Process:

1. **Feature obtaining:**

   We 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(Vegetation) = (B08 - B04) / (B08 + B04)

   NDBI(Building) = (B11 - B08) / (B11 + B08)

   Moisture Content = 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. **Buff area:**

   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)

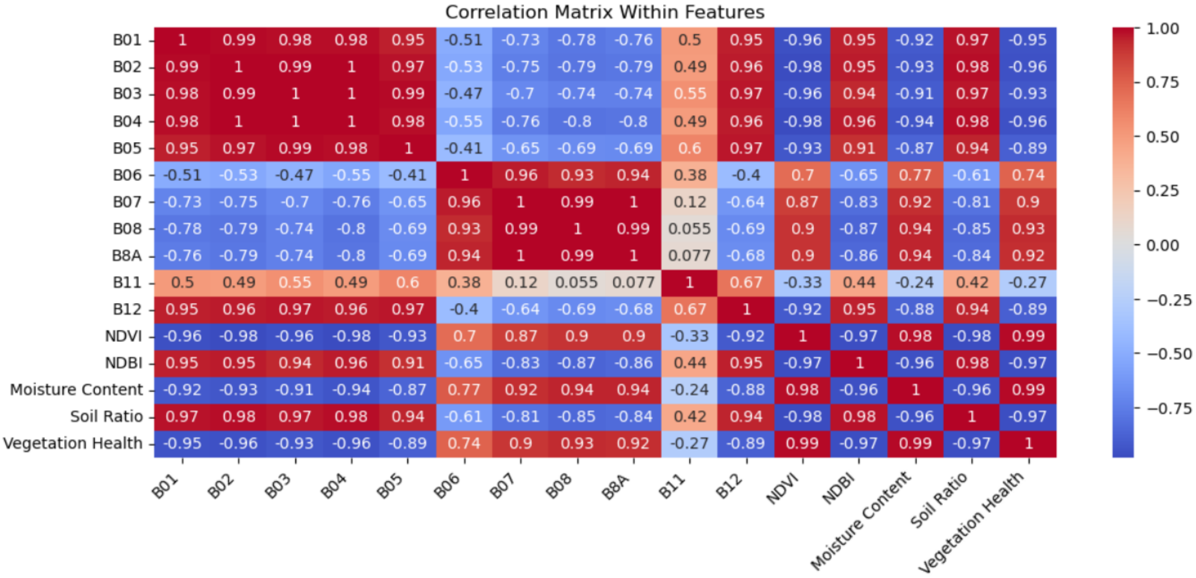
   ***Critical improvment: R-sqare (0.53 to 0.93)**

***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.**

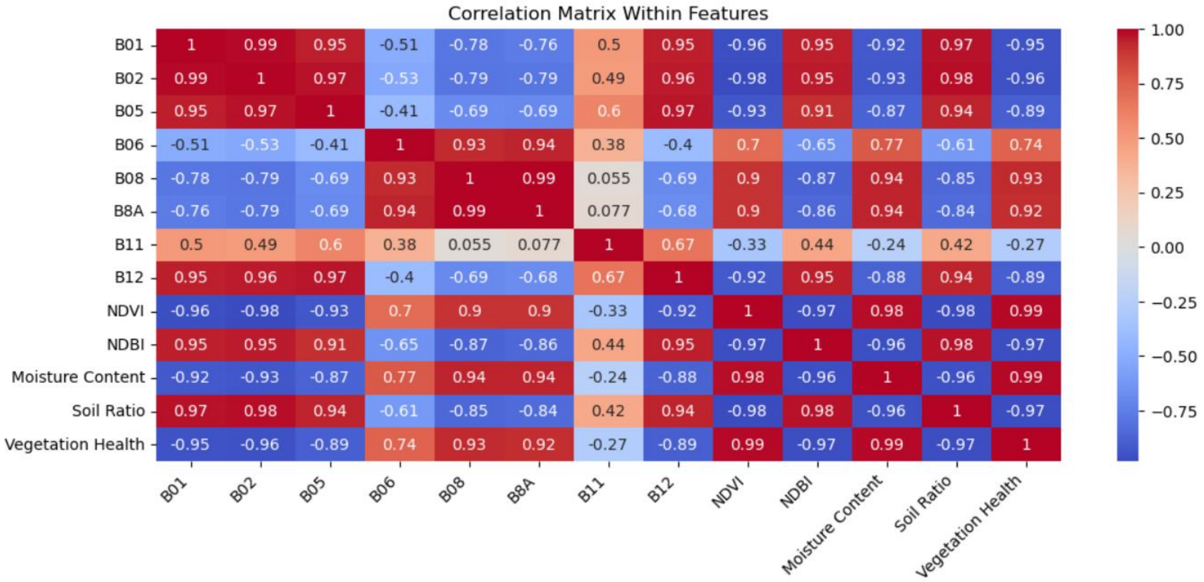
6. **Features Selection:**

   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:

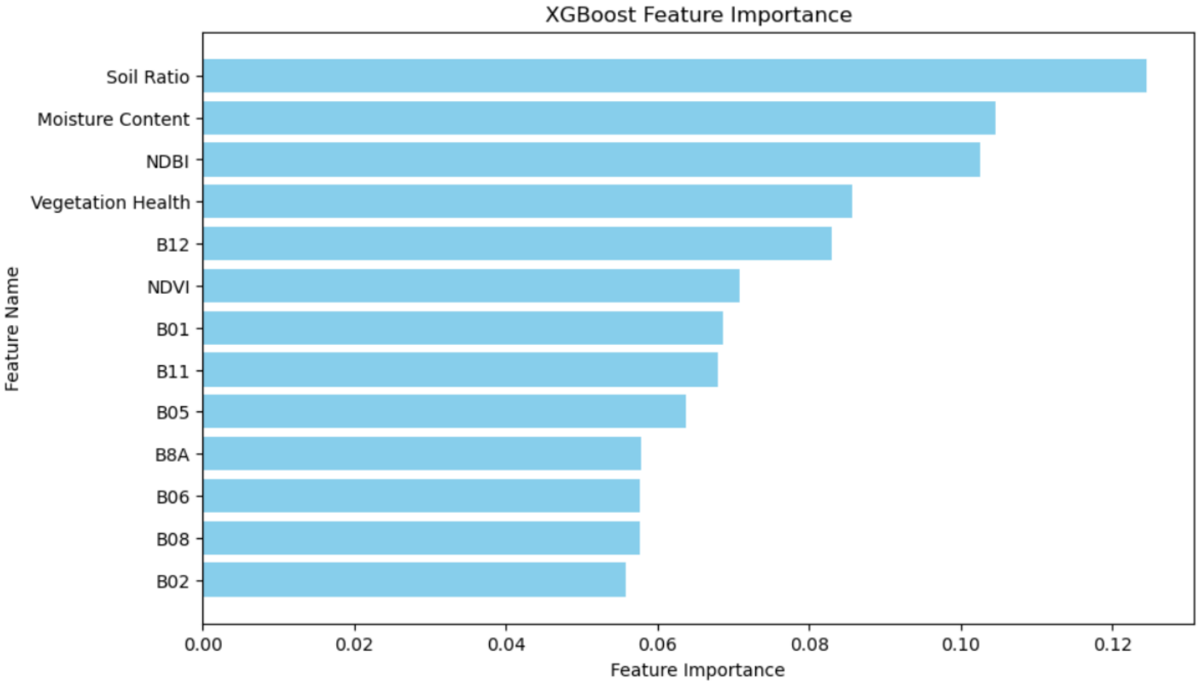


	B01	B02	B05	B06	B08	B8A	B11	B12	NDVI	NDBI	Moisture Content	Soil Ratio	Vegetation Health
B01	1	0.99	0.95	-0.51	-0.78	-0.76	0.5	0.95	-0.96	0.95	-0.92	0.97	-0.95
B02	0.99	1	0.97	-0.53	-0.79	-0.79	0.49	0.96	-0.98	0.95	-0.93	0.98	-0.96
B05	0.95	0.97	1	-0.41	-0.69	-0.69	0.6	0.97	-0.93	0.91	-0.87	0.94	-0.89
B06	-0.51	-0.53	-0.41	1	0.93	0.94	0.38	-0.4	0.7	-0.65	0.77	-0.61	0.74
B08	-0.78	-0.79	-0.69	0.93	1	0.99	0.055	-0.69	0.9	-0.87	0.94	-0.85	0.93
B8A	-0.76	-0.79	-0.69	0.94	0.99	1	0.077	-0.68	0.9	-0.86	0.94	-0.84	0.92
B11	0.5	0.49	0.6	0.38	0.055	0.077	1	0.67	-0.33	0.44	-0.24	0.42	-0.27
B12	0.95	0.96	0.97	-0.4	-0.69	-0.68	0.67	1	-0.92	0.95	-0.88	0.94	-0.89
NDVI	-0.96	-0.98	-0.93	0.7	0.9	0.9	-0.33	-0.92	1	-0.97	0.98	-0.98	0.99
NDBI	0.95	0.95	0.91	-0.65	-0.87	-0.86	0.44	0.95	-0.97	1	-0.96	0.98	-0.97
Moisture Content	-0.92	-0.93	-0.87	0.77	0.94	0.94	-0.24	-0.88	0.98	-0.96	1	-0.96	0.99
Soil Ratio	0.97	0.98	0.94	-0.61	-0.85	-0.84	0.42	0.94	-0.98	0.98	-0.96	1	-0.97
Vegetation Health	-0.95	-0.96	-0.89	0.74	0.93	0.92	-0.27	-0.89	0.99	-0.97	0.99	-0.97	1

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', 'B05', 'B06', 'B08', 'B8A', 'B11', 'B12', 'NDVI', 'NDBI', 'Moisture Content', 'Soil Ratio', 'Vegetation Health'

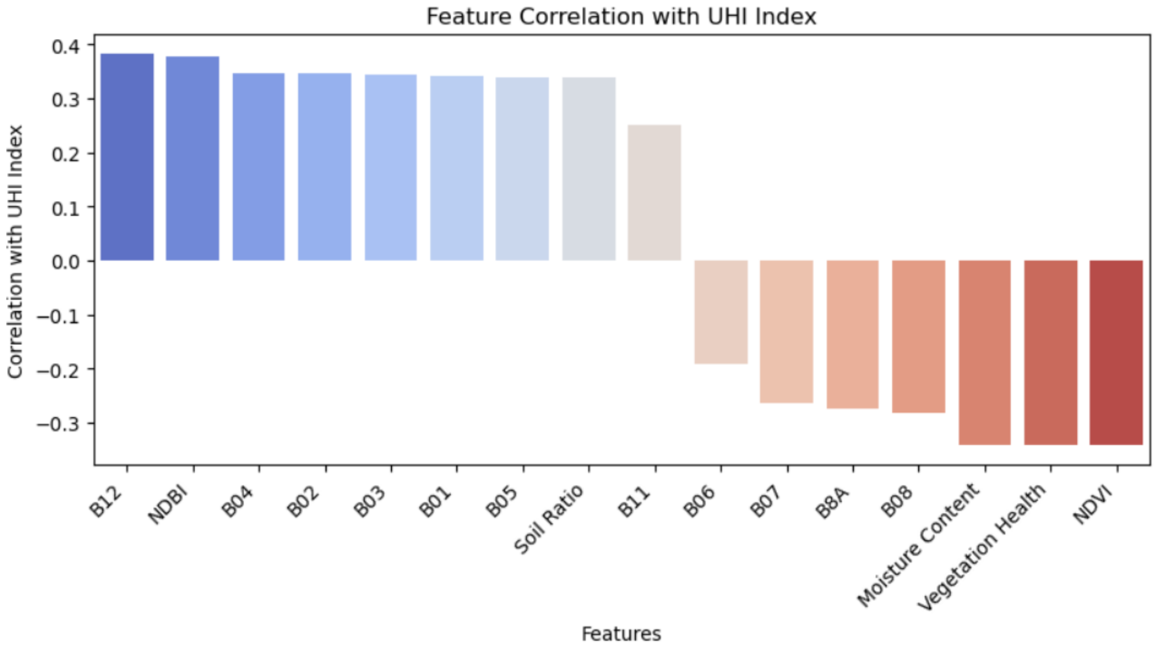
Also Check the feature importance to see which feature is more important to our model



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

Below are some features research:

Correlation coefficient of features and UHI index:



Critical Positive correlation features: Building (NDBI), Soil Ratio

Critical Negative correlation features: Moisture Content, Vegetation (NDVI), Vegetation Helth

Moisture Content (Water):

1. Water bodies **cool the surrounding air** through **evaporation**, where heat energy is used to convert water into vapor.

2. Water **absorbs and releases heat slowly** compared to land surfaces, this helps mitigate the UHI effect, leading to a negative correlation between water presence and UHI.

Soil Ratio(Bare soil coverage):

1. **Soil Stores and Releases Heat Efficiently**, dry soil has a **higher thermal capacity** than vegetated surfaces. It **absorbs heat during the day and releases it at night**, contributing to the UHI effect.

2. Bare Soil Has l**ow moisture** (less evaporative cooling), vegetated areas **release heat through evapotranspiration**. **Bare soil lacks** this **cooling process**, so more soil exposure leads to higher temperatures.

3. Soil has lower albedo, **exposed soil absorbs more heat** than forests, grasslands, or crops. This effect is especially strong in arid and semi-arid environments where soil surfaces dominate.

Building (NDBI):

1. Low Albedo of Urban Surfaces (More Heat Absorption): Urban materials like asphalt, concrete, and metal rooftops **absorb more solar radiation** and have a **low albedo (reflectivity)**, meaning they store and re-radiate heat, increasing surface and air temperatures.

2. Urban Geometry & Heat Trapping ("Canyon Effect"): High-rise buildings create urban canyons, **trapping heat and reducing wind circulation**. Narrow streets and densely packed buildings prevent heat from dissipating, keeping cities hotter than open landscapes.

Vegetation (NDVI):

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.