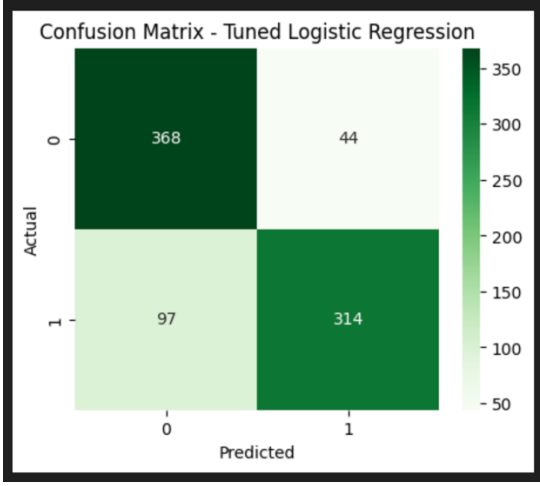
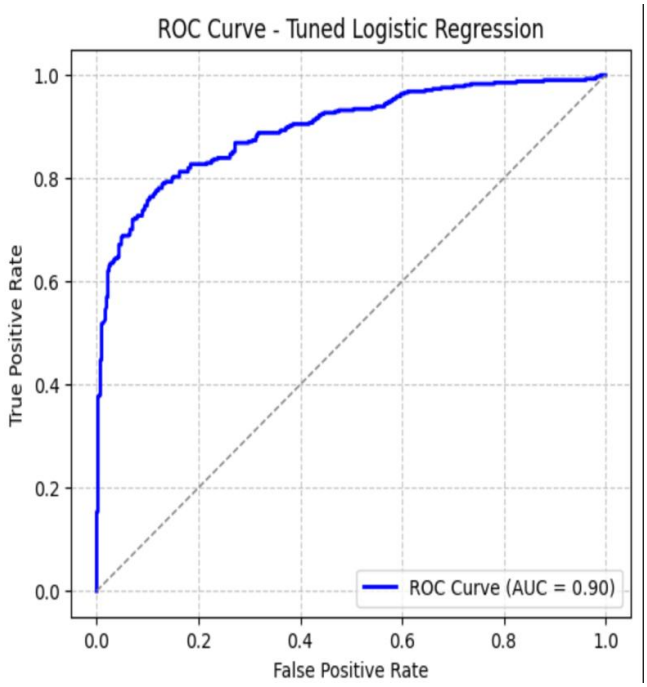
Models & Their Scores

Logistic Regression

We selected logistic regression because it's a simple and interpretable model that works well for binary classification problems like predicting student graduation or dropout.

```
Tuned Accuracy: 0.8287

Classification Report (Tuned):
              precision    recall  f1-score   support

           0       0.79      0.89      0.84       412
           1       0.88      0.76      0.82       411

    accuracy                           0.83       823
   macro avg       0.83      0.83      0.83       823
weighted avg       0.83      0.83      0.83       823
```

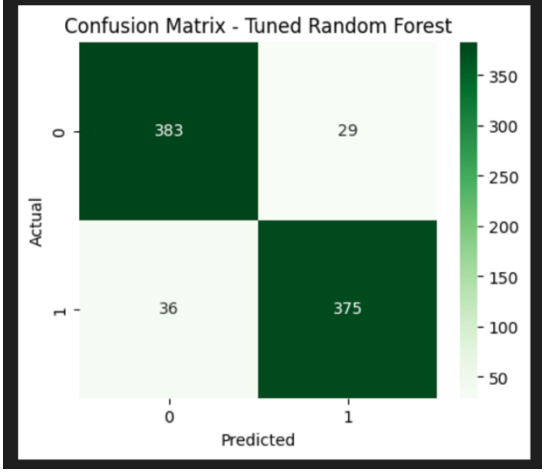


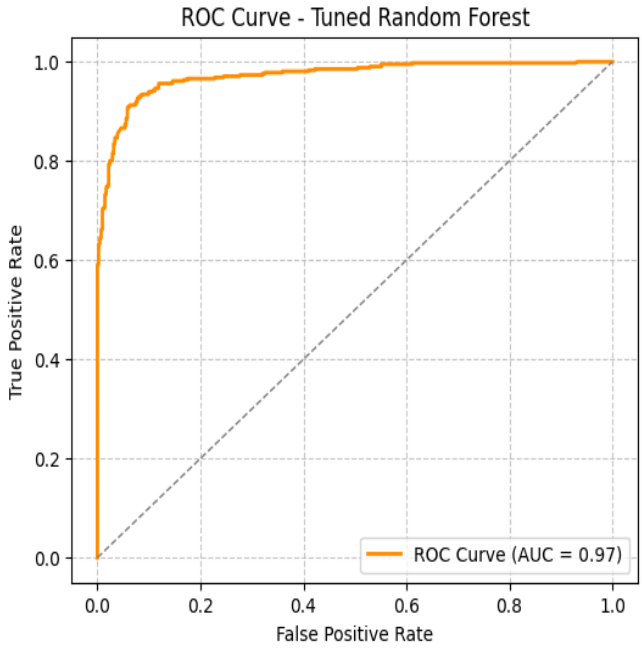


Random Forest

We chose random forest as it's an ensemble of decision trees that provides higher accuracy, reduces overfitting, handles numerical and categorical data effectively, and offers feature importance scores for student retention factors.

```
Tuned Test Accuracy: 0.9210

Classification Report (Tuned):
              precision    recall  f1-score   support

           0       0.91      0.93      0.92       412
           1       0.93      0.91      0.92       411

    accuracy                           0.92       823
   macro avg       0.92      0.92      0.92       823
weighted avg       0.92      0.92      0.92       823
```

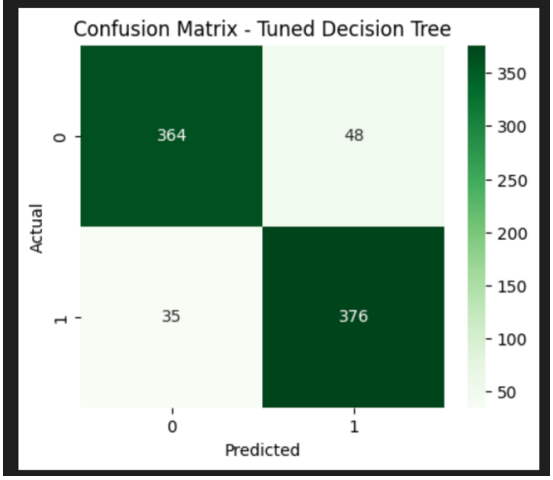


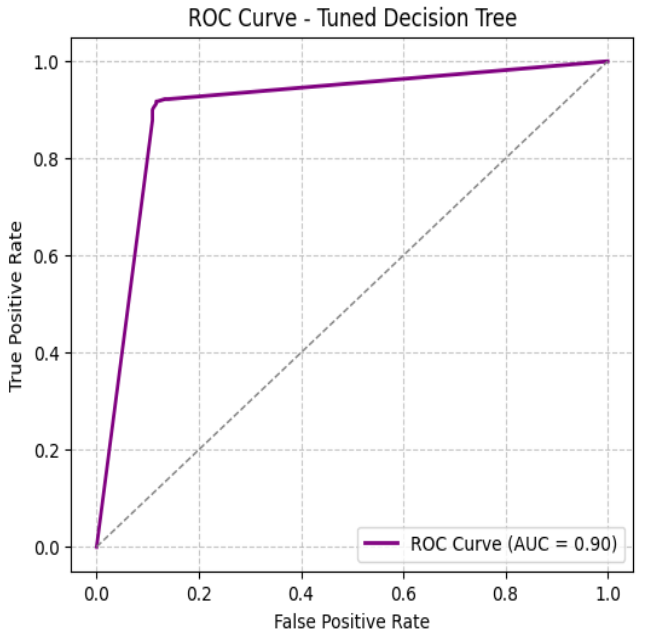


Decision Tree

We picked decision tree because it's easy to understand and visualize, helping identify key factors influencing student outcomes in this binary classification task.

```
Tuned Test Accuracy: 0.8991

Classification Report (Tuned):
              precision    recall  f1-score   support

           0       0.91      0.88      0.90       412
           1       0.89      0.91      0.90       411

    accuracy                           0.90       823
   macro avg       0.90      0.90      0.90       823
weighted avg       0.90      0.90      0.90       823
```

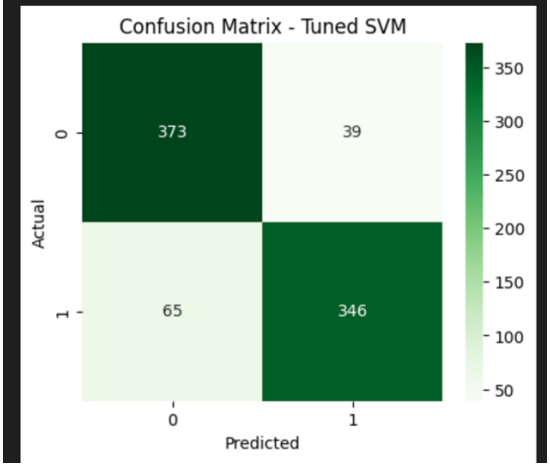


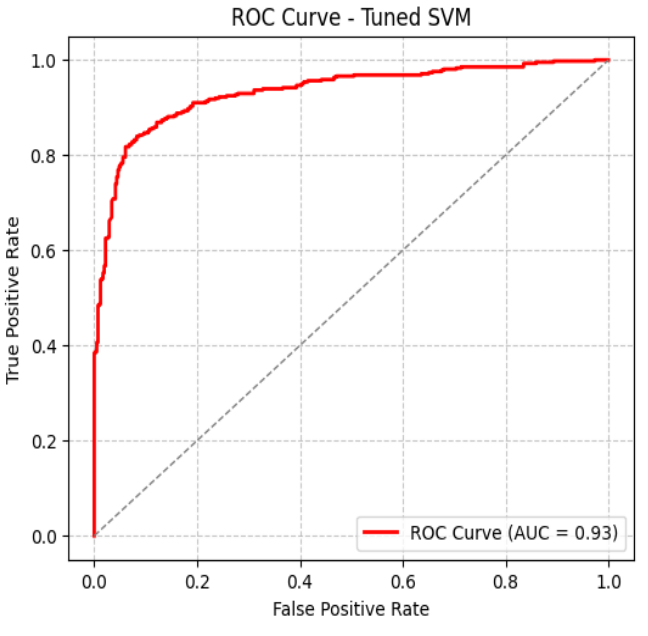


SVM (Support Vector Machine)

We selected SVM for its usefulness in binary classification, as it effectively separates classes with a hyperplane, working well with both linear and non-linear data when margins are clear.

```
Tuned Test Accuracy: 0.8736

Classification Report (Tuned):
              precision    recall  f1-score   support

           0       0.85      0.91      0.88       412
           1       0.90      0.84      0.87       411

    accuracy                           0.87       823
   macro avg       0.88      0.87      0.87       823
weighted avg       0.88      0.87      0.87       823
```

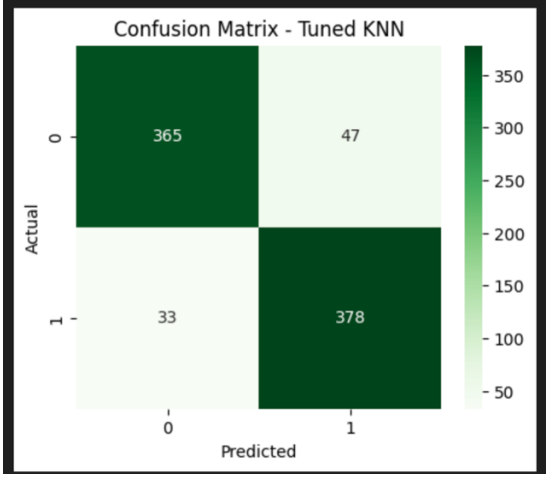


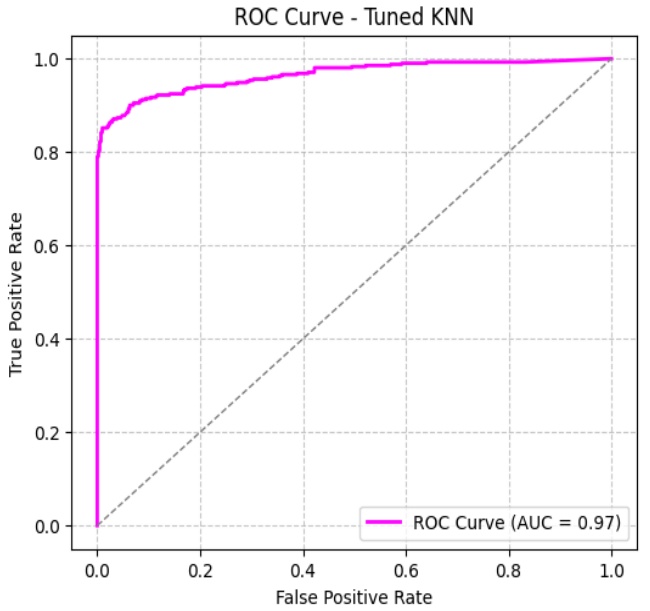


KNN (K-Nearest Neighbors)

We chose KNN as a simple instance-based method that classifies based on similar student profiles, useful for data with clear clusters or patterns in retention prediction.

```
Tuned Test Accuracy: 0.9028

Classification Report (Tuned):
              precision    recall  f1-score   support

           0       0.92      0.89      0.90       412
           1       0.89      0.92      0.90       411

    accuracy                           0.90       823
   macro avg       0.90      0.90      0.90       823
weighted avg       0.90      0.90      0.90       823
```

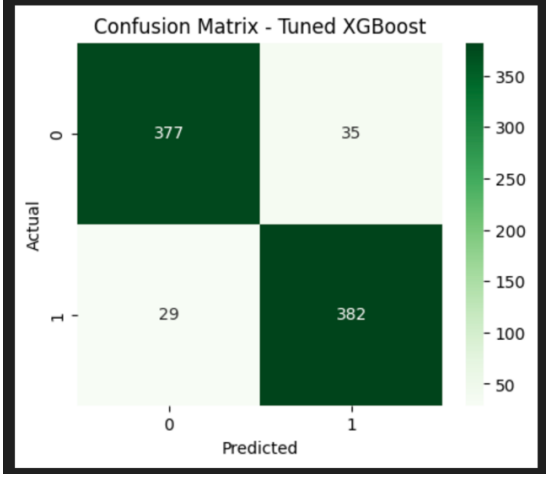


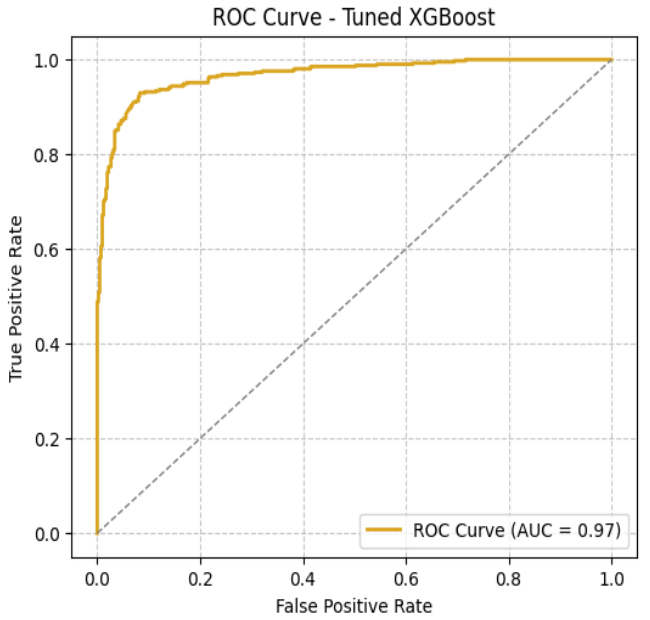


XGBoost

We selected XGBoost for its efficiency as a gradient boosting algorithm that delivers high accuracy on structured tabular datasets like ours for student outcome prediction.

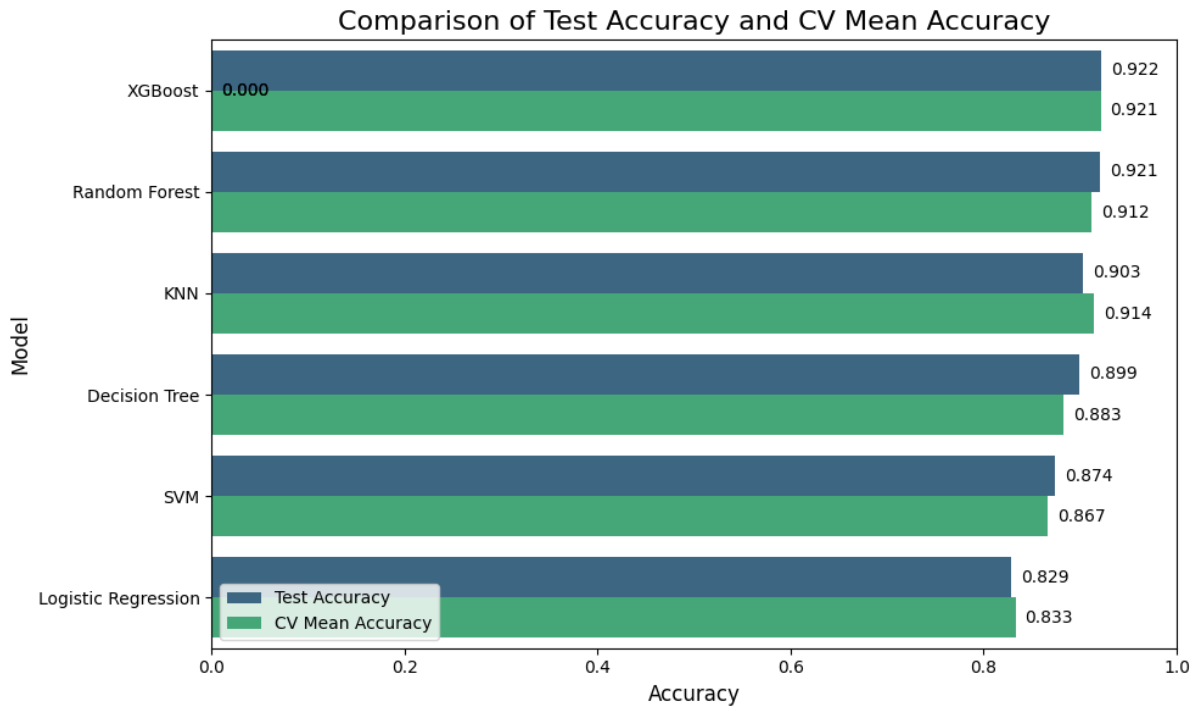
```
Tuned Test Accuracy: 0.9222

Classification Report (Tuned):
              precision    recall  f1-score   support

           0       0.93      0.92      0.92       412
           1       0.92      0.93      0.92       411

    accuracy                           0.92       823
   macro avg       0.92      0.92      0.92       823
weighted avg       0.92      0.92      0.92       823
```





Model Comparison

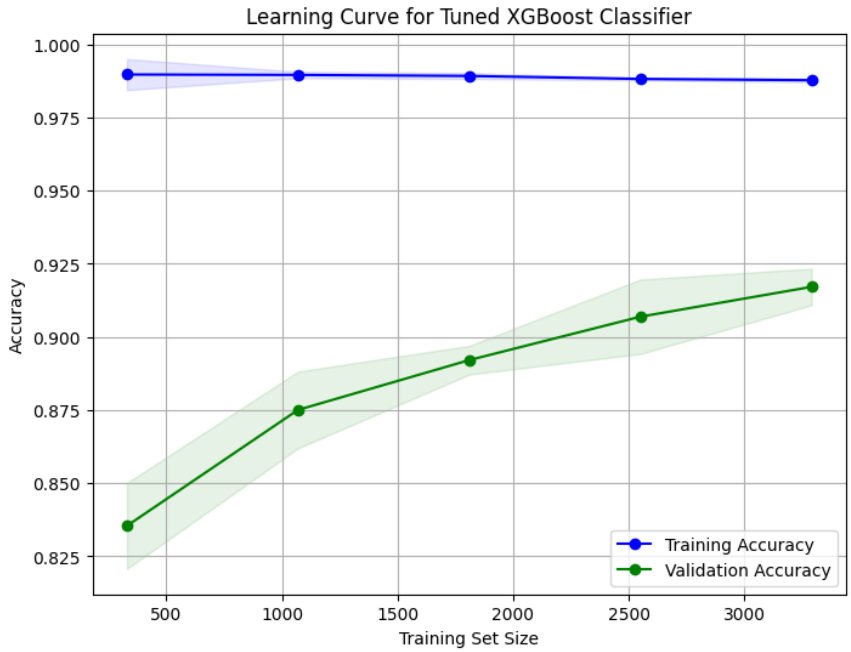
Model	Test Accuracy	CV Mean Accuracy	CV std
XGBoost	0.9222	0.9213	0.0089
Random Forest	0.9210	0.9116	0.0073
KNN	0.9028	0.9143	0.0054
Decision Tree	0.8991	0.8827	0.0126
SVM	0.8736	0.8669	0.0064
Logistic Regression	0.8287	0.8329	0.0129



Final Model Selected – XGBoost

After evaluating all developed models based on accuracy, cross-validation results, confusion matrices, and ROC curves, the **tuned XGBoost (Extreme Gradient Boosting)** model was selected as the final predictive model for student graduation and dropout classification.

XGBoost demonstrated the **highest overall accuracy** among all models, along with excellent **cross-validation stability** and a **high AUC value** in the ROC analysis. The confusion matrix showed that the model correctly identified most dropout and graduate cases, with minimal misclassifications.



The learning curve of the tuned XGBoost model further confirmed that it achieved a **good balance between bias and variance**, indicating strong generalization to unseen data without overfitting.

The main advantages of XGBoost that contributed to its superior performance include:

- Ability to handle complex nonlinear relationships in the data
- Built-in regularization techniques to prevent overfitting
- Efficient computation and scalability for large datasets

Therefore, the tuned XGBoost model was finalized as the most suitable and reliable algorithm for predicting student retention outcomes. Its strong predictive capability and stable performance make it an effective choice for practical use in educational analytics and early intervention systems.