CSC 642 - Statistical Learning with Applications

"PREDICTING STUNTING IN TODDLERS UNDER FIVE: A MACHINE LEARNING APPROACH USING HEIGHT, GENDER, AND AGE"

JONATHAN LATIM & MALENA FUENTES

Spring 2024

AGENDA OVERVIEW

01

BACKGROUND OF THE STUDY

02

STATE-OF-THE-ART

03

FRAMEWORK

04

METHODOLOGY

05

RESULTS

06

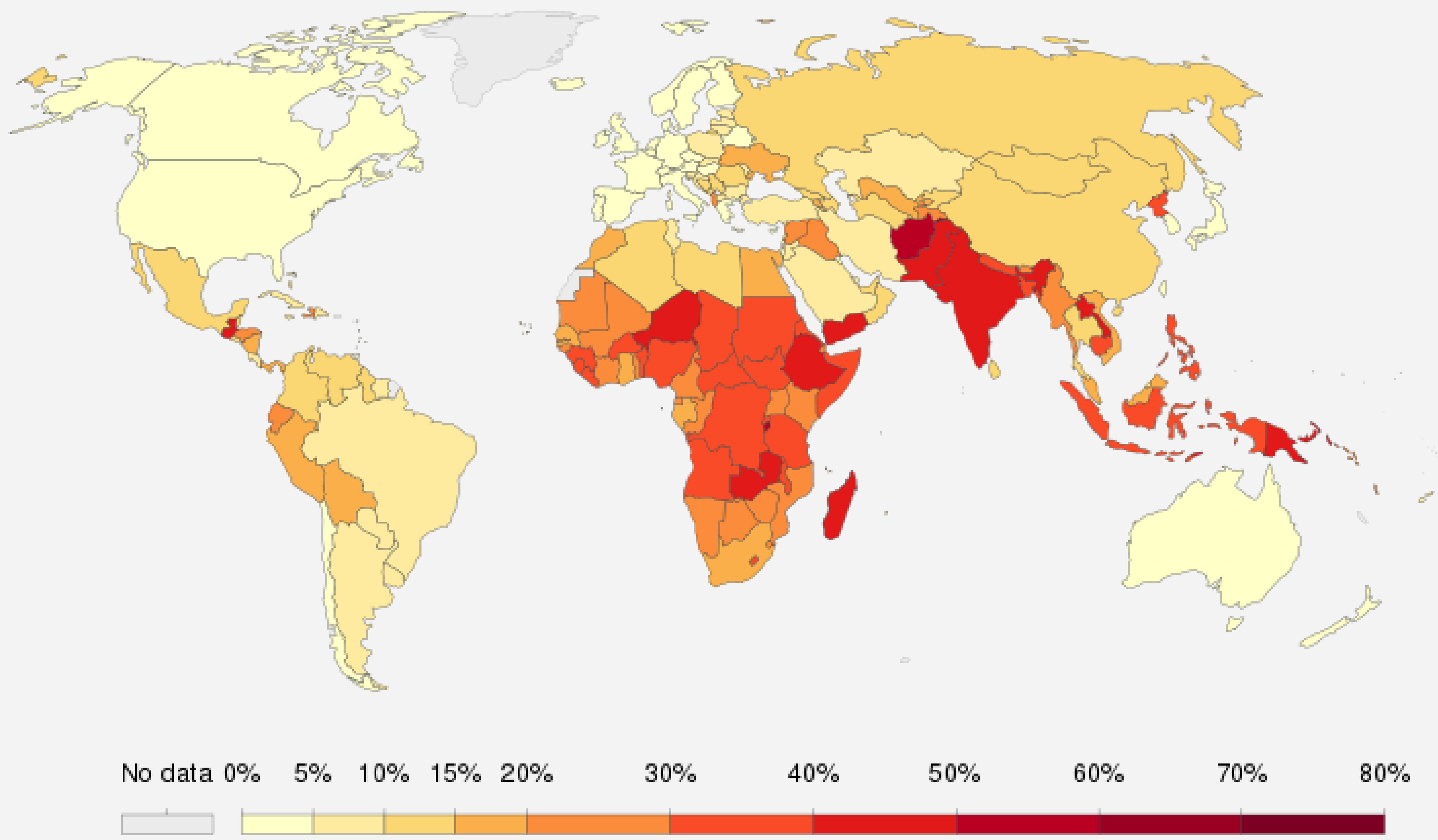
CONCLUSIONS AND FUTURE WORK

BACKGROUND OF THE STUDY

- **Stunting**, or chronic malnutrition, is a significant health issue affecting children under the age of five globally.
- Stunting is the **impaired growth and development** that children experience from poor nutrition, repeated infection, and inadequate psychosocial stimulation.
- As of 2020, an estimated **149 million children** under 5 years of age, are stunted worldwide. More than 85% of the world's stunted children live in Africa and Asia.

Share of children who are stunted, 2016

The share of children younger than five who are stunted – significantly shorter than the average for their age, as a consequence of poor nutrition and/or repeated infection.



Source: Institute for Health Metrics and Evaluation (IHME)
Note: Stunting in children is defined as being less than two standard deviations below the median height for their age.

STATE-OF-THE-ART

Machine learning has been used to predict stunting in toddlers based on factors like height, gender, and age.

INDONESIA

- In East Java Province, Indonesia, a random forest model was used but did not outperform multi-linear regression.

RWANDA

- In Rwanda, a gradient boosting classifier model achieved an 80.49% training accuracy, identifying predictors such as mother's height and education.

BANGLADESH

- In Bangladesh, logistic regression and machine learning classifiers were used to predict malnutrition risks, with random forest achieving the highest accuracy.

INDONESIA

- In Indonesia, a model using logistic regression and ROC Curve analysis was developed to prevent and delay stunting in toddlers by 64%.

PROBLEM STATEMENT

Our study aims to detect stunting in children under five years old

01

SCOPE OF THE STUDY

Identify the predictive factors of stunting in under-five children, using machine learning techniques to develop a model that can accurately classify stunting status.

02

RELEVANCE OF THE STUDY

Understanding predictive factors of stunting is crucial for effective interventions and policies to reduce stunting prevalence, improving child health.

03

RESEARCH QUESTION

How can machine learning techniques enhance early detection and intervention strategies for childhood stunting?

FRAMEWORK

80/20 split
121,000 rows of data

OVERVIEW OF DATASET

The "Stunting Baby/Toddler Detection" dataset uses WHO's z-score formula to detect stunting in children under five.

DATASET COLUMN DETAILS

AGE (MONTHS)
ages 0 to 60 months

GENDER
'male' and 'female'

HEIGHT
Recorded in centimeters

NUTRITION STATUS
categorized into 4 statuses - 'severely stunted', 'stunted', 'normal', and 'tall'.

FRAMEWORK

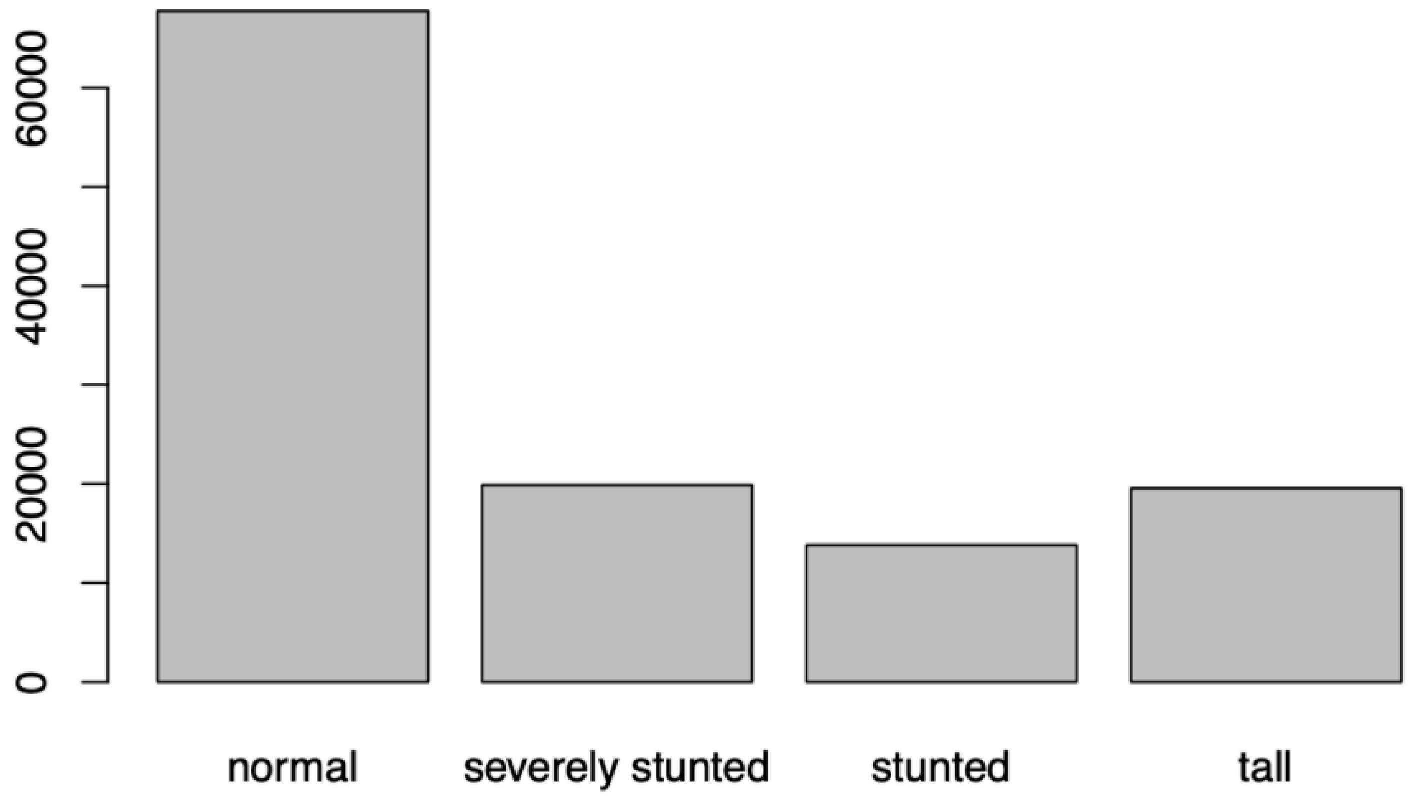
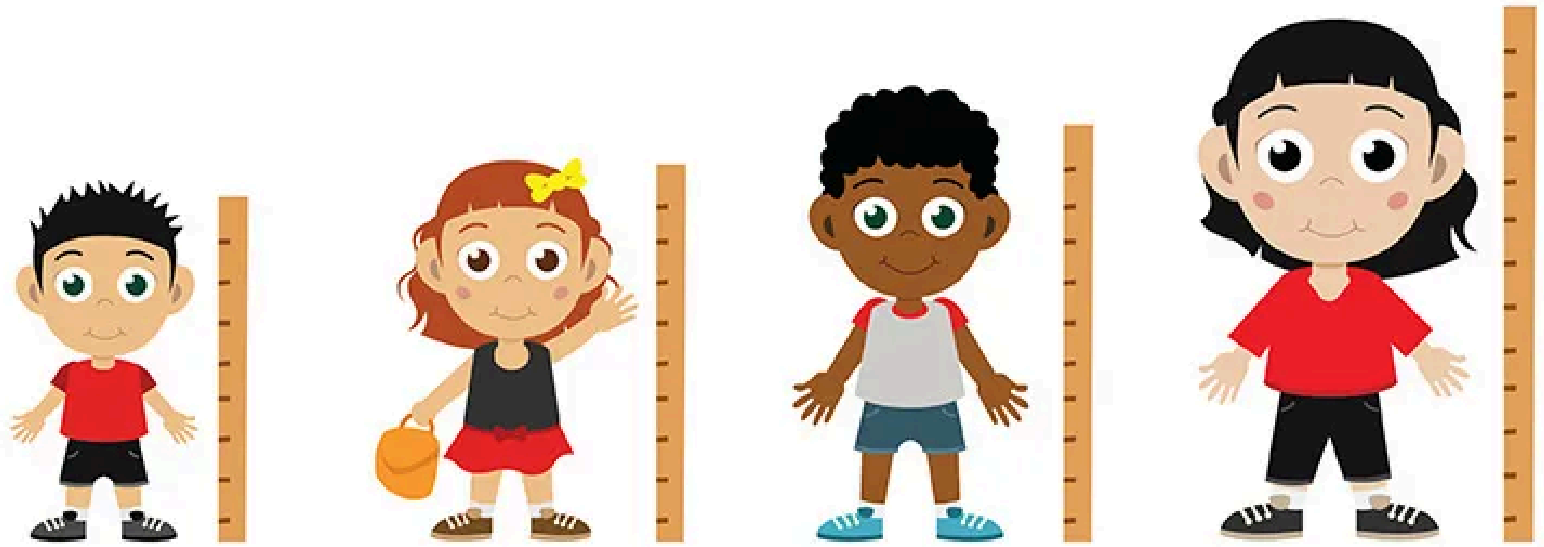
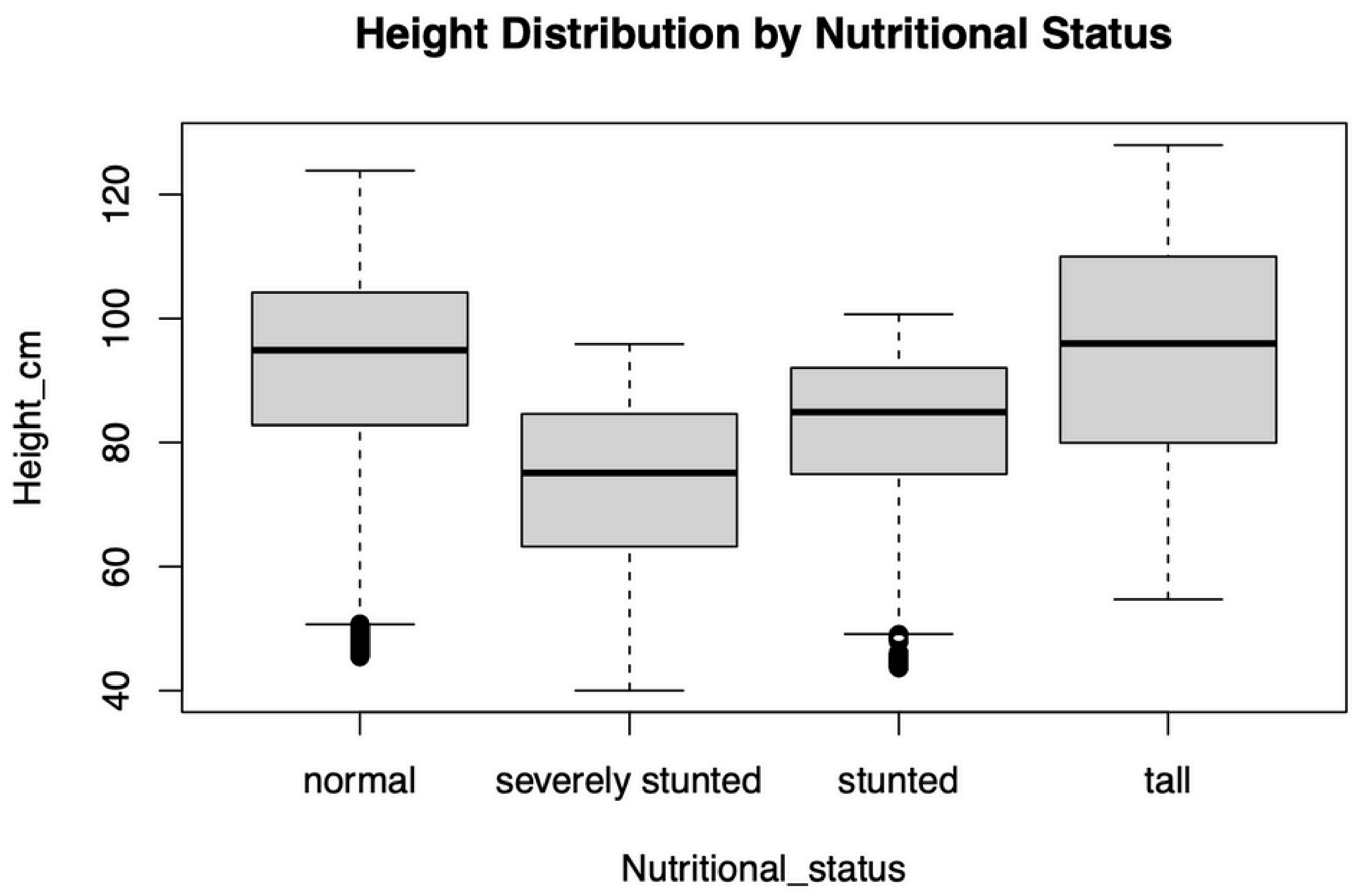


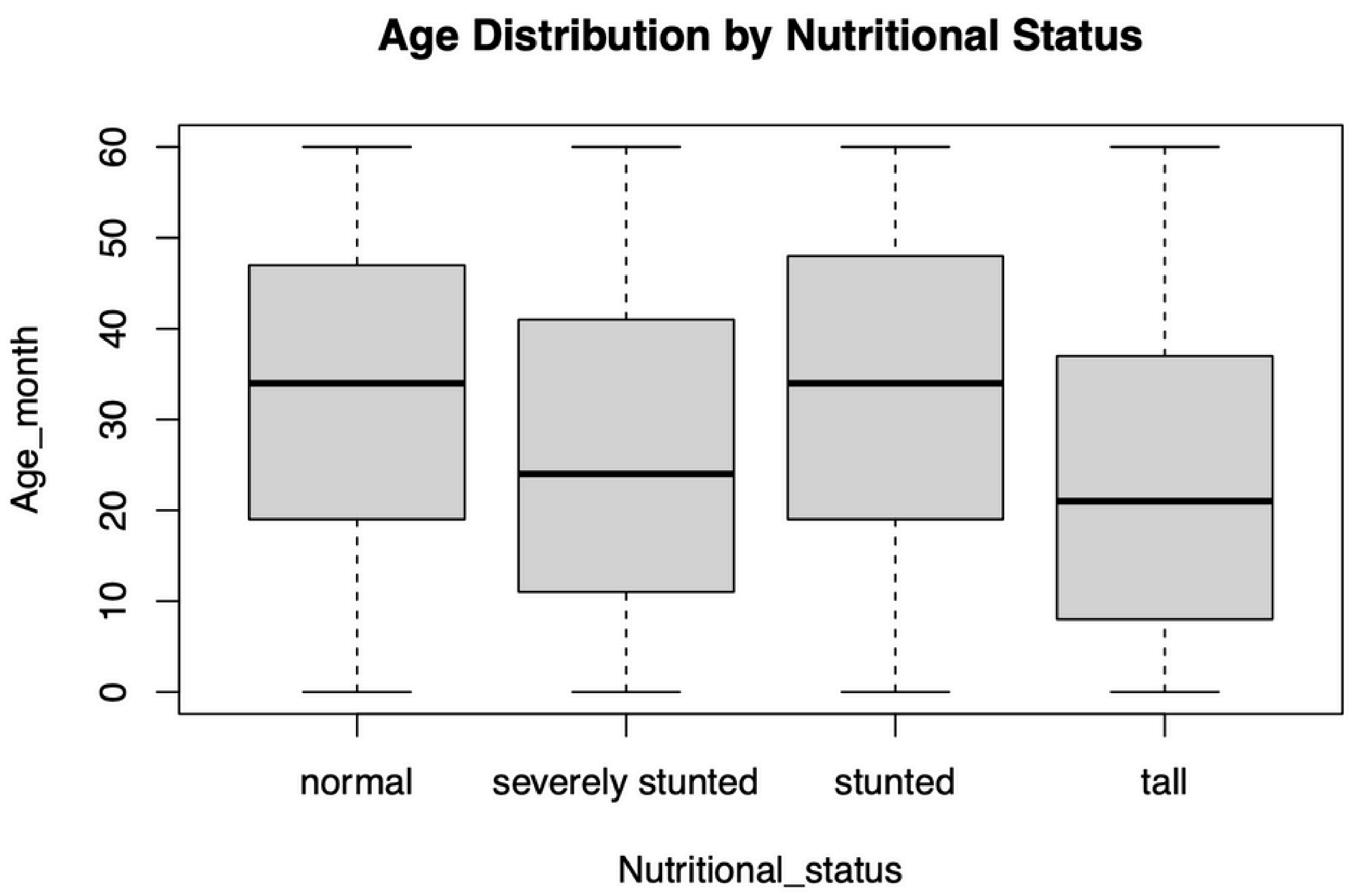
Figure 1: Distribution of Nutritional_status categories in the dataset, indicating the proportion of children classified as 'normal', 'severely stunted', 'stunted', and 'tall'.

Nutritional Status	Population Distribution
Normal	56%
Severely Stunted	16%
Stunted	11%
Tall	16%



FRAMEWORK





Distribution of children's height based on nutritional status providing insights into the **impact of nutrition on growth.**

Shows how ages of children are distributed across different nutritional categories. It helps visualize if there are **age differences among children with different nutrional status.**

METHODOLOGY

Logistic Regression	Interpretable, suitable for multiclass classification.
Decision Trees	Captures complex, non-linear patterns, easy to interpret.
Bagging	Reduces overfitting, improves generalization.
Random Forest	Further reduces overfitting, higher performance.
Support Vector Machines (SVM)	Effective for high-dimensional data, captures complex relationships.

RESULTS

The **accuracy** metric explains the overall proportion of correct predictions.

Kappa measures agreement between predictions and true labels, adjusted for chance.

ML Algorithms	Accuracy	Kappa
Logistic Regression	0.7767	0.6206
Decision Trees	0.7857	0.6446
Bagging	**0.9993**	**0.9989**
Random Forest	0.9647	0.9413
Support Vector Machines	0.9897	0.9832

Severely Stunted Class

Machine Learning Algorithms	Sensitivity	Specificity
Logistic Regression	0.8571	0.9384
Decision Trees	0.8498	0.9469
Bagging	0.9995	0.9998
Random Forest	0.9985	0.9985
Support Vector Machines	0.9919	0.9972

Stunted Class

Machine Learning Algorithms	Sensitivity	Specificity
Logistic Regression	0.19261	0.98009
Decision Trees	0.33161	0.97539
Bagging	0.9963	0.9998
Random Forest	0.80148	0.99967
Support Vector Machines	0.9627	0.9975

KEY TAKEAWAYS

NO “UNIVERSAL” BEST MODEL

Trade off between cost of running one model vs accuracy must be considered

Overall accuracy might not correlate to accuracy in predicting the preferred nutritional class.

DATA-DRIVEN GUIDANCE

Helps guide decision making and policy making

Targeted intervention is more effective

CONCLUSIONS AND FUTURE WORK

01

Oversampling: Replicate or synthesize data points from minority classes to increase their representation in the training set to counter bias and improve decision boundaries.

- Random Oversampling
- SMOTE (Synthetic Minority Oversampling Technique)

02

Collecting More Information: Include additional factors like family income and education levels to inform policy formation and target interventions more effectively.

03

Understanding Gender Differences: Investigate why females have a lower likelihood of being classified as 'severely stunted' compared to males for decision-making and lead to more targeted interventions.

THANK YOU

April 25, 2024