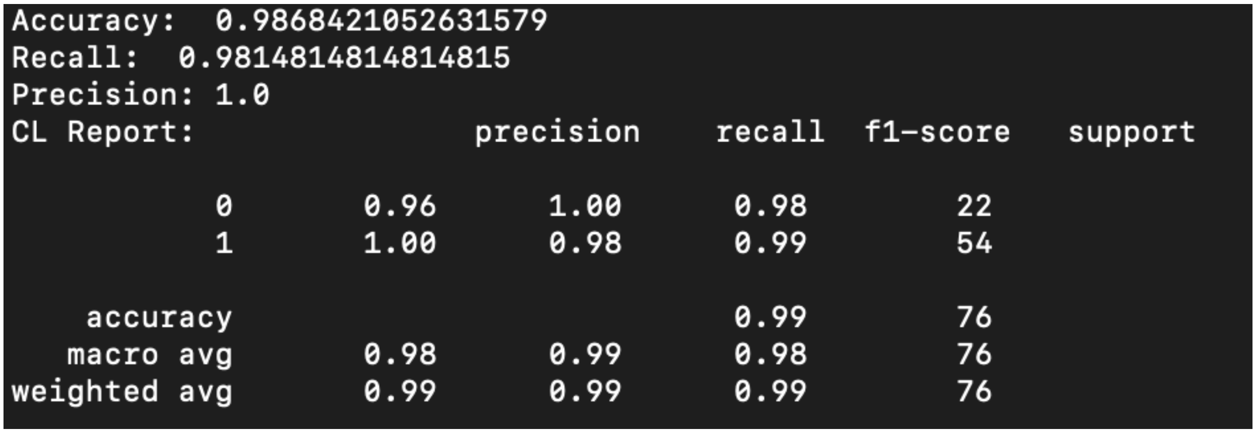
Data science test report
Kevin Ma

I viewed this challenge as a binary classification question: whether the patient would be alive after 12 months. I chose a logistic regression model and used chi-square test to find the features most related to the outcome. I have created the basic approach and will discuss results, shortfalls, and further work needed to create a more comprehensive model.

Summary of Methodology
First, I eliminated incomplete data columns, including Tumor Stage, Number of metastasis in LNs, Number of distant mets, and Tumor size. After fixing some typos (1B vs IB, Right vs Righ), I created a conditional column, "OneYearSurvival", to be true under this condition: "patient was alive at follow-up" OR "patient died but the follow-up was longer than one year later". This is the dependent variable, or outcomes, of the predictive model. Next, I separated the categorial labels (stage, histology, primary site) into true/false labels, as seen in attached output.csv.

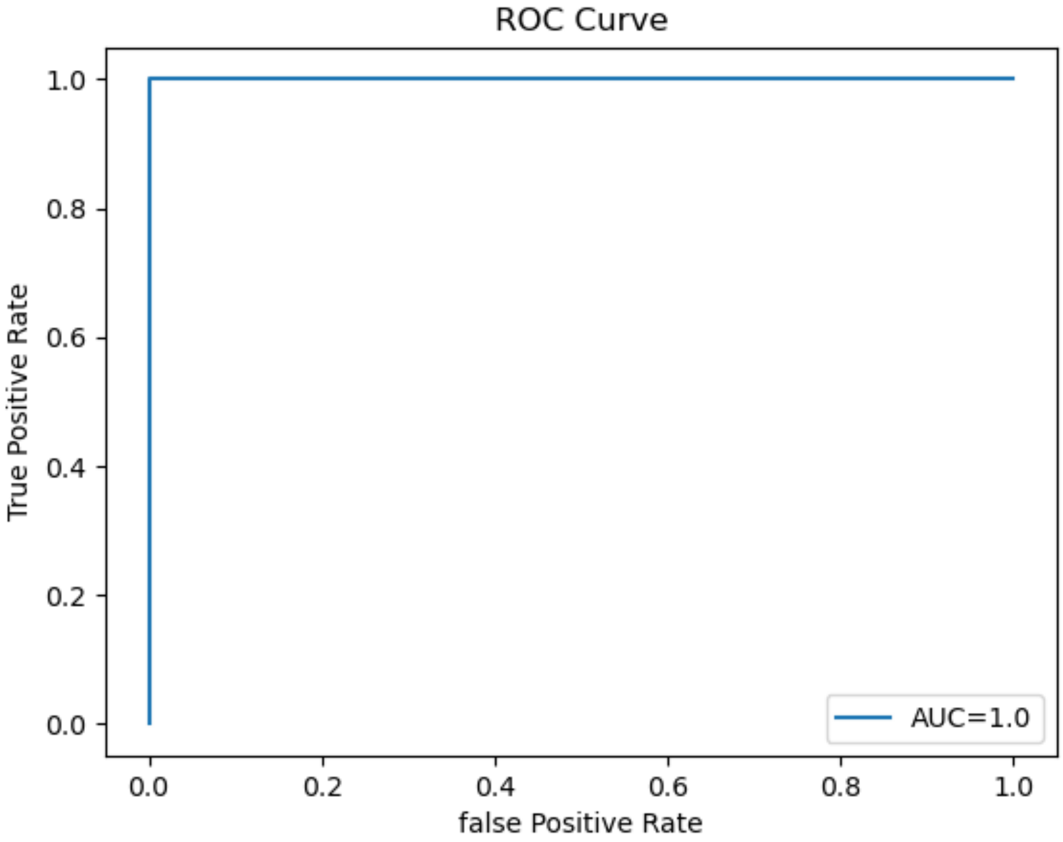
Next, I ranked the features by their correlation scores with the outcome. The scores are saved in features_scores.csv. I then dropped the continuous labels that are least correlated with outcome: age and number of mutations. I then split the data into 60% training and 40% testing, then created the logistic regression model. I then assessed the accuracy score, recall, precision, and classification report, and created the ROC graph, as shown below:



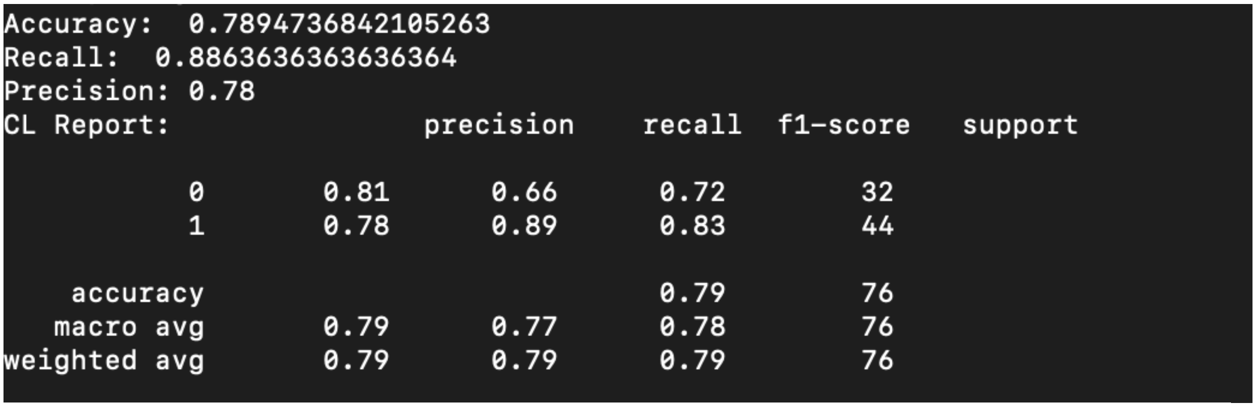
```
Accuracy:  0.9868421052631579
Recall:  0.9814814814814815
Precision: 1.0
CL Report:               precision    recall  f1-score   support

           0       0.96      1.00      0.98        22
           1       1.00      0.98      0.99        54

    accuracy                           0.99        76
   macro avg       0.98      0.99      0.98        76
weighted avg       0.99      0.99      0.99        76
```



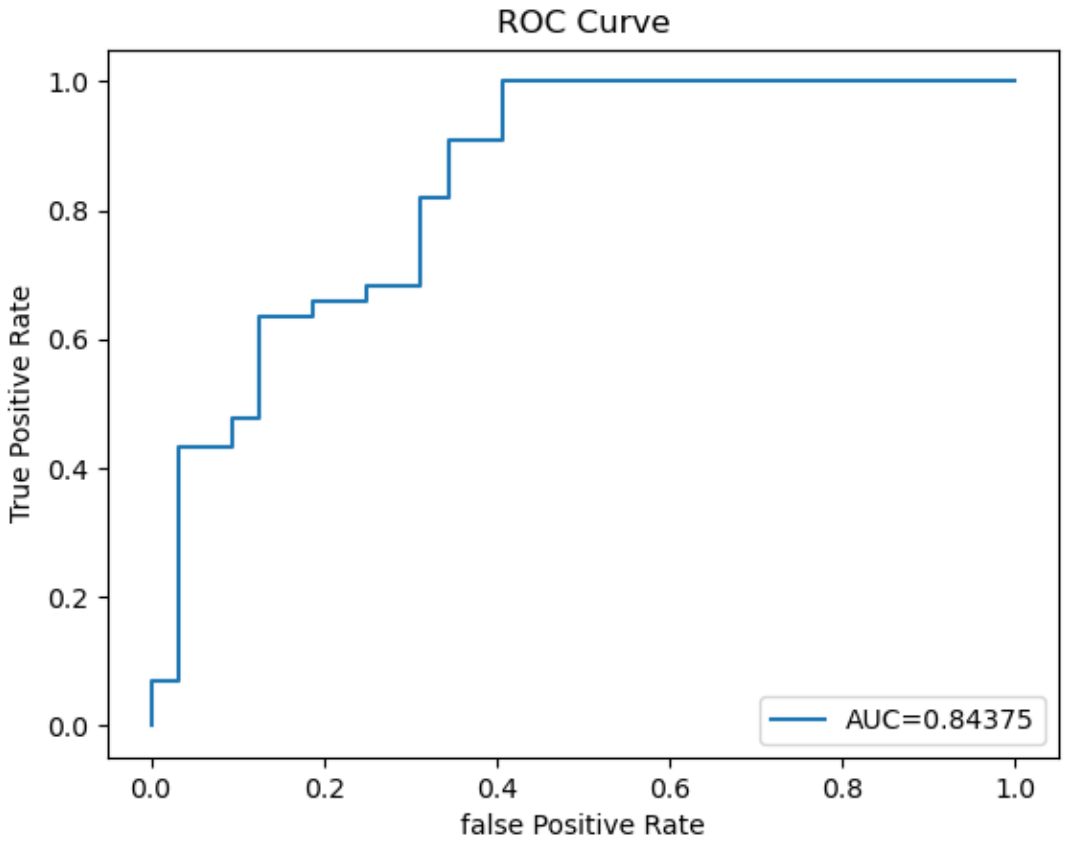
After evaluating the data, I found that my outcome criteria created a data set with 140 “alive” and 50 “dead” labels. I changed my criteria to “follow up was longer than 16 months” and reduced the number of positive to 104, versus 86 negatives. The second model output results are shown here and saved under “output2.csv” and “features_scores2.csv”.



```
Accuracy:  0.7894736842105263
Recall:  0.8863636363636364
Precision: 0.78
CL Report:               precision    recall  f1-score   support

           0       0.81      0.66      0.72        32
           1       0.78      0.89      0.83        44

    accuracy                           0.79        76
   macro avg       0.79      0.77      0.78        76
weighted avg       0.79      0.79      0.79        76
```



Discussion

The initial test achieved high classification scores and high accuracy; however, this could be because of bias in outcome. All predictive metrics are reduced after the outcome condition was changed from 12 months to 16 months. White altering outcome condition from “alive after 12 months” to “alive after 16 months” creates a more balanced outcome classification, this is not a permanent solution. Other methods such as data augmentation and weight distribution may be better in normalizing the prediction model.

This model does not include the genomic data due to lack of time. Presence and absence of certain genomic labels may affect the outcome prediction. Another methodology to correct the incomplete data in excluded columns can also be considered to create a more robust model.