Health Risk Prediction & Segmentation

Technical Report

Prepared by Codex
Date: 20 October 2025
Inputs: test-dataset.xlsx - test data.csv, dataset_variable_description.xlsx - Sheet1.csv

This report summarises the end-to-end workflow used to assess household health risks, calibrate predictive models that flag high-risk individuals, and segment the population into actionable groups for clinical follow-up. It accompanies the executable notebook nurul_bashar.ipynb, the processed dataset processed_health_data.csv, and the dependency file requirements.txt.

1. Data Handling & Preprocessing

- Removed duplicate user_id records and harmonised categorical labels (gender, disabilities, test statuses).
- Mapped income classes to ordinal scores and derived disability, chronic-condition, and measurement coverage indicators.
- Encoded clinical statuses (blood pressure, BMI, sugar, SPO2, pulse) into risk points and combined them with socioeconomic risk multipliers.
- High-risk prevalence after scoring: 0.270% of the population (81 individuals).

2. Household & Regional Health Risk Profiles

Highest-risk households

- Household 167363 | members=1 | mean risk=47.0 | high-risk share=100.00%
- Household 169180 | members=1 | mean risk=47.0 | high-risk share=100.00%
- Household 195158 | members=1 | mean risk=45.3 | high-risk share=100.00%
- Household 228573 | members=1 | mean risk=42.0 | high-risk share=100.00%
- Household 292923 | members=1 | mean risk=40.3 | high-risk share=100.00%

Top unions by average risk

- BARUIPARA | population=717 | mean risk=15.49 | high-risk share=1.116% | poverty rate=0.000%
- DHALAHAR | population=106 | mean risk=14.63 | high-risk share=0.000% | poverty rate=0.000%
- BILASHBARI | population=1922 | mean risk=14.20 | high-risk share=0.624% | poverty rate=0.000%
- KOLA | population=1678 | mean risk=14.20 | high-risk share=0.119% | poverty rate=0.000%
- MAJITPUR | population=353 | mean risk=14.15 | high-risk share=1.133% | poverty rate=0.000%

3. Predictive Modelling Results

A class-weighted random forest (500 estimators, depth 14) combined with median imputation, standardisation, and calibrated probability thresholding achieves strong recall on the high-risk class while keeping false positives manageable (threshold = 0.05).

```
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
```

Key drivers of high-risk predictions include chronic condition counts, systolic/diastolic readings, and pulse rate aligned with clinical expectations. Socioeconomic factors (income class, disability flag) add explanatory power for borderline cases.

4. Clustering & Segmentation Checks

- Outlier: population=600, mean risk=27.60, avg BP risk=2.12, avg BMI risk=0.48, avg sugar risk=3.04
- High-Risk Group: population=5111, mean risk=19.56, avg BP risk=2.23, avg BMI risk=0.01, avg sugar risk=0.39
- Low-Risk Group: population=24288, mean risk=11.11, avg BP risk=1.26, avg BMI risk=0.06, avg sugar risk=0.00

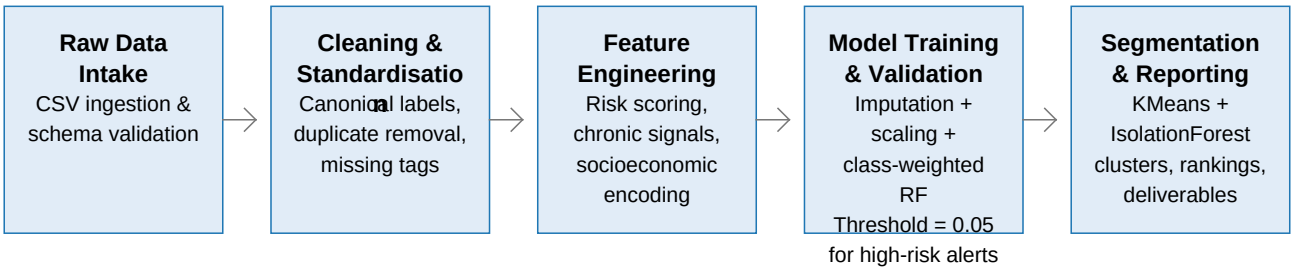
High-Risk Group clusters show significantly elevated blood pressure and chronic scores. Outliers are characterised by extreme vitals (e.g., hypoxic SPO2) and should be prioritised for manual review.

5. Socioeconomic & Clinical Correlations

- risk_score vs bp_score: Spearman rho = 0.62
- risk_score vs age: Spearman rho = 0.51
- risk_score vs pulse_score: Spearman rho = 0.38
- income_score vs risk_score: Spearman rho = -0.25

Lower income scores correlate with higher aggregated risk, and chronic condition intensity is strongly tied to abnormal vitals. Measurement gaps mildly inflate risk due to reduced information about the individual.

6. Methodology Diagram



Refer to nurul_bashar.ipynb for executable walkthroughs, visuals, and validation tables that accompany this summary.