Coding Challenge Write-ups

Analyzing COVID-19 Case Categories Using Random Forest Classification

1. Background Information

The COVID-19 pandemic significantly impacted global health, economies, and daily life, triggering unprecedented public health responses. Governments around the world implemented various measures to control the spread, such as vaccination campaigns and social restrictions. This analysis uses country-level COVID-19 data, including new cases, vaccination counts, and government stringency measures, to understand how these factors relate to the intensity of outbreaks.I specificallt chose vaccination counts and government stringency measures because I think these two are the most prominent parameters that influence the COVID cases, vaccination for preventative purpose and stringency measures for emergent responses. This project attempts to categorize countries into two groups—those with high and low new case counts by using a Random Forest classifier.

2. Problem Statement

This project aims to predict whether a country experienced a **high** or **low number of COVID-19 cases** based on its **total vaccinations** and **stringency index** (a measure of government restrictions).

3. Hypothesis

It is hypothesized that countries with higher vaccination rates and stricter government interventions will have fewer COVID-19 cases, resulting in a low case category classification. The model aims to validate this relationship using the features available in the dataset.

4. Methods

Data Collection and Cleaning

We utilized a publicly available dataset from the Our World in Data COVID-19 project. The data included information on the number of new COVID-19 cases, vaccination rates, and the stringency of government responses.

- **Data Cleaning**:
  - Missing values were **forward-filled** to ensure completeness.
  - Column names were **standardized** by stripping spaces and converting them to lowercase.
  - Relevant columns selected: country, date, new_cases, total_vaccinations, stringency_index.
- **Feature Choice**
  - The data was **grouped by country**, summing new cases, taking the maximum vaccinations, and averaging the stringency index.
  - Data was normalized using the **MinMaxScaler** to ensure features are on a similar scale.
  - A **binary target variable** was created: case_category, where countries with new cases above the median were labeled as **1 (High)**, and others as **0 (Low)**.
- **Model Selection**:
  - I chose **Random Forest Classifier** for this task due to its ability to handle non-linear relationships and importance in feature selection.This is because I think COVID-19 cases is very

complicated so it requires various random subset of the data to make a more precise final prediction.

- The dataset was split into **80% training and 20% testing sets** to evaluate model performance.

5. Results and Discussion

The Random Forest Classifier achieved an accuracy of 53%, indicating moderate performance. The classification report shows that the precision, recall, and F1-scores for class 0 (low cases) are higher (precision: 0.53, recall: 0.63, F1: 0.58) compared to class 1 (high cases) (precision: 0.47, recall: 0.42, F1: 0.44), suggesting the model struggles more with identifying high-case countries. The macro and weighted averages reflect balanced, but suboptimal performance, with an F1-score around 0.52. From the confusion matrix, the model shows a tendency towards false negatives, misclassifying high-case countries as low-case more often, indicating potential issues with feature relevance or class overlap. Since the time limit, I didn't choose another model for this coding competition. Nevertheless, for future work, I may need to choose alternative algorithms like K-Nearest Neighbors (KNN) that may better capture non-linear relationships. Moreover, I may need to include more data inputs such as other factors that influece the COVID-19 cases to improve the classification performance of my model.