**Ethiopia Public Health Institute**

**National Data Management Center**

**Data Analytics, Modeling and Visualization Division**

**Part I: Overview**

1. **Objective of the data Challenge**

The objective of this data challenge is to explore the intern’s analytical skill and to imagine their problem-solving skill by providing them with real word problems organizations face on daily basis. In doing so, a data science modeling challenge that assesses students mathematical, programming and problem conceptualization skill has been prepared.

1. **Description of the CHAMPS Dataset**

In this data challenge, Child Health and Mortality Prevention Surveillance (CHAMPS) dataset has been provided to the interns. The purpose of the CHAMPS Surveillance dataset was to collect, analyze, and share data to help identify the causes of child deaths in areas with high child mortality. For the child mortality dataset, there are underlying cause of death from the infant side and also maternal factors contributing to this death. For ease of understanding, variable names and their possible values have been decoded in table 1 below

1. Dataset name: **Champs.csv**

|  |  |  |  |
| --- | --- | --- | --- |
| **S/N** | **Variable/Field Name** | **Field Label** | **Field Attributes (Field Type, Validation, Choices, Calculations, etc.)** |
| 1 | champs\_id | CHAMPS\_ID (Mortality) | text, Required |
| 2 | dp\_013 | **Case Type** | CH00716 :Stillbirth  CH01404 : Death in the first 24 hours  CH01405 : Early Neonate (1 to 6 days)  CH01406 : Late Neonate (7 to 27 days)  CH00718 : Infant (28 days to less than 12 months)  CH00719 : Child (12 months to less than 60 months) |
| 3 | dp\_108 | ***Underlying Cause*** *:*  Underlying cause or factor/ Main condition in fetus or infant  They are labeled "Undetermined" if no cause of death was determined by the panel | text, Required |
| 4 | dp\_118 | *Main maternal disease condition affecting the child/fetus* | text |

**Table 1.1**: Decoded Champs Dataset.

**Part II: Problem Solving Section:**

**Note:** While we understand that you might be comfortable with other data analytic languages and environment, for the sake of this data challenge, the center dictates you to use the python/R language and packages alone.

1. **Preprocessing and EDA**: Based on the given dataset(champs.csv) and the decoded variables in table 1, do the following preprocessing and Exploratory Data Analysis (EDA)
2. Read the dataset
3. How many rows and columns are they in the dataset
4. Enumerate the columns of the dataset
5. Rename the columns. **Example**: rename column **dp\_013** to **case\_type**
6. Rename values.  **Example:** rename **CH00716** to **Stillbirth.** Do the same for others too.
7. Show the proportion of null values in each column.
8. **Descriptive Data analysis**: looking back to the dataset above
9. What are the magnitude and proportion of each of the infant underlying cause for child death?
10. What are the proportion and magnitude of the maternal factors contributing for child death?
11. What are the proportion of the child death by the **case type**
12. **Correlation analysis**: Using correlation or Heat Maps, show how each of the infant under lying conditions and maternal factors are correlated to the top three causes of the child death identified above under 2(A)
13. **Feature engineering:** You are expected to select the top infant underlying causes and maternal factors(features) that would contribute to the top three causes of child death identified under 2(A) above. For this, you need to select the best and likely features. In doing so:
14. Select the classification models LogisticRegression, Support Vector Machine, AdaBoostClassifier, Random Forest Classifier , Gradient Boosting Classifier and XGBOOST and train each on the dataset
15. Import the appropriate package for each of the classification models above
16. Rank the features based on their importance for each of the top underlying causes of child death identified above under 2(A), for each of the classification algorithms under (A )
17. Model evaluation using the proper metrics
18. Import the appropriate evaluation metric packages
19. Using the appropriate n-fold cross validation and out of sample data, select the best preforming model from the candidate models under 4(A)
20. Ensemble the models and see the performance of the combination models on the data
21. Use Accuracy score metrics to evaluate the performance of the models above
22. Plot the AUC and ROC curve on the same graph to visualize and compare the performance of each of the models above
23. **Result Visualization**: Import the appropriate visualization package and:
24. Plot the feature importance in descending order for each of the models using horizontal bar chart
25. Plot the top five infant underlying causes of the child death
26. Plot the top five maternal factors contributing to the child death
27. Plot the child death based on the **case types**