Final Project Report
DATS 6103-10, Group Four
Benjamin Lee, Kristin Levine, Russell Moncrief

## 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.[1]

## Description of the Data Set

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.[2] A summary of the features of this data set are provided below.

| 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 |

[1] http://www.rfi.fr/en/science-and-technology/20200227-france-s-pasteur-institute-develop-coronavirus-vaccine-candidate

[2] https://www.wsj.com/graphics/school-measles-rate-map/

| | | |
|---|---|---|
| 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 | |

## Description of Algorithms

### KNN with State as Target

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.

### Vaccination Rate as Target

We also decided to run the decision tree, random forest, and KNN algorithms on this data in order to classify whether a not a specific school had reached 95% MMR vaccination. In other words, when provided values for city, state/county, enrollment, type of school, and exemption records, we wanted to classify whether or not the school would have an MMR vaccination rate >= the recommended 95%. In this case, we were using the MMR rate as the target. We also ran the data using a >= 90% target rate.

## Experimental Setup

### KNN with State as Target

Preprocessing for implementation of the k-nearest neighbors algorithm (KNN) using the state feature as the target variable was performed independently of the preprocessing for decision tree, random forest and regression. Imputation of missing values was completed using the state-specific mean value of each numerical feature. For example, the missing MMR values for Arizona were imputed using the mean of the existing values (92.70). It is important to note that this procedure could not be completed for all missing data. For example, the values of overall, xrel, xmed and xper were completely missing for Arkansas; state-specific imputation of values for these features would make no sense because no values exist. Because KNN cannot work with missing data, we must impute all missing values. Therefore, a decision was made to impute values for this kind of situation using the mean of all values for a given feature. Although different states are bound to have different values for the numerical features, there is no way to accurately impute in a state-specific manner when the state has no values for a feature. I believe this was the best course of action to maintain low levels of bias that could be exaggerated by eliminating observations with missing data. It is also important to note that in the file containing data for all states, -1 was used as a replacement for some missing values. Upon further inspection, -1 was used as a placeholder for values that were missing in individual states' data files. For example, the missing MMR values for Arizona showed up blank in Arizona's data file, but in the data set containing all states the same data showed up with -1 for the missing values. Because certain features in the combined data set containing all states were excluded from individual state sets, there were a mix of -1's and missing values. For example, Arizona had missing MMR values but no 'overall' feature in its individual file, so in the combined states file the missing MMR values were -1's and all values for 'overall' were blank. For the sake of ease, every instance of -1 was replaced with NaN. After handling NaN values in this way, the replacement was conducted. The state-specific mean values for each numerical feature were computed, whereby every row was indexed and NaN values replaced for the respective state and feature.

### Vaccination Rate as Target

Preprocessing for implementation of decision tree, random forest, and KNN was also performed using MMR as the target variable. While WSJ was looking for a complete list of all schools, we wanted to look at schools for which we had data. The first course of action was to eliminate schools that possessed no enrollment data.

```
# Eliminate schools with no enrollment data
```

```
m = m.loc[(m[‘enroll’ > 0)]
```

Because states report vaccination data in different ways, there were two columns in the original file: MMR vaccination rate and overall vaccination rate (includes MMR and other diseases). Sometimes both values were provided and sometimes only one was given. Because we were focused specifically on MMR rates, we chose to take MMR when available and overall vaccination rate otherwise.

```
# Combine MMR and overall vaccination rates; use MMR if given and overall otherwise

m[‘vac_rate’] = m[‘mmr’]

m[‘vac_rate’] = m[‘mmr’].where(m[‘mmr’] > 0, m[‘overall’])
```

The WSJ appeared to use -1 as a replacement for missing values in certain places. We eliminated all schools with a negative vaccination rate; it is impossible to have a negative vaccination rate (see next page for code).

```
# Eliminate schools with a negative vaccination rate

m = m.loc[(m[‘vac_rate’] >= 0)]
```

Many sources set the necessary vaccination rate to prevent disease at 95%, while others set the value to 90%. For our target variable, we created new columns for both cases and set the variable type to Boolean.

```
# Is the vac_rate >= 95%?

m[‘at_least_95’] = (m[‘vac_rate’] >= 95)

# Is the vac_rate >= 90%?

m[‘at_least_90’] = (m[‘vac_rate’] >= 90)
```

The state column was grouped to compute the mean vaccination rate for each state.

```
# Find the mean vac_rate per state
state_mean = m[['state', 'vac_rate']]
state_mean = m.groupby('state').agg({'vac_rate': 'mean'}).reset_index()
state_mean = state_mean.rename(columns= {'vac_rate': 'state_mean'})
state_mean = state_mean.round(decimals=1)
```

Then, we wanted to substitute the mean values computed above back into the original frame to take the place of the state names. This was performed by converting state_mean to a dictionary and using replace().

```
# Add state_mean to original df
# First need to convert state_mean df to a dictionary, then use replace()
sm_dict = dict(zip(state_mean['state'], state_mean['state_mean']))
m['state_mean'] = m['state']
m = m.replace({'state_mean': sm_dict})
```

The same action was performed for the city and county features; if missing values were given, they were replaced with zeros (see next page for code).

```
# Fill county NaN values with zero
m['county_mean'] = m['county_mean'].fillna(0)
```

Next, we looked at the type feature. A dictionary was created for values in the type feature, using zero for missing values.

```
# Rename type column
```

```
m = m.rename(columns={'type': 'type_of_school'})

# Enter school type variables as binary

ts_dict = {'Public': 1, 'Charter': 2, 'Private': 3, 'Kindergarten': 4}

m = m.replace({'type_of_school': ts_dict})

# Fill type_of_school NaN values with zero

m['type_of_school'] = m['type_of_school'].fillna(0)
```

In the file containing data for all states, the exemption data is separated into religious, medical and personal exemptions (xrel, xmed and xper, respectively). For the purpose of simplification, missing values were replaced with zeros and the exemption features were combined to measure total exemptions.

```
# Fill exemption values with zero

m['xrel'] = m['xrel'].fillna(0)

m['xmed'] = m['xmed'].fillna(0)

m['xper'] = m['xper'].fillna(0)

# Add all different types of exemptions together

m['xtotal'] = m['xrel'] + m['xmed'] + m['xper']
```

The state mean, city mean, county mean, school type, enrollment, xtotal, at least 95 and at least 90 features were selected and converted to binary for use in decision tree and random forest analysis.

```
# Select columns to use for DT

m_tree = m[['state_mean', 'city_mean', 'county_mean', 'type_of_school', 'enroll',
'xtotal', 'at_least_95', 'at_least_90']]

# Check to see if df has any NaN values
```

```
print(m_tree.isnull().sum())
print(m_tree.dtypes)
# Converting variables to binary for use in analysis
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
m_tree['state_mean'] = le.fit_transform(m_tree['state_mean'])
m_tree['city_mean'] = le.fit_transform(m_tree['city_mean'])
m_tree['county_mean'] = le.fit_transform(m_tree['county_mean'])
m_tree['type_of_school'] = le.fit_transform(m_tree['type_of_school'])
m_tree['enroll'] = le.fit_transform(m_tree['enroll'])
m_tree['xtotal'] = le.fit_transform(m_tree['xtotal'])
m_tree['at_least_95'] = le.fit_transform(m_tree['at_least_95'])
m_tree['at_least_90'] = le.fit_transform(m_tree['at_least_90'])
```

## Results

### KNN State as Target 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.

### Visualization: Mapping Using GeoPandas

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.

### Vaccination Rate as Target Results

### Decision Tree Results 95% target rate

In our test dataset of size 0.3, there were actually 6422 schools with an MMR rate >= 95 and 3166 schools with an MMR rate < 95.

This model found 6339 of the schools with 95% MMR rates and 1116 of the schools with lower vaccination rates. It mislabeled 83 schools with rates >= 95% as not meeting the 95% threshold; it also mislabeled 2050 schools as meeting the 95% threshold when they did not.

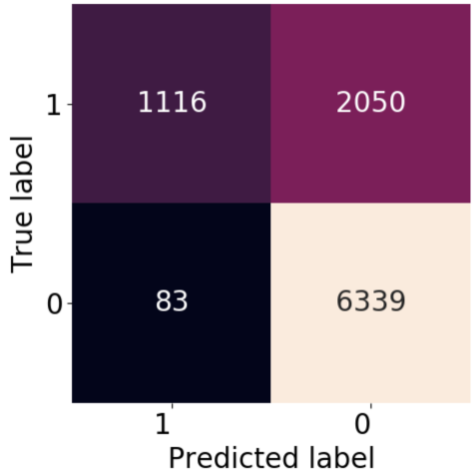
This model had more False Positive (Type I errors) than False Negative (Type II errors.)

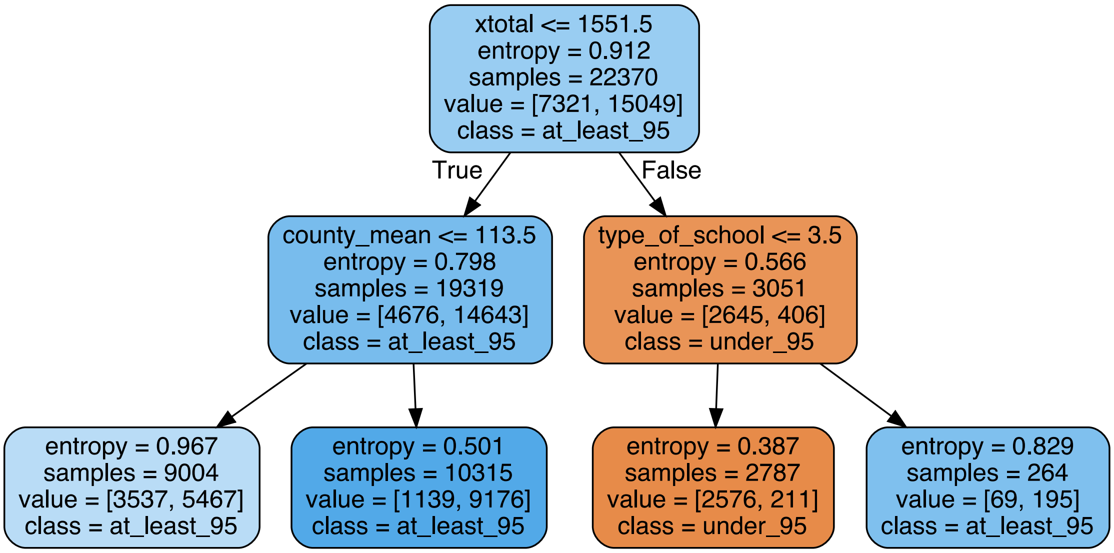
Results Using Entropy:

Classification Report:

```
          precision    recall  f1-score   support

       0       0.93      0.35      0.51      3166
```

```
           1       0.76      0.99      0.86      6422

    accuracy                           0.78      9588
   macro avg       0.84      0.67      0.68      9588
weighted avg       0.81      0.78      0.74      9588

Accuracy :  77.75344180225282
```





## Decision Tree Results 90% target rate

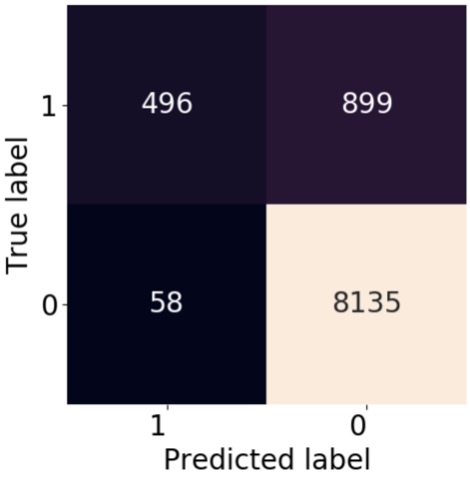
If we lowered the target rate to >= 90, there were actually 8193 schools that met this target, and 1395 that did not.

In this case, the model found 8135 of the schools that met the target rate and correctly labeled 496 schools that did not.

It mislabeled 58 schools with rates over 90% as not meeting the threshold; it also mislabeled 899 schools as meeting the 90% threshold when they did not.

This model had more False Positive (Type I errors) than False Negative (Type II errors.)

```
--------------------------------------------------------

Results Using Entropy:

Classification Report:

              precision    recall  f1-score   support

           0       0.90      0.36      0.51      1395

           1       0.90      0.99      0.94      8193

    accuracy                           0.90      9588

   macro avg       0.90      0.67      0.73      9588

weighted avg       0.90      0.90      0.88      9588

Accuracy :  90.01877346683355
```



**Random Forest Results 95%**

For our Random Forest Algorithm, the results were similar: this model found 5655 of the schools with 95% MMR rates and 2072 of the schools with lower vaccination rates. It mislabeled 767 schools with rates >= 95% as not meeting the 95% threshold; it also mislabeled 1094 schools as meeting the 95% threshold when they did not.

Compared to the Decision Tree, the Random Forest reduced the number of Type 1 errors, however, it increased the number of Type II errors. However, overall, this model still had more False Positive (Type I errors) than False Negative (Type II errors.)

Results Using All Features:

Classification Report:

```
              precision    recall  f1-score   support

           0       0.73      0.65      0.69      3166

           1       0.84      0.88      0.86      6422

    accuracy                           0.81      9588

   macro avg       0.79      0.77      0.78
9588

weighted avg       0.80      0.81      0.80
9588
```

Accuracy :  80.70504797663747

ROC_AUC :  86.5299946114637

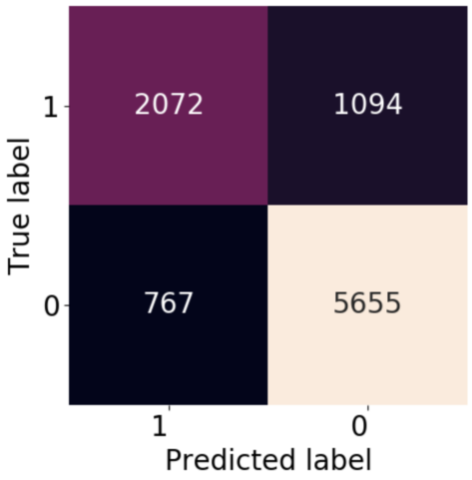
Results Using K features:

Classification Report:

```
              precision    recall  f1-score   support

           0       0.73      0.65      0.69      3166

           1       0.84      0.88      0.86      6422

    accuracy                           0.81      9588

   macro avg       0.78      0.77      0.77
9588

weighted avg       0.80      0.81      0.80
9588
```

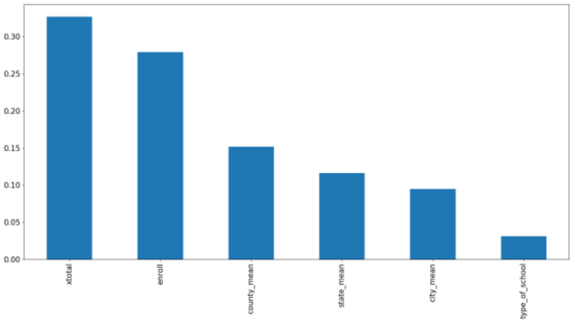
Accuracy :  80.59032123487692

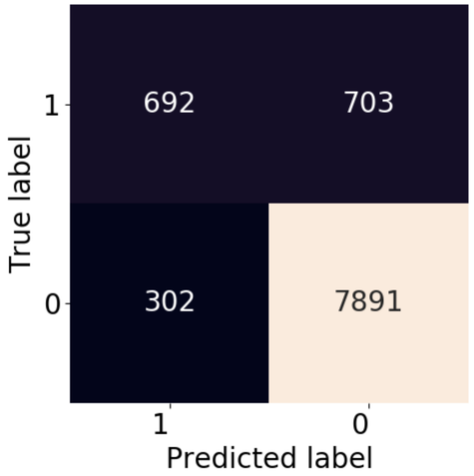
ROC_AUC :  86.63533813507853



**Random Forest 90%**

With a 90% threshold, the results were very similar. It’s interesting to note, that the most important feature was the total exemption rate; enrollment was second.  The type of school was the least important.





Results Using All Features:

Classification Report:

```
              precision    recall  f1-score   support

           0       0.69      0.50      0.58      1395

           1       0.92      0.96      0.94      8193

    accuracy                           0.89      9588

   macro avg       0.80      0.73      0.76
9588

weighted avg       0.88      0.89      0.89
9588
```

Accuracy :  89.3825615352524

ROC_AUC :  87.01679508733525

Results Using K features:

Classification Report:

```
              precision    recall  f1-score   support

           0       0.70      0.50      0.58      1395

           1       0.92      0.96      0.94      8193

    accuracy                           0.90      9588

   macro avg       0.81      0.73      0.76
9588

weighted avg       0.89      0.90      0.89
9588
```

Accuracy :  89.51814768460575

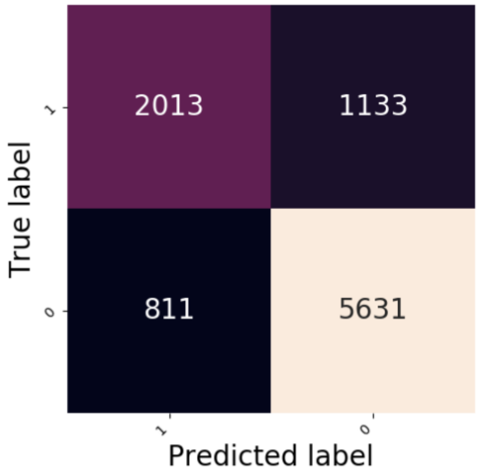
ROC_AUC :  86.84485882038474

The KNN results were very similar to the Random Forest

**KNN 95%**

Classification Report:

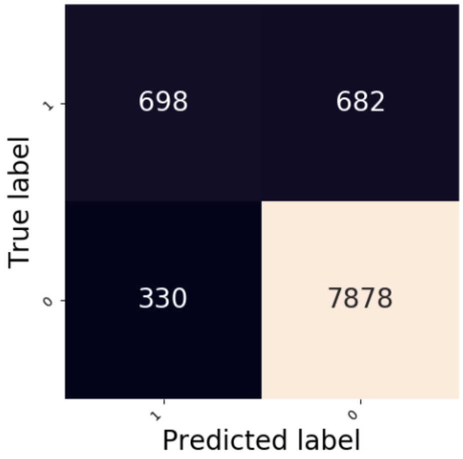
```
              precision    recall  f1-score   support

           0       0.71      0.64      0.67      3146

           1       0.83      0.87      0.85      6442

    accuracy                           0.80      9588

   macro avg       0.77      0.76      0.76
9588

weighted avg       0.79      0.80      0.79
9588

Accuracy :  79.72465581977471
```



**KNN 90%**

Classification Report:

```
              precision    recall  f1-score   support

           0       0.68      0.51      0.58      1380

           1       0.92      0.96      0.94      8208

    accuracy                           0.89      9588

   macro avg       0.80      0.73      0.76
9588

weighted avg       0.89      0.89      0.89
9588

Accuracy :  89.44513975803086
```



In fact, all three models (Decision Tree, Random Forest, and KNN) were quite similar. With the 95% threshold, our accuracy ranged from 77.8% to 80.7% across the three models. With the 90% threshold, our accuracy ranged from 89.4% to 90.0%

## Comparing Accuracy of All Three Models

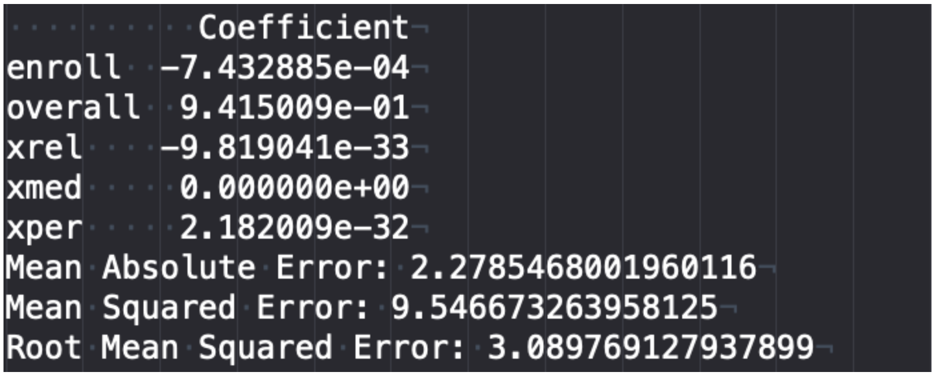
| | **95% Threshold** | **90% Threshold** |
|---|---|---|
| **Decision Tree** | 77.8 | 90.0 |
| **Random Forest** | 80.7 | 89.4 |
| **KNN** | 79.7 | 89.4 |

## Linear Regression

In the case of regression, we performed linear regression to investigate which variables influence the MMR vaccination rate among different states and whether a linear model can predict the MMR rate among different states. We subset the data based on state name from user input and regressed mmr on enroll, overall, xrel, xmed and xper. The linear model performed well with some states and performed poorly with others. We suspect that highly imputed data in several states caused the OLS estimator to fail because of low variability among the data. Several states where OLS did perform well are listed on the next page.

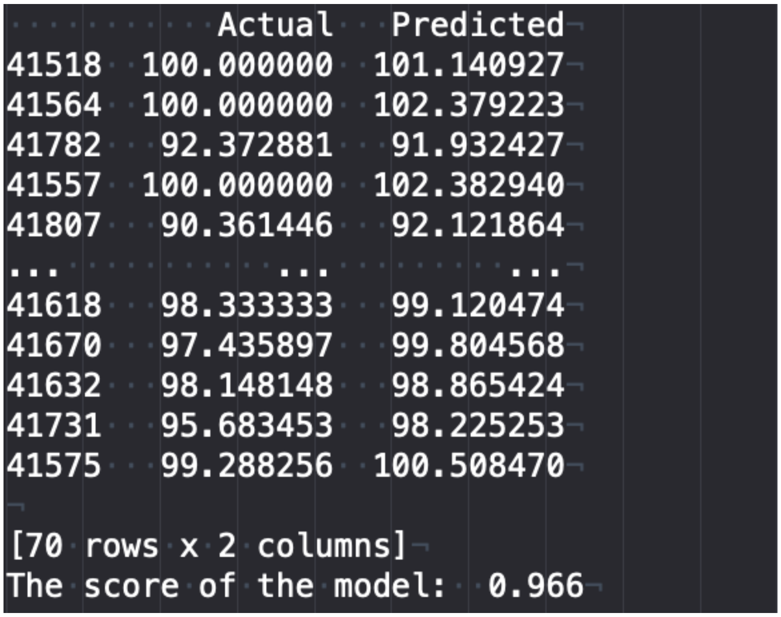
### *Vermont*

The coefficient on enroll suggests a one student increase in total enrollment decreases the MMR rate by -0.000743 percent. The coefficient on the overall MMR rate increases the MMR rate by 0.0941 percent. The coefficients on religious and personal exemptions greatly decrease the MMR rate. Exemptions due to medical reasons may lack variation because of imputation; its coefficient is zero.



```
           Coefficient
enroll  -7.432885e-04
overall  9.415009e-01
xrel    -9.819041e-33
xmed     0.000000e+00
xper     2.182009e-32
Mean Absolute Error: 2.2785468001960116
Mean Squared Error: 9.546673263958125
Root Mean Squared Error: 3.089769127937899
```

*Coefficient results of linear regression on Vermont*



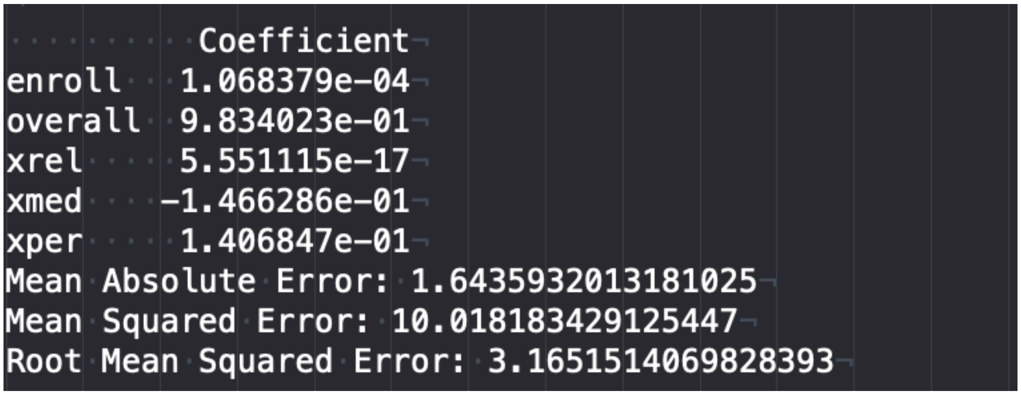
```
           Actual   Predicted
41518  100.000000  101.140927
41564  100.000000  102.379223
41782   92.372881   91.932427
41557  100.000000  102.382940
41807   90.361446   92.121864
...           ...         ...
41618   98.333333   99.120474
41670   97.435897   99.804568
41632   98.148148   98.865424
41731   95.683453   98.225253
41575   99.288256  100.508470

[70 rows x 2 columns]
The score of the model:  0.966
```

*Linear model results of linear regression on Vermont*

The linear model performed well. The mean error statistics are relatively low, and the model score is 96.6 percent.

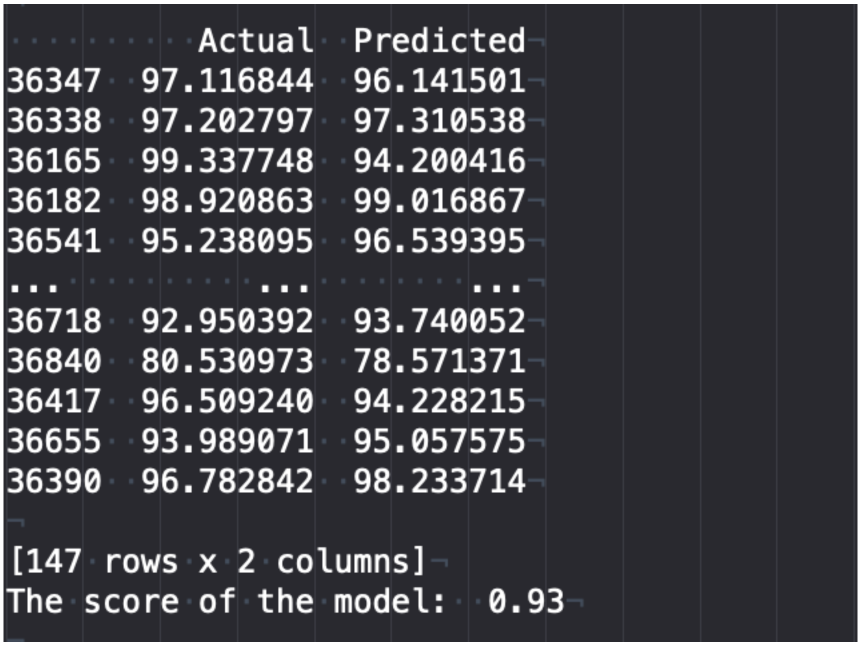
*Oregon*

The coefficient on enroll suggests a one student increase in total enrollment increases the MMR rate by 0.000106percent. The coefficient on the overall MMR rate increases the MMR rate by 0.0983 percent. The coefficient on xmed suggests a one student increase in medical exemption decreases the MMR rate by -0.146 percent. The signs of the coefficients on xrel and xper do not makes sense and suggest poor data and heavy bias for those variables. Despite probable bias in xrel and xper, the model performs well and displays a score of 93 percent.



```
          Coefficient
enroll   1.068379e-04
overall  9.834023e-01
xrel     5.551115e-17
xmed    -1.466286e-01
xper     1.406847e-01
Mean Absolute Error: 1.6435932013181025
Mean Squared Error: 10.018183429125447
Root Mean Squared Error: 3.1651514069828393
```

*Coefficient results of linear regression on Oregon*



```
          Actual  Predicted
36347  97.116844  96.141501
36338  97.202797  97.310538
36165  99.337748  94.200416
36182  98.920863  99.016867
36541  95.238095  96.539395
...          ...        ...
36718  92.950392  93.740052
36840  80.530973  78.571371
36417  96.509240  94.228215
36655  93.989071  95.057575
36390  96.782842  98.233714

[147 rows x 2 columns]
The score of the model:  0.93
```

*Linear model results of linear regression on Vermont*

Overall, the linear model is able to predict the MMR rate given a dataset that is not heavily imputed.

## Summary and Conclusions

In the case of decision tree, random forest and KNN using vaccination rates as the target, one interesting conclusion was that the total number of exemptions was the most important feature when looking at whether or not a vaccination rate would be over a certain threshold. While this isn’t exactly surprising, it does provide evidence that state laws probably DO influence overall vaccination rates.  If a state only allowed medical exemptions, they would probably have a higher vaccination rate than a state that made it easy to request a religious or personal exemption. For future research, we are interested in tracking down state laws and see if this is indeed the case. Additionally, it would be interesting to track down vaccination data from states not listed in this sample and see how that data works with our model. Also, while the data looking at vaccination rates over 95% was fairly balanced, when we changed the cutoff point to over 90% it became much less balanced.  Finally it would be interesting to look at school with a lower vaccination rate (perhaps under 80% ) to see what they have in common.

The linear model is able to predict MMR vaccination rates by state given a dataset that is not heavily imputed.  In states where the model performed well, the model achieved an accuracy of 90 percent or greater.  The linear model identified which variables most affect MMR vaccination rates.  Exemptions for religious, personal and medical reasons most affect MMR vaccination rates.  From a policy standpoint, to increase participation in MMR vaccinations efforts should be focused on state policy exemptions.

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.

## References

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

Info about French Vaccine

http://www.rfi.fr/en/science-and-technology/20200227-france-s-pasteur-institute-develop-coronavirus-vaccine-candidate

Data Mining Class GitHub

https://github.com/amir-jafari/Data-Mining

Wall Street Journal Article

https://www.wsj.com/graphics/school-measles-rate-map/

Wall Street Journal Dataset

https://github.com/WSJ/measles-data