**Machine Learning Project Documentation**

**Group Members:**

1. **Samiullah Gulzar**
2. **Hashmatullah Asady**
3. **Huma Akrami**
4. **Baiqra Muradi**
5. **Fardaws Karimi**

**Model Refinement**

## 1. Overview

The model refinement phase is crucial for enhancing the performance of the machine learning model aimed at early detection of common disease like respiratory infection in Afghanistan. This phase focuses on optimizing the model's accuracy, reducing overfitting, and ensuring that the model generalizes well to unseen data.

## 2. Model Evaluation

Initial results of the Random Forest Classifier were very precise, with a classification report displaying precision, recall, and F1-scores very close to 1. However, there are certain categories where recall has room for improvement. Visualization has been done by creating confusion matrices and feature importance plots that show the areas of improvement.

## 3. Refinement Techniques

To refine the model, several techniques were employed:

* **Hyperparameter Adjustment**: Various hyperparameters of the Random Forest Classifier were tuned to find the optimal settings.
* **Ensemble Methods**: Consideration was given to combining multiple models to improve predictions.
* **Algorithm Exploration**: Logistic Regression was also tested as a comparative baseline to assess improvements.

## 4. Hyperparameter Tuning

Hyperparameter tuning was performed using grid search and cross-validation techniques. Notable adjustments included:

* Increasing the number of estimators in the Random Forest.
* Modifying the maximum depth of the trees to prevent overfitting.

These changes led to improved model performance, as evidenced by higher cross-validation scores.

## 5. Cross-Validation

The cross-validation strategy was enhanced by implementing k-fold cross-validation, which allowed for a more robust evaluation of the model's performance across different subsets of the data. This approach provided a clearer picture of how the model would perform in real-world scenarios.

## 6. Feature Selection

Feature selection methods, such as recursive feature elimination, were applied to identify the most significant features impacting model predictions. This process helped in reducing the dimensionality of the dataset and improving model efficiency without sacrificing accuracy.

# Test Submission

## 1. Overview

The test submission phase involved preparing the trained model for deployment and evaluation on a separate test dataset. This phase ensured that the model was ready for practical application in detecting respiratory infections.

## 2. Data Preparation for Testing

The test dataset was carefully prepared, ensuring that it mirrored the training data in terms of preprocessing steps, such as handling missing values and encoding categorical variables. Special attention was given to maintain consistency in data representation.

## 3. Model Application

The trained Random Forest model was applied to the test dataset using the following code snippet:

# Applying the trained model to the test dataset

y\_test\_pred = rf\_model.predict(X\_test)

This step allowed for the evaluation of the model's performance on new, unseen data.

## 4. Test Metrics

Metrics including accuracy, precision, recall, and F1-score were calculated for the test dataset. Comparisons were made with training and validation metrics to assess any discrepancies and understand the model's generalization capability.

## 5. Model Deployment

Steps towards deploying the model included creating an API for real-time predictions and integrating the model with healthcare systems in Afghanistan. This deployment aims to assist healthcare professionals in early detection and intervention for respiratory infections.

## 6. Code Implementation

Relevant code snippets from both the model refinement and test submission phases are as follows:





A screenshot of a graph

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A white rectangular object with text

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

## Conclusion

The results of the model refinement and testing phases showed significant improvements in the accuracy and reliability of the model. One challenge was that data for common diseases in Afghanistan was not available. We requested data from the DHS website but have not received it yet. Other challenges include addressing class imbalance and ensuring robust data preprocessing. Ultimately, the model achieved a high level of performance, making it a valuable tool for early detection of respiratory infections.

## References:

Scikit-learn documentation (<https://scikit-learn.org/>)

Pandas and NumPy libraries

<https://data.humdata.org/dataset/afghanistan-malnutrition-prevalence>

Momand et al. (2023). "Machine Learning for Predicting Dehydration in Children."

Gupta et al. (2022). "AI-Driven Tuberculosis Prediction for Resource-Constrained Settings."

World Health Organization (WHO). Health Statistics Database.

Afghanistan Demographic and Health Survey (DHS).