Health Risk Assessment for UIU Project

Nurul Bashar

October 22, 2025

Executive Summary

An exploratory and modelling workflow was executed inside the Jupyter notebook `nurul_bashar.ipynb`. The goal was to understand health risks captured in the screening dataset, engineer interpretable features, train a high-recall high-risk classifier, and segment the population. All data preparation and analytics code now lives inside the notebook; no standalone Python scripts are required.

1 Data Preparation

- Raw files (`test-dataset.xlsx - test data.csv` and the variable description) were relocated to `data/`.
- The notebook now loads data via reusable helpers that normalise income labels, canonicalise test result strings, harmonise categorical encodings, and engineer ordinal scores (income, disabilities, chronic conditions, measurement coverage).
- The same helpers compute a composite risk score (0–100) using blood pressure, BMI, blood sugar, SPO_2, pulse rate, chronic conditions, and socioeconomic penalties. Risk levels are classified as Low (< 20), Moderate (20–35), and High (≥ 35).
- The fully enriched dataset (including cluster labels and model scores) is saved to `data/processed_health_dat`

2 Health Risk Analysis

Population-level risk is highly skewed toward low scores, yet targeted hotspots emerge:

Table 1: Risk Level Distribution (n = 29,999)

Risk Level	People	Share
Low	27,426	91.4%
Moderate	2,492	8.3%
High	81	0.27%

Table 2: Highest-Risk Households

Household ID	Members	Mean Risk	High-Risk Share
169180	1	47.0	100%
167363	1	47.0	100%
195158	1	45.3	100%
228573	1	42.0	100%
177040	1	40.3	100%

Table 3: Top Unions by 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

Socioeconomic drivers remain evident: lower income classes and poverty flags correlate with elevated risk scores, and measurement sparsity increases the "missing penalty" component.

3 Predictive Modelling

- A class-weighted random forest (500 trees, max depth 14, threshold 0.05) delivers 0.812 recall on the rare high-risk class while preserving 0.203 precision.
- Aggregate metrics (validation fold, 6,000 rows):

```
              precision    recall  f1-score   support

Low/Moderate      0.999     0.991     0.995      5984
High              0.203     0.812     0.325        16

accuracy                            0.991      6000
macro avg      0.601     0.902     0.660      6000
weighted avg   0.997     0.991     0.994      6000
```

Chronic condition burden, systolic/diastolic readings, pulse rate, income score, and disability indicators dominate feature importance, aligning model logic with clinical intuition.

4 Clustering and Segmentation

- KMeans (k=2) paired with an isolation forest surfaces three actionable cohorts:
  - High-Risk Group: mean risk 19.6 with hypertensive profiles and chronic comorbidities.
  - Low-Risk Group: mean risk 11.1, consistently normal vitals.

  - Outliers: mean risk 27.6, sparse but extreme observations (e.g., severe hypertension or hypoxic SPO_2).

- Cluster labels and anomaly flags are stored in `data/processed_health_data.csv` for downstream targeting.

5 Key Insights & Recommendations

- Less than 1% of the population is categorised as high-risk, yet those individuals cluster in specific unions and households, enabling focused outreach.
- Blood pressure remains the most complete and discriminative vital; investment in expanding BMI/SPO_2 coverage will improve early detection of metabolic and respiratory issues.
- Economic vulnerability (poverty status, lower income class) magnifies risk, implying any clinical programme should be paired with social protection measures.
- The calibrated forest errs on the side of recall—appropriate for triaging households for follow-up—while the segmentation step filters likely false positives.

Reproducibility Notes

- Activate the project environment (`.venv`) and run the notebook in order: this regenerates all analysis outputs and the processed dataset.