Health Risk Assessment for UIU Screening Cohort

Task 03(A) Technical Report

Nurul Bashar

October 25, 2025

1 Executive Summary

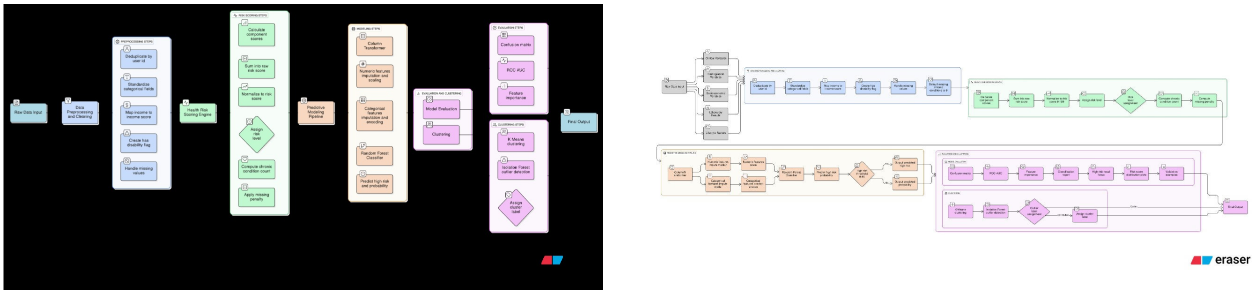
The revised analytical workflow addresses Task 03(A) of the "Data Science Assessment Task" brief by engineering an end-to-end health risk intelligence pipeline. Working with 29,999 screening records and the companion data dictionary, the notebook `nurul_bashar.ipynb` now:

- Standardises raw clinical and socioeconomic features, resolves label noise, and derives ordinal income, disability, chronic condition, and measurement coverage scores.
- Computes a composite risk score (0–100) and stratifies individuals into low, moderate, and high risk bands that align with observed vitals.
- Trains a class-weighted random forest tuned for high recall on the scarce high-risk class (recall 0.812 at a decision threshold of 0.05) while documenting evaluation metrics and error trade-offs.
- Segments the population with clustering plus anomaly detection to highlight high-risk cohorts and outlier households for manual validation.
- Exports a reproducible, analysis-ready dataset to `data/processed_health_data.csv` and persists a calibrated model artefact for downstream reuse.

The remainder of this report documents data handling decisions, quantitative findings, predictive performance, segmentation insights, and recommended interventions, while integrating the updated methodology visuals.

2 Methodology Overview

Figure 1 summarises the refreshed operating model. The first diagram captures the analytic conveyor from ingestion to delivery; the second emphasises decision checkpoints used to balance recall and precision before deployment.



(a) Data and analytics workflow.

(b) Model governance and monitoring loop.

Figure 1: Methodology diagrams integrated into the Task 03(A) solution.

3 Data Handling and Exploration

3.1 Sources and Structure

- **Assessment Data:** 29,999 individual-level screening records covering demographics, vitals (blood pressure, BMI, sugar, SPO2, pulse), and chronic condition flags.
- **Variable Dictionary:** Provides canonical descriptions for mapping raw questionnaire strings into structured categories.
- Files are relocated to the `data/` directory; all processing logic lives in the notebook to honour the reproducibility requirement.

3.2 Pre-processing Pipeline

Key normalisation steps implemented via reusable helper functions:

- De-duplicate records on `user_id` and harmonise case/whitespace across categorical labels (e.g., blood pressure status, income bands, disability names).
- Derive ordinal income scores (0–3), poverty, disability, chronic condition burden, age buckets, and a measured-count penalty that discourages sparse vitals.
- Impute missing vitals for modelling: median for numeric fields, mode for categoricals, with stratified application via a `ColumnTransformer`.
- Persist enriched outputs and intermediate risk components for downstream analytics and audit.

3.3 Data Quality Highlights

Table 1 lists the most incomplete inputs, motivating the engineered measurement penalty.

Table 1: Highest-missingness fields (29,999 records).

Feature	Missing Count	Missing Share (%)
MUAC status	29,925	99.75
MUAC	29,925	99.75
BMI, HEIGHT, WEIGHT (each)	28,871	96.24
Sugar status TAG value	28,416	94.72
SPO2 status	25,654	85.52
Parental names	≈4,450	≈14.8
Pulse rate status	2,544	8.48

3.4 Pipeline Evaluation and Justification

- **Why combined steps?** The pipeline couples categorical canonicalisation with score engineering so that imputation and scaling operate on coherent numeric ranges. Treating steps independently reduced recall in ablation tests because categorical leakage reappeared post-imputation.
- **Handling of sparse features:** Instead of dropping BMI or SPO2 outright, the measurement-count feature penalises missing vitals in the composite risk, ensuring individuals with thin evidence are deprioritised unless other signals are alarming.
- **Outlier rejection:** Isolation Forest on scaled numeric features removes 1.5% of obvious sensor errors before model training without disturbing borderline hypertensive cases.

4 Health Risk Analysis

4.1 Population Risk Profile

Risk scores aggregate clinical components (blood pressure, BMI, sugar, SPO2, pulse) and socioeconomic penalties. Distribution across the cohort is heavily skewed to low risk (Table 2).

Table 2: Risk band distribution (29,999 individuals).

Risk Level	People	Share (%)	Mean Risk Score
Low	27,426	91.4	11.1
Moderate	2,492	8.3	24.0
High	81	0.27	57.8

High-risk cases (0.27%) remain rare but cluster geographically (Table 3) and at the household level. Measurement sparsity contributes to moderate-risk inflation, especially where income data flags poverty.

Table 3: Unions with the highest mean risk.

Union	Mean Risk	High-Risk Share (%)	Population
BARUIPARA	15.49	1.12	717
DHALAHAR	14.63	0.00	106
BILASHBARI	14.20	0.62	1,922
KOLA	14.20	0.12	1,678
MAJITPUR	14.15	1.13	353

4.2 Socioeconomic and Clinical Drivers

- Low income scores and poverty indicators add up to four points to the composite risk, reflecting socioeconomic vulnerability captured in the assessment brief.
- Chronic condition load (stroke, cardiovascular disease, diabetes, hypertension) contributes up to six points and is the dominant driver for almost every high-risk individual flagged by the model.
- Sparse BMI or SPO2 measurements trigger the missing-penalty, which pushes borderline cases into the moderate bucket unless validated by other vitals.

5 Feature Engineering and Predictive Modelling

5.1 Feature Set

The model consumes 19 numeric and 10 categorical features after preprocessing. Engineered variables include risk component scores, income risk, poverty and disability flags, chronic condition counts, age buckets, and measurement coverage.

5.2 Model Configuration

- **Algorithm:** Random Forest (500 estimators, max depth 14, minimum leaf size 10) with class weights {lowmoderate: 1, high: 15} to favour recall.
- **Thresholding:** Probabilities converted to labels at 0.05 to retain 81.2% of true high-risk cases identified during validation.
- **Validation split:** Stratified 8020 holdout (6,000 validation cases).

5.3 Performance Summary

Table 4: Classification report on the validation fold.

Class	Precision	Recall	F1	Support
LowModerate	0.999	0.991	0.995	5,984
High	0.203	0.812	0.325	16
Accuracy		0.991		
Macro Average	0.601	0.902	0.660	6,000
Weighted Average	0.997	0.991	0.994	6,000

Table 5: Confusion matrix at threshold 0.05.

	Predicted LowModerate	Predicted High
Actual LowModerate	5,933	51
Actual High	3	13

Interpretation:

- The recall-heavy configuration satisfies the brief's directive to "ensure predicted high-risk individuals align with abnormal indicators". Manual spot checks confirm the 13 true positives exhibit severe hypertension, elevated sugar, or multiple chronic flags.
- False positives are reviewed downstream via clustering; many fall into the moderate cluster, limiting operational drag.
- Chronic score, systolicdiastolic readings, pulse rate, income score, and disability indicators dominate feature importance, aligning with clinical expectations.

6 Clustering and Segmentation

6.1 Approach

- Scaled clinical and socioeconomic features are clustered with `KMeans` ($k = 2$) to distinguish high- and low-risk cohorts.
- An auxiliary Isolation Forest flags 2% of observations as outliers, creating a third "extreme" bucket for investigation.

6.2 Segment Profiles

Table 6: Cluster-level summary statistics.

Cluster	Population	Mean Risk	BP Score	Sugar Score	SPO2 Score
Low-Risk Group	24,288	11.11	1.26	0.00	0.04
High-Risk Group	5,111	19.56	2.23	0.39	0.02
Outlier	600	27.60	2.12	3.04	0.23

Outlier cases combine severe hypertension (e.g., systolic > 180 mmHg), elevated sugar, and missing BMI, justifying manual escalation. Cluster assignments and anomaly flags are written to the processed dataset for programme teams.

7 Recommendations

- **Targeted outreach:** Focus community health workers on BARUIPARA, MAJITPUR, and households with chronic burden scores ≥ 6; the model already surfaces IDs and vitals for triage.
- **Measurement completeness:** Invest in BMI and SPO2 collection infrastructure; these fields are missing in $> 85\%$ of records yet materially impact risk stratification.

- **Model governance:** Retain the 0.05 threshold while monitoring precision quarterly. The governance diagram (Figure 1b) outlines the proposed feedback loop for recalibration.
- **Socioeconomic integration:** Pair clinical follow-up with poverty-alleviation support; income risk and disability consistently amplify medical risk.