Recall, and calibration**. AUC is widely used to compare classifiers across thresholds and has strong support in recent methodology literature (Li, Zhao and Xu, 2024).

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. 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).