

# A machine learning approach to map and predict inequalities in COVID-19 vaccine uptake in Scotland

Sohail Ferdous

2026-01-01

## Background

Coronavirus disease 2019 (COVID-19) is a respiratory illness caused by the SARS-CoV-2 virus, and it primarily spreads through respiratory droplets. Common symptoms include fatigue, cough, muscle aches; and severe infections may lead to pneumonia, acute respiratory distress syndrome, and death. COVID-19 was declared as a pandemic by the World Health Organization March 08, 2020, following which the UK imposed a nation-wide lockdown on March 23, 2020.<sup>1</sup> This was followed by large-scale disruption of society which led to devastating social and economic consequences, affecting almost every aspect of daily life.<sup>2</sup> Almost 6 years since the initial declaration, the long-term societal impacts of COVID-19 are still being studied, with multiple population studies increasing their frequency of data collection to study the effect of the pandemic on the lives of their participants.<sup>3</sup>

The COVID-19 pandemic also induced an unprecedented acceleration in vaccine development and testing. The Scottish Government introduced a national vaccination programme for COVID-19 on December 08, 2020, the first phase providing two doses of the vaccine to all adults, young people aged 12 to 17 suffering from comorbidities putting them at high-risk of COVID-19, and young people aged 12 years and over who lived with immunocompromised persons.<sup>4</sup>

As of December 07, 2025, there have been a total 779 million COVID-19 infections globally, 25.1 million of which are from the UK.<sup>5</sup> Currently, the disease has been downgraded to endemic status, and, for autumn 2026-spring 2027 season, the Joint Committee on Vaccination and Immunisation (JCVI) recommends vaccination to adults 75 years and older, care home residents for older adults, and individuals over 6 months age who are immunosuppressed.<sup>6</sup>

This report serves as a **proof-of-concept mini-project** for the PhD project “Developing Novel Data-Driven Tools and Methodologies to Understand Inequalities in Maternity Vaccination Uptake in Scotland”, which aims to investigate variations in maternal vaccine uptake in Scotland. The objectives of this report include exploring existing patterns of COVID-19 vaccine uptake in Scotland according to socioeconomic status and ethnicity as of June 18, 2023; and to train a random forest machine learning algorithm which would predict areas of low vaccine uptake based on population size, health board, and socioeconomic status/ethnicity. Through these objectives, this report attempts to justify the role of artificial intelligence tools when studying inequalities in maternal vaccine uptake, and to demonstrate the author’s statistical, coding, and scientific communication skills.

## Methods

This project follows an ecological (population-level) design using publicly available COVID-19 vaccine uptake data in Scotland. Health board level vaccination count by ethnicity and Scottish Index of Multiple Deprivation (SIMD) deciles are made available by Public Health Scotland under the UK Open Government

Licence (OGL) and the data is available for download from their website ([link](#)).<sup>7</sup> Vaccination counts were available at two time points: 2023-01-29, and 2023-06-18. Since both datasets (SIMD decile and ethnicity) were available at health board level, linking them was not possible, and hence, two identical analyses have been conducted in parallel.

Analysis was conducted using R statistical programming language (4.5.2)<sup>8</sup> using the RStudio<sup>9</sup> graphical user interface. This report was generated using R Markdown.<sup>10</sup>

## The analysis

### 1. Data import

Data files and relevant libraries were imported to R. This was followed by a quick visual inspection of the available data.

```
# 1.1 Loading libraries =====
library(tidyverse)
library(tidymodels)
library(ranger)
library(janitor)
library(lubridate)

# 1.2 Importing data =====
data_simd <- read.csv("covid_simd_sb_20231806.csv")
data_ethnicity <- read.csv("covid_ethnicity_sb_20231806.csv")

# 1.3 Checking imported data =====
view(data_simd)
glimpse(data_simd)

view(data_ethnicity)
glimpse(data_ethnicity)
```

### 2. Data preparation

Data preparation consisted of the following steps:

- Renaming columns
- Standardising date formats - date was reported in both yyyy-mm-dd and yyyy-dd-mm formats, this step standardises it to yyyy-mm-dd, and transforms variable type to date.
- Removing rows with aggregated Scotland data, and unknown SIMD deciles or ethnicity values
- Transforming variables and calculating number of unvaccinated people, and total proportion of vaccine uptake per row.
- Creating subsets with columns relevant to the analysis.

```
# 2.1 Cleaning column names =====
data2_simd <- clean_names(data_simd)
data2_ethnicity <- clean_names(data_ethnicity)

# 2.2 Date is reported in both yyyymmdd and yyyyddmm formats - standardising to yyyymmdd
# format and changing variable type to date =====
data2_simd <- data2_simd %>%
```

```

    mutate(date_new = parse_date_time(date, orders = c("Ymd", "Ydm")),
           date_new = as.Date(date_new))

data2_ethnicity <- data2_ethnicity %>%
  mutate(date_new = parse_date_time(date, orders = c("Ymd", "Ydm")),
         date_new = as.Date(date_new))

# 2.3 Removing rows for incompatible and unknown health board, ethnicity, and simd
# variables =====
clean_simd <- data2_simd %>%
  filter(hb_name != "Scotland",
         hb_name != "Unknown",
         simd_decile != "Not Known")

clean_ethnicity <- data2_ethnicity %>%
  filter(hb_name != "Scotland",
         hb_name != "Unknown",
         ethnicity != "Not Known")

# 2.4 Factorising variables and calculating vaccination counts =====
clean_simd <- clean_simd %>%
  mutate(hbname = factor(hb_name),
         simd_decile = factor(simd_decile, levels = c("10", "9", "8", "7", "6", "5", "4", "3",
                                                       "2", "1"), ordered = TRUE),
         n_vaccinated = count,
         n_unvaccinated = population - count,
         prop_vaccinated = n_vaccinated / population)

clean_ethnicity <- clean_ethnicity %>%
  mutate(hbname = factor(hb_name),
         ethnicity =
           factor(ethnicity,
                  levels = c("African, African Scottish or African British",
                            "Other African",
                            "Bangladeshi, Bangladeshi Scottish or Bangladeshi British",
                            "Chinese, Chinese Scottish or Chinese British",
                            "Indian, Indian Scottish or Indian British",
                            "Other Asian, Asian Scottish or Asian British",
                            "Pakistani, Pakistani Scottish or Pakistani British",
                            "Black, Black Scottish or Black British",
                            "Caribbean, Caribbean Scottish or Caribbean British",
                            "Other Caribbean or Black",
                            "Any mixed or multiple ethnic groups",
                            "Arab, Arab Scottish or Arab British",
                            "Other ethnic group",
                            "Gypsy/Traveller",
                            "Irish",
                            "Other British",
                            "Other white ethnic group",
                            "Polish",
                            "Scottish"))),
         n_vaccinated = count,
         n_unvaccinated = population - count,

```

```

    prop_vaccinated = n_vaccinated / population)

# 2.5 Creating subsets only with required variables and renaming the date_new variable ===
subset_simd <- subset(clean_simd, select = c(date_new, hb_name, simd_decile, population,
                                              n_vaccinated, n_unvaccinated,
                                              prop_vaccinated))

subset_simd <- subset_simd %>%
  rename(date = date_new)

subset_ethnicity <- subset(clean_ethnicity, select = c(date_new, hb_name, ethnicity,
                                                       population, n_vaccinated,
                                                       n_unvaccinated, prop_vaccinated))

subset_ethnicity <- subset_ethnicity %>%
  rename(date = date_new)

```

### 3. Patterns of vaccine uptake (as of June 18, 2023)

- Patterns of COVID-19 vaccine uptake across Scotland stratified by SIMD deciles and ethnicities were visually examined using **histograms**.
- **One-way ANOVA tests** were conducted to check for differences in mean uptake values across different ethnicities and SIMD deciles.
- For ANOVA tests, assumptions of normality were visually inspected by plotting test residuals in histograms.

```

# 3.1 Plot - Vaccine uptake by SIMD deciles/ethnicity =====
#
## 3.1.1 Preparing data ~~~~~
subset_simd_plot <- subset_simd %>%
  group_by(simd_decile) %>%
  summarise(total_vaccinated = sum(n_vaccinated),
            total_population = sum(population),
            prop_vaccinated = total_vaccinated / total_population,
            .groups = "drop")

subset_ethnicity_plot <- subset_ethnicity %>%
  group_by(ethnicity) %>%
  summarise(total_vaccinated = sum(n_vaccinated),
            total_population = sum(population),
            prop_vaccinated = total_vaccinated / total_population,
            .groups = "drop")

## 3.1.2 Plot code ~~~~~
simd_plot <- ggplot(subset_simd_plot, aes(x = simd_decile, y = prop_vaccinated)) +
  geom_col() +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1)) +
  labs(title = "Figure 2: COVID-19 vaccine uptake by SIMD decile (as of June 18, 2023)",
       x = "SIMD decile (10 = least deprived)",
       y = "Uptake") +
  theme_minimal() +
  theme(plot.title = element_text(size = 11))

```

```

ethnicity_plot <- ggplot(subset_ethnicity_plot,
                         aes(x = reorder(ethnicity, -prop_vaccinated),
                             y = prop_vaccinated)) +
  geom_col() +
  scale_y_continuous(labels = scales::percent_format(accuracy = 1)) +
  labs(title = "Figure 3: COVID-19 vaccine uptake by ethnicity (as of June 18, 2023)",
       x = "Ethnicity",
       y = "Uptake") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(size = 11))

# 3.2 ANOVA testing for mean vaccine uptake and simd deciles/ethnicity =====
simd_aov <- aov(prop_vaccinated ~ simd_decile, data = subset_simd)
ethnicity_aov <- aov(prop_vaccinated ~ ethnicity, data = subset_ethnicity)

# Checking for assumption of normality by plotting histogram of ANOVA test residuals
simd_resid <- residuals(simd_aov)
# hist(simd_resid) # Remove '#' to generate histogram

ethnicity_resid <- residuals(ethnicity_aov)
# hist(ethnicity_resid) # Remove '#' to generate histogram

```

#### 4. Machine learning analysis

- For this demonstration, low vaccine uptake was arbitrarily defined as uptake **less than or equal to the 25th percentile** (i.e., within the first quartile).
  - A **random forest** machine learning algorithm was built for the analysis. Breiman defined random forests as “a classifier consisting of a collection of tree-structured classifiers where random choices are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input  $x$ .<sup>11</sup> Essentially, random forests are a combination of decision trees and bagging, with the added benefit of feature randomness, which helps it create uncorrelated decision trees, thus ultimately reducing variance and improving predictive performance.
- In simple terms, it is a machine learning algorithm that is made of multiple decision trees, and it reaches a single result with the help of majority voting (Figure 1).<sup>12</sup>

# Random Forest

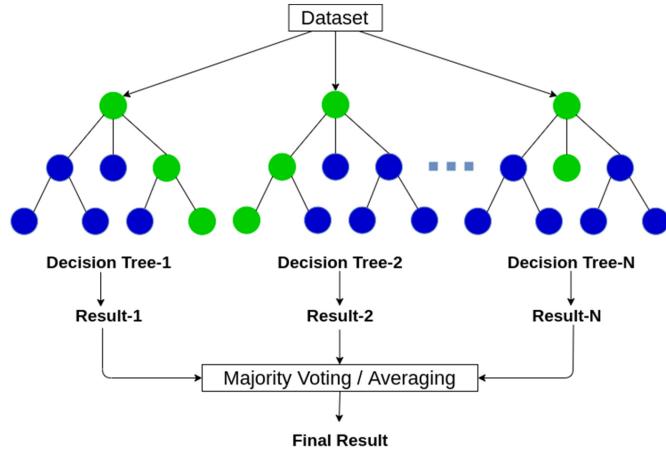


Figure 1: A visual depiction of a random forest algorithm. Source: <https://anasbrital98.github.io/assets/img/20/random-forest.jpg>

- Tree count and ranger implementation of the algorithm were selected after consulting resources on machine learning.<sup>13,14</sup> In general, higher tree counts reduce error rate, however, they may prove unnecessarily inefficient for large datasets. The mode was set to “classification” due to the categorical nature of the outcome variable (yes/no for low uptake).
- A **supervised learning** approach was selected for this analysis. This involves providing the algorithm a “ground truth” to adjust its parameters until the algorithm has been fitted appropriately.<sup>15</sup> This is done by labelling training dataset with correct outcomes - in our scenario, this refers to classifying each row as yes/no for low vaccine uptake.
- Vaccine uptake data were available for two dates: January 29, 2023, and June 08, 2023. Based on this, (SIMD and ethnicity) datasets were split into **training data (2023-01-29)**, and **test data (2023-06-18)**. The training data would be used to train the machine learning algorithm, and the test data would be used to test algorithm performance.
- Following splitting of datasets, algorithm parameters were defined, and the algorithms were trained on the training datasets.
- After training, the algorithms were asked to calculate probabilities and make predictions for low vaccine uptake on the test dataset (output displayed on page 12).

```

# 4.1 Calculating low coverage threshold =====
threshold_simd <- subset_simd %>%
  summarise(q25 = quantile(prop_vaccinated, 0.25, na.rm = TRUE)) %>%
  pull(q25)

print(threshold_simd) # 71.07%

threshold_ethnicity <- subset_ethnicity %>%
  summarise(q25 = quantile(prop_vaccinated, 0.25, na.rm = TRUE)) %>%
  pull(q25)
  
```

```

print(threshold_ethnicity) # 44.18%

# 4.2 Labelling low uptake rows =====
data_ml_simd <- subset_simd %>%
  mutate(low_uptake = prop_vaccinated < threshold_simd,
        low_uptake = factor(low_uptake, levels = c(TRUE, FALSE), labels = c("yes","no")))

data_ml_ethnicity <- subset_ethnicity %>%
  mutate(low_uptake = prop_vaccinated < threshold_ethnicity,
        low_uptake = factor(low_uptake, levels = c(TRUE, FALSE), labels = c("yes","no")))

# 4.3 Splitting datasets based on date =====
set.seed(123) # Specifying the random seed to make the analysis reproducible

simd_train <- data_ml_simd %>%
  filter(date == "2023-01-29")

simd_test <- data_ml_simd %>%
  filter(date == "2023-06-18")

ethnicity_train <- data_ml_ethnicity %>%
  filter(date == "2023-01-29")

ethnicity_test <- data_ml_ethnicity %>%
  filter(date == "2023-06-18")

# 4.4 Creating a recipe which clearly defines the explanatory and outcome variables =====
simd_recipe <- recipe(low_uptake ~ simd_decile + hb_name + population, data = simd_train)

ethnicity_recipe <- recipe(low_uptake ~ ethnicity + hb_name + population,
                           data = ethnicity_train)

# 4.5 Specifying algorithm specifications =====
algorithm_spec <- rand_forest(trees = 500) %>%
  set_engine("ranger") %>%
  set_mode("classification")

# 4.6 Creating workflows (could consider these untrained algorithms) by combining recipes
# and algorithm specifications =====
simd_ml_workflow <- workflow() %>%
  add_model(algorithm_spec) %>%
  add_recipe(simd_recipe)

ethnicity_ml_workflow <- workflow() %>%
  add_model(algorithm_spec) %>%
  add_recipe(ethnicity_recipe)

# 4.7 Training the algorithms =====
simd_fit <- simd_ml_workflow %>%
  fit(data = simd_train)

ethnicity_fit <- ethnicity_ml_workflow %>%
  fit(data = ethnicity_train)

```

```

# 4.8 Testing trained algorithms =====
#
## 4.8.1 Calculating probabilities for low_uptake ~~~~~
simd_prob <- predict(simd_fit, simd_test, type = "prob")
ethnicity_prob <- predict(ethnicity_fit, ethnicity_test, type = "prob")

## 4.8.2 Calculating binary yes/no predictions for low_uptake ~~~~~
simd_cate <- predict(simd_fit, simd_test, type = "class")
ethnicity_cate <- predict(ethnicity_fit, ethnicity_test, type = "class")

# 4.9 Combining the predictions with test data to view the output (this might appear a bit
# complicated as I am trying to rearrange and bind columns in one step) =====
simd_prediction <- simd_test %>% select(hb_name, simd_decile, low_uptake) %>%
  bind_cols(simd_cate) %>%
  bind_cols(simd_prob) %>%
  bind_cols(simd_test %>% select(date))

ethnicity_prediction <- ethnicity_test %>% select(hb_name, ethnicity, low_uptake) %>%
  bind_cols(ethnicity_cate) %>%
  bind_cols(ethnicity_prob) %>%
  bind_cols(ethnicity_test %>% select(date))

```

## 5. Algorithm performance and diagnostics

- Algorithm performance in predicting low vaccine uptake in the test dataset was evaluated using standard tests for classifier models.<sup>16,17</sup>
- Confusion matrices** were generated to view the raw ground truth vs model performance metrics.
- For quantitative estimates of performance, **accuracy** and **Cohen's kappa statistic** were generated. Accuracy refers to the proportion of correct predictions, whereas, kappa evaluates model performance by comparing algorithm predictions to predictions expected by random chance.
- Receiver operating characteristic (ROC) curves** were plotted to visually evaluate the algorithms' probability to estimate true positives.

```

# 5.1 Confusion matrices =====
simd_confusion <- simd_prediction %>%
  conf_mat(truth = low_uptake, estimate = .pred_class)

ethnicity_confusion <- ethnicity_prediction %>%
  conf_mat(truth = low_uptake, estimate = .pred_class)

# 5.2 Accuracy and kappa values =====
simd_perf <- simd_prediction %>%
  metrics(truth = low_uptake, estimate = .pred_class)

ethnicity_perf <- ethnicity_prediction %>%
  metrics(truth = low_uptake, estimate = .pred_class)

# 5.3 Area under curve calculation and ROC curves =====
simd_auc <- roc_auc(simd_prediction, truth = low_uptake, .pred_yes)
simd_roc <- simd_prediction %>%
  roc_curve(truth = low_uptake, .pred_yes) %>%
  autoplot() +

```

```

labs(title = "Figure 4: ROC curve for random forest predictions of low uptake for SIMD decile",
     subtitle = paste0("AUC = ", round(simd_auc$estimate, 3)),
     x = "1 - specificity",
     y = "Sensitivity") +
theme_minimal() +
theme(plot.title = element_text(size = 9),
      plot.subtitle = element_text(size = 9))

ethnicity_auc <- roc_auc(ethnicity_prediction, truth = low_uptake,.pred_yes)
ethnicity_roc <- ethnicity_prediction %>%
  roc_curve(truth = low_uptake, .pred_yes) %>%
  autoplot() +
  labs(title = "Figure 5: ROC curve for random forest predictions of low uptake for ethnicity",
       subtitle = paste0("AUC = ", round(ethnicity_auc$estimate, 3)),
       x = "1 - specificity",
       y = "Sensitivity") +
  theme_minimal() +
  theme(plot.title = element_text(size = 9),
        plot.subtitle = element_text(size = 9))

```

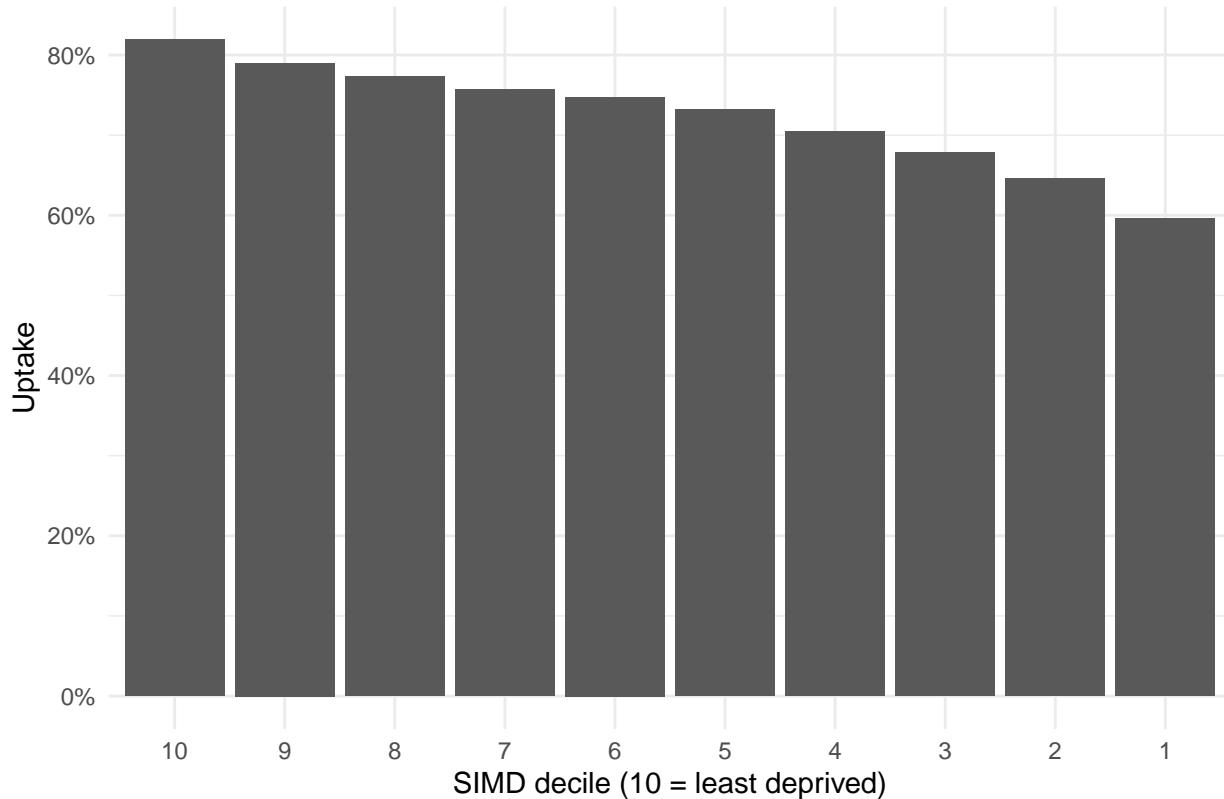
## Results and interpretation

### 1. Trends in COVID-19 vaccine uptake (as of June 18, 2023)

#### 1.1 SIMD deciles

Visual inspection of the histogram depicting vaccine uptake by SIMD levels revealed a consistent inverse relation between COVID-19 vaccine uptake and socioeconomic status (Figure 2). This association was confirmed to be statistically significant by the one-way ANOVA test ( $p<0.001$ ).

Figure 2: COVID-19 vaccine uptake by SIMD decile (as of June 18, 2023)



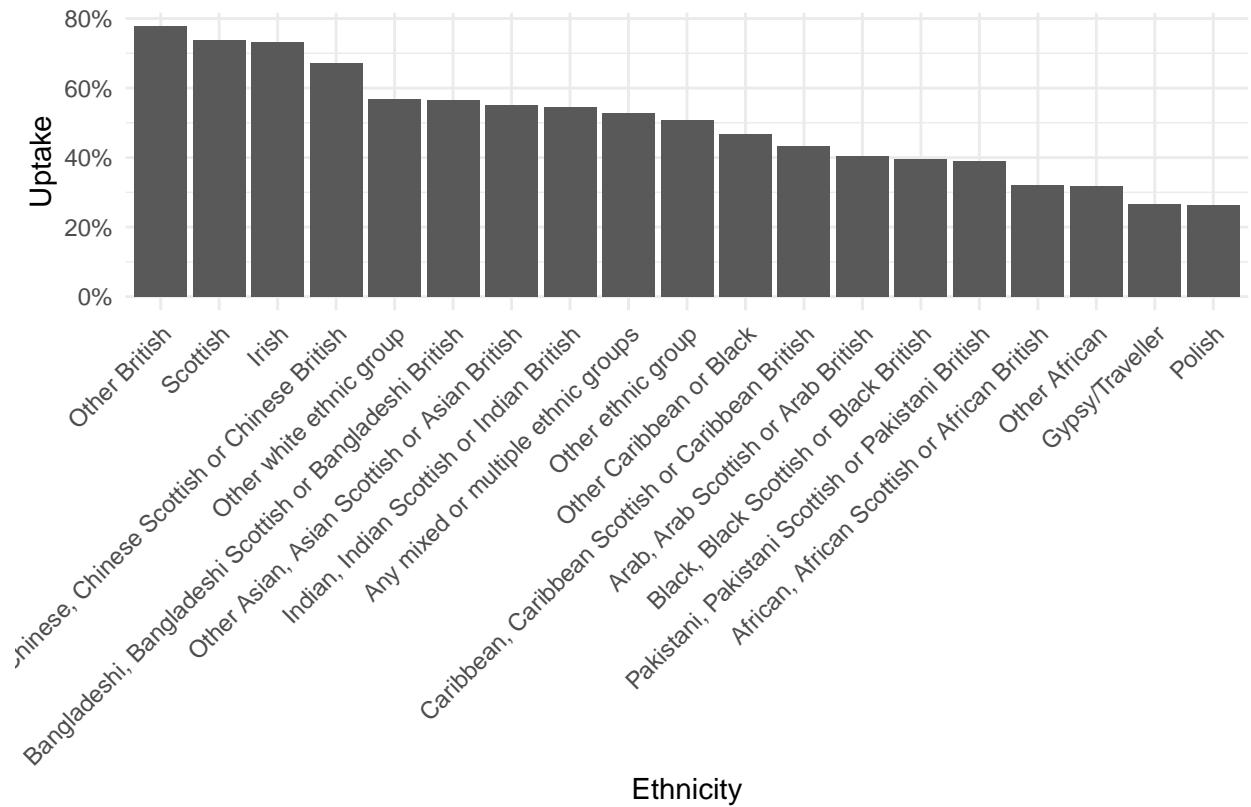
```
##
## ANOVA test
##
##           Df Sum Sq Mean Sq F value Pr(>F)
## simd_decile    9 0.7328 0.08142   50.42 <2e-16 ***
## Residuals    240 0.3876 0.00162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## 1.2 Ethnicity

As of June 18, 2023, Other British and Scottish ethnicities showed highest uptake of the COVID-19 vaccine, whereas Polish and Gypsy/Traveller ethnicities were among groups with lowest uptake (Figure 3). Differences in uptake among different ethnicities were statistically significant ( $p<0.001$ ).

Further investigation over possible causes for these inequalities is beyond the scope of this report due to the nature of available data. This particularly applies to ethnicity, where groups are not as directly comparable as SIMD deciles, and more granular data is required to capture complex relationships between different groups. For example, individual-level data may allow us to look at the interaction between demographic data, socioeconomic status and ethnicity, and allow us to derive more definitive sources of inequalities.

Figure 3: COVID–19 vaccine uptake by ethnicity (as of June 18, 2023)



```
##
## ANOVA test

##           Df Sum Sq Mean Sq F value Pr(>F)
## ethnicity    18  8.296  0.4609   60.51 <2e-16 ***
## Residuals   343  2.612  0.0076
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## 2. Predictions by random forest algorithm on test data

This output demonstrates predictions made by the machine learning algorithm. Column `low_uptake` represents “ground truth”, i.e., it was manually calculated using real data. Columns `.pred_yes` and `.pred_no` are probabilities for low vaccine uptake calculated by the trained algorithm. Column `.pred_class` gives the algorithm’s final prediction for low vaccine uptake. Only first few rows of the test dataset are displayed for brevity.

It should be noted that the ground truth (`low_uptake`) is **not required** in the test data - it is only required in the training data for supervised learning. I have included `low_uptake` column for test data for demonstrative purposes.

## 2.1 Random forest algorithm predictions on test data ( SIMD deciles)

```
##          hb_name simd_decile low_uptake .pred_class .pred_yes .pred_no      date
## 1 NHS Ayrshire and Arran           1       yes       yes 0.76863907 0.2313609 2023-06-18
## 2 NHS Ayrshire and Arran          10      no        no 0.07749391 0.9225061 2023-06-18
## 3 NHS Ayrshire and Arran           2      no        yes 0.73218272 0.2678173 2023-06-18
## 4 NHS Ayrshire and Arran           3      no        no 0.41127697 0.5887230 2023-06-18
## 5 NHS Ayrshire and Arran           4      no        no 0.19020532 0.8097947 2023-06-18
## 6 NHS Ayrshire and Arran           5      no        no 0.05580305 0.9441970 2023-06-18
```

## 2.2 Random forest algorithm predictions on test data (ethnicity)

Columns .pred\_yes, .pred\_no and date have not been printed to preserve readability.

```
##          hb_name          ethnicity low_uptake .pred_class
## 1 NHS Ayrshire and Arran African, African Scottish or African British no      yes
## 2 NHS Ayrshire and Arran Chinese, Chinese Scottish or Chinese British no      no
## 3 NHS Ayrshire and Arran Indian, Indian Scottish or Indian British yes      no
## 4 NHS Ayrshire and Arran Other Asian, Asian Scottish or Asian British no      no
## 5 NHS Ayrshire and Arran Pakistani, Pakistani Scottish or Pakistani British yes      no
## 6 NHS Ayrshire and Arran Any mixed or multiple ethnic groups no      no
```

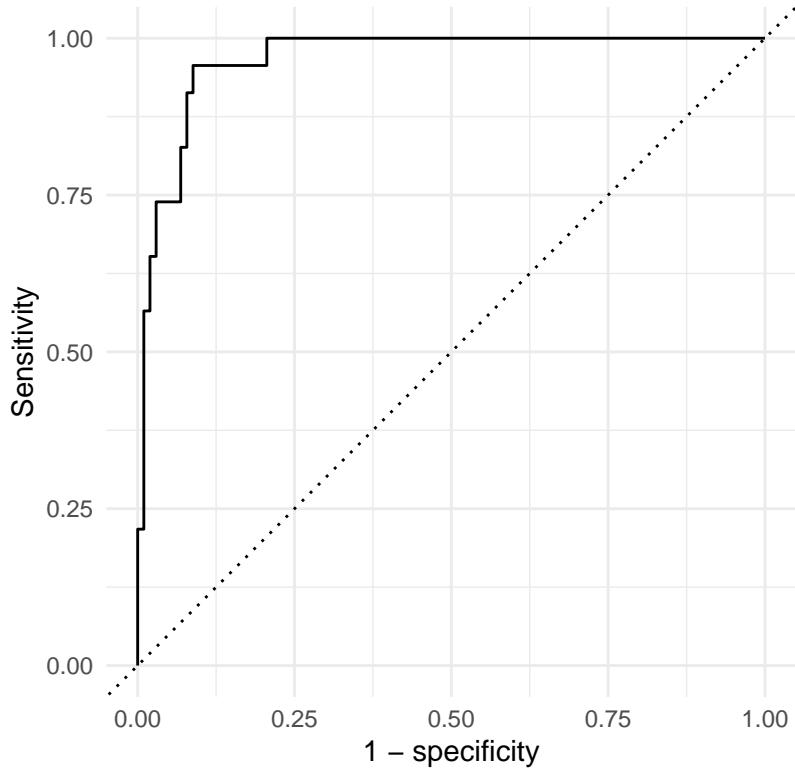
## 3. Algorithm performance

### 3.1 SIMD random forest algorithm

Overall, the SIMD random forest algorithm demonstrated encouraging performance metrics. Visual inspection of the confusion matrix show a majority of true positive and true negative predictions. This was quantified by high accuracy value (92%). Cohen's kappa (0.765) showed "substantial agreement" with ground truth beyond chance.<sup>18</sup> Finally, high area under curve (96.7%) suggested that the algorithm showed a high probability of correctly identifying true positives (Figure 4).

```
##  
## Confusion matrix  
  
##          Truth  
## Prediction yes no  
##       yes   22   9  
##       no    1 93  
  
##  
## Performance metrics  
  
## # A tibble: 2 x 3  
##   .metric  .estimator .estimate  
##   <chr>    <chr>        <dbl>  
## 1 accuracy binary     0.92  
## 2 kap         binary     0.765
```

Figure 4: ROC curve for random forest predictions of low uptake for SIMD decile  
AUC = 0.967



### 3.2 Ethnicity random forest algorithm

The ethnicity random forest model also demonstrated positive predictive capabilities, however, performance was weaker compared to the SIMD algorithm. The ethnicity algorithm showed 80% accuracy, and kappa (0.432) showed “moderate” agreement with the ground truth. Area under curve (82.1%) suggested a reasonably high probability of correct true positive predictions (Figure 5).

A few reasons could be hypothesised behind the discrepancy in model performances. Firstly, imbalanced group sizes may be responsible (especially relevant for kappa - the second paradox described by Feinstein and Cicchetti<sup>19</sup> ). Visual inspection of data frames **subset\_simd\_plot** and **subset\_ethnicity\_plot** reveal clear imbalances within ethnic group sizes, which might have affected performance. Secondly, there may be greater heterogeneity within ethnic groups in the real world, which may affect the algorithm’s predictive power. Larger, high quality (bias-free) training datasets at individual-level would likely benefit model performance. Finally, a person’s socioeconomic status may be a stronger predictor of vaccine uptake than their ethnicity in the real-world.

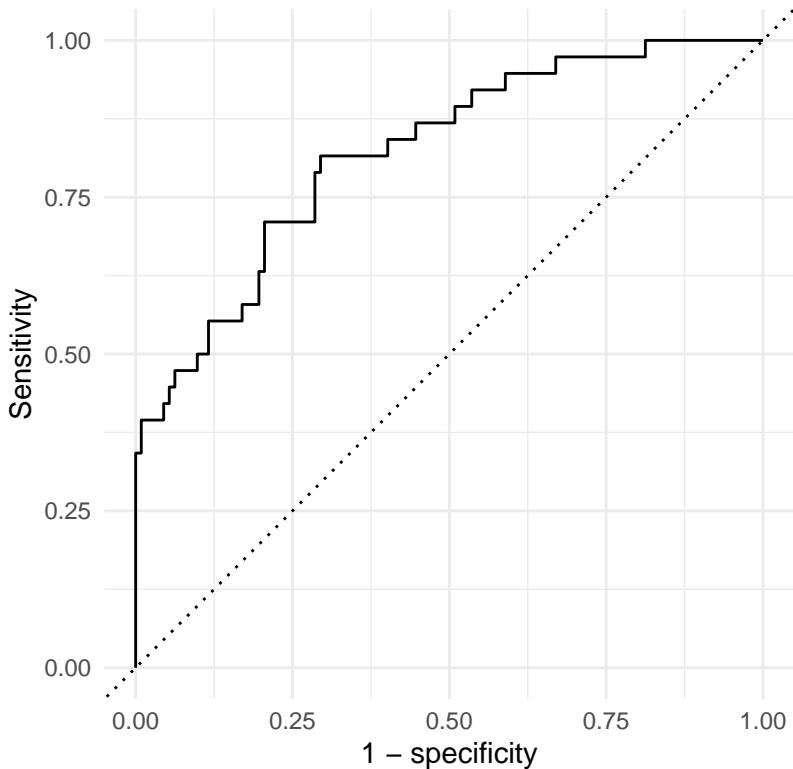
```
##  
## Confusion matrix  
  
##          Truth  
## Prediction yes no  
##      yes 19 11  
##      no   19 101  
  
##  
## Performance metrics
```

```

## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>        <dbl>
## 1 accuracy binary      0.8
## 2 kap       binary     0.432

```

Figure 5: ROC curve for random forest predictions of low uptake for ethnicity  
AUC = 0.821



## Conclusion

This report explored existing patterns of COVID-19 vaccine uptake in Scotland as of June 18, 2023, and described algorithms to predict low uptake. Greater socio-economic deprivation was generally associated with lower uptake, and ethnic groups showed significant differences in vaccine uptake.

Although models performed reasonably well, more detailed, high-quality data is required to reach definite conclusions and provide recommendations based on these results. Real-world causal chains (or rather, webs) are much more complicated than three explanatory variables, and thus mathematical models require a wide variety of data points to be sufficiently representative. Subsequently, as variety increases, so does the requirement of a large amount of data to maintain statistical power.

This mini-project aimed to provide a quantitative proof-of-concept regarding the use of artificial intelligence tools (namely, machine learning) in the study of vaccine uptake. Its shortcoming is where I would like to link the PhD project “Developing Novel Data-Driven Tools and Methodologies to Understand Inequalities in Maternity Vaccination Uptake in Scotland.” Individual-level health data available for this PhD project would intuitively overcome the main limitations faced by the ecological design of this mini-project. Additionally, qualitative interactions with primary stakeholders (mothers), care-givers, and healthcare providers would allow for insights and perspective over the data analysis which only a mixed-methods project could provide.

Recently, a population-based study in Scotland reported an uptake rate of 50.4% for the respiratory syncytial virus (RSV) maternal vaccine, and an effectiveness rate of 82.9% against RSV-related hospitalisation.<sup>20</sup> This moderate uptake, combined with promising effectiveness is a clear example of the untapped potential of vaccines, and the need to better understand underlying causes which affect their uptake.

## Notes

This work was independently created by the author with no inputs from the prospective supervisors.

The author would like to acknowledge the use of LLM tools (ChatGPT 5.1-Thinking) (23 December 2025) during the brainstorming phase of this project. All subsequent steps are the author's own contribution.

## References

1. UK Health Security Agency. COVID-19: The green book chapter. In: The Green Book [Internet]. UK Government; 2025. Available from: <https://www.gov.uk/government/publications/covid-19-the-green-book-chapter-14a>
2. British Academy. The COVID decade: Understanding the long-term societal impacts of COVID-19. In 2021.
3. CLOSER COVID-19. COVID-19 Longitudinal surveys [Internet]. [cited 2025 Dec 29]. Available from: <https://www.ukri.org/who-we-are/how-we-are-doing/research-outcomes-and-impact/esrc/covid-19s-impact-on-our-lives/>
4. Scottish Government. Inclusive vaccinations: Phase one of the COVID-19 vaccination programme [Internet]. The Scottish Government; 2022 [cited 2025 Dec 29]. Available from: [https://www.gov.scot/binaries/content/documents/govscot/publications/progress-report/2022/02/vaccine-inclusion-phase-one-covid-19-vaccination-programme/documents/inclusive-vaccinations-phase-one-covid-19-vaccination-programme/govscot%3Adocument/inclusive-vaccinations-phase-one-covid-19-vaccination-programme.pdf?forceDownload=true](https://www.gov.scot/binaries/content/documents/govscot/publications/progress-report/2022/02/vaccine-inclusion-phase-one-covid-19-vaccination-programme/documents/inclusive-vaccinations-phase-one-covid-19-vaccination-programme/inclusive-vaccinations-phase-one-covid-19-vaccination-programme/govscot%3Adocument/inclusive-vaccinations-phase-one-covid-19-vaccination-programme.pdf?forceDownload=true)
5. World Health Organization. WHO COVID-19 dashboard [Internet]. World Health Organization; 2025 [cited 2025 Dec 29]. Available from: <https://data.who.int/dashboards/covid19/cases?n=c>
6. Department of Health & Social Care. JCVI statement on COVID-19 vaccination in autumn 2026 and spring 2027 [Internet]. UK Government; 2025 [cited 2025 Dec 29]. Available from: <https://www.gov.uk/government/publications/covid-19-vaccination-in-autumn-2026-and-spring-2027-jcvi-advice-16-july-2025/jcvi-statement-on-covid-19-vaccination-in-autumn-2026-and-spring-2027>
7. Public Health Scotland. Flu & COVID vaccinations - Scottish Health and Social Care Open Data [Internet]. [cited 2025 Dec 29]. Available from: <https://www.opendata.nhs.scot/dataset/flu-covid-vaccinations>
8. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2025. Available from: <https://www.R-project.org/>
9. Posit team. RStudio: Integrated development environment for r [Internet]. Boston, MA: Posit Software, PBC; 2025. Available from: <http://www.posit.co/>

10. Allaire J, Xie Y, Dervieux C, McPherson J, Luraschi J, Ushey K, et al. Rmarkdown: Dynamic documents for r [Internet]. 2025. Available from: <https://github.com/rstudio/rmarkdown>
11. Breiman L. Random Forests. Machine Learning [Internet]. 2001 Oct;45(1):5–32. Available from: <https://doi.org/10.1023/A:1010933404324>
12. Kavlakoglu. What Is Random Forest? | IBM [Internet]. 2021 [cited 2025 Dec 29]. Available from: <https://www.ibm.com/think/topics/random-forest>
13. Random Forests · AFIT Data Science Lab R Programming Guide [Internet]. [cited 2025 Dec 29]. Available from: [https://afit-r.github.io/random\\_forests](https://afit-r.github.io/random_forests)
14. Kuhn M, Silge J. Tidy Modeling with R [Internet]. Available from: <https://www.tidyverse.org/>
15. Belcic I, Stryker C. What Is Supervised Learning? | IBM [Internet]. 2025 [cited 2025 Dec 29]. Available from: <https://www.ibm.com/think/topics/supervised-learning>
16. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. Scientific Reports [Internet]. 2024 Mar;14(1):6086. Available from: <https://doi.org/10.1038/s41598-024-56706-x>
17. Bajaj A. Performance Metrics in Machine Learning [Complete Guide] [Internet]. neptune.ai. 2022 [cited 2025 Dec 30]. Available from: <https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide>
18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977 Mar;33(1):159–74.
19. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology [Internet]. 1990;43(6):543–9. Available from: <https://www.sciencedirect.com/science/article/pii/089543569090158L>
20. McLachlan I, Robertson C, Morrison KE, McQueenie R, Hameed SS, Gibbons C, et al. Effectiveness of the maternal RSVpreF vaccine against severe disease in infants in Scotland, UK: A national, population-based case-control study and cohort analysis. The Lancet Infectious Diseases [Internet]. [cited 2025 Dec 30]; Available from: [https://doi.org/10.1016/S1473-3099\(25\)00624-3](https://doi.org/10.1016/S1473-3099(25)00624-3)