Project Overview: Predicting COVID-19 Vaccination Status in Canada Using Machine Learning

1. Project Objective

This project aims to develop a predictive model that classifies individuals' COVID-19 vaccination status using socio-demographic and behavioral features extracted from a Canadian survey dataset. The focus is on distinguishing between "fully vaccinated" and "not fully vaccinated" individuals using a supervised learning approach, specifically the **K-Nearest Neighbors (KNN)** algorithm.

2. Data Sources

- **Primary Dataset**: COVID-19BehaviorData_CAN2022.csv
  A large-scale survey dataset capturing behavioral, attitudinal, and demographic data related to COVID-19 among Canadian residents.
- **Supplementary Metadata**: ins.xlsx
  An instruction file containing mappings of categorical responses to numeric values and metadata for valid values.

3. Data Preprocessing and Cleaning

Initial Steps:

- Removed columns with more than 55% missing values (NaN).
- Replaced non-standard missing value indicators ("", "__NA__", "Don't know") with NaN.
- Stripped all white space from string values to ensure consistent formatting.

Categorical Encoding:

- Used custom value mappings based on the ins.xlsx file.
- Replaced string values with numerical codes as specified in the instructions file using a purpose-built function (filler()).

Imputation of Missing Values:

- Imputed remaining NaN values using random sampling from existing values within the same column. This approach maintains the original distribution of values, avoiding bias introduced by constant or mean/mode imputation.

Standardization of Categorical Data:

- Provinces in the region column were encoded with unique integers due to inconsistencies found during inspection (UK regions instead of Canadian provinces).

4. Exploratory Data Analysis (EDA)

- Conducted frequency distribution analysis on the vac (vaccination status) variable. The class distribution was found to be imbalanced.
- Re-categorized vaccination status into two classes:
    - Class 1: “Not fully vaccinated”
    - Class 2: “Fully vaccinated”
- Generated bar plots and stacked bar plots of key predictor variables (vac7, r1_8, vac_man_1, vac_man_4, vac_man_5, vac2_7, vac2_3) against the response (vac) using cross-tabulations.
- Analyzed data distributions using histograms, box plots, and correlation heatmaps to:
    - Assess variable distributions (non-normal variables identified)
    - Confirm absence of significant outliers
    - Identify multicollinearity

5. Feature Selection

Selected predictors were based on domain relevance and correlation analysis:

- vac2_3, r1_8, vac2_7, vac_man_1, vac_man_4, vac_man_5

These predictors were retained after evaluating pairwise correlations and visualization-based inspections to reduce redundancy.

6. Model Development: K-Nearest Neighbors (KNN)

Data Splitting:

- Train-Test Split: 70-30
- Stratified sampling used to maintain class distribution
- Seed fixed for reproducibility

Feature Scaling:

- Applied z-score normalization (StandardScaler) to input features to ensure optimal KNN performance.

Model Selection:

- Performed 10-fold cross-validation with a grid search over k values (1 to 80) to identify the optimal number of neighbors.
- Best performing value: k = 42

Model Training and Evaluation:

- Final model trained using optimal k=42
- Evaluation metrics:
    - Accuracy Score
    - Precision, Recall, F1-Score via classification_report
    - Confusion Matrix and Summary Statistics via dmba.classificationSummary

7. Results and Insights

- The final KNN model achieved competitive performance in classifying vaccination status.
- Predictive features such as opinions about vaccines (vac_man_*), behavioral tendencies (vac7, r1_8), and past vaccine behaviors (vac2_3, vac2_7) showed significant influence on the response variable.
- The normalization of predictors and stratification in train/test splitting contributed to model stability despite class imbalance.

8. Conclusion

This project demonstrates a structured machine learning pipeline, from raw data processing, extensive data cleaning, and feature engineering to model development and evaluation, on a real-world COVID-19 behavioral dataset. The final KNN model provides an interpretable and statistically sound tool for predicting vaccination behavior, which could inform public health strategies and targeted interventions.

9. Recommendations for Future Work

- Apply SMOTE or ADASYN for handling class imbalance.
- Explore other classifiers like Logistic Regression, Random Forest, or Gradient Boosting for comparison.
- Perform feature importance analysis using model-agnostic techniques such as SHAP or LIME.
- Expand to a multi-class classification task if future data includes more granular vaccination statuses.