Disaster Relief Project: Part 1

DS-6030 – Group 12

Ampoyo, Rheyan (xcs5hg)

Belle, Camisha L. (fbv2ub)

Cuevas Rodriguez, Dalila (zfd9aj)

Tran, Karmen Victoria (aqq2ex)

March 3, 2025

Table of Contents

Introduction 3
Exploratory Data Analysis 3
Training Dataset 3
Holdout Dataset 6
Methodology 10
Results 12
Training Dataset Analysis 12
Holdout Dataset Analysis 14
Conclusions 16
Conclusion 1: Logistic Regression Outperforms LDA and QDA 16
Conclusion 2: Enhance Logistic Regression with Regularization and Ensemble Models 17
Conclusion 3: Geospatial Intelligence for Faster, Smarter Aid Delivery 18

Introduction

On January 12, 2010, a catastrophic magnitude 7.0 earthquake struck Haiti near the capital, Port-au-Prince. It caused widespread devastation, killing an estimated 230,000 people, injuring 300,000 more, and displacing over 1.5 million.[i] The earthquake destroyed much of the capital's infrastructure, including homes, hospitals, and government buildings. The disaster overwhelmed Haiti's limited resources prompting an international humanitarian response.

In the wake of this disaster, rescue workers faced significant operational difficulties delivering aid due to destroyed communications, impassable roads, and the vastness of the affected area. To aid rescue efforts, a team from the Rochester Institute of Technology (RIT) collected high-resolution, geo-referenced imagery of the region. The imagery revealed that displaced individuals, whose homes were destroyed, were using blue tarps to construct temporary shelters.

These blue tarps had the potential to serve as crucial indicators of displaced persons' locations. But locating them amidst thousands of images collected daily was an overwhelming task. The solution to this challenge was in data-mining algorithms, which could search and analyze the images more efficiently and accurately than manual methods.

This project applies statistical techniques to address this historical data-mining challenge. The objective of this project is to apply the techniques taught in the University of Virginia's Data Science 6030 course (Statistical Learning) to the imagery data from the Haiti relief effort. The primary aim is to identify the most effective method to accurately locate displaced persons.

The project will be completed in two parts, with Part I focused on setting up a modeling workflow to build and evaluate three classification models—Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and logistic regression—using cross-validation and performance metrics. This report summarizes Group 12's work to complete Part 1 of the Disaster Relief Project in partial completion of the requirements of the course.

Exploratory Data Analysis

Training Dataset

The model's dense training data, based on the RIT Haiti Earthquake Lidar dataset, has 63,241 observations and four (4) variables with no missing values. The four (4) variables include: one categorical variable ("Class") and three numeric variables ("Red", "Green", and "Blue").

Each observation is associated with a specific class (Blue Tarp, Rooftop, Soil, Various Non-Tarp, Vegetation) and corresponding RGB (Red, Green, Blue) color value. The numeric variables ("Red", "Green", and "Blue") represent the intensity values for each respective color

channel, with each value ranging from 0 to 255. The "Class" represents different types of objects or materials in the captured images. See below for density plots of the RGB values by Class.

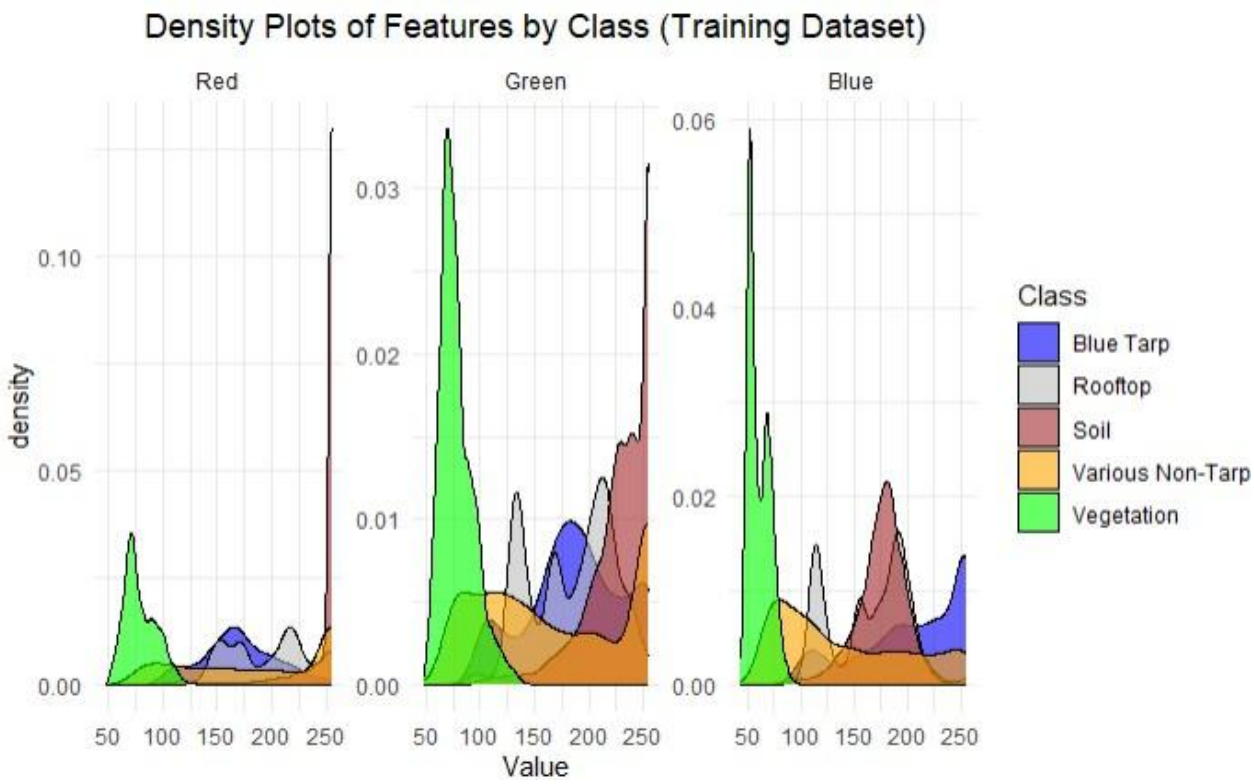

Figure 1: Density Plots by Class (Training Data) - density plots of the Red, Green, and Blue (RGB) features, grouped by the original class labels in the training dataset. Displays strong variability in the feature data distributions across classes of the response variable "Class".

The features exhibit a wide range of values across different classes, indicating strong variability in the feature distributions. This suggests that the dataset contains meaningful differences in spectral reflectance patterns. The RGB features may serve as strong discriminators between classes making the dataset well-suited for predictive modeling. The presence of non-overlapping regions in the distributions indicates that a classification model should be able to effectively learn patterns that differentiate between the categories, potentially leading to high predictive accuracy.

Group 12 discussed whether to convert the RGB color model to an alternative such as HSL or HEX. RGB corresponds to the way photographic images capture and render color, making it more relevant for analysis involving raw imagery. HEX is a compact hexadecimal representation of RGB values, but less suited for data analysis. As a base-16 system, performing mathematical operations with HEX requires conversion to a different format, such as RGB or HSL. HSL, which defines colors based on hue, saturation, and lightness, can be more intuitive for certain analyses, but conversion was unnecessary to achieve valid results. RGB was deemed sufficient for this exercise and leads to better interpretability.[ii]

To determine the location of blue tarps in RIT's images, Group 12 transformed the "Class" response variable into a bivariate output, "Blue_Tarp". The new variable was assigned a value of "Blue Tarp" for observations with a class of "Blue Tarp" (n = 2,022; 3.2%) with all other values coded as "Not Blue Tarp" (n = 61,219; 96.8%). T-tests indicate that all three features ("Blue", "Green", and "Red") have statistically significant differences between the binary "Blue_Tarp" classes. The p-values for all features are extremely small ($p < 0.05$). The color values remain highly effective in distinguishing between the categories after reclassifying "Class" into the binary "Blue_Tarp" variable as displayed in the charts below.

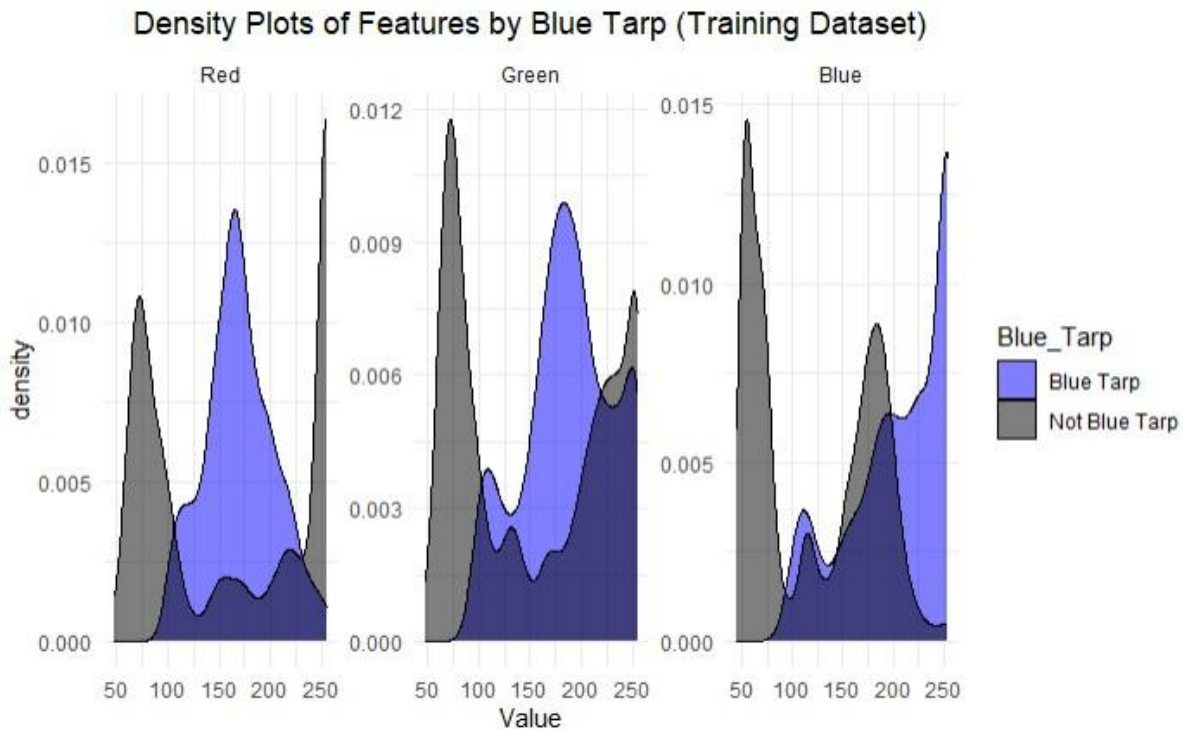

Figure 2: Density Plots by Blue Tarp (Training Data) - Like the previous plot, but with the transformed output variable which is grouped into two categories: "Blue Tarp" and "Not Blue Tarp."

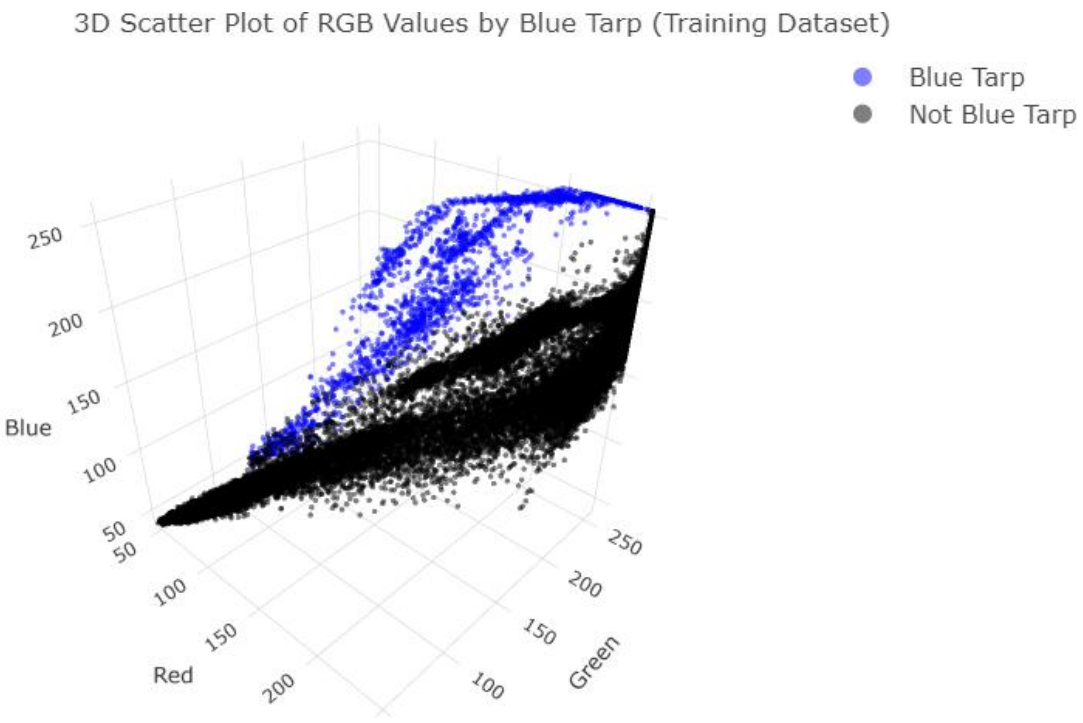

Figure 3: 3D Scatter Plot of RGB (Training Data) - In R, this is an interactive 3D scatter plot showing the distribution of "Blue Tarp" and "Not Blue Tarp" based on Red, Green, and Blue values. It is a visual depiction of the separation between the classes.

In summary, the dataset is well-suited for tasks involving color-based classification analysis. The “Blue-Tarp” variable, derived from the transformed "Class" variable, provides the output labels, while the "Red", `Green`, and `Blue` variables serve as the features. The variability in the color channels suggests that the dataset captures a diverse range of color intensities, which is useful for training machine learning models to distinguish between different classes.

Holdout Dataset

The holdout dataset was comprised of eight individual text and three JPG files. Each text file represents different subsets of data associated with various conditions. Specifically, the following files were included:

- Text Data Files
  - orthovnir078_ROI_NON_Blue_Tarps.txt
  - orthovnir078_ROI_Blue_Tarps.txt
  - orthovnir069_ROI_NOT_Blue_Tarps.txt
  - orthovnir069_ROI_Blue_Tarps.txt
  - orthovnir067_ROI_NOT_Blue_Tarps.txt
  - orthovnir067_ROI_Blue_Tarps.txt
  - orthovnir067_ROI_Blue_Tarps_data.txt
  - orthovnir057_ROI_NON_Blue_Tarps.txt
- JPG Image Files

- orthovnir071_makeshift_villiage1.jpg
- orthovnir071_makeshift_villiage2.jpg
- orthovnir078_makeshift_villiage1.jpg

Each Text Data Files in the holdout dataset describes an analysis of a Region of Interest (ROI) within a remote sensing dataset. The dataset consists of an image with dimensions, indicating a large-scale image. The specific ROI is analyzed, and the region is visually marked with the RGB color value. The data for each pixel in the ROI is organized in columns, with each row representing one pixel. The columns provide several key pieces of information: ID, a unique identifier for each pixel, X and Y (denotes the pixel's coordinates in the image's 2D place), Map X and Map Y (convert the pixel coordinates to a mapped geographic reference system, Lat and Lon (indicating the real-world location of the pixel, and B1, B2, B3 columns that contain the intensity values for the pixel corresponding to the spectral bands such as Red, Green, and Blue.

The holdout dataset, which contains 2,008,623 observations, also includes a categorical variable called "Class". Like the training dataset, this data does not contain any missing values. Unlike the training dataset, this variable already classifies the data as either "Blue Tarp" (18,926 observations representing only 0.94% of the data) or "Not Blue Tarp" (1,989,697).

The data processing procedure for the holdout dataset involved several key steps. First, for each individual dataset (e.g., Data_057_NonTarp, Data_067_Tarp2, etc.), the columns of interest were isolated, specifically columns 8, 9, and 10. These columns were renamed as follows: column 8 was designated "Red," column 9 was renamed "Green," and column 10 was renamed "Blue."

After exploring all possible RGB combinations using an RGB calculator, we determined that interpreting column 8 as Red, column 9 as Green, and column 10 as Blue produced scatterplots with colors that closely matched the provided example images. Any other combination resulted in unnatural hues of green, purple, or turquoise, indicating an incorrect mapping of the color channels.

It is important to note that the position of the figures in the plot was not a factor in our analysis; this approach was solely used to verify the correct order of the RGB channels.

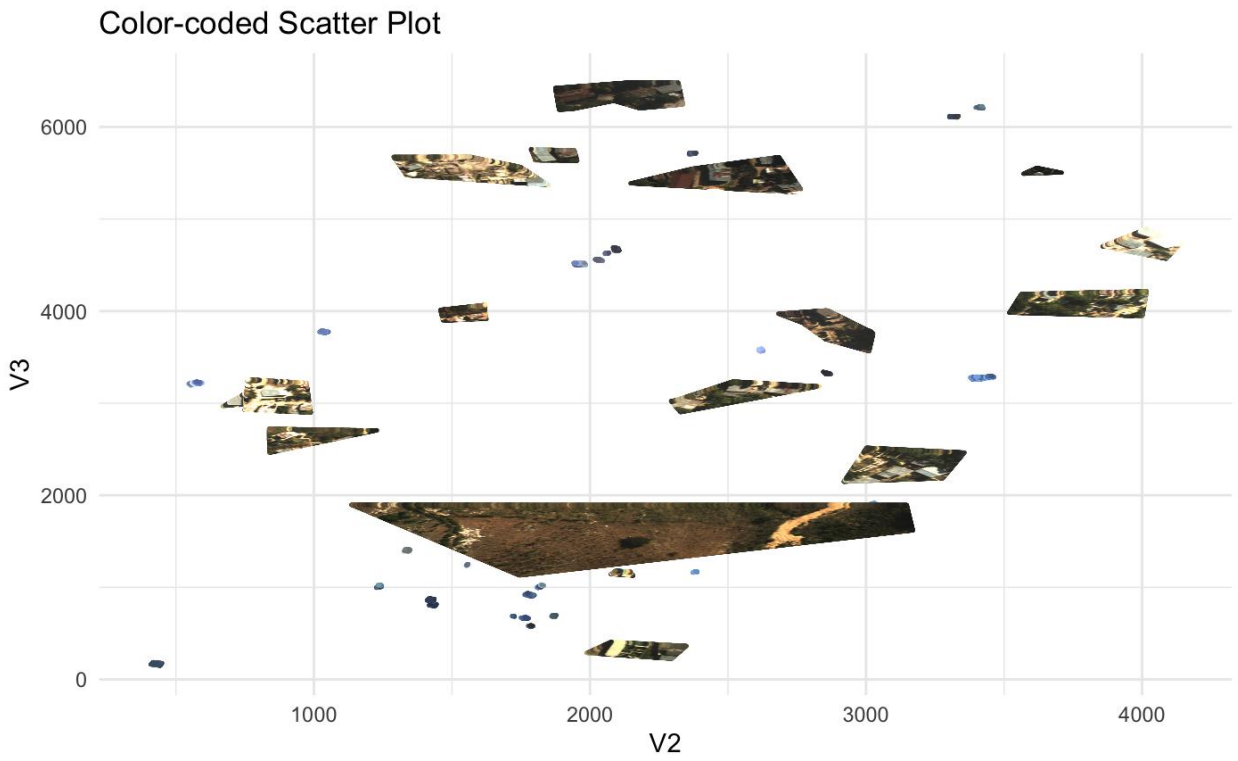

Figure 4: Color-coded Scatter Plot of an Image File (Holdout Data) - A visual depiction of Group 12's exploration of all possible RGB combinations using an RGB calculator. The position of the figures in the plot was not a factor in our analysis; this approach was solely used to verify the correct order of the RGB channels.

Comparison of boxplots from each dataset confirms the columns were assigned correctly. While the whiskers and interquartile ranges for the datasets vary due, in part, to more outliers in the holdout dataset, the relationship between the median values for images with blue tarps versus those without is consistent between the training and holdout datasets.

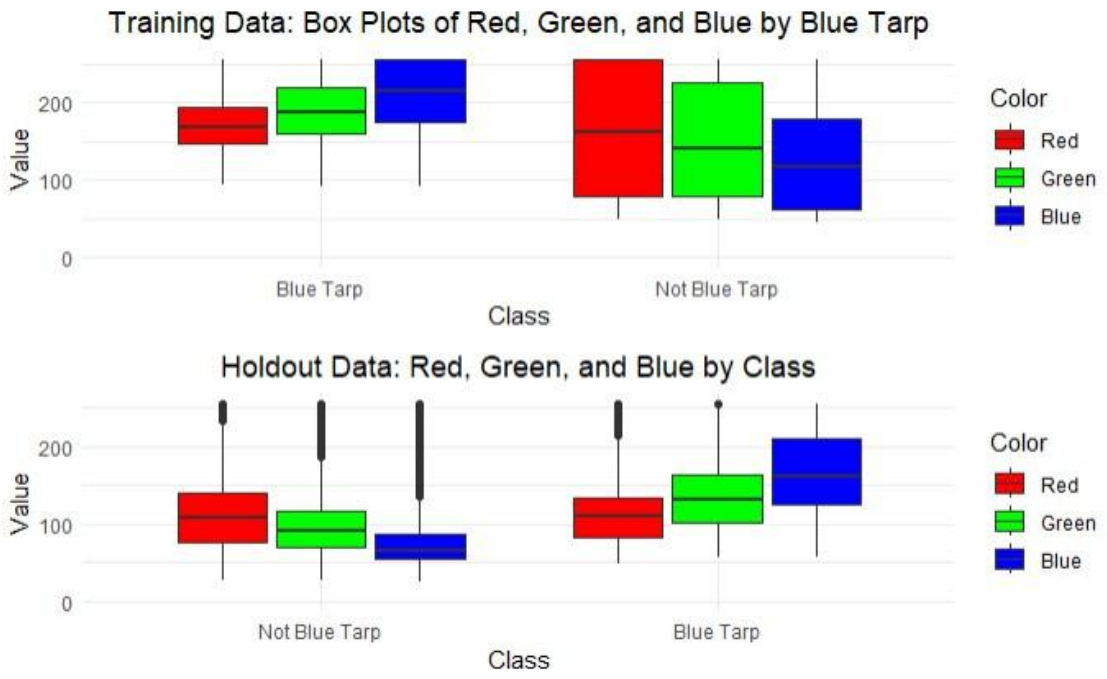

Figure 5: Box Plots of RGB Values (Training and Holdout Data) - Comparative boxplots generated to compare comparing "Blue Tarp" and "Not Blue Tarp" for both the training and the holdout datasets.

The RGB boxplots for the holdout dataset exhibit significant differences from the training dataset, suggesting variations in the distribution of color intensities between the two sets. However, upon closer inspection of the 3D scatter plot generated from a random sample of the holdout data, Group 12 observes a clear separation in the RGB values between images containing blue tarps and those without. This distinct clustering indicates that, despite the differences in density distributions, the fundamental color-based distinction between the two classes still exists.

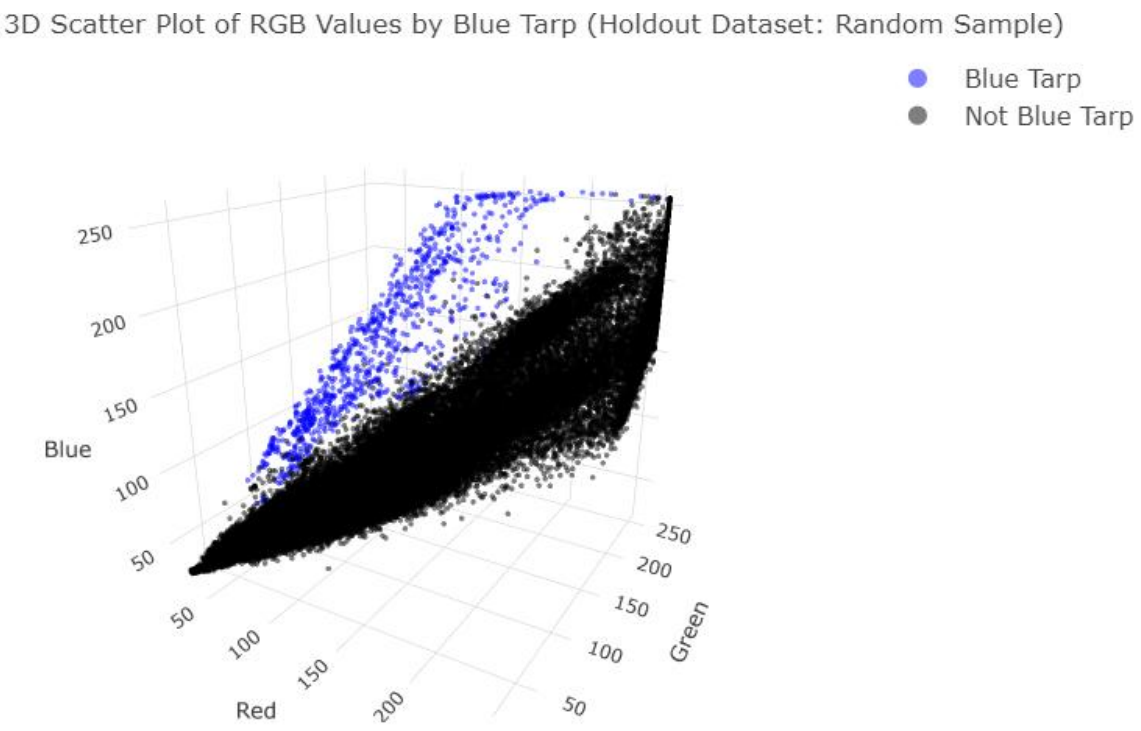

Figure 6: 3D Scatter Plot of RGB (Holdout Data, Sampled) - In R, this chart renders as an interactive 3D scatter plot of the holdout dataset, using a random sample of 100,000 points to reduce complexity. Within this paper, the image represents a visual depiction that the separation between the classes exists within the holdout dataset as it did in the training dataset.

A new categorical column, "Blue_Tarp," was subsequently introduced to the data. This column served as a factor, indicating whether the dataset corresponded to a blue tarp ("Blue Tarp") or a non-tarp condition ("Not Blue Tarp").

Following the modification of column names and the addition of the "Blue_Tarp" factor, the datasets were merged using the bind_rows function. This process resulted in the consolidation of the individual datasets (Data_057_NonTarp, Data_067_Tarp, Data_067_Tarp2, Data_067_NonTarp, Data_069_Tarp, Data_069_NonTarp, Data_078_Tarp, Data_078_NonTarp) into a single, unified data frame, hereafter referred to as holdout data.

Finally, the levels of the "Blue_Tarp" factor within the holdout data frame were modified by reordering them, with "Not Blue Tarp" (non-tarp) being set as the reference level. This adjustment ensures that "Not Blue Tarp" serves as the baseline category in any subsequent analyses involving the "Blue_Tarp" variable.

Methodology

For this project, Group 12 selected R as the primary software for statistical computing, data visualization, and model development. The tidymodels framework was used to streamline the

modeling workflow, ensuring reproducibility and scalability. Additionally, Microsoft Word was used to draft the final report, enabling efficient collaboration and documentation of findings.

To ensure the reliability and generalizability of the models, 10-fold cross-validation was used on the training dataset, allowing the models to be trained and tested on different subsets of the data to mitigate overfitting. The final model was evaluated on a holdout dataset. This validation strategy ensures that performance metrics reflect the model's ability to classify new, unseen data effectively. In addition, given that the dataset is highly imbalanced, group 12 evaluated whether to employ imbalance handling techniques.

The 10-fold stratified cross-validation approach is well-suited for this project because it ensures effective use of the dataset while addressing the significant class imbalance. Each instance is used for both training and validation at different stages, providing a more reliable estimate of model performance. Stratification ensures that each fold maintains the same proportion of the minority and majority classes, minimizing bias and improving generalizability.

The choice of classification threshold significantly impacts model performance by influencing the trade-off between sensitivity and specificity. In Group 12's study, the threshold was optimized based on the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC). The threshold was adjusted to balance the True Positive Rate (TPR) and False Positive Rate (FPR), ensuring the best trade-off between correctly identifying blue tarps and minimizing false positives. A threshold of 0.7 was chosen for most models, as it provided a high level of classification accuracy while reducing unnecessary misclassifications.

Multiple metrics were used to assess model effectiveness:

- **Accuracy:** Measures the overall correctness of the model's predictions.
- **Brier Score:** Evaluates the accuracy of probabilistic predictions by penalizing confidence in incorrect classifications.
- **Area Under the ROC Curve (AUC-ROC):** Assesses the model's ability to distinguish between classes across different threshold values.
- **Precision:** Indicates how many predicted positives were actually correct, crucial for minimizing false alarms in disaster response scenarios.
- **True Positive Rate (TPR):** Measures how effectively the model identifies blue tarps, ensuring minimal missed detections.
- **False Positive Rate (FPR):** Evaluates the proportion of incorrect positive classifications, helping mitigate unnecessary false alerts.

These metrics collectively provide a comprehensive evaluation of the model's predictive power and real-world applicability.

In summary, team 12 prepared the dataset for this supervised machine learning task by ensuring that the target variable was in the correct format, defining the relationship between the target and predictors, and setting up a preprocessing pipeline to normalize numeric predictors. This

approach is modular and follows best practices in machine learning workflows, particularly when using the tidymodels ecosystem. It ensured a reproducible, clear, and scalable process for building and evaluating the model.

Results

Training Dataset Analysis

Model Training Cross-validation Performance Metrics			
Model	**Accuracy**	**Brier-Class**	**AUC**
Logistic regression	0.886	0.093	0.924
LDA	0.856	0.104	0.912
QDA	0.901	0.080	0.933

Table 1: Cross-validation Performance Metrics (Training Data) - Summary of the statistics generated from the 10-fold stratified cross-validation approach applied to the training data to assist with model evaluation between Logistic Regression, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA) models.

Based on the cross-validation performance metrics (see Table 1), the Quadratic Discriminant Analysis (QDA) model emerges as the best-performing model among the three options: Logistic Regression, Linear Discriminant Analysis (LDA), and QDA. This conclusion is supported by its superior performance across all evaluation metrics: accuracy, Brier Score, and AUC.

QDA assumes that the decision boundary between classes is quadratic. It can curve to separate classes in a more flexible manner than linear classifiers. This assumption arises from QDA's underlying model, which assumes different covariance structures for each class, allowing it to capture the complex relationships between the variables. In our 3D rendering (see the chart below), we observe that the model effectively creates a quadratic decision boundary that adapts to the shape of the data, neatly partitioning the training set into distinct regions based on the predicted class probabilities. This visualization reinforces the QDA model's flexibility in modeling class boundaries that fit the true distribution of the data.

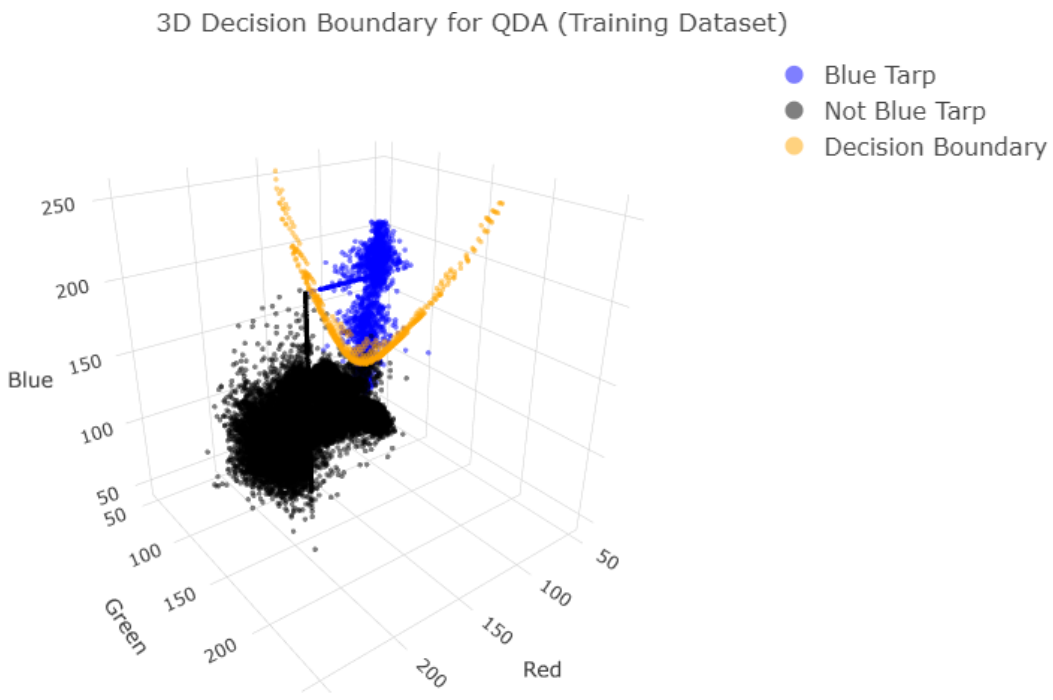

Figure 7: 3D Scatter Plot of RGB by Blue Tarp Classification with Decision Boundary (Training Data; QDA Model) - 3D chart displaying that the model effectively creates a quadratic decision boundary that adapts to the shape of the data.

QDA achieves the highest accuracy (0.901), meaning it correctly classifies approximately 90.1% of the instances in the dataset. In comparison, Logistic Regression has an accuracy of 0.886, and LDA has the lowest accuracy at 0.856. While higher accuracy suggests that QDA is better at making correct predictions overall, it is important to consider the impact of class imbalance on this metric.

The dataset exhibits a significant class imbalance, with one class (“Blue_Tarp” = “Not Blue Tarp”) dominating the observations. Because of this, accuracy can be misleading, as the model could achieve high accuracy simply by predicting the majority class most of the time, without effectively distinguishing between the minority class ("Blue_Tarp" = “Blue Tarp”). While QDA's high accuracy is promising, it is essential to evaluate its performance using additional metrics to ensure that it is not merely benefiting from the class imbalance and is genuinely effective at identifying the minority class. This is particularly critical in disaster relief scenarios, where correctly identifying the minority class (blue tarps in images) is of utmost importance.

The Brier Score, which measures the accuracy of probabilistic predictions, further supports QDA's superiority. A lower Brier Score indicates better performance, and QDA has the lowest score at 0.080. This suggests that its predicted probabilities are closer to the actual outcomes compared to Logistic Regression (Brier Score of 0.093) and LDA (Brier Score of 0.104). QDA's lower Brier Score demonstrates its ability to provide more reliable probability estimates, which is crucial for tasks requiring confidence in predictions.

The AUC metric, which evaluates the model's ability to distinguish between classes, also favors QDA. QDA achieves the highest AUC value of 0.933, indicating it is better at differentiating the classes (blue tarp, rooftop, soil, various non-tarp, and vegetation). Logistic Regression follows with an AUC of 0.924, and LDA has the lowest at 0.912. QDA's higher AUC confirms its stronger ability to separate classes, making it most effective for classification tasks based on this metric.

In conclusion, QDA outperforms both Logistic Regression and LDA across all three metrics: accuracy, Brier Score, and AUC. Its higher accuracy reflects better overall classification performance, its lower Brier Score indicates more accurate probability predictions, and its higher ROC-AUC demonstrates superior class separation. While Logistic Regression and LDA are strong models, QDA is the best choice for this dataset based on the cross-validation performance statistics.

However, it is important to consider additional factors when selecting a model. If the dataset is large or the relationship between features and the target is linear, Logistic Regression or LDA might still work best when applied to new unseen data. Additionally, Logistic Regression is often preferred when interpretability is a priority, as it provides clear coefficients for each feature. In summary, while QDA is the best model in the training environment based on the given metrics, the final model decision should also consider the specific context of the problem as well as data-related attributes such as the dataset size, feature relationships, and the need for interpretability.

Holdout Dataset Analysis

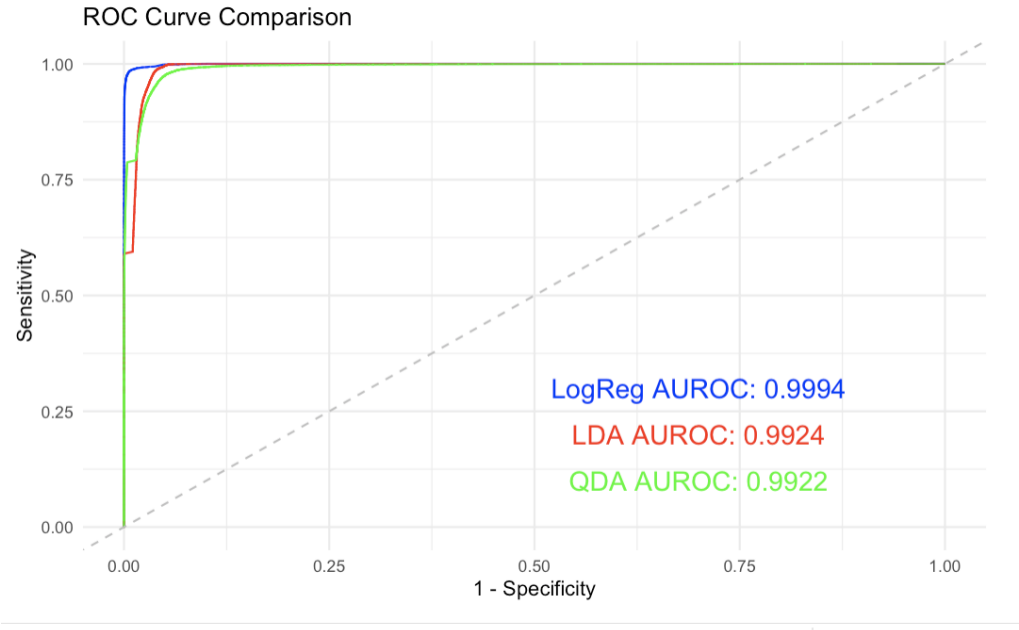

Figure 8: ROC Curve Comparison (Holdout Data) - A graphical representation of a classification model's performance across different threshold values for each of the models examined by Group 12. Shows Logistic Regression model (blue curve) exhibits an AUROC of 0.9994, which is very close to indicating a perfect classification performance in distinguishing between the two classes predicting whether the Blue_Tarp variable is "Blue Tarp" or "Not Blue Tarp."

As displayed in the chart above, the ROC curve comparison of three classification models: Logistic Regression, LDA, and QDA. The Logistic Regression model (blue curve) exhibits an AUROC of 0.9994, which is very close to indicating a perfect classification performance in distinguishing between the two classes predicting whether the Blue_Tarp variable is "Blue Tarp" or "Not Blue Tarp". A high AUROC suggests that the Logistic Regression model has an extremely high true positive (sensitivity) and a very low false positive rate (1 – specificity). It is capable of correctly identifying positive and negative cases with remarkable precision.

The LDA model demonstrates a strong performance with an AUROC of 0.9924, which is still quite high but slightly lower than that of Logistic Regression. While not as perfect as Logistic Regression model, this AUROC value still indicates that LDA is highly effective at discriminating between the classes. LDA works by finding a linear combination of features that best separate the two classes, and its performance here suggests that the classes are relatively well-separated in the feature space. However, the slight difference in AUROC compared to the logistic regression could be due to the LDA's assumption of normally distributed data with equal variances across classes, which might not hold true in all cases. See the chart above.

QDA's AUROC score of 0.9922 is very similar to LDA, which suggests that QDA is quite capable of distinguishing between the classes. QDA differs from LDA by relaxing the assumption of equal variance for the two classes, allowing for more flexibility in the classification decision. However, despite this flexibility, QDA's performance here is almost identical to that of LDA. This could indicate that the assumption of equal variances is not a significant limitation for this holdout dataset, and both models are able to achieve similar performance. The very slight difference in AUROC scores between LDA and QDA suggests that for this specific problem, the added complexity of QDA may not provide a substantial improvement in classification accuracy.

Holdout Data Set Performance Metrics						
Model	**AUROC**	**Threshold**	**Accuracy**	**TPR**	**FPR**	**Precision**
QDA	0.9922	0.7	0.9954	0.9986	0.3502	0.9967
LDA	0.9924	0.9	0.9835	0.9859	0.2664	0.9974
Logistic Regression	0.9994	0.7	0.9947	0.9948	0.0195	0.9998

Table 2: Model Performance Metrics (Holdout Data) - Summary of the statistics generated from applying the Logistic Regression, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA) models to the heretofore unseen holdout data. Logistic Regression exhibited the highest AUROC of 0.9994, indicating exceptional discriminatory ability.

The table summarizes the hold-out performance results of three classification models on the holdout data set: QDA, LDA, and Logistic Regression (Log Reg). Performance evaluation of these models was conducted using several key metrics, including the Area Under the Receiver Operating Characteristic Curve (AUROC), classification threshold, accuracy, True Positive Rate (TPR), False Positive Rate (FPR), and precision.

Logistic Regression exhibited the highest AUROC of 0.9994, indicating exceptional discriminatory ability. The Logistic Regression operated with a 0.7 threshold and performed

exceptionally well in balancing TPR and FPR. It also demonstrated a robust balance between accuracy (0.9947) and TPR (0.9948), suggesting the model effectively identified most positive instances. Furthermore, the FPR for Logistic Regression was notably low at 0.0195, highlighting its ability to minimize false positives compared to the other models. Its precision score of 0.9998 further corroborates its high reliability in making positive predictions.

QDA also performed admirably, achieving an AUROC of 0.9922 and the highest accuracy (0.9954) among the models. The selected threshold is 0.7 leading to an exceptionally high TPR (0.9986), indicating that almost all positive instances are correctly identified. However, its relatively higher FPR (0.3502) suggests a substantial number of negative instances were incorrectly classified as positive. In comparison, LDA exhibited slightly lower accuracy (0.9835) than QDA but had a lower FPR (0.2664). LDA's AUROC (0.9924) was comparable to QDA, and its precision (0.9974) was marginally higher. The threshold for LDA was higher at 0.9, which resulted in a slightly lower TPR (0.9859). This indicates that some positive cases are missed due to the stricter classification criterion.

Due to the small minority class in both the training and holdout data, Group 12 discussed whether to employ imbalance handling techniques like oversampling the minority class, undersampling the majority class, or Synthetic Minority Over-sampling Technique (SMOTE). All models, particularly Logistic Regression, handle the extreme class imbalance well, achieving high AUROC (>0.99), recall (~99%), and precision (~99.98%), with Logistic Regression maintaining an exceptionally low False Positive Rate (1.95%). Given these strong performance metrics, additional imbalance handling techniques are unnecessary.[iii] QDA’s higher False Positive Rate (35.02%) is a concern, but overall, the models effectively distinguish Blue Tarp cases without needing further adjustments.

In summary, Logistic Regression emerged as the top-performing model due to its superior AUROC, lower FPR, and excellent overall classification performance. It outperforms both QDA and LDA, offering more reliable results for the given dataset by maintaining high accuracy and precision while minimizing false positives. These findings suggest that Logistic Regression is the most effective classifier among the models evaluated.

Conclusions

After review of model results and subsequent analysis, Group 12 offers the following conclusions from work related to Part I of the Disaster Relief Project for the University of Virginia’s Data Science 6030 course (Statistical Learning).

Conclusion 1: Logistic Regression Outperforms LDA and QDA

Despite Quadratic Discriminant Analysis (QDA) performing best on the cross-validation dataset, Logistic Regression outperformed Linear Discriminant Analysis (LDA) and QDA on the holdout

dataset. The Logistic Regression model achieved the highest AUROC (0.9994), the lowest False Positive Rate (0.0195), and a near-perfect precision (0.9998). This makes it the most reliable classifier for identifying blue tarps. Furthermore, these results suggest that while QDA may work well in training data, it is not as robust when applied to new unseen data. Overall, Logistic Regression's performance on the holdout dataset indicates that it's better at generalizing, making it the most suitable model for real-world implementation.

The results suggest that while QDA performed well in training data, it failed to generalize effectively to new, unseen data. One of the challenges experienced with complex models is overfitting to the training dataset. Overfitting may have likely contributed to QDA's high false positive rate (0.3502) in the holdout dataset.

Additionally, Logistic Regression provides significant advantages with interpretability. Unlike LDA and QDA, which rely on assumptions about data distributions, Logistic Regression directly estimates the class probabilities. Ergo, making it easier to analyze model confidence in real-world applications. This is crucial for disaster response efforts, where rescue teams need clear, reliable predictions for prioritizing aid.

With its strong generalizability, low false positive rate, and high precision, Logistic Regression is the best model for real-world implementation in identifying displaced persons based on aerial imagery.

Conclusion 2: Enhance Logistic Regression with Regularization and Ensemble Models

Previously, it was established that Logistic Regression outperformed other models in the holdout dataset. However, additional techniques could be explored to further enhance performance. One recommendation would be to incorporate regularization methods, such as L1 (lasso) or L2 (ridge) penalties, to potentially improve generalization and reduce overfitting. Regularization would help in preventing excessive reliance on specific color intensity values (Red, Green, Blue), which may vary due to lighting or environmental conditions. Also, exploring ensemble methods, like Random Forest or Boosting methods, may yield stronger classification results. This can help in capturing complex relationships within the data.

Another area for improvement is utilizing feature selection to extract additional predictive attributes from the RGB values, such as spatial and geospatial information (Map_X, Map_Y, Lat, Lon). By using feature selection this could further enhance classification accuracy and to potentially identify spatial patterns in tarp locations. Spatial clustering techniques could help identify high-density tarp areas, providing additional context for accurate classification. Random Forest and Boosting methods can capture nonlinear relationships between features, which may provide a more nuanced classification approach. Ensemble models can leverage multiple decision trees or learners, making them less sensitive to noise and individual misclassifications compared to single classifiers like Logistic Regression.

Fine-tuning of classification thresholds could also be used to enhance the model. By adjusting the probability threshold used to classify blue tarps, it could further improve precision and recall. Since Logistic Regression demonstrates high precision but a slightly lower recall, optimization of the threshold could help capture more true positives without significantly increasing false positives. By implementing the recommended optimization, classification performance could be improved, providing a more accurate, reliable, and efficient approach to detecting people who are displaced in disaster-stricken areas.

Conclusion 3: Geospatial Intelligence for Faster, Smarter Aid Delivery

There is transformative potential of machine learning in humanitarian aid efforts, specifically in the automation of the identification of displaced persons using aerial imagery. Considering the datasets inclusion of spatial and geospatial features (Map_X, Map_Y, Lat, Lon), there is an opportunity to leverage location-based intelligence for relief operations.

Considering one of the challenges in disaster relief, like quickly identifying and prioritizing areas where aid might be needed more, machine learning can rapidly analyze thousands of images. By having this ability, rescue teams can pinpoint high-density tarp clusters without having to manually inspect. With the integration of geospatial data, models can provide real-time mapping of tarp locations, further helping aid workers navigate through damaged infrastructure and optimize resource distribution.

In conclusion, the inclusion of geospatial analysis and integration of machine learning would offer a scalable and data driven approach to disaster relief. This leads to faster response times and effective resource allocation to those impacted by natural disasters.

---

[i] United Nations Office for the Coordination of Humanitarian Affairs (OCHA), *Evaluation of OCHA Response to the Haiti Earthquake* (New York: OCHA, 2010), accessed [insert date], https://www.unocha.org/sites/unocha/files/dms/Documents/Evaluation%20of%20OCHA%20Response%20to%20the%20Haiti%20Earthquake.pdf

[ii] Wikipedia contributors. (n.d.). *RGB color model – Numeric representations*. Wikipedia, The Free Encyclopedia. Retrieved February 28, 2025, from https://en.wikipedia.org/wiki/RGB_color_model#Numeric_representations

[iii] van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J Am Med Inform Assoc. 2022 Aug 16;29(9):1525-1534. doi: 10.1093/jamia/ocac093. PMID: 35686364; PMCID: PMC9382395.