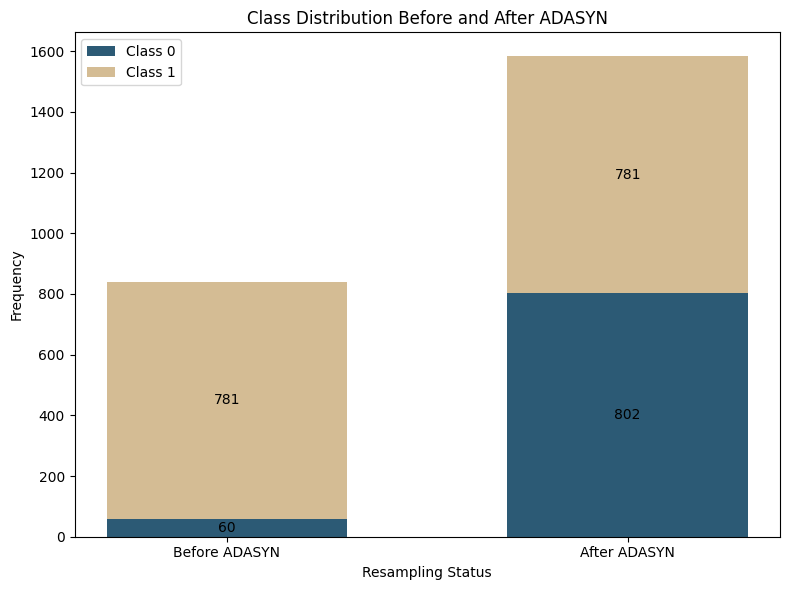
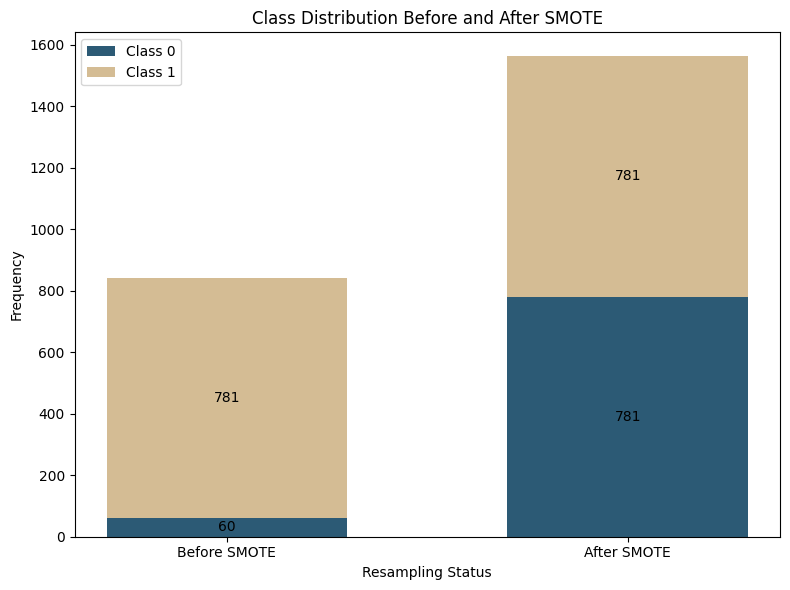
# Class Imbalance graph

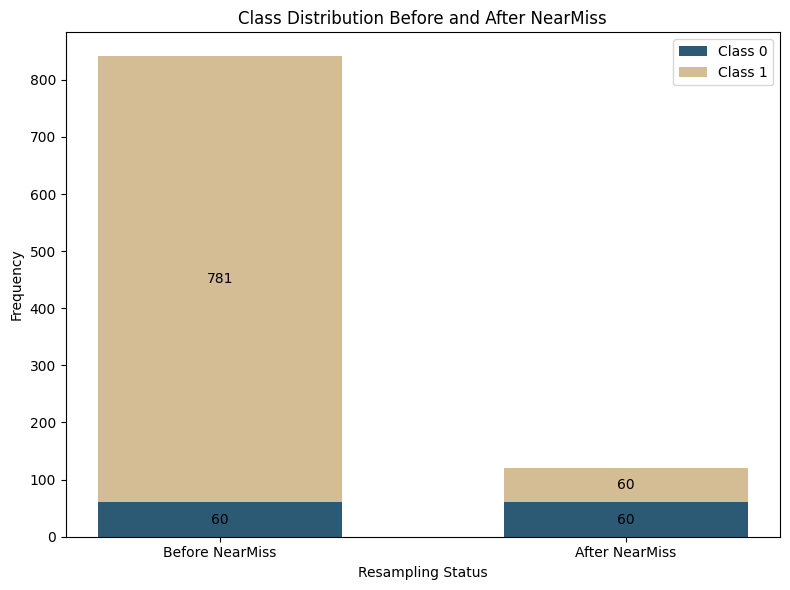
* ADASYN



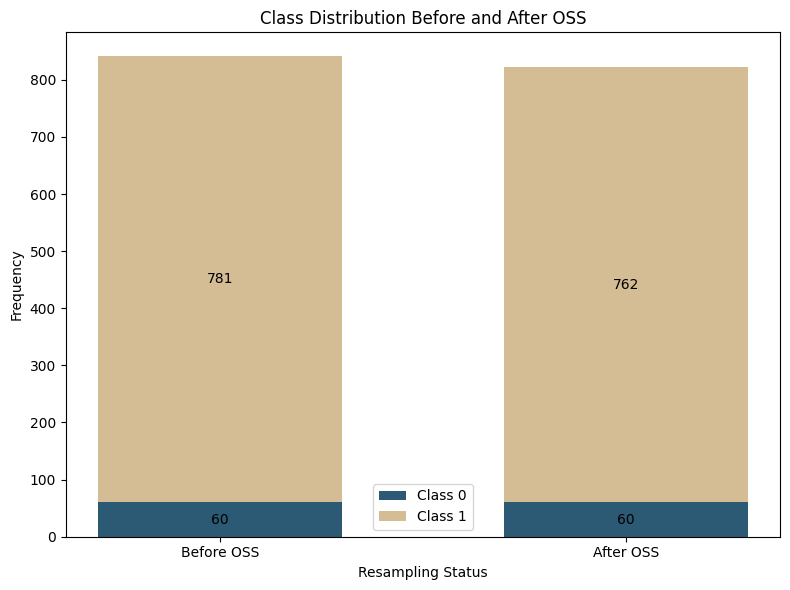
* SMOTE



* Near Miss (make train data smaller than test data (120 vs 400)

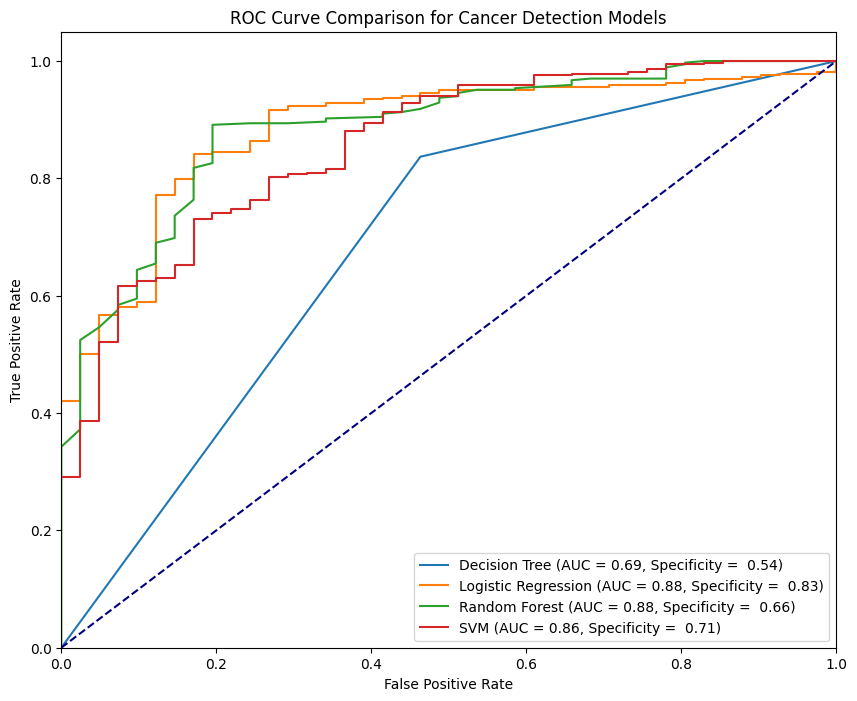
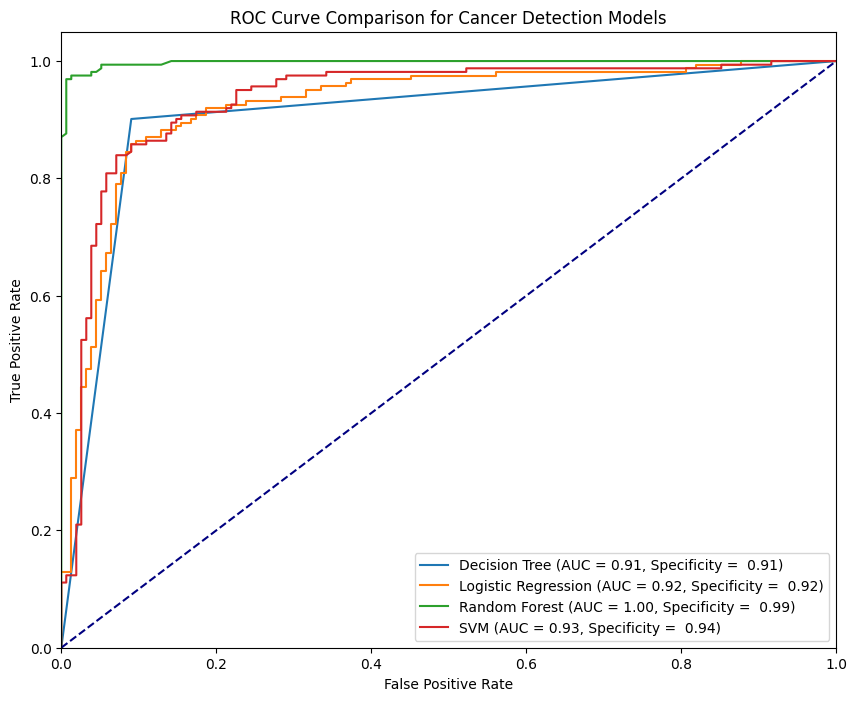


* OSS (failed at it because it is still imbalanced. This is due to OSS nature)

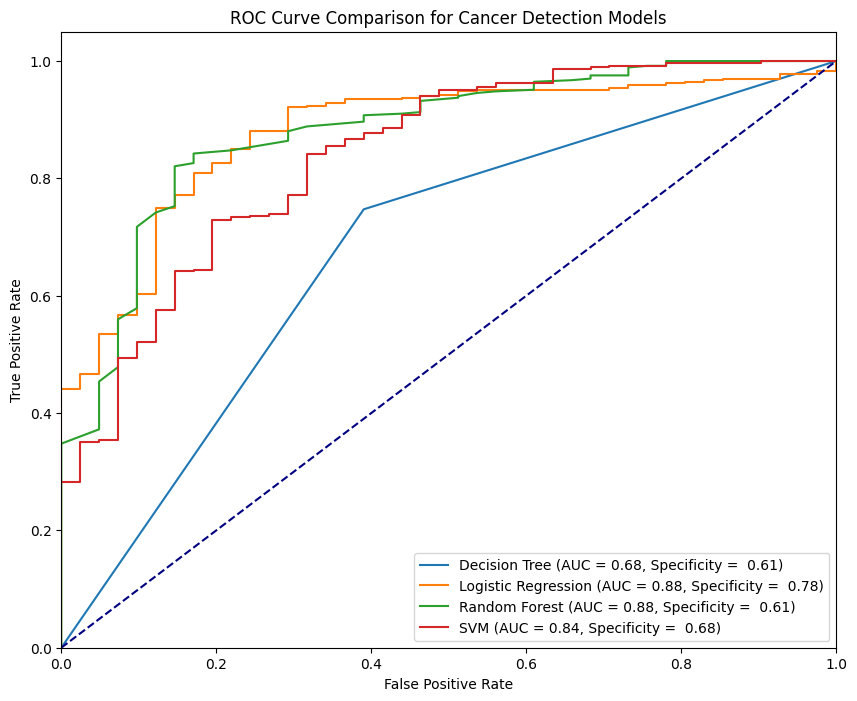
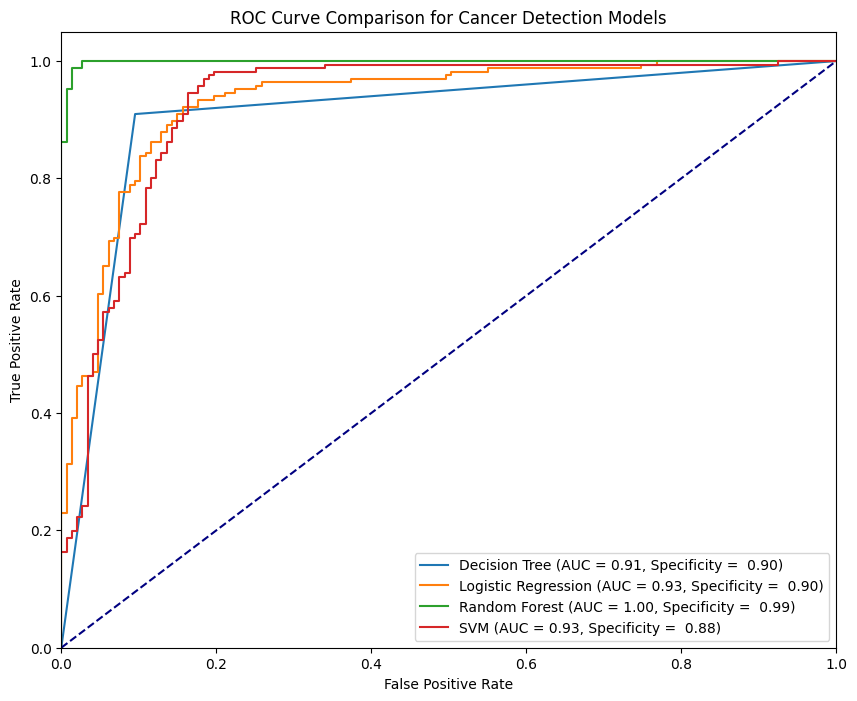


# Baseline

* ADASYN (left: validation, right: test)



* SMOTE (left: validation, right: test)



Sign of overfitting: test against validation data is high but when against the test data is quite low (especially random forest).

# Hyperparameters tuning

* ADASYN

1. DT

-> best parameter: {'dt\_\_class\_weight': 'balanced', 'dt\_\_criterion': 'entropy', 'dt\_\_max\_depth': 10, 'dt\_\_max\_features': None, 'dt\_\_min\_samples\_leaf': 1, 'dt\_\_min\_samples\_split': 10}

-> specificity against validation data using the best parameter: 0.68

1. LR

-> best parameter:{'lr\_\_C': 1, 'lr\_\_penalty': 'l1', 'lr\_\_solver': 'liblinear'}

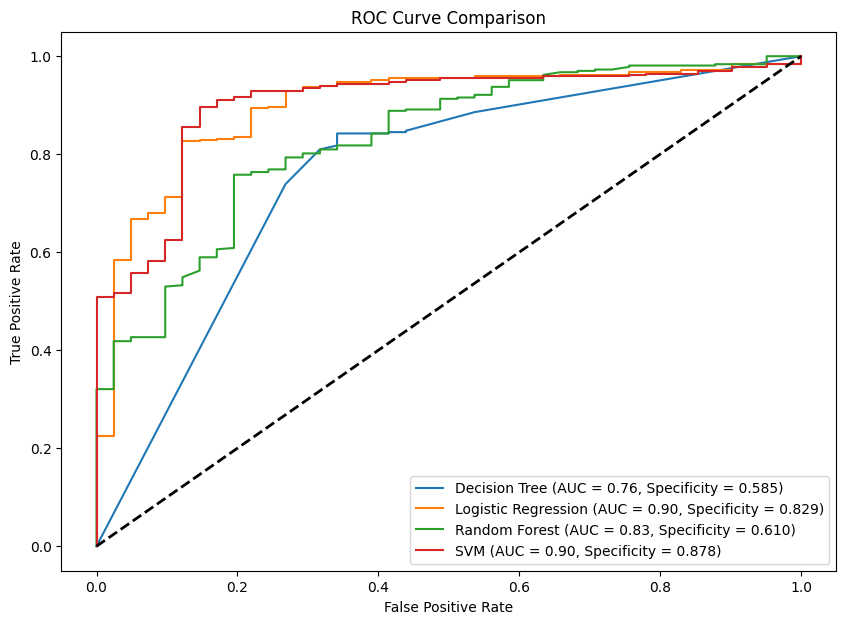
-> specificity against validation data using the best parameter: 0.92

1. RF

-> best parameter: {'rf\_\_bootstrap': True, 'rf\_\_max\_depth': 2, 'rf\_\_max\_features': 0.75, 'rf\_\_min\_samples\_leaf': 5, 'rf\_\_min\_samples\_split': 15, 'rf\_\_n\_estimators': 150}

-> specificity against validation data using the best parameter: 0.78

Test against the test data using the best parameters:



Best: SVM (0.88)

* SMOTE

1. DT

-> best parameter: {'dt\_\_class\_weight': 'balanced', 'dt\_\_criterion': 'gini', 'dt\_\_max\_depth': 20, 'dt\_\_max\_features': None, 'dt\_\_min\_samples\_leaf': 5, 'dt\_\_min\_samples\_split': 10}

-> specificity against validation data using the best parameter: 0.67

1. LR

-> best parameter: {'lr\_\_C': 1, 'lr\_\_penalty': 'l1', 'lr\_\_solver': 'liblinear'}

-> specificity against validation data using the best parameter: 0.88

1. RF

-> best parameter: {'rf\_\_bootstrap': True, 'rf\_\_max\_depth': 2, 'rf\_\_max\_features': 0.75, 'rf\_\_min\_samples\_leaf': 5, 'rf\_\_min\_samples\_split': 15, 'rf\_\_n\_estimators': 150}

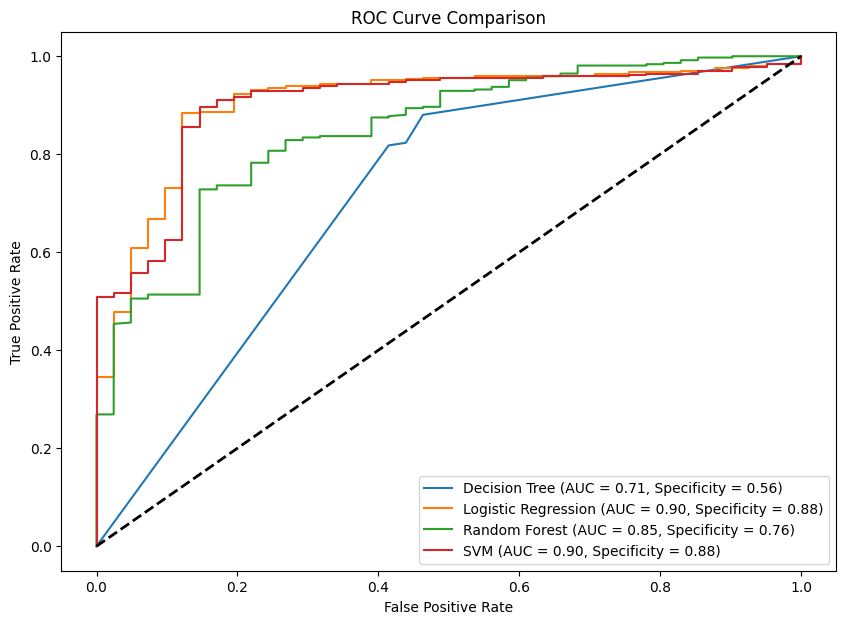
-> specificity against validation data using the best parameter: 0.77

1. SVM

-> best parameter:{'C': 1, 'kernel': 'linear'}

-> specificity against validation data using the best parameter: 0.85

Test against the test data using the best parameters:



Best: Logistic Regression + SVM (0.878)