## AIM

The goal of this analysis is to predict the likelihood of individuals receiving two types of vaccines: the XYZ vaccine and the seasonal flu vaccine. The dataset includes various features related to demographic information, behavioral factors, and opinions about the vaccines. The problem is formulated as a multilabel classification task where we predict two probabilities: one for receiving the XYZ vaccine and another for receiving the seasonal flu vaccine. The evaluation metric for this task is the Area Under the Receiver Operating Characteristic Curve (ROC AUC).

## Data Overview

The dataset comprises 36 columns, with the first column being a unique identifier (`respondent_id`) and the remaining 35 columns containing various features. The target variables are ‘`xyz_vaccine`’ and ‘`seasonal_vaccine`’.

## Data Preprocessing

### Handling Missing Values

I began by addressing missing values in the dataset. The strategy for handling missing values depended on the proportion of missing data for each column. Columns with less than 5% missing values were imputed using the mode of the respective column. For columns with more significant missing values, we also imputed using the mode, considering the need to preserve the dataset's structure for further analysis.

### Encoding Categorical Variables

Categorical variables were encoded using one-hot encoding. This method was chosen because it is straightforward and ensures that the model can interpret categorical data effectively. The categorical columns encoded included demographic information such as age group, education, race, and employment details.

## Feature Engineering

### Polynomial Features

To enhance the model's predictive power, I created interaction and polynomial features. Interaction features capture the combined effect of two features, which can provide additional insights that are not apparent when considering features individually.

### Correlation Analysis

A correlation matrix was computed to identify the features most strongly associated with the target variables. We selected the top 10 features for each target variable based on their correlation with `xyz_vaccine` and `seasonal_vaccine`.

## Model Training

### Data Splitting

The dataset was split into training and testing sets using an 80-20 split. This ensures that the model's performance can be evaluated on unseen data, providing a robust measure of its predictive capabilities.

### Model Selection

I chose the RandomForestClassifier for this task due to its robustness and ability to handle both numerical and categorical data effectively. Random forests are also less prone to overfitting compared to other models, especially given the high-dimensional nature of our dataset.

## Model Evaluation

The model's performance was evaluated using the ROC AUC metric. This metric is suitable for binary classification tasks and provides a measure of how well the model can distinguish between the two classes. I calculated the ROC AUC for both target variables and averaged them to get the overall score.

ROC AUC for xyz_vaccine: 0.851183092554774

ROC AUC for seasonal_vaccine: 0.8432563615243414

Mean ROC AUC: 0.8472197270395576