**Cancer Survival Prediction Model Using Haberman Dataset**

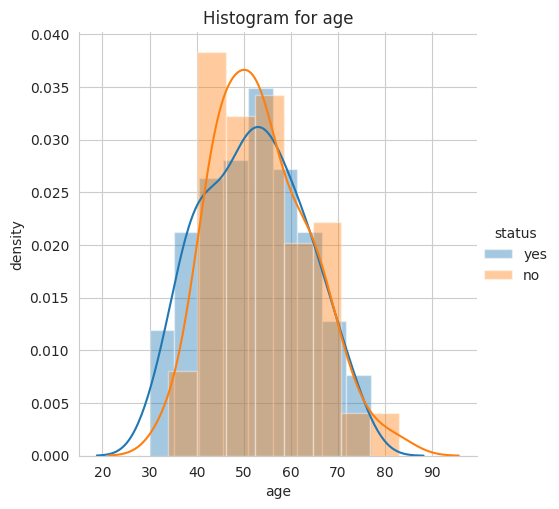


**Project Objectives**:

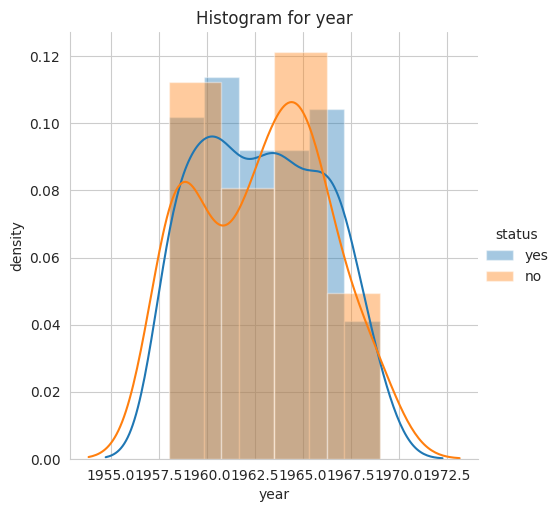
The specific objectives of this project are to:

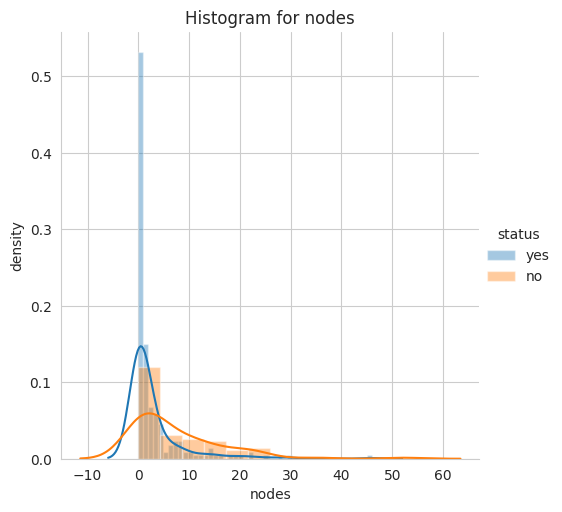
* Collect and preprocess a dataset of historical patient data.
* Develop a machine learning model that can predict the survival of cancer patients.
* Evaluate the performance of the model on a held-out dataset.
* Deploy the model in a production environment.

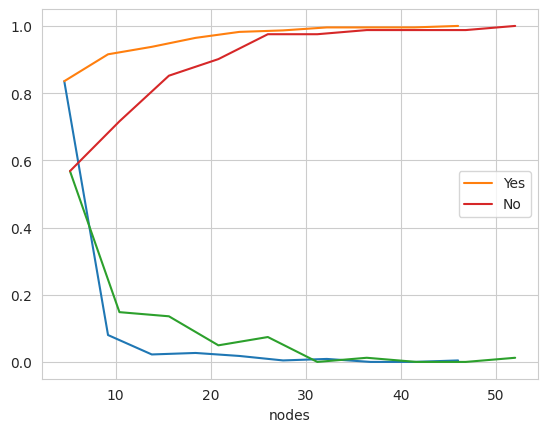
**Exploratory Data Analysis:**



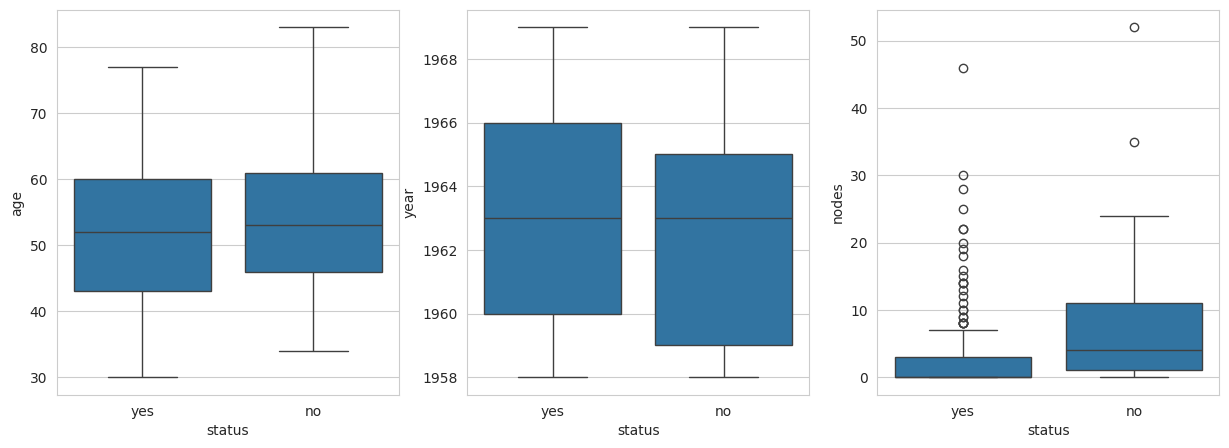
* Major overlapping is observed, which suggests that age isn’t a major determining factor in the patients likihood of survival.
* Differences between the age of the yes dataset and no dataset are barely observable given the amount of overlap in the PDF.
* Perhaps another statistical method can unearth a pattern between age and survival status.
* Model Insights: Ages 30–40 had a higher chance of survival, whereas ages 40–60 did not. For ages 60+ the chances of survival were about 50/50.

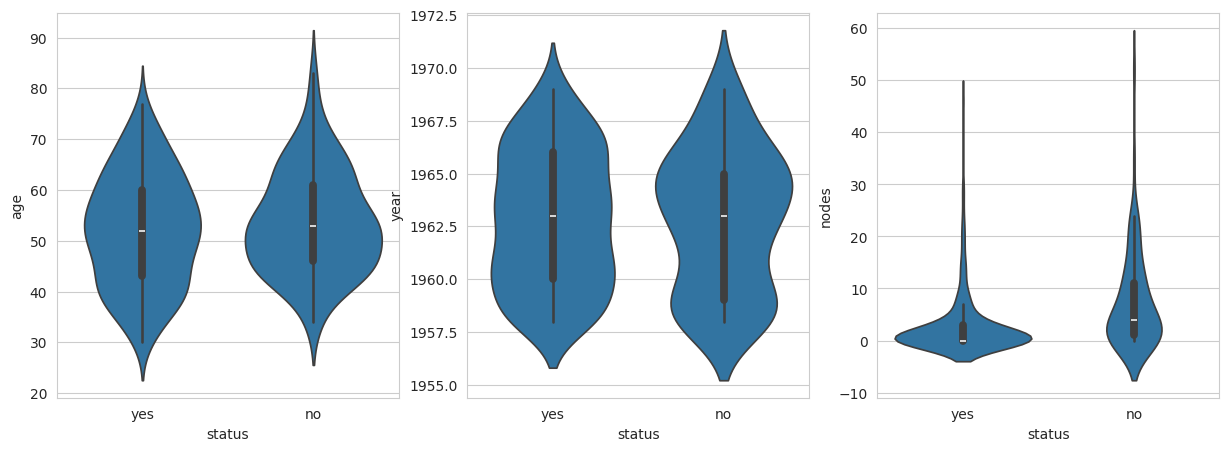


* Major overlapping continues again suggesting that the year of the patient’s surgical procedure did not affect their survival rate/outcome after 5 years.
* There was a spike in the death rate for patients whose surgery was in year 1965 and a decrease for procedures done in 1960. Patient’s likelihood of survival was up between 1960–1962.  
    
    
  
* Complete separation would be ideal to distinguish the exact number of nodes for patients who survived.
* Patients with 0 nodes or 1 node are more likely to survive. There are very few chances of surviving if there are 25 or more nodes.
* This plot has shown that the number of nodes seem to influence the survival rate of patients more so than age and year of operation.
* Model Insight: Patient non-survival increasingly likely after 5 nodes.

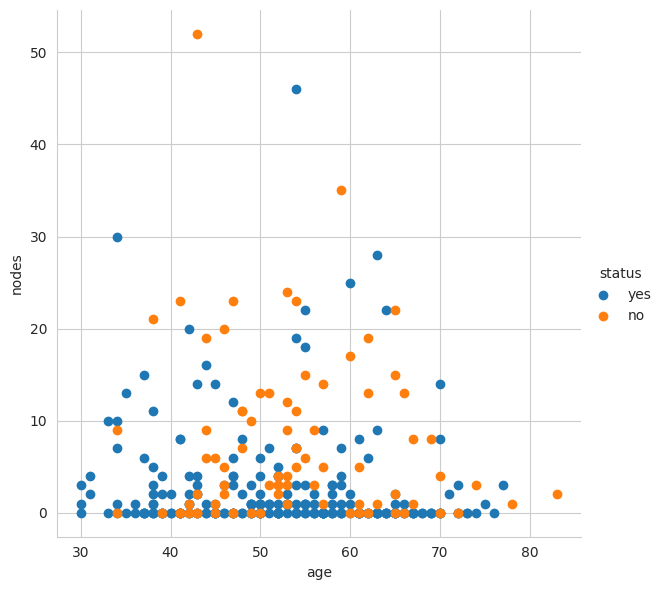


* Approximately 83.55% of patients who survived had nodes in the 0 to 4.6 range as per the CDF summary stats.

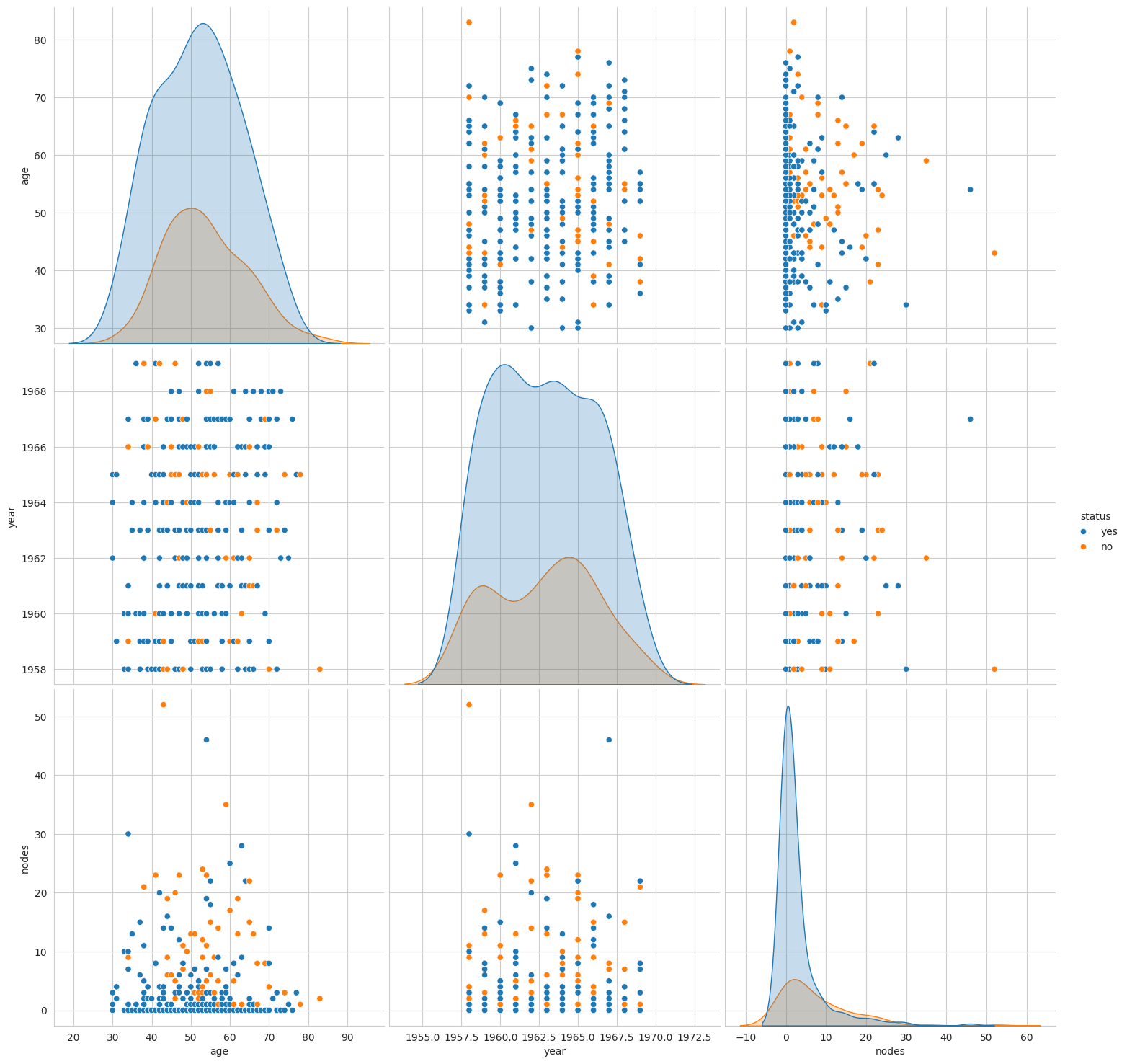


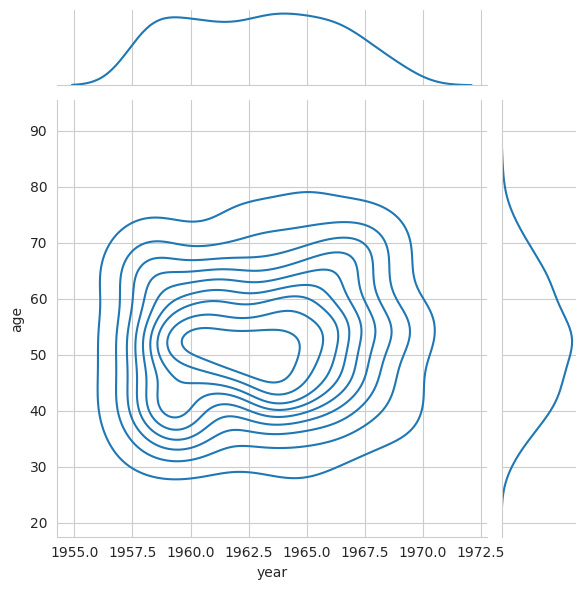
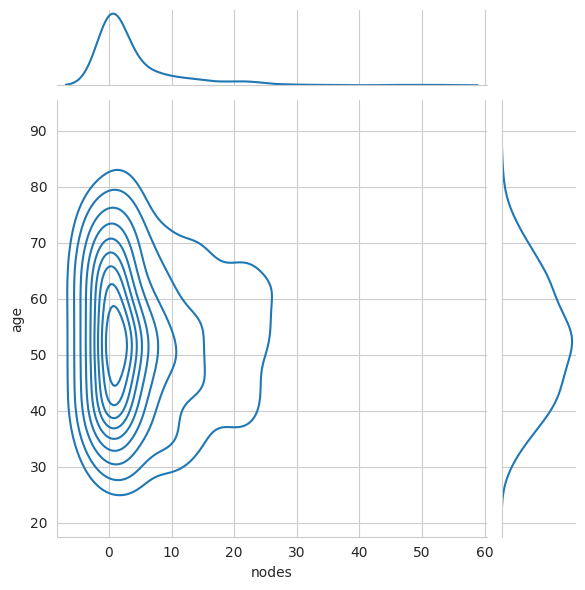


* There were comparatively more people in the age group 45 to 65 who did not survive. However, patient age alone is not an important parameter in determining their survival.
* There were more people operated on in the year 1965 that did not survive beyond 5 years.
* Patients with more than 1 node are less likely to survive. Generally, more nodes hint toward a decreased survival rate. A large percentage of patients who survived had 0 nodes. Yet there is a small percentage of patients who had 0 positive axillary nodes, who died within 5 years of operation, thus an absence of positive axillary nodes cannot guarantee survival.
* The box plots and violin plots for age and year parameters give similar results with a substantial overlap of data points. The overlap in the box plot and the violin plot of nodes is less compared to other features but the overlap still exists and thus it is difficult to set a threshold to classify both classes of patients.



* Patients with 0 to 15 nodes aged between 30 and 40 are more likely to survive. Irrespective of age patients with 0 to 1 nodes are more likely to survive. Patients aged 50+ with 5+ nodes are more likely to have not survived.



* In the pair-plot between age vs nodes, class variables can be differentiated between each other with two overlapping bell-shaped curves. For further analysis age and nodes attributes can be considered for pattern finding as you can determine, however vaguely, a relationship. However, all other attributes did not show much difference in their PDF.  
    
    
   
* From about 1960 to 1964, there were more surgical procedures completed on the patients between the ages 45 and 55. The majority of patients with 0 to 1 nodes were between 45 to 65 years of age.

**Conclusions**

Patient’s age and operation year alone is not the lone deciding factors for their survival. Yet, individuals of ages less than 35 years have an increased chance of survival. The likelihood of survival seems to be inversely proportional to the number of positive axillary nodes discovered in each patient. However, the data has shown that an absence of positive axillary nodes does not always guarantee survival. The objective of classifying the survival status of a new patient is difficult and will likely produce unreliable results because the dataset needs additional features and attributes.

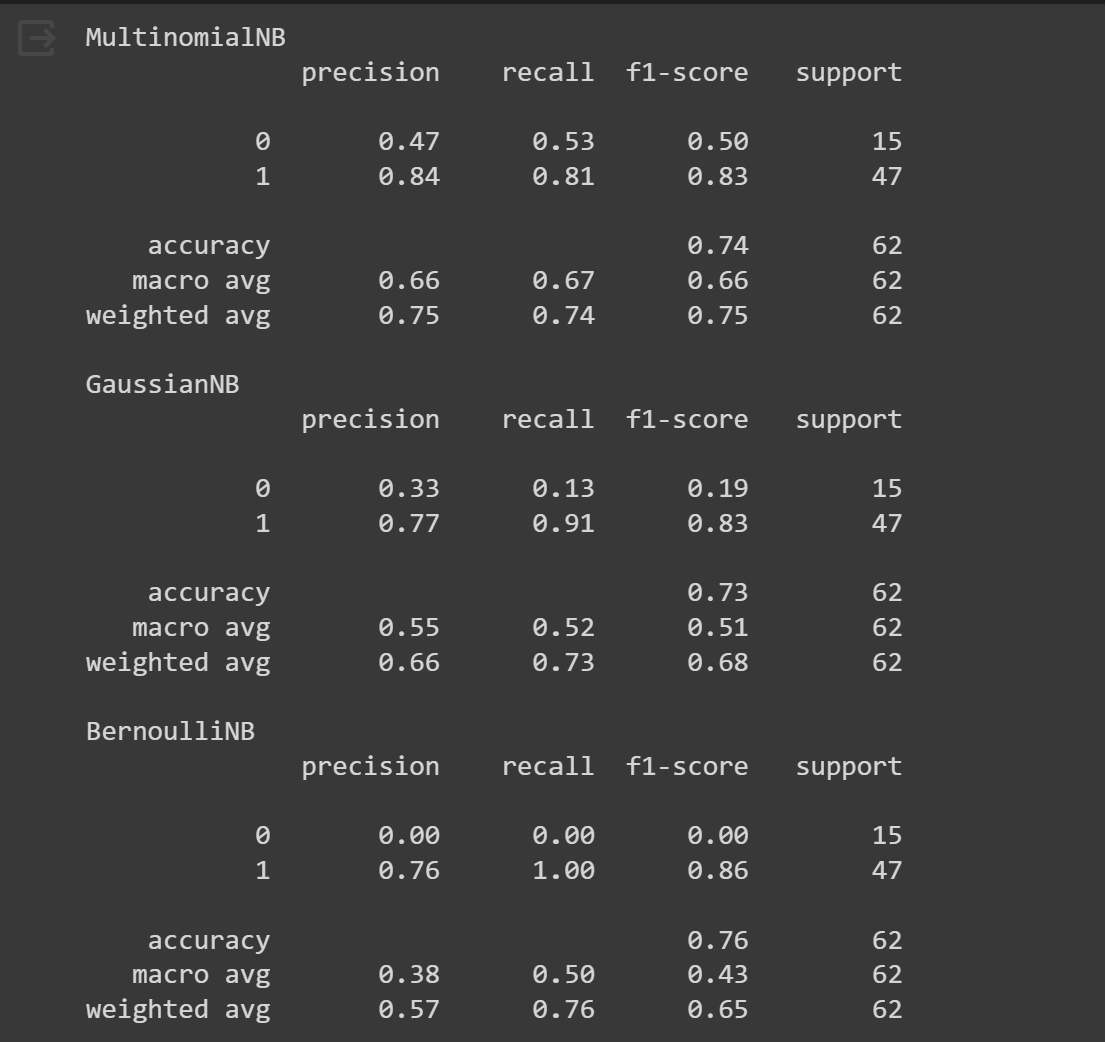
**Based on the data I observed, the number of nodes and age were the biggest determining factor in whether a patient lived or died. We could add additional features to this model to increase its accuracy.**

**Applying Various Supervised Machine Learning Model**

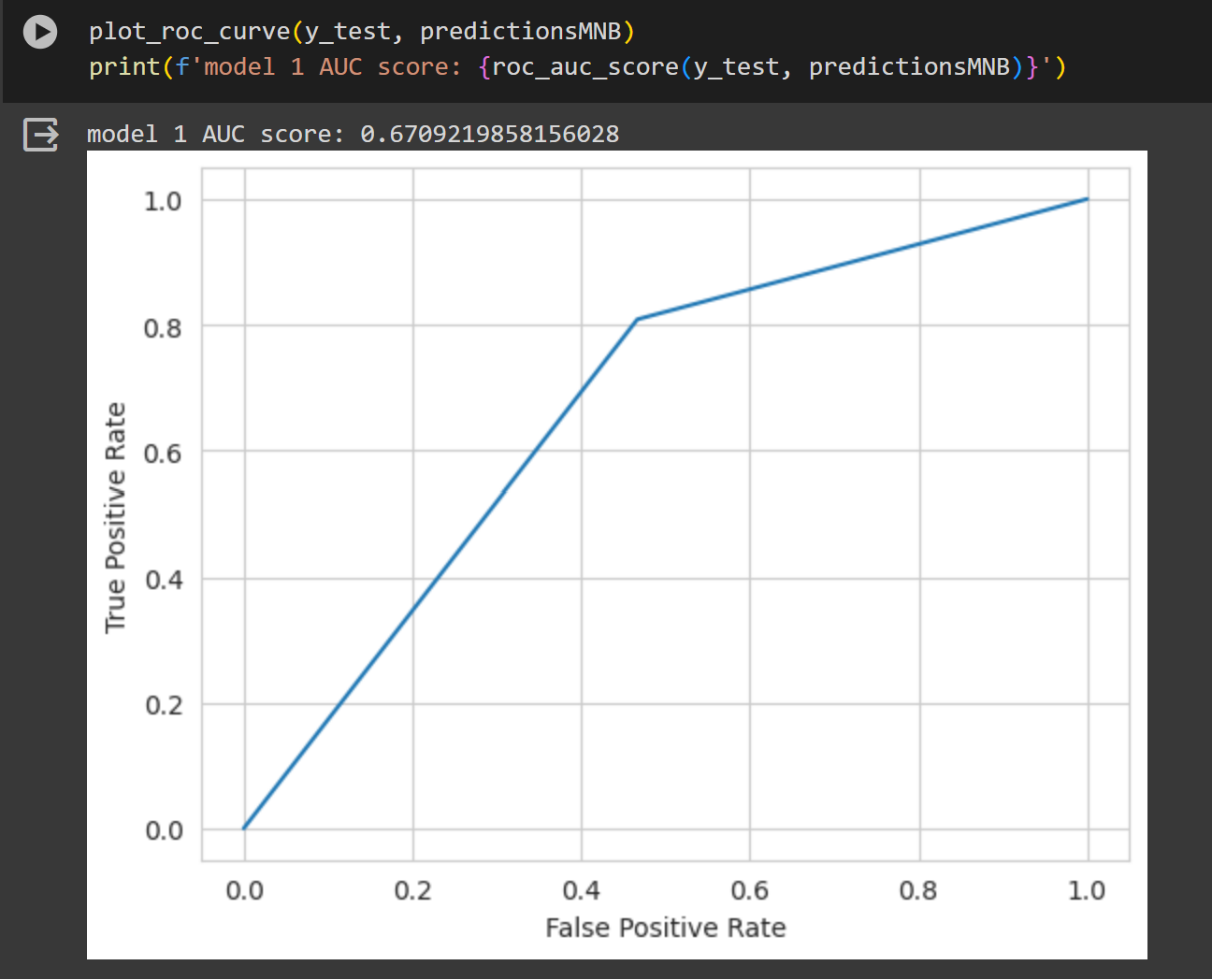
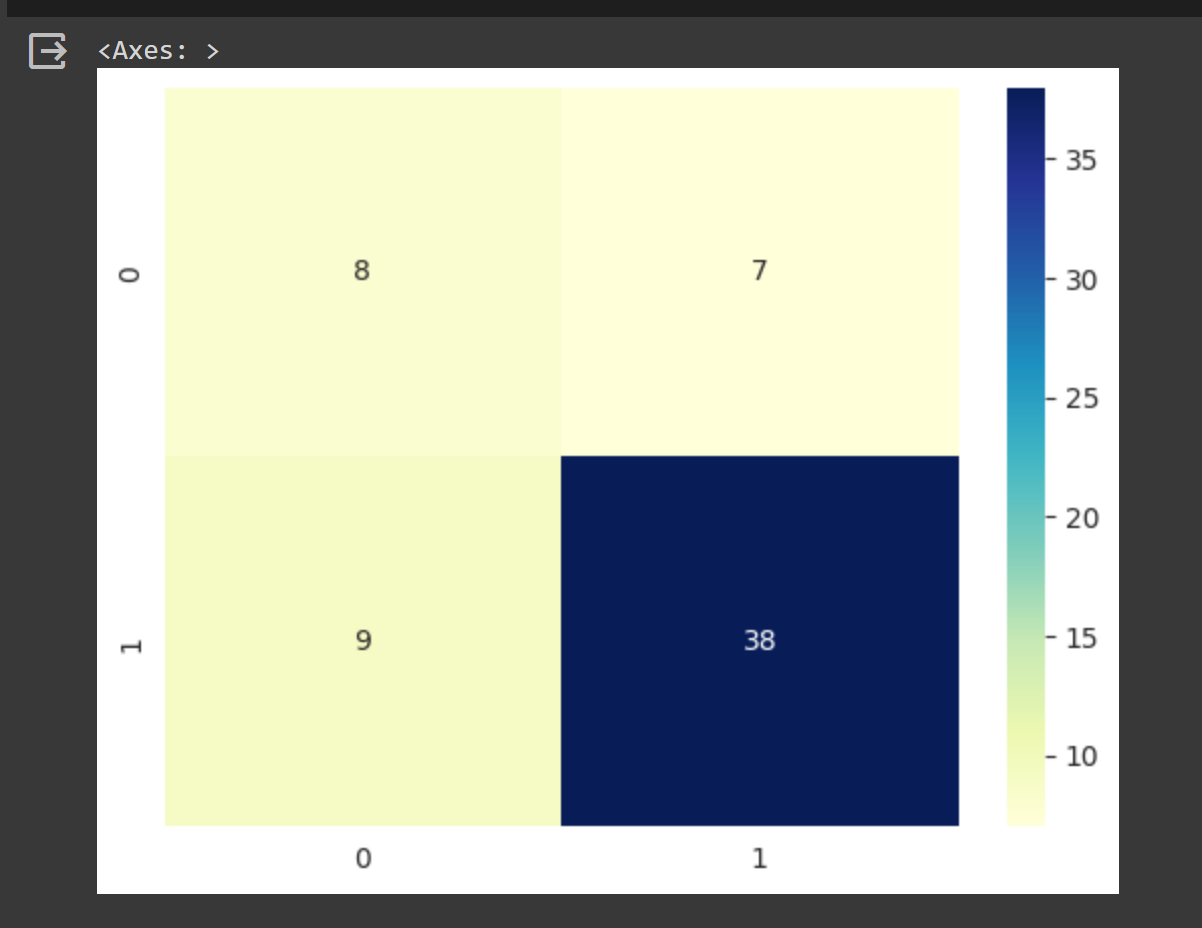
Naive Bayes

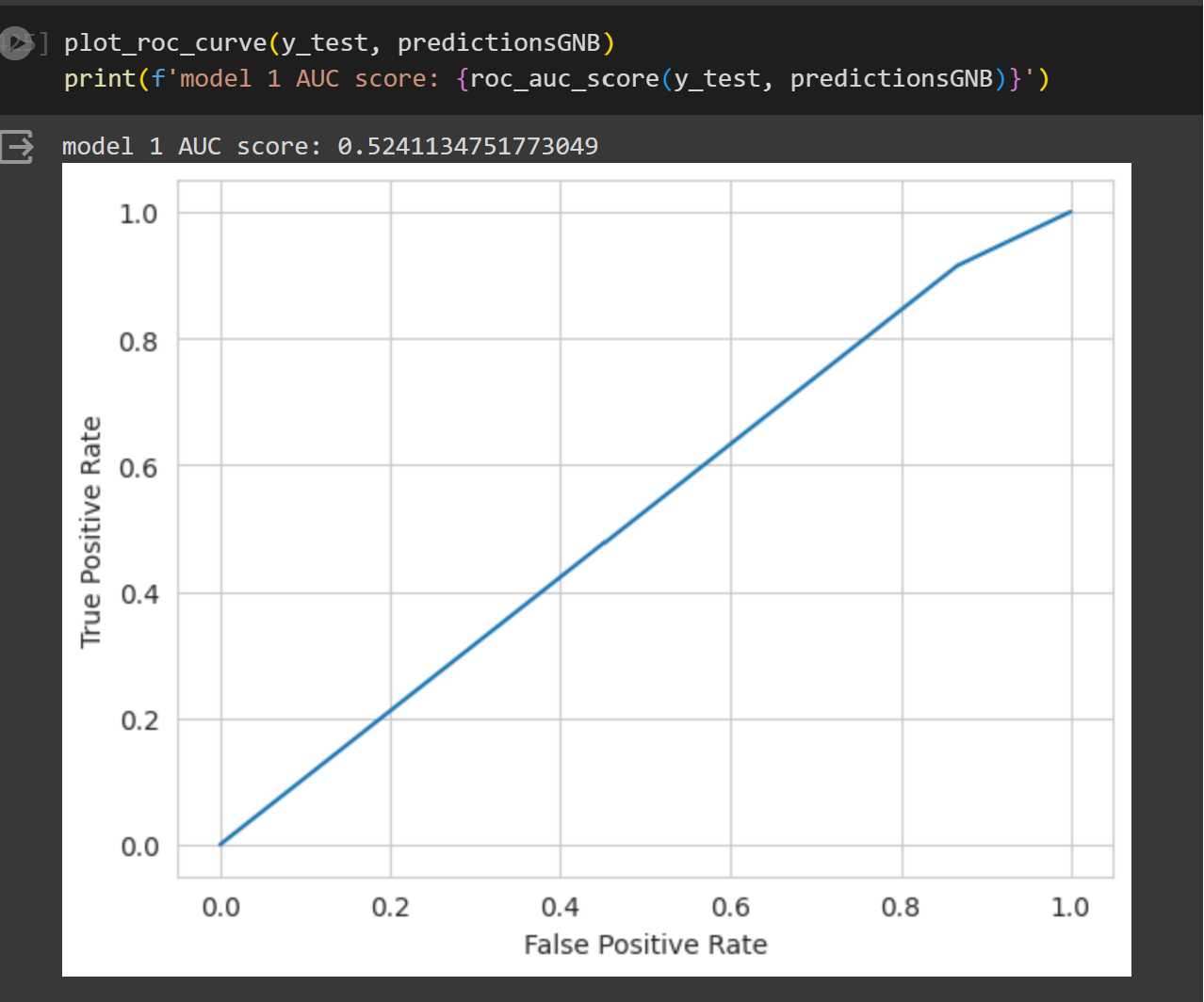
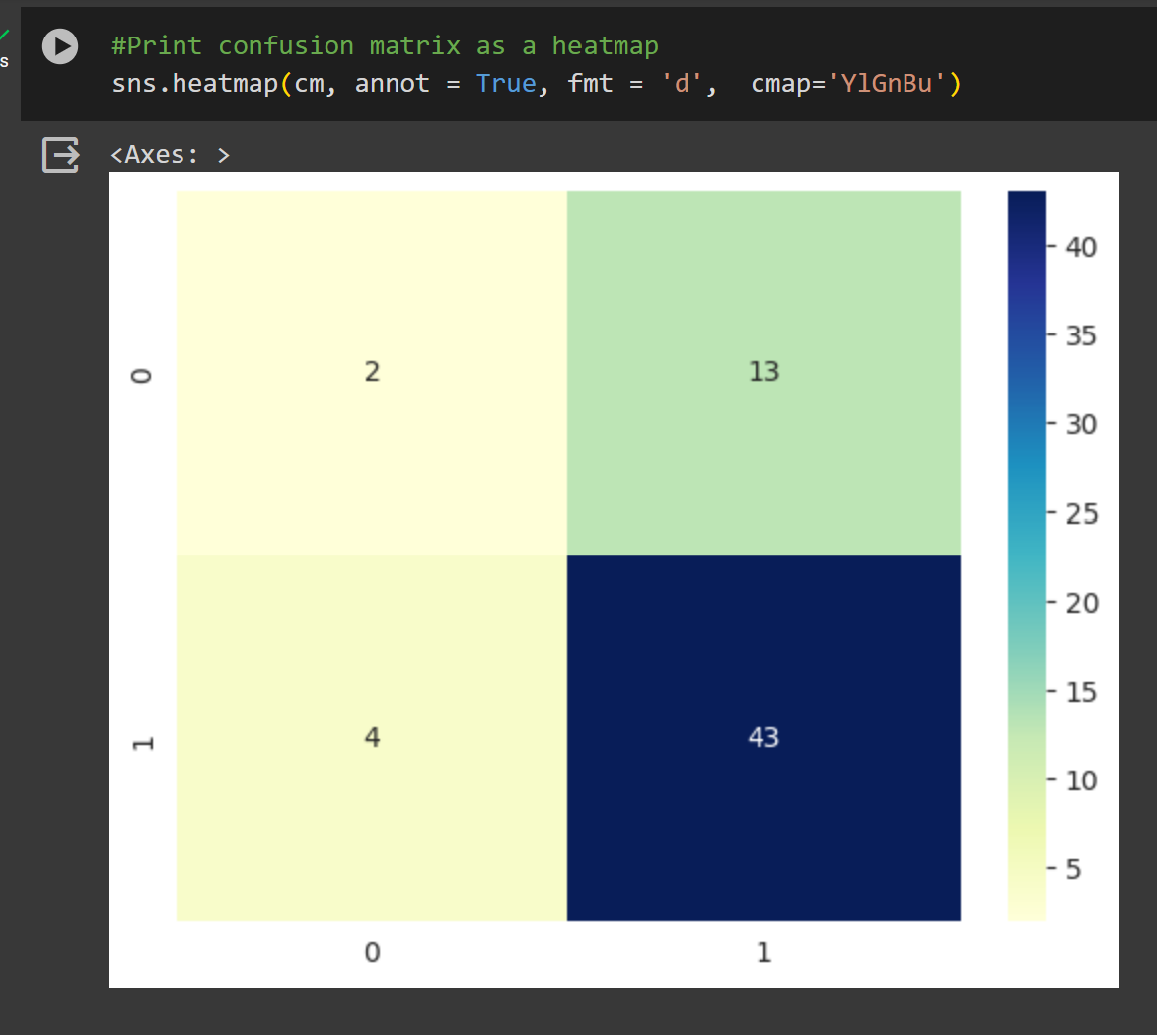
Applied below naive bayes algorithm to the data set

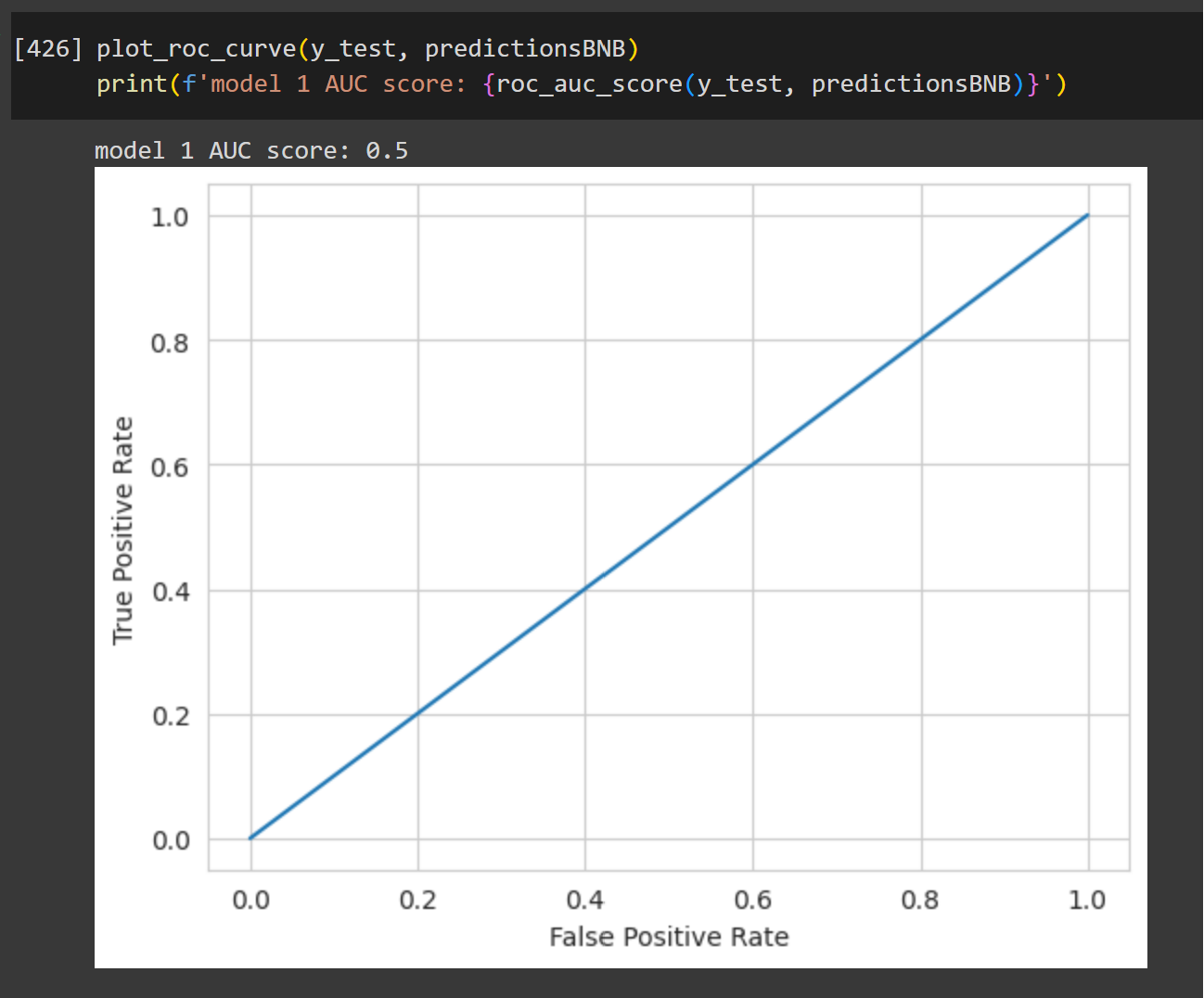
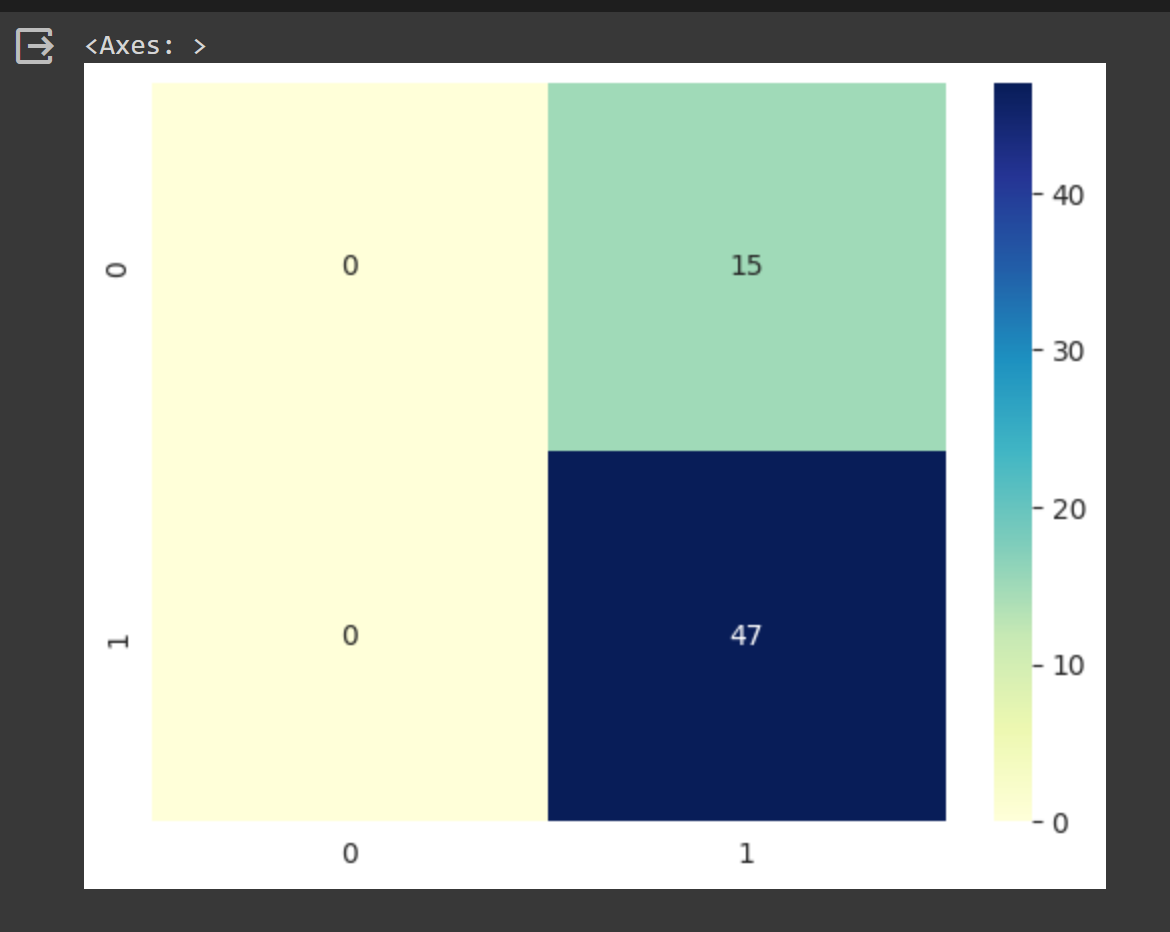
* Multinomial
* Gaussian
* Bernoulli

below is the snippet of performance metrics **

**AUC Curve Confusion matrix**

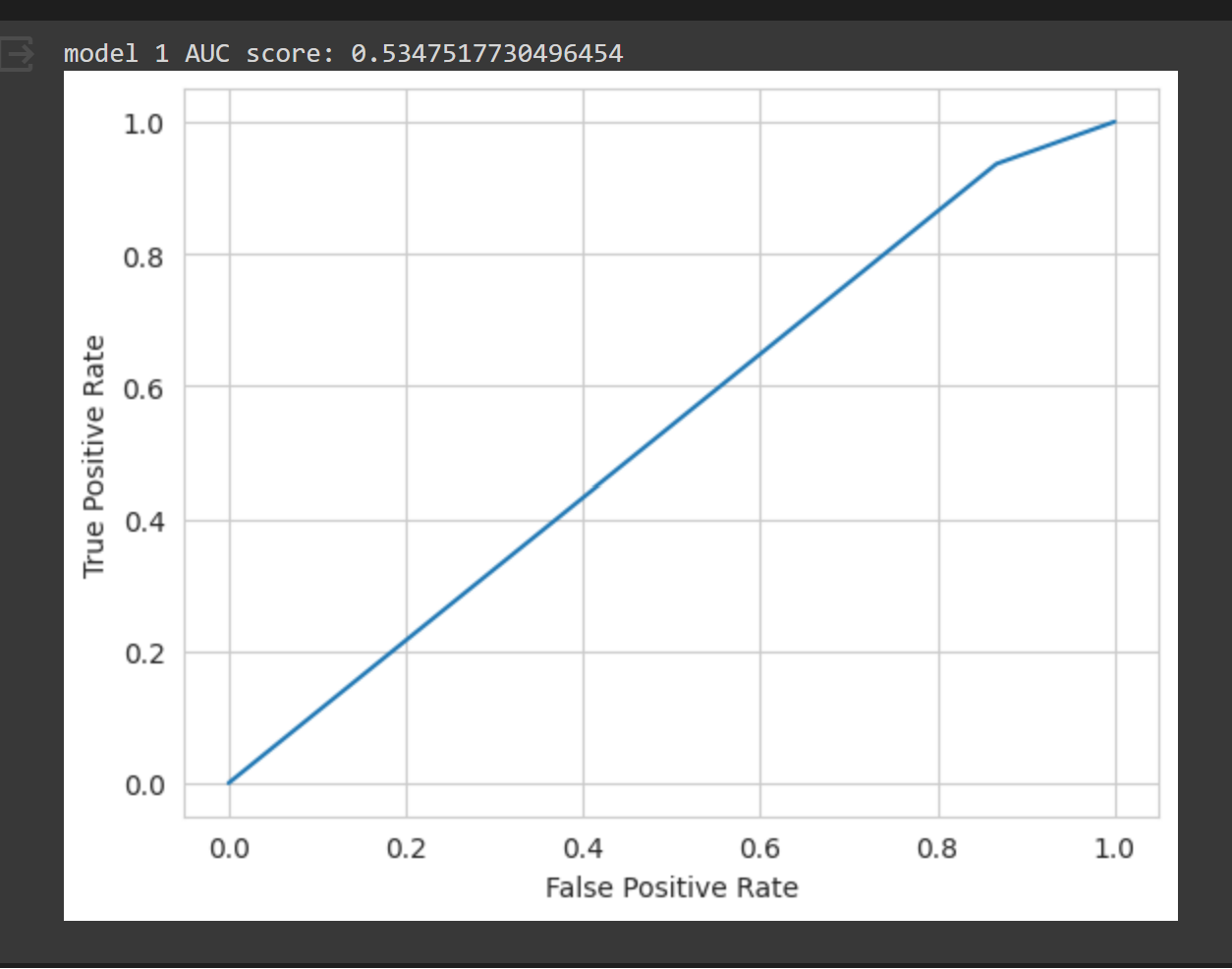
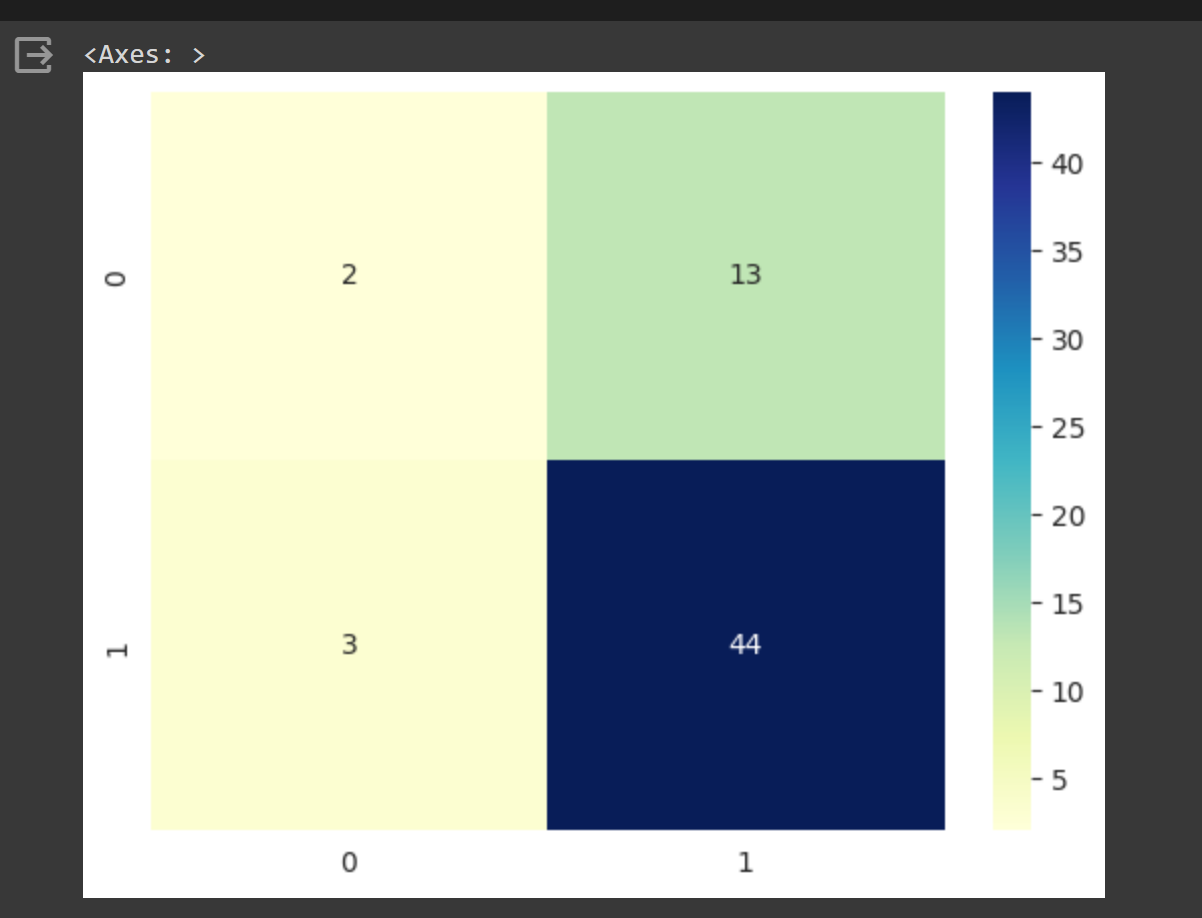
Since our objective is to identify persons who can survive after 5 years of surgery based on age and number of nodes, so that healthcare professional will give better treatment to increase their life expectancy.

* For this we need a model which has good accuracy along with high recall, also AUC score of >0.5 .
* By observing above performance metrics Multinomial Naïve Bayes predicting very well with good accuracy and recall.

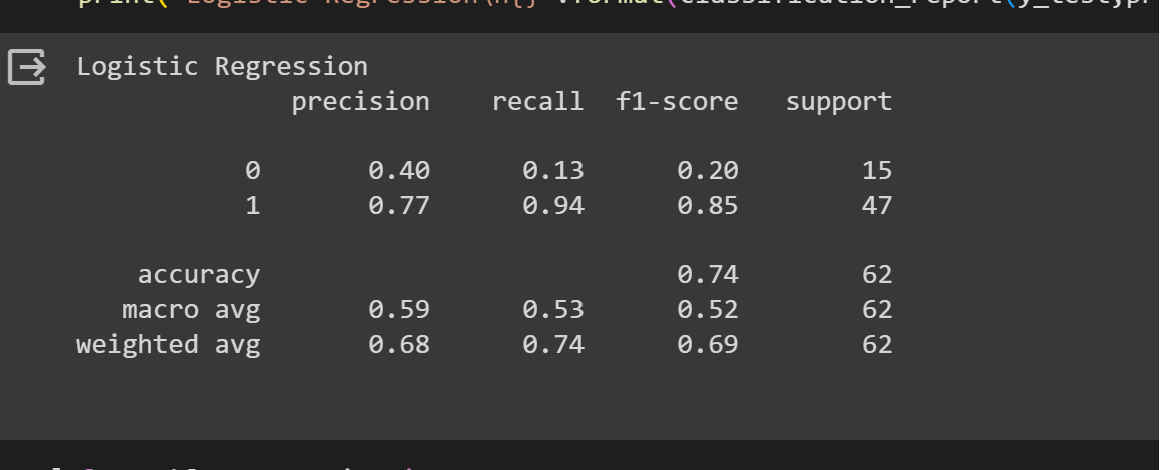
|  |  |  |  |
| --- | --- | --- | --- |
| Naive Bayes classifiers | Accuracy | Recall | AUC Score |
| Multinomial | **0.74** | **0.81** | **0.67** |
| Gaussian | 0.73 | 0.91 | 0.52 |
| Bernoulli | 0.76 | 1.00 | 0.50 |

Logistic regression

**Confusion matrix AUC Curve**



**Performance metrics**



|  |  |
| --- | --- |
| Logistic regression | Accuracy |
| Accuracy | **0.74** |
| Recall | 0.94 |
| AUC Score | 0.53 |