# Analyzing the Impact of Social Inequality on COVID-19 Vaccination Rates in the US

Conor Doyle, Juan Aguilar, Xueyao Zhao

Northeastern University DS3000 Fundamentals of Data Science

## Abstract

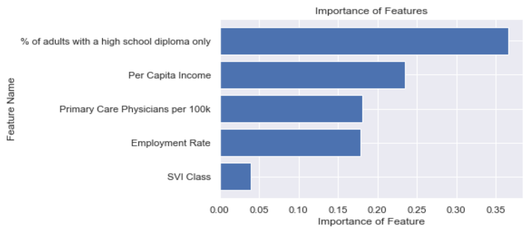
We aim to determine how various factors regarding social inequality affect COVID-19 vaccination rates in the US. The data was derived through multiple sources, and we tested 3 models to predict vaccination rates. After tuning each model with hyperparameter tuning and feature engineering, we determined the Random Forest Regressor was the best model to predict vaccination rates. We discussed how our findings can be used to provide more care to certain communities.

## Introduction

Vaccinations play a significant role in ensuring people's safety against the pandemic. According to the Centers for Disease Control and Prevention, "more than 670 million doses of COVID-19 vaccine" have been distributed since Dec 2020 (CDC, 2023). Although millions of vaccinations have been given, there are still disparities that can affect vaccination rates.

We wanted to explore the social issues affecting different US counties and how they affect vaccination. To represent these social issues, we found data on real income, SVI (Social Vulnerability Index), education rate, physician ratio, and unemployment rate.

We analyzed the relationship between the above variables across different US counties with vaccination rates to see how different attributes of a population would affect the vaccination rate per county. Understanding how social inequality can cause the disparity in vaccination rate will help us better understand how to distribute the vaccines and thus contribute to reducing the risk of having severe illness by ensuring all populations can get the vaccines they need. We aim to investigate if there is any correlation between the rate of people receiving vaccinations and their social status, and we used various parameters to represent social inequality. We are using this project to look at a social issue and its impact on one aspect relating to COVID-19. This project will allow us to look at the impact social inequality can have on vaccination rates and may allow us to see trends that can be used to investigate further. The results could potentially be used to determine if a certain area requires more resources and assistance.




*The importance scores for the features we selected.*

## Related Work

One example of related work to our analysis is an article posted in the Journal of Medical Internet Research, "Predictive Modeling of Vaccination Uptake in US Counties: A Machine Learning–Based Approach." The article can be found at this link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8623305/. Within this study, researchers from various institutions developed a model that can predict COVID-19 vaccinations across US counties with 62% accuracy. The model uses a variety of socioeconomic and demographic variables. The model is based on a Decision-Tree Regressor sourced from XGBoost, an open-source programming library with machine learning models.

## Methodology

**Data Acquisition**

To start off the analysis, we acquired data that are frequently considered social measurements from various sources. There are other factors that are interesting to explore, however, due to the unavailability to acquire the data since they might not be on the county level, we ended up using the data of:

Unemployment Rate

We got unemployment rates for each county from the US Bureau of Labor Statistics.

Education Rate

To quantify the education rate, we used data on the percentage of adults with a high school diploma only from 2017-21 from the U.S. Department of Agriculture.

Physician Data

To quantify physician data, we looked at the Primary Care Physicians Ratio from County Health Rankings.

SVI (Social Vulnerability Index)

SVI is a measure created by the CDC "which uses U.S. Census data on categories like poverty, housing, and vehicle access to estimate a community's ability to respond to and recover from disasters or disease outbreaks".

Vaccination Rates

To quantify vaccination rates, we used data from the CDC, and we found the percentage of the total population with at least one dose of the COVID-19 vaccination.

**Data Preparation**

After collecting the data, we first normalized the data to make sure they are on the same scale of measurement (the higher the better for all features):

- We calculated the Employment Rate by using 1 minus the Unemployment Rate
- We structured the Physician Data from the Physician ratio to the number of primary care physicians per 100k citizens
- We created dummies for SVI from 0-3, indicating high to low SVI classes

Besides the adjustment we made on data, we also conducted EDA on each data frame by removing duplicate records, unwanted rows like states' names (filtered out by using a list that contains names for all states) and introduction lines, and empty values, we merged all the data frames on the counties' names and repeated the same process to create the data frame we used to train and test models.

**Models**

We were looking to determine how various variables can affect the vaccination rate of a US county. To evaluate this, we used 3 regression models to determine vaccination rates. Initially, we selected "Per capita Income", "SVI Class", "employment rate", "number of primary care physicians per 100k citizens", and "percent of adults with a high school diploma" as our features. We used these features to predict our target, the percent of population with at least one COVID-19 dose. First, we used a random tree regressor model. Then we used a SVR (Support Vector Regression) model. For this model, we normalized the data using the StandardScaler from sklearn. The third model we used was the Decision Tree Regressor model from sklearn.

We decided to use MSE as our metric of evaluation as it well explains the overall quality of the regression model. For each model, we partitioned the data into a training and test set, where we used 30% of the data for the training set and then 70% of the data for the test set.

Random Forest Regressor

We tuned the random forest regressor model by utilizing feature selection and grid search. We first removed the SVI class (which has the lowest feature importance score of only 4%, while the education rate is the most important feature with an importance score of about 36%). After observing a minor decrease in MSE, we decided to proceed to perform grid search with and without SVI class to tune the model by testing various values on n estimators, max depth, min samples split, and min samples leaf, and then refine our model using the returned best parameters.

SVR

We used grid search to tune the model. Initially, we ran a model with minor parameter adjustments, and then we tested different values for kernel, gamma, and C. After we obtained the best model/parameters and the best score, we used these parameters to adjust our model.

Decision Tree Regressor

We tuned our Decision Tree Regressor model by instituting a grid search that tested different max depth, min sample split, and min sample leaf hyperparameters. The best parameters were chosen, and the model was adjusted accordingly.

## Results and Evaluation

Below is the MSE (mean square error) for each model before and after hyperparameter tuning.

| MSE | Random Forest Regressor | SVR | Decision Tree Regressor |
|---|---|---|---|
| MSE before hyperparameter tuning | 149.32 | 1565994475 9.51 | 247.69 |
| MSE after hyperparameter tuning | 143.10 | 223.32 | 154.77 |

Although the SVI class seems unimportant in the Random Forest Regressor as it has a feature importance score of only 4%, we found that the grid search performs slightly better with it included, so we decided to keep it in our model. The best parameters generated from grid search were a max depth of 7, min samples split of 5, n estimators of 200, and random state of 7.

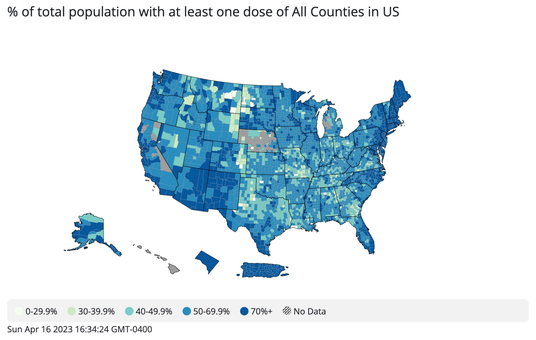
We suggested that the extremely high value for the MSE before model tuning for SVR is because we first ran the model with a linear kernel. After we adjusted its hyperparameters to have the RBF kernel, the model performed much better. From using grid search, we determined the best parameters were a C value of 5, a gamma of 0.1, and an RBF (radial basis function) kernel.

We added a grid search to our baseline Decision Tree model and determined the best parameters were a Max Depth of 3, Min Samples Leaf of 2, and Min Samples Split of 5. This gave an MSE of 154.77, and the difference in MSE before and after tuning the model exceeds the one for Random Forest Regressor. However, it is still not as good as our Random Forest Regressor, which initially performs better than all the other models.

Since the Random Forest Regressor performs more balanced before and after hyperparameter tuning and it has the best initial untuned MSE, we determined that it is the best model for predicting vaccination rates, having a MSE of 143.10 after tuning. We predicted and confirmed the education rate and per capita income have higher feature importance scores since these are general measurements of social well-being. However, SVI only having a score of 4% is unexpected. It might indicate that social vulnerability does not have a strong correlation with the vaccination rate. We supposed this might be due to the SVI index measures a picture too broad and general to be detail considered.

## The Impacts

This project allowed us to investigate the effect social inequality has on COVID-19 vaccination rates, which can allow us to identify potential solutions to address disparities in vaccination rates. The results could potentially be used by communities to determine what areas need more assistance with healthcare. It can also be used to see how improving an attribute of a county can lead to more accessibility to necessary healthcare such as vaccinations. Our results may contribute to improving access to other healthcare in the United States. If a county has a low vaccination rate, it is possible that there is other issues affecting the community.




## Conclusion

In conclusion, this project demonstrates how different social factors can impact how a community accesses and attains healthcare services. We found there is various barriers that may prevent how individual and its community access necessary healthcare.

By evaluating three different models, we determined that the Random Forest Regressor was the best model for predicting vaccination rates. However, its MSE is still comparatively high, indicating that the model might be missing explanatory features and requires more investigation into other possible social measurements. The models can be improved in the future by implementing more data of vaccination rates over time. It is also possible that there exists a better model that can better predict the vaccination rates which would better help identify opportunities for healthcare improvement. Overall, this project can be a valuable resource for healthcare professionals and law officials to help improve healthcare accessibility and equity.

## References

CDC. "Safety of COVID-19 Vaccines - Safety of COVID-19 Vaccines." Center for Disease Control and Prevention, 13 February 2023, https://www.cdc.gov/coronavirus/ 2019-ncov/vaccines/safety/safety-of-vaccines.html.

Education Rate: https://www.ers.usda.gov/data-products/county-level-data-sets/county-level-data-sets-download-data/

General Vaccine Trends: https://covid.cdc.gov/covid-data-tracker/ - vaccine-delivery-coverage

Physician Data: https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation

SVI: https://covid.cdc.gov/covid-data-tracker/ - vaccination-equity

Unemployment Rate: https://www.bls.gov/lau/tables.htm - cntyaa

www.PosterPresentations.com