Machine Learning Model Report

Rana Hamed, Heba Ibrahim, Beshoy Ageeb, Eman Mansour

**1. Introduction**

This report outlines the process and results of building machine learning models to predict the target variable lumpy using a dataset with various features. The models evaluated in this study include Random Forest and XGBoost. The steps taken in this project include data preprocessing, handling class imbalance, feature selection, model training and validation, and visualizations of the results.

**2. Data Preprocessing**

* The dataset was imported from a CSV file named preprocessed\_data.csv.
* The following columns were dropped due to irrelevance to the analysis: Longitude, Latitude.
* The features were defined as all columns except for the target variable lumpy.

**3. Handling Imbalanced Data**

The class distribution of the target variable was imbalanced. To address this issue, SMOTE (Synthetic Minority Over-sampling Technique) was used to oversample the minority class:

**4. Data Splitting**

The resampled dataset was split into training (60%), validation (20%), and test (20%) sets:

**5. Feature Selection**

Feature selection was performed using the SelectKBest method with ANOVA F-test, selecting the top 3 features:

The selected features were:

* Cloud\_Cover\_Percentage
* Precipitation\_Amount
* Mean\_Temp

**6. Model Training**

* **6.1 Random Forest**
* Hyperparameter tuning for the Random Forest model was conducted using GridSearchCV:
* **6.2 XGBoost**
* Similarly, hyperparameter tuning for the XGBoost model was performed:

**7. Model Validation**

Both models were validated using the validation set, and their accuracy scores were obtained:

**Random Forest Validation**

* **Validation Accuracy**: 0.9534

**XGBoost Validation**

* **Validation Accuracy**: 0.9510

**8. Model Testing**

The models were then tested on the test set:

**Random Forest Testing**

* **Test Accuracy**: 0.9515
* **Classification Report**:
  + Precision: 0.95
  + Recall: 0.95
  + F1-score: 0.95

**XGBoost Testing**

* **Test Accuracy**: 0.9474
* **Classification Report**:
  + Precision: 0.94
  + Recall: 0.95
  + F1-score: 0.95

**9. Confusion Matrix**

The confusion matrices for both models were visualized using heatmaps, providing insight into the true positive, false positive, true negative, and false negative counts.

**Random Forest Confusion Matrix:**

A graph with blue squares and white squares

Description automatically generated

**XGBoost Confusion Matrix:**

A blue squares with white text

Description automatically generated

**10. Additional Visualizations**

**10.1 ROC Curves**

ROC curves were plotted for both models to visualize the trade-off between true positive rates and false positive rates.

A graph of a curve

Description automatically generated

**10.2 Feature Importance**

The feature importance plots indicated which features contributed most to the model predictions for both Random Forest and XGBoost.

A graph of blue rectangular bars

Description automatically generated with medium confidence

A graph of blue rectangular objects

Description automatically generated with medium confidence

**10.3 Classification Report Heatmaps**

Classification report heatmaps provided a visual summary of precision, recall, and F1-score for each class.

A graph showing a number of different colored squares

Description automatically generated

**Description of the Heatmap for Random Forest Classification Report**

This heatmap visualizes the classification report for the Random Forest model, providing insights into its performance across different metrics for the two classes (0 and 1). Below are key observations drawn from the heatmap:

1. **Precision**:
   * The precision for both classes (0 and 1) is consistently high at **0.95**. This indicates that the model is very effective at identifying true positive instances, with a low rate of false positives.
2. **Recall**:
   * Similar to precision, the recall for both classes is also **0.95**. This suggests that the model successfully captures a significant proportion of true positives for both classes, demonstrating a strong ability to identify actual instances of the positive class.
3. **F1-Score**:
   * The F1-score is a harmonic mean of precision and recall, which is also **0.95** for both classes. This reinforces the model's overall effectiveness, as it balances the trade-off between precision and recall.
4. **Support**:
   * The support value represents the number of actual occurrences of each class in the test dataset. In this case, both classes have substantial support, which is indicative of a well-balanced dataset.
5. **Macro Average**:
   * The macro average scores are also shown in the heatmap. The precision, recall, and F1-score for the macro average are **0.95**. This average treats all classes equally and is useful for assessing the model's performance across imbalanced datasets.

Overall, this heatmap demonstrates that the Random Forest model performed exceptionally well across all metrics, indicating its reliability in classifying the instances correctly.

**Additional Insights from the Classification Report**

* The high performance metrics across all evaluated areas indicate that the Random Forest model is suitable for this classification task.
* Given the consistency in scores between both classes, it is clear that the model is not biased towards one class, thereby reducing the risk of misclassification in practical applications.
* The high support values for both classes imply that the dataset used for training and validation is robust and adequately represents the underlying distribution of the target variable.

A graph of a graph showing the number of heatmap

Description automatically generated

Finally

Our main goal is to catch as many diseased buffalo as possible to prevent the spread of disease. **Recall** tells us how good the model is at finding all the buffalo that are actually diseased.

If **recall** is high, it means we’re catching most of the buffalo that have the disease, which is critical for effective intervention.

In short: we want to minimize the chance of missing any diseased buffalo, and that's why **recall** is the most important metric for this study case.