Allan Wanjala Wafula

Technical Assignment - Land Cover Classification

CONTENTS

1. Exploring Data -------------------- 1

2. Processing data -------------------- 2
    2a. Handling Nulls Rows -------------------- 2
    2b. Handling Duplicates -------------------- 2

3. Finding Correlations -------------------- 2
    3a. Correlations with buildings -------------------- 2
    3b. Correlations with Cropland -------------------- 3
    3c. Correlations with Wood cover -------------------- 3

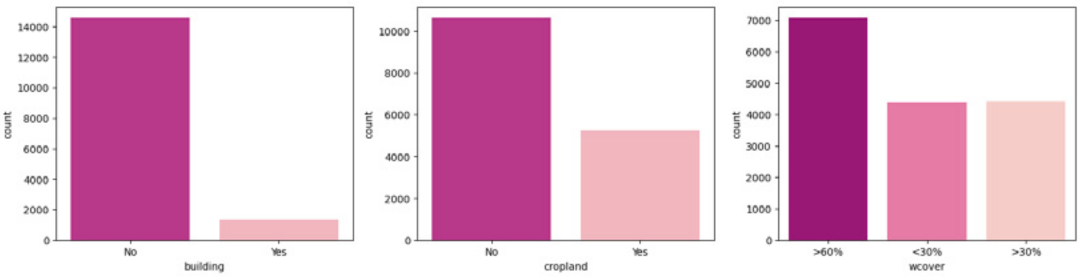
4. Machine Learning Modelling -------------------- 4
    4a. Categorical Column Encoding -------------------- 4
    4b. Data scaling -------------------- 4
    4c. Random Forest Classifier -------------------- 5
    4d. Model Predictions Sample -------------------- 5
    4e. Class Probabilities Sample -------------------- 5

5. Critical findings -------------------- 6

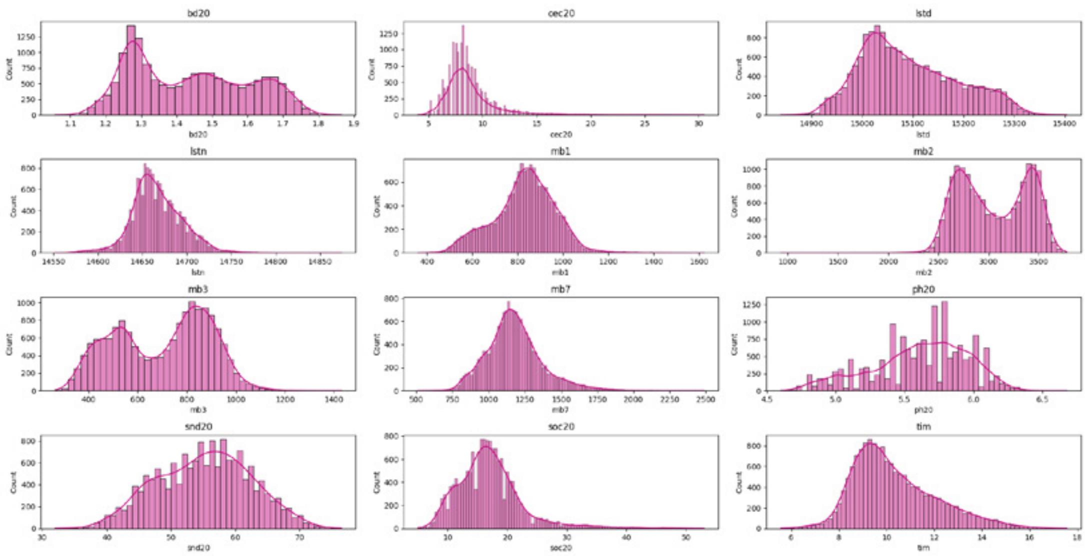
6. Recommendations -------------------- 6

1. Exploring Data

The categorical features in this dataset include Building presence, Cropland and Woody Vegetation cover. Building and Cropland are binary variables, each having two categories: Yes and No. Woody Vegetation Cover is an ordinal variable with three categories: greater than 60%, greater than 30%, and less than 30%.



The rest were all numerical sets most of which were continuous in nature having a Normal distribution (A bell-shaped curve, symmetrical around the mean), a Skewed distribution (Data is concentrated more on one side, with a long tail on the other or a Uniform distribution (Data is evenly distributed across the range of values.)



2. Processing data

2a. Handling Nulls Rows

Tobler's First Law of Geography, which states: "Everything is related to everything else, but near things are more related than distant things".

This principle was applied by using geopandas spatial join techniques plus extra post processing. This effectively imputed null values by borrowing information from the nearest available data points, under the assumption that geographically proximate locations exhibit greater similarity.

```Python
# Perform spatial join to fill nulls with nearest Spatial neighbors
joined = gpd.sjoin_nearest(empties, gdf, how='inner')
```

2b. Handling Duplicates

There we no duplicated Rows or columns after the post processing.

3. Finding Correlations

A correlation analysis was conducted (Pearson Correlation) to identify what factors are likely to have linear relationships with the occurrence of building, cropland and wood cover classes.

3a. Correlations with buildings

Variable	lcc21	bcount	mb2	bio15	npps	dor2
Correlation Strength	0.466	0.460	0.142	-0.111	-0.114	-0.118

lcc21 = Habitat humains et infrastrutures (pixel count)
mb2 = Average MOD13Q1 band 2 reflectance (2001-2021)
bio15 = Mean rainfall seasonality (CV, 1979-2013)
*npps = SD MOD17A3HGF NPP (gC/m^2/yr *0.1; 2000 - 2021)*
dor2 = Distance to any known road (km)

A positive correlation of 0.466 between "Habitat humains et infrastructures" (human settlements and infrastructure, typically referring to areas with human-built structures such as roads, buildings, etc.) and building occurrence can be explained by the fact that the presence of infrastructure generally indicates the development or expansion of human settlements.

3b. Correlations with Cropland

Variable	mb7	lstd	bio7	soc20	cec20	dor1
Correlation Strength	0.222	0.175	0.171	-0.153	-0.154	-0.169

mb7 = Average MOD13Q1 band 7 reflectance (2001-2021)
lstd = Average day-time land surface temp. (deg. C , 2001-2020)
*bio7 = Mean annual temperature range (deg. C * 10, 1979-2013)*
soc20 = Predicted topsoil (0-20 cm) organic carbon content (g/kg)
cec20 = Predicted topsoil cation exchange capacity (cmol/kg)
dor1 = Distance to major road (km)

The positive correlation of 0.222 between "Average MOD13Q1 band 7 reflectance (2001-2021)" and cropland can be understood by considering what band 7 in MODIS satellite data represents and how it relates to vegetation, particularly crops.

A positive correlation of 0.222 suggests a weak but noticeable relationship between the reflectance values from this band and the presence of cropland. In areas where croplands are present, the reflectance in band 7 tends to be higher, as crops reflect more in the near-infrared spectrum compared to non-vegetated or sparsely vegetated areas.

3c. Correlations with Wood cover

Variable	mb7	lstd	bio7	soc20	cec20	dor1
Correlation Strength	0.222	0.175	0.171	-0.153	-0.154	-0.169

fpara = Average fAPAR (2000-2021)
fpars = SD fAPAR (2000-2021)
mb2 = Average MOD13Q1 band 2 reflectance (2001-2021)
mb7 = Average MOD13Q1 band 7 reflectance (2001-2021)
lstd = Average day-time land surface temp. (deg. C , 2001-2020)

4. Machine Learning Modelling

4a. Categorical Column Encoding

Categorical columns, unlike numerical ones, represent qualities or categories (e.g., colors, types of fruit). Machine learning algorithms typically work with numbers, so we need to convert these categories into numerical representations. Ordinal encoding was applied to convert categorical features into numerical representations. Each unique category was assigned an integer, and the order of the integers reflects the inherent order of the categories.

Categorical Data

Building	Cropland	Wood cover
No	No	>60%
Yes	Yes	>30%
		>30%

Encoded Data

Building	Cropland	Wood cover
0	0	2
1	1	0
		1

4b. Data scaling

Data scaling was performed to enhance numerical stability and accelerate model convergence. By scaling features to a comparable range, we facilitate the comparison of their relative importance within the model. Specifically, a Min-Max Scaler was employed to transform the data, confining all feature values to the interval between 0 and 1.

Original Sample

snd20	soc20	tim
66.75	12.25	8.079082
51.50	14.25	9.549431
47.00	14.50	10.523131

Scaled Sample

scaled_snd20	scaled_soc20	scaled_tim
0.780899	0.151042	0.210365
0.438202	0.192708	0.332834
0.337079	0.197917	0.413936

4c. Random Forest Classifier

Random Forests, being ensemble methods of decision trees, can capture complex, non-linear relationships between features and the target variable. This was selected for this classification task. The unique identifier column, subid, was excluded from model training as it provides no predictive value. The model was trained using 45 columns (X_train) and the target variable as 3 classes (y_train), which consisted of three distinct classes.

4d. Model Predictions Sample

subid	Predicted Building	Predicted Cropland	Predicted Wcover
1548905	No	No	>60%
1548829	No	No	>60%
1548811	No	No	>60%
1548806	No	Yes	>60%
1548798	No	Yes	>60%
1548770	No	No	>60%
1548755	No	No	>60%
1548698	No	No	>60%
1548622	No	Yes	>60%
1548587	No	Yes	>60%

4e. Class Probabilities Sample

subid	building_prob	cropland_prob	wcover_prob
1548905	0.08	0.20	0.72
1548829	0.16	0.31	0.53
1548811	0.14	0.22	0.64
1548806	0.14	0.31	0.55
1548798	0.18	0.26	0.56
1548770	0.09	0.38	0.53
1548755	0.11	0.42	0.47
1548698	0.16	0.35	0.49
1548622	0.22	0.30	0.48
1548587	0.17	0.37	0.46

5. Critical findings

Based on the analysis of the data, the areas in question show distinct characteristics that contribute to their environmental and land use patterns.

Low Building Density

These areas have far fewer buildings compared to urbanized or developed regions. This suggests that they are likely to be rural or undeveloped regions with minimal human infrastructure. The low building occurrence could be due to factors such as limited urbanization, lower population density, or preservation of natural landscapes.

Limited Cropland Coverage

Cropland is also relatively scarce in these regions. With a low proportion of land dedicated to agriculture, these areas may not be suitable for extensive farming, either due to environmental factors (e.g., soil quality or water availability) or economic reasons. The low cropland presence suggests that the land is used for other purposes, possibly for conservation, forestry, or other non-agricultural activities.

Predominantly Woodland

The majority of these areas are over 60% covered by woodland. This indicates that these regions are primarily forested, which could be a result of favorable climatic conditions, land protection policies, or low human activity in the area.

6. Recommendations

Given the extensive woodland coverage, it is crucial to prioritize the conservation and protection of these forested areas. These regions may be important for biodiversity, carbon sequestration, and water regulation. Establishing protected areas, national parks, or nature reserves could help preserve the natural ecosystem.

In areas with limited cropland, there is an opportunity to encourage sustainable land use practices that respect the natural environment. Practices like agroforestry or small-scale organic farming could allow for the development of agriculture without harming the surrounding woodland.