Data Analytics (Fall'23) Assignment - 7

WeiJun Li
liw18@rpi.edu, RIN : 662006326, Level : 4000

December 3th, 2023

Exploring the Socioeconomic Drivers of COVID-19 Vaccine Acceptance

1 Abstract and Introduction:

Covid-19 is serious pandemic around the world and the vaccine is the best way people can get the protect. However, the rumor and physical circumstance affect the people to get the vaccine and the booster shoot to protect himself. So, I want to know what the factor is affect people get the vaccine or booster shoot. Then I go to the HealthData.gov to find the relate information. I found out a survey name "HSS COVID-19 Monthly Outcome Survey" which focus on the assess COVID-19 vaccine uptake as well as beliefs, intentions, and behaviors relevant to COVID-19 vaccination at a point in time. This survey is conduct with cross-section sample of approximately 5000 U.S. adults in each month [1]. This is important dataset which can help us to get more people affect people get the vaccine avoid the mass people die for the global pandemic. So, I choose the dataset in the July 31, 2023 which focus on the people get the booster shoot which it is more important people usually lost their intention about the pandemic after get the vaccine and they will not expect the Covid-19 happen to them.

This dataset have many feature which have dimension 5018 rows and 96 column which kind of complex dataset and all the feature is categorization which good for us to use the classification model like Random forest, Decision tree and so on. The data type of this dataset which have 2 type text and number. This is important which we should convert all the data characteristics data into the number or remove it from the dataset. My target value for this dataset is CAM5_VaccUptake which indicate people take the booster shoot or how many times they take the vaccine. To find out the most importance feature, I will do the PCA first to analysis how the Cumulative Variance distribution and I pick some important feature will strongly affect the people get the vaccine and booster shoot, such income, education, insurance, employment, and good behavior. Also, I will run the Random Forest and Decision tree, SVM and Kmean to find we can group the data cluster correctly. My hypothesis for this dataset is income, education, insurance and employment, and good behavior have strong impact for people willing to take the vaccine and booster shoot. In order to achieve this I will start the EDA and see how the distribution of data.

2. Data Description and Exploratory Data Analytics:

As we do the EDA, we can find out many different features are included in my dataset, they can split out with 2 group one is physical condition like income, education, employment and mental condition such as attitude about the Covid-19 and vaccine. Also, all the data their categorical data which it represents by different group of representation. It is good which we can apply the classification model to analysis what cause people to uptake the vaccine. So, I think the model such the Decision tree, Random Forest, Support Vector machine which supervised method and use the K-mean to find how many cluster should I take which it is a unsupervised method and not need to clean the data and see how many cluster is best fit for my dataset. I believe it is 3 cluster which my target value has 3 different class.

This is important because to have the dataset like it which can know how people get affect by the Covid-19 which we will have the better prepare for the next upcoming global pandemic. As the result of the Covid-19, it cause 1,146,793 people die for Covid-19 which bigger number than die for the WW2. Also, According to CDC(figure 1), as the graph below showing, the people die by sort by group, we can clear to see that older people between 55-74 group of people have bigger affect by it and this is why people should get the vaccine or the booster shoot. The vaccine and the booster shoot can significantly improve people die for the symptom for the Covid-19.

As the Case study base on the survey, we can analysis what physical condition affect people get the vaccine or the booster shoot and next time we have the better policy to help people who need most and more people can get the vaccine which only 80% of people get the vaccine according to the survey I choose. Also, higher vaccine rate can protect who is more vulnerable which call the herd immunity. In order to achieve this goal. The government should make some policy about it, such require people who work for the federal or medical industry should enforce them to take the vaccine or booster shoot which they are most easy to get infection.

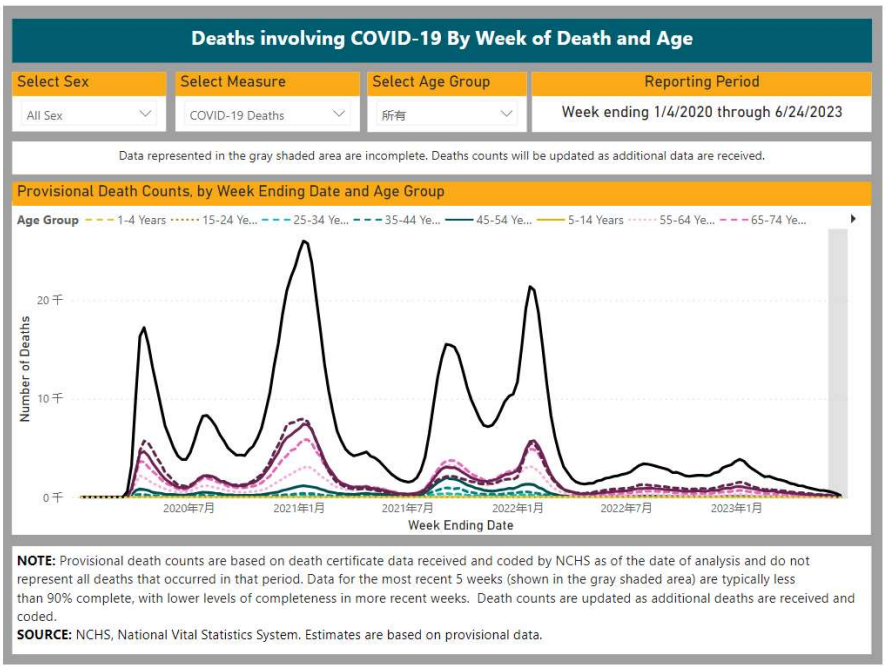


Figure 1

3. Analysis (5%)

For the analysis my data, we need to see the distribution of the feature we choice, such the income (Figure 2), education (Figure 3), insurance (Figure 4), employment (Figure 5). As we can see below, each variable have different distribution. As all the data is categorized which I cannot apply the Multivariate Regression to it which it is not the continuous variable. As we can see histogram about the income, each group have similar sample set which it is important to let each group have enough sample to do the calssification then have a good representation. The Employment information graph show the number 3 have a large number there which it represent most the people is employment and another big group is the people get retired and other status which it show that the complex of the world. Also, it shows a group of my data is represent all different age group and well represent how the different group of the people reacte which the Vaccine and booster shoot.

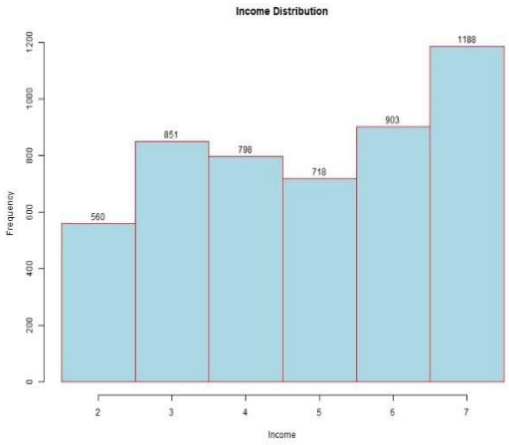


Figure 2

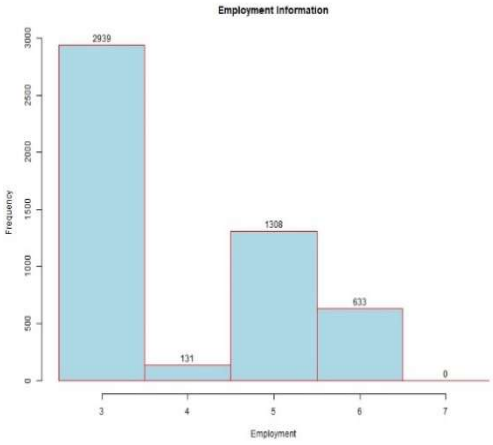


Figure 3

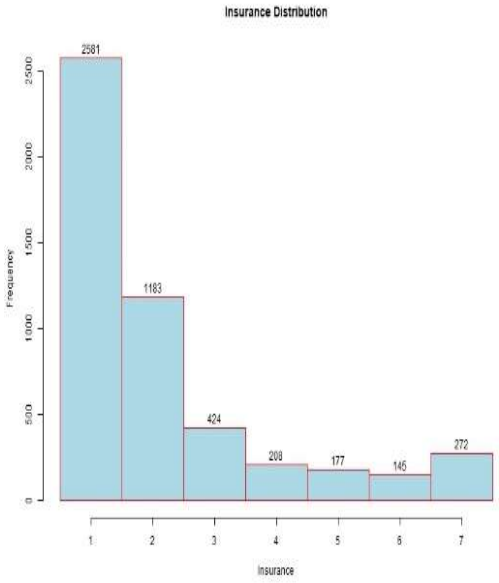


Figure 4

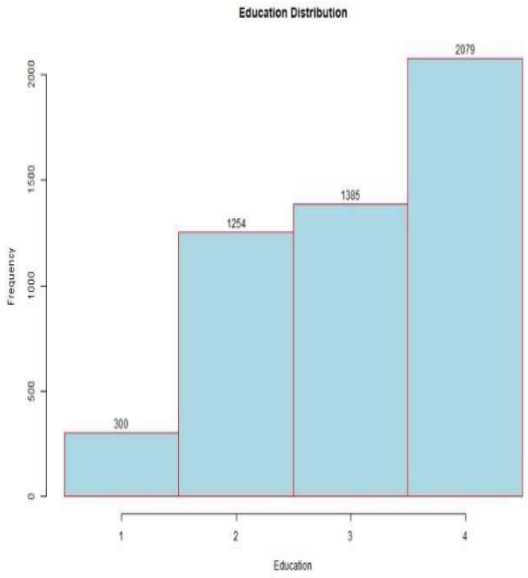


figure 5

As the education histogram show that most people take the survey is get a good education which number 1 represent less the high school and it only take 6% of my data. This is important which show my data may not represent well. According to the U.S. census there around 8.9% of people have not get the high school diploma which is this example may not well represent the people not getting the high school diploma and this is the factor we should consider in our model. For the insurance feature which it is the most affect people to get the vaccine which they need to pay the full price for the vaccine. According to my histogram show that only 94.58% of people have the different type of insurance (number 1-6) and only 5.32% of people(number 7) don't have the insurance. According to pgpf.org show that there have 7.9 % of people don't have the insurance and in our example is lower than their expect. So, it is where we should pay attention on like people don't have the high school diploma which less represent group of people in the U.S. Maybe we need to consider add different weight

of group of people to have more representation data. 、

Then I do the principal component analysis (PCA) to see how the feature distribution in dataset. The first thing I do remove the people refuse to answer how many vaccines they take which it is noise for our model and improve my model performance. Then next thing is removing the data type is text type, which is ID, Uptake_Date_1, Uptake_Date_2, and Uptake_Date_3 which it is not relative will my target value and if not remove cause the bigger noise for my model we create. Next step, I do the cumulative variance to analysis (Figure 6) to see each factor contributing to target value, as we can see that only 24 feature contributions over 80% of variance. This is indicated that many feature have not effect on the target value and it also strong to prove my hypothesis which only few features have important effect on the people get the vaccine or the booster shoot. Also, analysis the distribution of how many people take the vaccine (Figure 7), we can see the in the sample, there 3749 people take the booster shoot, 147 take at least one shoot and 893 people have no get any vaccine. Consider time which is July 31,2023. This is indicate that most of the people take at least one shoot vaccine. The is indicate most of the people have positive attitude for the exam and this is a god example for us to analysis which it is better to show why the people not take the vaccine and transfer our goal to make more people take the vaccine to protect vulnerable group of people. So let see how my hypothesis work for my dataset after I build the model to analysis.

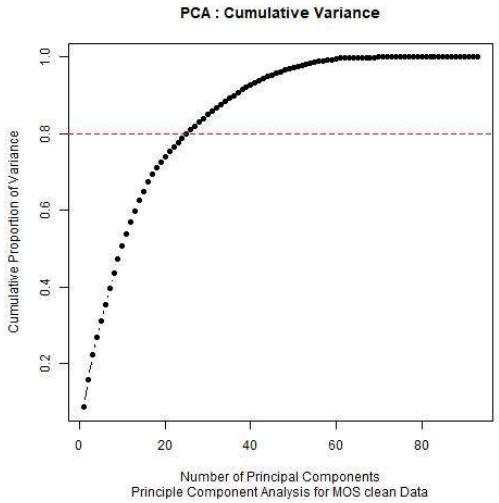


Figure6

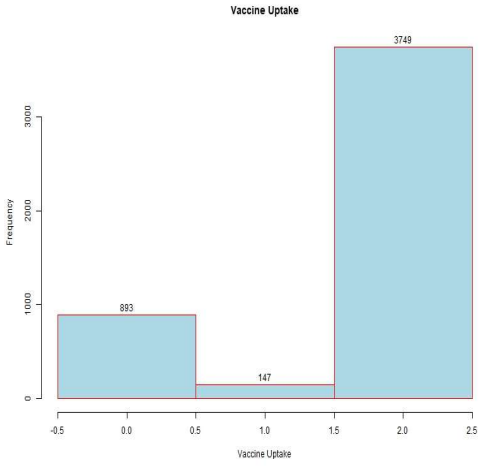


Figure 7

4. Model Development and Application of model(s):

The first model I use is K-mean which it is an unsupervised method does not need to modify the data we use, So, we can see how many cluster should this dataset have and I use the fviz_nbclust method (Figure 8) and clusGap method (Figure 9) to determine what is the best k we should contain the best result. Fviz_nbclust method is using the within-cluster sum of squares criterion as the condition also, known as elbow method. As I draw the line in the number of 3 is the point the total within sum of square decrease slowly which it is prove my point this dataset should set to 3 cluster which the derivative of the k-mean starting flatter with it is indicate it is the good choice to use 3 to in the k-mean model. Also, I use the clusGap method which computes the gap statistic and use the bootstrapping technique to compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data. As we can see it through the graph, I test the k value from 1 to 10 and the result indicates that although there is no obvious peak but we can see that when k = 3 there have a local peak point which it show we can choice to k = 3 as the possible value because since the Gap statistics of other k values are also relatively high and the error is large,

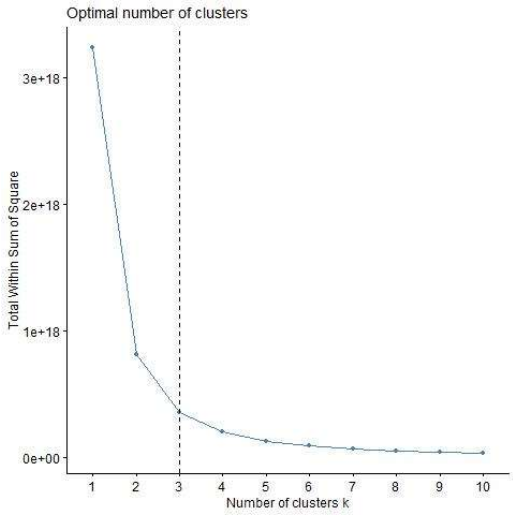


Figure 8

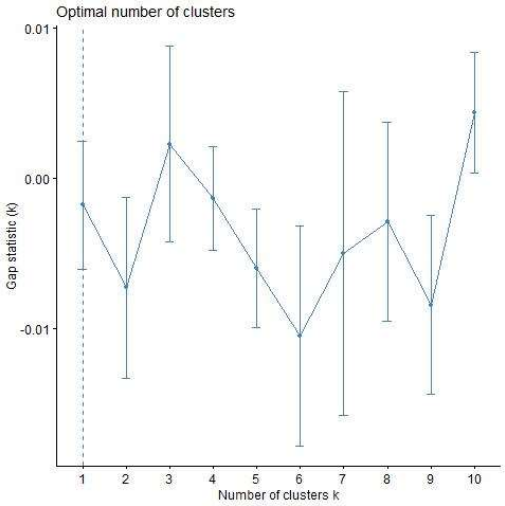


Figure 9

Because of all the evidence above I choose 3 as the k value as my K mean model, First, I split the 80% of data for train and 20% for test my k-mean model. As the model show below (Figure 10), we can clearly to see that there have many part is overlaps which it is indicate that it is not good performance of my model, As a good model, we need to see each cluster have a the differences within clusters are small and the difference between clusters is large. As I call the summary function of my K-mean model which we can see that (betweens_SS/ total__SS) = 13%. This is mean that 13%

of the total sum of squares is due to different between cluster and the similarity within cluster accounts for 87% of the total difference. This show that it is not good model which each point in the cluster is not close the center and not far away with other clusters. This is also meet our prediction, which the Gap-statistic indicate there have many noises we need to handle. As the result, we cannot use the K-mean model to predict the cluster because the model cannot find the correct number of the cluster and identify the correct group of people.

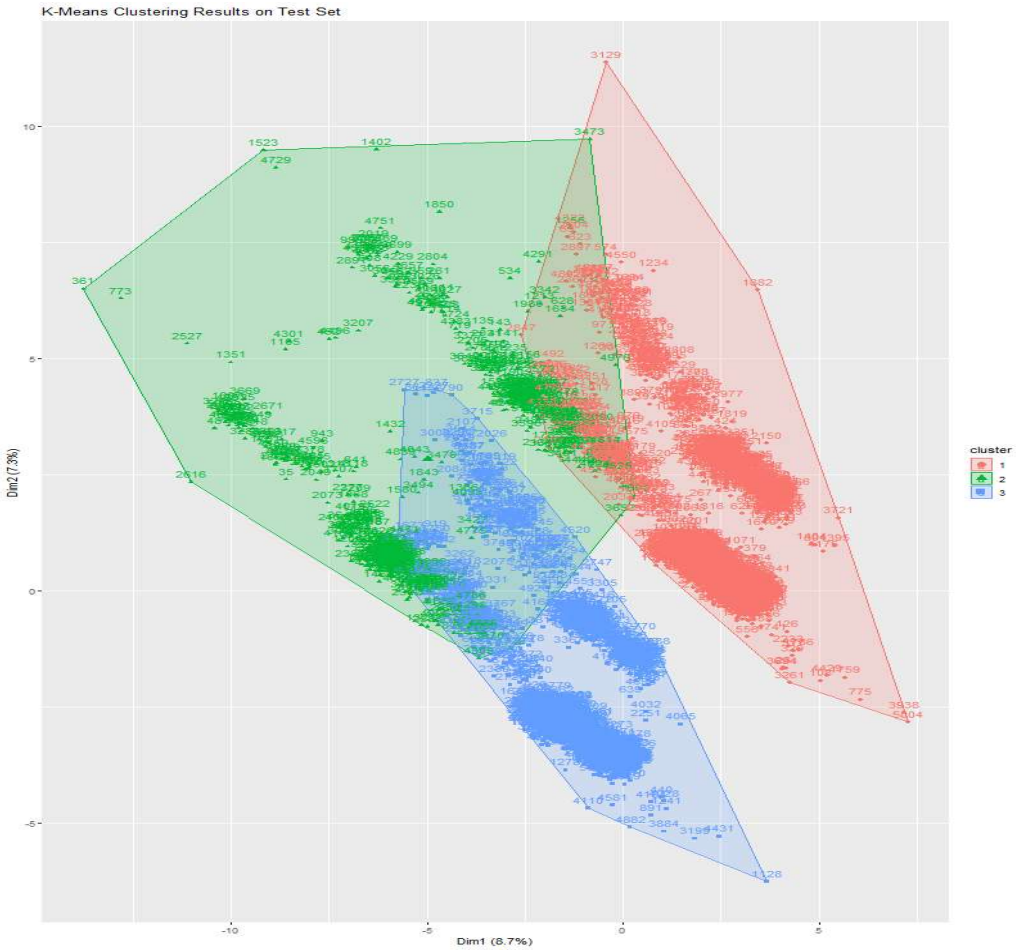


Figure 10

The next model I use is the Random Forest which it also not need to clean the data and have a good ability to handle the complex relationships. However, As the random forest design it is require taking all the features to analysis and have the better representation which it is not let me to pick only few features to analysis which it is make up with many different decision tree and each decision tree is not be connected. As the result, the random forest will choice from most appear result in all the decision tree as the result.

The first step I do is split the dataset in the 80% for train and 20% for test and use the k-cross validation (Figure 13) to find the best mtry parameter which each node choice how many features in the dataset and control the random level in the random forest. As the result show when the mtry is 31 will be get the best result and it indicate that 31 features is enough to predict dataset. As the graph (Figure 12) show, I finished the train the model and apply the model to the test set and use the confusion model to verify my result which have very good result which have 99.9 % on the test set. As the table show that it is very sensitive for each group which close to 1. Moreover, check it with the mtry with accuracy (Figure 12), when mtry equal to 31 have the best result but we can see that around the 5-10 features can have the similar result as the 31 which close to 99%. This is also proving my point only few features have strong effect on the target value which it is prove my point to choose only few features can be do the predict the result in the acceptable ratio. This is important which it is indicate that we only need to control few variables to significant increase the ratio people to get the vaccine. So, let move forward to see how the Decision Tree can help us to determine the target value in the correct group.

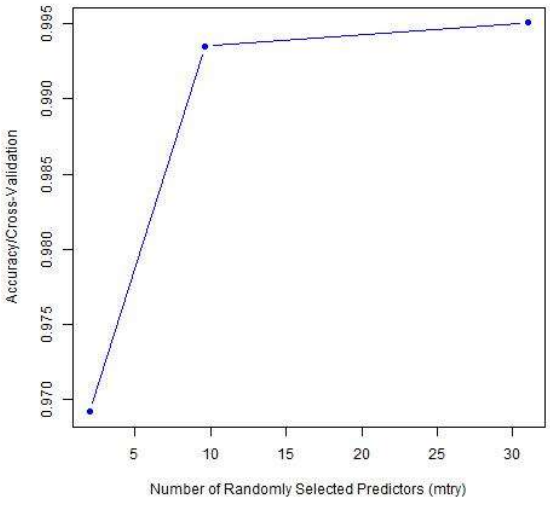


Figure 11

```
> print(confusion)
Confusion Matrix and Statistics

          Reference
Prediction  X0  X1  X2
        X0 178   1   0
        X1   0  28   0
        X2   0   0 749

Overall Statistics

               Accuracy : 0.999
                 95% CI : (0.9942, 1)
    No Information Rate : 0.7835
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.997

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: X0 Class: X1 Class: X2
Sensitivity             1.0000   0.96552    1.0000
Specificity             0.9987   1.00000    1.0000
Pos Pred Value          0.9944   1.00000    1.0000
Neg Pred Value          1.0000   0.99892    1.0000
Prevalence              0.1862   0.03033    0.7835
Detection Rate          0.1862   0.02929    0.7835
Detection Prevalence    0.1872   0.02929    0.7835
Balanced Accuracy       0.9994   0.98276    1.0000
```

Figure 12

```
> print(rf_model)
Random Forest

3833 samples
  92 predictor
   3 classes: 'X0', 'X1', 'X2'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3449, 3449, 3450, 3449, 3450, 3451, ...
Resampling results across tuning parameters:

  mtry       logLoss     AUC        prAUC      Accuracy   Kappa      Mean_F1
   2.000000  0.13434505  0.9963015  0.8904108  0.9692157  0.9111045        NaN
   9.591663  0.03336130  0.9992467  0.7068537  0.9934787  0.9814091  0.9544433
  31.000000  0.01654883  0.9995500  0.5406312  0.9950466  0.9859399  0.9685247
  Mean_Sensitivity  Mean_Specificity  Mean_Pos_Pred_Value  Mean_Neg_Pred_Value
  0.6666667         0.9862027               NaN            0.9897386
  0.9364834         0.9973791         0.9812452            0.9977199
  0.9629078         0.9980895         0.9761048            0.9981800
  Mean_Precision  Mean_Recall  Mean_Detection_Rate  Mean_Balanced_Accuracy
        NaN       0.6666667    0.3230719            0.8264347
  0.9812452       0.9364834    0.3311596            0.9669313
  0.9761048       0.9629078    0.3316822            0.9804987

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 31.
```

Figure 13

For the decision tree(Figure 16), this is a good model to do the classification which it matches up with our intuition. As before I split the 80% for train and 20% for test and do the K-cross validation to find the best model, I have which the parameter cp which prevent the model is too simple or overfit and control the decision tree grow. As the performance (Figure 14) indicators show that the model reaches approximately 78.65% to 78.93% accuracy, while the Kappa value fluctuates between 0.17 and 0.18. At the same time, the AUC value is between 0.66 and 0.67, which shows that the model's classification ability is in the acceptable ratio, but there is also room for improvement.

```
> print(dt_model)
CART

3833 samples
   8 predictor
   3 classes: 'X0', 'X1', 'X2'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3449, 3449, 3450, 3449, 3450, 3451, ...
Resampling results across tuning parameters:

  cp           logLoss    AUC       prAUC      Accuracy   Kappa      Mean_F1
  0.002400960  0.5953044  0.6663075  0.2981415  0.7865872  0.1709695  NaN
  0.002801120  0.5768089  0.6710998  0.2883962  0.7891962  0.1755122  NaN
  0.004201681  0.5889096  0.6333738  0.2211245  0.7837226  0.1187011  NaN
  Mean_Sensitivity  Mean_Specificity  Mean_Pos_Pred_Value  Mean_Neg_Pred_Value
  0.3801286         0.7111716         0.4280995            0.8005984
  0.3808748         0.7122070               NaN            0.8075010
  0.3650276         0.6975727               NaN            0.7859697
  Mean_Precision  Mean_Recall  Mean_Detection_Rate  Mean_Balanced_Accuracy
  0.4280995       0.3801286    0.2621957            0.5456501
        NaN       0.3808748    0.2630654            0.5465409
        NaN       0.3650276    0.2612409            0.5313002

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.00280112.
```

Figure 14

The next step is applied it on the test set which we get follow result, it have the similar result as the train set which 79.08% on accuracy. The p-value for McNemar's Test is less than 2e-16, which is very low and indicates that the difference in prediction accuracy between categories is significant. One thing I notice is this model is do very bad on the people only get one vaccine which the sensitivity of the X1 is 0 which this dataset has too less sample for the people get 1 shoot of vaccine. This is import that this the problem of my dataset which we only have 147 sample in the data set which this is one reason that our decision tree model performance not well which many different groups of the people are misclassification into the group 1 which only take one shoot of the vaccine. One reason can explain why the group of people take only one shoot of vaccine is the data we choice is close to end of the pandemic which it is in the 2023. This indicate that another factor we should consider about the time change more and more people will more will to take the booster shoot and the policy is change, such the people require to have the vaccine card to show they take at least 2 shoot of the vaccine to go work.

```
> print(confusion)
Confusion Matrix and Statistics

          Reference
Prediction  X0  X1  X2
       X0  31   2  24
       X1   0   0   0
       X2 147  27 725

Overall Statistics

               Accuracy : 0.7908
                 95% CI : (0.7636, 0.8162)
    No Information Rate : 0.7835
    P-Value [Acc > NIR] : 0.3068

                  Kappa : 0.1703

 Mcnemar's Test P-Value : <2e-16

Statistics by Class:

                     Class: X0 Class: X1 Class: X2
Sensitivity            0.17416   0.00000    0.9680
Specificity            0.96658   1.00000    0.1594
Pos Pred Value         0.54386       NaN    0.8065
Neg Pred Value         0.83648   0.96967    0.5789
Prevalence             0.18619   0.03033    0.7835
Detection Rate         0.03243   0.00000    0.7584
Detection Prevalence   0.05962   0.00000    0.9404
Balanced Accuracy      0.57037   0.50000    0.5637
```

Figure 15

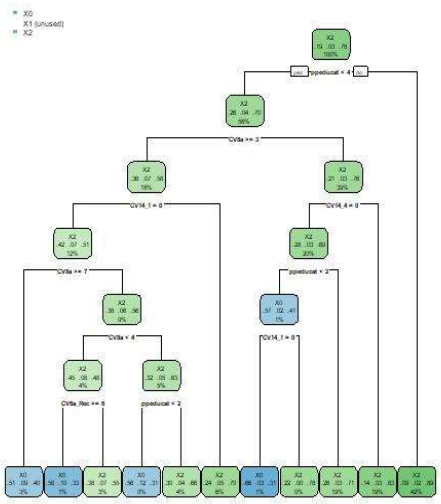


Figure 16

The last model I use is the SVM which Support Vector Machine. This is very strong tool to find the hyperplane in the data then split it to find the different class they belong to. First step is split the dataset in 80 – 20 and use the K-cross validation to find the best C and sigma which help the model have the best performance. Second, I choose 8 feature to run my SVM model and choose the best C and Sigma for train my model. The parameter C is a regularization parameter that controls how tolerant the model is to misclassified training samples. It is a trade-off parameter that balances the width of the classification interval with the classification error on the training set.

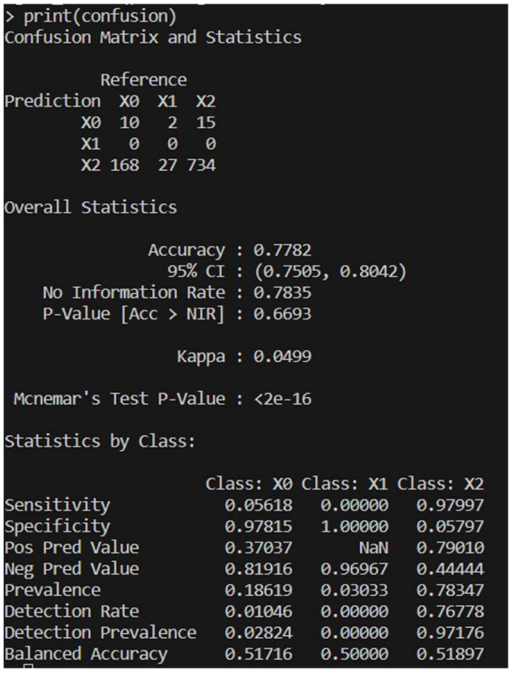
My SVM is using a Gaussian kernel. Sigma is a parameter of the kernel function that determines the extent to which a single training sample affects the shape of the decision boundary. All the parameters is use for avoid overfitting or underfitting of the data. As the Figure 17 show, When the sigma equal to 0.1336754 and C =0.25 will have the accuracy 78% which it can be identified the difference group in the dataset.

```
> print(svm_model)
Support Vector Machines with Radial Basis Function Kernel

3833 samples
   8 predictor
   3 classes: 'X0', 'X1', 'X2'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 3449, 3449, 3450, 3449, 3450, 3451, ...
Resampling results across tuning parameters:

  C     logLoss    AUC        prAUC      Accuracy   Kappa      Mean_F1
  0.25  0.5872341  0.6407227  0.4203355  0.7873685  0.1034306  NaN
  0.50  0.5872112  0.6489784  0.4199666  0.7855428  0.0933052  NaN
  1.00  0.5850874  0.6455521  0.4229884  0.7871088  0.1013211  NaN
  Mean_Sensitivity  Mean_Specificity  Mean_Pos_Pred_Value  Mean_Neg_Pred_Value
  0.3601912         0.6911886         NaN                  0.8111899
  0.3576411         0.6886990         NaN                  0.8057759
  0.3593568         0.6904782         NaN                  0.8089055
  Mean_Precision  Mean_Recall  Mean_Detection_Rate  Mean_Balanced_Accuracy
  NaN             0.3601912    0.2624562            0.5256899
  NaN             0.3576411    0.2618476            0.5231700
  NaN             0.3593568    0.2623696            0.5249175

Tuning parameter 'sigma' was held constant at a value of 0.1336754
Accuracy was used to select the optimal model using the largest value.
The final values used for the model were sigma = 0.1336754 and C = 0.25.
```

Figure 17

In order to test my SVM model, I test my SVM on the test set and use the confusion model to see how it performs. One question I notice is it have the similarities question with the Decision tree which I should be combine the people who take one shoot of vaccine with people who take the booster shoot. As the Figure 18 show there no X1 will be correct group which one reason is the example is too small and make other data misclassified. The accuracy of the model is 0.7782, which means that the probability that it correctly predicts the category is about 77.82%. The 95% confidence interval for the accuracy is between 0.7505 and 0.8042, which suggests that if the model is applied to other samples from the same aggregate, the model's accuracy is likely to fall within this interval. The p-value for McNemar's Test is less than 2e-16, which is very low and indicates that the difference in prediction accuracy between categories is significant. The SVM model is heavily biased towards predicting the X2 category, possibly at the expense of other categories. This may be due to an imbalance in the categories of the training data or problems with the choice

of model parameters.



```
> print(confusion)
Confusion Matrix and Statistics

          Reference
Prediction  X0  X1  X2
        X0  10   2  15
        X1   0   0   0
        X2 168  27 734

Overall Statistics

               Accuracy : 0.7782
                 95% CI : (0.7505, 0.8042)
    No Information Rate : 0.7835
    P-Value [Acc > NIR] : 0.6693

                  Kappa : 0.0499

 Mcnemar's Test P-Value : <2e-16

Statistics by Class:

                     Class: X0 Class: X1 Class: X2
Sensitivity            0.05618   0.00000   0.97997
Specificity            0.97815   1.00000   0.05797
Pos Pred Value         0.37037       NaN   0.79010
Neg Pred Value         0.81916   0.96967   0.44444
Prevalence             0.18619   0.03033   0.78347
Detection Rate         0.01046   0.00000   0.76778
Detection Prevalence   0.02824   0.00000   0.97176
Balanced Accuracy      0.51716   0.50000   0.51897
```

Figure 18

Conclusions and Discussion (3%):

In this project, I explored the factors that influence COVID-19 vaccination behavior. Through an in-depth analysis of the HHS COVID-19 Monthly Outcomes Survey dataset on HealthData.gov[1], I used a variety of classification models such as Random Forests, Decision Trees, Support Vector Machines (SVMs), and Kmean, and my findings revealed significant influences of key factors such as income, education, insurance, employment, and good behavior on vaccination. key factors have a significant impact on vaccination.

By applying and comparing different machine learning models, we found that Support Vector Machines (SVMs) performed well in dealing with the problem, despite the challenge of category imbalance. The worst was the Kmean model, as there were many places where it overlapped with other cluster classes and the intra-cluster distances differed too much.

Adjustments in data preprocessing and model selection during the project had a significant impact on the final results. Initial models were constructed based on simple assumptions, but as we gained a better understanding of the data, we introduced more complex features and models to capture the nuances of the data. In addition, we found that data imbalance had a significant impact on model

performance, which prompted us to consider the use of merging similar groups to achieve improvements in model performance

If we have the opportunity to continue this research in the future, we will explore more data balancing techniques and experiment with more advanced models, such as deep learning networks, to improve the capture of complex patterns. In addition, we plan to deeply analyze time-series data to understand influences over time, as well as implement more comprehensive feature engineering to reveal deeper connections hidden in the data.

In summary, while our research has yielded positive initial results, we recognize that more extensive research and analysis is needed to more fully understand the factors that influence vaccination behavior. In future work, we expect to be able to apply these insights to design effective public health strategies to promote widespread vaccination to combat the COVID-19 epidemic.

Reference

[1] U.S. Department of Health & Human Services.. HHS COVID-19 Monthly Outcome Survey Wave 28. HealthData.gov. Retrieved from https://healthdata.gov/Health/HHS-COVID-19-Monthly-Outcome-Survey-Wave-09/6itx-ccwh/about_data

[2] U.S. Census Bureau. (2022, March 17). Educational Attainment in the United States: 2021. Retrieved from https://www.census.gov/newsroom/press-releases/2022/educational-attainment.html

[3]Peter G. Peterson Foundation. (2023, November). The Share of Americans Without Health Insurance in 2022 Matched a Record Low. Retrieved from https://www.pgpf.org/blog/2023/11/the-share-of-americans-without-health-insurance-in-2022-matched-a-record-low

[4]Centers for Disease Control and Prevention. (n.d.). Weekly Updates by Select Demographic and Geographic Characteristics. Retrieved from https://www.cdc.gov/nchs/nvss/vsrr/covid_weekly/index.htm

Link to GitHub repository:
https://github.com/wei-jun7/DataAnalytics2023_WeijunLi