1. Overview

We are going to find some factors that influence people's willingness to get vaccinated. In the previous report, we proposed that people’s willingness is related to ICU capacity, mortality, social media sentiment, local policies, and pandemic trends. We did some basic statistical analysis on these factors and obtained the trend of the severity of the epidemic from histograms last time. Therefore, in this report, we cut the original data set based on this trend and give four more detailed hypotheses. Besides, we use those factors mentioned before as features and support the four hypotheses based on the traditional statistical hypothesis testing and machine learning predictive analytics.

2.Hypothesis

Our hypotheses are as follows：

1. The influence of people's tweeting mood on vaccine willingness on social media.Increased fear and sadness may increase people's willingness to get vaccinated.
2. The severity of the epidemic on people's willingness to vaccinate.The increase in the number of deaths and patients may increase people's willingness to vaccinate.
3. The impact of the degree of utilization of medical resources on people's willingness to vaccinate.The increasement in the number of people in hospitals and ICUs may increase people's willingness to be vaccinated.
4. When facing COVID-19, are there differences in the distribution of people’s emotions?

3. Data generated

In COVID ACT NOW dataset, the vaccination data starts from January 15, 2021, so we use the sentiment data range from January 15 to the end of August in 2021. We calculated the number and proportion of tweets dominated by each sentiment each day and saved them in units of days. For example, in the number dataset, the attribute called anger means the number of tweets that mainly express angry sentiment on a certain day. In the proportion dataset, the attribute called anger saved the proportion which is equal to the number of angry tweets divided by the total number of tweets in a day. We saved the number and proportion in two separate data sets because they are used separately in the regression and classification later.

In order to analyze the relation between the people’s willingness to get vaccination and the number of each sentiment each day in regression, we want to use the attribute ‘newIncreaseDoes’ in COVID ACT NOW dataset to reflect the willingness and use data in twitter-COVID-dataset (the number of different emotion’s tweets each day) and combine them together by date attribute.

In order to do supervised learning classification, we generated a new column of modified.csv called willingness. This feature stands for people's willingness to be vaccinated, and this is our label of supervised learning. We think that vaccination ratio can stand of people’ willingness to get vaccinated. So we bin vaccinationsCompletedRatio into three classes by the median 0.001 and the mean 0.0023. After this operation we have three labels: very willing to, indifferent, and very reluctant which value is 3, 2, 1. Hence, we generated a new file called labeled.csv. And we use this dataset to test our hypothesis later.

4.Parametric statistical tests

4.1 T-test

In the T-test we tested hypothesis 4, if there is a notable difference between the distribution of four different emotions. We use four attributes ‘happiness’, ‘fear’, ‘anger’ and ‘sadness ’from day_emotion.csv.

P Values generated by t-test are in the table below. We tried a t-test on each pair of emotions to find if their mean and variance are similar.

Table 4.1. p values of four emotions

p value	**happiness**	**fear**	**anger**	**sadness**
happiness	1	1.2808e-8	2.0000e-4	6.7145e0
fear	1.2808e-8	1	1.1719e-2	3.9128e-62
anger	2.0000e-4	1.1719e-2	1	1.7089e-107
sadness	6.7145e0	3.9128e-62	1.7089e-107	1

From the p values generated by t-test we found that there is uneven variance on each pair of data. So we tried the T’ test on it, however, the p values were still all less than 0.5.

That shows that when facing COVID-19, people’s emotions are distributed differently. And because of this atypism, these attributes may help regression perform better.

4.2 Regression

We want to find if there is a relation between people’s willingness to get vaccinations and the distribution of emotions in tweets. The emotional atmosphere of COVID-19 on twitter will affect people's view of COVID-19, thereby influencing their decision to vaccinate, because not everyone knows the virus very well, so people are subject to social media to decide whether to vaccinate.

We guess it is possible if most people are afraid of COVID-19, then there might be more people willing to get vaccination. It might have a different influence on people due to the different emotions. Based on this hypothesis, we implemented linear regression to find each emotion’s influence .First of all, we combine two datasets (twitter emotion dataset and Covid-19 act now dataset) into one dataset. Using anger', 'fear', 'happiness', 'no specific emotion', 'sadness' the emotion’s attribute as X-values, using changes (difference between the number of vaccinations on the day and the previous day) as Y-value to train linear regression(the data is range from 2021-1-15 to 2021-8-30 amount to 230 pieces of data).After that, we get a linear regression line and a curve between predicted value and test value :

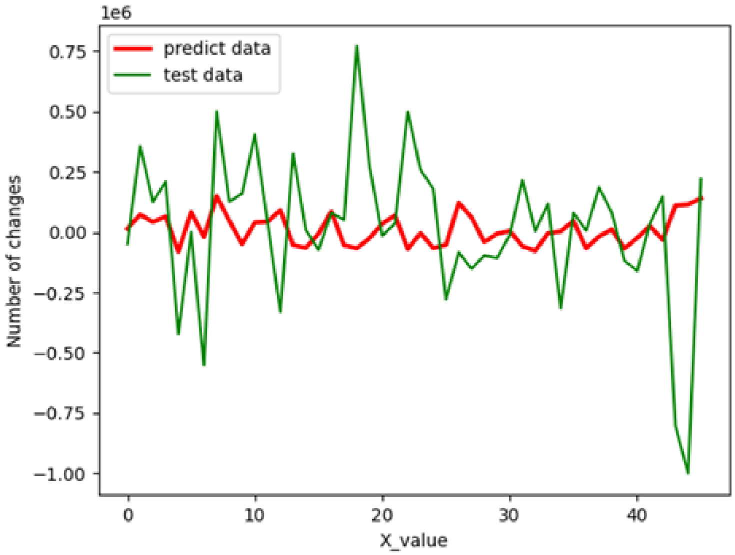


Figure 4.1 A curve between predicted value and test value

Regression line : Y = 103919.12 + 4.25 * anger - 11.63 * fear - 4.22 * happiness - 4.22 * no specific - 19.03 *sadness

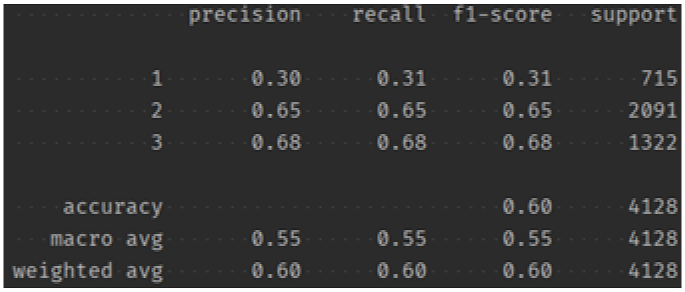
The regression line couldn’t predict well, it is affected by the small amount of data, which also affects the results of the regression line. Thus, we select a relatively good result in the report. Even though its performance is not good, it can partly reflect the trend of changes. So, we can say there is a relation between them.

Based on the regression line, we think that we can get some indication about the emotion’s effect. The coefficient of fear, happiness and sadness is negative. So we guess a happy atmosphere may make people less worried about COVID-19. But it is weird that fear and sadness are also negative, it might because people will ignore some negative information. And when most people are angry about COVID-19, most people might be aware of the seriousness of the virus and then they are willing to vaccinate.

Due to the limited amount of data(there is only eight month’s data), we couldn’t make a good prediction, thus the conclusion above might not be true. All of them are just guesses.

5. Predictive models

5.1 Decision tree and Random Forest

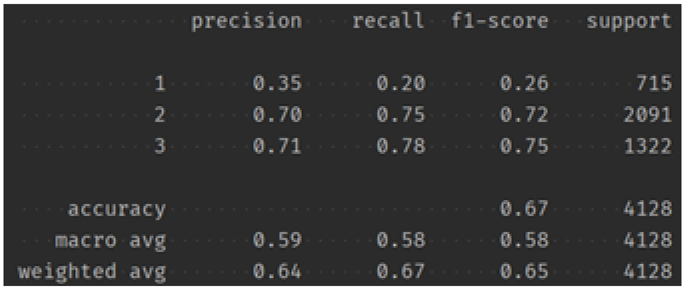


```
              precision    recall  f1-score   support

           1       0.30      0.31      0.31       715
           2       0.65      0.65      0.65      2091
           3       0.68      0.68      0.68      1322

    accuracy                           0.60      4128
   macro avg       0.55      0.55      0.55      4128
weighted avg       0.60      0.60      0.60      4128
```

Figure 5.1. Accuracy of decision tree model



```
              precision    recall  f1-score   support

           1       0.35      0.20      0.26       715
           2       0.70      0.75      0.72      2091
           3       0.71      0.78      0.75      1322

    accuracy                           0.67      4128
   macro avg       0.59      0.58      0.58      4128
weighted avg       0.64      0.67      0.65      4128
```

Figure 5.2 Accuracy of random forest model

We ran the decision tree model and the random forest model together to compare the results. Generally, random forest has more precise results.

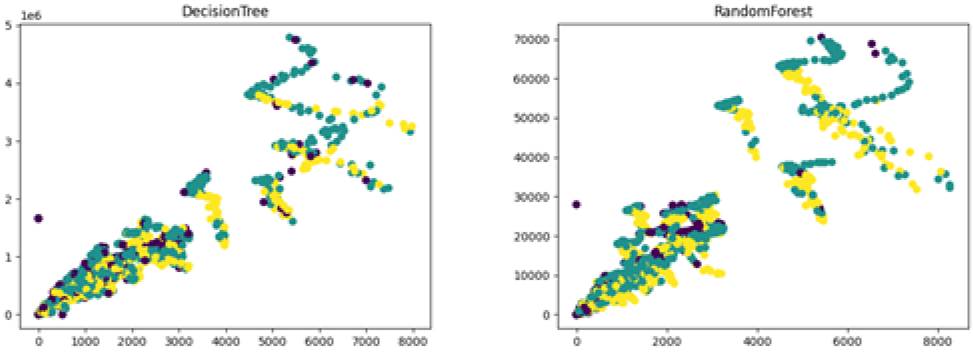


Figure 5.3 Test data for decision tree and random forest

This figure shows the test data of the decision tree and the random forest. The picture of the decision tree is about vaccination ratio and ICU beds usage, the picture of the random forest is about new cases and vaccination distribution number.

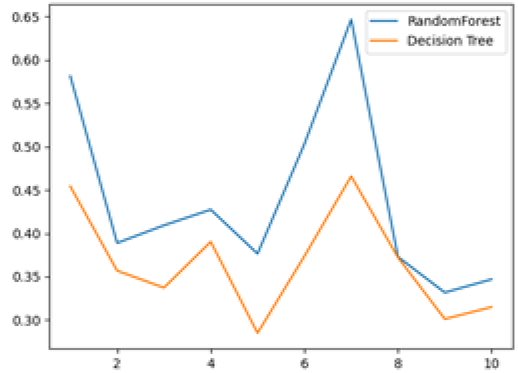


Figure 5.4 Cross validation of decision tree and random forest

This figure shows the result of 10-fold cross validation of decision tree and random forest. It’s obvious that the random forest model always has a better precision.

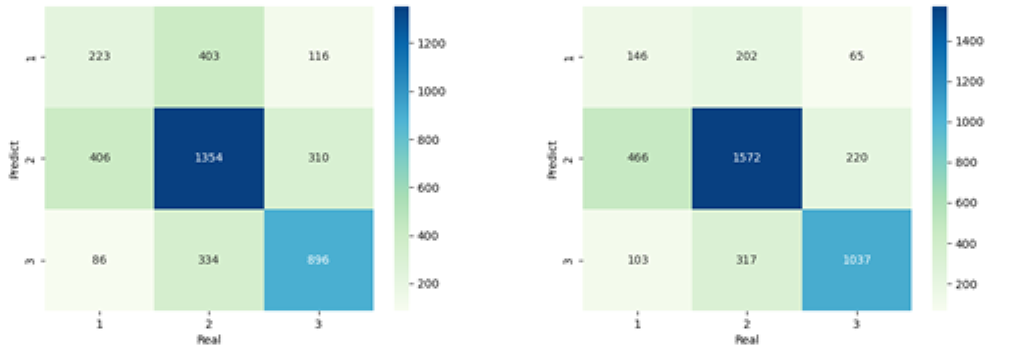
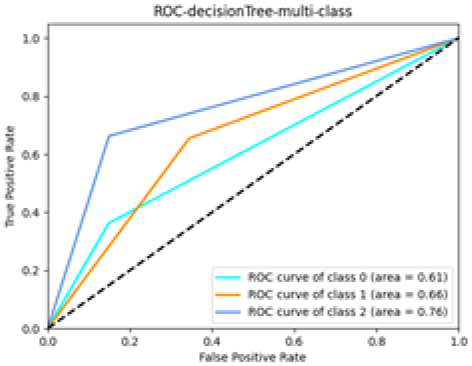


Figure 5.5 Confusion metrics of decision tree and random forest

In this figure, the table on the left represents the confusion metric of the decision tree, and the right table refers to the random forest model.



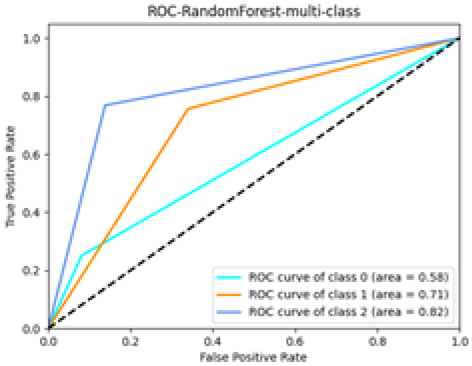


Figure 5.6 ROC curve of decision tree and random forest

This figure shows the result of the ROC curve of decision tree model and random forest model.

In summary, we implement the decision tree model and random forest model to support our second and third hypothesis. All the results are based on all of the attributes in the dataset COVID ACT NOW. The results support our hypotheses that when the usage of medical resources and the number of new cases increases, it will lead to people’s higher willingness of vaccination.

5.2 Lazy Learner Method

We built a KNN model to test hypothesis 2. Knn is one of the simplest classification algorithms, which is very easy to understand and implement. It selects the k samples that are the closest to the point, in which k samples have more categories, and classifies k into this category. Before considering using more advanced technology, trying this algorithm is a good benchmark method. We want to find based on the current severity of the epidemic, such as the number of deaths and confirmed cases, can we predict people's willingness to vaccinate? Here we believe that the vaccination rate can represent people's willingness to vaccinate.

In the KNN model, the independent variables we use are number of positives, ratio of positives, ratio of negatives and number of deaths. The label is the people's willingness to vaccinate, we have three labels : very willing to, indifferent and very reluctant.

We use ten-fold cross validation in this hypothesis, the accuracy is shown as follows(Table x). The mean accuracy of hypothesis 2 in the KNN model is 58.2%. This accuracy is not high, we think there may be two reasons. First is the problem of the model, when the samples are unbalanced, the category of the new sample is biased toward the dominant category in the training sample, which can easily lead to prediction errors. In our data, the indifferent number is the largest, the very reluctant number is the smallest. The ratio of very willing to, indifferent and very reluctant is 17%, 50% and 32%.So this imbalance is likely to cause our prediction accuracy to drop.The second reason may be that the relationship between the severity of the epidemic and people's willingness to vaccinate is not so direct，58.2% accuracy is reasonable of this reason .

The figures below show the confusion matrix and the ROC curve of KNN.

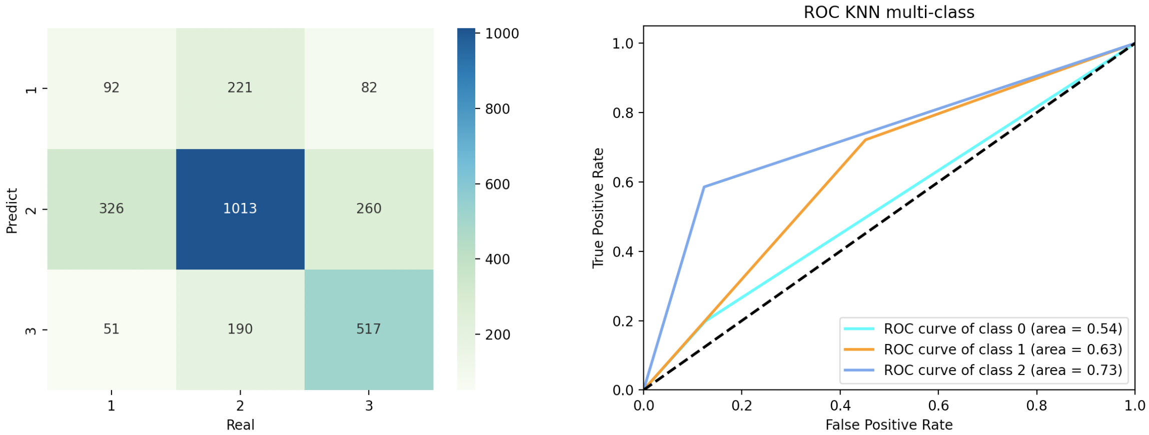


Figure 5.7 Confusion matrix and ROC curve for hypothesis 2 using KNN

It can be seen from the roc curve that the prediction accuracy rate of people have a strong willingness to vaccinate is the highest, it reaches 73%.We believe that this shows that the severity of the epidemic may be related to the willingness to be vaccinated, but the willingness to not be vaccinated may be related to more factors.

5.3 Naïve Bayes

We built a three-category Naive Bayes model to divide the data into three levels of vaccination willingness. Naive Bayes method is a classification method based on Bayes' theorem and the assumption of independence of feature conditions. In this report, we use Multinomial Naive Bayes to implement hypothesis 2 and 3.

The labels of these two hypotheses are the level of people's willingness. For hypothesis 2, we use six features to reflect the severity of the epidemic: testPositivityRatio, caseDensity, infectionRate, deaths, positiveTests and negativeTests. For hypothesis 3, we use two features to reflect the utilization of medical resources: hospitalBeds currentUsageCovid and icuBeds currentUsageCovid.

We use ten-fold cross validation in the two hypotheses, and the accuracy of these two are shown as follows(Table x). The mean accuracy of hypothesis 2 is 40.7% and the mean accuracy of hypothesis 3 is 42.1%. We guess that because the features we use are not completely independent of each other, they may not fit the Naive Bayes model very well.

Table 5.1 The accuracy of hypothesis 2 and 3 using ten-fold cross validation

Hypothesis	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Fold 6**	**Fold 7**	**Fold 8**	**Fold 9**	**Fold 10**
2	0.402	0.430	0.419	0.392	0.498	0.426	0.411	0.400	0.396	0.395
3	0.403	0.426	0.439	0.422	0.420	0.408	0.425	0.452	0.402	0.414

The figures below show the confusion matrix and the ROC curve of two hypotheses. Multinomial Naive Bayes does better when classifying utilization of medical resources than the emotion from tweets. The main reason may be that the vaccination started in January 2021, and after the promotion of the vaccine, people’s emotions on Twitter were more stable than before. Therefore, when we only used emotional states after the vaccine was promoted, the lack of emotional fluctuations caused an impact on the classification.

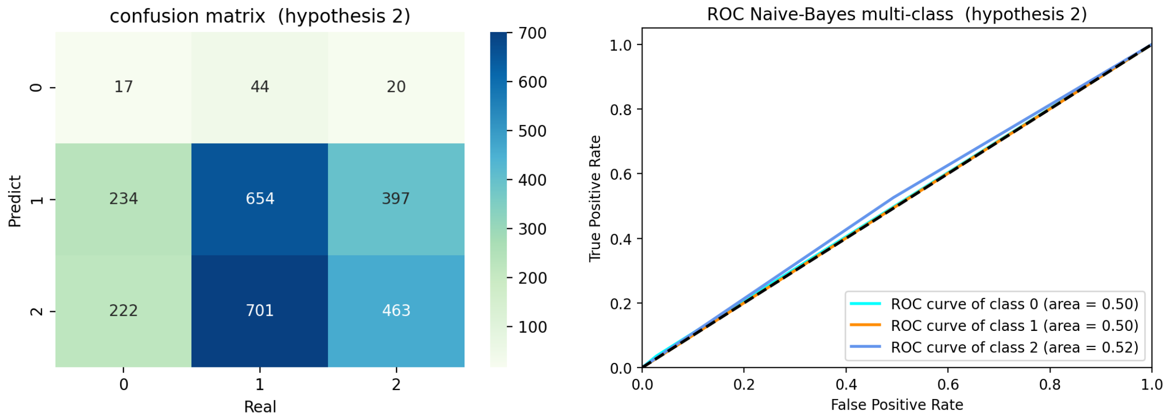


Figure 5.8 Confusion matrix and ROC curve for hypothesis 2 using Naive Bayes

From the confusion matrix, especially when we pay attention to the use of medical resources, we can see that the model does better for class 1 and class 2 classification, which means this model can better classify situations that will lead to strong vaccination willingness

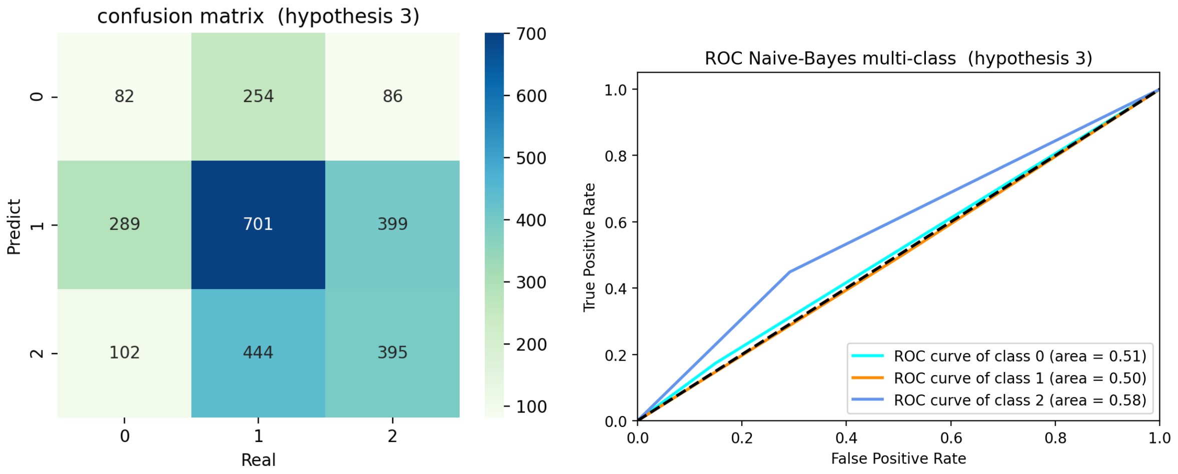
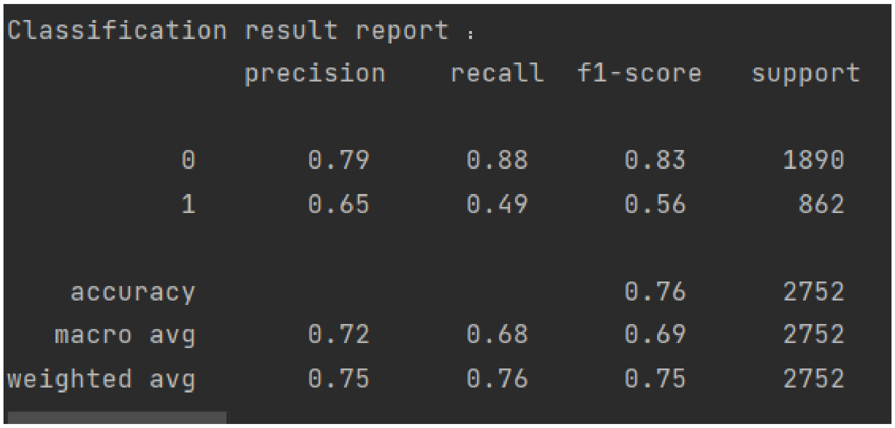


Figure 5.9 Confusion matrix and ROC curve for hypothesis 3 using Naive Bayes

5.4 SVM

We built a SVM model to find the impact of the degree of utilization of medical resources on people's willingness to vaccinate.Tests(hypothesis 3). We use these features to reflect the utilization of medical resources: hospital Bed capacity, hospitalBedscurrentUsageTota, hospitalBedscurrentUsageCovid, and we split willingness into two classes to fit the SVM model. Class 0 (inoculation change rate from 0 to 0.023, which reflecting indifferent attitude), Class 1 (inoculation change rate from 0.023 to 0.068, which reflecting very willing to)



```
Classification result report :
              precision    recall  f1-score   support

           0       0.79      0.88      0.83      1890
           1       0.65      0.49      0.56       862

    accuracy                           0.76      2752
   macro avg       0.72      0.68      0.69      2752
weighted avg       0.75      0.76      0.75      2752
```

Figure 5.10 Classification of SVM model

We ran the model, and the classification results show above. In order to find that best parameter of SVM, we set two parameter of svm model, C in [0.5, 1, 3, 5, 7, 9, 11] and gamma in [0.000001,0.00001, 0.0001, 0.001, 0.1, 1, 10]. We don’t modify the parameter kernel and we set it as ‘rbf’. Then we can get a matrix of the accuracy of different parameters:

Table 5.2. The accuracy of different parameter

C\gamma	**0.000001**	**0.00001**	**0.0001**	**0.001**	**0.1**	**1**	**10**
0.5	0.719	0.736	0.740	0.710	0.686	0.686	0.686
1	0.722	0.746	0.749	0.723	0.687	0.687	0.687
3	0.725	0.756	0.740	0.727	0.689	0.687	0.686
5	0.725	0.739	0.724	0.724	0.689	0.687	0.686
7	0,728	0.746	0.739	0.723	0.689	0.687	0.686
9	0.728	0.747	0.739	0.720	0.689	0.687	0.686
11	0.729	0.743	0.738	0.720	0.689	0.687	0.686

Based on this matrix, we find when C = 3, gamma = 0.00001 get the best accuracy, Thus. our later work on SVM is based on these two values.We use 10-fold cross validation on SVM, the mean accuracy is 0.7435 and std is 0.010418. The standard deviation is pretty low so that we can say And we also draw the confusion matrix and ROC curve of this model:

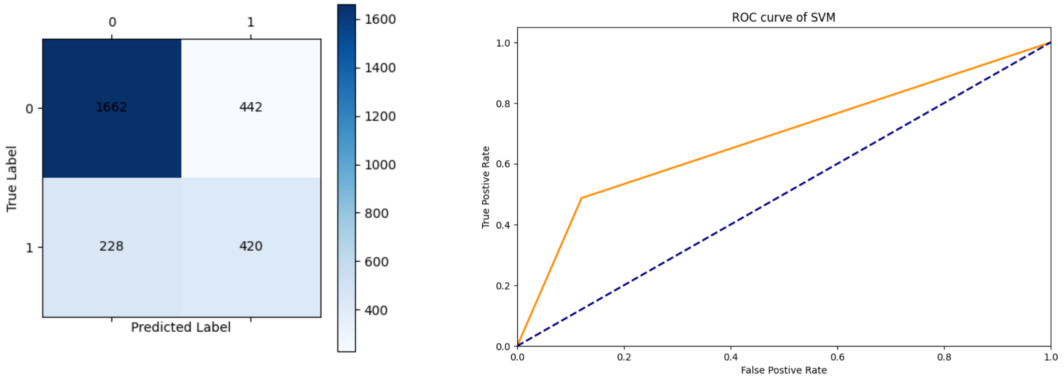
- 

Figure 5.11. Confusion matrix and ROC curve for hypothesis 3 using SVM

It can be seen from the roc curve that the prediction accuracy is relatively high and it reaches 74%. At first, we believe that this shows that hospital situations may be related to the willingness to be vaccinated. But when we see the results of classification in the figure classification of SVM model, the accuracy to predict indifference willingness(0.79) and a relatively low accuracy on high willingness(0.65), which means the model got a good prediction when people is reluctant to vaccinate but did worse on the situation people are very willing to vaccinate. We thought it indicates that information about hospital resources might not promote people to vaccinate. The relation between medical resources and vaccination ratio might be caused by the fact that the more room in hospital for vaccination, the more people can get vaccinated. If there were many empty rooms for vaccination, it couldn’t motivate people to get vaccinated.

6.Fairness

In addition, for the SVM model we also apply p_precent score to evaluate the fairness of this model. From the features we used, we think the ‘hospitalBeds capacity’ will cause bias because some small hospital resources are limited which means they can’t provide vaccination for COVID-19. Thus, the ratio of small capacity and large capacity might be extremely low. So, we implement P score on our SVM model. Because hospitalBeds capacity feature is not a binary feature, thus, we use the mean of this feature to split them into two classes(0 and 1). And set it as the model’s sentiment feature. Finally, we get p_percent_score= 0.459. It is pretty low and indicates that there might be some bias for the small capacity hospital..

7.Conclusion

For the four hypotheses we proposed this time, in the t-test we found that the four sentiment distributions in hypothesis four are not very consistent. We originally expected that positive emotions might be inconsistent with negative emotions, but the distribution of negative emotions might be more consistent. However, this result shows that people's attitudes towards the new crown are still very diverse and complicated. Then we implement regression to find the relation between emotions and people’s willingness to vaccinate. And we found that there is some relationship between them.

Regarding the hypothesis that features affect people’s willingness to vaccinate, we predicted both hypothesis two and three in the Naïve Bayes , but the results are not very good. However, hypothesis 2 performs well in KNN model, so we think it may be that the Naïve Bayes model itself is not so suitable for our data. In KNN, we observed that people’s strong willingness to vaccinate is positively correlated with the severity of the epidemic. The accuracy of predicting this class is relatively high, but the accuracy of predicting people’s unwillingness to vaccinate is significantly lower. We think there may be other factors that affect the accuracy of the forecast. Therefore, in the decision tree and random forest, we combined hypotheses 2 and 3 together, used both the severity of the epidemic and the medical conditions to predict. The result was still that the "very willing to" class was more accurate than "indifferent" class, the accuracy of "indifferent" class is more than "very reluctant".So our current conclusion is that we can predict people’s willingness to be vaccinated more accurately based on the severity of the epidemic and the situation in the hospital, but people’s willingness to not want to be vaccinated may be affected by more complicated other factors. Will try to solve this problem in the follow-up process.