Capstone 3 Final Report

In this data set we sought out to learn how to better predict thyroid cancer in patients. Considering that this was a data set entirely made up of categorical data (except for patient age), we needed to approach the data set in a different way than if it were numerical.

This data set didn't require any cleaning because the data set that was gathered from Kaggle was already clean. So our next step was to move on to finding correlation within the data set. We first did one hot encoding to get dummies for our data set so all columns had a numerical value associated with them. Next we performed a t-test with the age column as it related to the other columns in the data set to see if age was correlated within our data set, which it was. Next we performed a chi squared test with our dependent variable (Recurred_Yes) with our other variables within our data set. The results of our chi squared test showed that several columns were correlated.

Next, we moved on to exploratory data analysis. In this step we chose to plot some of the key variables that we felt related heavily to the prediction of thyroid cancer. We also at this point decided on our null and alternative hypothesis. Ho: Recursion is only due to T, N, M, Adenopathy, and Age. Ha: Recursion is not only due to T, N, M, Adenopathy, and Age.

Preprocessing was performed earlier in the analysis where we used one hot encoding to get dummies for our data set. Next, we performed our modeling analysis for the data set. Here we chose to go with a Naïve Bayes, Support Vector Machine (SVM), and a Decision Tree analysis. We chose these models based on researching what modeling systems are used in predictive modeling analysis of cancer in patients.

For our first model we went with Naïve Bayes and plotted only the variables that were heavily correlated. In this approach we wanted to see if just using the heavily correlated variables would give a better model performance, as well as precision, recall, and AUC. In our first model we were able to calculate an a recall of 78.6%, precision of 88%, and AUC of 87%. Figure 1 you can see the confusion matrix from our model and then following that image you can see our ROC curve for the model as well. This model did perform well, but it was not our best performing model and in our next model we used every column that was correlated to our data set.

Within our next model we had a better performing model where we were able to calculate a recall of 78.6%, a precision of 91.6%, and a AUC of 88%. In Figure 3 our confusion matrix we predicted 22 true negative and 66 true positives. This model performed better than our last model so we chose to continue to use the same variables for the rest of our modeling.

For our next model, we chose to go with a Support Vector Machine(SVM). In researching which models were used the most in cancer prediction, support vector machine was that model. So we had high hopes for this models performance with our data set. In using SVM we were able to calculate a recall of 89.3% and a precision of 96.2%. So we captures more positive values within this model as we can see from our recall score. In Figure 5 the confusion matrix we predicted 25 true negative and 67 true positive. We also had an AUC of 94%.

For our final model we chose to go with a decision tree. In this model we were able to calculate a recall of 85.7%, a precision of 96%, and a AUC of 92%. In Figure 7 our confusion matrix, we were able to correctly predict 24 true negative and 67 true positive.

Finally, we chose to perform hyper parameter tuning on our decision tree to increase performance and results. With hyper parameter tuning we were able to achieve a recall score of

78.6%, precision of 100% and an AUC of 88%. Figure 9 our confusion matrix we were able to predict 22 true negative and 66 true positive.

From the above results we can see that SVM was our best performing model. This is great because it's a very simple and straight forward model to implement that doesn't require hyper parameter tuning.

Insights that can be gained from this data set is that we can now better predict thyroid cancer in patients due to this analysis. Also doctors can implement this machine learning model to help in predicting new cases of thyroid cancer. Health insurance providers can use this model to help clients if the patient has the chance of getting thyroid cancer and being able to tell the patient ahead of time to seek medical attention. Depending on the age of the patient, prescreening can be done to see if they are at risk of thyroid cancer.

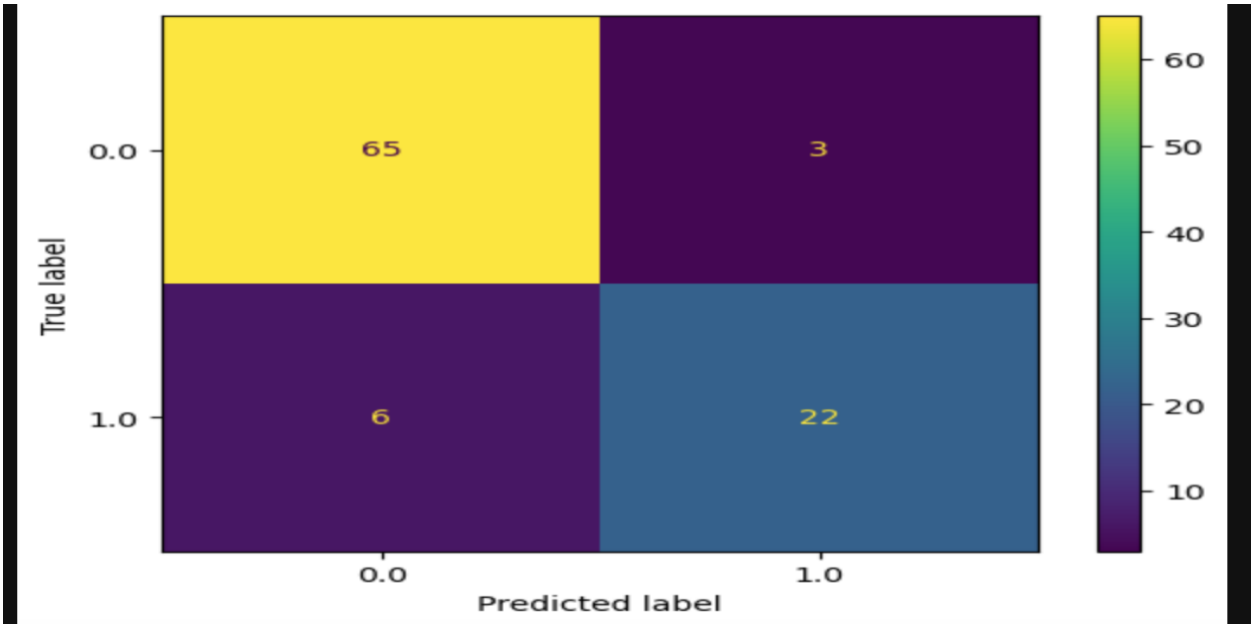

Figure 1 Confusion Matrix for Naive Bayes of Heavily Correlated Columns

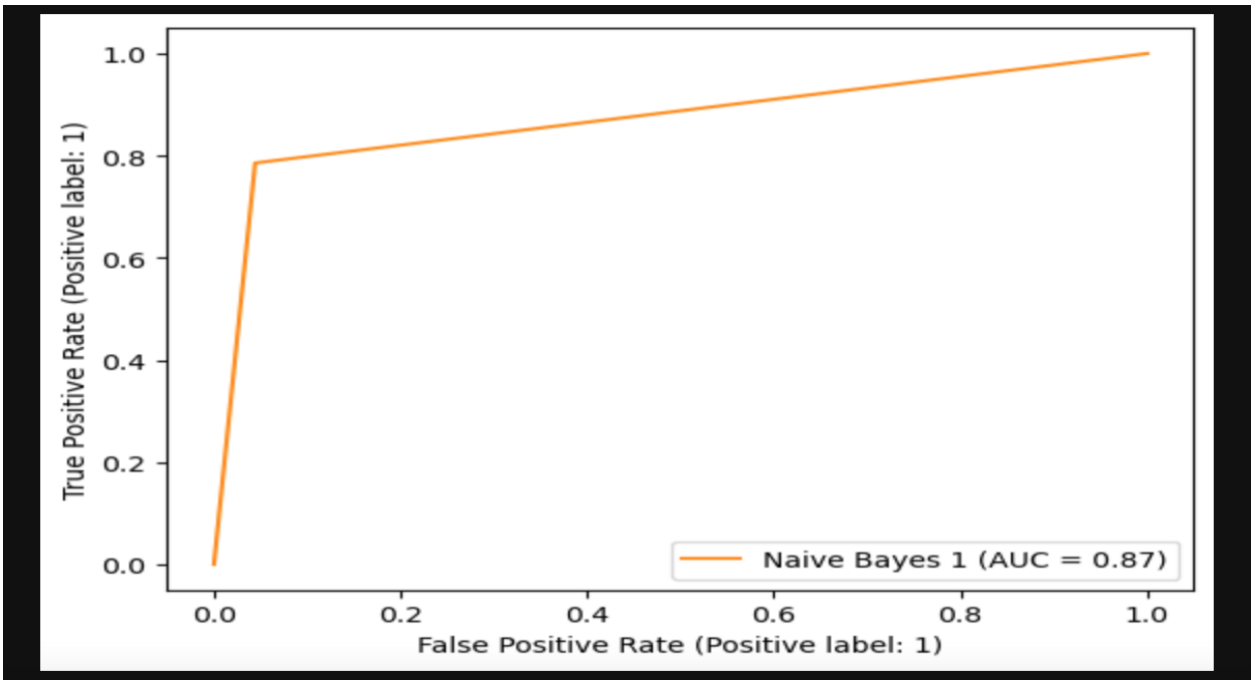

Figure 2 ROC Curve

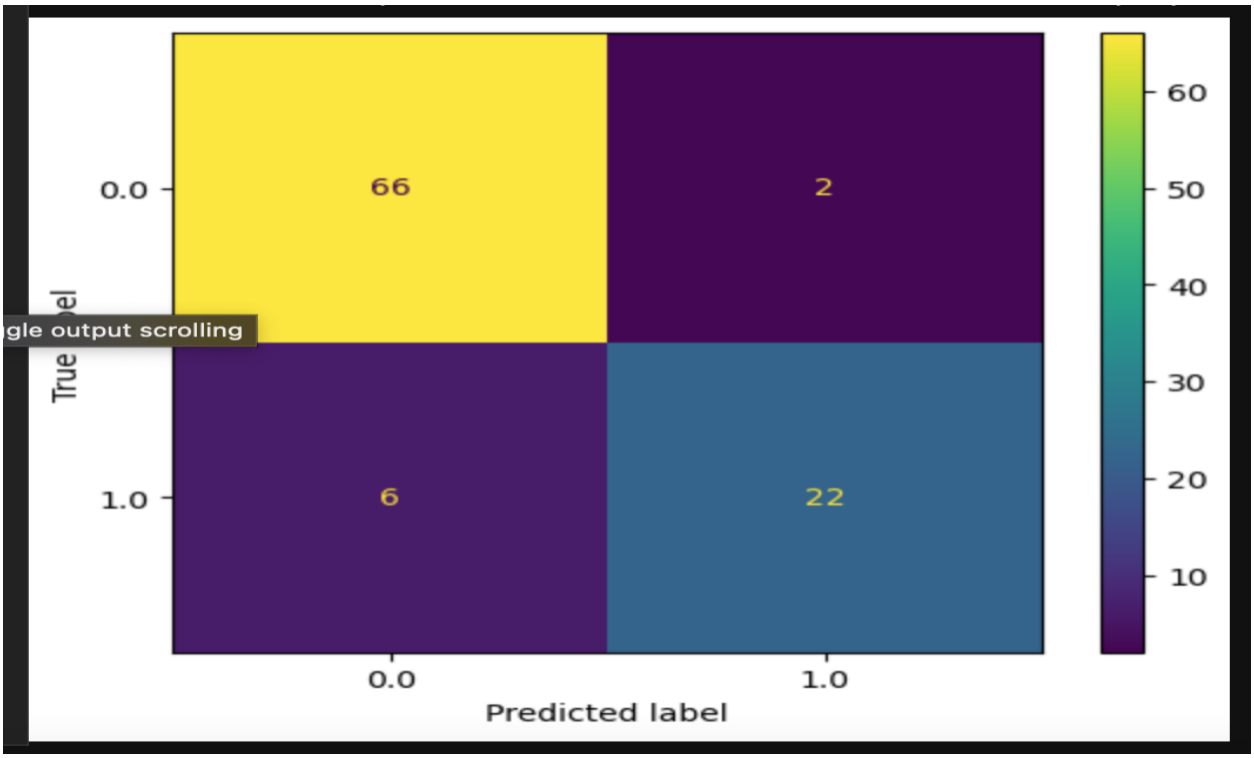

Figure 3 Naive Bayes of all correlate columns Confusion Matrix

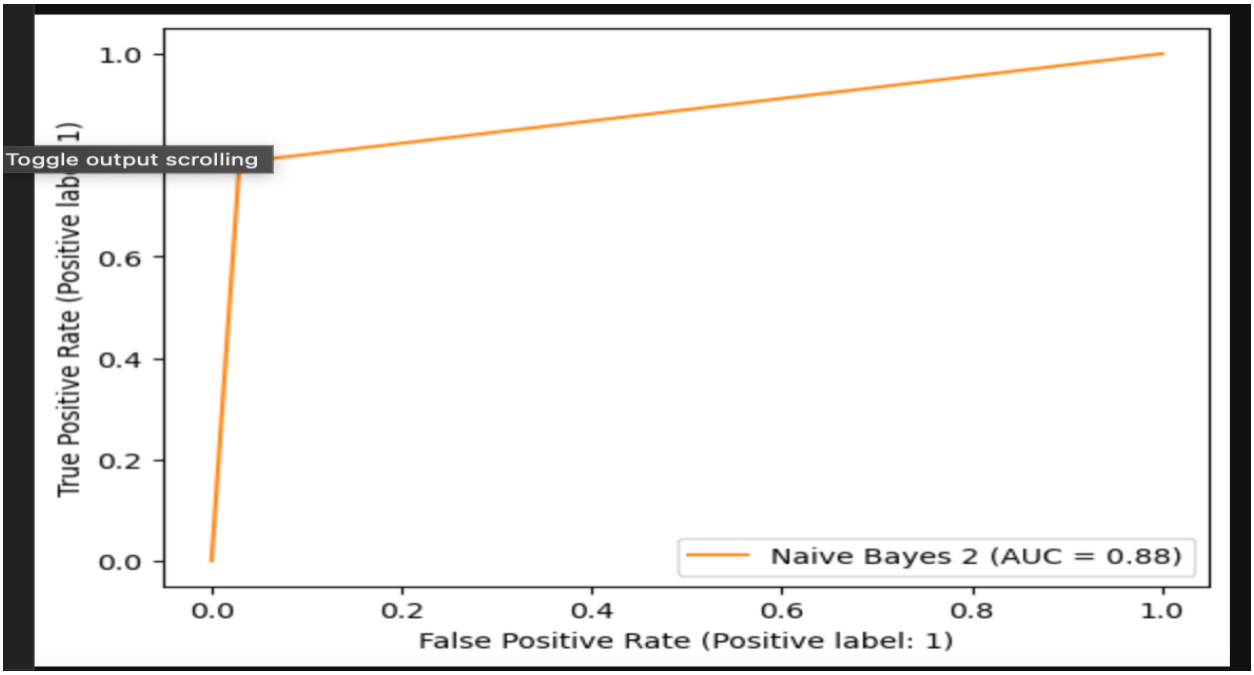

Figure 4 ROC Curve

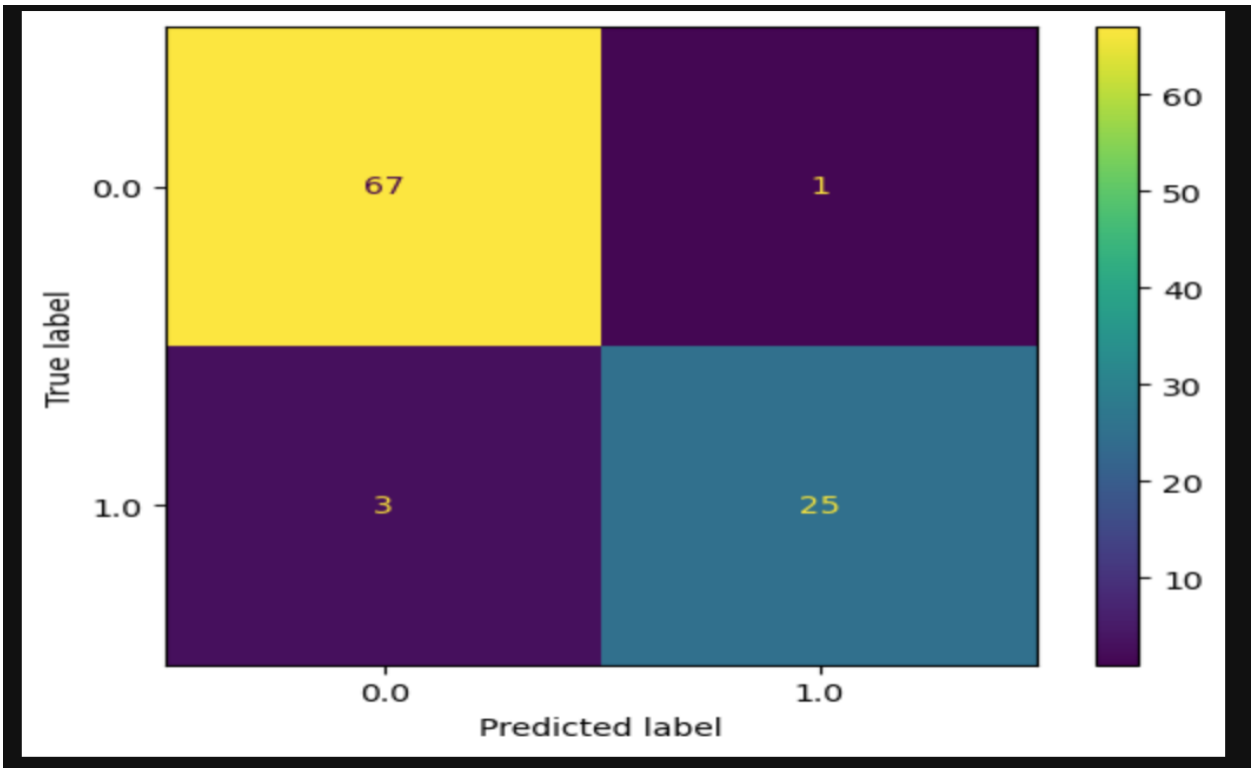

Figure 5 SVM Confusion Matrix

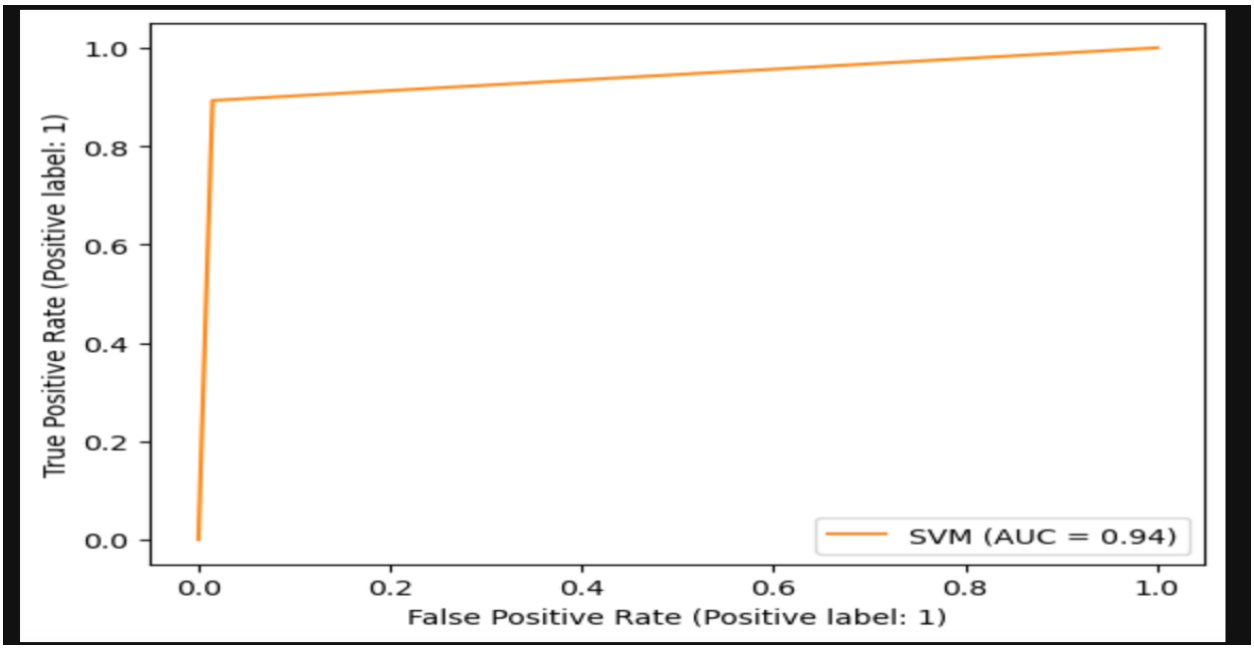

Figure 6 ROC Curve

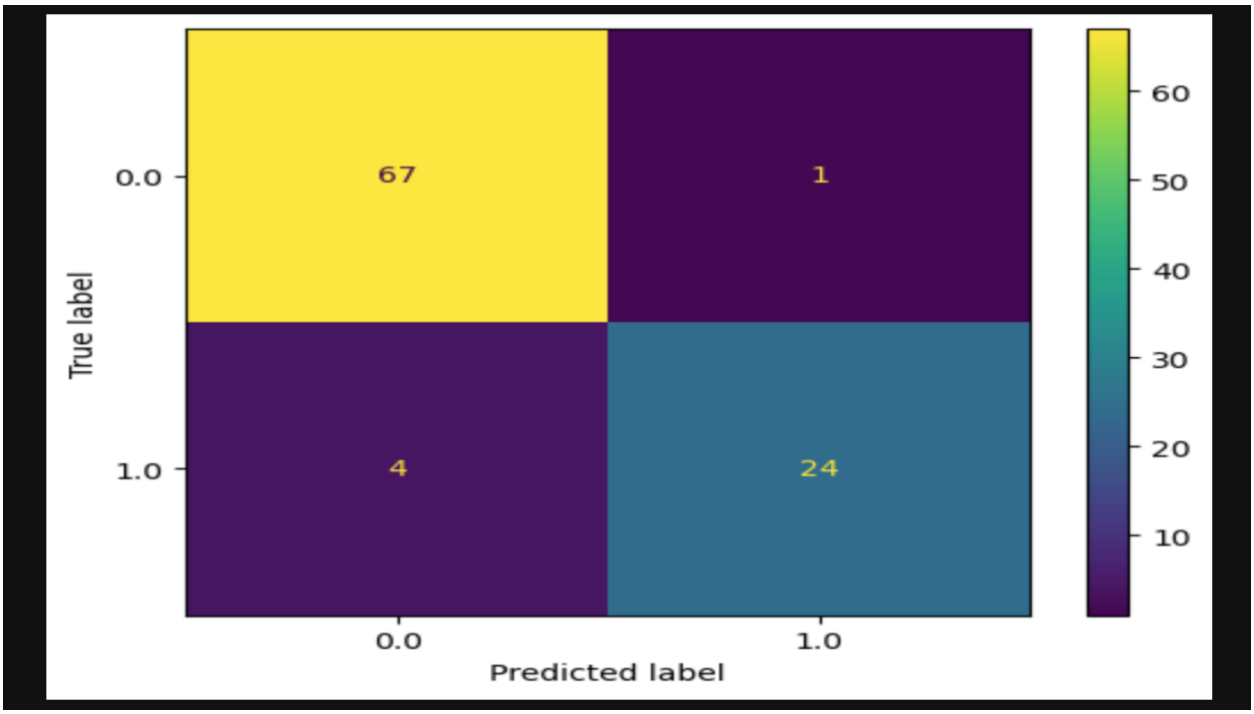

Figure 7 Decision Tree Confusion Matrix

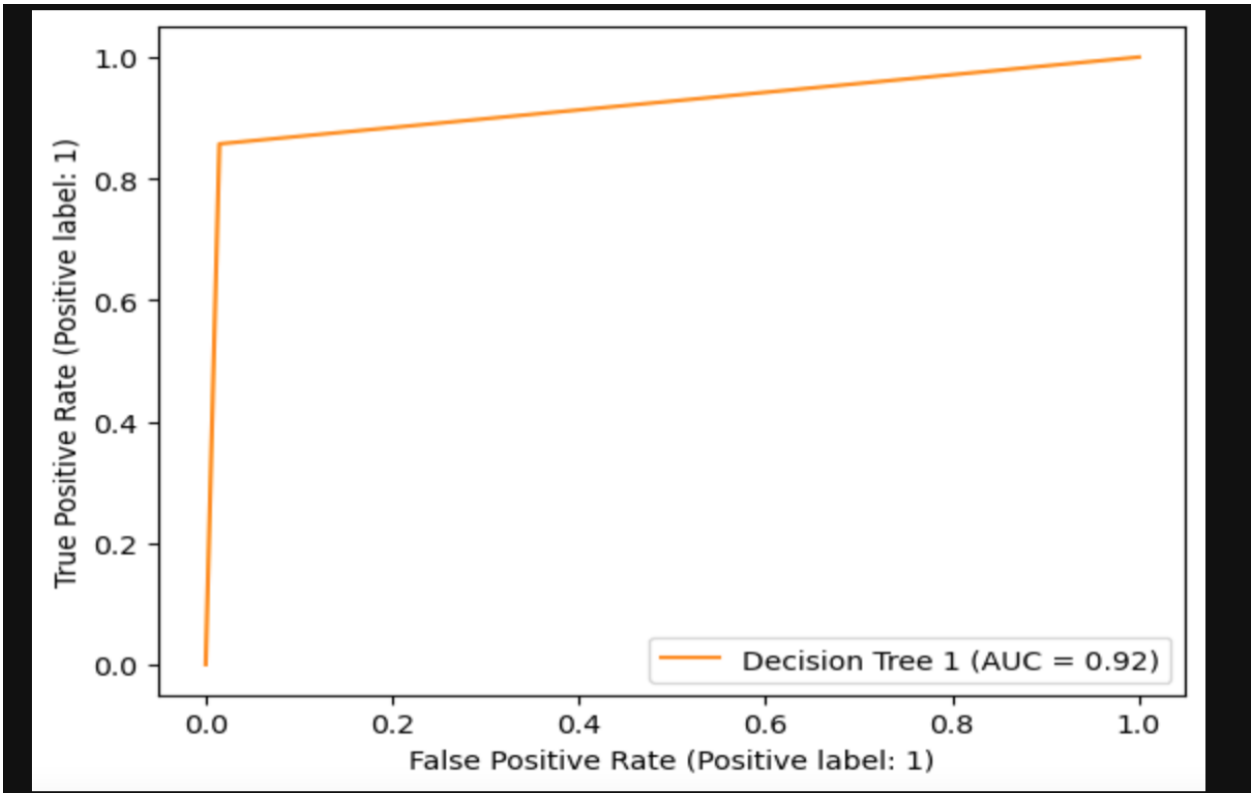

Figure 8 ROC Curve

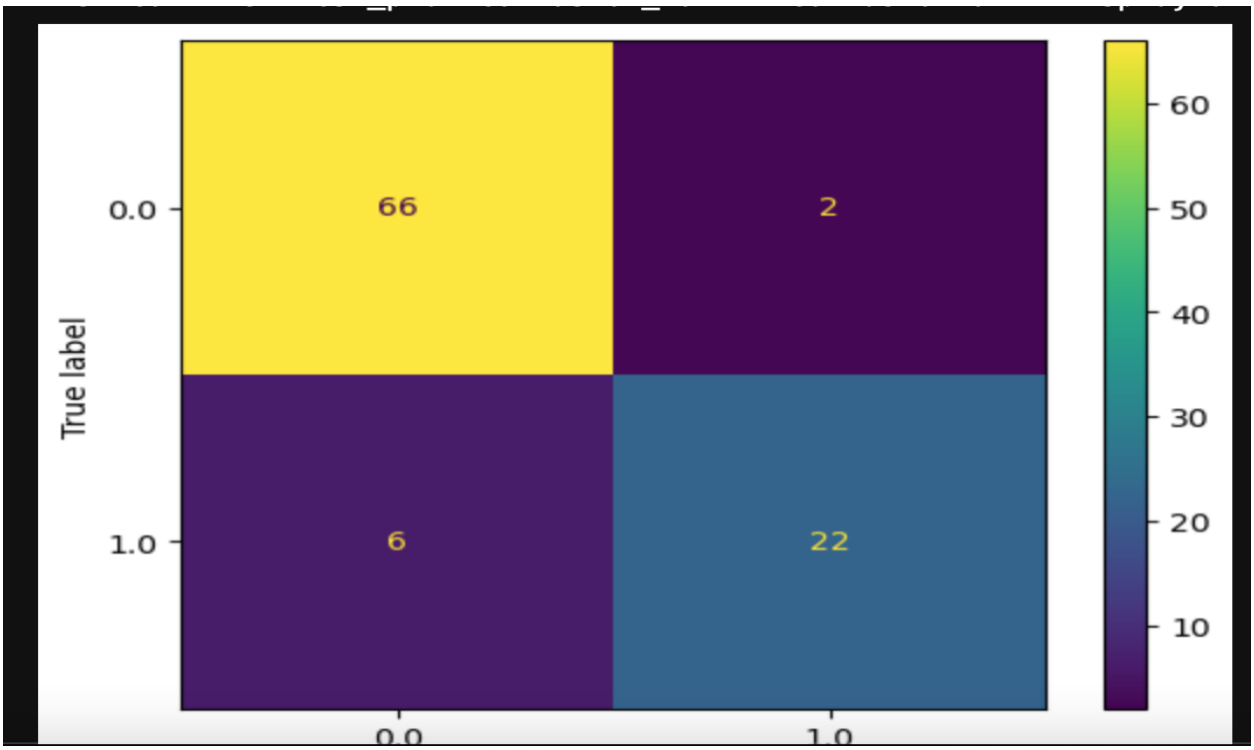

Figure 9 Hyper parameter tuning Confusion Matrix

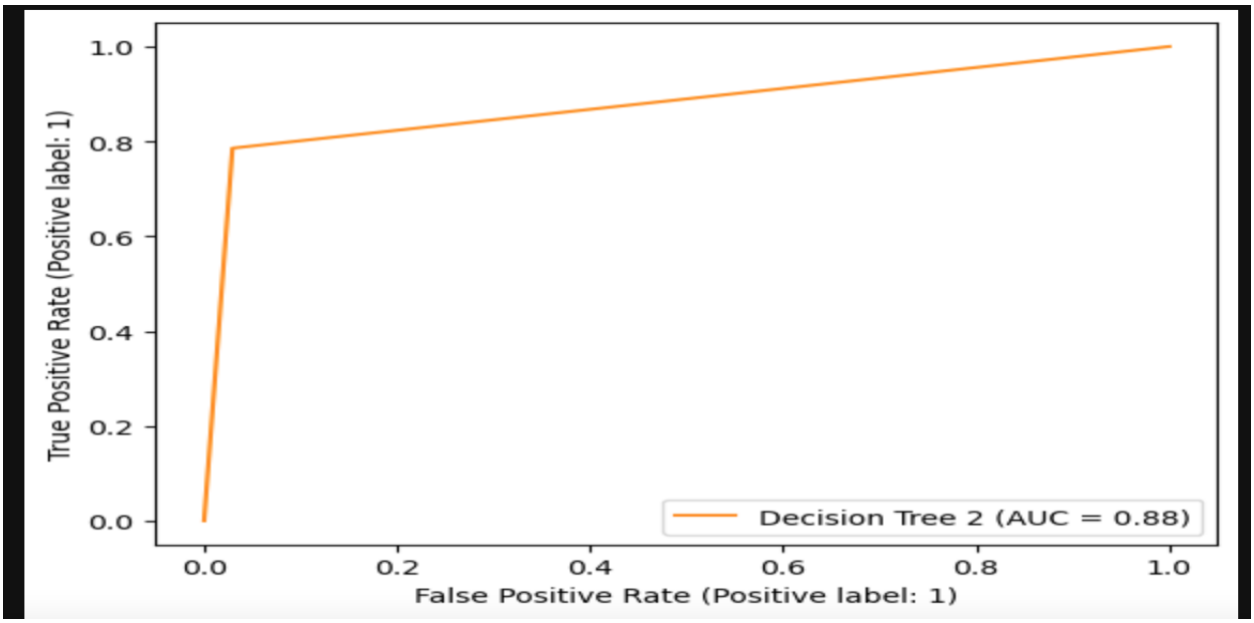

Figure 10 ROC Curve