Title: Predicting Child Mortality Using Machine Learning Techniques through Analysis of Infant Underlying Causes and Maternal Factors Data using Ethiopian Public Health Institute (EPHI)

1. Introduction

Child mortality remains an intimidating challenge globally, with millions of children surrendering to preventable causes before reaching the age of five. Despite notable strides in healthcare and socio-economic development, disparities persist, disproportionately affecting vulnerable populations in low- and middle-income countries. Addressing child mortality requires a multifaceted approach that encompasses understanding the complex interplay of factors contributing to these deaths. This study endeavors to delve into the underlying causes and maternal factors associated with child mortality, aiming to inform targeted interventions and policies to mitigate this pressing public health issue.

1.2.Statement of Problem

The persistence of high child mortality rates underscores the urgency of understanding the factors driving these deaths. Socio-economic inequities, inadequate healthcare access, maternal health disparities, and environmental factors all play pivotal roles in determining a child's survival chances. Despite concerted efforts to curb child mortality, challenges such as limited resources, infrastructure gaps, and systemic barriers impede progress. Identifying the specific drivers of child mortality and elucidating actionable insights from data are critical steps toward implementing effective interventions and achieving sustainable reductions in child mortality rates.

1.3.Methodology

1.3.1. Data Collection

A robust dataset encompassing demographic information, case types, infant underlying causes, and maternal factors was collected by the Ethiopia Public Health Institute (EPHI) with the National Data Management Center from Ethiopia to provide comprehensive insights into child mortality determinants.

13.2. Data Analysis

A. Data cleaning

Employing a blend of data Preprocessing, descriptive analytics, data visualization, advanced machine learning techniques, and model evaluation, the dataset was meticulously analyzed to uncover patterns, correlations, and predictive models associated with child mortality in terms of infant underlying cause and maternal factors. After reading the data I cleaned the data using mean and mode imputation techniques.

```
df= pd.read_csv("CHAMPS.csv",na_values=cc)
```

```
# Count the number of rows and columns
num_rows, num_cols = df.shape
print(f"Number of rows: {num_rows}")
print(f"Number of columns: {num_cols}")
```

```
Number of rows: 444
Number of columns: 381
```

Naming columns

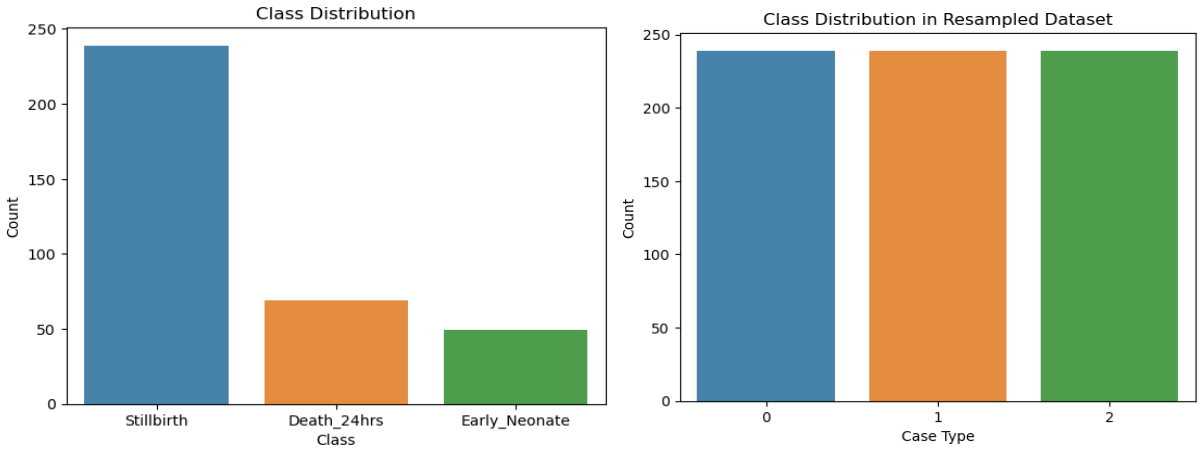
```
# Rename the columns
df = df.rename(columns={
    'dp_013': 'case_type',
     'dp_108':'infant_underlying_cause',
    'dp_118': 'maternal_factors'
})
```

B. Descriptive statistics

```
# Count the frequency of each underlying cause
cause_counts = df['infant_underlying_cause'].value_counts()

# Calculate the proportion of each underlying cause
cause_proportions = cause_counts / cause_counts.sum() * 100

# Display the results
print("Underlying Causes and Their Frequencies:")
print(cause_counts)
print("\nProportion of Each Underlying Cause (%):")
print(cause_proportions)
```

```
Proportion of Each Underlying Cause (%):
infant_underlying_cause
Intrauterine hypoxia                    33.333333
Birth asphyxia                           7.432432
Undetermined                             6.306306
Severe acute malnutrition                5.405405
Craniorachischisis                       3.603604
```

C. Data Balancing



Data balancing before SOMTE and After SMOTE

D. Correlation analysis

```python
# Create a new DataFrame with resampled data
df_resampled = pd.DataFrame(X_smote, columns=['infant_underlying_cause', 'maternal_factors'])
df_resampled['case_type'] = Y_smote

# Calculate correlation matrix
correlation_matrix = df_resampled.corr()

# Plot heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap')
plt.show()
```

E. Classification Modeling

After well data cleaning, data balancing, correlation analysis feature engineering, and train test split, the next steps are developing a machine learning model for predicting child mortality. Using a classification algorithm including Logistic Regression, Support Vector Machine, AdaBoostClassifier, Random Forest Classifier, Gradient Boosting Classifier, and XGBOOST, predictive models were developed to discern the most salient predictors of child mortality. Feature

engineering and selection methodologies were applied to identify key features driving mortality outcomes.

1.4.Results and Discussion

```
# Define a dictionary containing the classifiers
classifiers = {
    'Logistic Regression': LogisticRegression(),
    'Support Vector Machine': SVC(),
    'AdaBoost Classifier': AdaBoostClassifier(),
    'Random Forest Classifier': RandomForestClassifier(),
    'Gradient Boosting Classifier': GradientBoostingClassifier(),
    'XGBoost': XGBClassifier()
}

# Train each classifier and print the accuracy
for name, classifier in classifiers.items():
    classifier.fit(Xsmote_train, ysmote_train)
    accuracy = classifier.score(Xsmote_test, ysmote_test)
    print(f"{name} Accuracy: {accuracy}")
```

```
Logistic Regression Accuracy: 0.375
Support Vector Machine Accuracy: 0.5277777777777778
AdaBoost Classifier Accuracy: 0.5833333333333334
Random Forest Classifier Accuracy: 0.6527777777777778
Gradient Boosting Classifier Accuracy: 0.6805555555555556
XGBoost Accuracy: 0.6666666666666666
```

A. Ranked the features using the developed model

The analysis underscored the pivotal role of 'infant underlying cause' and 'maternal factors' in shaping child mortality outcomes across various classification models. These findings highlight the intricate interplay between maternal health, infant care practices, and child survival rates.

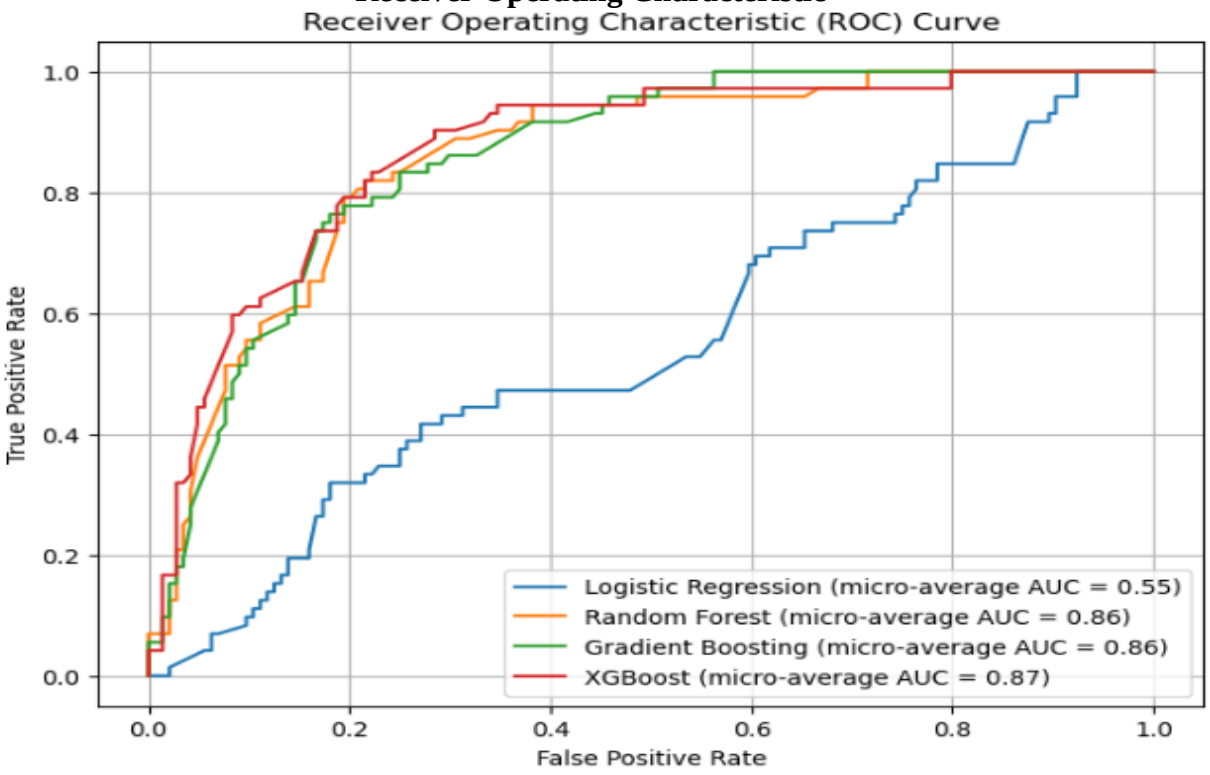
```
Ranked Features:
                        AdaBoost Classifier  Random Forest Classifier  \
infant_underlying_cause                 0.72                   0.68338
maternal_factors                        0.28                   0.31662

                        Gradient Boosting Classifier   XGBoost  \
infant_underlying_cause                      0.787576  0.655602
maternal_factors                             0.212424  0.344398

                        Mean Importance
infant_underlying_cause          0.71164
maternal_factors                 0.28836
```

Any concerned bodies, academicians, and reactionaries should pay attention to these factors to prevent child mortality.

B. Feature Importance Analysis

After training the machine learning models on the prepared data, the next steps are identifying the most important features for child mortality prediction using the developed model.

Receiver Operating Characteristic

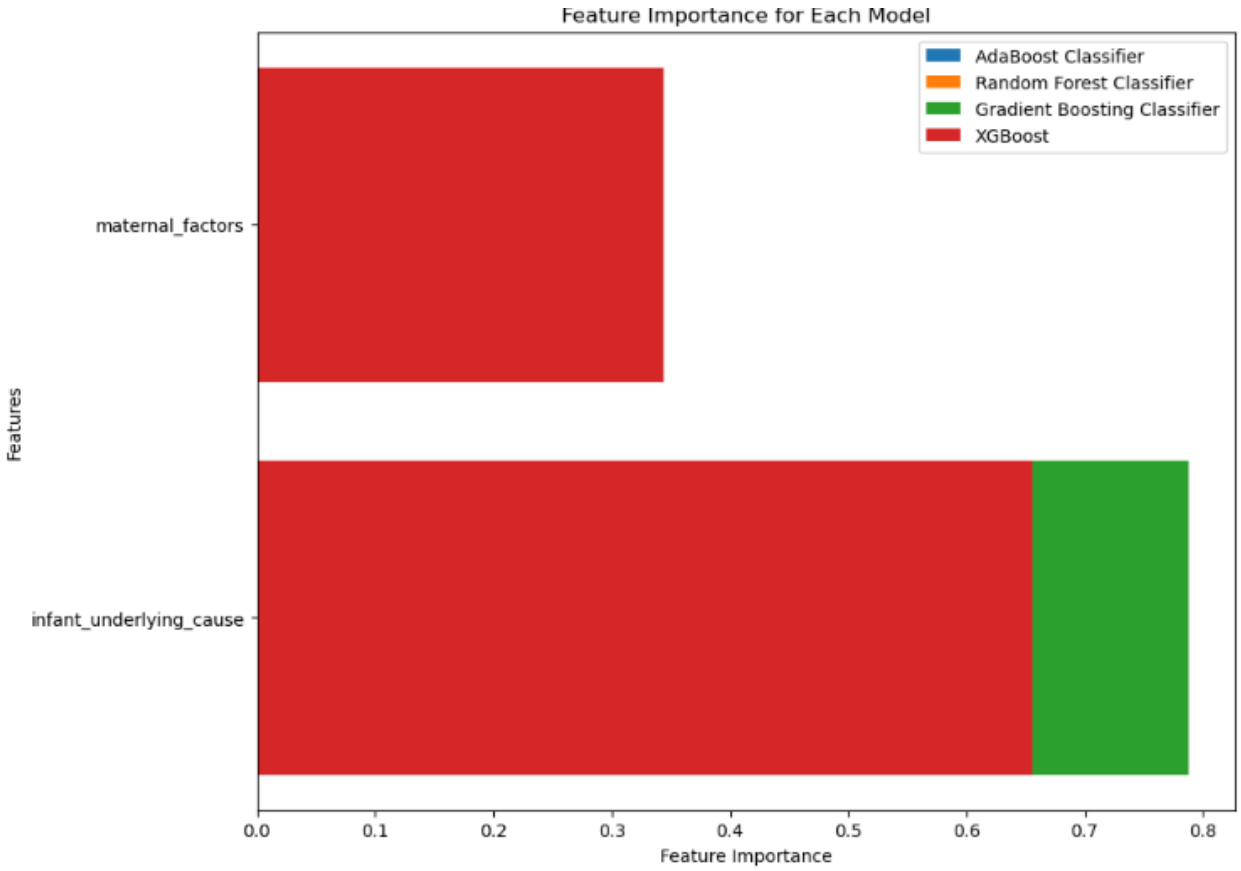


Parameter Tuning

Fine Tune the best parameter of the selected machine learning algorithm with grid search to increase the Performance of the model.

C. Top Underlying Causes and Maternal Factors

Visualization techniques elucidated the top five infant underlying causes and maternal factors contributing to child mortality. Insights gleaned from these visualizations serve as invaluable guideposts for crafting targeted interventions and policy frameworks aimed at averting child deaths.



2. **Conclusions and Future Considerations**

In conclusion, this study explains the nature of child mortality and underscores the imperative of holistic approaches to address this pressing public health concern. By prioritizing investments in maternal and child healthcare, bolstering healthcare infrastructure, and implementing evidence-based interventions, tangible progress can be made in reducing child mortality rates. Future research endeavors should focus on evaluating the efficacy of intervention strategies, monitoring progress toward global targets, and fostering cross-sectoral partnerships to advance child survival agendas on a global scale.

Note

The full implementations of this Project are here.

Note for the following

I am eager to share more details about my skills, past work experiences, and project involvement to demonstrate how I could contribute positively to your team as a Data Science Intern.

Please explore my professional profiles on Github, Medium, and Linkedin for comprehensive insights into my capabilities and achievements.