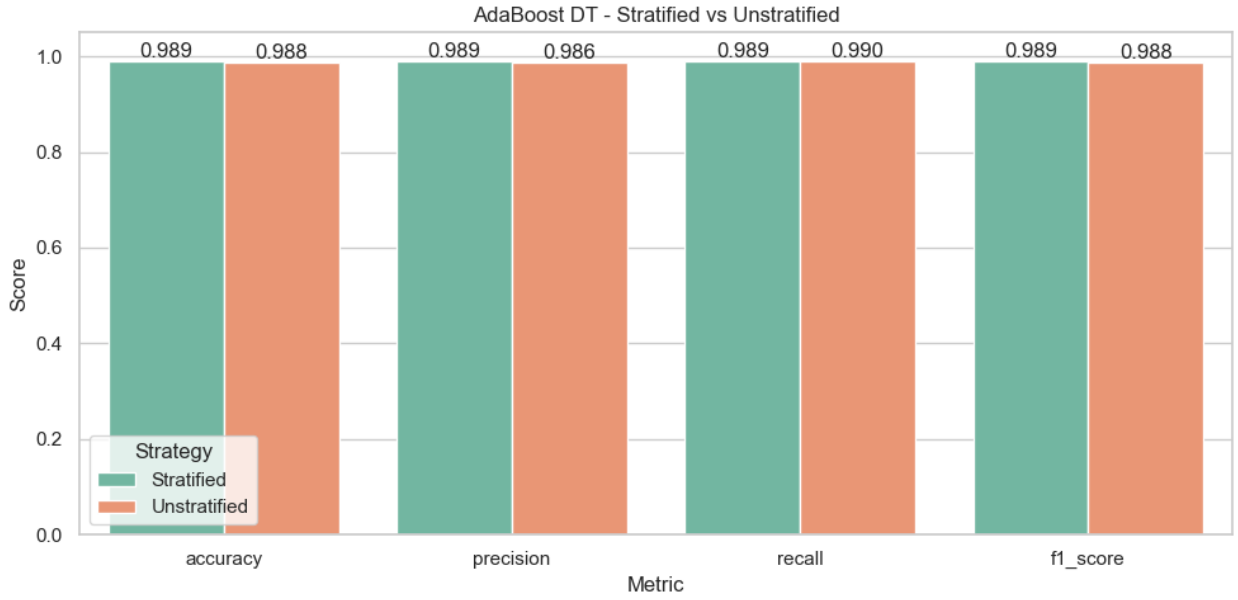
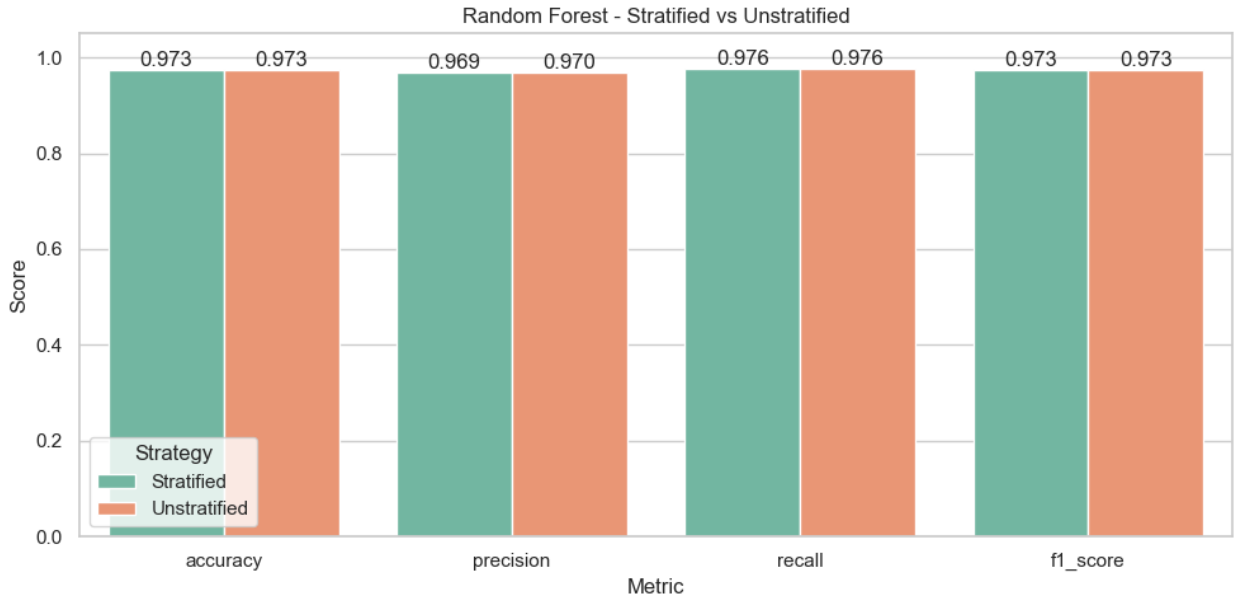
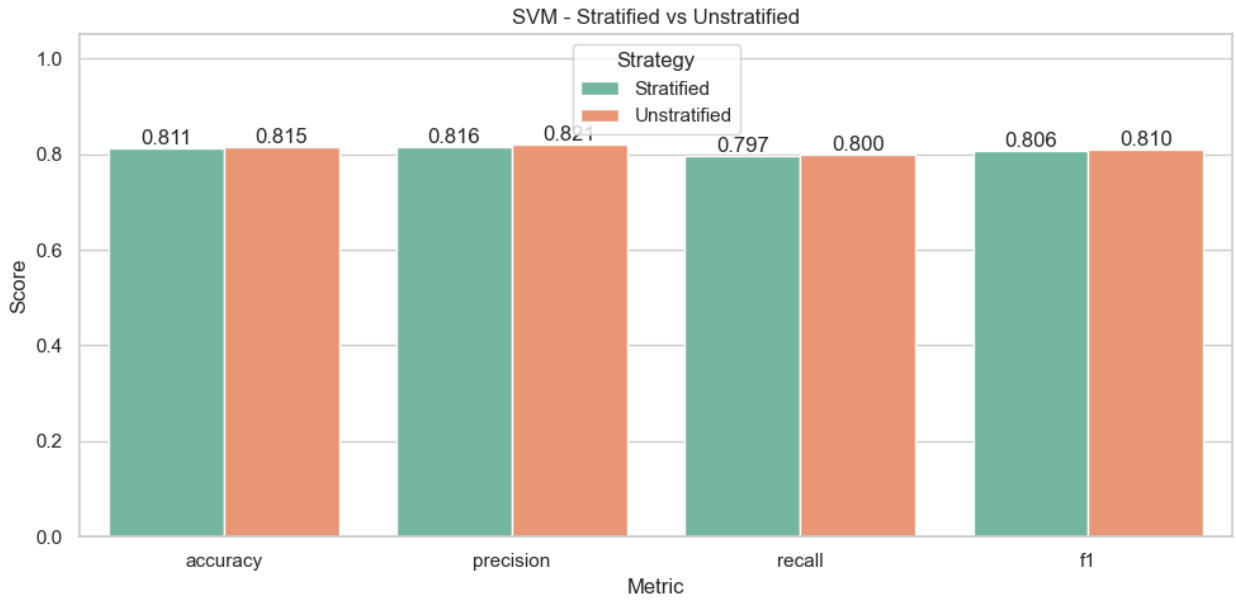
Cameron Lee
Vaibhav Garg
CS5644 Machine Learning with Big Data
29 April 2025

Assignment 2

**Compare classification performance of Bagging, Boosting, and SVM using accuracy, precision, recall, and F1-score. Plot the graphs representing each of these metrics**


AdaBoost DT - Stratified vs Unstratified
0.989
0.988
0.989
0.986
0.989
0.990
0.989
0.988
1.0
0.8
0.6
0.4
0.2
0.0
Score
Strategy
Stratified
Unstratified
accuracy
precision
recall
f1_score
Metric


Random Forest - Stratified vs Unstratified
0.973
0.973
0.969
0.970
0.976
0.976
0.973
0.973
1.0
0.8
0.6
0.4
0.2
0.0
Score
Strategy
Stratified
Unstratified
accuracy
precision
recall
f1_score
Metric




**Discuss which evaluation metric is most suitable given the dataset's class distribution**
Looking through the datasets provided, each is generally balanced with roughly 50% labeled as authentic and 50% labeled as AI generated. Because of this, accuracy is the most suitable evaluation metrics for the datasets.

**For each of the models, compare the performance between stratified 5-fold and unstratified 5-fold (only using the most suitable metric). Explain the intuition behind such difference in the performance. if one is higher than the other, you should explain why**
The table below shows the accuracy performance of the different stratified and unstratified models. Within each model, there does not seem to be much difference between the stratified and unstratified versions with them being within 0.4% of each other. I believe this is mostly due to the fact that the datasets were mostly balanced. What's interesting is that 2 of the three models – Random Forest and SVM – had the unstratified approach had a higher accuracy score than the stratified approach. It was only in the Adaboost Decision Tree model where the stratified accuracy score was higher than the unstratified approach. In most cases I would assume that stratified cross-validation would generally be preferred regardless of class imbalance. However, since the difference between the two is so minimal – within 0.1-0.4% - I may attribute this to the statistical noise that can occur with the 5-fold random variation.

| | Stratified Avg Accuracy | Unstratified Avg Accuracy |
|---|---|---|
| Adaboost Decision Tree | 98.91% | 98.81% |
| Random Forest | 97.29% | 97.34% |
| SVM | 81.15% | 81.54% |

**Explain why one model outperformed other models. Provide insights on strengths and weaknesses of each model based on your results**

Out of the different models, the Adaboost Decision Tree models performed the best, even across all other metrics. This model performed the best because it was able to learn the nuanced differences between the authentic and AI generated images by analyzing both textual features and image characteristics. Each learner corrected itself based on the previous learner's mistakes, As a result, this boosting ensemble method allowed this DT to become a more accurate classifier. The Random Forest classifier did not perform as well but still had a relatively high accuracy score – 97%. Random Forest is great to get a generalized score because it applied different independent decision trees to get their scores and then calculate the average/majority, which helps to avoid overfitting. However, it likely fell short to Adaboost DT because the trees in Random Forest were all independent from each other and didn't learn from the errors. Lastly, even though SVM can be effective in high dimensions, SVM performed the worst likely because the separation between our two classes (authentic and AI generated images) was non-linear. The complex relationship between the textual and image features couldn't be captured by the hyperplane separation as effectively, resulting in a lower performance.