Process:

1. Feature obtaining: I obtain every spectral band of satellite images (B01, B02,..B12) in the .tiff file in order to include as much information as we can.

2. Feature Engineering: Trying to transform the feature and add some additional information to improve the data performance.
   The feature we add:
   NDVI = (B08 - B04) / (B08 + B04)
   NDBI = (B11 - B08) / (B11 + B08)
   Building_Index = B11 / B12
   Soil_Ratio = B06 / B07
   Vegetation_Health = B08 / B02

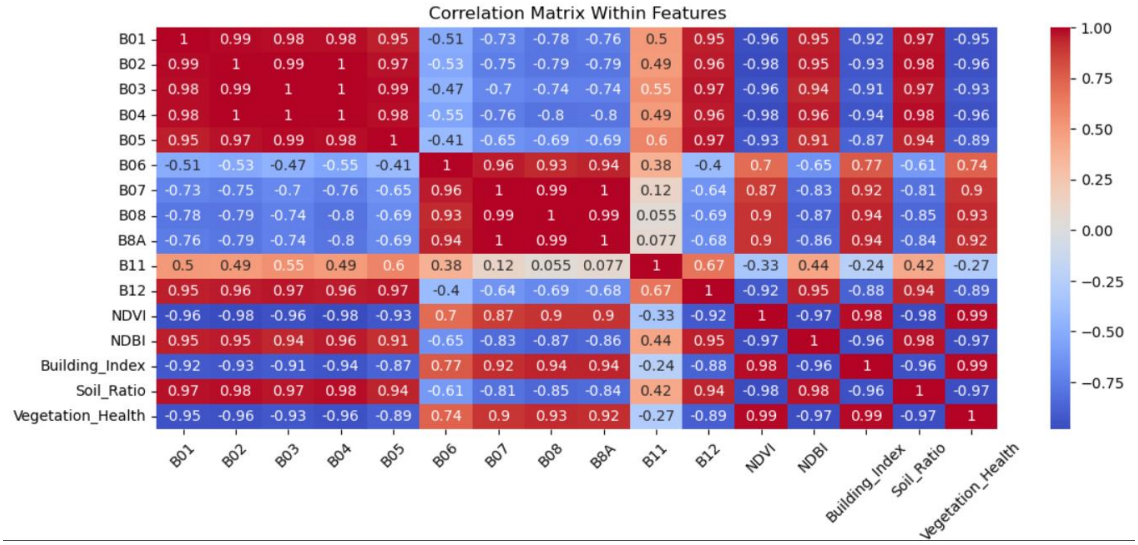
3. Model selection : we have tried random forest, XGBoost, neural network and linear regression and find that XGBoost give us the best performance of R-square on testing data, which also shows that this data is non-linear and XGBoost is good to handle non-linear data

4. Model parameter tuning:
   Important parameter: 1. n_estimators=1000, # Number of trees 2. learning_rate=0.005, 3. max_depth=20, # Tree depth (get the highest performance and also consider the time consuming while increase the complexity)
   Additionally, add the L2 regularization(Ridge) effect reg_lambda = 0.001 to avoid overfitting and mitigate the impact of redundant features

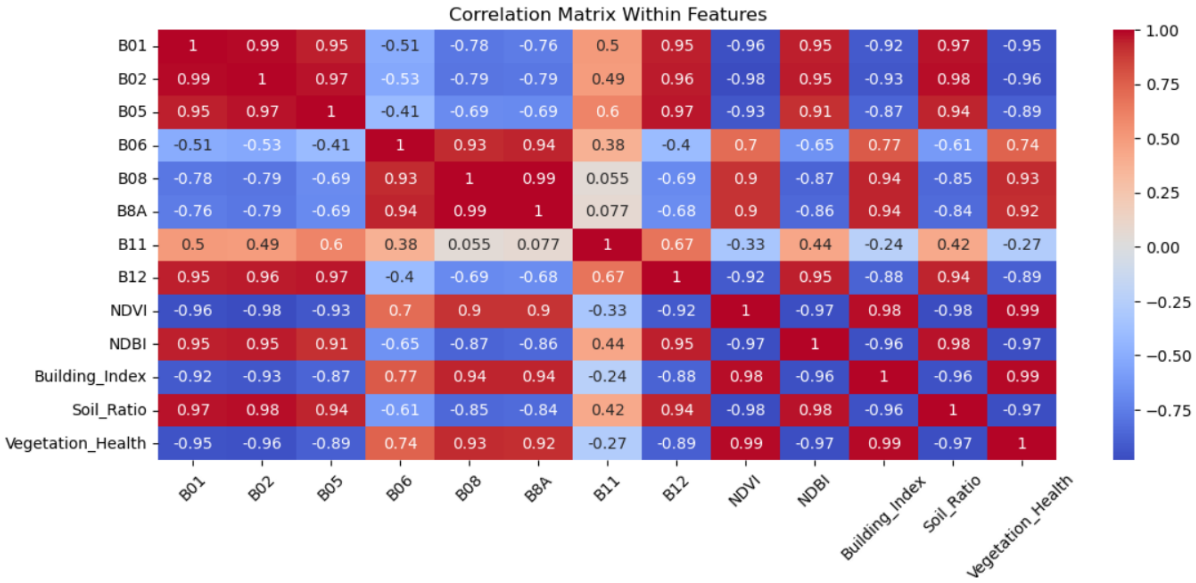
5. Using average pixels of specific range to create the buff area which are able to capture the information of surrounding area instead of relying on single pixel area, reducing the variability of pixel.(we take 2000 meter of aurrounding)
   ***Critical improvment: R-sqare (0.53 to 0.93)**
   ***Assumption: Since we take adventage of surrounding environment and UHI Index is affected by the whole area instead o single pixel. Therfore by taking the buff area, we can obtain the information of that whole environment.**

6. Features Selection: In our XGBoost model , we get the heatmap within features and we can found there are many features which has really high correlation, which may cause the Multicollinearity and undermine our model's performance(since too many redundant variables) , so we decide to take out some features from our model and inspect the r-square while doing this., make sure we do not take out the value which provide valuable information.

Before:



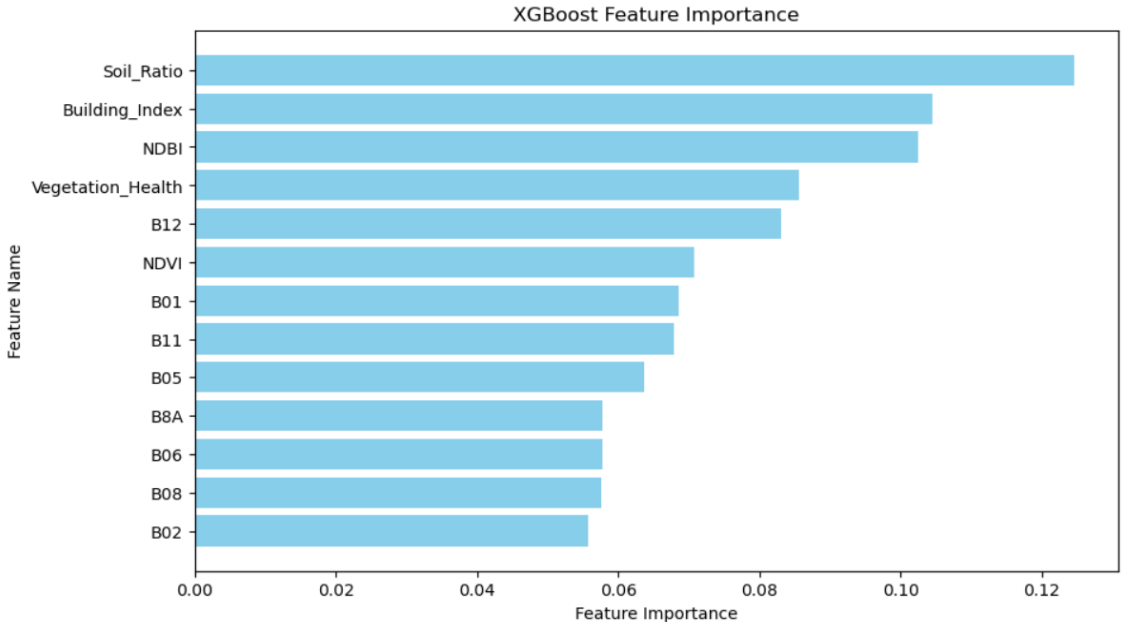
After:



advantage: by reducing the redundant variables we 1. slightly increase the R-square value(0.93 to 0.9414) and 2. improve the training time of our model(2 min to 1 min).

Final features: 'B01', 'B02','B03','B04', 'B05', 'B06', 'B07', 'B08', 'B8A', 'B11', 'B12', 'NDVI', 'NDBI', 'Building_Index', 'Soil_Ratio', 'Vegetation_Health'

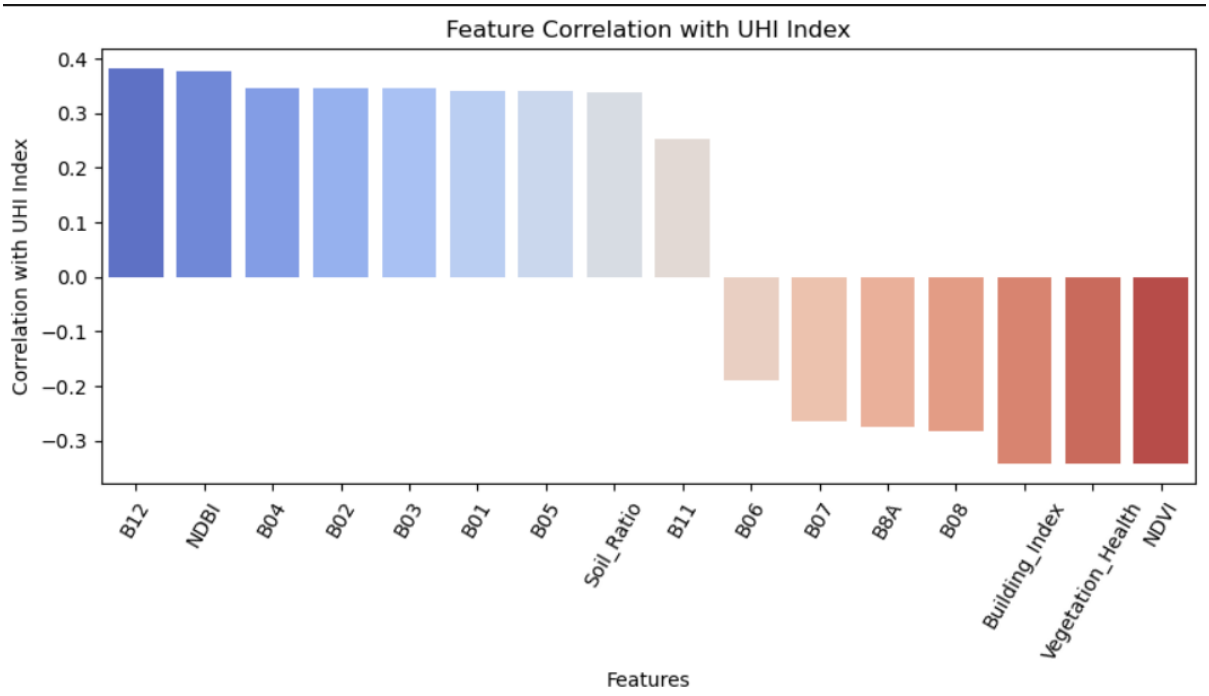
Also Check the feature importance to see which feature is more important to our model



Final Model: XGBoost, 2000 meters buff area(average of 2000 meters pixel)

Below are some feature research:

Correlation coefficient of features and UHI index



Positive correlation: water, density, building

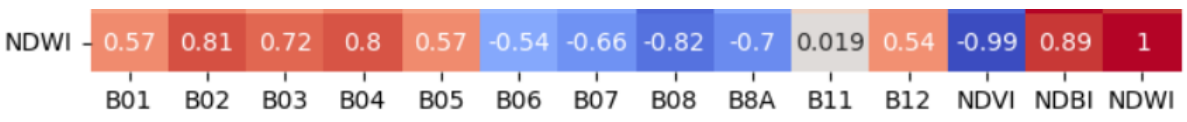
Negative correlation: vegetation

Water:

1. Water surfaces generally have a **lower albedo (reflectivity)** than light-colored urban surfaces, meaning they **absorb more solar radiation** and contribute to warming nearby areas.

2. **Cities often develop near water** sources for historical and economic reasons.

Proof:



Correlation coefficient between NDWI and NDBI is pretty high (0.89)

3. While water can provide evaporative cooling, its effectiveness depends on environmental conditions. In humid climates or areas with low wind, evaporation is less efficient, and water may not significantly cool the surrounding area.

Population Density (building as well):

1. Urban areas with **high population** density have more buildings, roads, and pavements, which **absorb and store heat** during the day and release it slowly at night.

2. These materials have **low albedo (reflect less sunlight)** and high heat capacity, making urban areas warmer than surrounding rural areas.

3. More people mean more energy consumption from **air conditioning, vehicles, industrial activities, and household appliances.** This **releases additional heat** into the environment, further increasing urban temperatures.

4. High-density areas have **tightly packed buildings**, **limiting natural airflow** and trapping heat. This "urban canyon" effect slows down cooling at night.

5. High-density urban areas tend to have **less exposed soil** or **water** surfaces, reducing evaporative cooling.

Vegetation:

1. Plants **release moisture** into the air through transpiration, which helps cool the surrounding environment.

2. Trees and vegetation **block direct sunlight**, reducing the heating of roads, pavements, and buildings.

3. Vegetation, especially green plants, **reflects more sunlight** compared to asphalt and concrete.

4. Vegetation can help break up urban heat islands by allowing better airflow and reducing heat trapping.