Mariam, Amira, Safia

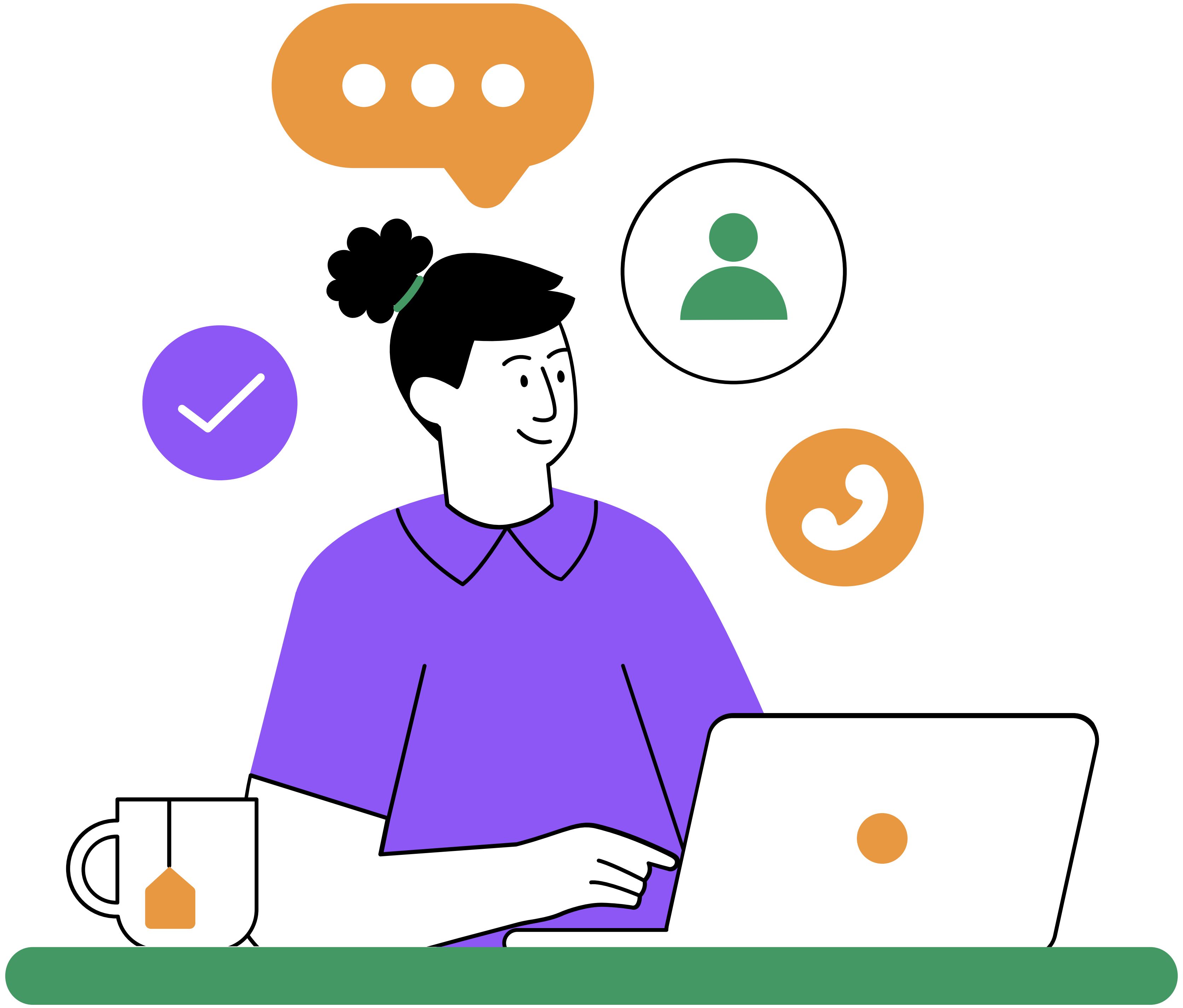
Child Malnutrition

Project Title:

Predictive Modeling of Child Malnutrition Using UNICEF Data

Team Members:

- Mariam Goda
- Amira Yasser
- Safia Mohamed



Why we chose this dataset

Relevance: Provides data across multiple countries, enabling cross-regional analysis.

Comprehensive Indicators: Includes key malnutrition metrics—stunting, wasting, underweight, and overweight.

Availability: Publicly accessible and well-structured

EDA Highlights

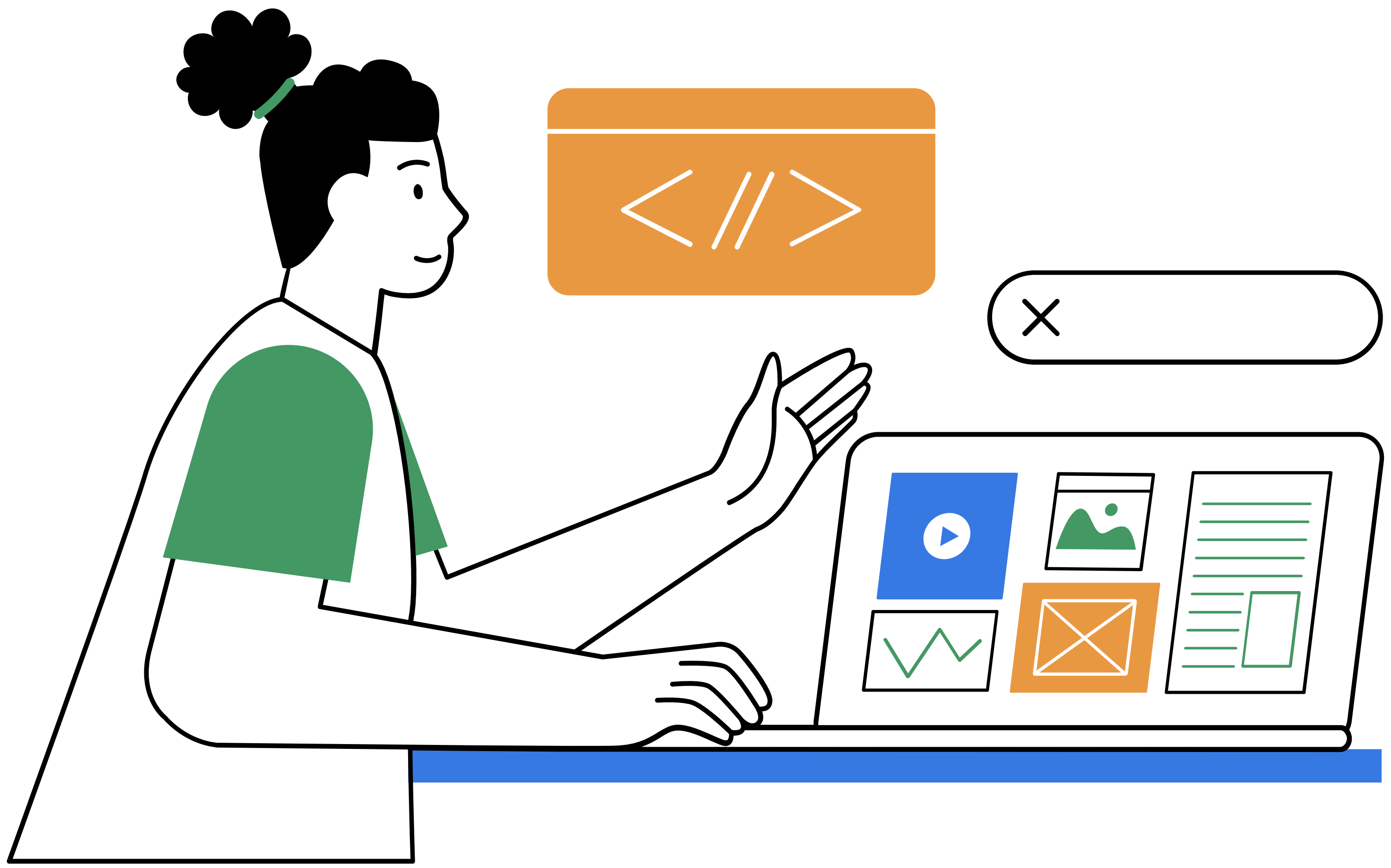
- Data Cleaning
- Missing Values
- Outlier Treatment
- Feature Relationships

Data Cleaning

- Dropped Irrelevant Columns: Removed columns like Report Author, ISO code, UNICEF Survey ID, and notes that don't affect prediction.

- Encoded Categorical Features: Used LabelEncoder for transforming text columns into numeric values.

- Scaling: Applied StandardScaler to normalize the numerical features for better model performance.



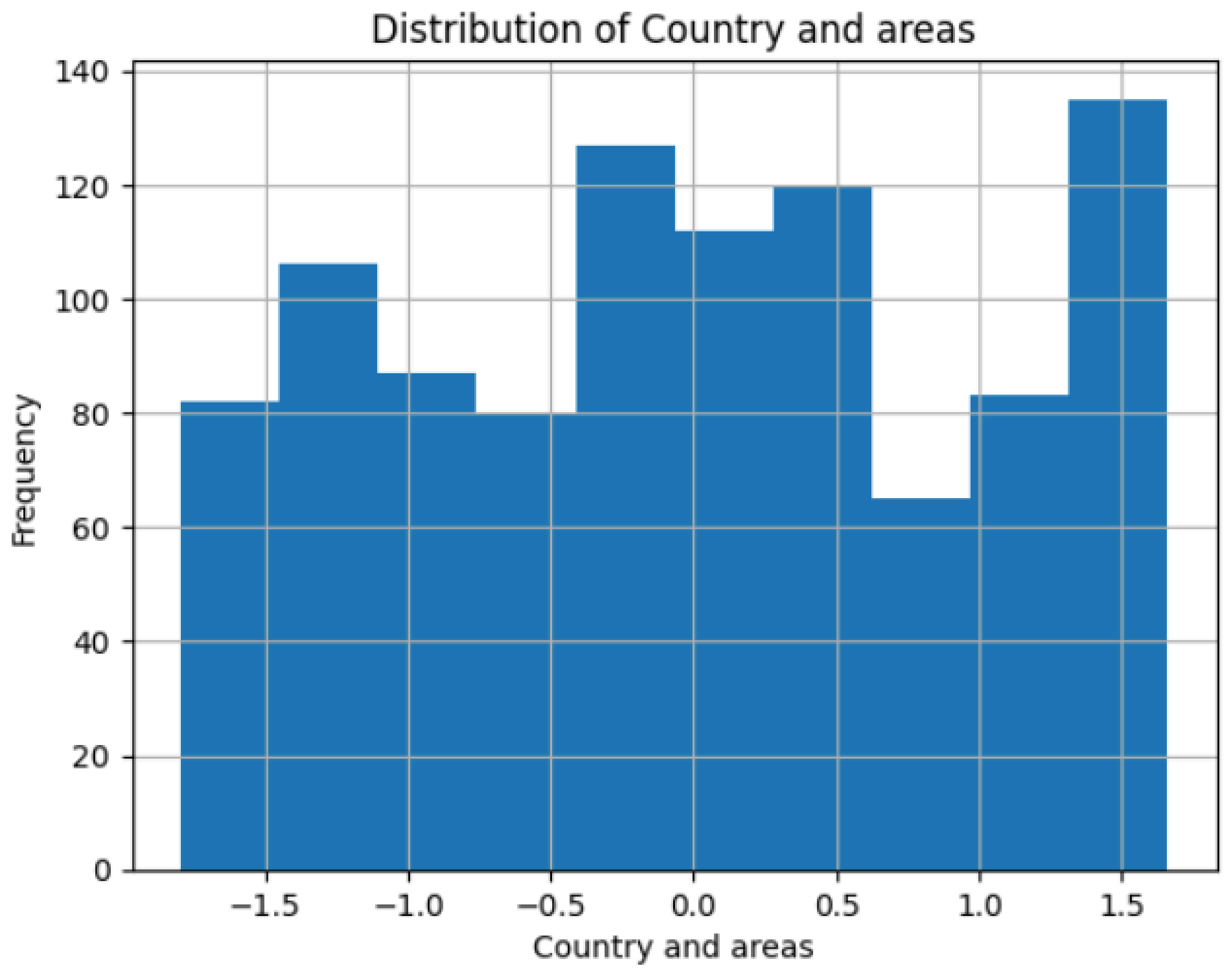
Missing values

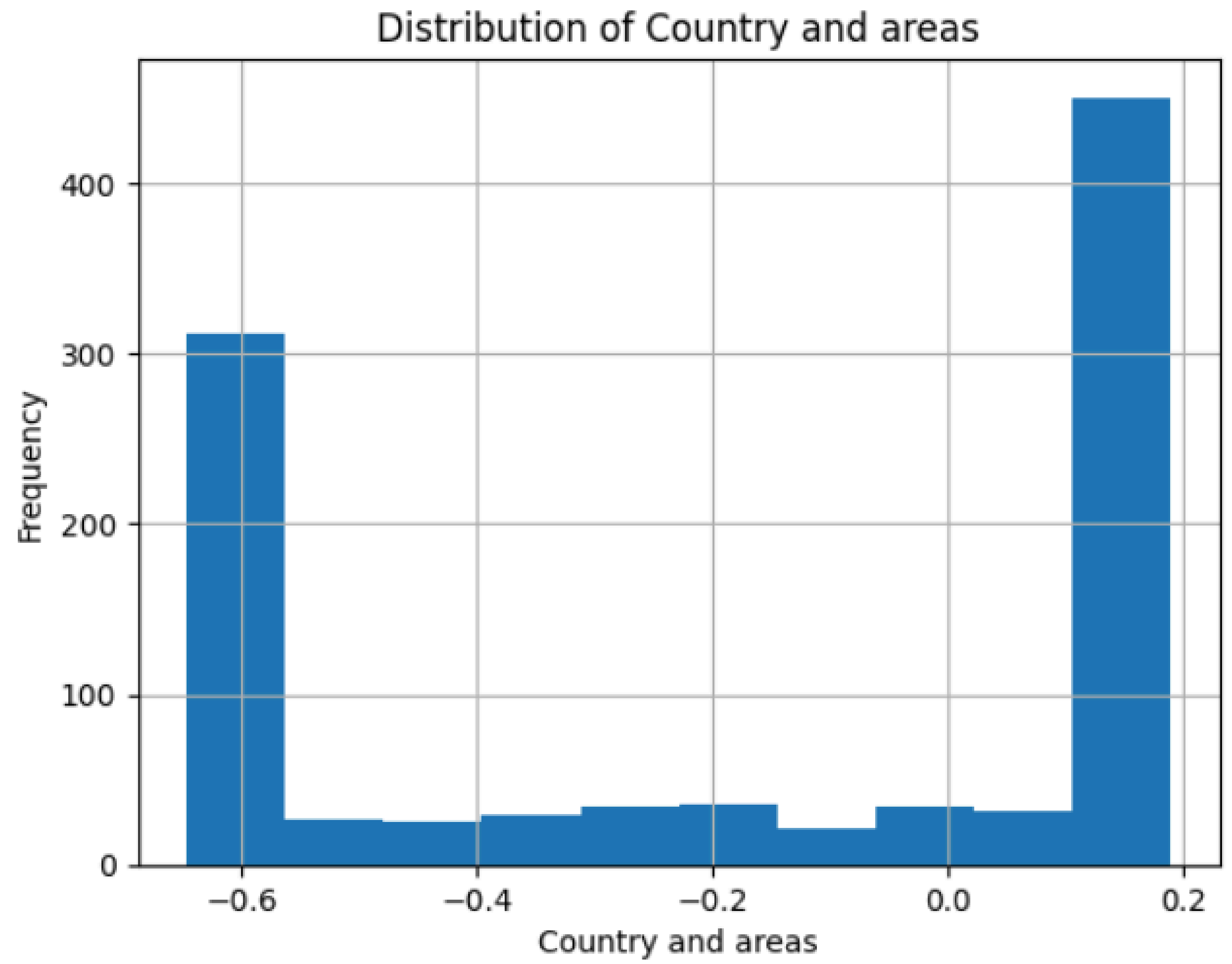
Identified columns with missing values and calculated their skewness:

- Used median imputation for highly skewed columns.
- Used mean imputation for nearly symmetric distributions.

Outlier Detection and Handling

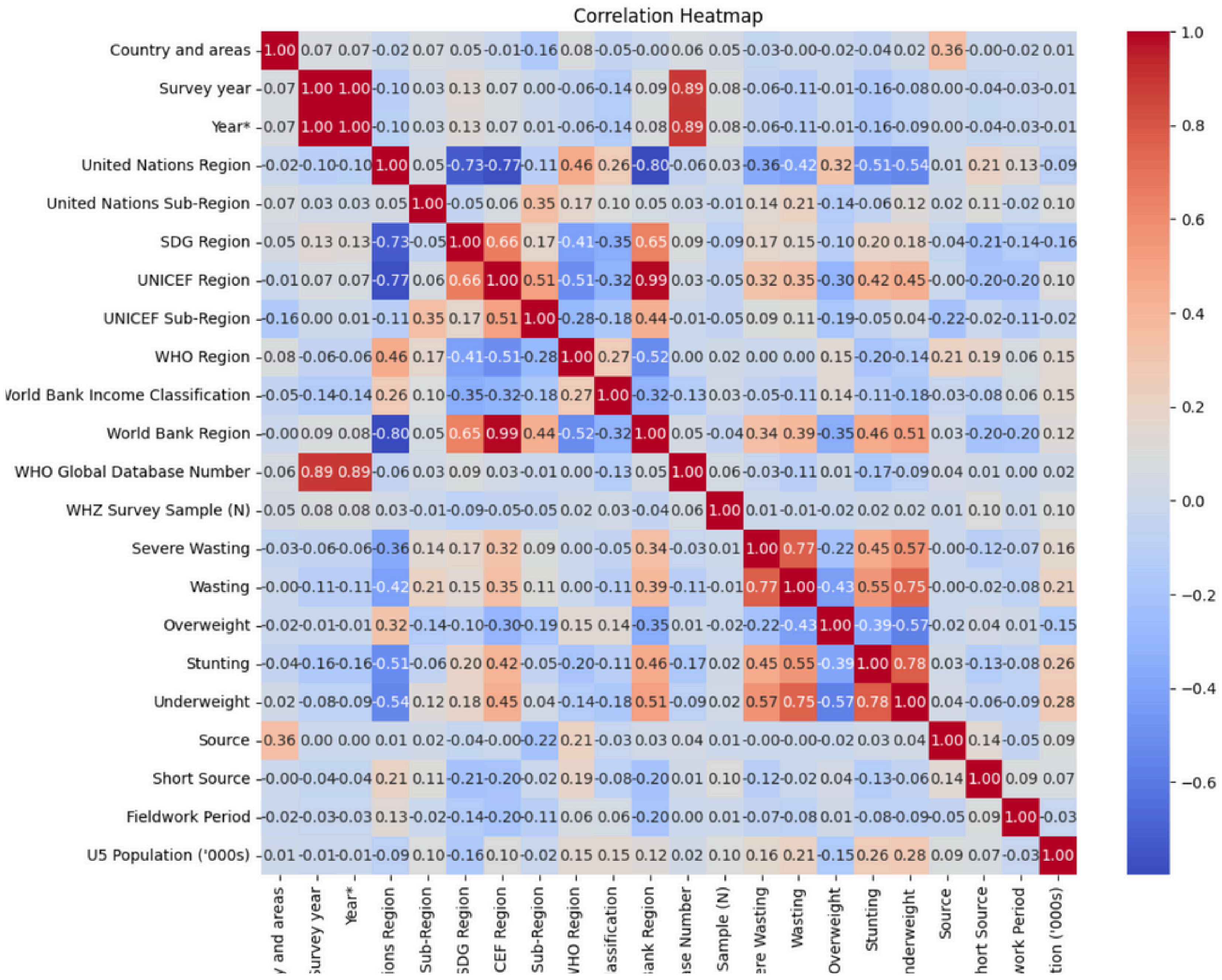
- Detected outliers using IQR.
- Applied clipping technique to reduce their impact.
- Visualized distributions before and after treatment (Box & Histogram Plots).



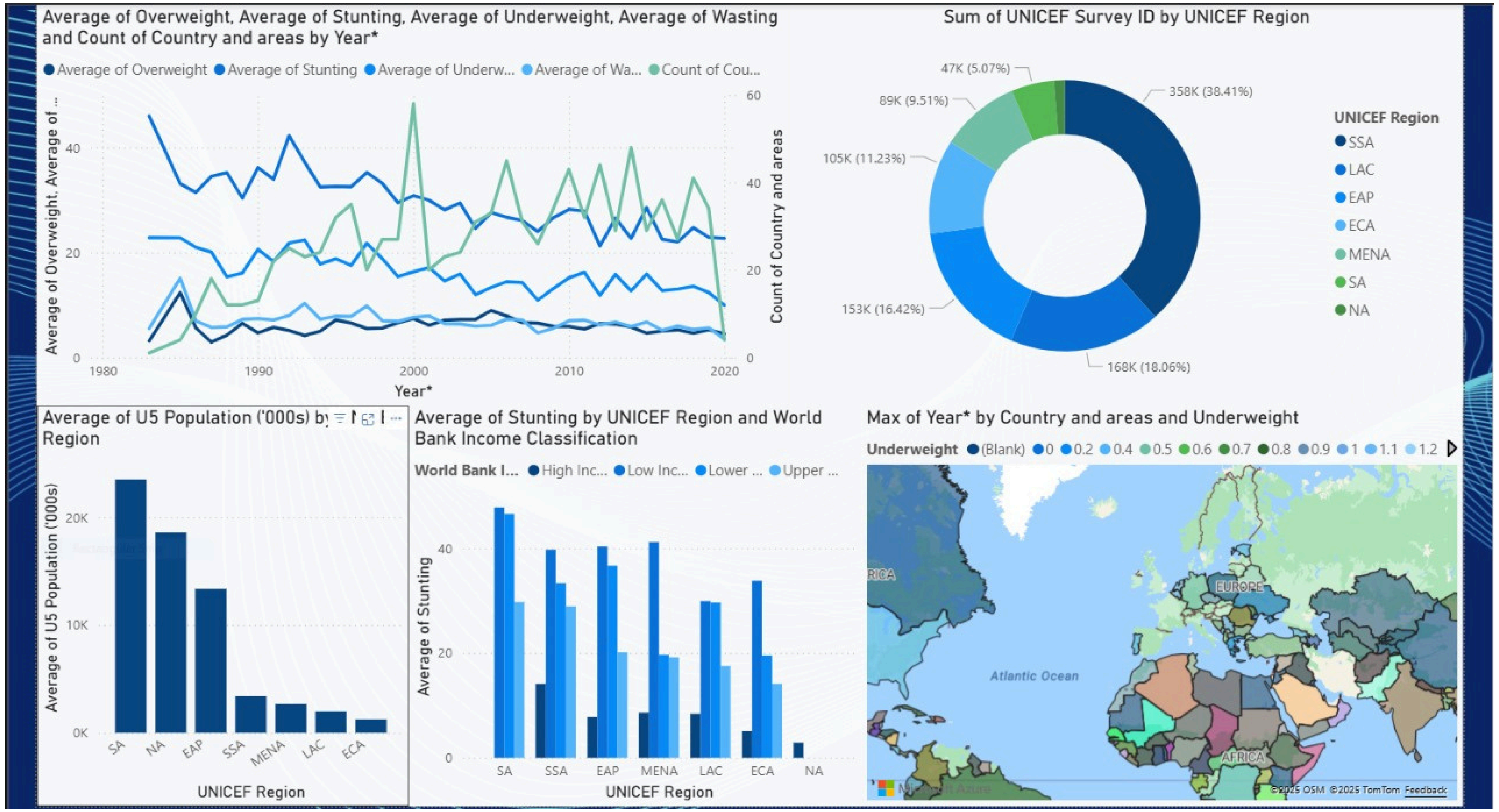


Exploratory Visualizations

- **Histogram plots:** Showed distribution of each feature.
- **Box plots:** Helped identify and visualize outliers.
- **Violin plots:** Provided combined view of data distribution and density.
- **Correlation Heatmap:** Revealed relationships between features.

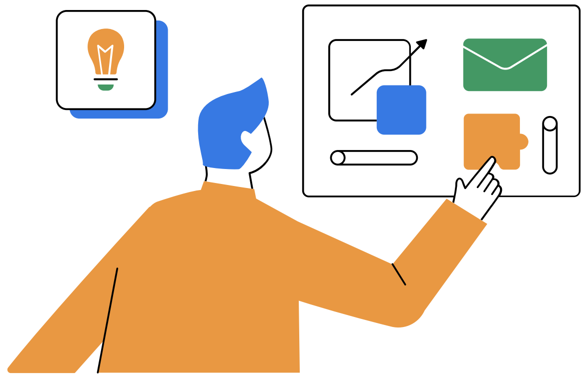


Dashboard



- The UNICEF Survey ID distribution shows Sub-Saharan Africa (SSA) dominates with 47% of survey entries, followed by East Asia and Pacific (18.06%) and Latin America and Caribbean (16.42%).
- The U5 Population graph reveals significant variations across regions, with Sub-Saharan Africa having the largest under-5 population.
- Stunting rates vary dramatically by World Bank income classification and UNICEF region, with lower-income regions showing higher stunting prevalence.
- The time series data demonstrates long-term trends in child nutrition metrics, showing some improvements and fluctuations over the decades.

Machine Learning



Step 01

Loads all necessary libraries for:

- Data processing: pandas, numpy
- Visualization: matplotlib, seaborn
- ML models: sklearn, xgboost
- Interpretability: SHAP, LIME

Step 02

Load and Explore Cleaned Data

- We begin by loading the cleaned survey dataset and ensuring the data is structured correctly.

Step 03

Feature Engineering

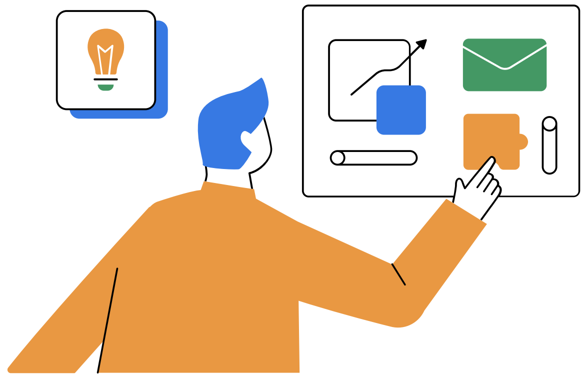
- Creates a new binary target Malnourished based on thresholds of indicators.
- Visualizes class imbalance.
- Creates an Avg_Malnutrition column to average the malnutrition measures.
- Computes a burden estimate using population.

```
df['Malnourished'] = ( (df['Severe Wasting'] > 0.1) |
                       (df['Wasting'] > 0.1) |
                       (df['Stunting'] > 0.1) |
                       (df['Underweight'] > 0.1)
                     ).astype(int)
```

```
print(df['Malnourished'].value_counts(normalize=True))
```

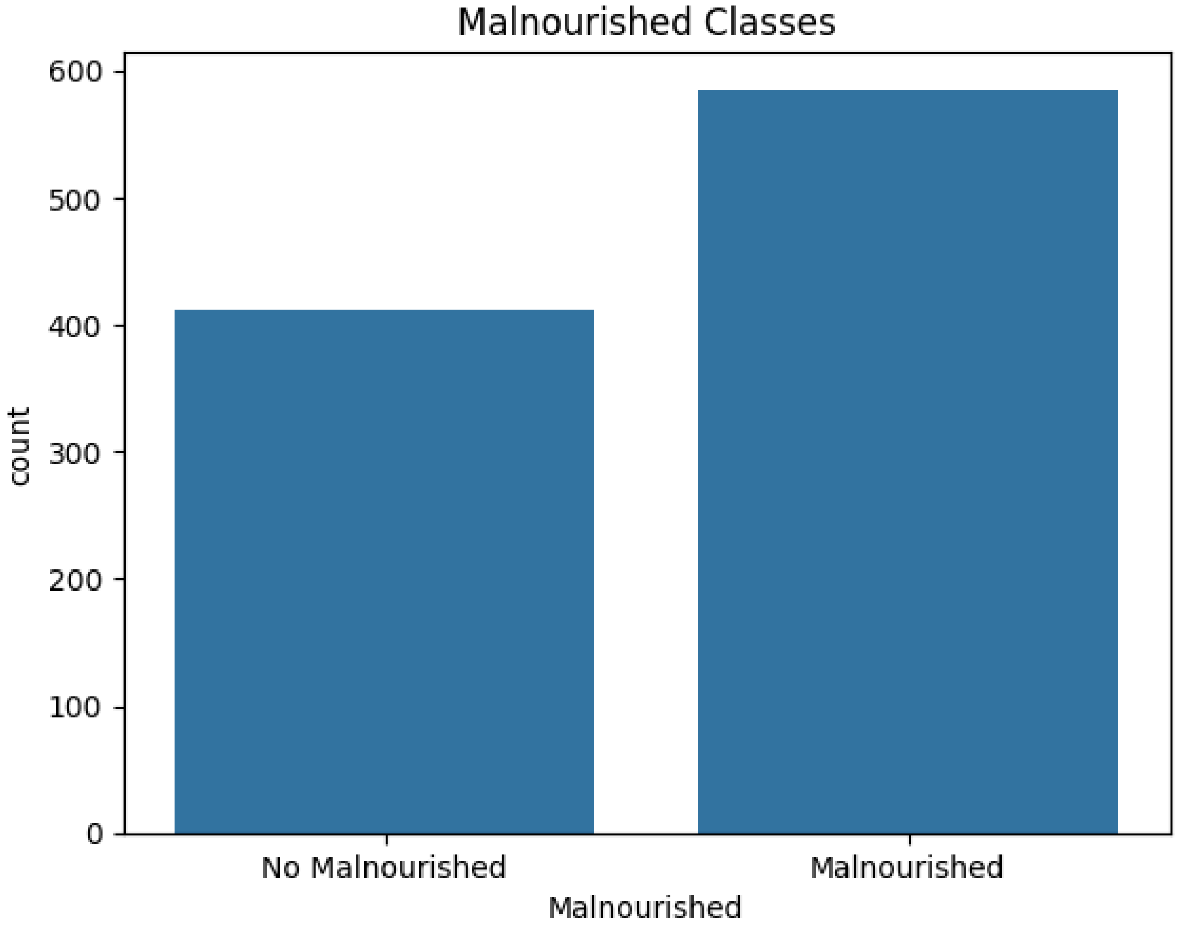
```
Malnourished
1    0.58676
0    0.41324
Name: proportion, dtype: float64
```

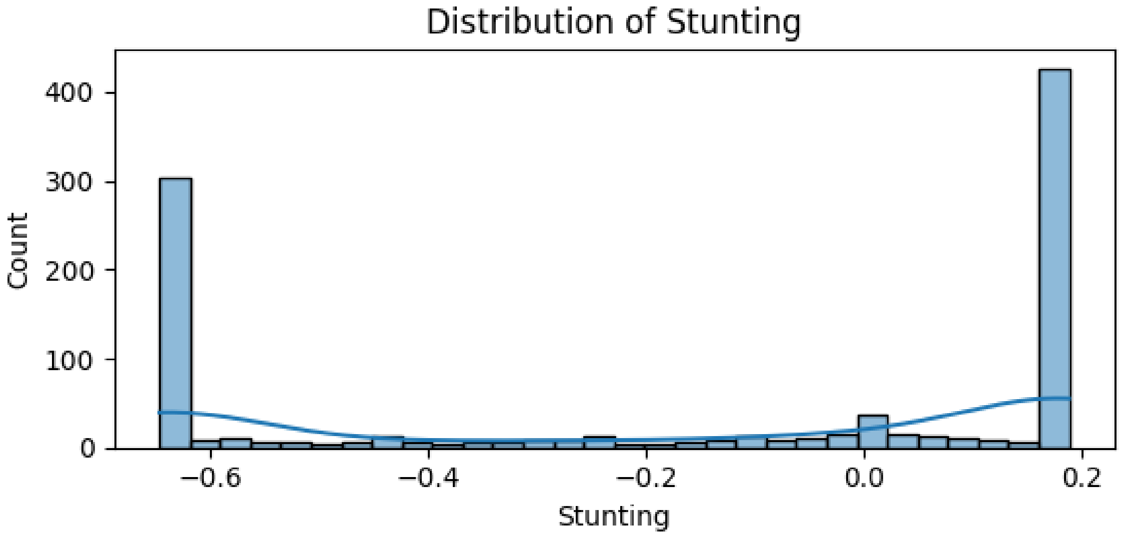
Machine Learning

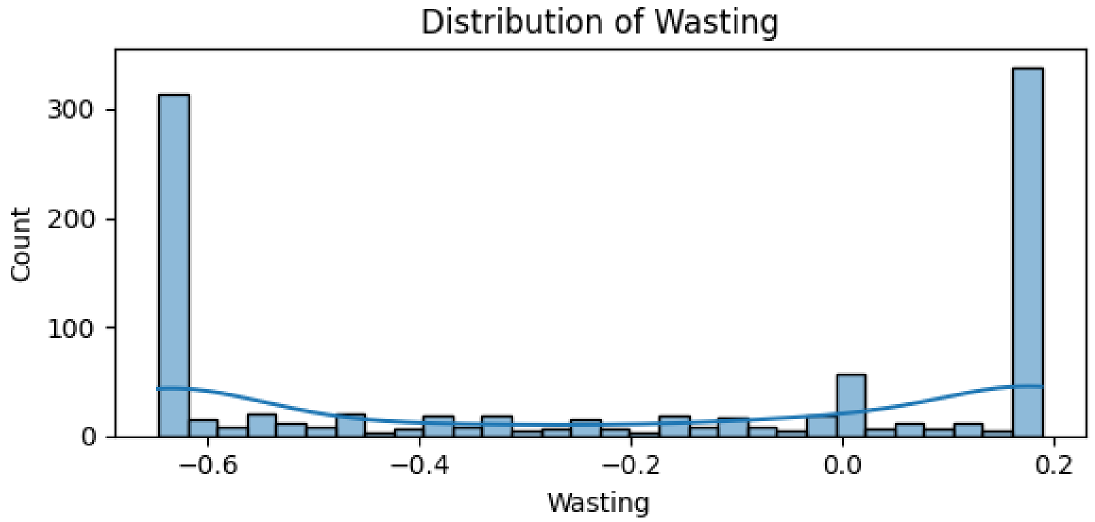


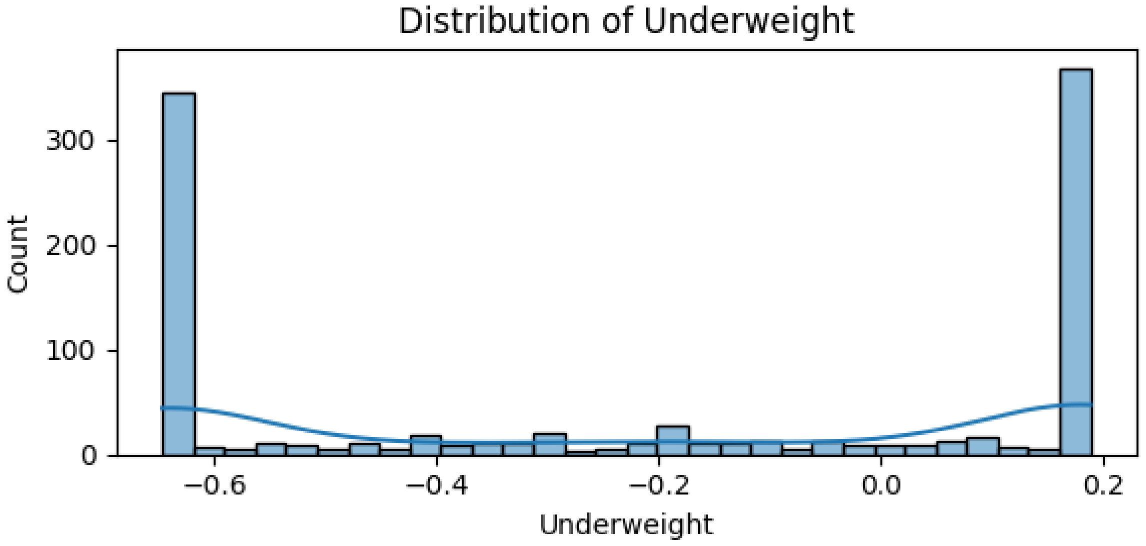
Step 04

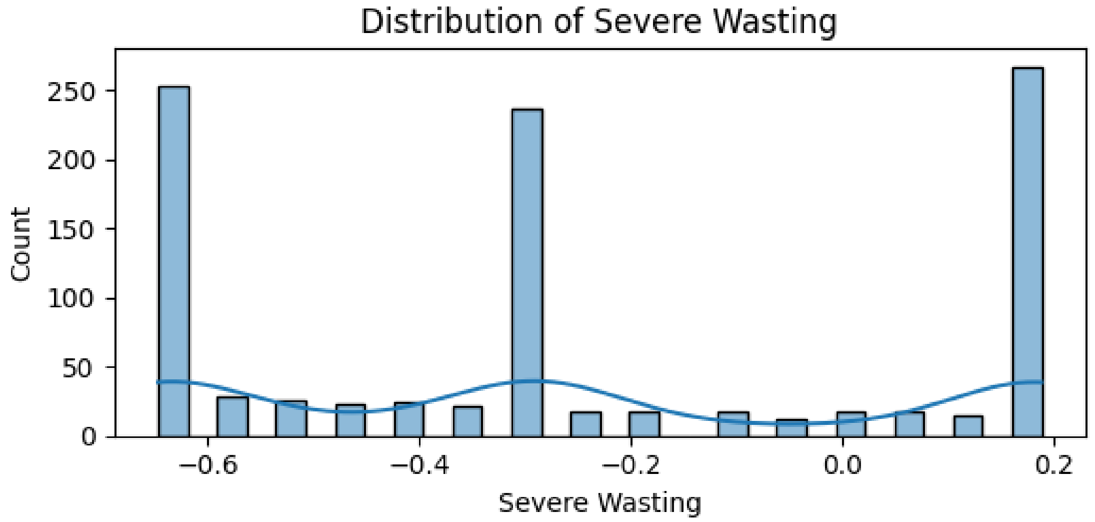
Data Visualization











❋ Machine Learning

```
[ ]  print("Classification report is : ",classification_report(y_test,y_log_pred))

Classification report is :                precision    recall  f1-score   support

           0       0.82      0.79      0.80        96
           1       0.81      0.84      0.82       104

    accuracy                           0.81       200
   macro avg       0.82      0.81      0.81       200
weighted avg       0.82      0.81      0.81       200
```

Step 05

Supervised Classification Models

Trains and evaluates the following models:

- Logistic Regression
- Random Forest
- XGBoost
- SVM
- K-Nearest Neighbors

```
print("Classification report is : ",classification_report(y_test,y_rfc_pred))

Classification report is :                precision    recall  f1-score   support

           0       0.81      0.82      0.82        96
           1       0.83      0.83      0.83       104

    accuracy                           0.82       200
   macro avg       0.82      0.82      0.82       200
weighted avg       0.83      0.82      0.83       200
```

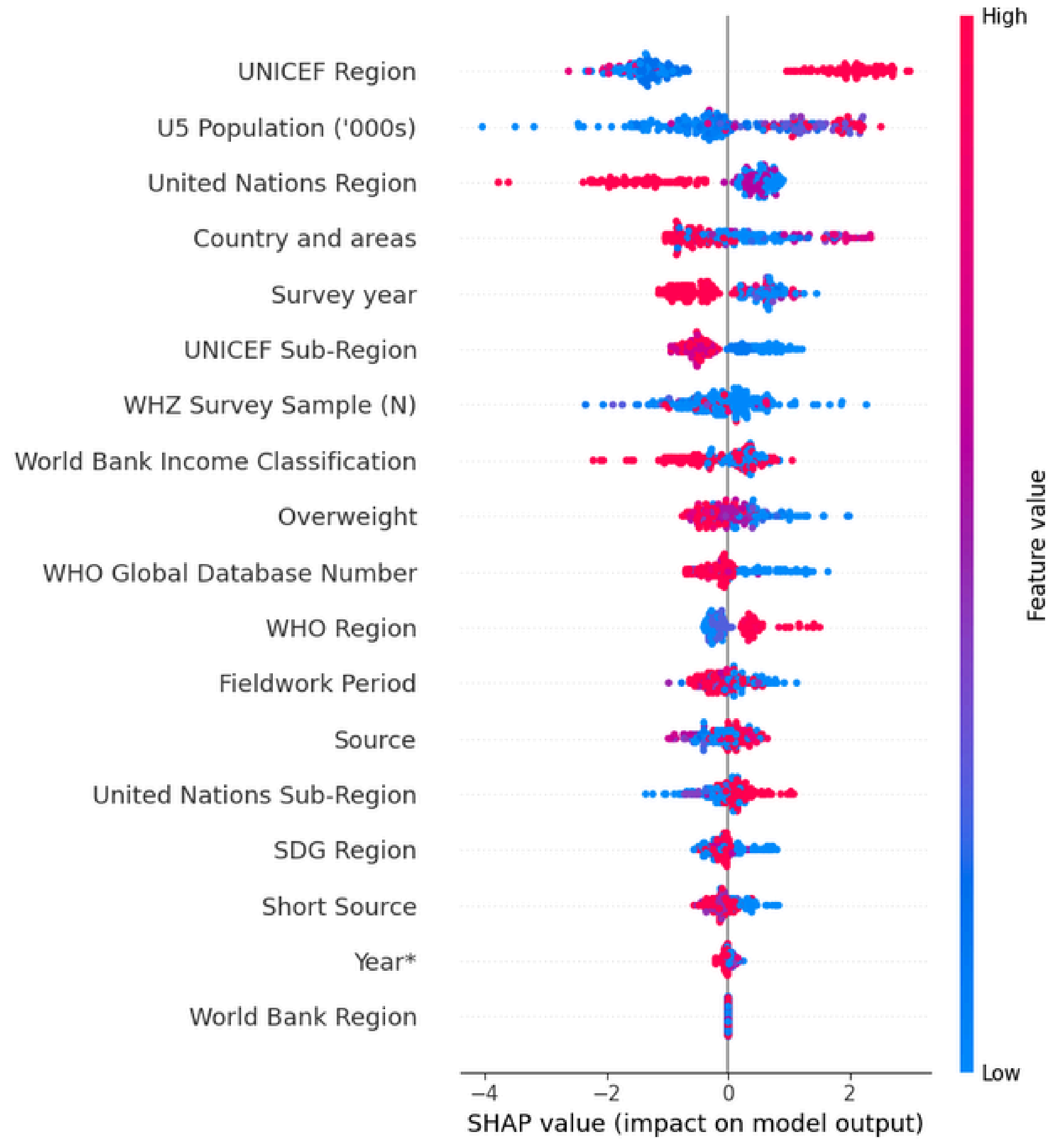
```
print("Classification report is : ",classification_report(y_test,y_svm_pred))

Classification report is :                precision    recall  f1-score   support

           0       0.86      0.78      0.82        96
           1       0.81      0.88      0.85       104

    accuracy                           0.83       200
   macro avg       0.84      0.83      0.83       200
weighted avg       0.84      0.83      0.83       200
```

Machine Learning



xgb

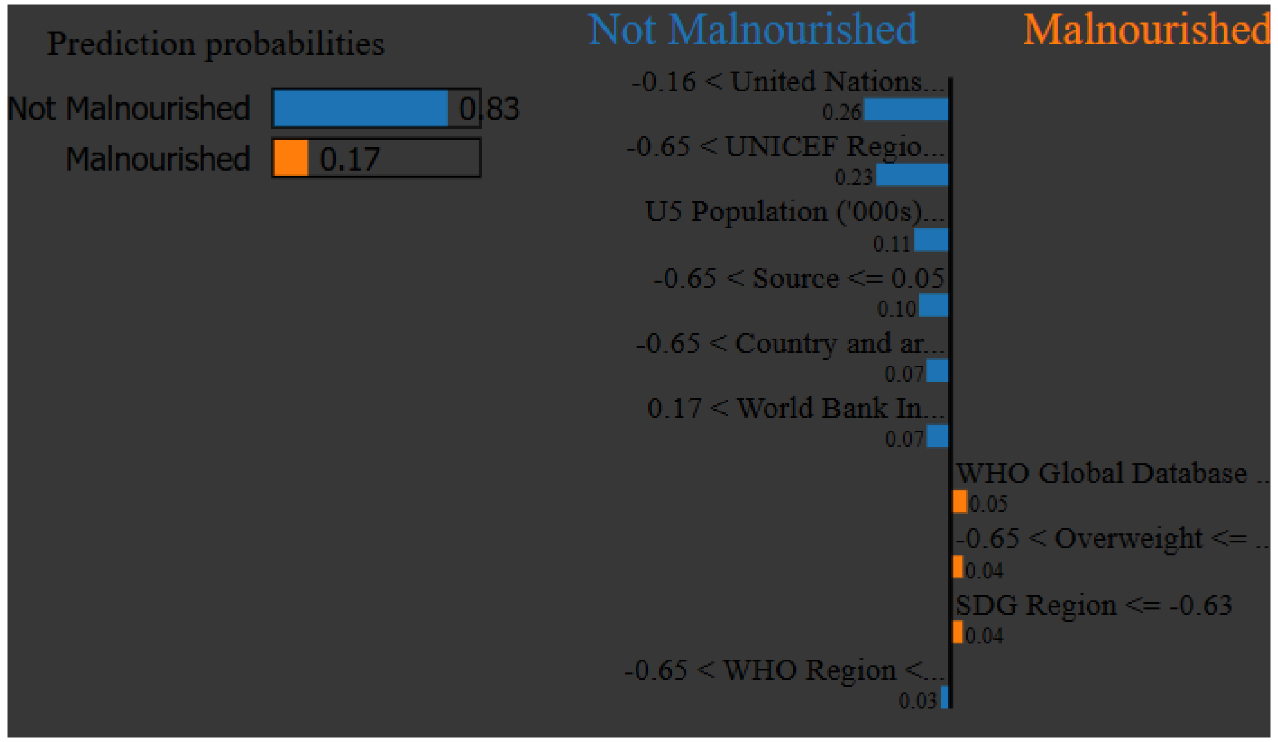
Step 05

Supervised Classification Models

Explainability Agnostic Models For XGBoost Classifier

local agnostic Models:

LIME
SHAD



Feature	Value
United Nations Region	0.19
UNICEF Region	-0.50
U5 Population ('000s)	-0.35
Source	-0.23
Country and areas	-0.29
World Bank Income Classification	0.19
WHO Global Database Number	-0.65
Overweight	-0.53
SDG Region	-0.63
WHO Region	-0.42

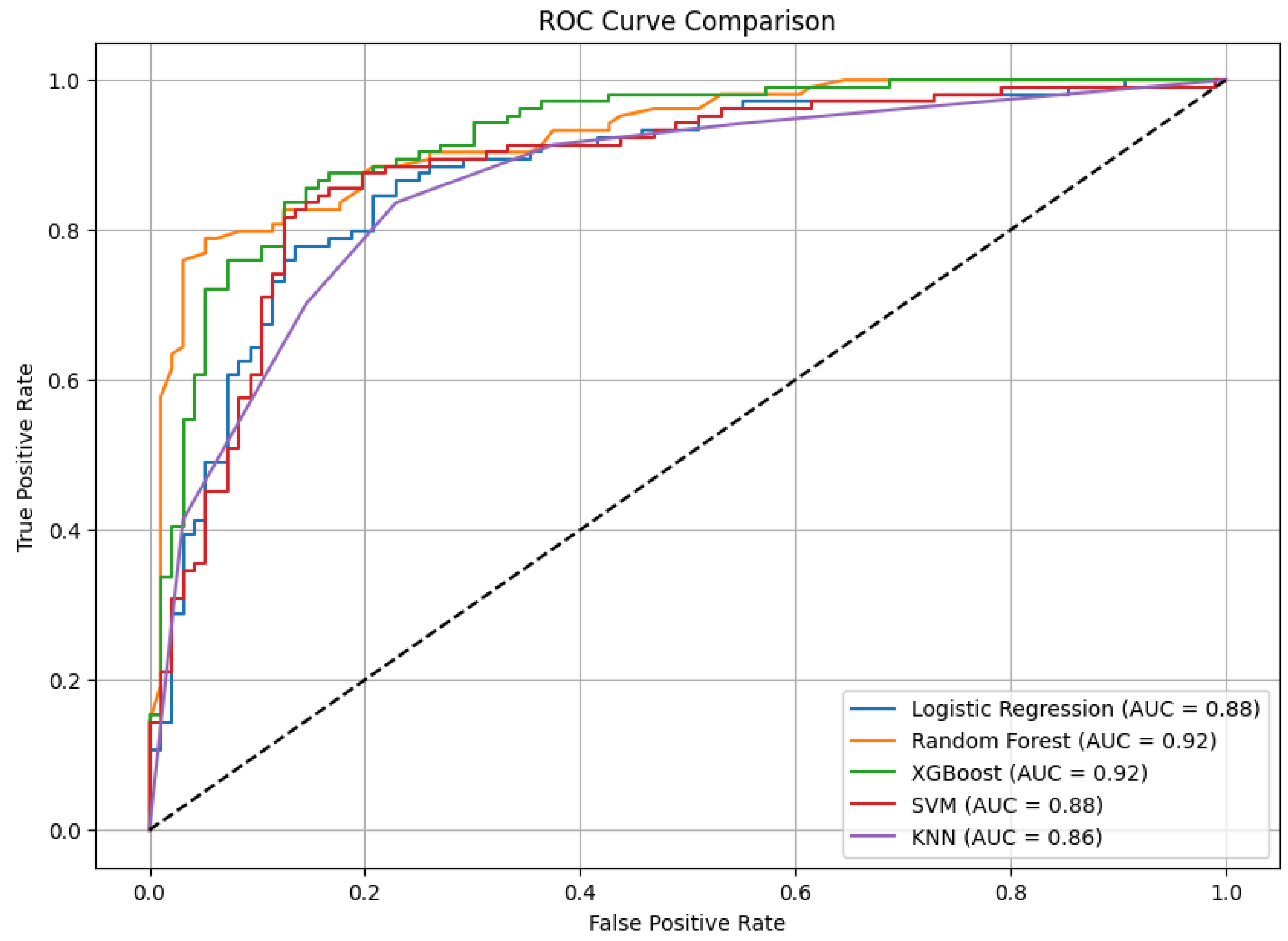
Machine Learning

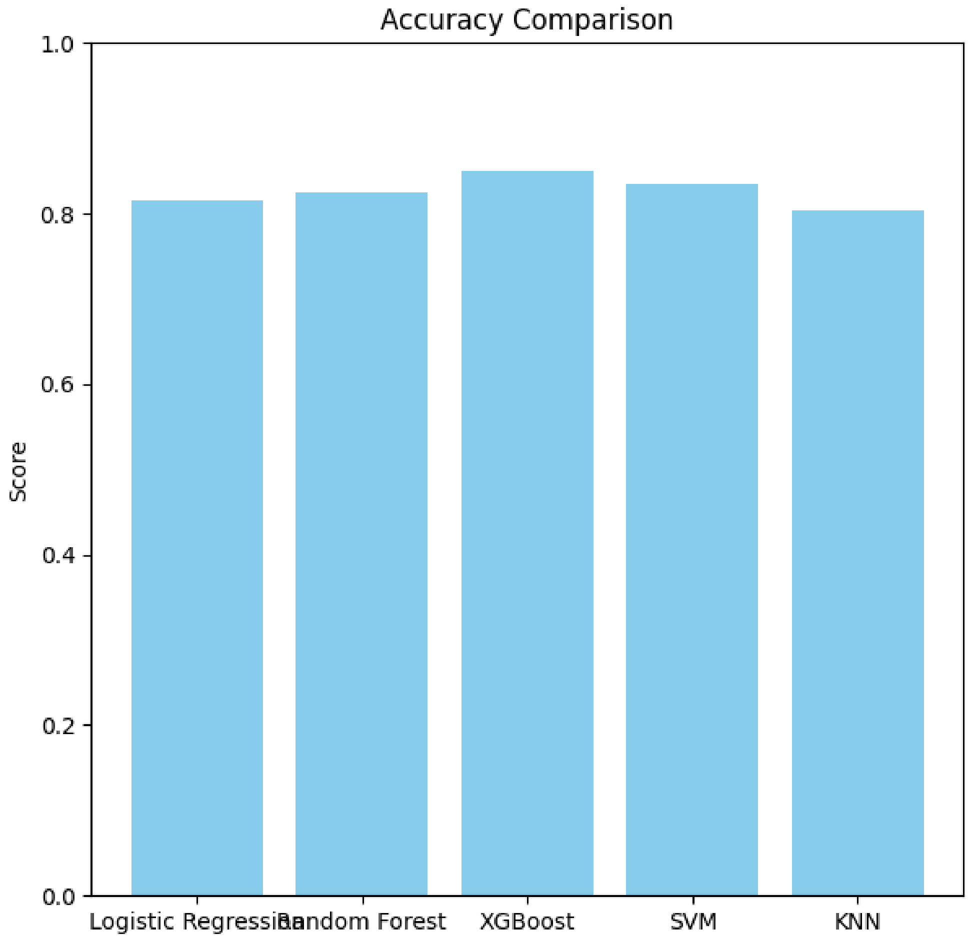
Step 05

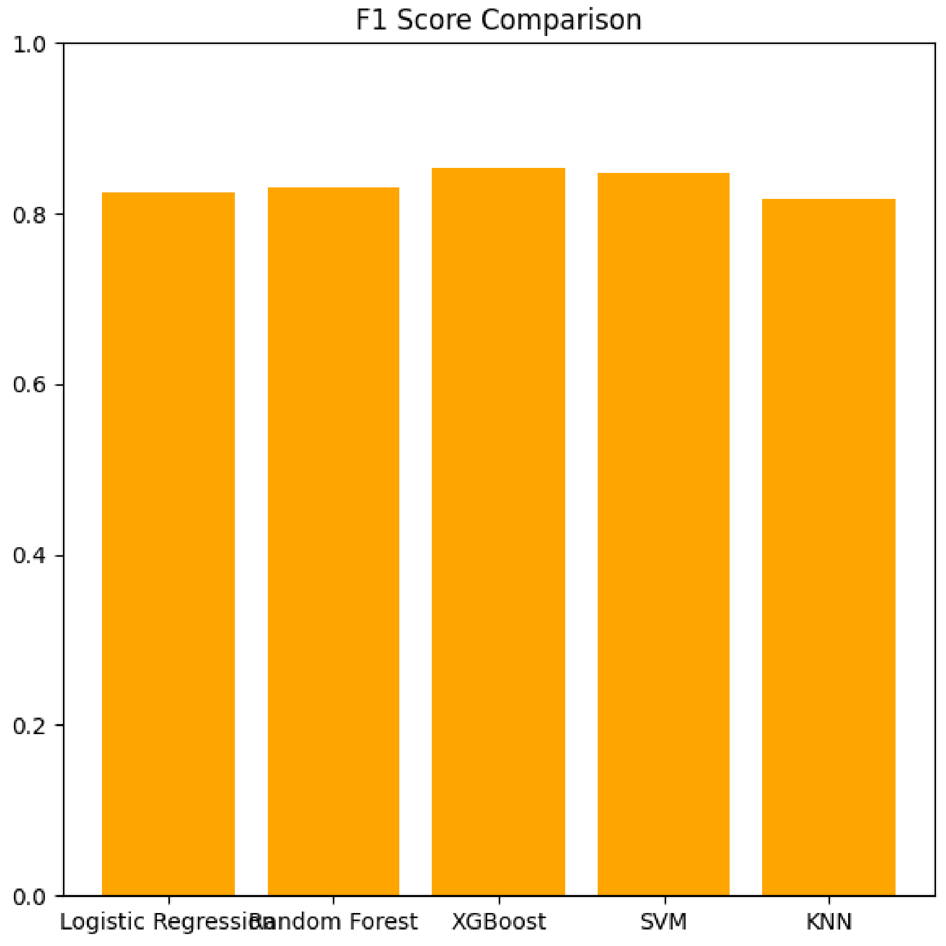
Supervised Classification Models

Evaluation Metrics Used:

- Accuracy
- Classification report
- Confusion matrix
- ROC curve
- F1, Precision, Recall







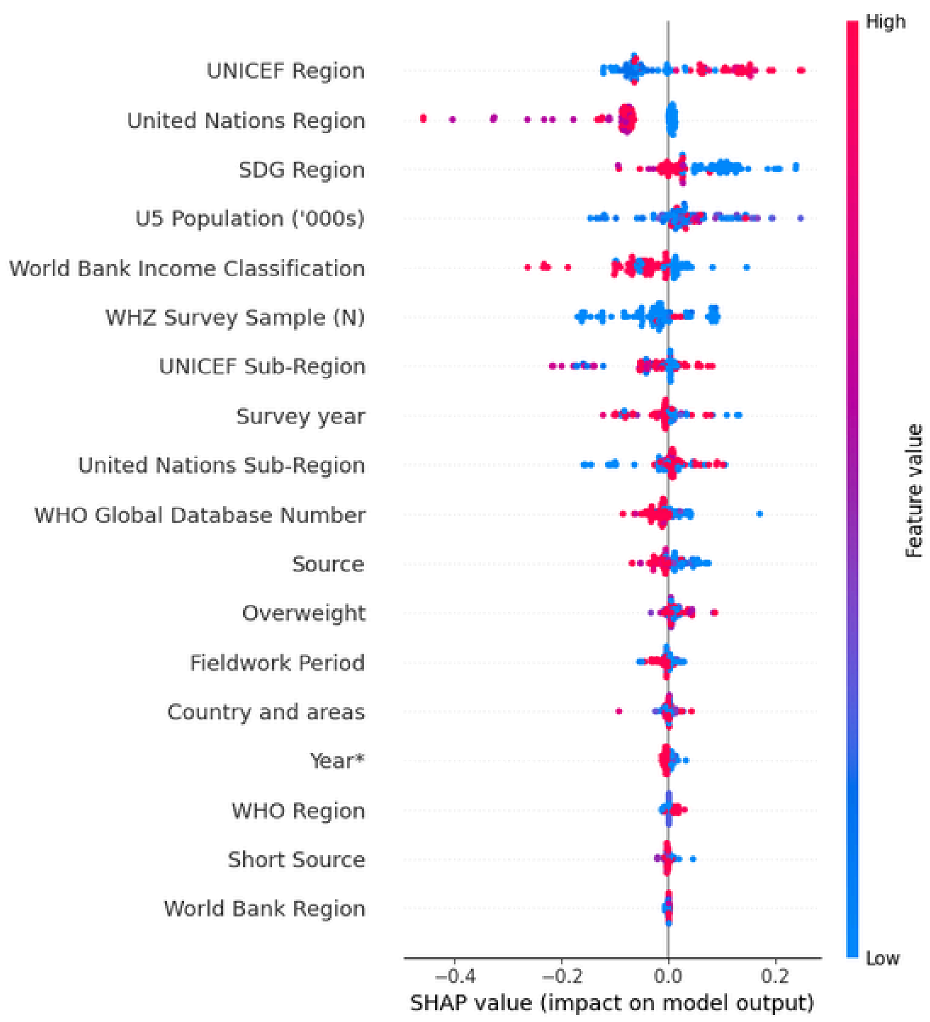
Machine Learning

Step 06

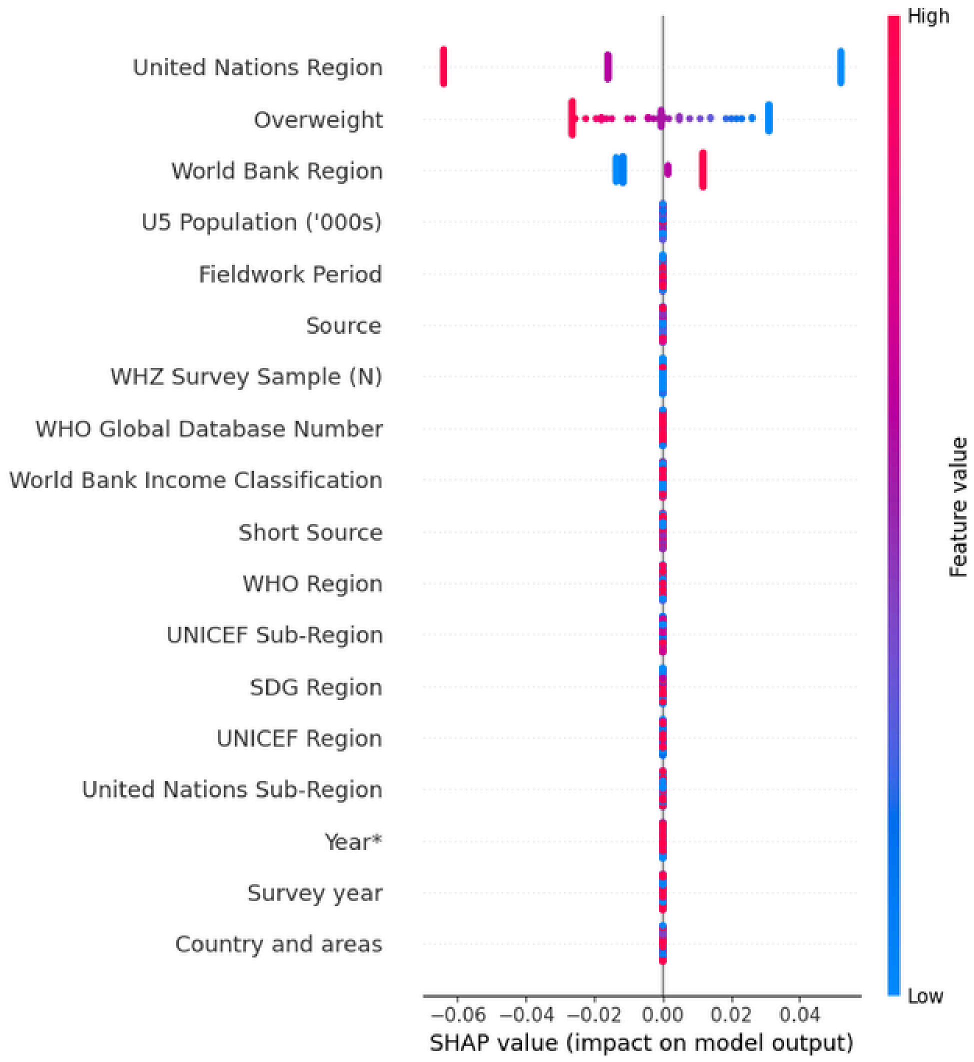
Supervised Regression Models

Models used to predict average malnutrition score:

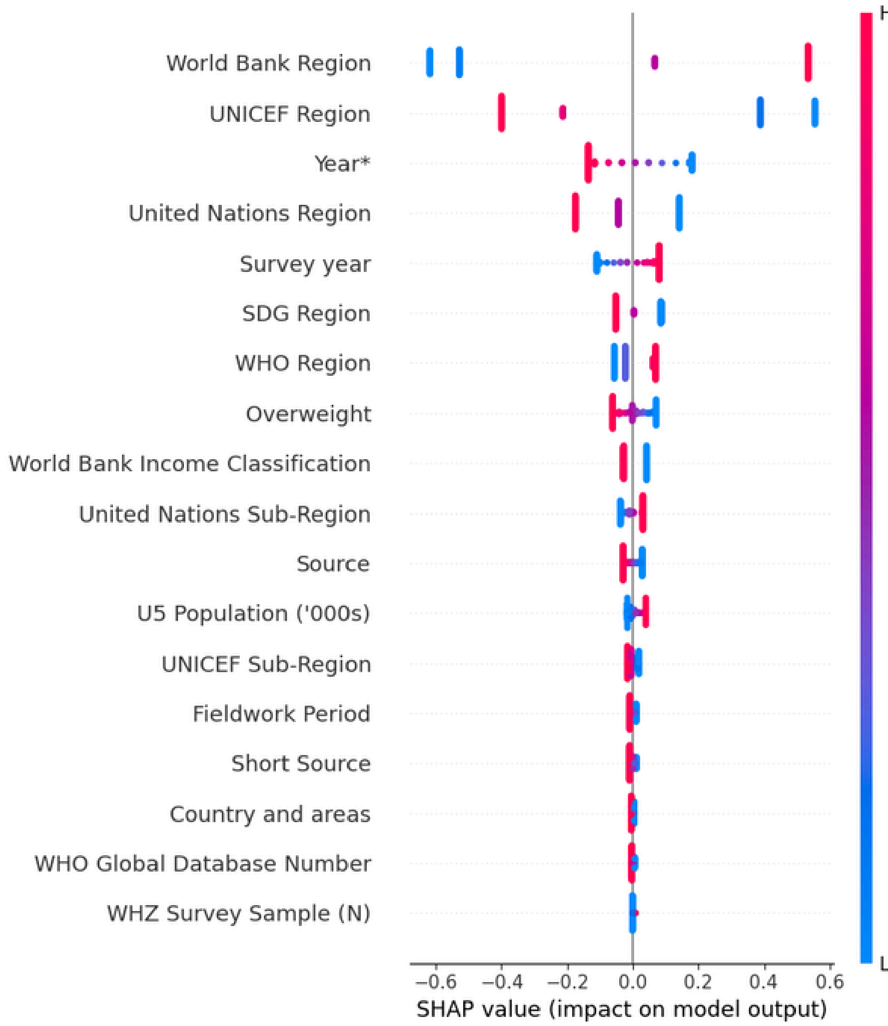
- Linear Regression
- Lasso Regression
- Decision Tree Regressor



Desicion Tree



Lasso



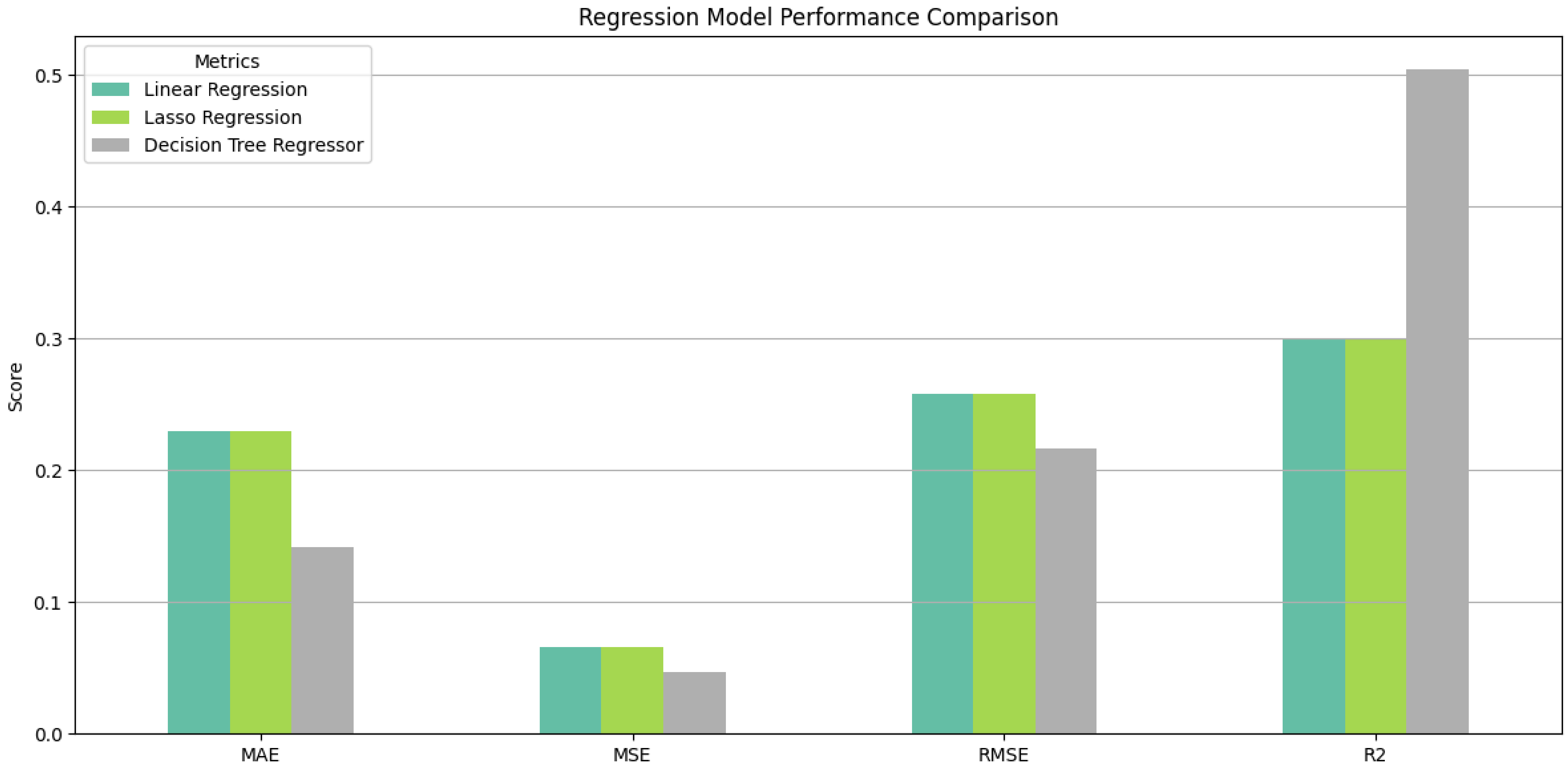
Linear Regression

Machine Learning

Step 06

- Metrics Used:

MAE, MSE, RMSE, R^2



Machine Learning

Step 07

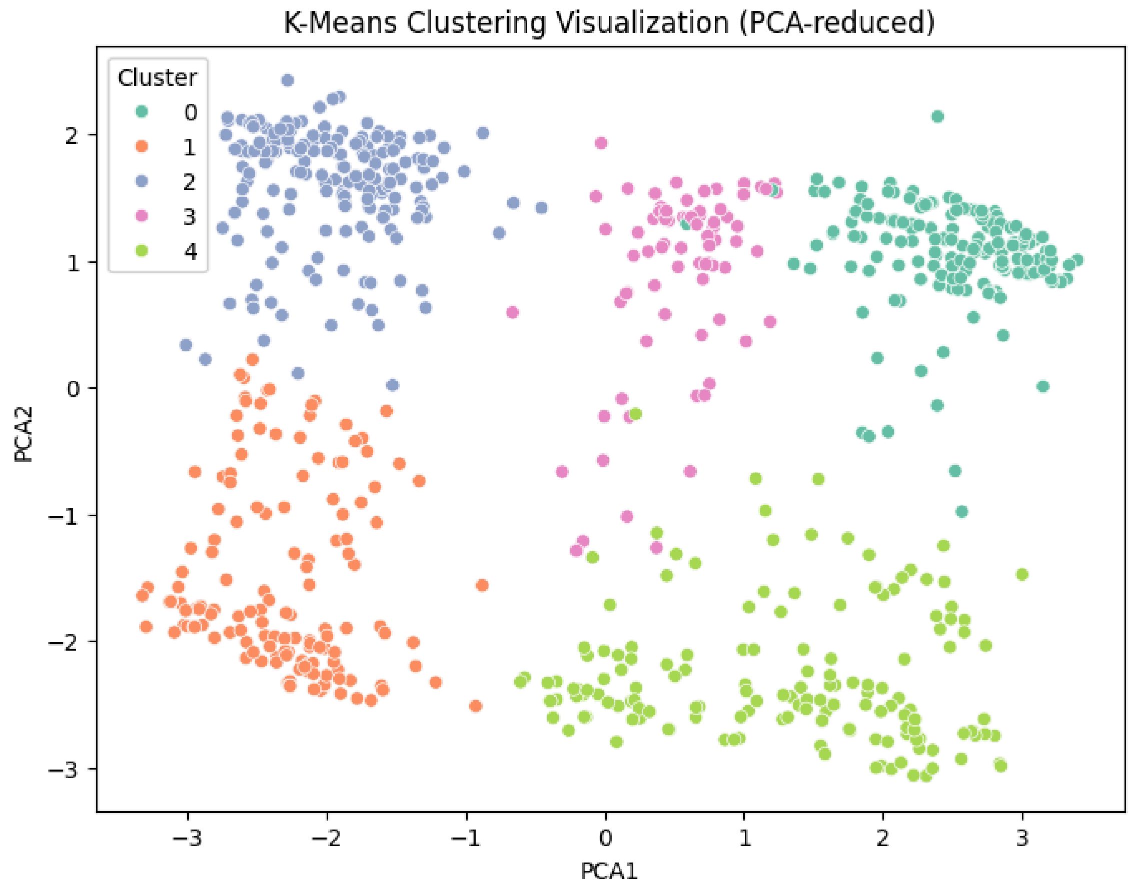
Unsupervised Clustering

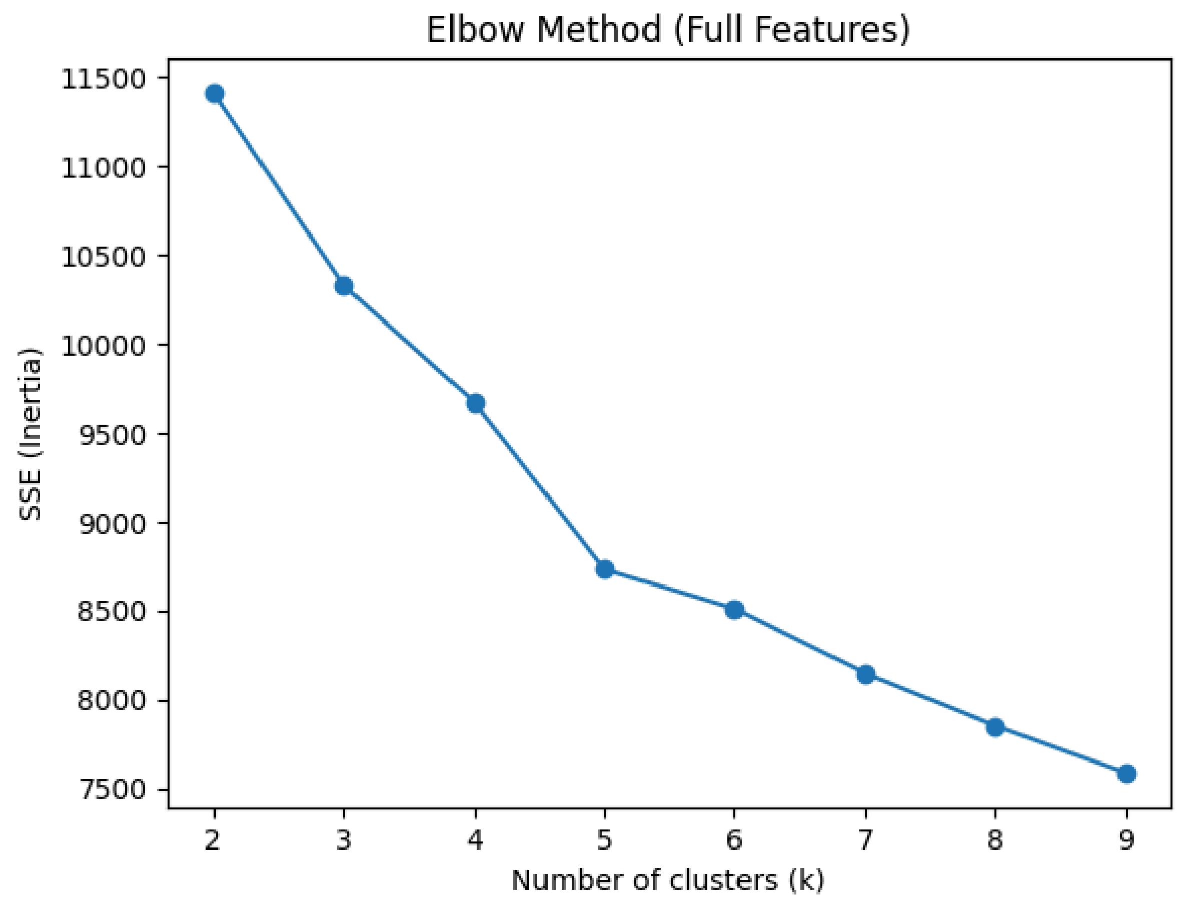
Method: K-Means

Elbow method used to determine optimal k

Clusters visualized using PCA

Silhouette Score used to assess quality

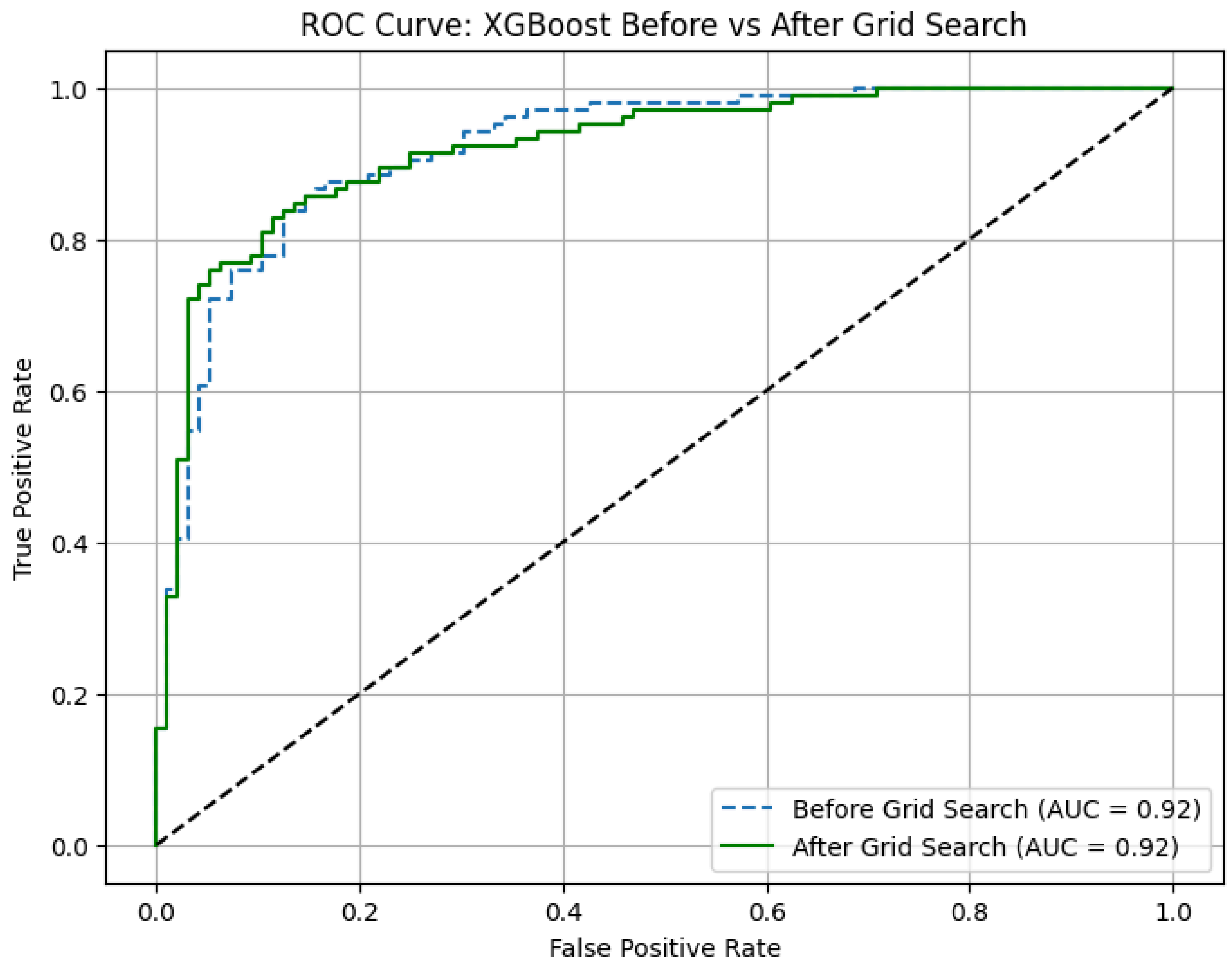


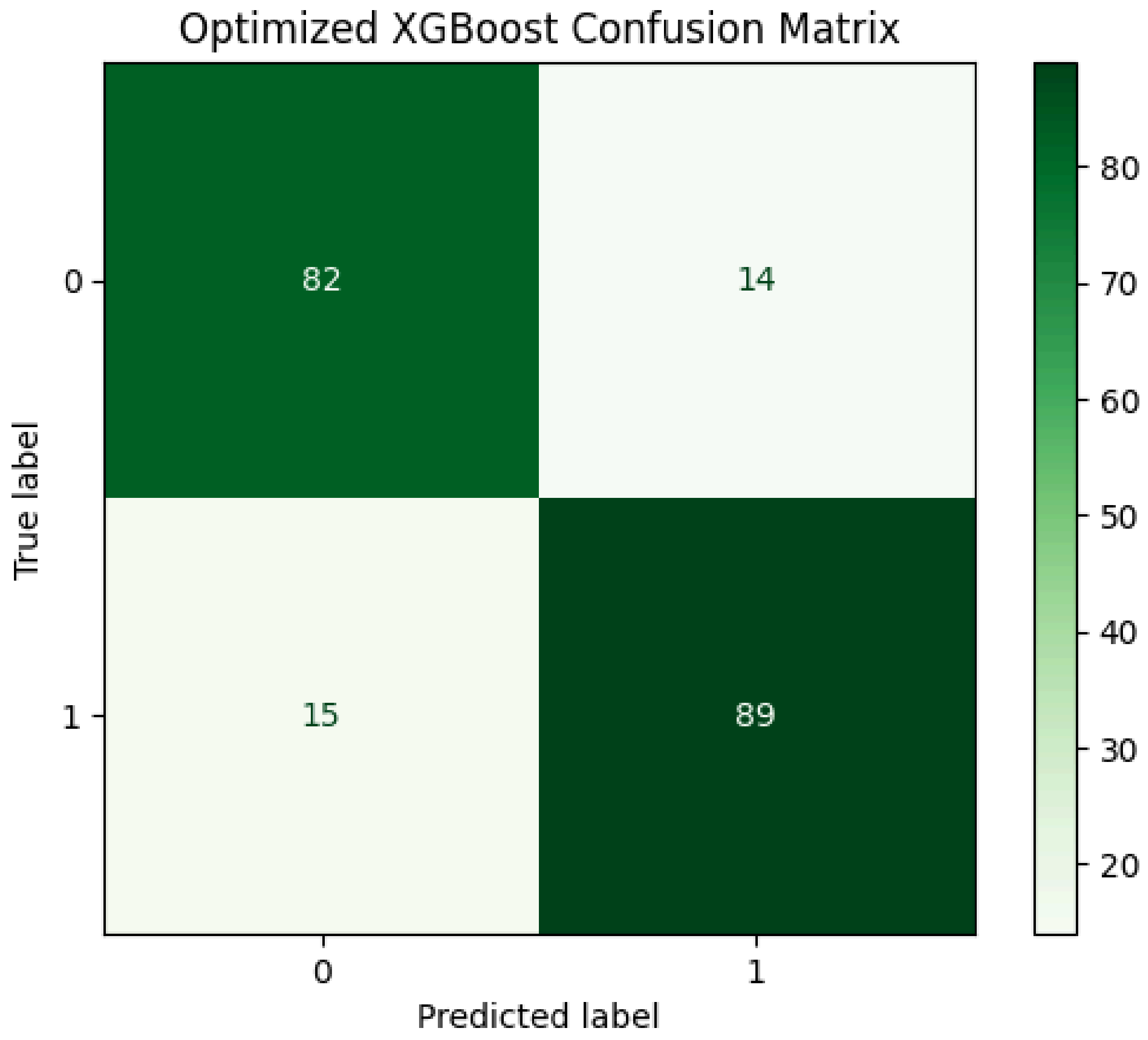


Machine Learning

Step 08

Model Optimization Using Grid Search





Thank You