Introduction:
This analysis explores chocolate and magic mushroom consumption using data from UCI's 2018 drug consumption database. The original multi-class categorization for both substances was converted into binary classification. We built six supervised learning models for each dataset: Single Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), Gradient Boosting (GB), Multi-Layer Perceptron (MLP), and K-Nearest Neighbors (k-NN).

The datasets differ in class balance—chocolate consumption data is highly skewed toward users, while the mushroom dataset is more balanced. To address class imbalance, we applied three rebalancing techniques: undersampling with Tomek Links, oversampling with SMOTE, and a hybrid approach combining both. Each model was trained using the holdout method, with 67% of the data for training and 33% for testing, and evaluated using confusion matrices, precision, recall, ROC curves, and area under the curve (AUC).

Algorithm Behavior and Results:
The chocolate dataset models achieved high precision and recall due to the skewed data, which resulted in overfitting. Users were frequently classified as recent consumers, inflating false positives. Random Forest, SVM, Decision Tree, and Gradient Boosting achieved perfect recall, but their AUC values hovered around 0.5, indicating limited ability to distinguish between users and non-users. The SVM model achieved the highest AUC of 0.64, though this still reflected poor generalization.

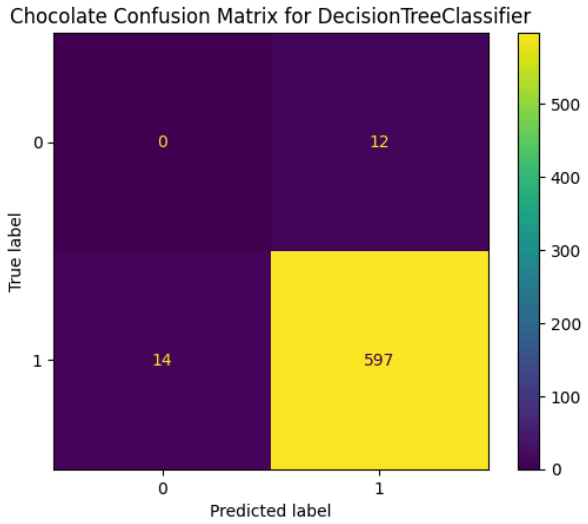
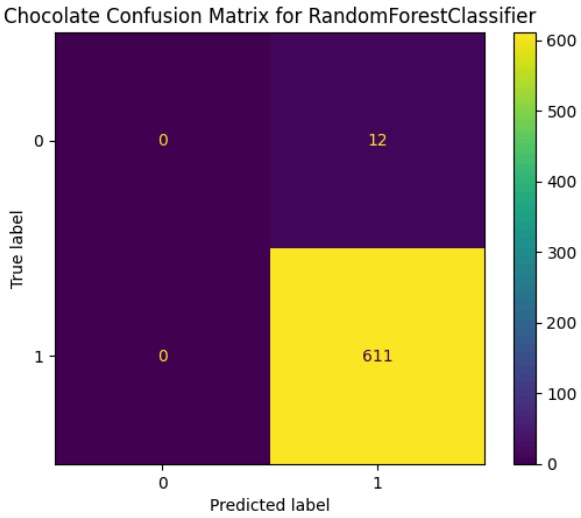
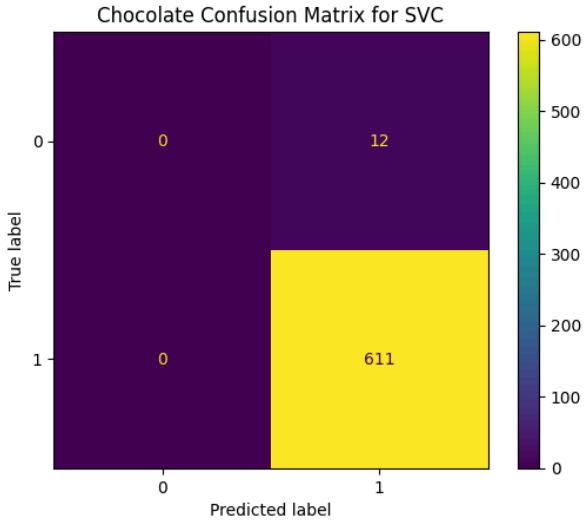
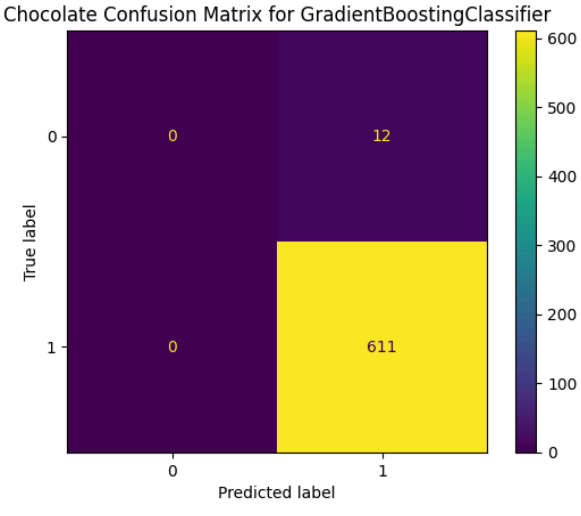
The mushroom dataset's balanced distribution enabled better generalization, with fewer false positives. While precision and recall scores were lower than those for chocolate, most models achieved AUC values between 0.67 and 0.80. SVM performed particularly well, achieving an AUC of 0.8, precision of 0.65, and recall of 0.66, demonstrating its effectiveness. KNN struggled with nuanced patterns, showing lower precision (0.581) and recall (0.593), suggesting challenges in capturing user behavior accurately.

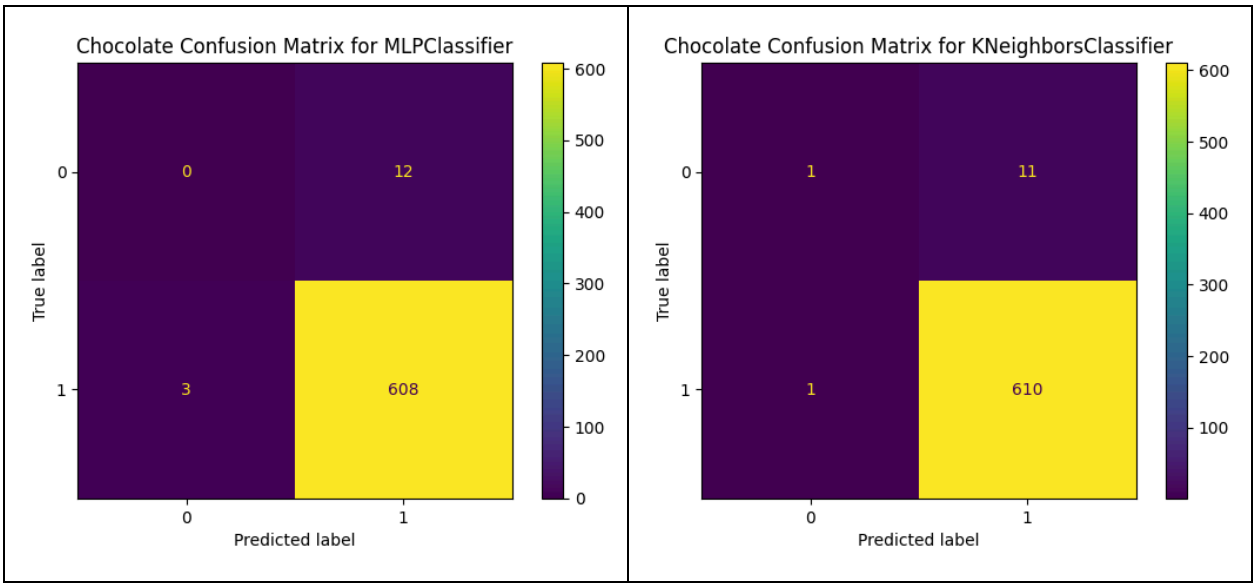
Impact of Rebalancing Techniques:
Rebalancing techniques had the most impact on the chocolate dataset due to its severe imbalance. Tomek Links removed only a few majority class points, resulting in minimal improvements. Precision and recall remained high across models, with SVM seeing a slight AUC increase of 0.02.

Oversampling with SMOTE was more effective, generating synthetic samples for the minority class and reducing overfitting. Models like KNN and SVM saw improved performance, with KNN achieving the highest AUC of 0.64 despite lower recall. The hybrid approach combining Tomek Links and SMOTE was particularly effective for Decision Trees and KNN, maintaining high recall while minimizing false positives.

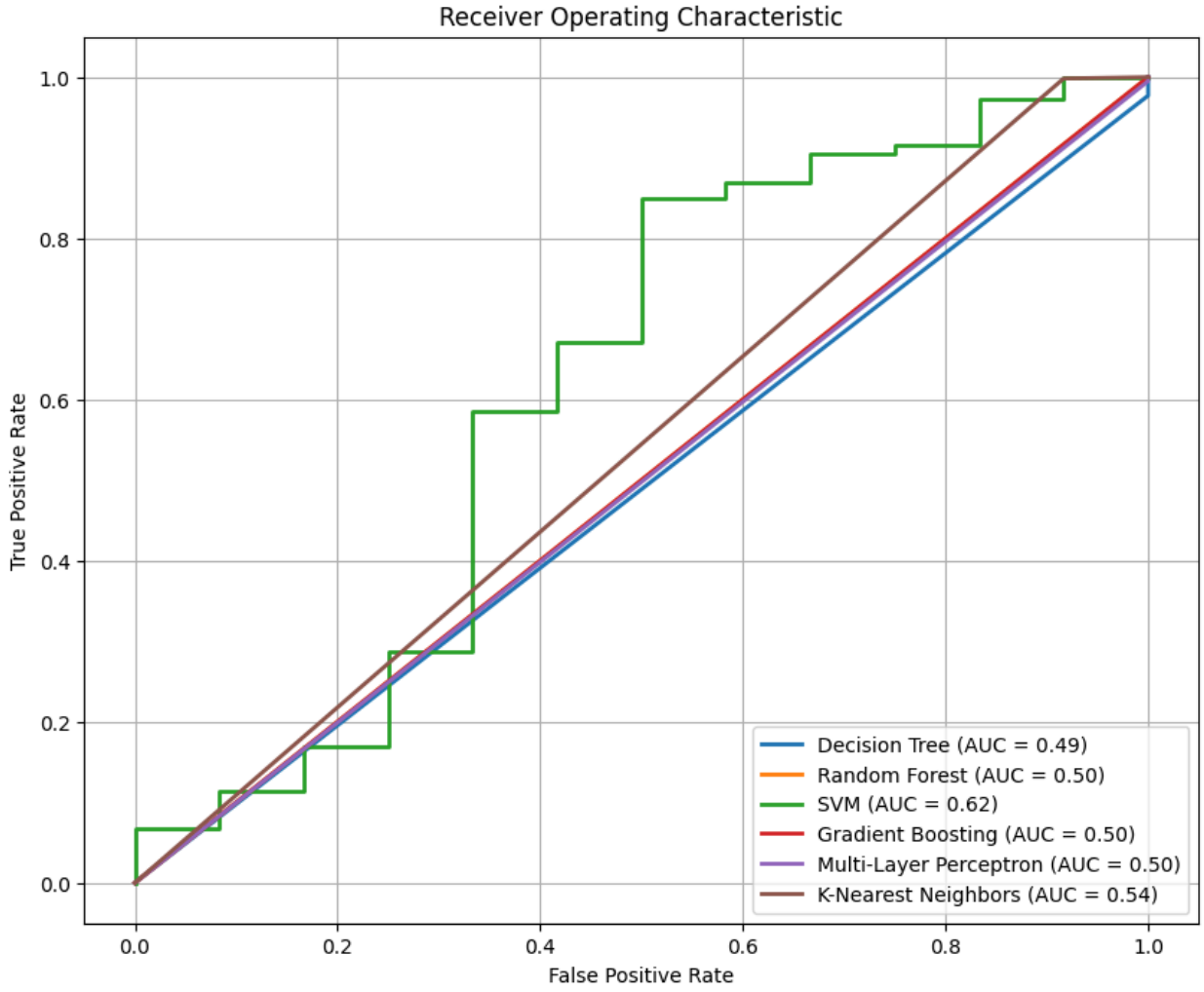
In the mushroom dataset, the impact of rebalancing was less significant due to its balanced nature. Tomek Links improved recall slightly but reduced precision, with Gradient Boosting and Random Forest achieving the best balance. AUC values changed minimally, typically by only 0.01. SMOTE further enhanced precision and recall, particularly for Gradient Boosting and Decision Trees, which achieved high AUC scores. The hybrid approach offered further improvements, with Random Forest and SVM emerging as top performers, achieving the highest AUC values across all models.

Model Confusion Matrix, Precision, and Recall for Chocolate:

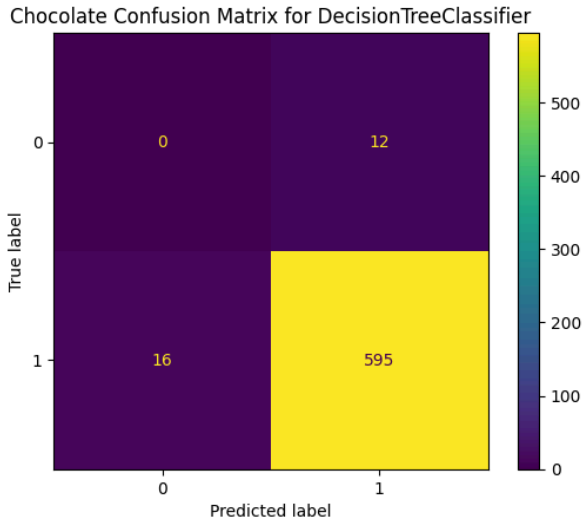
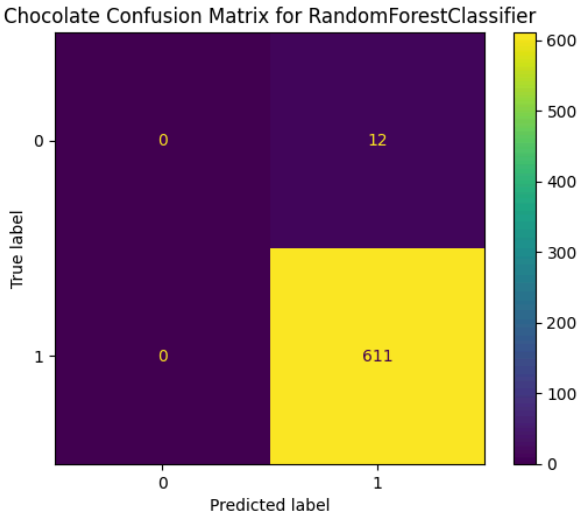
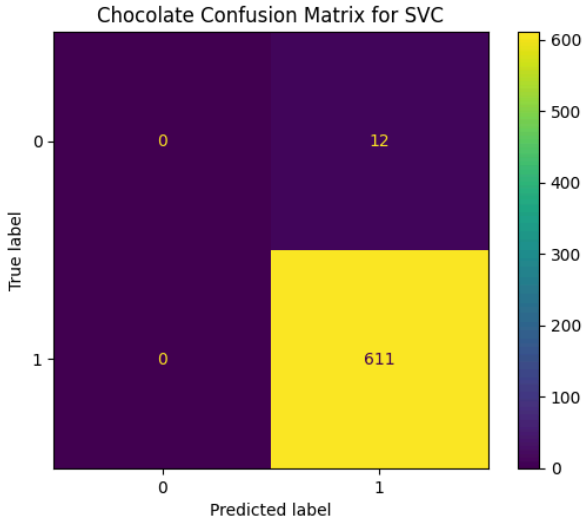
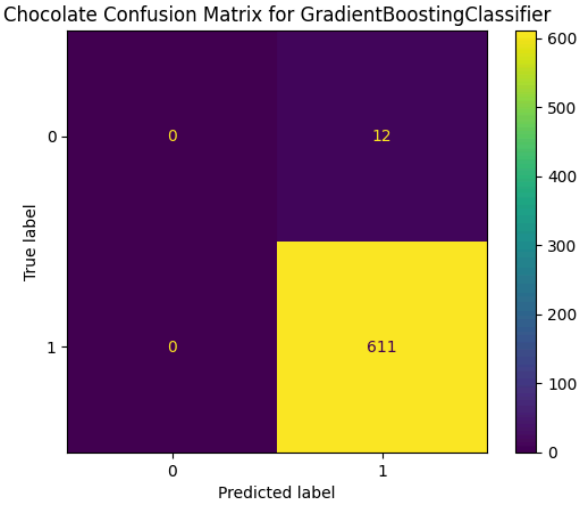
DecisionTreeClassifier Precision: 0.9802955665024631 Recall: 0.9770867430441899	**RandomForestClassifier** Precision: 0.9807383627608347 Recall: 1.0
	
SVC Precision: 0.9807383627608347 Recall: 1.0	**GradientBoostingClassifier** Precision: 0.9807383627608347 Recall: 1.0
	
MLPClassifier Precision: 0.9806451612903225 Recall: 0.9950900163666121	**KNeighborsClassifier** Precision: 0.9822866344605475 Recall: 0.9983633387888707

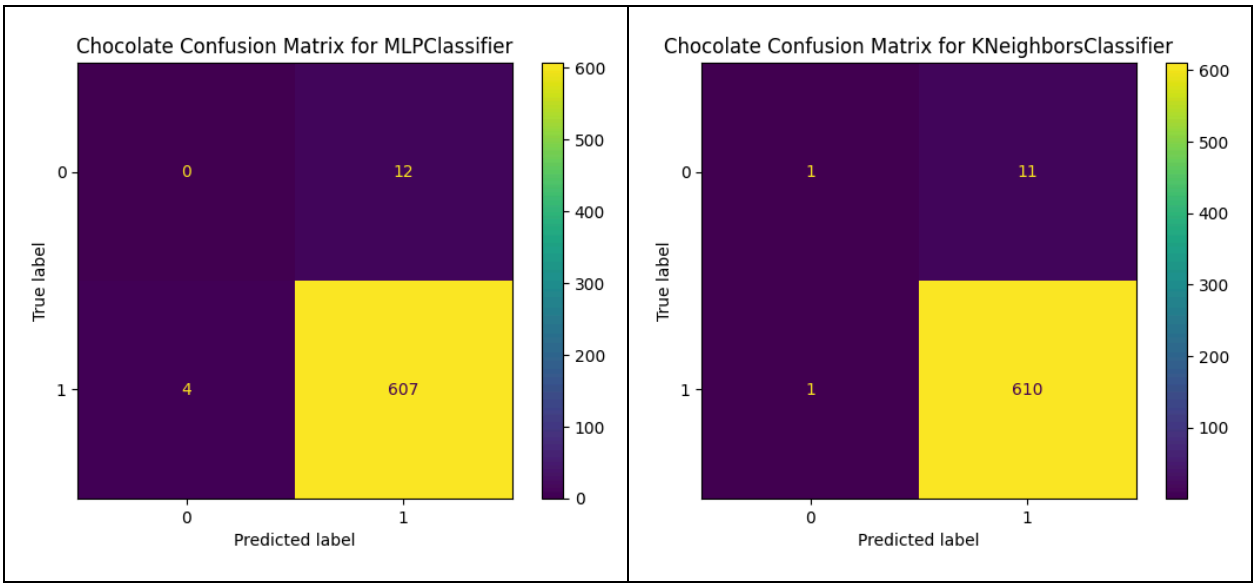


ROC Curve for Chocolate:

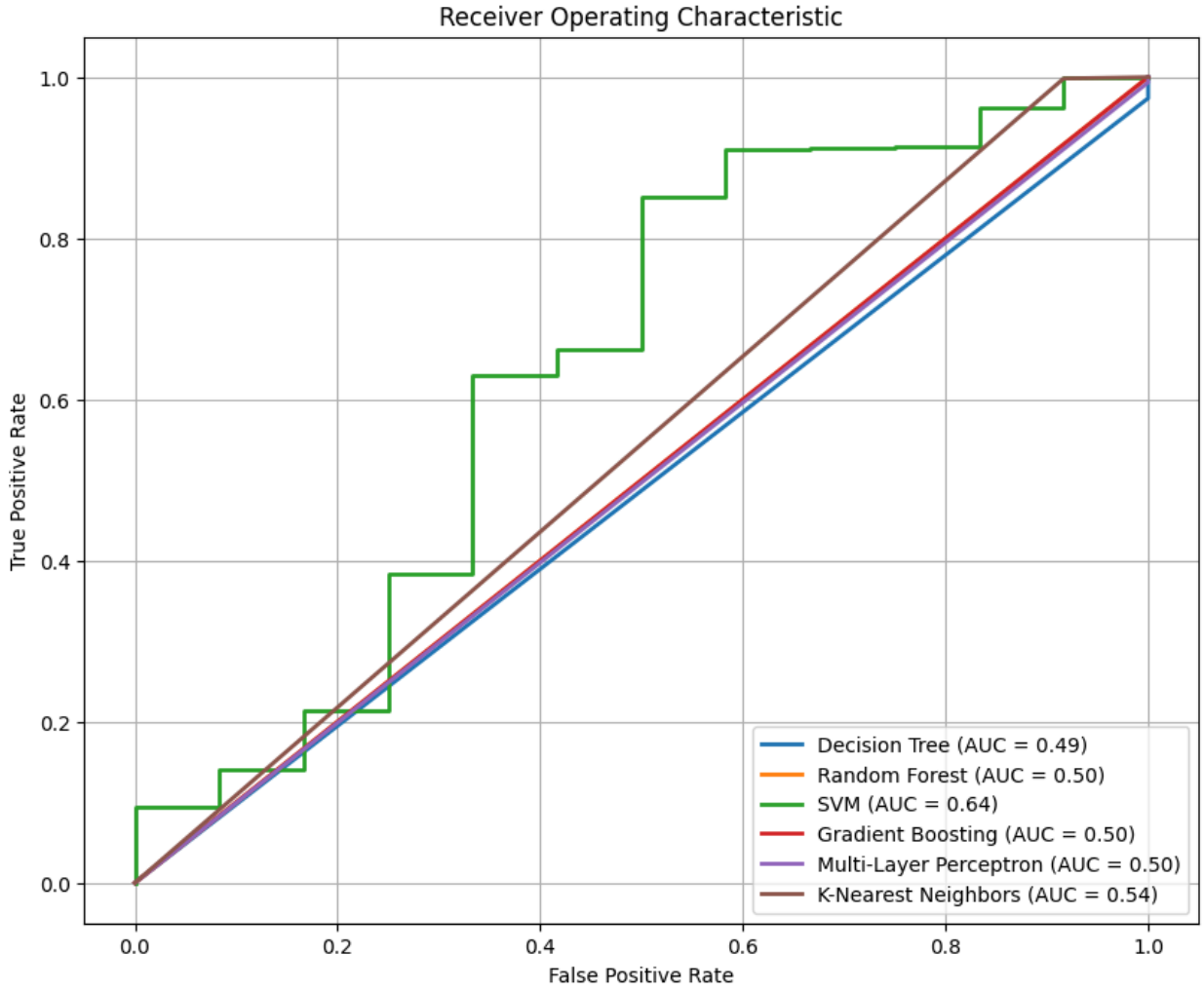


Model Confusion Matrix, Precision, and Recall for Undersampled Chocolate:

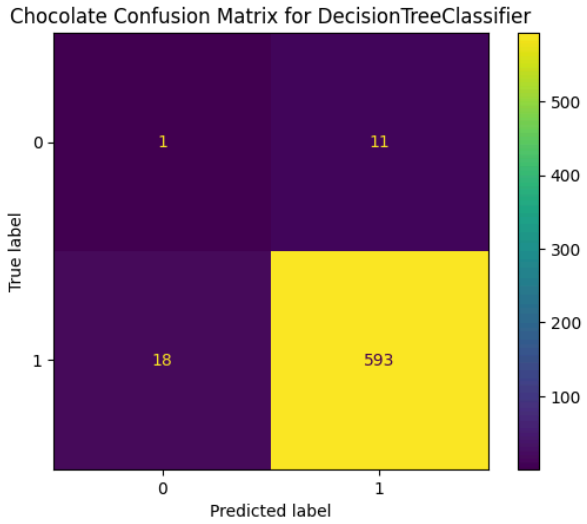
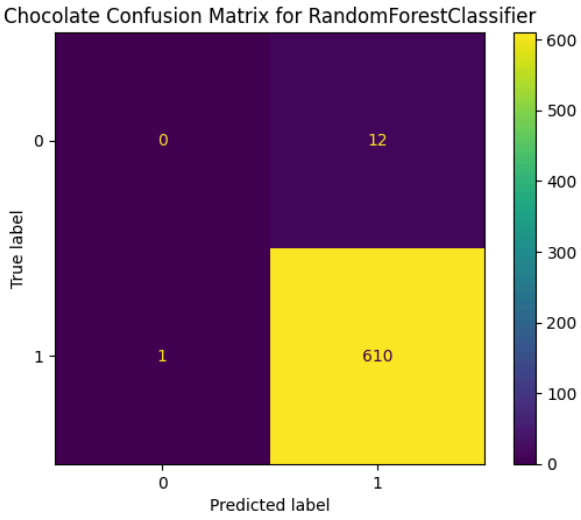
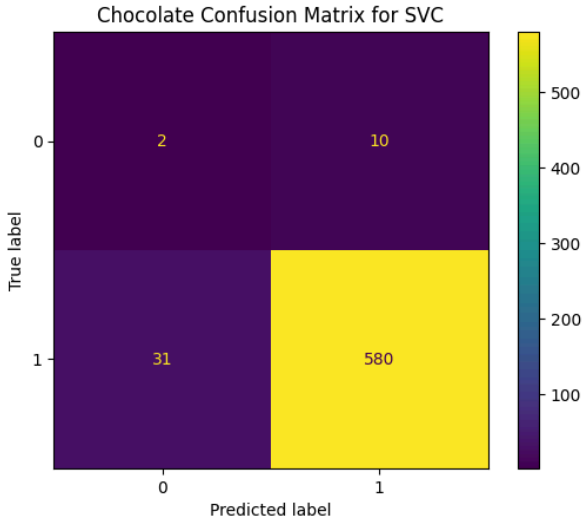
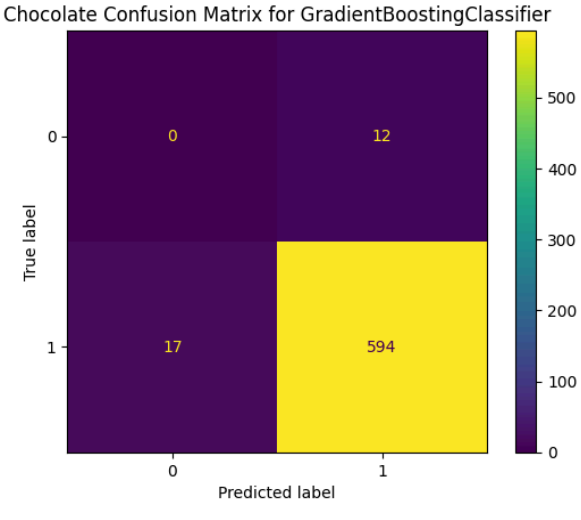
DecisionTreeClassifier Precision: 0.9802306425041186 Recall: 0.9738134206219312	**RandomForestClassifier** Precision: 0.9807383627608347 Recall: 1.0
	
SVC Precision: 0.9807383627608347 Recall: 1.0	**GradientBoostingClassifier** Precision: 0.9807383627608347 Recall: 1.0
	
MLPClassifier Precision: 0.9806138933764136 Recall: 0.9934533551554828	**KNeighborsClassifier** Precision: 0.9822866344605475 Recall: 0.9983633387888707

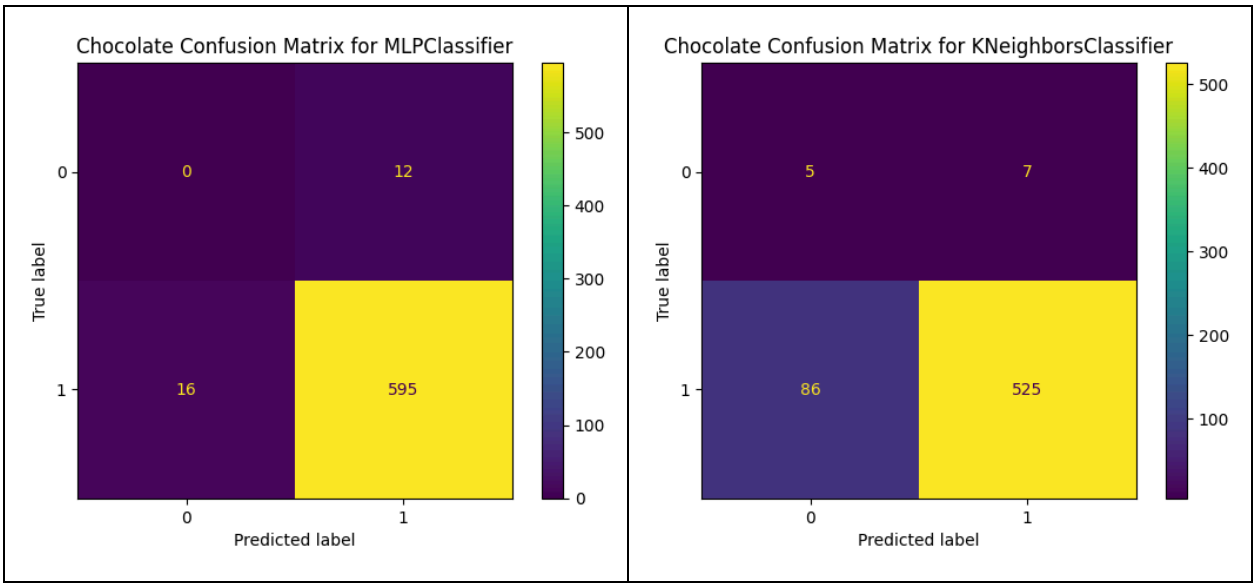


ROC Curve for Undersampled Chocolate:

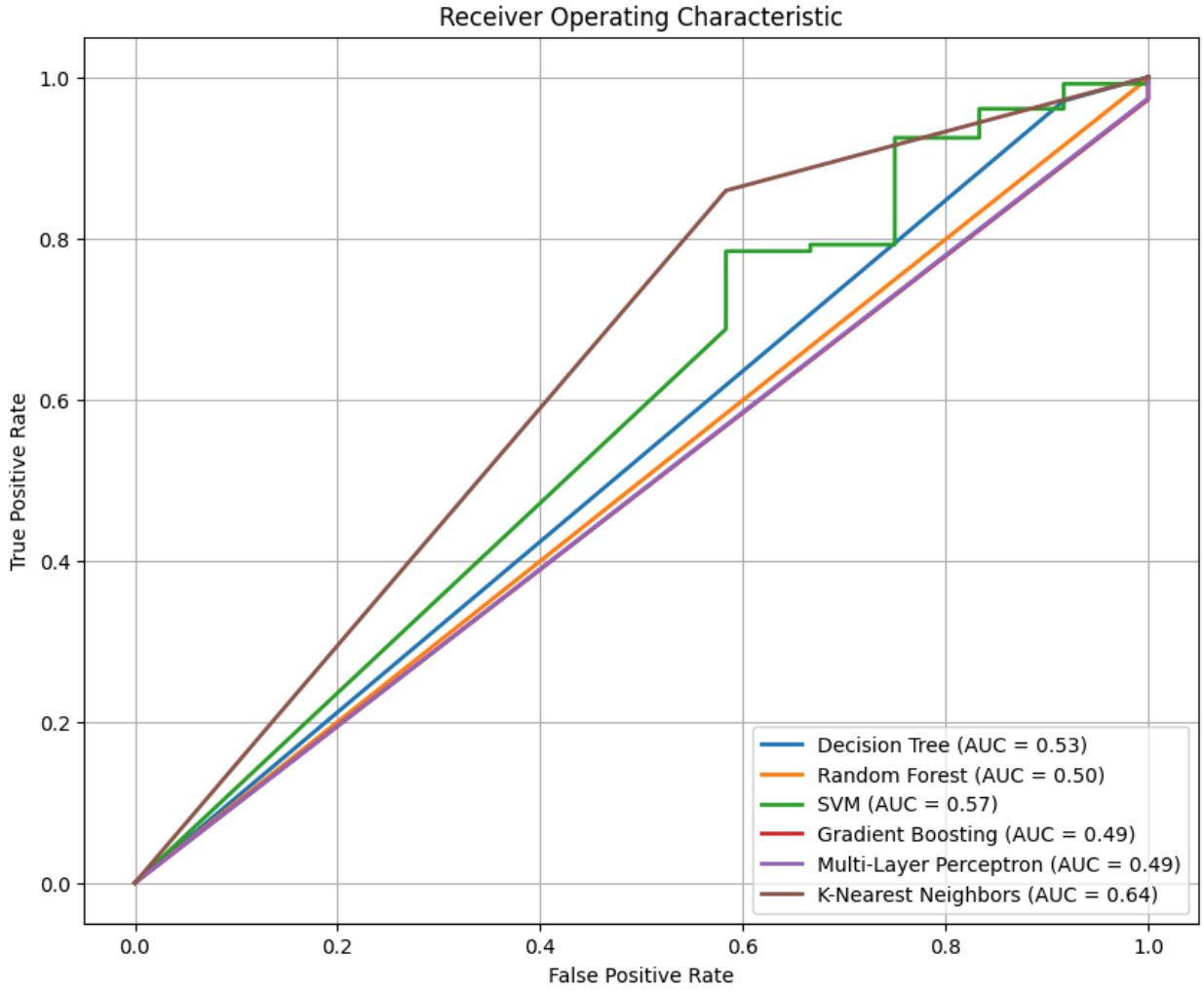


Model Confusion Matrix, Precision, and Recall for Oversampled Chocolate:

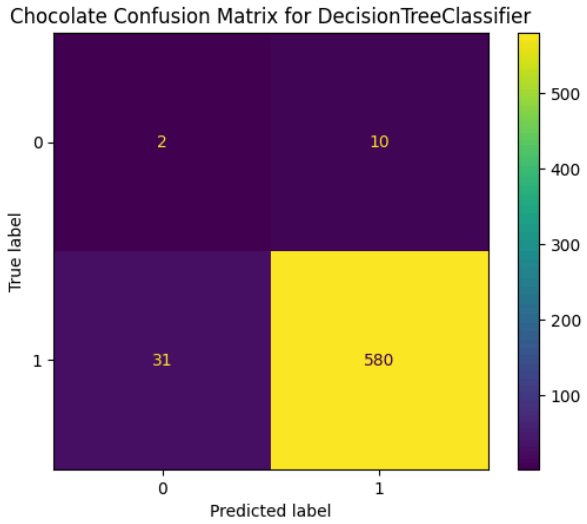
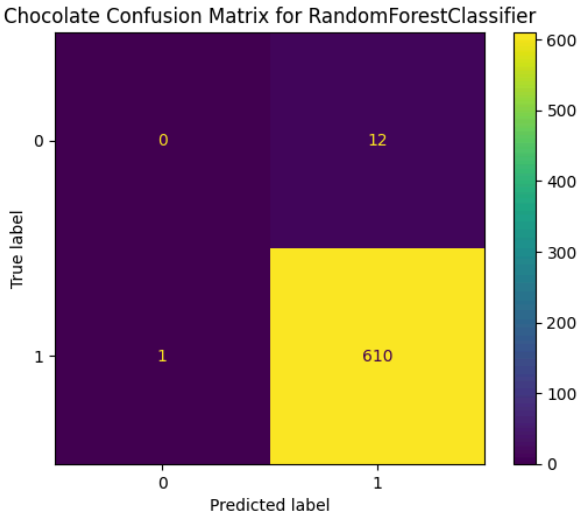
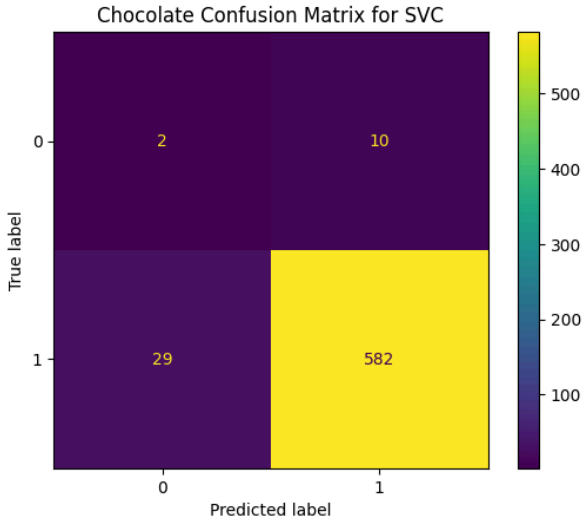
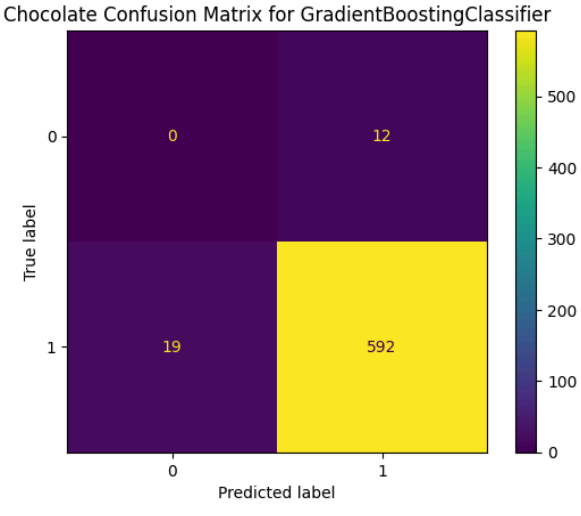
DecisionTreeClassifier Precision: 0.9817880794701986 Recall: 0.9705400981996727	**RandomForestClassifier** Precision: 0.9807073954983923 Recall: 0.9983633387888707
	
SVC Precision: 0.9830508474576272 Recall: 0.9492635024549918	**GradientBoostingClassifier** Precision: 0.9801980198019802 Recall: 0.972176759410802
	
MLPClassifier Precision: 0.9802306425041186 Recall: 0.9738134206219312	**KNeighborsClassifier** Precision: 0.9868421052631579 Recall: 0.8592471358428805

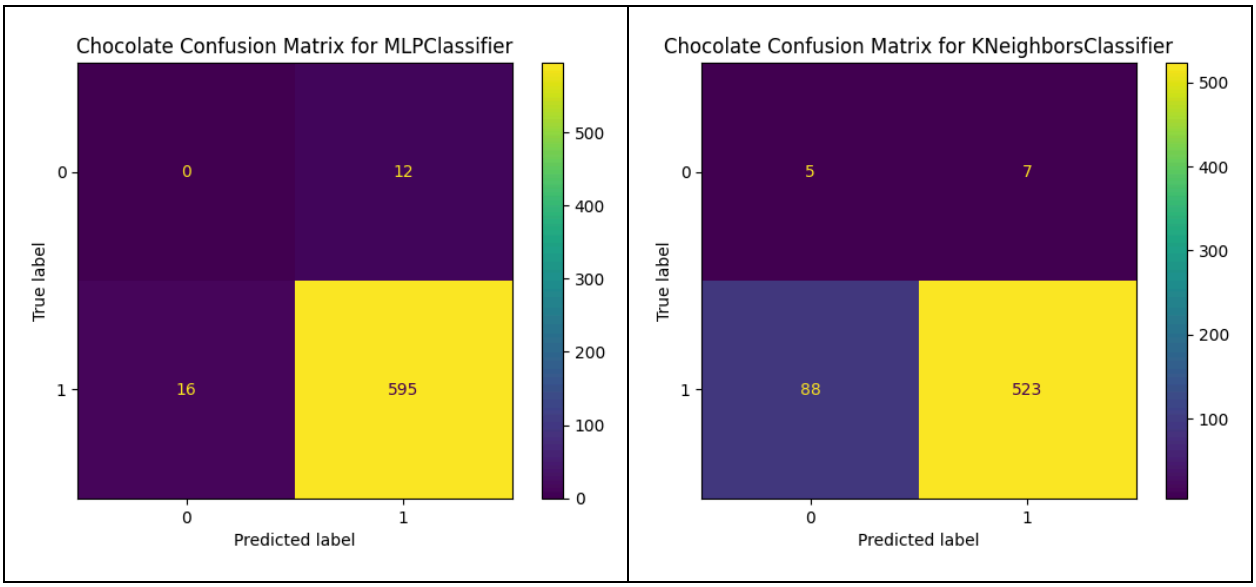


ROC Curve for Oversampled Chocolate:

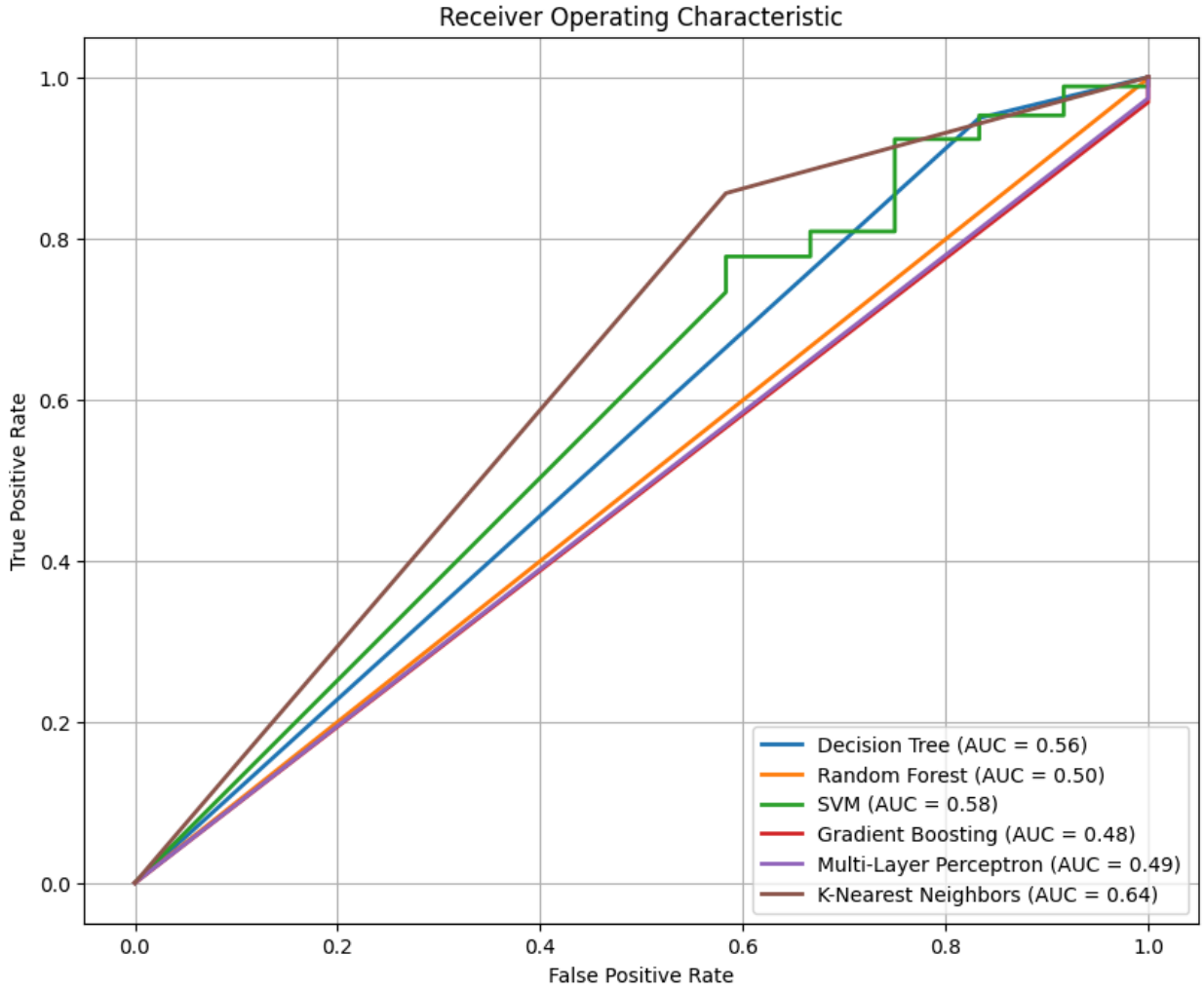


Model Confusion Matrix, Precision, and Recall for Combined Sampled Chocolate:

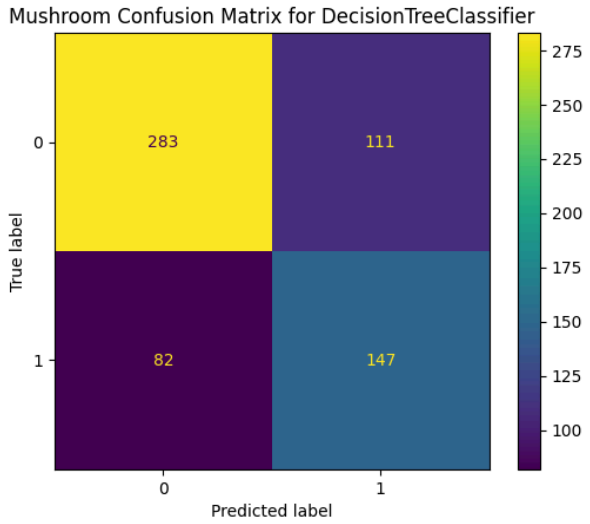
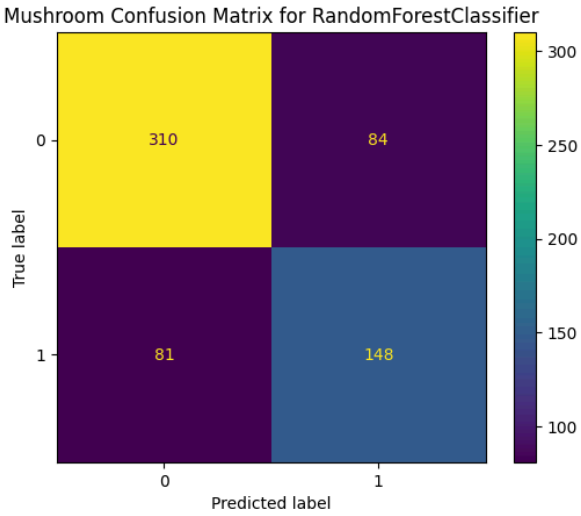
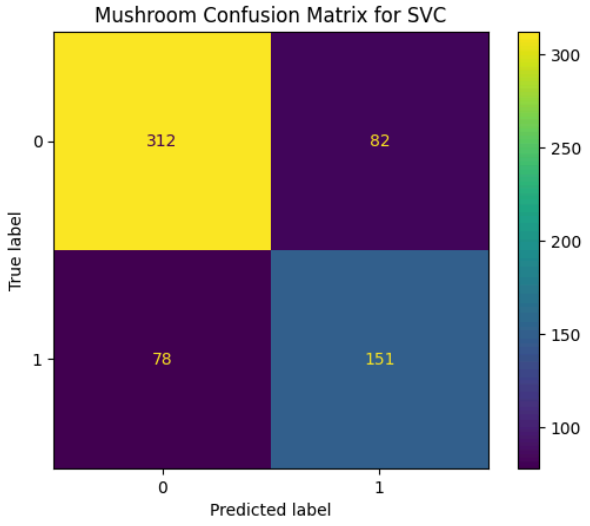
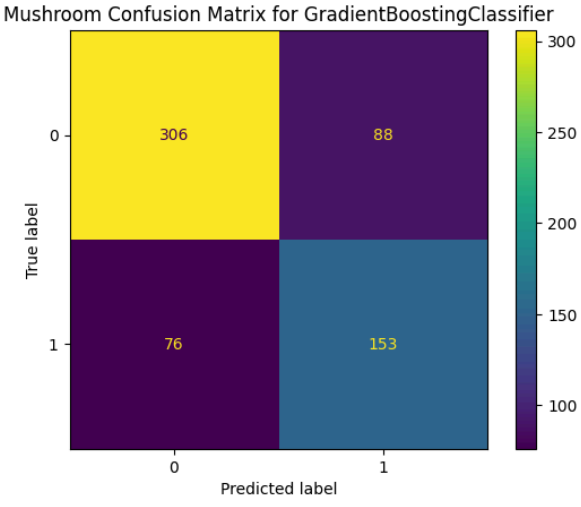
DecisionTreeClassifier Precision: 0.9830508474576272 Recall: 0.9492635024549918	**RandomForestClassifier** Precision: 0.9807073954983923 Recall: 0.9983633387888707
	
SVC Precision: 0.9831081081081081 Recall: 0.9525368248772504	**GradientBoostingClassifier** Precision: 0.9801324503311258 Recall: 0.9689034369885434
	
MLPClassifier Precision: 0.9802306425041186 Recall: 0.9738134206219312	**KNeighborsClassifier** Precision: 0.9867924528301887 Recall: 0.855973813420622

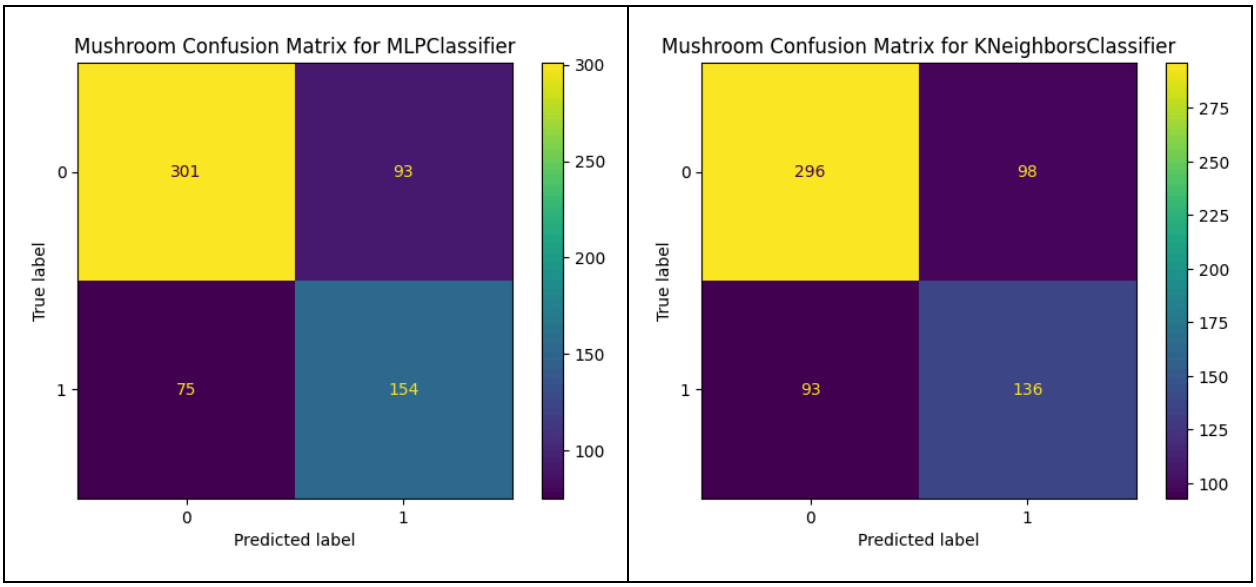


ROC Curve for Combined Sampled Chocolate:

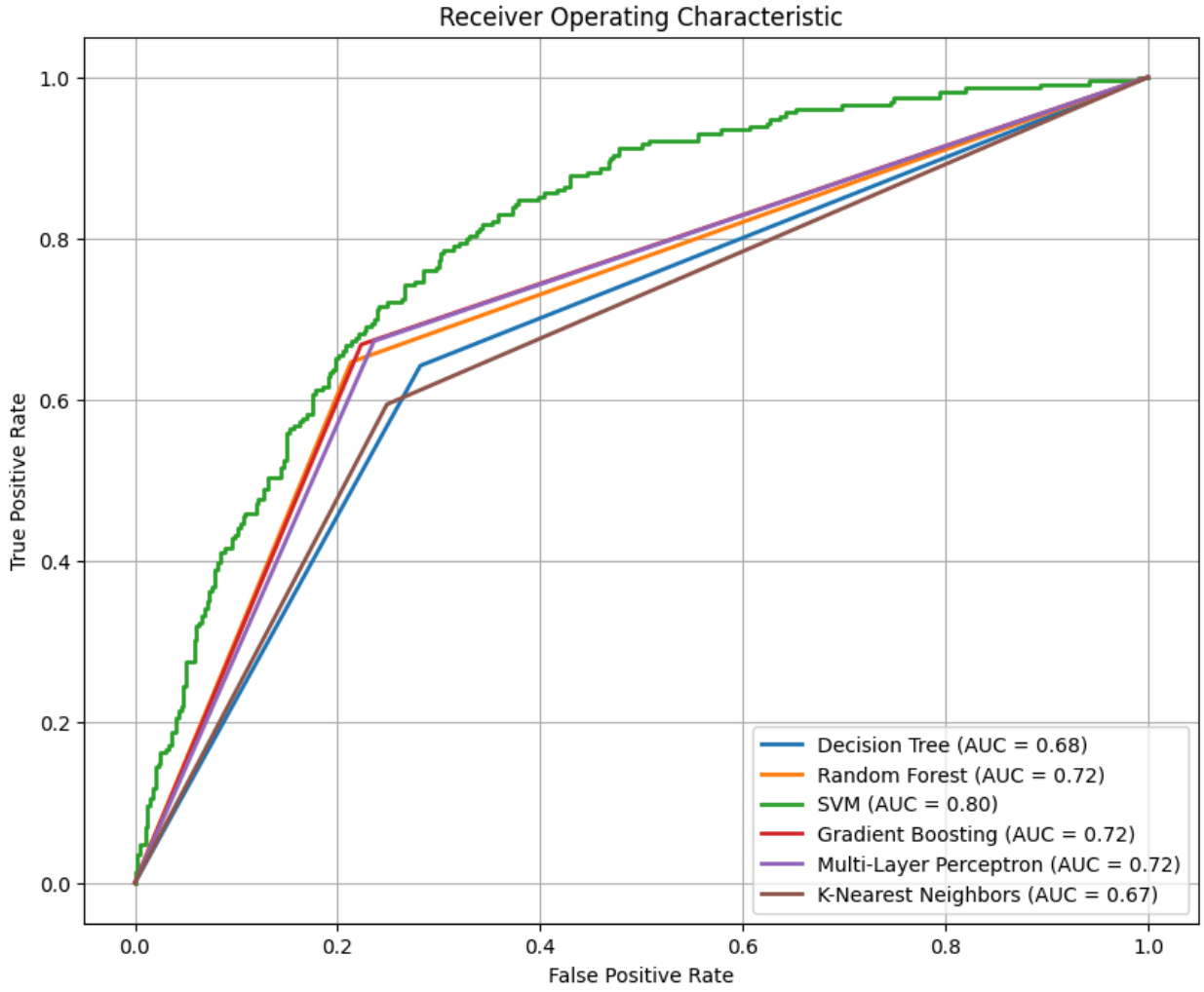


Model Confusion Matrix, Precision, and Recall for Mushrooms:

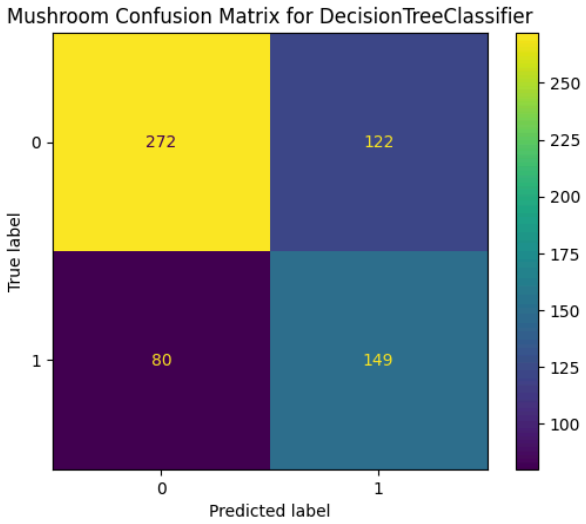
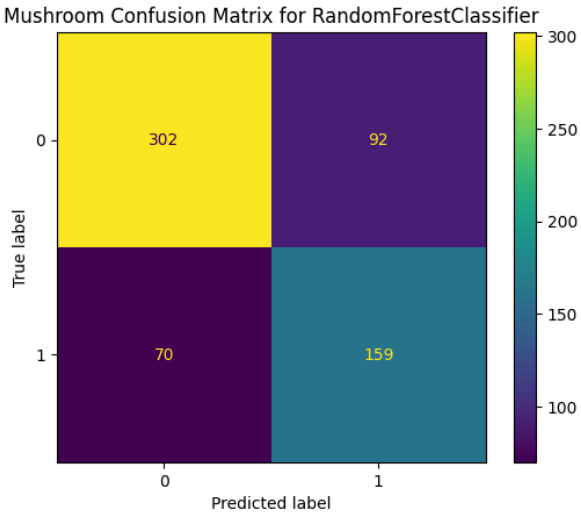
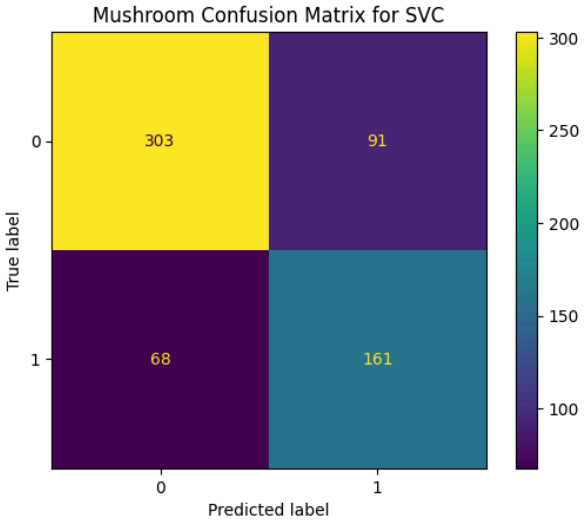
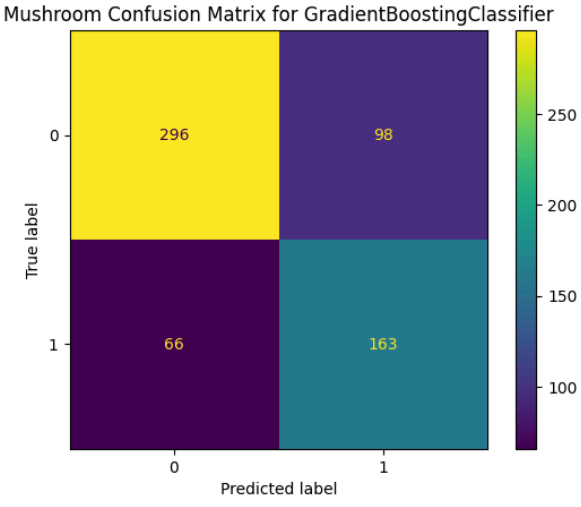
DecisionTreeClassifier Precision: 0.5697674418604651 Recall: 0.6419213973799127	**RandomForestClassifier** Precision: 0.6379310344827587 Recall: 0.6462882096069869
	
SVC Precision: 0.648068669527897 Recall: 0.6593886462882096	**GradientBoostingClassifier** Precision: 0.6348547717842323 Recall: 0.6681222707423581
	
MLPClassifier Precision: 0.6234817813765182 Recall: 0.6724890829694323	**KNeighborsClassifier** Precision: 0.5811965811965812 Recall: 0.5938864628820961

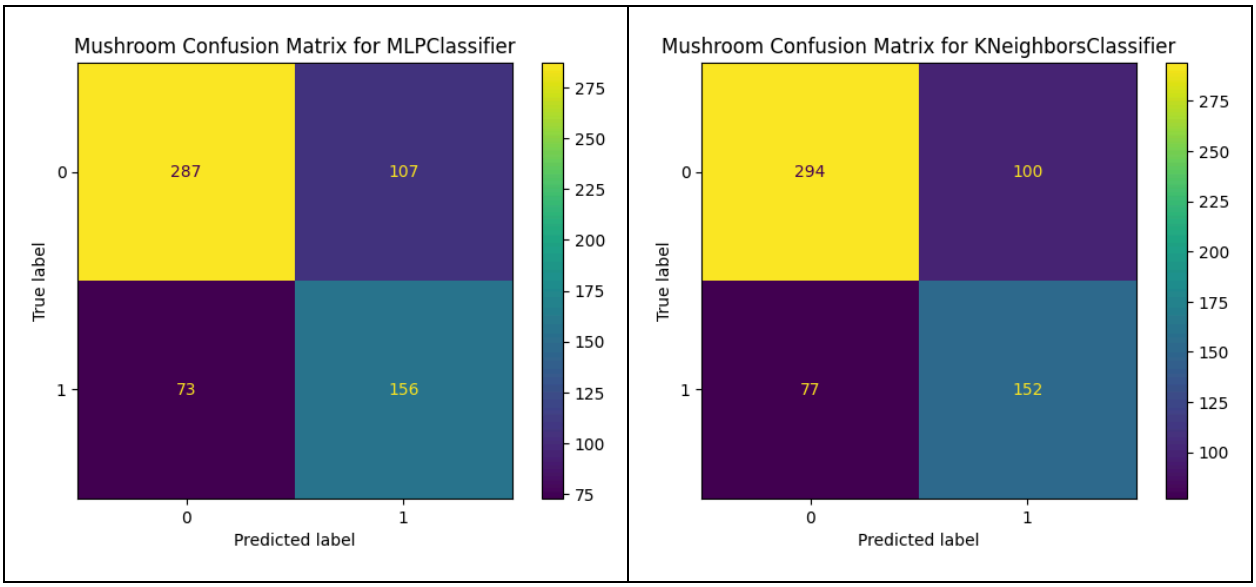


ROC Curve for Mushrooms:

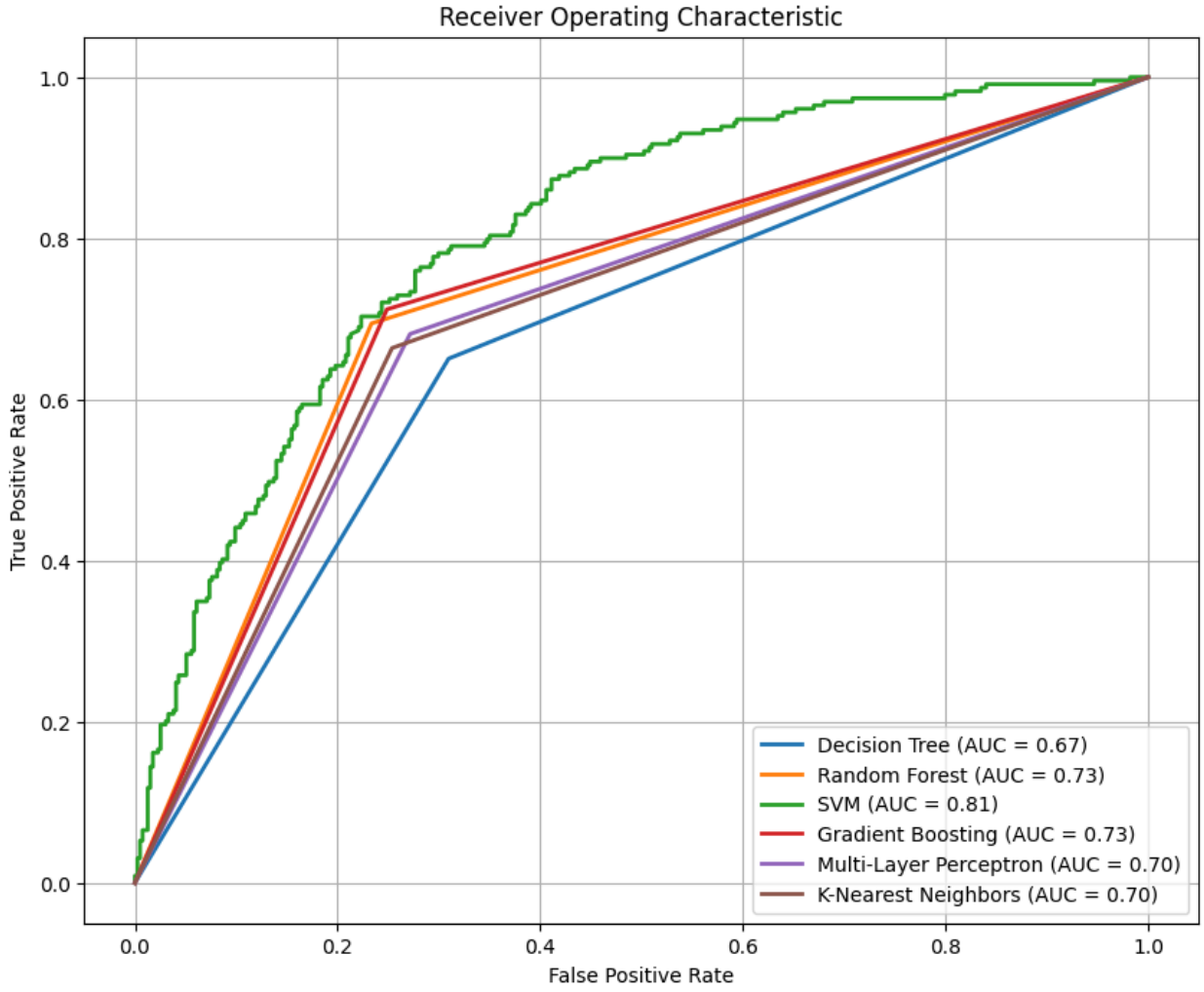


Model Confusion Matrix, Precision, and Recall for Undersampled Mushrooms:

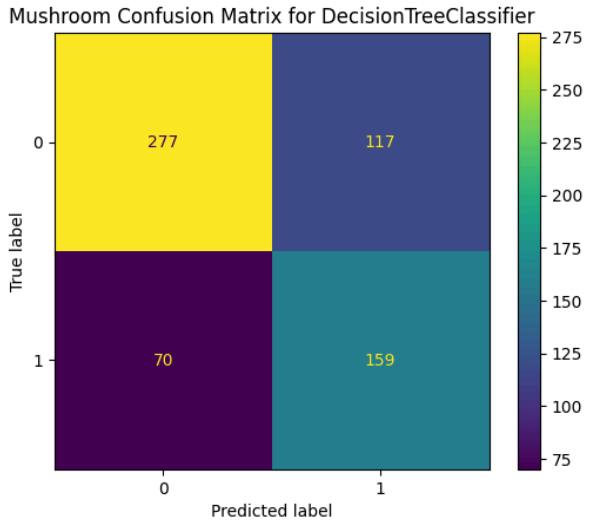
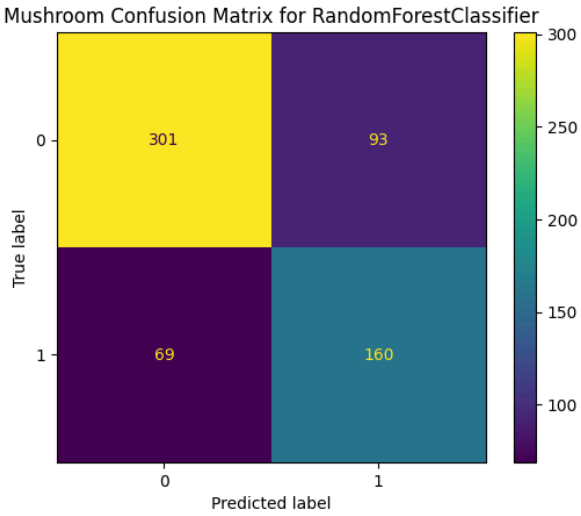
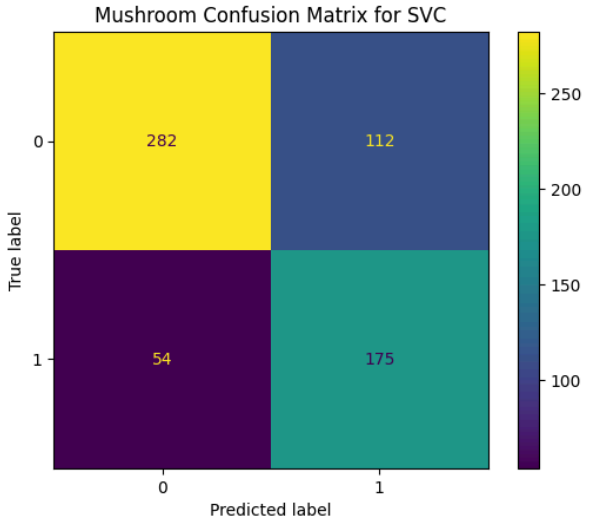
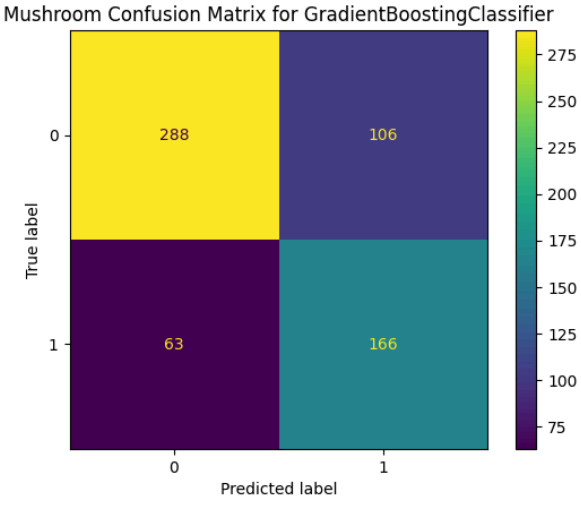
DecisionTreeClassifier Precision: 0.5498154981549815 Recall: 0.6506550218340611	**RandomForestClassifier** Precision: 0.6334661354581673 Recall: 0.6943231441048034
	
SVC Precision: 0.6388888888888888 Recall: 0.7030567685589519	**GradientBoostingClassifier** Precision: 0.6245210727969349 Recall: 0.7117903930131004
	
MLPClassifier Precision: 0.5931558935361216 Recall: 0.6812227074235808	**KNeighborsClassifier** Precision: 0.6031746031746031 Recall: 0.6637554585152838

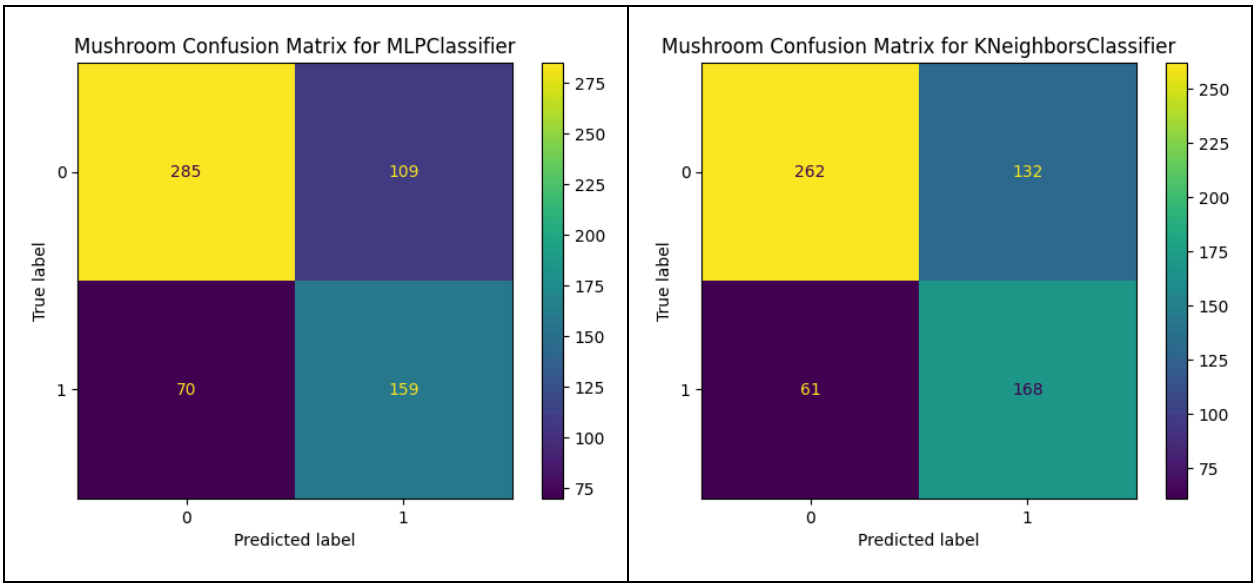


ROC Curve for Undersampled Mushrooms:

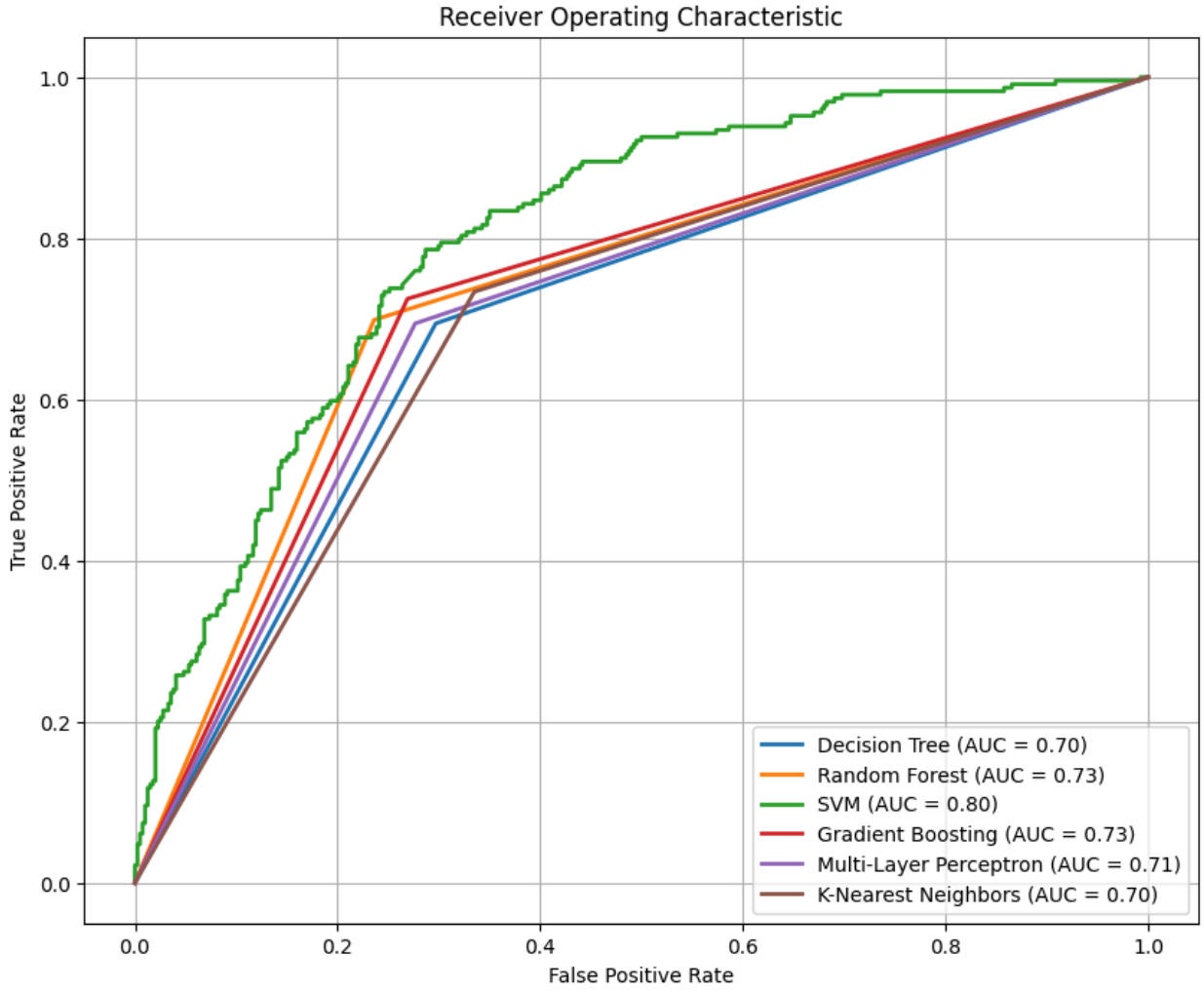


Model Confusion Matrix, Precision, and Recall for Oversampled Mushrooms:

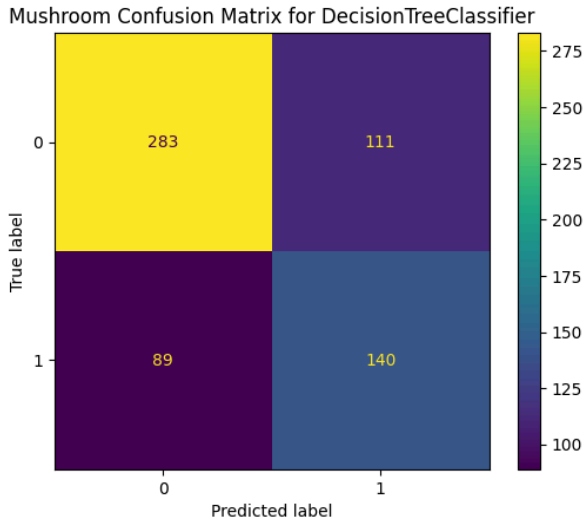
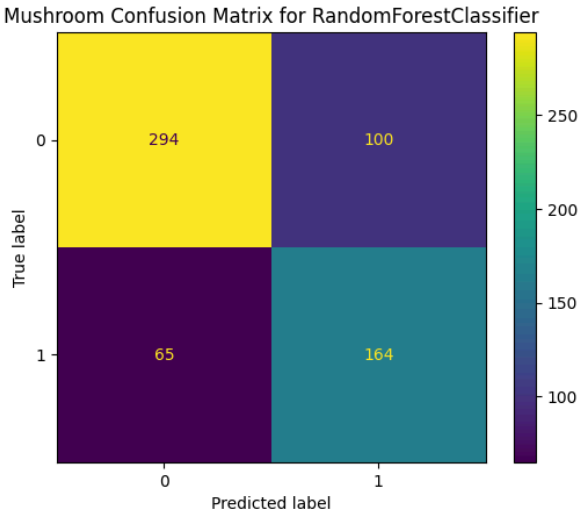
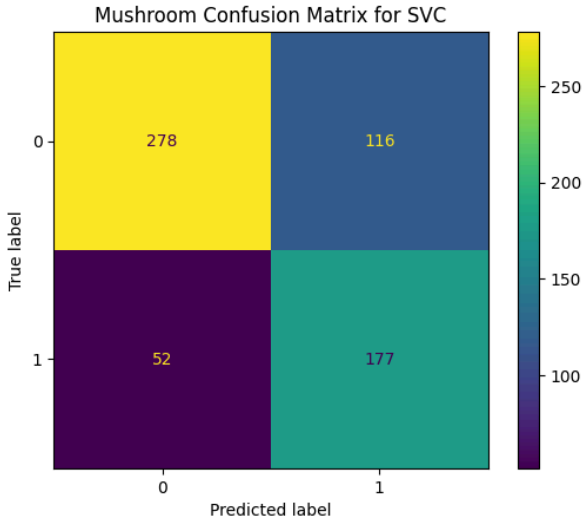
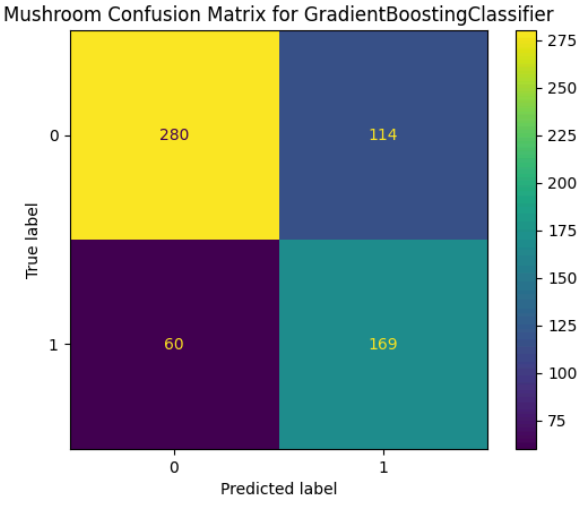
DecisionTreeClassifier Precision: 0.5760869565217391 Recall: 0.6943231441048034	**RandomForestClassifier** Precision: 0.6324110671936759 Recall: 0.6986899563318777
	
SVC Precision: 0.6097560975609756 Recall: 0.7641921397379913	**GradientBoostingClassifier** Precision: 0.6102941176470589 Recall: 0.7248908296943232
	
MLPClassifier Precision: 0.5932835820895522 Recall: 0.6943231441048034	**KNeighborsClassifier** Precision: 0.56 Recall: 0.7336244541484717

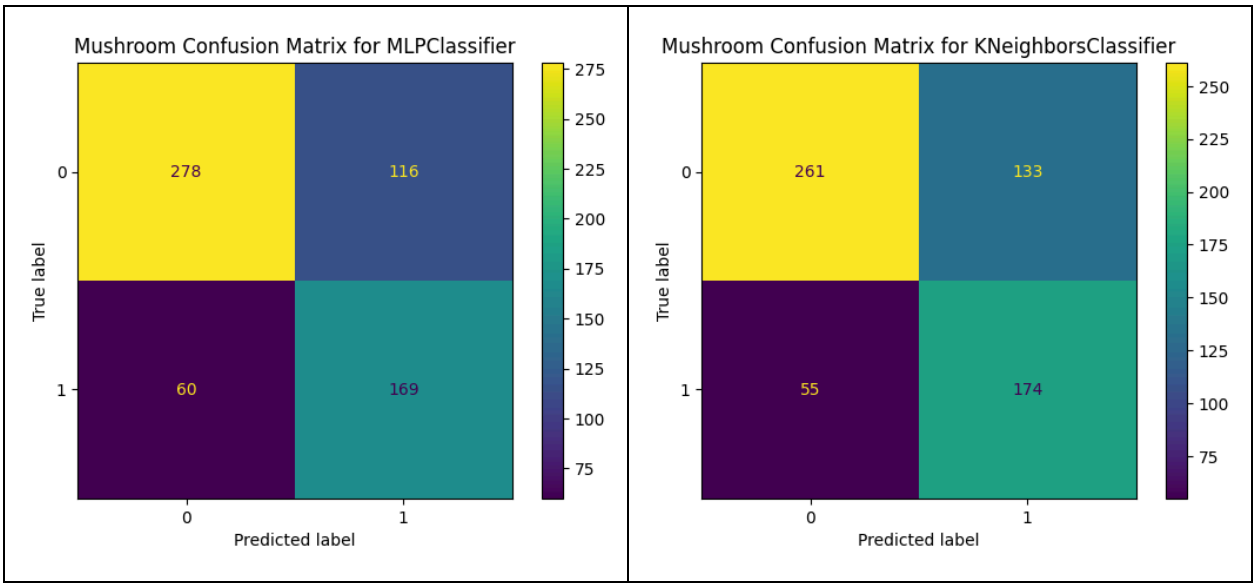


ROC Curve for Oversampled Mushrooms:



Model Confusion Matrix, Precision, and Recall for Combined Sampled Mushrooms:

DecisionTreeClassifier Precision: 0.5577689243027888 Recall: 0.611353711790393	**RandomForestClassifier** Precision: 0.6212121212121212 Recall: 0.7161572052401747
	
SVC Precision: 0.6040955631399317 Recall: 0.7729257641921398	**GradientBoostingClassifier** Precision: 0.5971731448763251 Recall: 0.7379912663755459
	
MLPClassifier Precision: 0.5929824561403508 Recall: 0.7379912663755459	**KNeighborsClassifier** Precision: 0.5667752442996743 Recall: 0.759825327510917



ROC Curve for Combined Sampled Mushrooms: