Technical Assessment Report

1. Introduction

Task involves working with geospatial data for a region experiencing rapid land-use changes and thus need to use the data to create a model that classifies various land covers for the region.

Approach used is the creation of a Multi Output Classifier to predict multiple target variables at once i.e. for the various targets, they also need to prediction on whether they are present
- whether there is a building or not
- whether there is cropland or not,
- whether woody vegetation cover is more than 60%

2. Objective:

To build a robust predictive model that can classify land cover into the three target categories i.e.
- Buildings, cropland and woody vegetation cover

3. Methodology

3.1 Data preparation and preprocessing

The data preparation and preprocessing for the task involved identifying the various features in the training set and the test test. This led to then finding the features not included in the test set to confirm the target as buildings, cropland and woody vegetation cover

```
pred = []
for name in train_feat:
    if name not in test_feat:
        pred.append(name)
print(pred)
```

```
['building', 'cropland', 'wcover']
```

Further preparation included identifying features that are duplicated as well as those not in the metadata. The following columns: x, y, lat and lon were identified as providing the same information i.e. coordinates of the various features and since they were not in the metadata, mlat and mlon were preferred for the task while they were among features not used in the training

Scaling the various features to standardize them thus eliminating large ranges was also implemented

3.2 Model definition & Training

The task was identified as a multi-output classification challenge involving three distinct classes as in the objective with each of the classes having two possible outputs with the model output requirement being the occurrence probabilities of the various distinct classes

Two tree based models were considered: RandomForest Classifier and xgBoost classifier with an addition of the multi-output classifier for the various classes

Model evaluation
Sklearn's accuracy score was used as the evaluation metric for the created model. The accuracy score was as follows:

- building Accuracy: 0.9426229508196722
- cropland Accuracy: 0.7831021437578815
- Woody vegetation cover Accuracy: 0.5772383354350568
-

4. Observations and Conclusion

The random forest classifier model performed better than the xgboost model thus use to make submissions on the test data
However, both model performed well on only one class of the three i.e. buildings thus suggesting that there is need to improve on it for the other two classes

Areas of improvement

- On dropping some of the columns whose metadata was unavailable, the random forest model accuracy dropped suggesting that the columns had data that is helpful to the final model
- Hyperparameters tuning was not performed for the models which can be an area to look into to improve the performance for the other two classes: woody vegetation and cropland
- The data balance distribution was not checked for the various classes which can be a source of the noted discrepancy in the model performance