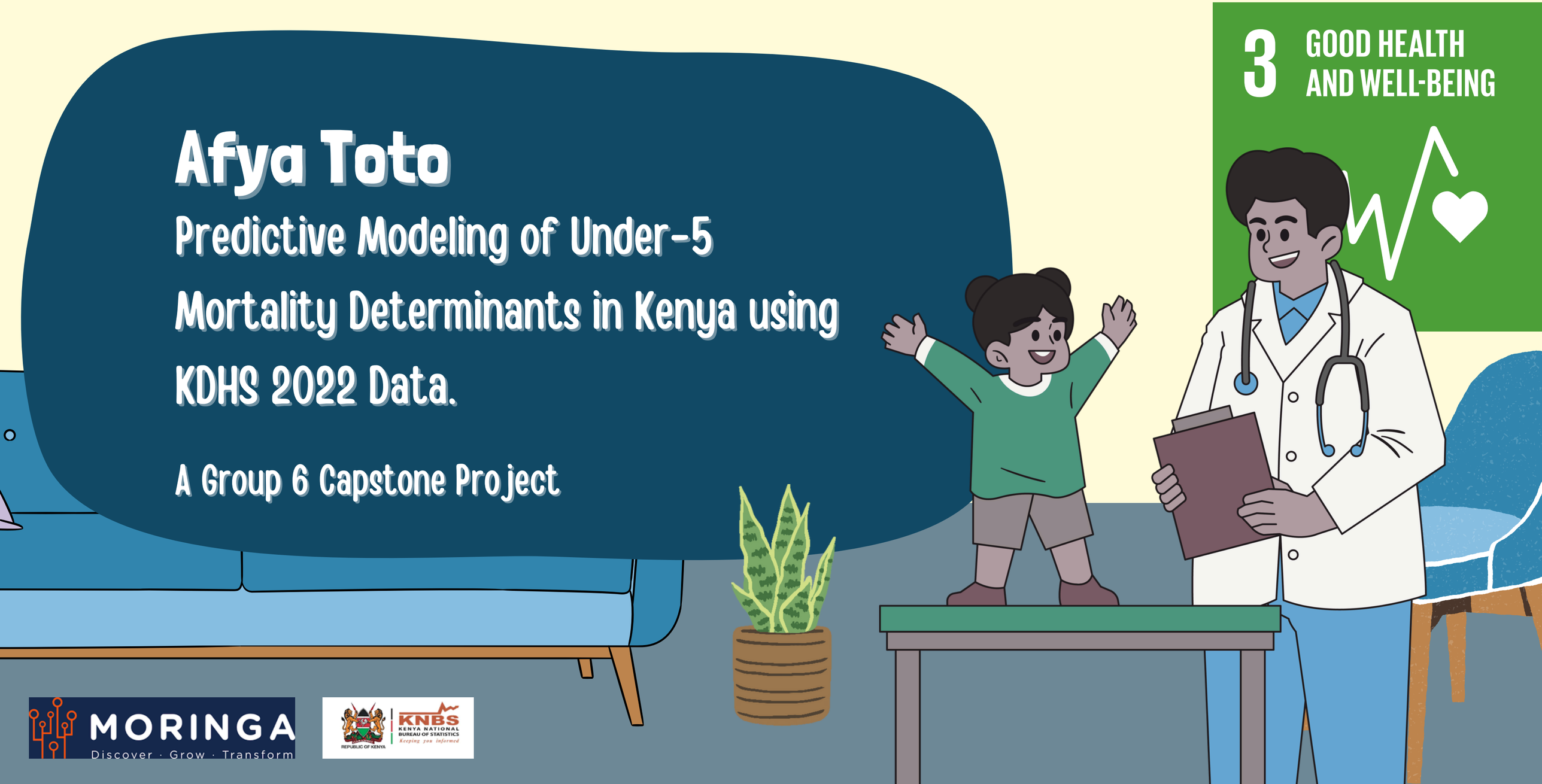
Afya Toto

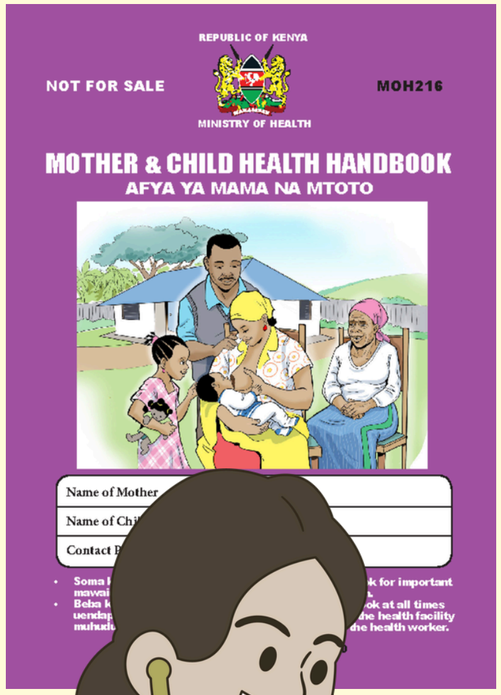
Predictive Modeling of Under-5 Mortality Determinants in Kenya using KDHS 2022 Data.

A Group 6 Capstone Project



MORINGA
Discover · Grow · Transform

REPUBLIC OF KENYA
KNBS
KENYA NATIONAL BUREAU OF STATISTICS
Keeping you informed



Our Team

Charity Mwangangi

Keith Tongi

Edgar Muturi

Jacob Abuon

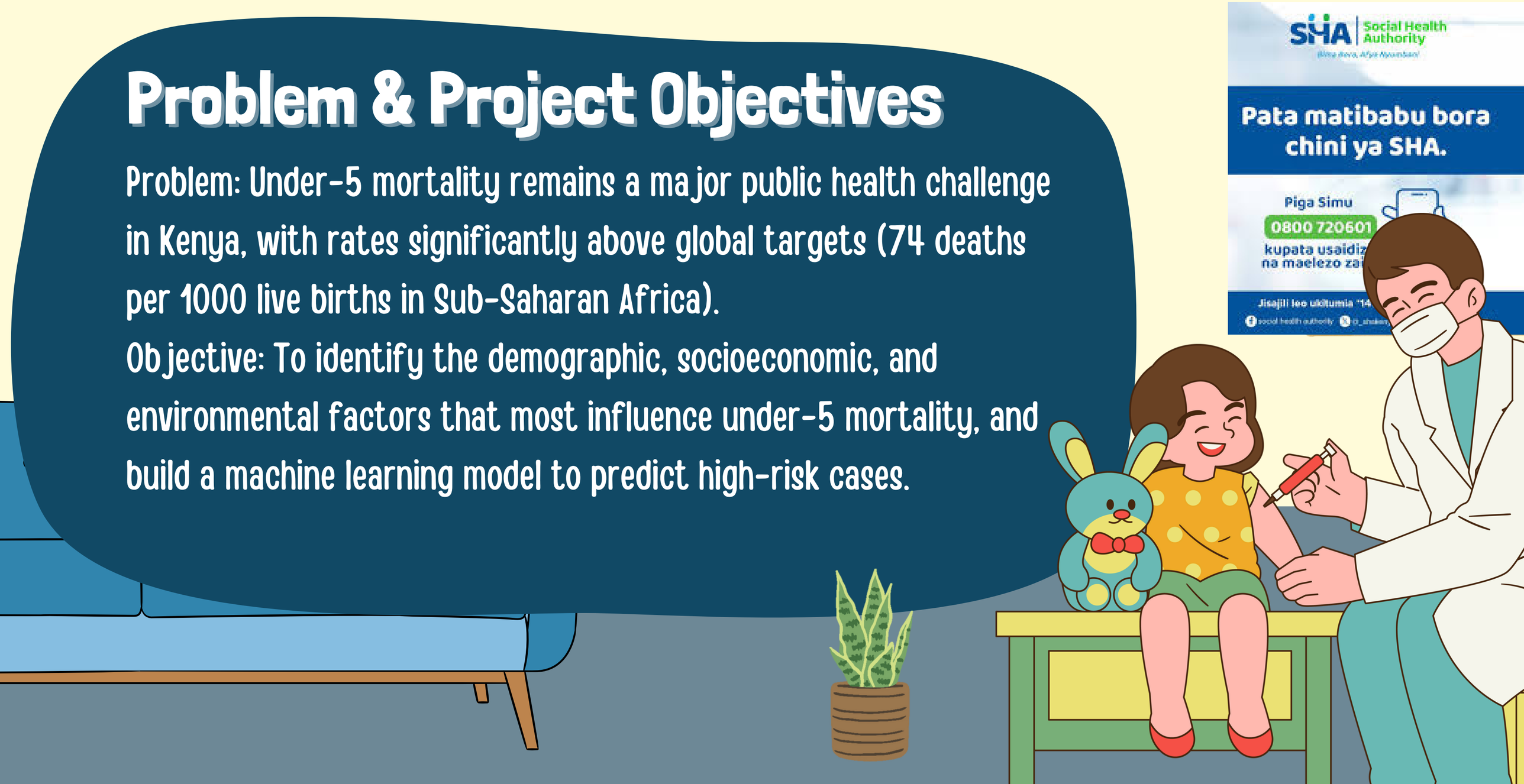
Edna Maina

Kevin Karanja

Problem & Project Objectives

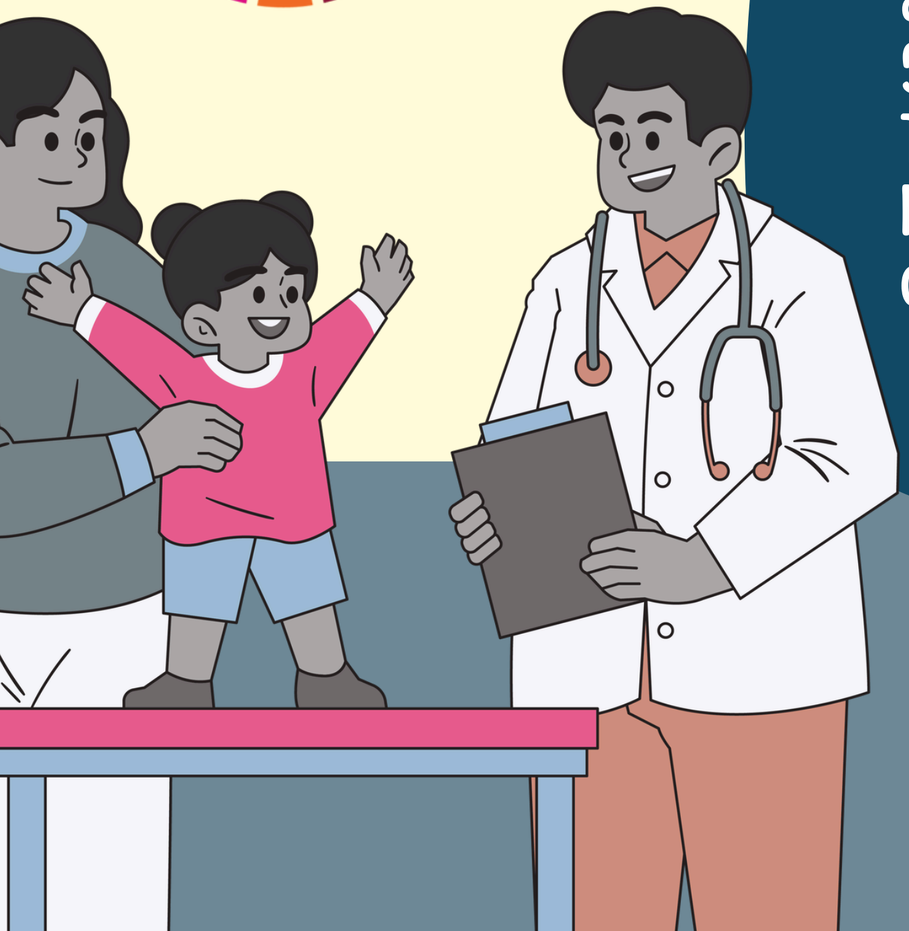
Problem: Under-5 mortality remains a major public health challenge in Kenya, with rates significantly above global targets (74 deaths per 1000 live births in Sub-Saharan Africa).

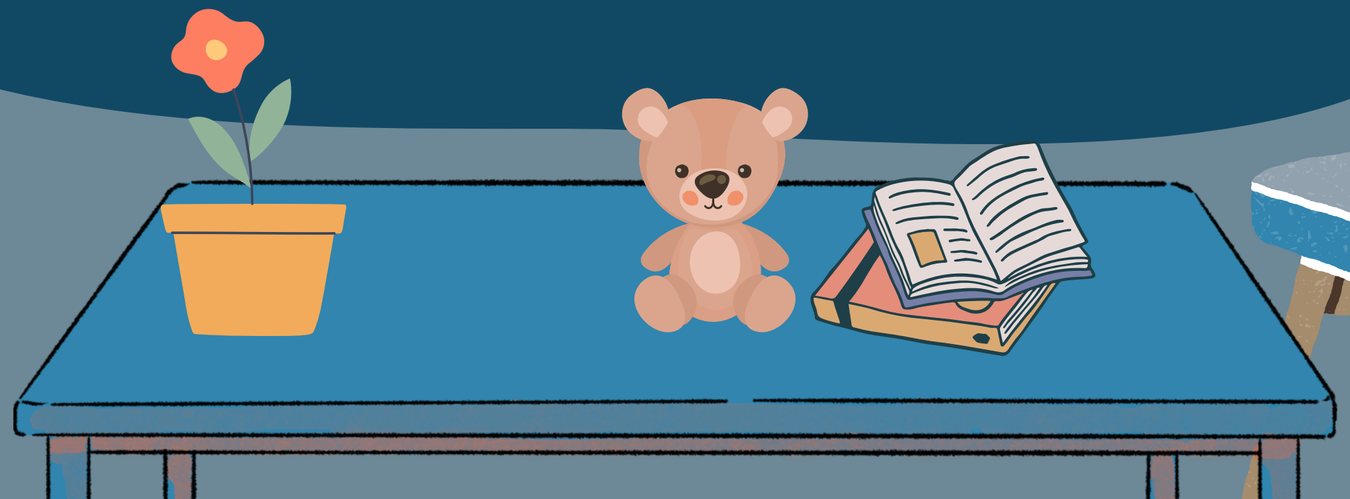
Objective: To identify the demographic, socioeconomic, and environmental factors that most influence under-5 mortality, and build a machine learning model to predict high-risk cases.



Implications & Target Metric

Value: The model can inform policymakers and public health agencies on where to allocate resources to achieve SDG target 3.2 (reducing child mortality).
Target Metric: Since mortality is a rare event, minimizing False Negatives (missed high-risk cases) is crucial. We will primarily optimize the F2 Score and Precision-Recall AUC (PR-AUC).





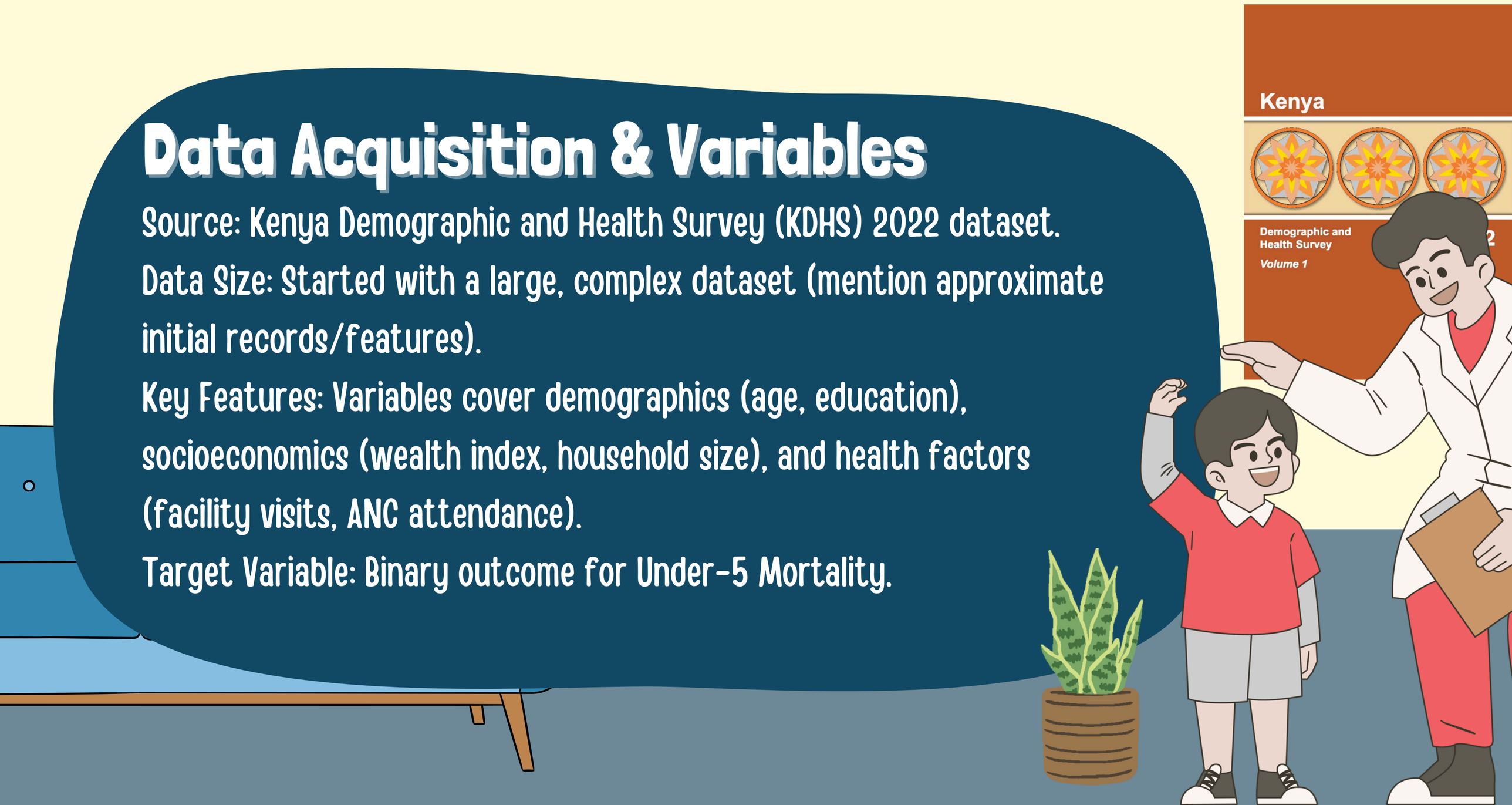
Data Acquisition & Variables

Source: Kenya Demographic and Health Survey (KDHS) 2022 dataset.

Data Size: Started with a large, complex dataset (mention approximate initial records/features).

Key Features: Variables cover demographics (age, education), socioeconomics (wealth index, household size), and health factors (facility visits, ANC attendance).

Target Variable: Binary outcome for Under-5 Mortality.



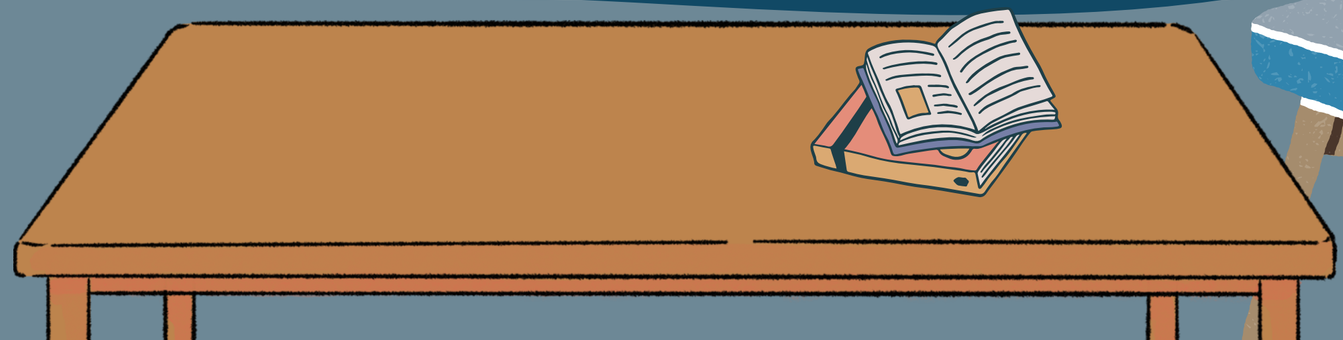
Data Cleaning & Feature Engineering

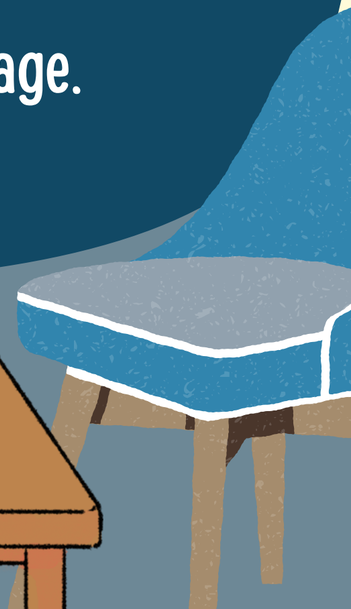
Cleaning:

- Handled a large number of missing values using appropriate imputation or by dropping features with >20% missingness.
- Converted raw categorical codes into meaningful binary or one-hot encoded variables.

Feature Engineering:

- Created new aggregate features that combine raw variables (e.g., a Health Access Index if applicable).
- Ensured all features are predictive and free of data leakage.





Target & Data Imbalance

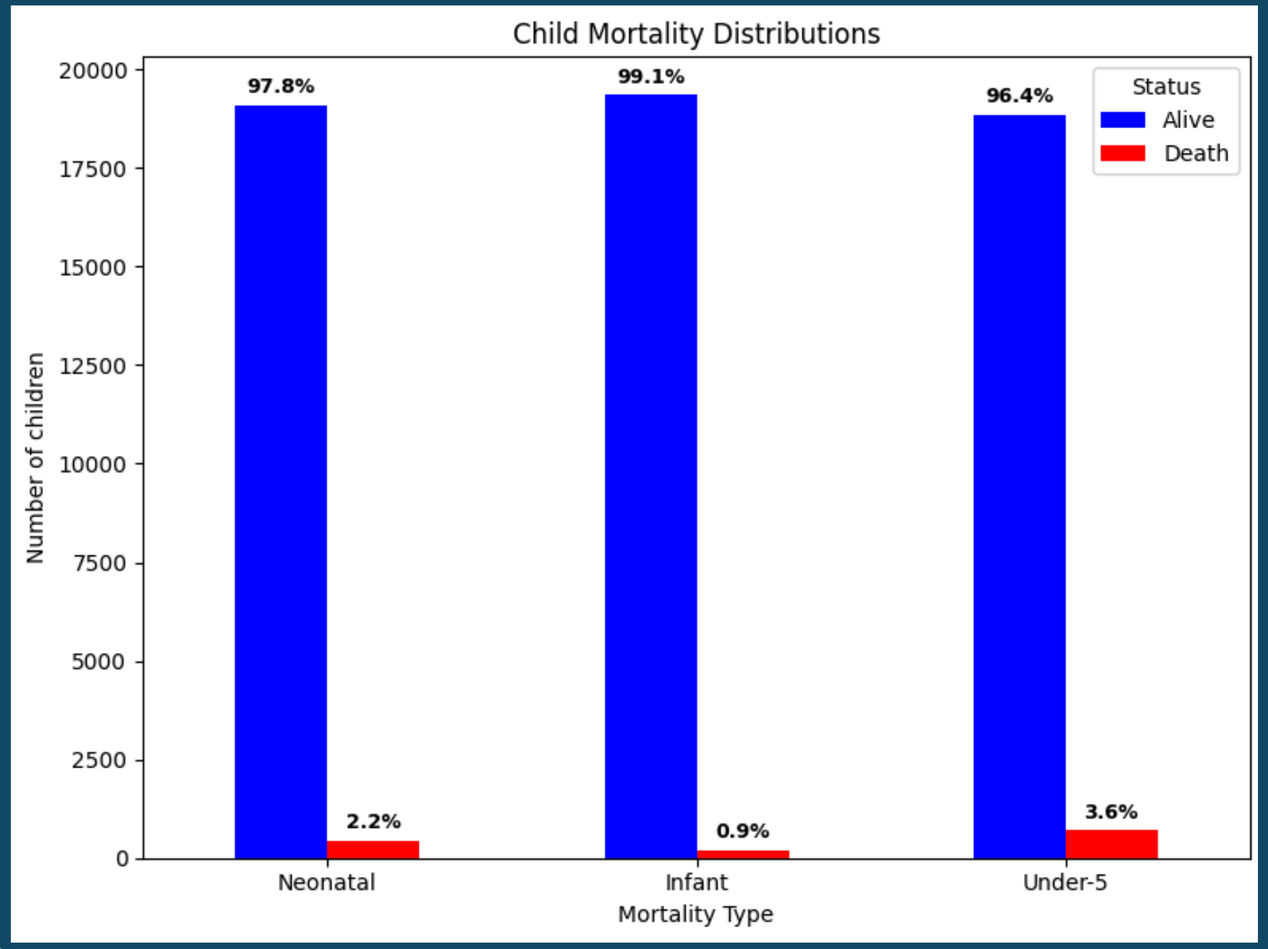


Chart: Bar plot showing the distribution of the Under-5 Mortality target variable.

- This confirms the significant class imbalance, which justifies the use of PR-AUC and F2 Score. ·
- We also visually establish that mortality is a rare event within the dataset.

Healthcare Access

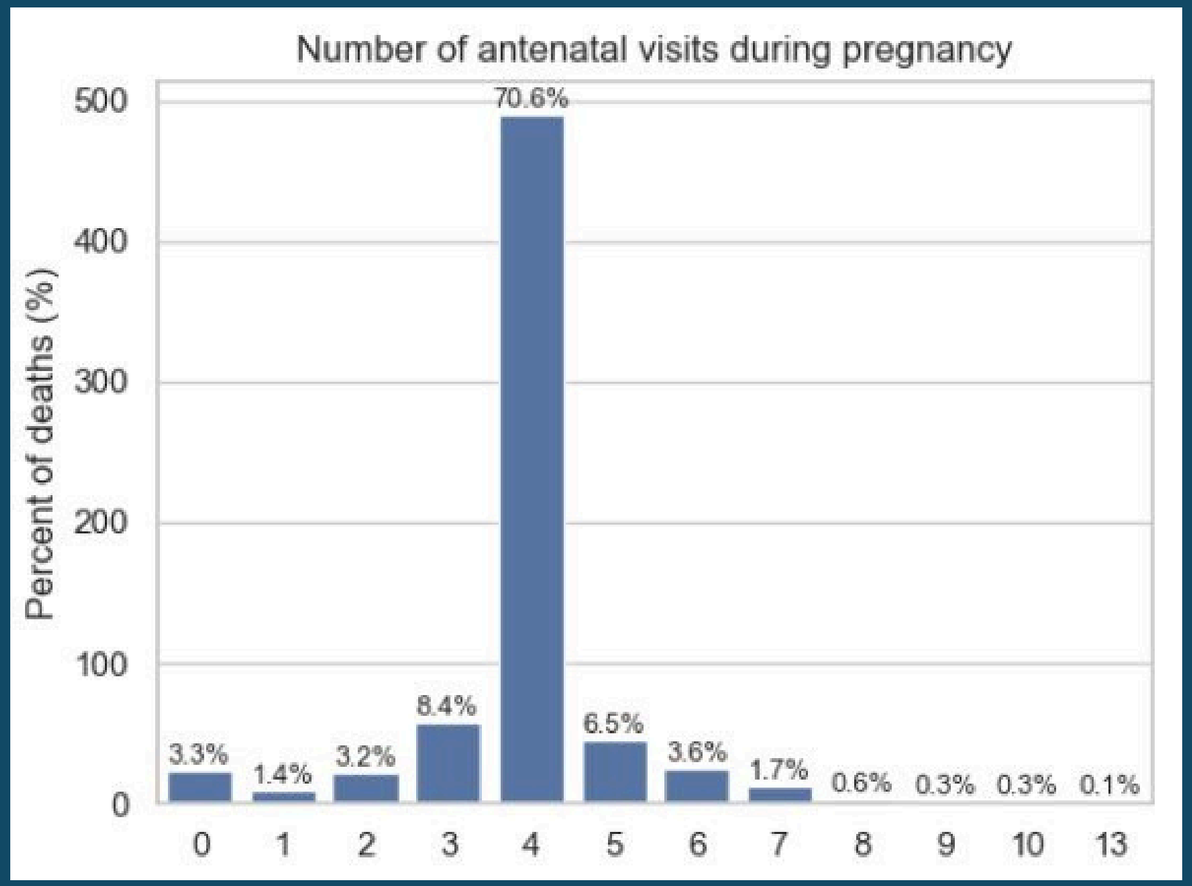


Chart: Number of antenatal visits vs Under-5 Mortality.

- Mortality is inversely correlated to the number of antenatal visits. The higher the frequency of visits, the lower the mortality.
- This highlights the importance of access and utilization of healthcare services.

Geographic/Distance Factor

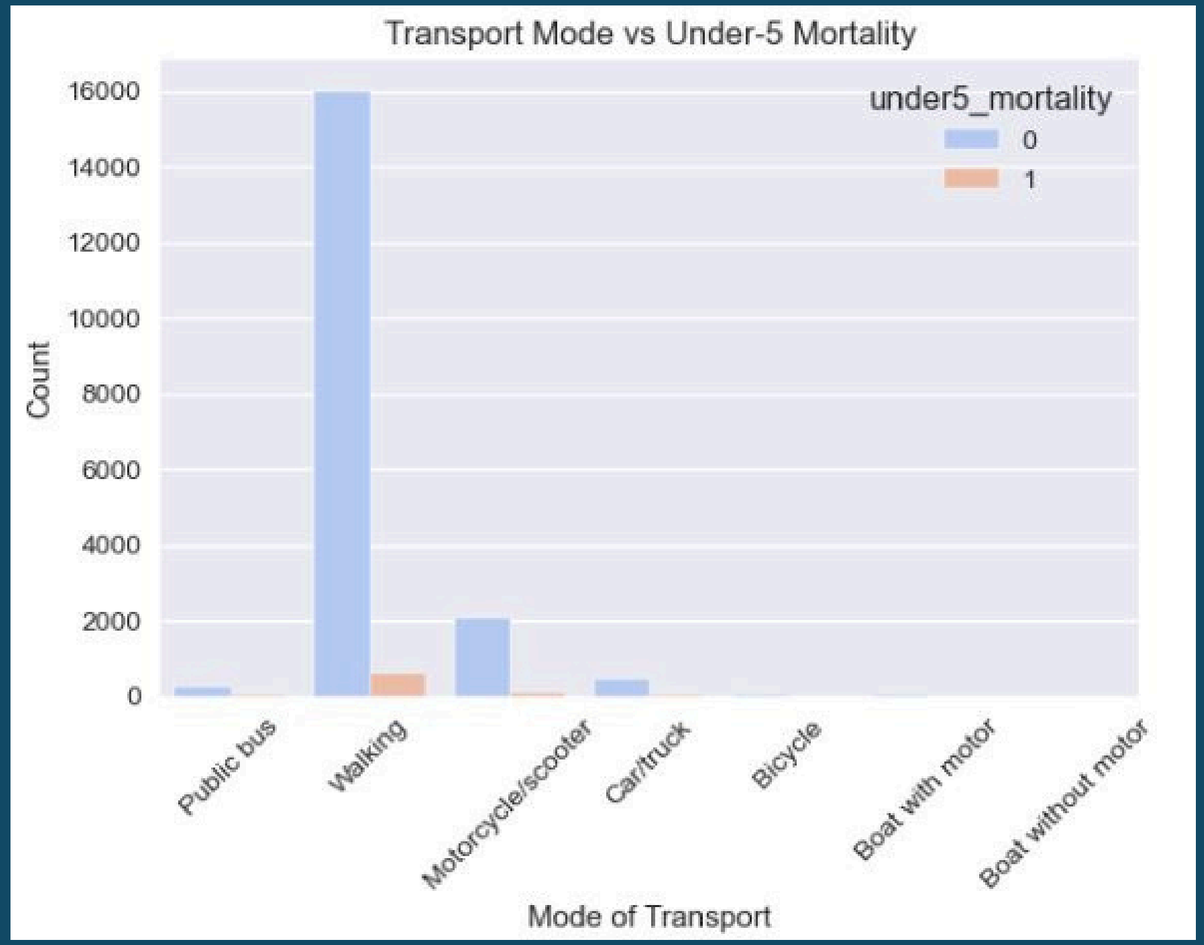


Chart: Plot showing "Transport Mode vs Under-5 Mortality".

- Transport mode is a proxy for distance. The group with a longer distance to a facility (walking) is visibly associated with a higher count of deaths.
- This confirms that geographical barriers are a major risk factor.

Socioeconomic Factors

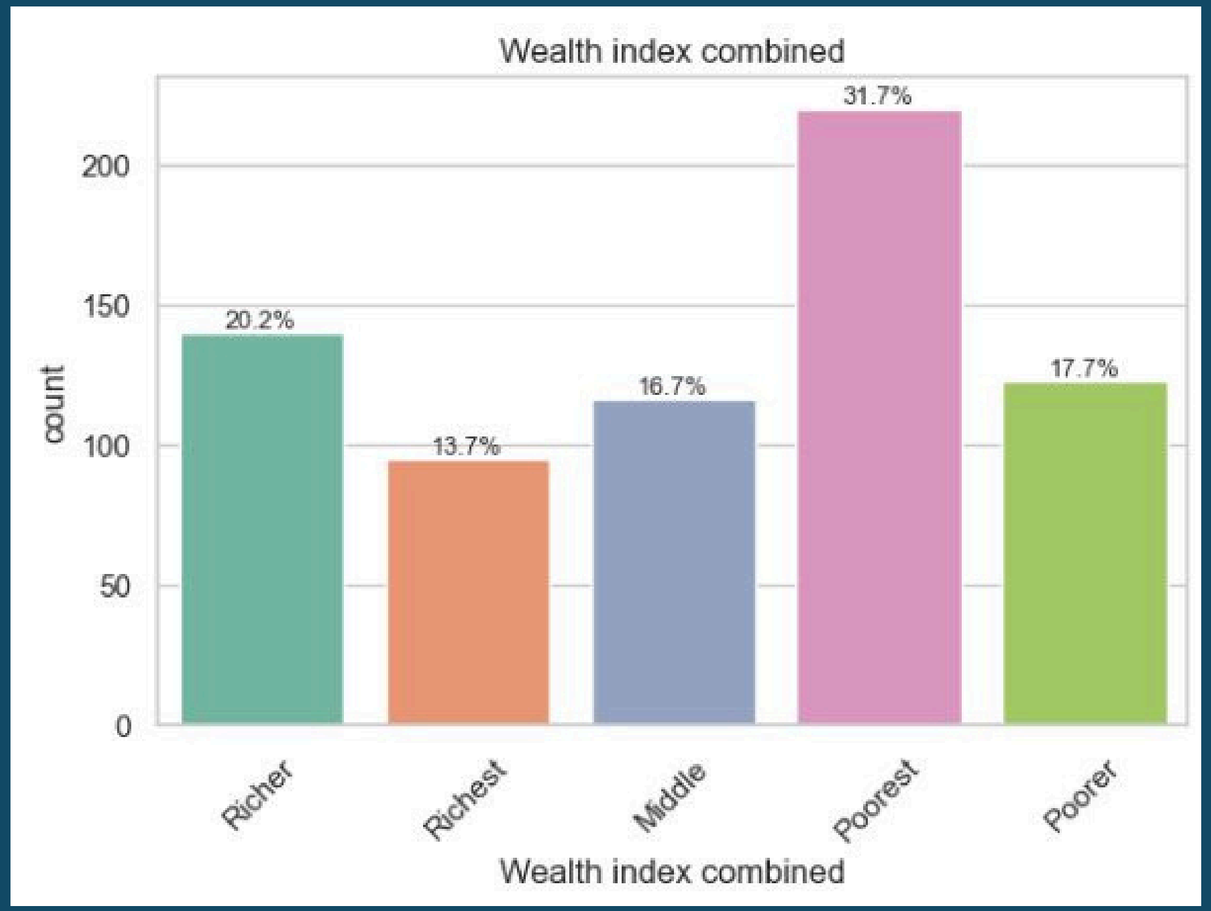
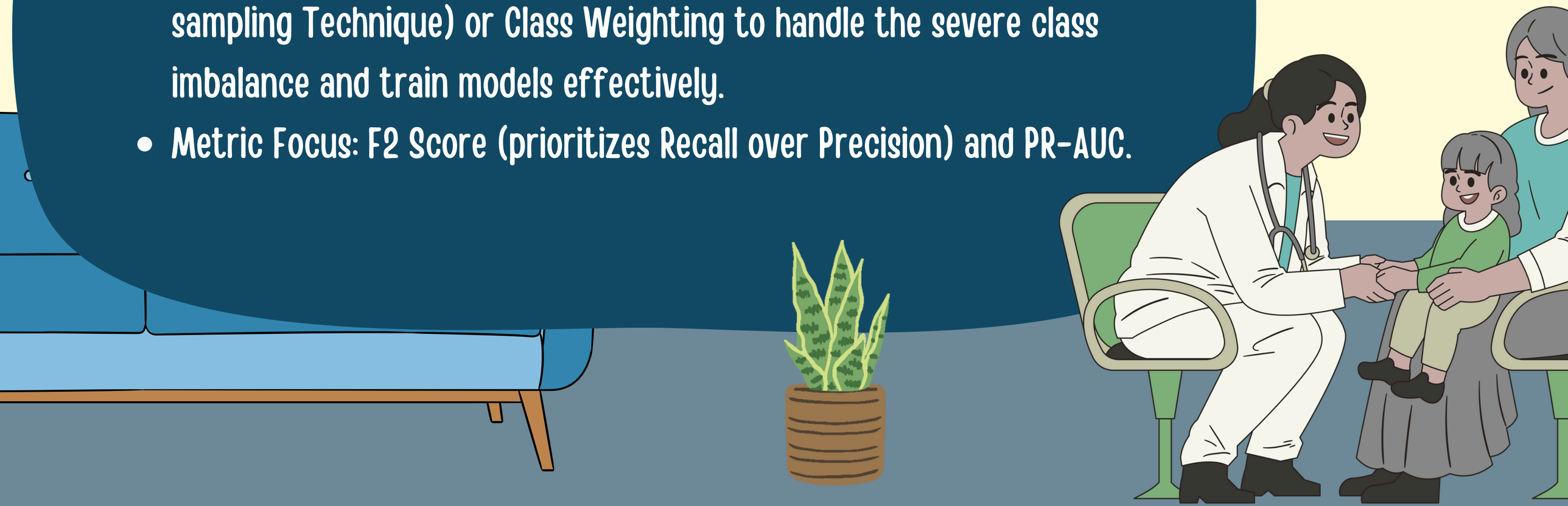


Chart: Plot showing "Wealth Index and distribution of Under-5 Mortality".

- Wealthier households experience fewer mortalities. Poorer ones experience more.
- This confirms that a relationship between wealth and mortality exists, and is a significant factor.

Modelling Strategy

- Models Tested: We tested a range of classifiers, including Logistic Regression, Gradient Boosting, XGBoost, and Random Forest.
- Method: We used techniques like SMOTE (Synthetic Minority Over-sampling Technique) or Class Weighting to handle the severe class imbalance and train models effectively.
- Metric Focus: F2 Score (prioritizes Recall over Precision) and PR-AUC.



Model Performance

Rank	Model	Target	F2 Score	Accuracy	PR-AUC
1	Stacking Ensemble	Under-5	0.968	0.994	0.938
2	Neural Network	Under-5 Mortality	0.902	0.992	0.916
3	Stacking Ensemble	Neonatal	0.855	0.982	0.845
4	LogReg + SMOTE	Under-5 Mortality	0.776	0.956	0.698
5	Logistic Regression	Under-5 Mortality	0.683	0.98	0.74
6	LogReg + SMOTE	Neonatal Mortality	0.693	0.959	0.635
7	Stacking Ensemble	Infant	0.665	0.978	0.439
8	Neural Network	Neonatal Mortality	0.664	0.986	0.715
9	Logistic Regression	Neonatal Mortality	0.499	0.985	0.646
10	LogReg + SMOTE	Infant Mortality	0.369	0.954	0.171
11	Neural Network	Infant Mortality	0.16	0.99	0.319
12	Logistic Regression	Infant Mortality	0.067	0.99	0.2

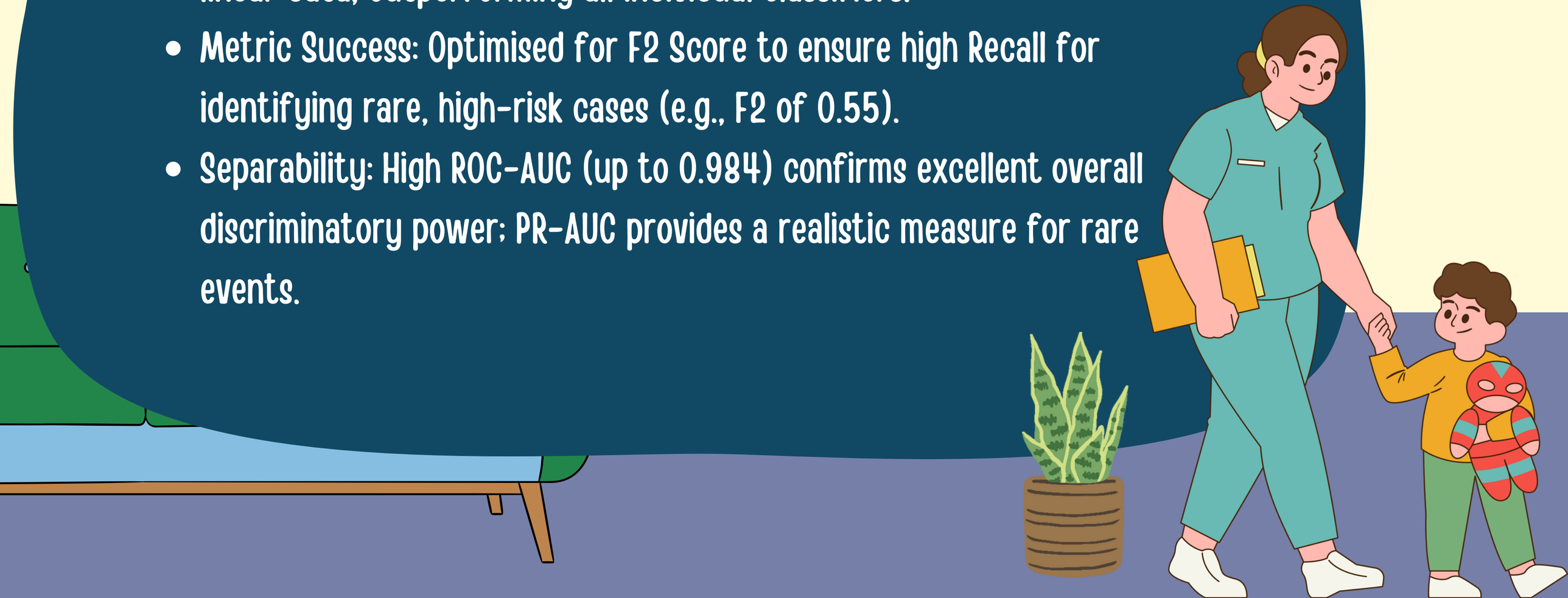
The best model across all three age distributions was the Stacking Ensemble Model which consistently produced the best F2 score, Accuracy and PR-AUC scores

Feature Importance

- Child survival is most influenced by previous child deaths in the family, child nutrition status, maternal/household characteristics (religion, ethnicity, birth order, breastfeeding practices), and access to healthcare and water.
- Maternal education mitigates risks associated with young or advanced maternal age. Policies aimed at maternal education would amplify protective effects.
- Child survival is driven by nutrition, prior family mortality history, and strong geographic disparities, indicating systemic inequities in health outcomes across regions.

Conclusion

- Top Performer: Stacking Ensemble model excelled at complex, non-linear data, outperforming all individual classifiers.
- Metric Success: Optimised for F2 Score to ensure high Recall for identifying rare, high-risk cases (e.g., F2 of 0.55).
- Separability: High ROC-AUC (up to 0.984) confirms excellent overall discriminatory power; PR-AUC provides a realistic measure for rare events.

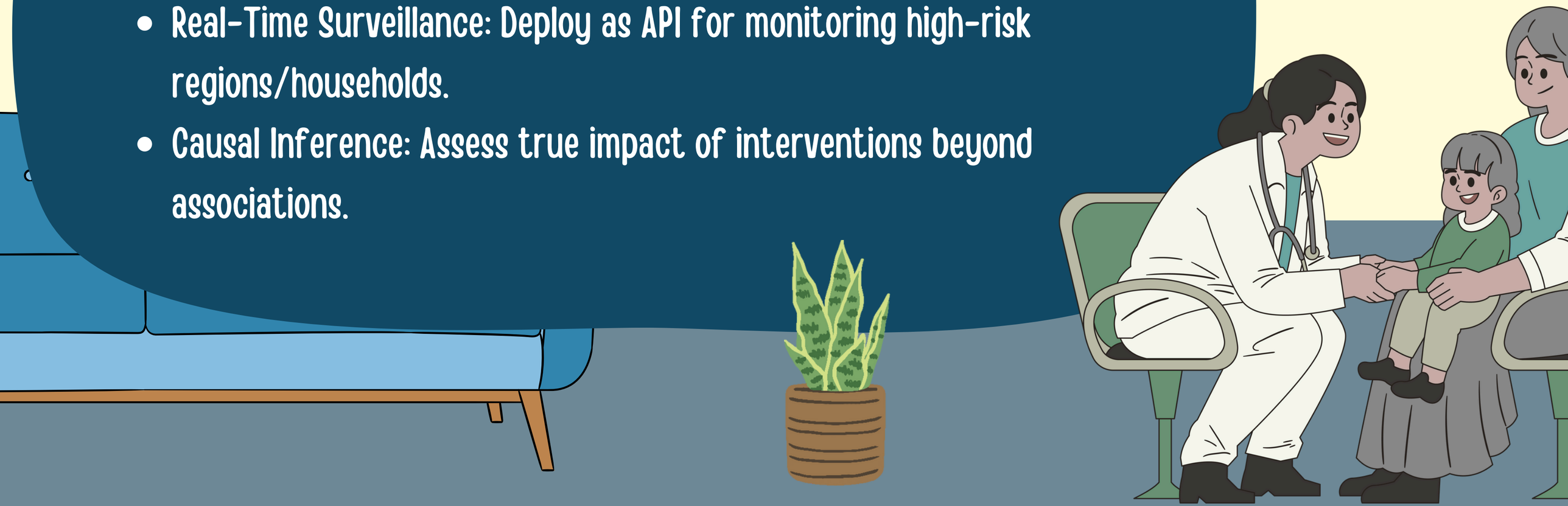


Recommendations

- Prioritise High-Risk Families: Focus on households with prior child deaths through targeted, recurring postnatal care.
- Invest in Maternal Education: Strengthen maternal education to reduce demographic risks and boost child survival.
- Address Basic Needs: Tackle nutrition, clean water, and housing via multi-sectoral policy solutions.
- Close Geographic Gaps: Reallocate resources to underserved, high-mortality regions with poor health access.

Next Steps

1. Model & Data Augmentation
    - Ensemble Refinement: Test Stacking Ensemble II with new meta-learners.
2. Deployment & Insights
    - Real-Time Surveillance: Deploy as API for monitoring high-risk regions/households.
    - Causal Inference: Assess true impact of interventions beyond associations.



Thank You
Questions?