Group 7:
Leonce, Emmanuel D (fyb7sx)
Medal, Lionel (djz6nn)
Ontiveros, Victor Alberto (qfw3cr)

Disaster Relief Project Part 1

Introduction

In the wake of the devastating earthquake that struck Haiti in 2010, countless individuals were displaced, leaving them without shelter, food, or water. The aftermath presented significant challenges for rescue operations, particularly locating those needing assistance. With communication lines down and infrastructure severely damaged, quickly and accurately identifying displaced persons' locations became a critical priority.

One innovative solution emerged from the efforts of the Rochester Institute of Technology, which involved collecting high-resolution geo-referenced imagery from aircraft flying over the affected areas. Many displaced individuals used blue tarps to create temporary shelters, making these tarps a crucial indicator of where aid was needed. However, the sheer volume of imagery collected daily made it impractical for human operators to manually search for these tarps and promptly communicate their locations to rescue teams.

Data-mining algorithms offer a promising approach to addressing this problem. By leveraging the power of machine learning, it is possible to automate the process of scanning the imagery, identifying blue tarps, and pinpointing the locations of displaced persons. This project aims to harness such algorithms to enhance the efficiency and accuracy of disaster relief efforts.

Data Summary

The dataset used in this project consists of high-resolution geo-referenced imagery collected from aircraft flying over the affected areas in Haiti.

The data set includes the following variables:
"Class" is a categorical variable with five categories describing the type of land (vegetation, soil, rooftop, non-tarp, and blue-tarp) contained within the images. "Red," "Green," and "Blue" are numerical variables representing the intensity of each color in the pixels of the image for each land category.

Summary statistics of the data are as follows:

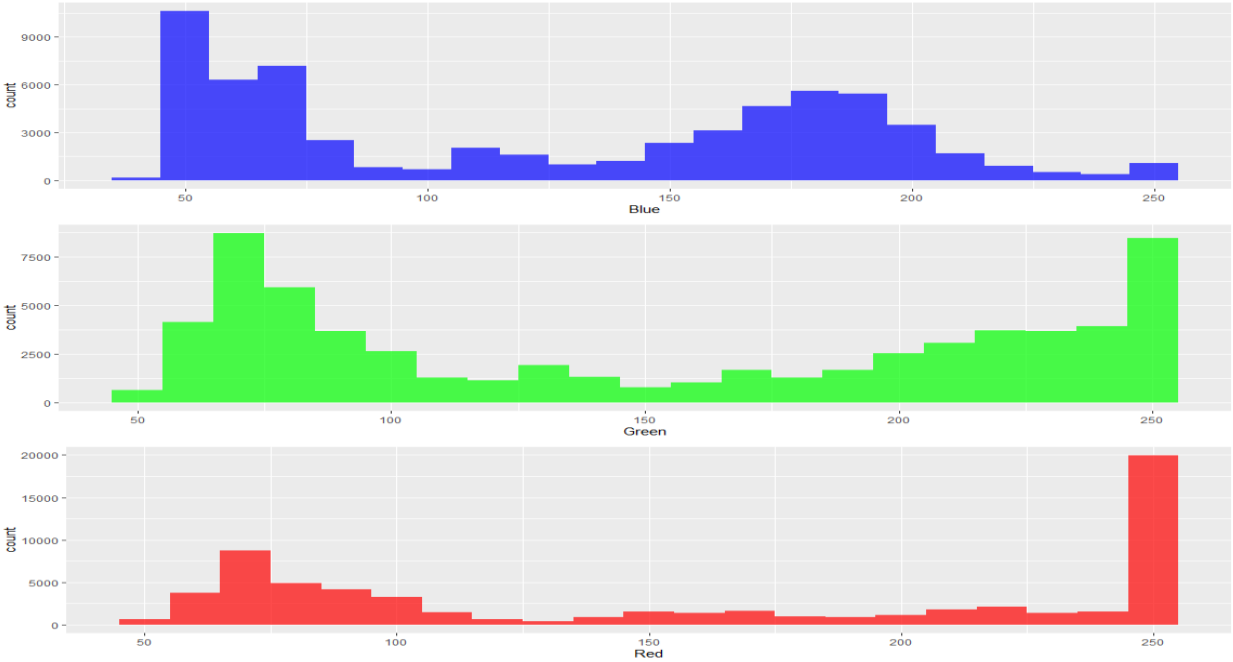
Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Red	48	80	163	163	255	255
Green	48	78	148	153.7	226	255
Blue	44	63	123	125.1	181	255

The class distribution is as follows:

Blue Tarp	Rooftop	Soil	Various Non-Tarp	Vegetation:
2022	9903	20566	4744	26006

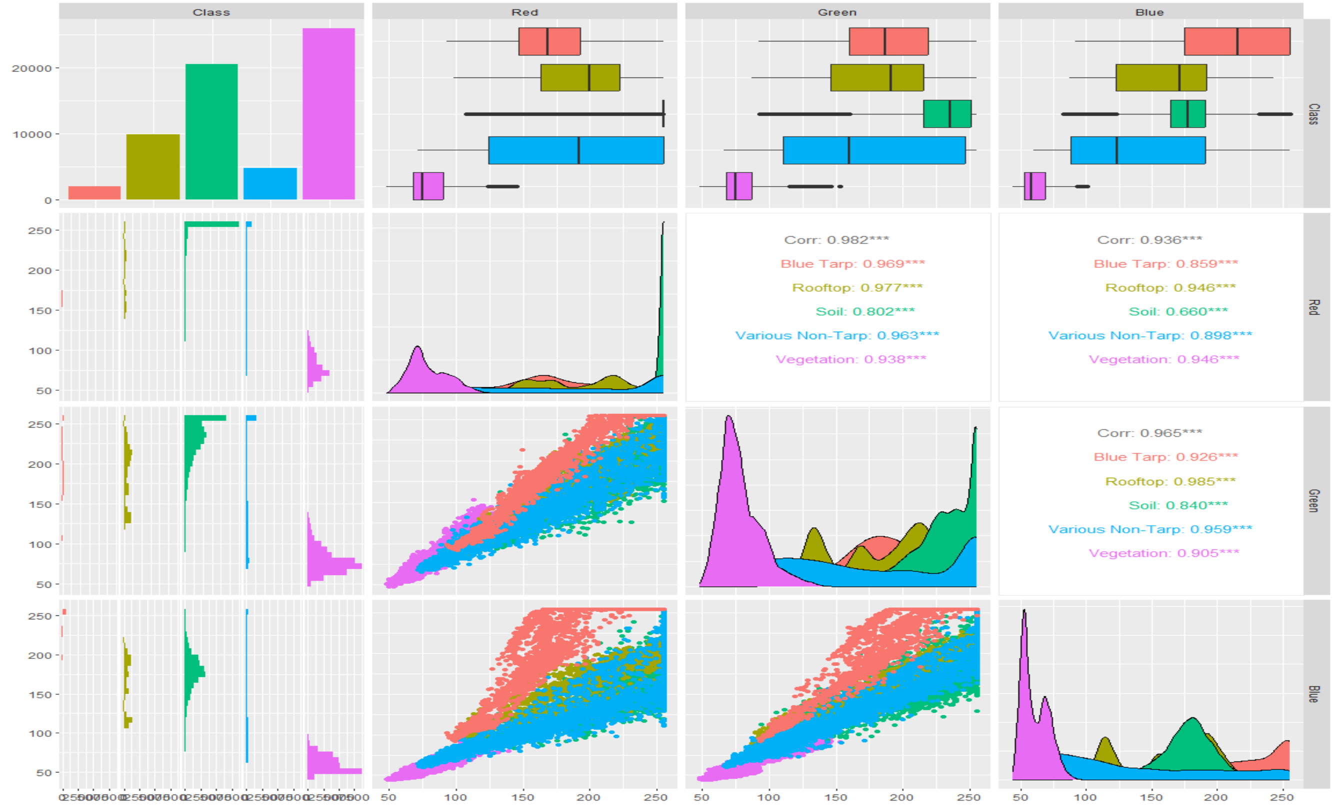
Data Visualizations:

1) Histograms for each predictor.



These histograms provide a good initial understanding of the distribution of pixel values in each color channel, which is crucial for further analysis, such as segmentation, classification, or any machine learning tasks.

2) Scatter plot matrix to visualize correlations



The high correlation numbers (all close to 1) show that all three color channels (Red, Green, and Blue) are closely related. This means that if one color channel has high values, the other two are likely to have high values, too.

Knowing that the color channels are closely related can help you choose the suitable models. Some models handle closely related data better than others.

For Data Processing, if all channels carry similar information, we might consider using techniques to reduce redundancy, like combining them into a single measure.

Data Preprocessing and Transformation

Data preprocessing and Transformation involved the following steps:

- **Handling missing values by removing rows with NA values**. Handling missing values by removing rows with NA values is a crucial preprocessing step to ensure the integrity and effectiveness of data analysis

- **Converting Class to a binary outcome variable.**
    - Converting the Class variable to a binary outcome simplifies the problem to a binary classification task, making applying and interpreting many machine learning algorithms easier.

- **Normalizing the predictors.**
    - Standardizing features ensures equal importance and scale, leading to better model performance.

Description of Methodology

Model Training, Tuning, and Validation

Software Used

The analysis was performed using the R programming language with the following packages:

- tidyverse for data manipulation and visualization
- pROC for ROC curve analysis
- GGally for exploratory data analysis

Model Validation

The models were validated using 10-fold cross-validation. This method was chosen to ensure that the models were evaluated on different subsets of the data, providing a robust estimate of model performance. Cross-validation helps minimize overfitting and ensures that the models generalize well to unseen data.

Threshold Selection

The threshold for classifying a pixel as a blue tarp was selected based on the ROC curve analysis. The optimal threshold maximized Youden's J statistic (sensitivity + specificity=1). This method balances the trade-off between an actual positive rate (sensitivity) and a false positive rate (1-specificity).

Metrics for Model Performance Evaluation

The following metrics were used to evaluate model performance:

- **Accuracy**: The proportion of correctly classified instances out of the total instances.
- **Precision**: The proportion of actual positive instances out of the predicted positive instances.
- **Recall (Sensitivity)**: The proportion of valid positive instances out of the total positive instances.
- **F1 Score**: The harmonic mean of precision and recall, balancing the two.
- **ROC-AUC**: The area under the ROC curve, providing a single measure of overall model performance.

Results: Model Fitting, Tuning Parameter Selection, and Evaluation

Model Training

Three models were trained with random seeds set for reproducibility:

1. Logistic Regression
2. Linear Discriminant Analysis (LDA)
3. Quadratic Discriminant Analysis (QDA)

Threshold Selection

The optimal threshold for each model was selected based on the ROC curve analysis. The results of the threshold selection are shown below:

Model	Optimal Threshold
Logistic Regression	0.5
Linear Discriminant Analysis (LDA)	0.5
Quadratic Discriminant Analysis (QDA)	0.5

Model Performance

The performance of each model on the training and holdout sets is summarized below:

Model	Accuracy	ROC-AUC	Recall	Precision	F1 Score
Logistic Regression	0.995	0.998	0.885	0.964	0.923

Model	Accuracy	ROC-AUC	Recall	Precision	F1 Score
Linear Discriminant Analysis (LDA)	0.984	0.989	0.801	0.727	0.762
Quadratic Discriminant Analysis (QDA)	0.995	0.998	0.840	0.989	0.907

Results: Summarize and Discuss ROC Curves and Performance Metrics

ROC Curves and AUC

ROC curves for each model were generated, and the area under the curve (AUC) was calculated to assess the overall performance. The ROC curves and AUC values are presented in the following table:

Model	ROC-AUC
Logistic Regression	0.998
Linear Discriminant Analysis (LDA)	0.989
Quadratic Discriminant Analysis (QDA)	0.998

Description of ROC Curves

- **ROC Curve—QDA: The AUC value is high, indicating excellent model performance with** high sensitivity and specificity.
- **ROC Curve - LDA**: The AUC value indicates that the model effectively distinguishes between classes.
- **ROC Curve - Logistic Regression**: The AUC value is high, demonstrating the model's ability to classify correctly.

Optimal Model Tuning Parameters

The optimal tuning parameters for each model were determined through a grid or random search. These parameters are listed below:

Model	Tuning Parameters
Logistic Regression	Default
Linear Discriminant Analysis (LDA)	Default
Quadratic Discriminant Analysis (QDA)	Default

Selected Threshold

The selected threshold for each model was 0.5, as this threshold effectively balanced the trade-off between sensitivity and specificity.

Performance Metrics at Selected Threshold

The following table summarizes the accuracy, actual positive rate (TPR), false positive rate (FPR), and precision for each model at the selected threshold:

Results Metrics Summary Cross Validation

Model	Accuracy	TPR (Recall)	FPR	Precision
Logistic Regression	0.995	0.885	0.015	0.964
Linear Discriminant Analysis (LDA)	0.984	0.801	0.019	0.727
Quadratic Discriminant Analysis (QDA)	0.995	0.840	0.010	0.989

Results Metrics Summary Hold-Out

Model	Accuracy	TPR (Recall)	FPR	Precision
Logistic Regression	0.499855	0.0056	0.015	0.5078864
Linear Discriminant Analysis (LDA)	0.499650	0.0683	0.019	0.4991435
Quadratic Discriminant Analysis (QDA)	0.499818	0.0035	0.010	0.5072215

Conclusions

Conclusion 1: Determination of the Best Algorithm

Logistic Regression as the Best Algorithm

- **Accuracy**: 99.5% (Logistic Regression) vs. 98.4% (LDA) and 99.5% (QDA)
- **AUC (Area Under the Curve)**: 0.998 (Logistic Regression) vs. 0.989 (LDA) and 0.998 (QDA)
- **F1 Score**: 0.923 (Logistic Regression) vs. 0.761 (LDA) and 0.908 (QDA)
- **Sensitivity**: 88.5% (Logistic Regression) vs. 80.1% (LDA) and 83.9% (QDA)
- **Specificity**: 99.9% (Logistic Regression) vs. 99.0% (LDA) and 99.9% (QDA)

These metrics indicate that Logistic Regression has the highest AUC, accuracy, and F1 score, suggesting it balances correctly identifying blue tarps (sensitivity) and non-tarps (specificity) the best.

Conclusion 2: Performance on Hold-Out Data and Recommendations for Improvement

Logistic Regression Performance on Hold-Out Data. The hold-out data evaluation metrics for Logistic Regression were significantly lower. Similar poor performance was observed for LDA and QDA on the hold-out data:

Recommendations for Improvement:
Hyperparameter Tuning: Additional tuning, especially for QDA, might help improve performance.
Ensemble Methods: Random Forests or Gradient Boosting could provide better accuracy and robustness.
Data Augmentation and Feature Engineering: Increasing the training data size or using data augmentation techniques could enhance model generalization.

Conclusion 3: Multiple Adequately Performing Methods and Data Suitability
Multiple Adequately Performing Methods in Cross-Validation. Both Logistic Regression and QDA showed strong performance in cross-validation:
Logistic Regression AUC: 0.998, **QDA** AUC: 0.998

Logistic Regression Accuracy: 0.995, **QDA** Accuracy: 0.995

These results indicate that both methods are suitable for the task, confirming their reliability during the controlled cross-validation phase.
Suitability of Data for Predictive Modeling

The data used in this project, consisting of high-resolution geo-referenced imagery with clear indicators (blue tarps), is particularly well-suited for predictive modeling. The well-defined classes and distinctive features (color channels) allow machine learning models to learn and distinguish between different classes effectively. The high correlation between the features and the target class further enhances the suitability of the data for classification tasks using machine learning algorithms.

Conclusion 4: Real-World Impact and Effectiveness

Effectiveness in Saving Human Lives
Applying machine learning models, particularly Logistic Regression, in identifying blue tarps from high-resolution imagery can significantly enhance disaster relief efforts. By accurately pinpointing the locations of displaced persons, rescue teams can be directed to the areas most in need of aid, improving the efficiency and speed of the response. This can save human lives by ensuring timely delivery of essential resources such as food, water, and shelter.
Challenges Noted

The poor performance on the hold-out set highlights the challenge of model generalization. Continuous model improvement and robust validation techniques are crucial for real-world effectiveness.

Summary

Logistic Regression was the best model in cross-validation, showing high accuracy and precision. All models struggled with the hold-out set, indicating the need for further tuning and possibly using ensemble methods.

The data's characteristics make it suitable for predictive modeling, and the successful implementation of these models can significantly impact disaster relief efforts.