Benjamin Lee

Professor Amir Jafari

Introduction to Data Mining

28 April 2020

Individual Final Report: MMR Vaccination Data

Introduction

With all the news about coronavirus and the search for a possible vaccine, we believed it would be interesting to consider the challenges of convincing an entire population to embrace a vaccine. The MMR vaccine (measles, mumps and rubella) has been around for over forty years, and yet there are still many areas around the country where at least 95% of school children are not vaccinated as recommended by the World Health Organization. Research into the MMR vaccine seems especially relevant because scientists in France are using the current MMR vaccine as the backbone of their vaccine candidate.

Description of Individual Work and Background Information

The dataset used for our final project is concerned with MMR vaccination data for 32 states in the US. The data was originally compiled by the Wall Street Journal (WSJ) for an article written on school vaccination rates. A summary of the features of this data set are provided on the next page.

Attribute	Description	Optional?
index		
state	School's state	
county	School's county	Y
district	School's district	Y
name	School's name	
type	Whether a school is public, private, charter	Y
enroll	Enrollment*	Y
mmr	School's Measles, Mumps and Rubella (MMR) vaccination rate	Y
overall	School's overall vaccination rate	Y
xmed	Percentage of students exempted from vaccination for medical reasons	Y
xper	Percentage of students exempted from vaccination for personal reasons	Y
xrel	Percentage of students exempted from vaccination for religious reasons	Y
lat	School's latitude	(Only in individual state files)
lng	School's longitutde	(Only in individual state files)

**Depending on the state, enrollment is for kindergarten only or may extend to include other grades.*

Data was available for the following states:

2017 - 2018 School Year	2018 - 2019 School Year	
Colorado	Arizona	North Carolina
Connecticut	Arkansas	Ohio
Minnesota	California	Oklahoma
Montana	Florida	Oregon
New Jersey	Idaho	Rhode Island
New York	Illinois	Tennessee
North Dakota	Iowa	Texas
Pennsylvania	Maine	Vermont
South Dakota	Massachusetts	Virginia
Utah	Michigan	Wisconsin
Washington	Missouri	

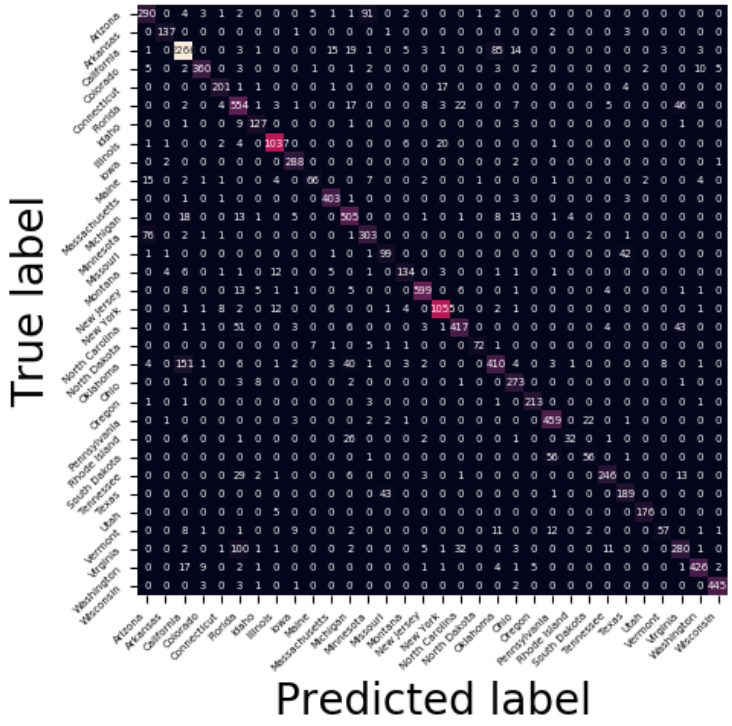
The decision to run the k-nearest neighbors algorithm (KNN) on this data set was made with the intent to classify the locational value of a data point based on its numerical features. In other words, when provided values for enrollment, MMR, overall, xmed, xper, and xrel we wanted to classify where in the US this datapoint is located. This method of classification helps us understand where vaccination rates in the US tend to be the lowest and where issues would arise when trying to convince the population to embrace a vaccine.

Description of Individual Work in Detail

I was responsible for conducting pre-processing, visualization, running k-nearest neighbors algorithm, and cleaning up and modifying the final GUI file for our project. The extent of my pre-processing work involved the imputation of missing values for the numerical features. The project scope for my work applying KNN using the state feature as the target was on a state-scale, so I dropped the year, county and district features. I imputed missing values by computing the mean value of each numerical feature by state and replacing state-specific NaN values with the computed mean. After nearly 20 minutes of runtime, there were still a large number of missing values. This is because certain states in the data set contain no data for certain numerical features. For example, Arizona contains no data on overall vaccination rate; there isn't a single existing value for the overall feature for Arizona in this data set. Thus, I decided to impute missing values of this accord by computing the mean of *every value* of the numerical features. Although this can be somewhat inaccurate, I thought it was the best course of action to reduce bias. By dropping every row with a missing value after the first imputation, the only state left was Colorado; *every state other than Colorado contained completely missing values for one or more numerical features*. Because the results of the visualization were somewhat inconclusive, imputing using the overall mean of each feature seemed appropriate. For visualization, I used GeoPandas and a provided framework to construct a map of the US and fill using the mean of the different numerical features. I read in the cleaned data set, read in the mapping framework as a GeoPandas data frame, added a column for state abbreviations to my data set, and merged the two data sets together on the state abbreviation column. This resulted in a data frame containing the state, numerical features, geometry and state abbreviation data. I plotted this data and changed what was used as the fill to visualize the distribution of the different numerical features across the US. For KNN, I read in the data and transformed the target variable (state) using label encoder and fit_transform. After splitting the training and testing data in a 70:30 split, a confusion matrix was computed and displayed. I also computed the accuracy score and ran cross validation and GridSearch to discover the best value for k. For our GUI file, I was responsible for cleaning the code and deploying a section that checks whether every applicable package is installed in order to run the program. I cleaned up variables, added in others' code and formatted everything in such a way that the average coder could pick up this file and understand what's going on.

Results

After imputing values for the numerical features and running KNN using state as the target, the accuracy score returned was roughly 87.44%. The corresponding confusion matrix proved that with a k-value of 3, the model could accurately predict, when given a value for MMR, where in the United States that value is located. The choice of k for this model was the result of cross-validation scores, GridSearchCV and analysis in R. Upon finding cross-validation scores and using GridSearchCV to deduce the best value of k for this model, the value returned was 3. Upon importing our cleaned data set into R and graphing accuracy vs values of k (between 1 and 21), the trend revealed was a steady decline in accuracy as k values grow larger. With these results, we thought it best to use k = 3 for our model.



Confusion Matrix for KNN using State as the target variable

Data containing the results of the classification and accuracy score is on the next page.

Classification Report:

	precision	recall	f1-score	support
0	0.74	0.72	0.73	403
1	0.94	0.95	0.94	144
2	0.91	0.94	0.92	2420
3	0.94	0.91	0.93	396
4	0.91	0.89	0.90	225
5	0.69	0.82	0.75	673
6	0.85	0.89	0.87	142
7	0.96	0.97	0.97	1072
8	0.92	0.98	0.95	293
9	0.84	0.62	0.71	106
10	0.92	0.98	0.95	412
11	0.80	0.89	0.84	570
12	0.72	0.78	0.75	387
13	0.67	0.68	0.68	145
14	0.86	0.79	0.82	170
15	0.95	0.93	0.94	645
16	0.96	0.96	0.96	1094
17	0.87	0.79	0.83	530
18	0.97	0.82	0.89	88
19	0.78	0.64	0.70	641
20	0.83	0.94	0.88	289
21	0.97	0.97	0.97	220
22	0.85	0.93	0.89	491
23	0.86	0.46	0.60	69
24	0.68	0.49	0.57	114
25	0.91	0.82	0.86	301
26	0.77	0.81	0.79	233
27	0.98	0.97	0.98	181
28	0.84	0.54	0.66	105
29	0.73	0.64	0.68	440
30	0.95	0.91	0.93	470
31	0.98	0.98	0.98	455
accuracy			0.87	13924
macro avg	0.86	0.83	0.84	13924
weighted avg	0.88	0.87	0.87	13924

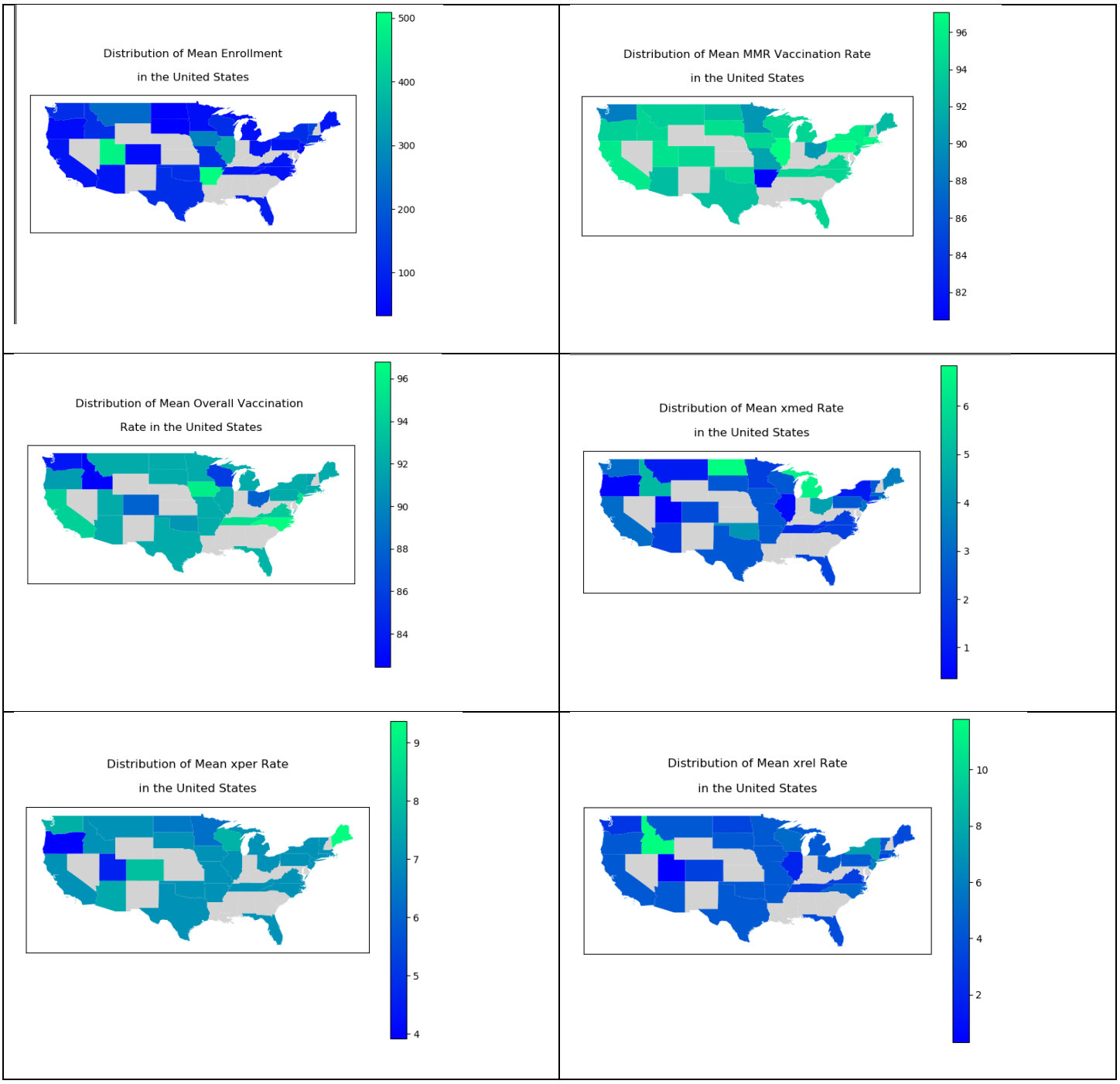
Classification report for KNN using State as the target variable; note the accuracy score of 0.87, or 87%

```
cv_scores = cross_val_score(model, X, Y, cv=5)
param_grid = {'n_neighbors': np.arange(1, 25)}
knn_gscv = GridSearchCV(model2, param_grid, cv=5)
knn_gscv.fit(X, Y)
knn_gscv.best_params_
```

Cross-validation and GridSearchCV code used to find the optimal k-value for KNN using State as the target variable

After viewing the maps constructed using GeoPandas, the results were very interesting. Of the 32 states provided in our data set, the distribution of mean MMR values tends to be higher in the northeast. Enrollment distribution tends to be even throughout the US, with some outliers in Arkansas and Utah. Overall vaccination rate tends to be lower in the northwest, with very low values in Washington and Idaho. The mean distribution of the percentage of children not vaccinated for medical reasons tends to be spread out with no noticeable pattern. The mean distribution of the percentage of children not vaccinated for personal reasons tends to be lower in the

northwest. The mean distribution of the percentage of children not vaccinated for religious reasons tends to be spread out, with some outliers in Idaho and Utah.



Visualization of Distribution of Mean Values for the Different Numerical Features

Summary and Conclusions

In the case of K-nearest neighbors using state as the target, I believe the choice for K is still debatable. The graph of accuracy vs k-values shows a steady decrease in accuracy as the k value increases. Although decreasing k could lead to overfitting and the induction of noise, it provided greater accuracy ratings than larger values of K. However, larger values of k may lead to overfitting and the ignorance of important patterns in our data. Specific to our data, the results show that we can correctly predict where in the US a given MMR value is located 87% of the time. This is important in developing strategies for convincing others to embrace a vaccine because it helps us understand where in the US we will likely have the most difficulty convincing others. An area with low vaccination rates may contain several anti-vaxxers. Throughout this project, I gained a better understanding of the application of classes and how they are called using different functions. I've also learned how to use GeoPandas to create visualizations and apply KNN in Python. I have also learned how to construct a GUI using PyQt5. However, perhaps the most important thing I've learned is that working with missing data can be messy. The reality of practical data sets is that they are very messy and contain incorrect or completely missing data. The data sets I have worked on in classes prior have been very clean and easy to understand. This data set was very hard to work with. An improvement I could make is to apply matplotlib and seasborn in PyQt to generate my visualizations directly instead of screenshotting from other code and displaying the screenshots. It was very confusing for me to try and incorporate plotting into the GUI.

Code Break-up

For my personal KNN file, 30 of the 148 lines of code were used from the Internet and 12 of these lines were modified; the credit for this code goes to Amir Jafari and Eijaz Allibhai of Towards Data Science (credit in References section). The percentage of code found/modified from the internet is as follows:

30 – 12 / 30 + 118 = 12.16%

For my visualization file, 27 of the 243 lines of code were used from the Internet and 20 of these lines were modified; the credit for this code goes to Stack Overflow, Erik G of Medium.com and GeoPandas (credit in References section). The percentage of code found/modified from the internet is as follows:

27 – 20 / 27 + 216 = 2.8%

For my imputation file, 3 of the 134 lines of code were used from the Internet and none of these lines were modified; the credit for this code goes to Stack Overflow (credit in References section). The percentage of code found/modified from the internet is as follows:

3 – 0 / 3 + 135 = 2.17%

References

France using MMR Vaccine for Coronavirus Vaccine Candidate: http://www.rfi.fr/en/science-and-technology/20200227-france-s-pasteur-institute-develop-coronavirus-vaccine-candidate

WallStreet Journal Data Set and Article: https://github.com/WSJ/measles-data, https://www.wsj.com/graphics/school-measles-rate-map/

Geoff Boeing, States Files for Visualization, GitHub.com: https://github.com/gboeing/beer-locations/tree/master/data-analysis/visualization/shapefiles/states_21basic

Erik G., Mapping using GeoPandas, Medium.com: https://medium.com/@erikgreenj/mapping-us-states-with-geopandas-made-simple-d7b6e66fa20d

GeoPandas: https://geopandas.org/mapping.html

StackOverflow: https://stackoverflow.com/