**Introduction**

The COVID-19 global pandemic is a problem that has been impacting the world for over a year now. A problem that started of containing, testing, minimizing death, has now become an issue on how to create herd immunity. The problem now is the number of vaccines to help obtain immunity. Do socioeconomic indicators impact the number of people that are receiving the vaccine? This report outlines a study conducted to see if socioeconomic impacts are increasing the number of people vaccinated and is motivated by the impact COVID-19 has had on our lives. Starting with data collection and exploration, data preprocessing, data mining algorithms used for analysis, conclusions, followed by an appendix of modeling outputs.

**Data Collection & Exploration**

Initial dataset is a vaccination dataset by country sourced was downloaded from Kaggle[[1]](#footnote-1) and comes from the ‘Our World in Data’ GitHub. The dataset is a daily aggregate of vaccinations whenever there is an update. We loaded it into a SQL database to get an aggregate of the most up to date value by country. The socioeconomic data was sourced from World Bank and included an education[[2]](#footnote-2), GDP[[3]](#footnote-3), mortality[[4]](#footnote-4) , and population[[5]](#footnote-5) datasets. The clean country data set includes, country name, people fully vaccinated per hundred, 2019 GDP, 2016 Bachelors, 2019 Mortality Rate, and 2019 population. Previous years were used in case the more recent years’ data was not complete from the World Bank. Final dataset is relatively small for those reporting vaccination numbers.

Due to concerns of too small of a dataset to perform data mining algorithms, a United States county dataset was collected. The CDC provides county level reporting for COVID-19 vaccinations. Socioeconomic data at the county level was sourced from the Economic Research Service[[6]](#footnote-6) and sourced similar attributes as the country dataset. Poverty, education, death rates, and population. Clean dataset includes 2019 information for county name, percent of people fully vaccinated, percent bachelors, percent poverty, population, and death percentage. This dataset is what we have used for testing data mining algorithms because of the increased number of instances.

Figure

Figure

**Methodologies**

Data Preprocessing

This process transforms raw data into an understandable format. It improves accuracy by removing duplicates and inconsistencies in data. A major part of preprocessing this data set involved removing noisy data and ignoring the missing values. The following data preprocessing techniques were used:

Aggregation - Aggregated data not only reduces variability but requires less memory and improves processing times too. For this project, both the country and county were both aggregated as the most recent date.

Attribute Selection - Many algorithms work better when the dimensionality is lower. Hence, the dimensions of the dataset can be reduced by removing irrelevant and redundant data. For instance, features like ‘human\_development\_index’ and cardiovasc\_death\_rate’ were irrelevant to the prediction of impact of socioeconomic factors on COVID-19 vaccination. For this project, the irrelevant features were removed through domain knowledge and no specific filter was used.

Discretization - Both the dependent and independent variables being a numeric, Discretization was not required as it was the perfect fit for running linear regression algorithms.

Preprocessing methods in WEKA and in a Python Jupyter Notebook were similar and only varied on how we handled missing values.

Data Mining

**Linear Regression**

The two data mining algorithms we have chosen for this study were linear regression and regression trees. The linear regression was chosen as an ideal candidate for our dataset because we are curious if our input features that are numeric can predict the number of vaccines that is also numeric. Had there of been different nominal datatypes and solving a vaccine Y/N question, decision trees may be stronger candidates for modeling.

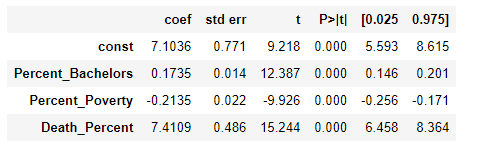
Using Python, we were able to achieve more detail about our models than what we receive from WEKA but have included the WEKA analysis as well.

Python

The clean dataset described earlier in this paper were read into a Jupyter Notebook. Nulls were dropped and the first algorithm used was a multiple linear regression model with education, poverty, and mortality as the features targeting percent of total vaccinations per hundred. The parameters used were a part of the ‘sklearn’ directory in Python and tested ‘train\_test\_split’ and random states cross validation.

Different models were ran using tests of 10%/20%/30%/40%achieved the highest fit, R2. 30% test and 70% training received the best results with a .16 R2 (figure 3). This is not a particularly high fit so the model is not the strongest, however we will need to do further analysis to see if our features are predictive for this model.

 Another parameter that was used was utilizing ‘random\_state’ which randomizes that data that is tested on like k-folds method. Testing the test/train split as well as adjusting the ‘random\_state’, this performance tuning method increased R2 to .21. Despite the overall model not being a great fit, we were able to tune the model to a better fit.

To analyze if socioeconomic features are predictive of the number of vaccines, we used another regression method ordinary least squares in python ‘stats.models’. This method minimizes the errors to help form the linear relationship. The interesting note from this model is that is summarizes all statistics. Shows that the P values are very small (figure 4), showing that we do have features that help predict an outcome. However, the fit is small. In this case it is better to have this model on the dataset is than not having a model at all and we can reject the null hypothesis.

Figure

Figure

WEKA

Loaded the same clean dataset into WEKA and replaced missing values. This differed than the Python preprocessing because it takes the mean to fill missing values rather than the dropping performed above. WEKA also gives us R instead of R2 . R is the correlation of predicted values while R2 indicates percentage of variation of our model that we explained above as not a good fit. Root mean squared error (RMSE) was also used a performance measure because it assists how close the data are to the model predicted values. Lower values indicate a better fit. In every test case there was a high RMSE, which in turn shows that not only is the model not a good fit, but each point was far away from the actual value. This model is not a good performer.

The model tuning performed was testing different test splits as well as using cross validation with various folds. The ideal outcome in both cases in WEKA was using lowest number of folds (3) and lowest number of test (10%) (figure 5).

Figure

**Tree**

Python

The clean dataset was loaded into a Jupyter Notebook. Nulls were dropped and the second algorithm tested was a regression tree. We used the same features of education, poverty, and mortality with a target of percent of total vaccinations per hundred. The parameters used were part of the ‘sklearn’ directory and utilized ‘train\_test\_split’ to fit a decision tree.

Using this method, we tested the same parameters of test 10%/20%/30%/40%. The accuracy was highest at 7.3% with 30% test and 70% train (figure 6). Again, not the results you are looking for in a predictive model, but still better than using no model at all.

Performancing measures we utilized changing the trees max depth. Max depth of 3 gave the best performance tuning results of 9.2% (figure 6). Notice when the tree does not grow high enough the accuracy is lower because there is too little flexibility to notice the patterns. Too high accuracy also goes down because it is overfit and the error increases.

Figure

WEKA

Loaded the same clean data set into WEKA and replaced missing values. Selected REPTree which is the linear regression decision tree. The same test cases were used with this algorithm, test 10%10%/20%/30%/40% as well as cross validation with different number of folds. The highest performance was using 7 folds with R of .3 and RMSE of 5.5 (figure 7). This is still a very poor model and has shown the predictability is not high in every test we performed.

Figure

**Logic of Problem**

Working through this analysis while using the logic of the problem toolset has been a positive experience. Before any data was collected, we identified a problem that is the COVID-10 pandemic. At the time of this class vaccine rollout was ramping up and sparked interest to see if there as an even distribution of vaccinations worldwide. Starting with a problem so large our first data checkpoint contained too many variables that were redundant. We sifted through those but were still running into issues come checkpoint three. It was apparent that our logic of problem needed to be updated with a larger dataset than countries which put us back at the information element of thought. We collected United State County data and were able to get right back on track. Key question and purpose stayed align in do socioeconomic factors increase the number of people vaccinated but using these steps we could refocus with more information and proceed through our analysis.

Pros of using the logic of problem framework is that is straight-forward in the process of solving a problem. For example, you start with a key question, purpose or motivation, information, then touch on what point of view or bias are you adding in. If you go back to gathering information you move down through the same question again and can use this framework to keep your goal in mind while solving a problem.

Two recommendations to be added to the logic of problem are what has been done regarding this problem already and to add measurability. Many issues are being tackled by many different people and it would be interesting as a step to gather information that others have analyzed already toward the question and purpose. It would also help justify writing the question in the first place to study something slightly different that what has already been produced. Measurability would assist during this framework to see if our goals of solving the problem are measurable or not.

**Conclusions**

While running through both linear regression and decision trees the same issues were apparent with the features that we have selected to use to predict vaccinations. Those issues are that we had a high variance of predicted outcomes that lead to a very poor fitting and low accuracy model. A pattern that showed up was the larger the test group we tested, the higher the variability on a large scale. This is expected but with a poor fitting model it was more apparent.

The Python models worked better in particular because they can be controlled and touch on more statistics that what WEKA outputs as a summary. Between the actual model’s, multiple linear regression using the sklearn package had the highest R2 values. This is not a surprised because all our data was numeric and linear regression is a good choice to help predict a numeric outcome.

At a United States county level, using education (% of pop that obtained a bachelor’s degree), poverty (% of poverty for that county), and mortality (% death per year), used to predict the total number of vaccinations per hundred turned to not show high predictability. The model had poor fitting using linear regression and less than 10% accuracy while using decision trees. That said, we still must reject the null hypothesis because the p values assigned to each of these three variables were very low. They still showed statistical significance as features in our model. Using this model with a poor fit is still better than no model at all.

If we were to continue this project, it would be exciting to find more rich datasets that included non-aggregated data and an individual level. Knowing socioeconomic variables at an individual level comparing their socioeconomic status to where they live could help truly answer the question of, “do socioeconomic factors increase the number of people getting vaccinated?” Even more interesting would have data of a specific home state and create a decision tree of who got vaccinated Y/N. Many questions and answers can be used to help us as we navigate towards ending this pandemic the next year and these studies would be very interesting with the data mining techniques we have learned in this course.

Works Cited

1. Preda, Gabriel. “COVID-19 World Vaccination Progress.” Kaggle, 14 Mar. 2021, www.kaggle.com/gpreda/covid-world-vaccination-progress.

2. “Education .” World Bank, World Bank, data.worldbank.org/topic/4.

3. GDP per Capita (Current US$).” World Bank, World Bank, data.worldbank.org/indicator/NY.GDP.PCAP.CD.

4. “Mortality Rate, Neonatal (per 1,000 Live Births).” World Bank, World Bank, data.worldbank.org/indicator/SH.DYN.NMRT.

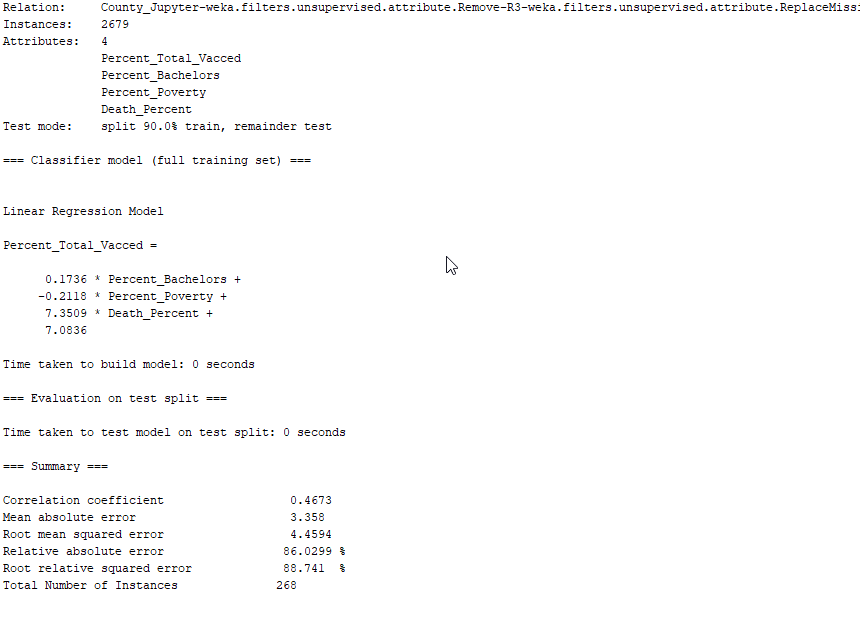
5. “CDC COVID Data Tracker.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, covid.cdc.gov/covid-data-tracker/#vaccinations.

6. “Download Data.” USDA ERS - Download Data, www.ers.usda.gov/data-products/county-level-data-sets/download- data/.

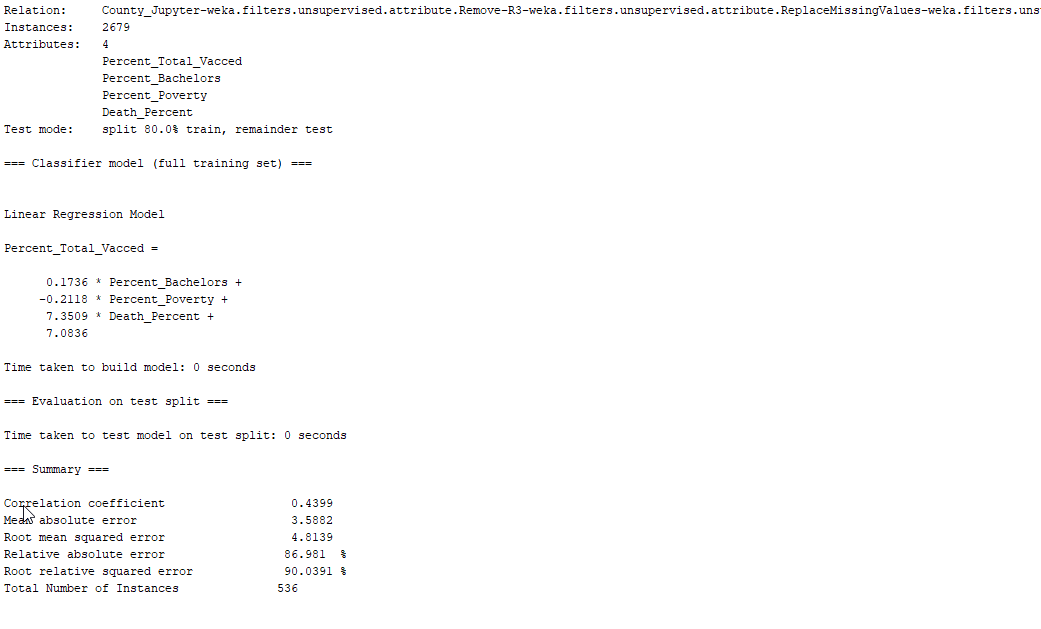
Appendix

WEKA Linear Regression

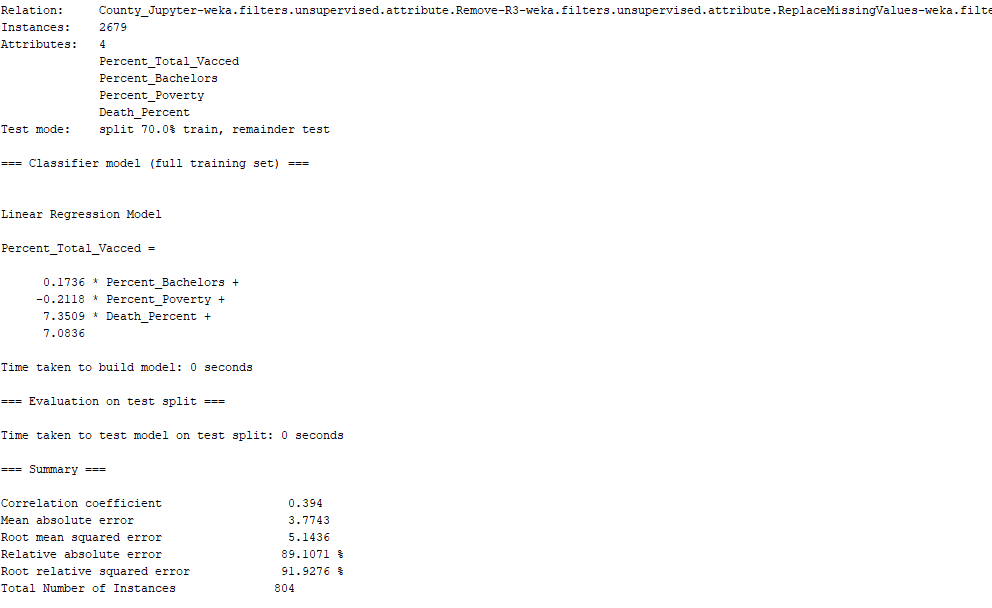
10% Test



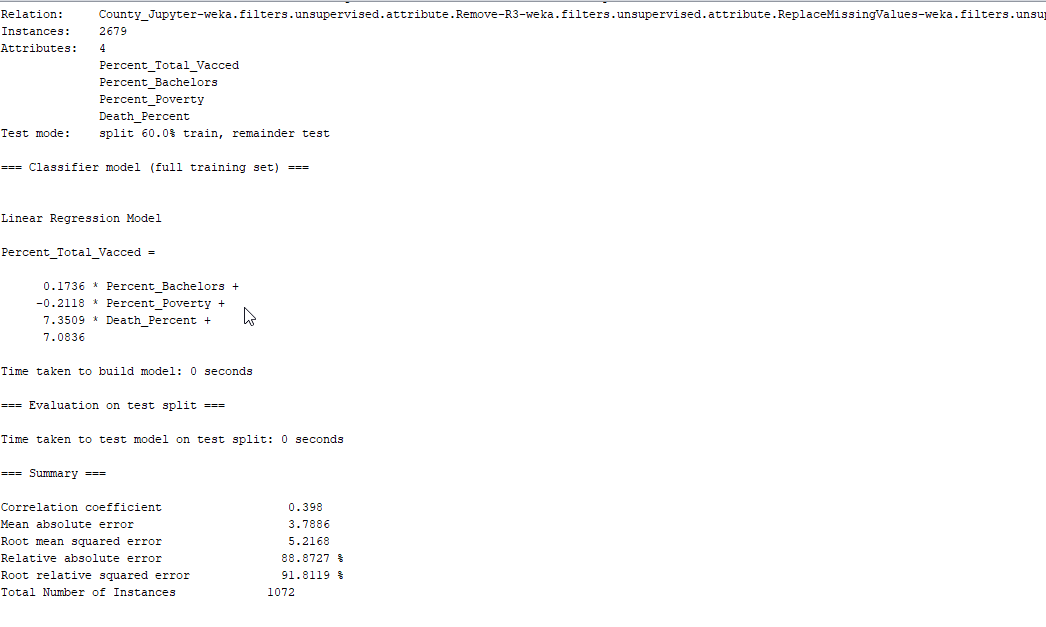
20 % Test



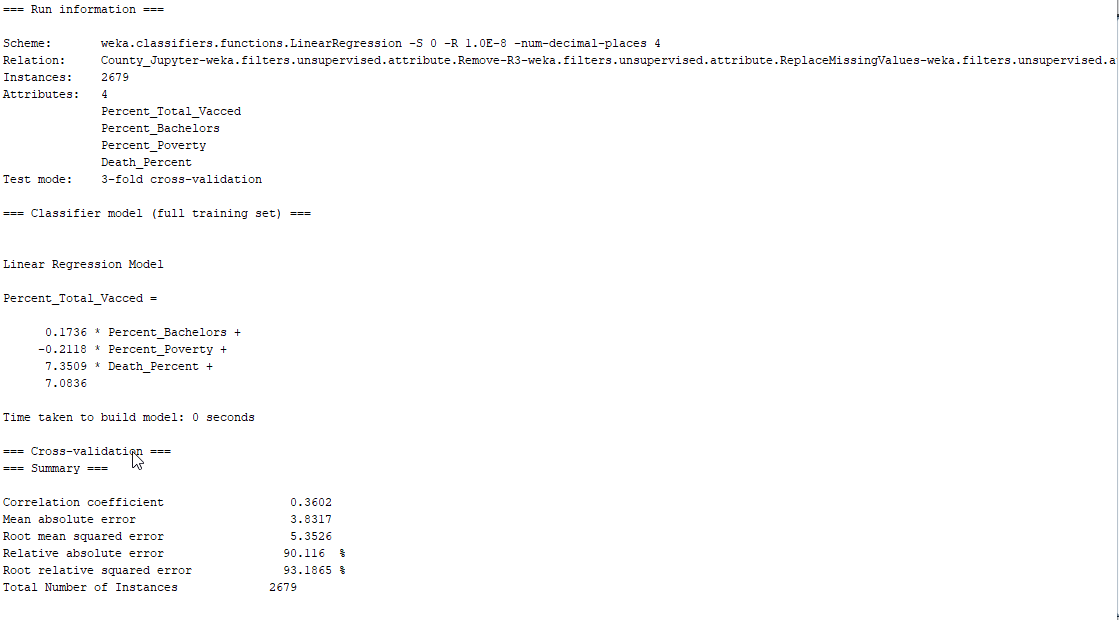
30% Test



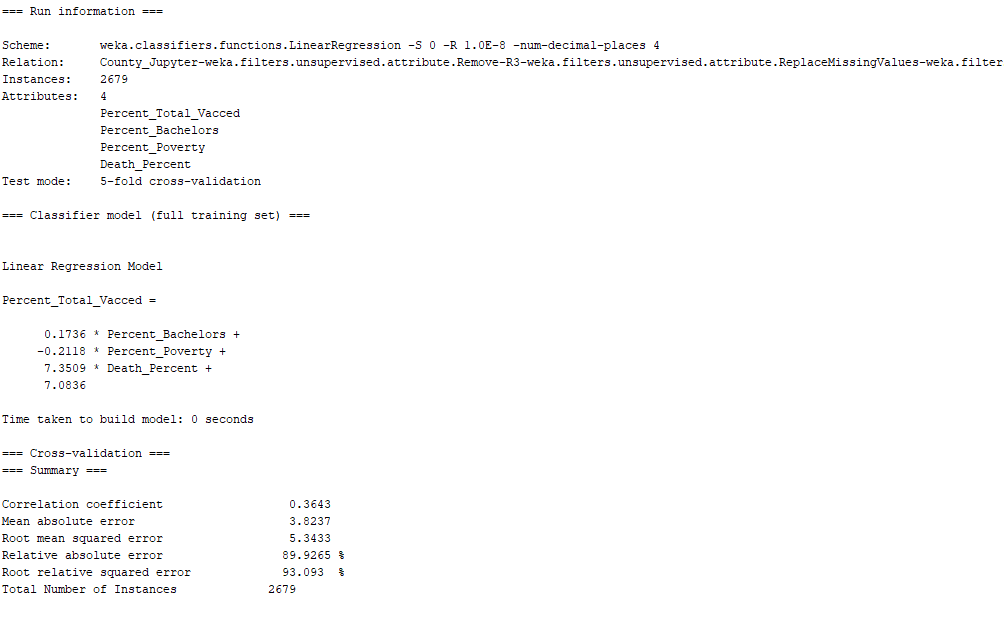
40% Test



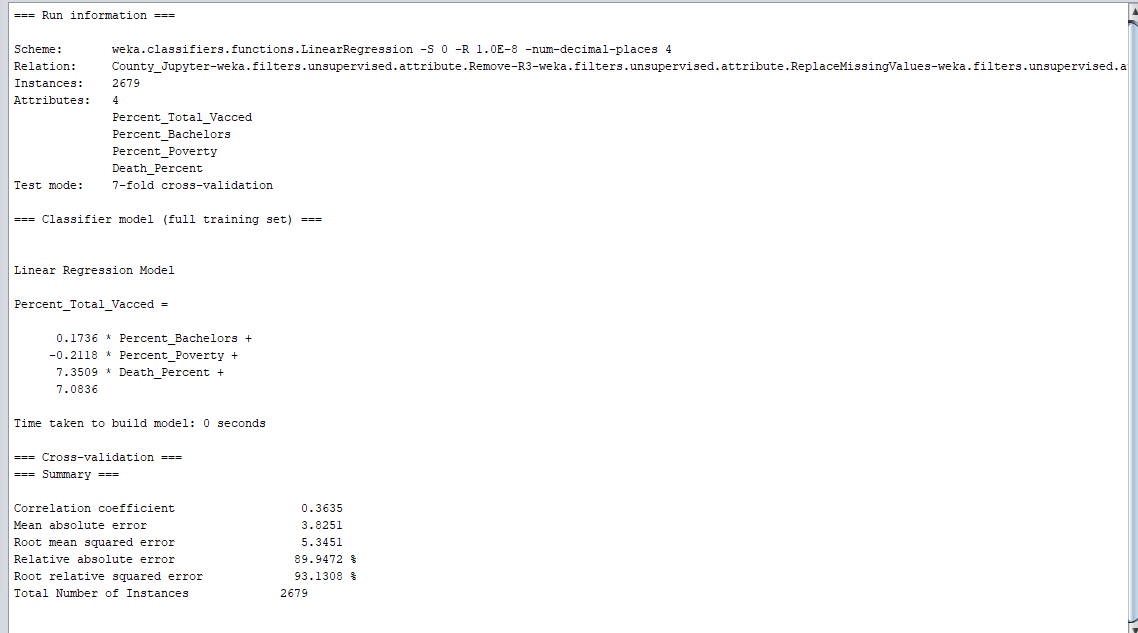
3 Folds



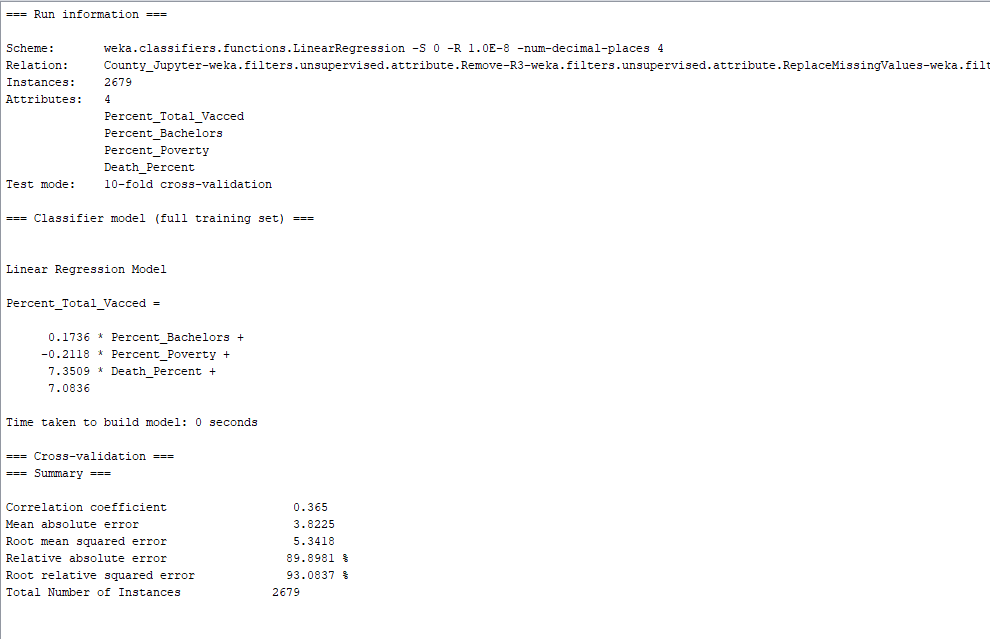
5 folds



7 folds

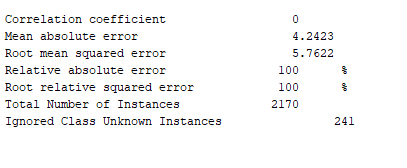


10 folds

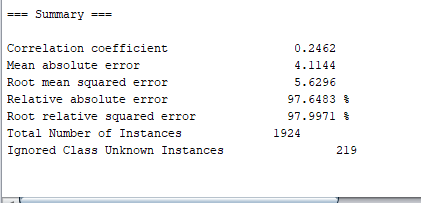


WEKA Regression Tree

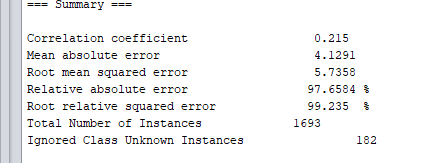
10% Test



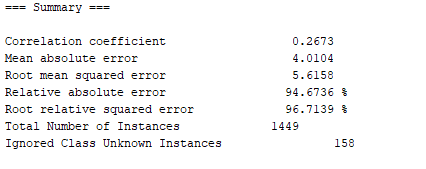
20% Test



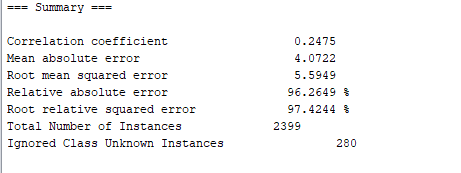
30% Test



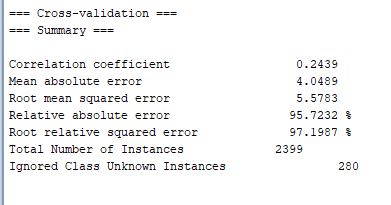
40% Test



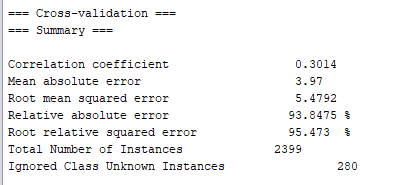
3 Folds



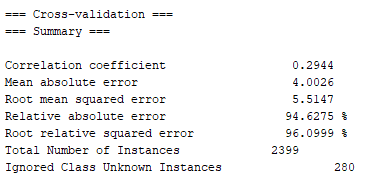
5 Folds



7 Folds

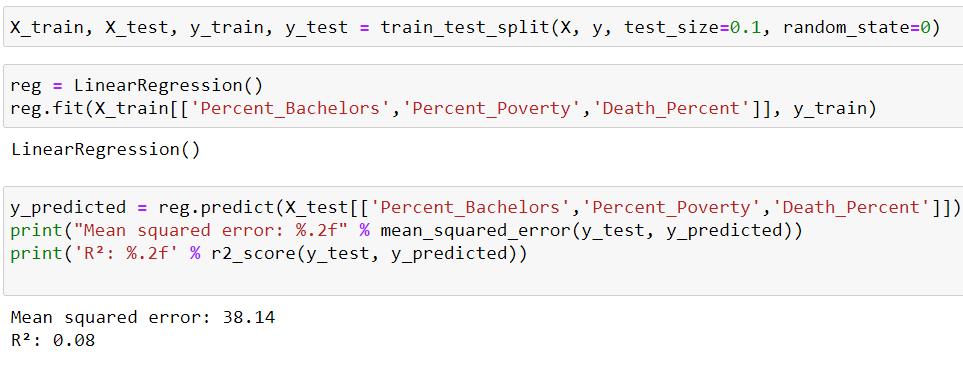


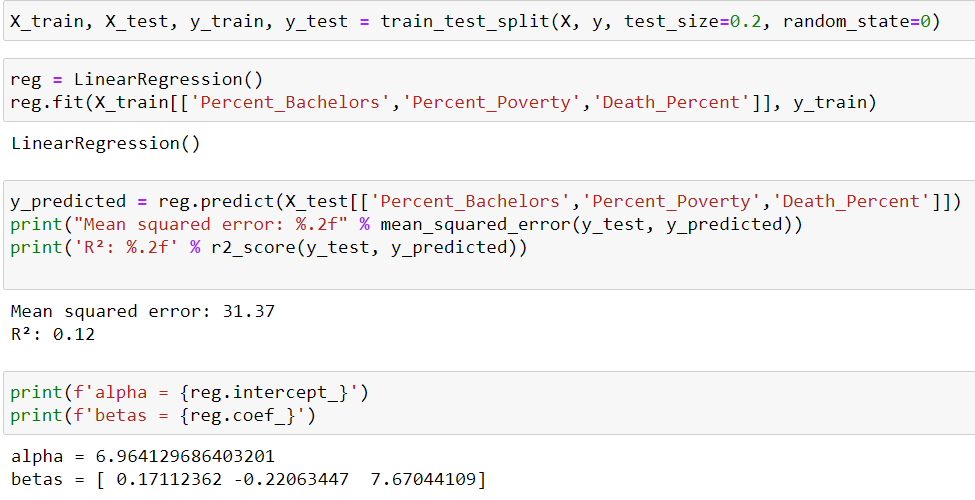
10 Folds



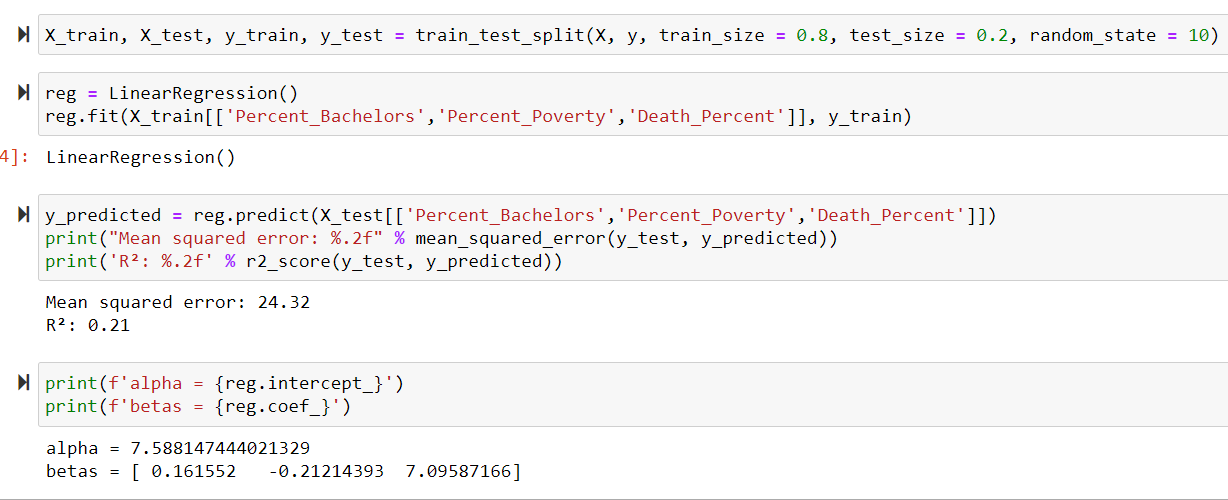
Python Jupyter Notebook

Linear Regression

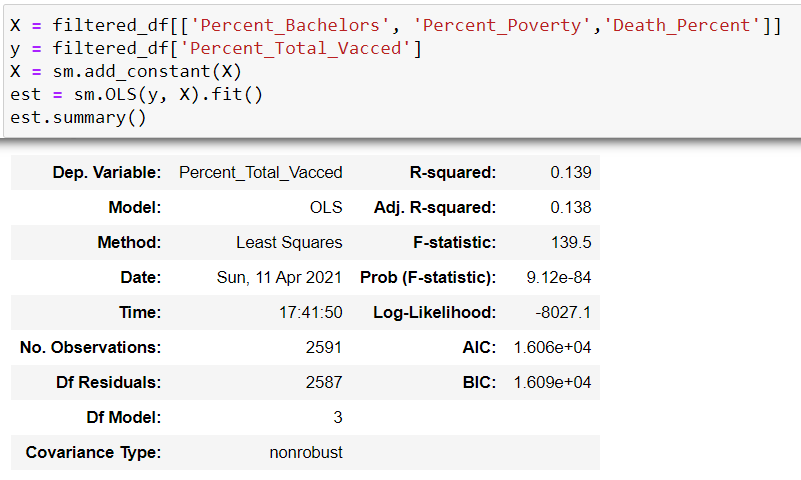


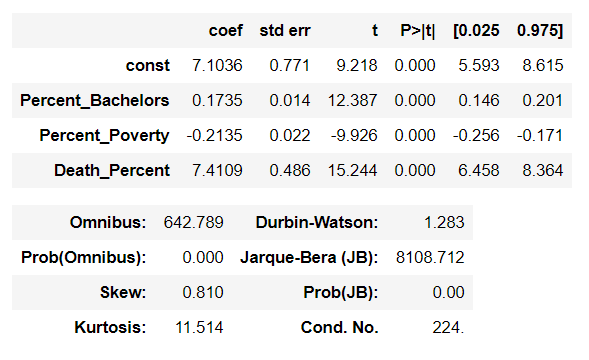


Linear regerssion random state

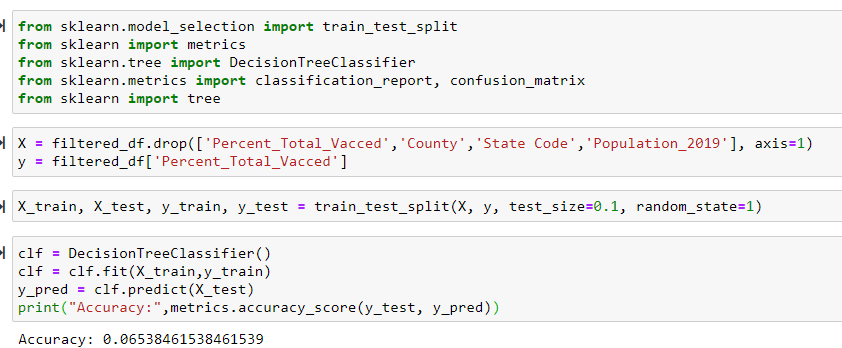


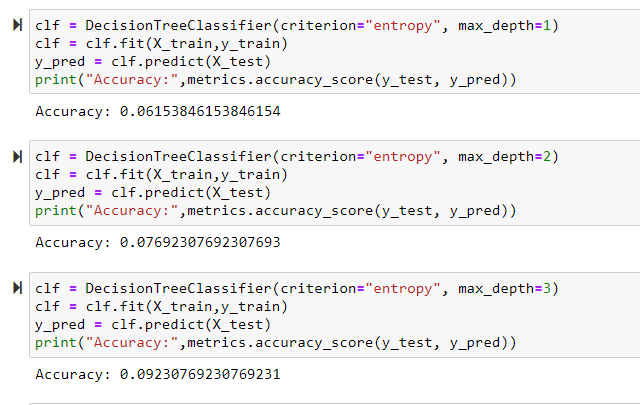
Ordinary least squares





Tree





1. Preda, Gabriel. “COVID-19 World Vaccination Progress.” Kaggle, 14 Mar. 2021, www.kaggle.com/gpreda/covid-world-vaccination-progress. [↑](#footnote-ref-1)
2. “Education .” World Bank, World Bank, data.worldbank.org/topic/4. [↑](#footnote-ref-2)
3. “GDP per Capita (Current US$).” World Bank, World Bank, data.worldbank.org/indicator/NY.GDP.PCAP.CD. [↑](#footnote-ref-3)
4. “Mortality Rate, Neonatal (per 1,000 Live Births).” World Bank, World Bank, data.worldbank.org/indicator/SH.DYN.NMRT. [↑](#footnote-ref-4)
5. “Population, Total.” Data, data.worldbank.org/indicator/SP.POP.TOTL. [↑](#footnote-ref-5)
6. “County-Level Data Sets.” USDA ERS - County-Level Data Sets, www.ers.usda.gov/data-products/county-level-data-sets/. [↑](#footnote-ref-6)