# Predicting COVID-19 Outcomes and Vaccine Uptake Using Behavioral and Belief Indicators

# 1 Introduction

This project focuses on predicting COVID-19 outcomes (test positivity rates) and vaccine uptake using behavioral and belief indicators. The goal is to determine whether these indicators can effectively predict important health outcomes.

Predicting test positivity rates and vaccine uptake is crucial for public health planning. Accurate predictions can help optimize resource allocation, improve vaccination outreach, and better manage COVID-19 spread.

# 2 Dataset Overview

The dataset contains survey data with around 4,000 observations and multiple features related to individuals' behaviors and beliefs.

Key features include :-

- smoothed_wcli: Percentage of individuals reporting COVID-like symptoms.
- smoothed_wtested_14d: Percentage of individuals tested for COVID-19 in the last 14 days.
- smoothed_wvaccine_likely_ (e.g., government, friends, politicians): Willingness to vaccinate.
- smoothed_wpublic_transit_1d, smoothed_wshop_1d, smoothed_wrestaurant_1d: Daily behavioral activities.

Target variables :-

- smoothed_wtested_positive_14d: Percentage of positive COVID-19 tests in the past 14 days.
- smoothed_wcovid_vaccinated: Percentage of individuals vaccinated for COVID-19.

Since there are 2 target variables, we have made two copies of the dataset, one with each of the target variables and all the features in each.

# 3 Exploratory Data Analysis

## 3.1 Data Cleaning

For smoothed_wtested_positive_14d :-

- ➢ Rows with missing values in the target were dropped. There were too many missing values to impute reliably, so we chose to remove them.
- ➢ Missing feature values were handled with mean imputation.

For smoothed_wcovid_vaccinated :-

- ➢ Rows with missing values in the target were again dropped.
- ➢ Missing feature values were handled with median imputation.

We initially tried iterative imputation, but this led to some data leakage and ultimately caused overfitting in the models. This is why we reverted to the simpler mean and median imputation methods, which did not introduce such issues.

## 3.2 Correlation Analysis and Feature Selection

We calculated the correlation between each feature and the target variable, visualizing them with heatmaps (see Appendix A). Features like smoothed_wcli and smoothed_wpublic_transit_1d showed strong correlations with both targets. We didn't remove weaker correlated features because regularization techniques like Ridge and Lasso handled multicollinearity.

When splitting the dataset, we removed the time_value and geo_value columns to avoid overfitting, as they could correlate with the targets but don't reflect actual behaviors or beliefs. This ensured the models focused on relevant features.

Feature Scaling: We applied standardization (Z-score normalization) to the features to ensure they were on the same scale, which is important for models like Neural Networks.

# 4 Baseline Model

For this project, we chose linear regression as our baseline model to predict both vaccine uptake (smoothed_wcovid_vaccinated) and COVID test positivity rate (smoothed_wtested_positive_14d). Linear regression was chosen because it provides a simple and interpretable model that assumes a linear relationship between the features and the target variables. We also applied Ridge and Lasso regularization to the linear regression models to handle multicollinearity and prevent overfitting, which is especially important given the potential correlation between many features.

- ➢ Ridge Regression: Uses L2 regularization to shrink the coefficients of the features, helping reduce the impact of multicollinearity and prevent overfitting.
- ➢ Lasso Regression: Uses L1 regularization, which can help in feature selection by forcing some coefficients to zero, thus removing irrelevant features.

These regularization techniques were chosen because they are well-suited for high-dimensional data with multicollinearity.

# 5 Methods

Beyond the baseline model, we implemented several other machine learning models to improve prediction performance.

## 5.1 Random Forest Regressor

Random Forest is an ensemble method that combines multiple decision trees to produce a more robust model. It can capture complex, non-linear relationships between features and the target variable, which is especially useful when the data has interactions that a simple linear model may not capture.

For both targets (smoothed_wcovid_vaccinated and smoothed_wtested_positive_14d) we used the Random Forest Regressor with default parameters.

Key Parameters :-

- n_estimators : Default value of 100 trees.
- max_depth : Default value is None, meaning the trees expand until leaves are pure or contain fewer than the minimum samples to split.
- min_samples_split and min_samples_leaf: Default values are 2 and 1, respectively.

## 5.2 Neural Network (MLP Regressor)

Neural networks are capable of learning complex patterns in the data, especially with a large number of features. We chose MLP Regressor to see if the model could capture interactions between features that simpler models like regression might miss.

**For smoothed_wtested_positive_14d (COVID Positivity Rate) :-**

The MLP Regressor was used to predict the COVID positivity rate. The model's architecture included 3 hidden layers with sizes (128, 64, 32) and used the ReLU activation function.

Max iterations were set to 3000 to ensure convergence, and the model was trained on the original training data without logarithmic transformation.

Key Parameters:
- hidden_layer_sizes = (128, 64, 32): Three hidden layers with decreasing sizes.
- activation = 'relu': Using ReLU to add non-linearity.
- max_iter = 3000: Training for a maximum of 3000 iterations.
- random_state = 42: To ensure reproducibility.

**For smoothed_wcovid_vaccinated (Vaccine Uptake) :-**
Log Transformation: Due to the distribution of vaccine uptake values, log transformation was applied to the target variable (y_vaccinated_train and y_vaccinated_test) before training.
Key Parameters :-
- hidden_layer_sizes = (256, 128, 64, 32) : Four hidden layers to better capture the complexity in vaccine uptake data.
- activation = 'relu ': Using ReLU to add non-linearity.
- max_iter = 3000 : Training for a maximum of 3000 iterations.
- early_stopping = True: Stops training if the validation score doesn't improve for a certain number of iterations.
- random_state = 42 : Ensures results are reproducible.

## 5.3 XGBoost Regressor

XGBoost is a powerful gradient boosting algorithm that performs well on a wide range of regression problems, especially when the relationships between features and the target are complex and non-linear. We included this model to see if it could outperform Random Forest and Neural Networks.

**For smoothed_wtested_positive_14d (COVID Positivity Rate) :-**

Key Parameters :-
- objective = 'reg:squarederror' : This is the regression objective, used for regression tasks.
- n_estimators = 100 : The number of boosting rounds (trees).
- max_depth = 5 : The maximum depth of each tree.
- learning_rate = 0.1 : The step size used in each boosting round.
- subsample = 0.8 : The fraction of the training data to be used for each boosting round.
- colsample_bytree = 0.8: The fraction of features used for each tree.
- random_state = 42 : Ensures results are reproducible.

**For smoothed_wcovid_vaccinated (Vaccine Uptake) :-**

Key Parameters :-
- objective = 'reg:squarederror' : This is the regression objective, used for regression tasks.
- n_estimators = 450 : The number of boosting rounds (trees).
- max_depth = 15 : The maximum depth of each tree.
- learning_rate = 0.1 : The step size used in each boosting round.
- subsample = 0.8 : The fraction of the training data to be used for each boosting round.
- colsample_bytree = 0.8 : The fraction of features used for each tree.
- random_state = 42 : Ensures results are reproducible.

## 5.4 Models Discarded

We experimented with a Support Vector Machine (SVM) for regression, but its performance was poor on this dataset. The model struggled to capture the non-linear relationships and did not improve with parameter tuning.

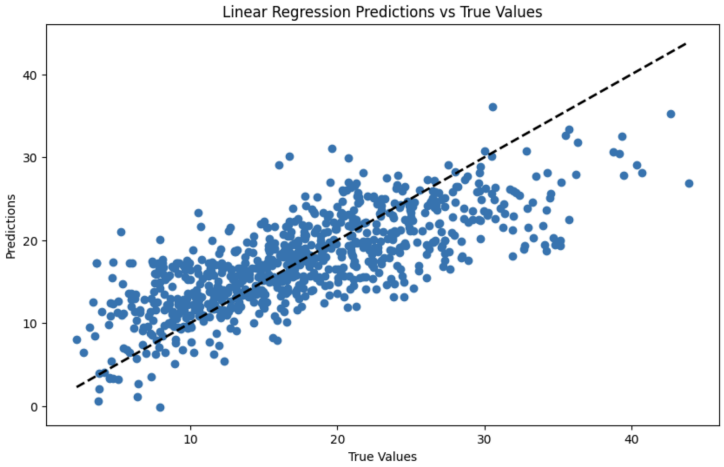
We also explored K-Nearest Neighbors (KNN) but found that it was slow to train and did not perform well with the large number of features in the dataset.

# 6 Results

Let us now evaluate the performances of each model. We have used three metrics to assess the model's performance - Mean Square Error (MSE), Mean Absolute Error (MAE), and $R^2$, which tells how much of the variance in the data is explained by the model.

## 6.1 Linear Regression

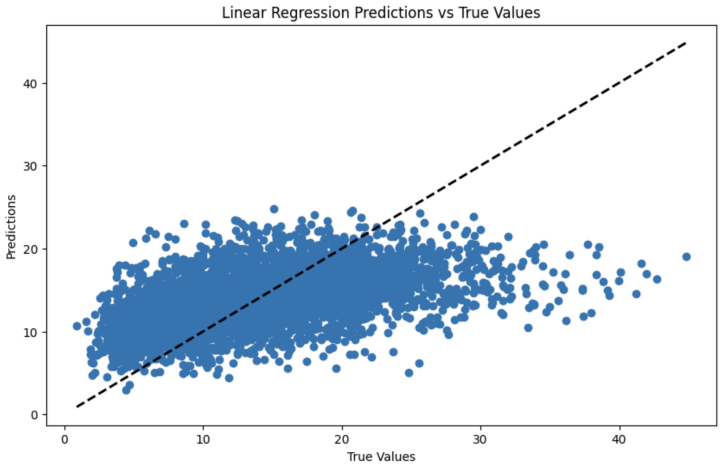
**COVID Test Positivity Rate (smoothed_wtested_positive_14d)**




Model's performance metrics :-

MAE : 3.889
MSE : 25.494
$R^2$ : 0.569

**Vaccine Uptake (smoothed_wcovid_vaccinated)**




Model's performance metrics :-

MAE : 4.443
MSE : 32.796
$R^2$ : 0.280

For COVID Test Positivity Rate, the Linear Regression model performed moderately well with an $R^2$ of 0.569, meaning it explained 57% of the variance in the target variable. The MAE and MSE indicate acceptable errors, but there's room for improvement.

For Vaccine Uptake, the model struggled more with an $R^2$ of 0.280, explaining only 28% of the variance. The higher error metrics (MAE and MSE) suggest that Linear Regression is less effective in predicting vaccine uptake with the current set of features.

While we also tested Lasso and Ridge regression models, they performed worse than the standard Linear Regression model. Despite their ability to handle multicollinearity, the regularization techniques didn't improve performance in this case, possibly due to the dataset's characteristics and the nature of the relationships between features and the target variables.

## 6.2 RandomForest Regressor

**COVID Test Positivity Rate (smoothed_wtested_positive_14d)**
Model's performance metrics :-

- MAE : 2.655
- MSE : 13.380
- $R^2$ : 0.774

**Vaccine Uptake (smoothed_wcovid_vaccinated)**
Model's performance metrics:

- MAE : 3.503
- MSE : 21.966
- $R^2$ : 0.518

For COVID Test Positivity Rate, the model performed well with an $R^2$ of 0.774, explaining 77% of the variance. The error metrics (MAE and MSE) show good predictive accuracy.
For Vaccine Uptake, the model's performance was weaker, with an $R^2$ of 0.518, indicating that the model explained only 52% of the variance. The higher MAE and MSE suggest room for improvement.

## 6.3 Multi-Layered Perceptron (MLP)

**For COVID Test Positivity Rate (smoothed_wtested_positive_14d):**
Model's performance metrics:

- MAE : 2.012
- MSE : 7.164
- $R^2$ : 0.879

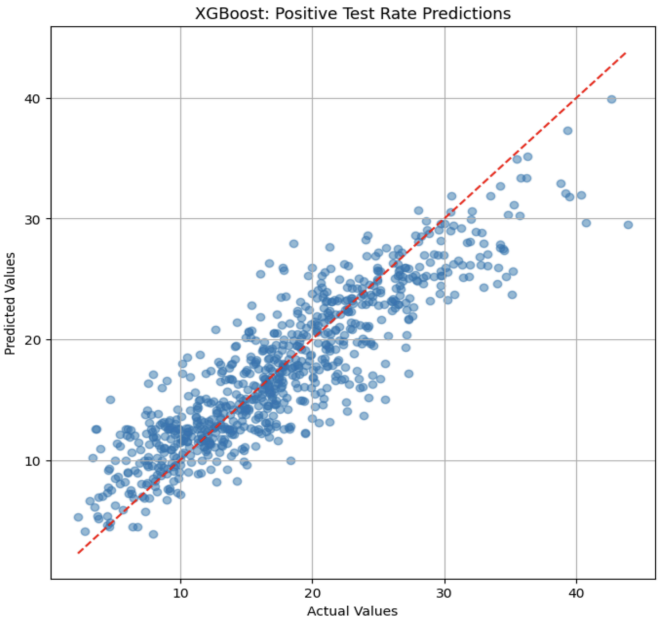
**For Vaccine Uptake (smoothed_wcovid_vaccinated):**
Model's performance metrics:

- MAE : 2.815
- MSE : 16.430
- $R^2$ : 0.639

For COVID Test Positivity Rate, the model showed excellent performance with an $R^2$ of 0.879, explaining 88% of the variance. The low MAE and MSE indicate a strong fit to the data.
For Vaccine Uptake, the MLP model performed moderately with an $R^2$ of 0.639, explaining 64% of the variance. Although better than linear regression, the model still has room for improvement in predicting vaccine uptake.

## 6.4 XGB Regressor

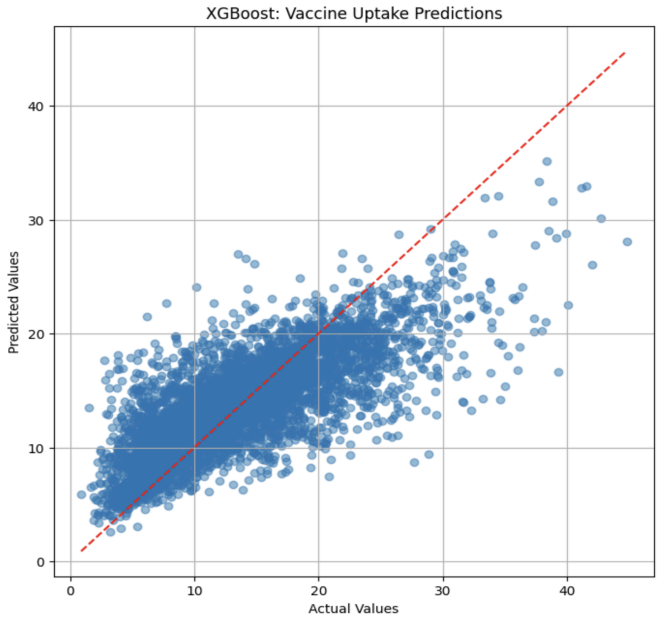
**For COVID Test Positivity Rate (smoothed_wtested_positive_14d)**




Model's performance metrics:

- MAE : 2.730
- MSE : 12.704
- $R^2$ : 0.785

**For Vaccine Uptake (smoothed_wcovid_vaccinated):**




Model's performance metrics:

- MAE : 3.2687
- MSE : 19.3538
- $R^2$ : 0.5750

For COVID Test Positivity Rate, the model performed excellently with an $R^2$ of 0.785, explaining 78% of the variance. The MAE and MSE values indicate that the model's predictions are close to the actual values.
For Vaccine Uptake, the model's performance was decent, with an $R^2$ of 0.575, indicating the model explained 57% of the variance. The MAE and MSE are relatively higher, suggesting that the model can still be improved for predicting vaccine uptake.

# 7 Analysis and Policy Recommendation :

## 7.1 Critical Analysis

### 7.1.1 Overview of Methods and Results

In this project, we used regression models (Ridge, Lasso, Neural Networks, Random Forest, and XGBoost) to predict COVID-related outcomes. The models showed promising results, with XGBoost performing best for both COVID test positivity and vaccine uptake, followed by MLP.

For COVID test positivity, XGBoost performed the best, with an $R^2$ score of 0.785. For vaccine uptake, XGBoost also performed well, with an $R^2$ score of 0.645. XGBoost and MLP were particularly effective because of their ability to model non-linear relationships and interactions between features.

### 7.1.2 Data Quality and Preprocessing Assumptions

We assumed mean imputation for the target variable smoothed_wtested_positive_14d and median imputation for smoothed_wcovid_vaccinated, as many features had missing data.

Feature selection: Even though some features had low correlation with the target, they were retained because they might provide subtle insights for prediction, as demonstrated by Random Forest’s performance.

### 7.1.3 Model Limitations and Assumptions

Model Complexity : While Random Forest and MLP showed high performance, these models have high complexity and might not generalize well with limited training data.

Overfitting Risk : Some models like MLP might overfit, but we mitigated this using techniques like early stopping.

### 7.1.4 Conditions or Scenarios Where Models Might Be Ineffective

Changes in Public Behavior : If behaviors shift due to new policies or health guidance, the models may not adapt well, especially if unrepresented behaviors were missed in training data.

Generalization Issues : Predicting both COVID test positivity and vaccine uptake solely based on behavioral indicators could overlook key factors like healthcare access, trust in government, and health disparities.

## 7.2 Answering the Key Questions

Q1: Can vaccine uptake be accurately predicted using behavioral and belief indicators?

Answer : **Yes**, The XGBoost model showed a moderate predictive performance, with an $R^2$ of approximately 0.64. While this indicates that behavioral and belief indicators can explain a significant portion of the variance in vaccine uptake, there is still room for improvement. Therefore, there is a lot of other data as well that needs to be explored in order to accurately predict vaccine uptake.

Q2: Can COVID-related outcomes, particularly the percentage of positive cases, be reliably predicted from these indicators?

Answer : **Yes**, MLP was able to reliably predict test positivity, with $R^2$ ~0.88, suggesting that COVID-related outcomes can be predicted using survey data on behaviors and beliefs.

Q3: How accurately can vaccine uptake and COVID cases be predicted using historical survey data?

Answer : The models achieved excellent accuracy for both tasks, suggesting that historical survey data is highly effective for predicting health outcomes.

## 7.3 Insights and Conclusion

Model Behavior: XGBoost performed the best for both tasks, followed by MLP. This indicates that more complex models capture non-linear relationships in the data better than simpler ones like Linear Regression.

Impact on Health Policy: The insights from these models suggest that behaviors such as testing frequency, mask-wearing, and attitudes toward vaccination can help predict vaccine uptake. This can guide targeted interventions, especially in underrepresented populations.

# Appendix

## Appendix A: Correlation Heatmaps

To provide a deeper understanding of the feature relationships with the target variables, we included correlation heatmaps for both COVID Test Positivity Rate and Vaccine Uptake. These heatmaps were used to visually explore the correlations among the features and targets, helping inform our feature selection process.

Figure A1: Correlation Heatmap for COVID Test Positivity Rate

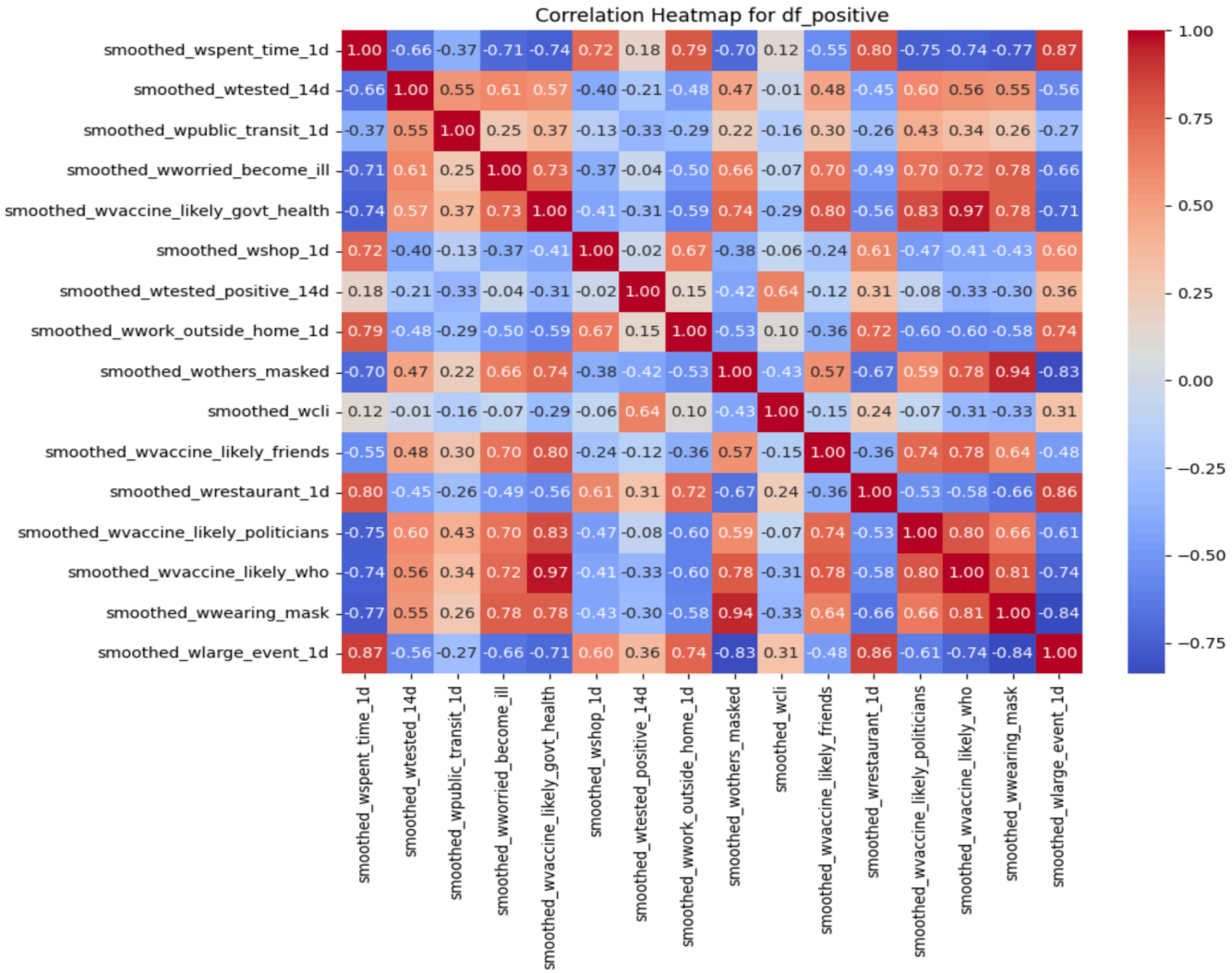
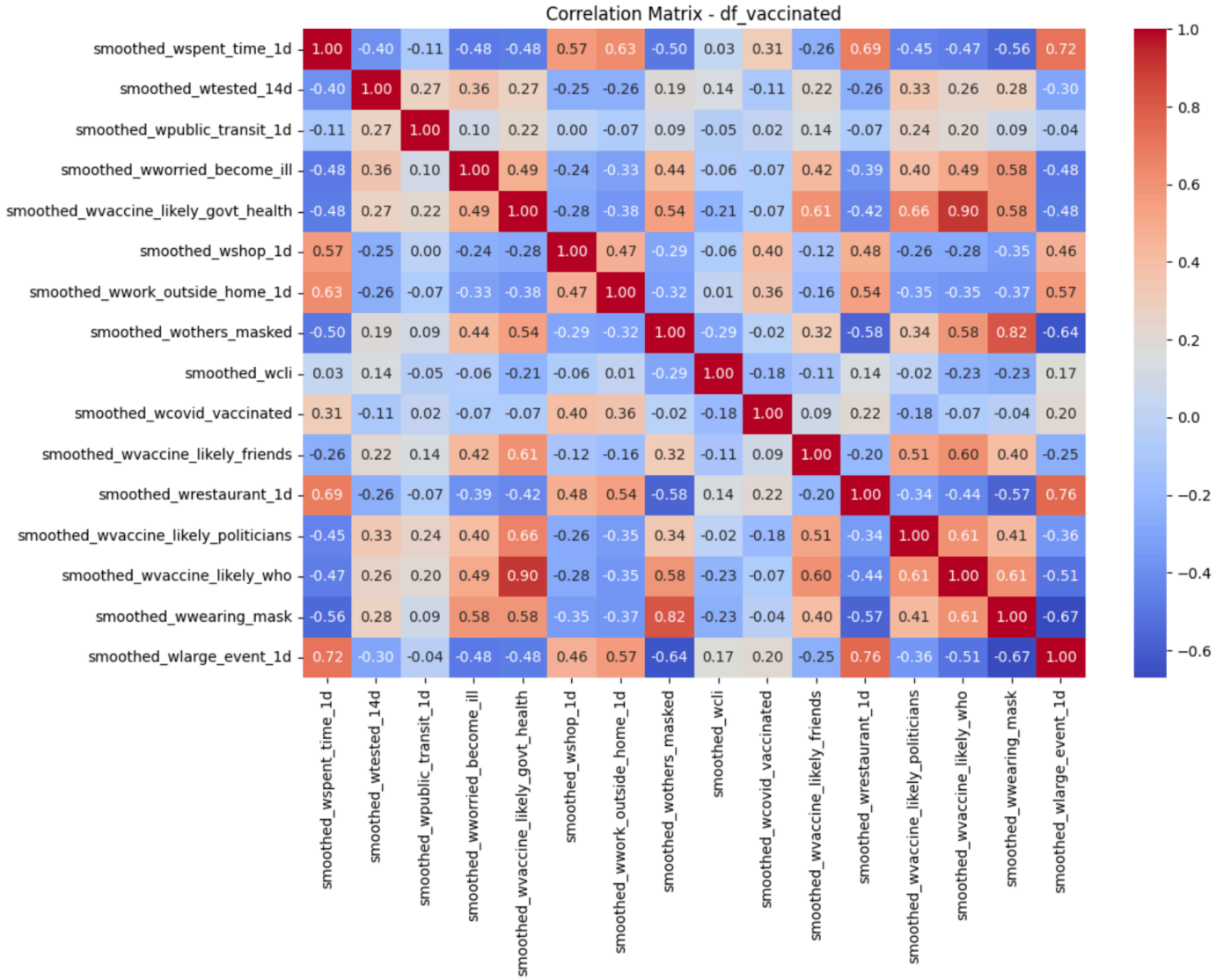


Figure A2: Correlation Heatmap for Vaccine Uptake



# References


- https://www.geeksforgeeks.org/ml-linear-regression/
- https://scikit-learn.org/stable/modules/preprocessing.html
- https://www.geeksforgeeks.org/neural-networks-a-beginners-guide/
- Popescu, Marius-Constantin & Balas, Valentina & Perescu-Popescu, Liliana & Mastorakis, Nikos. (2009). Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems. 8.
- https://www.datacamp.com/tutorial/multilayer-perceptrons-in-machine-learning
- Chen, T., & Guestrin, C. (2016). XGBoost: a Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, 1(1), 785–794. https://doi.org/10.1145/2939672.2939785
- http://machinelearningmastery.com/xgboost-for-regression/