Phase I Report - COMP6721 Project

1. Introduction and Problem Statement

In this project, we resolve a venue classification problem with image data. The purpose is to classify indoor scenes as belonging to one out of three categories: Museum, Library, or Shopping Mall. The classification task is resolved by blending supervised learning models (Support Vector Machine and Random Forest) with a semi-supervised learning model (Decision Tree with iterative pseudo-labeling). The data were sourced from the MIT Places2 dataset with 5000 images in each class, though a subset was utilized for evaluation and testing.

Some of the main challenges faced include dealing with a smaller labeled set to be used during the semi-supervised stage and making the models generalize well even though they are employed using flattened raw pixel features that could be less than discriminative.

2. Proposed Methodologies

Image Preprocessing:
All images were resized to 64x64 pixels and converted to RGB format. Each image was flattened into a 12288-dimensional vector (64x64x3), and pixel values were normalized to the [0, 1] range.

Model 1: SVM
A Support Vector Machine with RBF kernel was used. Hyperparameters: C=1.0, gamma='scale'.

Model 2: Random Forest
A Random Forest classifier was trained with 100 estimators, default depth.

Model 3: Semi-Supervised Decision Tree
Started with 20% labeled data. In each iteration, a Decision Tree was trained on the labeled set, and pseudo-labels were generated for the remaining data. Only predictions with confidence >= 0.85 were added back to the training set. Iteration continued until no more confident predictions remained.

3. Solving the Problem

Models were evaluated on the same stratified test set (20%).

Results Summary:

- SVM: Accuracy = 0.53, Macro F1 = 0.53
- Random Forest: Accuracy = 0.52, Macro F1 = 0.52
- Semi-Supervised DT: Accuracy = 0.35, Macro F1 = 0.34

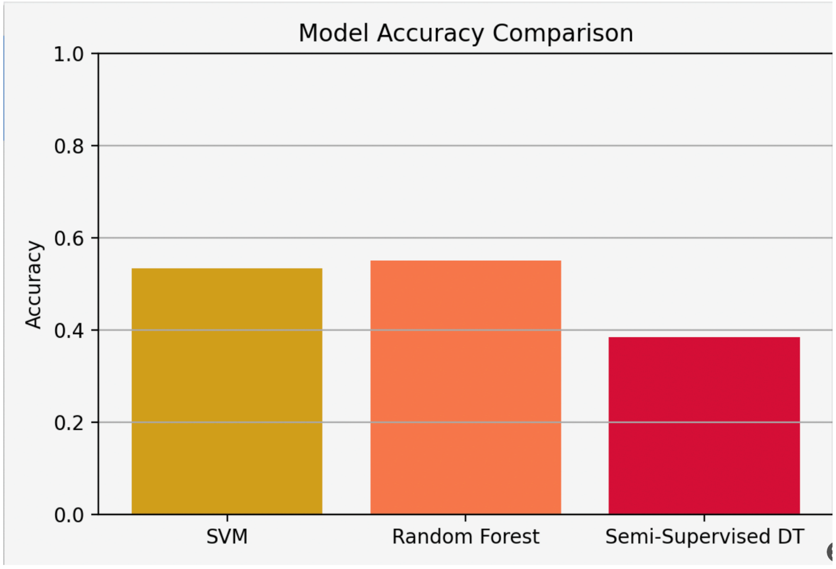

Figure: Accuracy Comparison

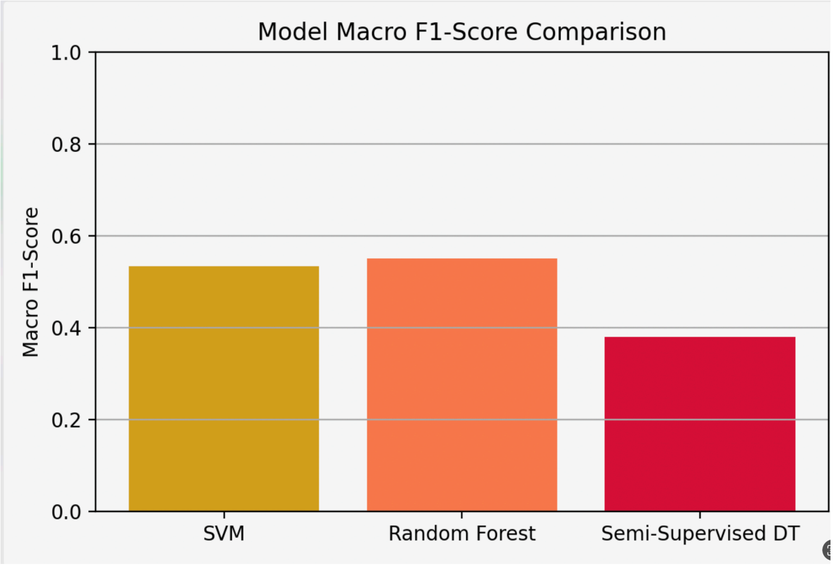

Figure: Macro F1 Score Comparison

The SVM marginally outperformed the Random Forest in both metrics. The Semi-Supervised Decision Tree performed significantly worse, likely due to overconfident early pseudo-labeling that introduced noise into training.

4. Future Improvements

- Perform hyperparameter tuning using grid search or cross-validation.
- Use feature extraction methods (e.g., color histograms, edge features) instead of raw pixels.
- Increase the confidence threshold or limit number of pseudo-labeled samples per iteration.
- Try ensemble methods combining multiple weak learners.

5. References

[1] scikit-learn Documentation - https://scikit-learn.org/

[2] MIT Places2 Dataset - http://places2.csail.mit.edu/

[3] COMP6721 Summer 2025 Project Guidelines