This report will be analyzing the tumor sample dataset and identify which classification method is the most suitable at predicting whether the tumor is benign (class 0) or malignant (class 1).

Preprocessing

All datapoints in dataset are integers. As such no further processing was applied for categorical information. Likewise, missing values were not found within the dataset and no action was taken to handle them.

The dependent variable distribution was imbalanced (444 class 0 vs. 239 class 1) as such stratified method was used for test split to maintain the class distribution ratio of samples. The 7:3 train-test split was done for the original dataset. Then, Synthetic Minority Oversampling Technique (SMOTE) was applied for train samples.

Training Models

The train sample from the preprocessing was then used to train 5 different classification models.

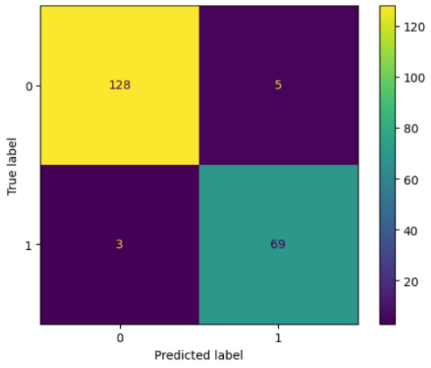
The logistic regression model was used as a base, then support vector machine, decision tree, and random forest models were trained using 5 fold cross validation with varying hyperparameters to find the optimal model. Then these models were used to predict the test sample from the 7:3 split for comparison.

Lastly, simple stacking classifier was trained using all 4 methods with logistic regression method as the final estimator.

Data

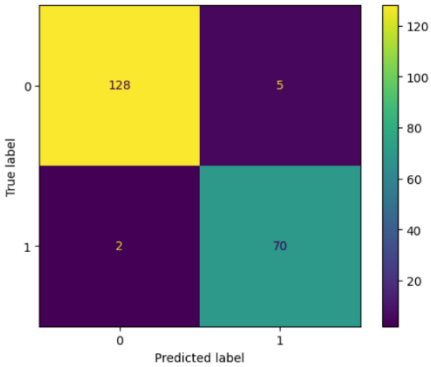
Logistic Regression

True Positive: 69
True Negative: 128
False Positive: 5
False Negative: 3
Accuracy is: 0.96
Precision is: 0.93
Recall is: 0.96
Fscore is: 0.96
AUC is: 0.96



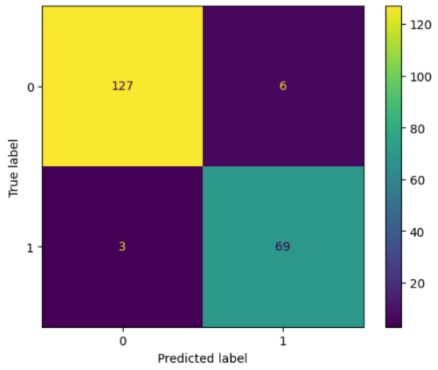
Support Vector Machine

True Positive: 70
True Negative: 128
False Positive: 5
False Negative: 2
Accuracy is: 0.97
Precision is: 0.93
Recall is: 0.97
Fscore is: 0.97
AUC is: 0.97



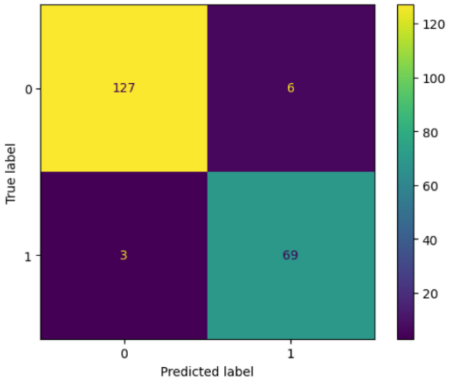
Decision Tree

True Positive: 69
True Negative: 127
False Positive: 6
False Negative: 3
Accuracy is: 0.96
Precision is: 0.92
Recall is: 0.96
Fscore is: 0.96
AUC is: 0.96



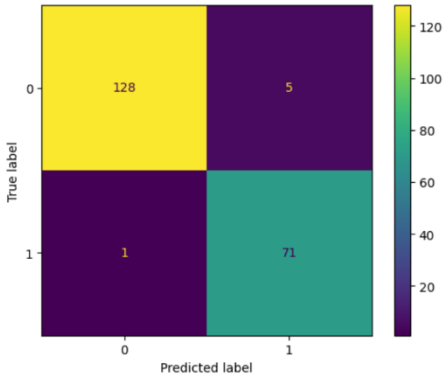
Random Forest

True Positive: 69
True Negative: 127
False Positive: 6
False Negative: 3
Accuracy is: 0.96
Precision is: 0.92
Recall is: 0.96
Fscore is: 0.96
AUC is: 0.96



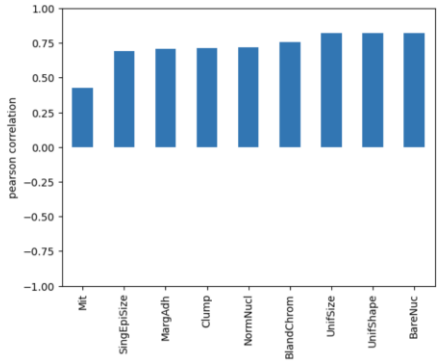
Stacking

True Positive: 71
True Negative: 128
False Positive: 5
False Negative: 1
Accuracy is: 0.97
Precision is: 0.93
Recall is: 0.99
Fscore is: 0.98
AUC is: 0.97



Findings

Overall, there seem to be very small differences between the model performances. All models performed with < .95 fscore, which is very high. However, the best predictor seems to be the stacking model with the highest fscore and accuracy followed by the support vector machine. As such stacking method seems to be the best suited for predicting the class of tumor based on these features.



The high performance across all models as well as correlation between features and dependent variable may suggest that the data may have been too simplistic in nature. It is also possible that the model may be too skewed for this specific dataset as the size of sample was fairly small (~ 700). It is also possible that there may be features that were not captured for this dataset as well as more granular layer to existing features, which could cause underfitted model comparatively to the real life use cases. To generate more accurate model, increasing the size as well as the features would be recommended.