COMP 47350 ASS2-Individual Report
Student name: Xuhui An
Student number: 20211294

1. Personal contribution

I did part 1-3 and prepared dataset in this assignment.

1.1 Data understanding and prep:

Preparation of data: I merged 2 original dataset as both of us didn't do very well in the first assignment, so we need to repeat the process of cleaning the data again following the professor's instructions: cleaned all duplicate features and rows, change all null and unknown value into Missing, checked if there is constant column…Besides that I did logistic test, which I didn't done in the last assignment. Lastly, I prepare a data quality plan.

Split the dataset into test and training set: before splitting the dataset, there's still few things to do: checking if there is null value, I found there's still one null value in state_fips_code column, I just delete it for one row has nearly no impact on the whole dataset. Here I used integer-encoding, the reason is: before building the predictive model, we don't know the relationship between all features and the target feature, and there are lots of values in this dataset. If I use one-hot-encoding here, that will lead to too many columns which means too many variables, which is not beneficial to model building. Then shuffling the dataframe and removing the target data-"death_yn" from the dataset.

Then is plotting all features in the dataset with target features to see if we need them in the later model building: as we don't have continuous features in this dataset, so I only plot categorical features with "death_yn". Here I dropped 4 features: 'county_fips_code', 'current_status', 'ethnicity' and 'race', the reason is: 'county_fips_code' is like a more specific version of 'state_fips_code', thus they are duplicates. Two values in 'current_state ' have tiny differences with the target feature, so I deleted them. The amount of two values in 'race' has a huge gap, 14000 vs. 2, which cannot support reliability for model building. For the same reason, I deleted the 'ethnicity' feature.

1.2 Predictive modelling: Linear regression

Linear regression model: first thing is to train a linear regression model with the descriptive feature that I prepared in the previous part. Then print out all coefficients learned by the model. From the result I see that 'age_group' has the strongest relationship(weight=0.2197), whereas 'state_fips_code' has the weakest relationship(weight=0.00299).

Printing 100 predicted target features and evaluating the prediction: we evaluate the model with train data, test data and cross validation, all of them showing the model is more good at predicting navigate cases, which makes sense as there are more negative values in the original dataset. After cross validation the results are similar to previous: accuracy 88.16%, precision 73.58%, recall 74.85% and f1 74.20%.

1.3 Predictive modelling: Logistic regression

Logistic regression model: same steps with linear regression. From the result I see that 'age_group' has the strongest impact (weight=2.50), whereas 'state_fips_code' has the weakest impact(weight=0.0285) on the target feature.

Printing 100 predicted target features and evaluating the prediction: same steps with linear regression. After cross validation the results are similar to previous: accuracy 88.50%, precision 73.94%, recall 76.80% and f1 75.34%.

2. What did you learn from the project? (500 words)

First thing is understanding the data, which is actually the content of the first assignment. I did not do very well in the first assignment, leaving lots of meaningless values and features in the dataset, compared to removing them from the dataset totally. I just ignore them when plotting, which is not processing data. Moreover I didn't do any logistic test of the dataset. Thus after the professor launched the reference of how to process all the features, I followed the instructions step by step. The first thing is: when two features are telling the same thing, for example res_state and state_fips_code, we can delete one of them from the dataset. Second, if there's over 50% null value in a feature and the feature has a tiny relationship with the target feature, just drop it no matter how diverse the rest values are, like "process" column. A logistic test is essential, when dealing with the dataset, it's only important to consider the relationship of features with target features. The relationship of features is also important. For example, the logic of "hosp_yn" and "icu_yn" is when a person fills "Yes" in the "icu_yn" column. He must fill "Yes" in the "hosp_yn" column too, or it will be illogical, then this row will be meaningless.

Prepare dataset for predictive model: As all features in our dataset are categorical, so I need to transfer them to "int64" type, or in the plotting process the x-axis will only show 0. When I decided to keep which features, I learnt that firstly, I need to consider the difference in the count of individual values in a feature. For example, in the 'race' feature, the 'Native Hawaiian/Other Pacific Islander' has only two samples, and both of them survive, which does not indicate that the race has a 100% survival rate in the covid. So the 'race' column can be removed.

Secondly, consider different encoding types for different features, it's not mean one type of encoding is appropriate for all features. Sometimes we need to check the relationship of this feature with the target feature then decide how to encode it.

Predictive modelling: I learnt that the coefficient in linear and logistic regression stands for different meaning, in linear regression coefficient can be seen as the relationship of this feature with the target feature while in logistic regression, higher coefficient means this changes in this feature will have higher impact on predictive target feature.

COMP 47350 ASS2-Individual Report
Student name: Xuhui An
Student number: 20211294

3. Anything else related to this project that was not covered in the previous 2 parts (500 words)

I'd like to talk about the encoding type we chose, which is a hybrid way. Although all of us know that we should use one-hot-encoding for features in linear regression, for the reasons I mentioned in both Part 1 and notebook, I still chose integer-encoding firstly.
In the final improvement section in Part 5, we decided to use both integer and one-hot-encoding in this dataset. The result is unexpected: Compared to using one-hot-encoding on all features, hybrid encoding could improve models in general. Although for linear regression models the accuracy and precision decreased a little, it still works very well on other models. So we can say in this dataset, hybrid encoding is better than only one-hot-encoding.
The reason might be that: values of features in this dataset is many(there are 34 unique values in case_month feature), use one-hot-encoding on all of them means add so many variables in model building, which might cause:

1.Overfitting: When a model contains a large number of variables, it will fit the training data more easily, but may perform poorly on new data.

2.Increased computation time: If the number of variables is large, the computer needs to spend more time to train the model, which may lead to a significant increase in training time.

3.Effect of noisy variables: If the model contains a large number of irrelevant variables, these may interfere with the predictive power of the model and may even lead to a reduction in the model's performance.

4.Reduced explanatory power: When the number of variables is large, the model becomes more complex and this may reduce the explanatory power of the model..*[1-2]*

Reference:

[1] *Predictive Models Using Regression*-Richard V. McCarthy, Mary M. McCarthy & Wendy Ceccucci

[2] *Chowdhury MZI, Turin TCVariable selection strategies and its importance in clinical prediction modellingFamily Medicine and Community Health 2020;8:e000262. doi: 10.1136/fmch-2019-000262*

COMP 47350 ASS2-Individual Report
Student name: Xuhui An
Student number: 20211294