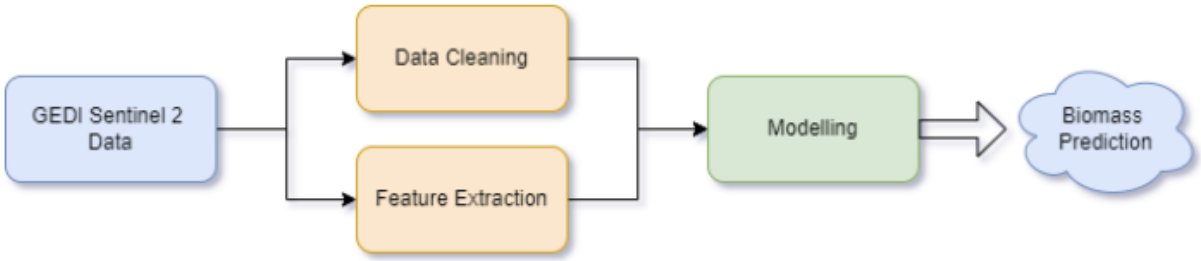
Subject: Biomass estimation of an area through satellite images.
Objective: Design an AI model to predict the biomass of an area.

Solution:



Solution Pipeline

1. Data Cleaning:

For every machine learning project or real life problem cleaning data is an important step to make the models more robust and less sensitive to outliers,
For that the best approach based on experiments was to eliminate samples that have biomass lower than 45 which helped the model to focus on the areas that contain higher values of biomass which mean extract relevant features and patterns.

2. Feature Engineering:

Sentinel 2 Data is very rich in terms of information thanks to the presence of 12 different bands. This abundance of bands provides a wide range of combinations that can be utilized to extract the most relevant features. For instance, These indices are derived from specific combinations of the available spectral bands and serve as valuable indicators for analyzing vegetation health, water content, urban areas, and other characteristics of the Earth's surface.

Vegetation indices, such as NDVI (Normalized Difference Vegetation Index), EVI (Enhanced Vegetation Index), and SAVI (Soil-Adjusted Vegetation Index), are widely used to assess vegetation density, vigor, and overall health. These indices leverage the reflective properties of different bands to quantify the presence and abundance of vegetation. They provide valuable insights into parameters like photosynthetic activity, biomass, and ecosystem dynamics.

Water-related indices, such as NDWI (Normalized Difference Water Index) and MNDWI (Modified Normalized Difference Water Index), are used to detect and monitor water bodies, analyze changes in water content, and assess water quality. These indices utilize specific band combinations sensitive to water absorption and scattering properties.

Indices related to urban areas, such as NDBI (Normalized Difference Built-Up Index) and SIPI (Structure Insensitive Pigment Index), help identify built-up regions, analyze urban

expansion, and assess urban land cover characteristics. These indices leverage the distinctive spectral signatures of urban features to provide insights into urbanization patterns.

Other indices mentioned in the code snippet, such as ARVI (Atmospherically Resistant Vegetation Index), TCARI (Transformed Chlorophyll Absorption Ratio Index), and PSRI (Plant Senescence Reflectance Index), offer specialized perspectives on vegetation characteristics, including sensitivity to atmospheric interference, chlorophyll content, and plant senescence.

Overall, these indices allow researchers and analysts to extract valuable information from Sentinel 2 data by leveraging the different spectral bands and their combinations. They contribute to a better understanding of vegetation dynamics, water resources, urbanization processes, and other important environmental factors.

3. Modelling:

For this particular task, we employed a Tree-based model called LightGBM. One advantage of using such models is that they do not require scaled data, eliminating the need for any scaling methods in our preprocessing pipeline.

To ensure robustness and avoid overfitting, we implemented a 5-fold cross-validation approach. This methodology allows our model to train and evaluate on different subsets of the data, ensuring that it gains exposure to the entire dataset. The final prediction is obtained by taking the average of the predictions from each fold, providing a more reliable and generalized outcome.

To fine-tune the model's performance, we focused on parameter optimization. Specifically, we experimented with modifying the maximum depth of the tree structure. By doing so, we aimed to address the issue of obtaining low training results while the validation results differed. Additionally, we incorporated an early stopping parameter, which terminates the training process if no significant improvement is observed, thereby mitigating the risk of overfitting.

Throughout the training process, we employed the Mean Squared Error as our loss function. This choice allowed us to measure the dissimilarity between predicted and actual values, providing a quantitative evaluation of the model's performance.

The results of our model, which reflect its predictive capabilities, are summarized in the table below. These outcomes provide insights into the accuracy and effectiveness of our trained LightGBM model.

Train MSE	Val MSE	Public MSE	Private MSE
60.25	78.37	53.64	72.58

We can see that our results on the validation data isn’t so far from the private data.