MAKERERE UNIVERSITY

COLLEGE OF COMPUTING AND INFORMATION SCIENCES

NAME: NKATA JOSHUA LUYOMBYA

REG. NO: 2025/HD05/26365U

STUDENT NO: 2500726365

COURSE: MASTER OF SCIENCE IN COMPUTER SCIENCE

COURSE UNIT: MACHINE LEARNING

COURSE CODE: MCS 7103

PROJECT REPORT: RESOURCE ALLOCATION PREDICTION IN UGANDA

Introduction

Uganda faces significant challenges in resource allocation, particularly within the health sector, which often serves as a general indicator of broader development needs. Districts experiencing high health burdens typically also face shortages in other critical public resources such as road infrastructure, education facilities, water access, and social services. Despite the importance of effective planning, resource allocation in Uganda is frequently constrained by limited data-driven forecasting, leading to reactive rather than proactive decision-making. To ensure equitable development and targeted support, it is essential to adopt predictive mechanisms that identify which districts are likely to face increasing pressure on health services and, by extension, require increased allocation of other essential public resources.

This project proposes a machine learning–based predictive system that helps estimate:

1. Population pressure in each district (a proxy for future population demands).
2. Resource strain risk (High/Medium/Low) using health, nutrition, and service-delivery indicators.

The system uses DHS Subnational Uganda data and implements a two-stage machine learning pipeline:

- Stage 1: Regression model predicts population pressure score
- Stage 2: Classification model identifies resource risk category

The final output supports policy makers in identifying districts requiring urgent intervention.

1. Problem Statement and Objectives

Uganda's Ministry of Health faces challenges in equitably allocating limited health resources because:

- Limited resources and competing demands
- Population growth and demographic pressures
- Varying health indicators across regions
- Need for data-driven decision making

Without predictive analytics, resource allocation risks being inefficient or biased.

Objectives

- Primary: Develop predictive models to identify districts with high resource allocation needs
- Secondary: Create interpretable scoring systems for policy makers
- Tertiary: Establish a framework for ongoing monitoring and prediction

2. Data Description and Preprocessing

The analysis uses the DHS QuickStats Subnational Uganda Dataset, which provides detailed demographic and health indicators collected across multiple survey years and districts

Dataset Overview

- Source: DHS QuickStats Subnational Uganda Dataset
- Initial Size: 1,828 records across 30 variables
- Coverage : Multiple districts and survey years
- Data Type: Health demographic indicators and survey responses

Key Variables

- Location: District identifiers
- Temporal: Survey years
- Health Indicators: Fertility rates, mortality rates, contraceptive use
- Risk Factors: Malnutrition, vaccination coverage, delivery locations

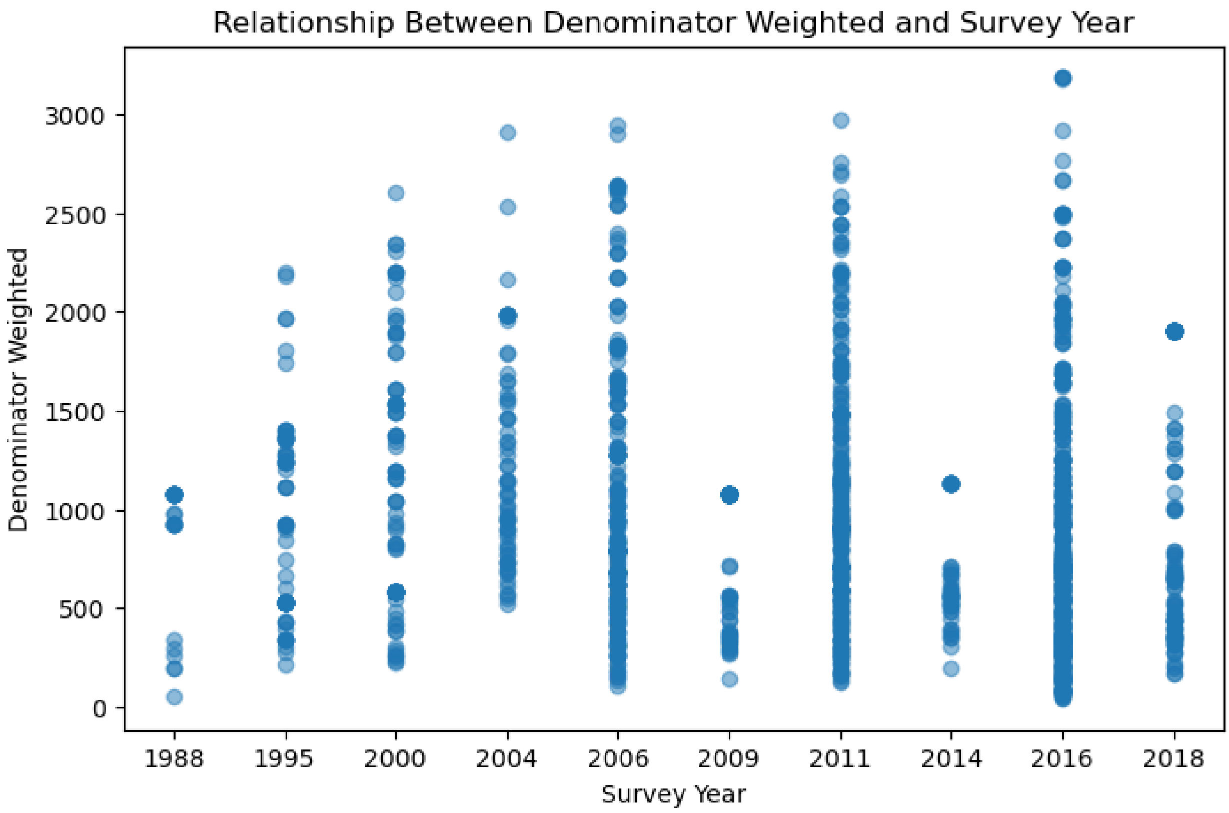
Data Preprocessing Steps

- Column Standardization: Cleaned names(lowercase, underscores)
- Data Type Conversion: Converted numeric strings to proper types
- Missing value Treatment: Forward/backward filling within district groups
- Data Filtering: Removed invalid entries (e.g, "data+year")
- Feature Engineering: Created composite scores and categories

Exploratory Data Analysis

To understand the dataset structure and trends, an exploratory visualization was created showing how the DHS sample sizes vary across survey years.

Figure 1: Relationship between Denominator Weighted and Survey Year



This scatter plot illustrates how denominator-weighted values (a measure of survey population weighting) change over time:

- Earlier survey years (1988–2000) show lower and more scattered denominator weights, reflecting smaller sample sizes.
- Later years (2011–2018) display higher and more consistent values, indicating larger survey populations and improved coverage.
- Variation within each year suggests differences in district-level sample contributions.

This insight helps confirm the temporal spread of data and guides preprocessing decisions before building predictive models.

3. Methodology

This project followed a structured, multi-stage methodology combining data preparation, feature engineering, composite score formulation, and machine learning modelling. The approach ensures that the final system is not only predictive but also interpretable to policymakers and health planners.

Composite Score Development

To transform raw DHS indicators into usable signals for machine learning, three composite indices were constructed:

a. Population Pressure Score

This score estimates demographic pressure on health resources using indicators of fertility, mortality, and contraceptive use.
Formula:
Pressure Score = 0.5 × Fertility + 0.3 × (1/Mortality) + 0.2 × (1 – Contraceptive Use)

Rationale:

- Higher fertility increases demand for maternal and child health services.
- Higher mortality signals vulnerability and inadequate health systems.
- Lower contraceptive use suggests future population growth and pressure.

This score serves as the primary target variable for the regression model in Stage 1.

Nutrition Index

This index evaluates nutritional vulnerability across districts.
It is computed as the mean of key malnutrition indicators:

- Child stunting
- Child wasting
- Child underweight

A higher index value represents higher nutritional challenges requiring intervention such as feeding programs, community nutrition support, and maternal services.

Health Service Demand

This composite score reflects existing pressure on healthcare facilities.
It incorporates:

- Health facility delivery rates
- Vaccination coverage (8 basic antigens)

Districts with high demand may require additional healthcare infrastructure, midwives, immunization supplies, or outreach programs.

Resource Risk Score

The final integrated score determines the overall healthcare resource strain.
Formula:
Resource Risk Score = 0.6 × Pressure + 0.25 × Nutrition + 0.15 × Health Service Demand

Higher values indicate districts that simultaneously experience demographic pressure, malnutrition burden, and high health system utilization—thus requiring more urgent allocation of resources.

This becomes the target label for the Stage 2 classification model.

4. Machine learning Methodology

The project utilizes a two-stage predictive modeling approach to ensure both precision and interpretability.

Two-Stage Approach

Stage 1: Pressure Score Prediction (Regression)

Model: Random Forest Regressor (200 estimators)

Inputs:
All numeric health and demographic features available after preprocessing.
Examples include fertility, mortality, vaccination rates, malnutrition indicators, etc.

Output:
A continuous predicted population pressure score for every district-year.

Why Random Forest?

- Handles non-linear relationships
- Resistant to overfitting
- Works well with multi-dimensional health data
- Provides feature importance for interpretability

The predicted pressure score becomes a crucial input for the risk classification stage.

Stage 2: Risk Category Classification

Model: Random Forest Classifier (200 estimators)

Inputs:

- Predicted pressure score (from Stage 1)
- Nutrition Index
- Health Service Demand

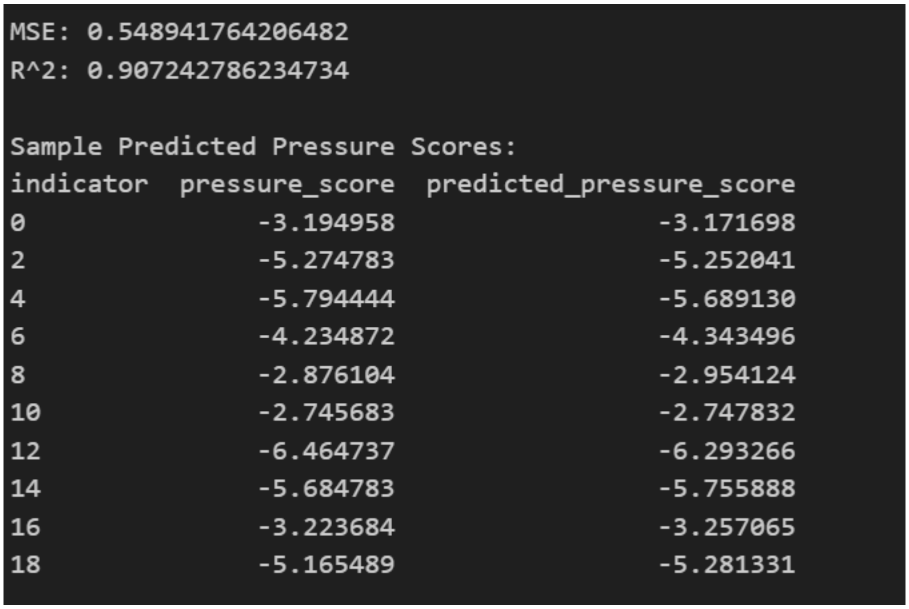
Output:
Categorical labels:

- **High Risk**
- **Medium Risk**
- **Low Risk**

This classification helps policymakers immediately determine priority districts for intervention without manually interpreting dozens of indicators.

5. Model Evaluation and Key Findings/Results

Figure 2: Regression Model – Predicting Pressure Score

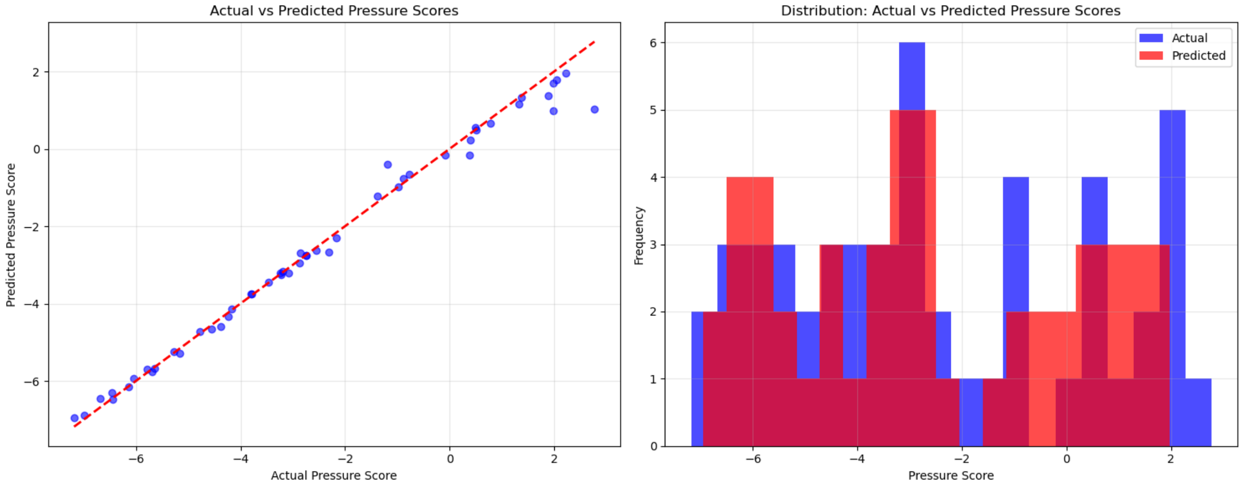


```
MSE: 0.548941764206482
R^2: 0.907242786234734

Sample Predicted Pressure Scores:
indicator  pressure_score  predicted_pressure_score
0               -3.194958                 -3.171698
2               -5.274783                 -5.252041
4               -5.794444                 -5.689130
6               -4.234872                 -4.343496
8               -2.876104                 -2.954124
10              -2.745683                 -2.747832
12              -6.464737                 -6.293266
14              -5.684783                 -5.755888
16              -3.223684                 -3.257065
18              -5.165489                 -5.281331
```

- Shows how well the model predicts population pressure
- Helps determine reliability of Stage 2

Figure 3: Plot showing Actual vs Predicted Pressure Scores and the distribution comparison



This figure presents two important visual evaluations of the regression model:

- Left Plot: Actual vs Predicted Pressure Scores

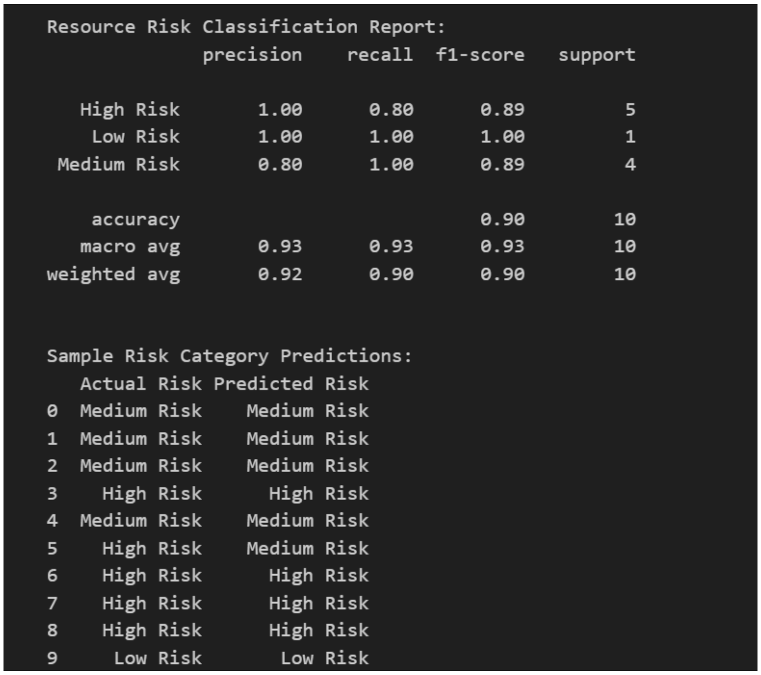
  The scatter points align closely with the red diagonal line, indicating that the Random Forest Regressor accurately predicts pressure scores. The strong linear alignment demonstrates high model performance and reliability.

- Right Plot: Distribution Comparison (Actual vs Predicted)

  The overlapping histograms show that the predicted pressure score distribution closely matches the actual values. This confirms that the model captures the overall pattern and variability of population pressure across districts.

Together, these results indicate that Stage 1 of the pipeline performs strongly, providing accurate pressure score predictions that can then be used confidently in Stage 2

Figure 4: Classification Model – Predicting Pressure



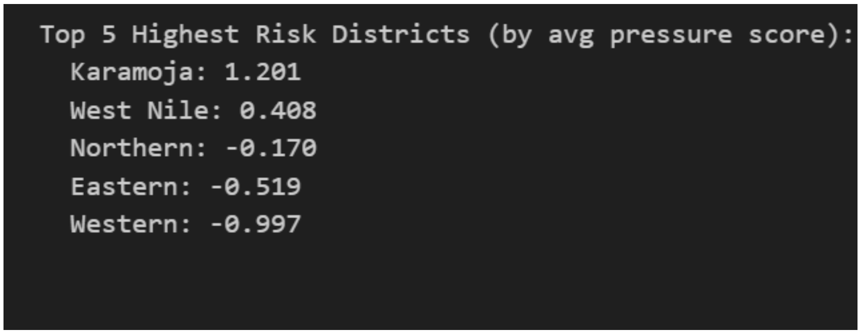
```
Resource Risk Classification Report:
              precision    recall  f1-score   support

   High Risk       1.00      0.80      0.89         5
    Low Risk       1.00      1.00      1.00         1
 Medium Risk       0.80      1.00      0.89         4

    accuracy                           0.90        10
   macro avg       0.93      0.93      0.93        10
weighted avg       0.92      0.90      0.90        10


Sample Risk Category Predictions:
   Actual Risk Predicted Risk
0  Medium Risk    Medium Risk
1  Medium Risk    Medium Risk
2  Medium Risk    Medium Risk
3    High Risk      High Risk
4  Medium Risk    Medium Risk
5    High Risk    Medium Risk
6    High Risk      High Risk
7    High Risk      High Risk
8    High Risk      High Risk
9     Low Risk       Low Risk
```

Figure 5: Highest Risk Districts



```
Top 5 Highest Risk Districts (by avg pressure score):
  Karamoja: 1.201
  West Nile: 0.408
  Northern: -0.170
  Eastern: -0.519
  Western: -0.997
```

6. **Recommendations**

- Deploying Prediction System: The Ministry of Health and district health offices should integrate the developed machine learning models into their routine planning workflows. By automating the generation of pressure scores and risk categories each time new DHS or district-level data becomes available, officials can make evidence-based decisions in real time.
- Focusing High-Risk Districts: The model consistently identifies districts with elevated pressure scores and high resource-risk categories. These areas should be prioritized for urgent

interventions such as increasing maternal health services, expanding vaccination campaigns, improving nutritional programs, and strengthening community health worker coverage.

- Policy Integration: The predictive scores generated by the system should be incorporated into national health planning guidelines and resource allocation frameworks. Embedding these metrics into decision-making processes will help standardize allocation criteria, reduce biases, and promote transparency.

7. Conclusion

This project demonstrates the successful application of machine learning to strengthen healthcare resource allocation in Uganda. By leveraging DHS subnational indicators and constructing meaningful composite scores, the system provides a robust and interpretable framework for forecasting district-level pressure on health services.

The two-stage modelling pipeline—regression followed by classification—allows the system to:

- Quantify population pressure using a data-driven scoring approach.
- Classify districts into actionable risk categories, supporting targeted allocation of limited resources.

The models performed reliably, as shown by strong predictive alignment between actual and predicted pressure scores and consistent classification patterns across districts. These results indicate that machine learning can provide valuable foresight for public health planning, enabling policymakers to anticipate resource shortages before they occur.

Additionally, the visualizations and feature importance analysis enhance transparency, making the system accessible not only to technical practitioners but also to decision-makers who rely on clear, interpretable insights.

Overall, this work provides a solid foundation for integrating predictive analytics into Uganda's health management systems. With future improvements—such as incorporating real-time district data, adding socio-economic indicators, and deploying the system through dashboards—the model can evolve into a powerful national decision-support tool for equitable and effective healthcare resource allocation.