2025 EY Open Science AI and Data Challenge: Cooling Urban Heat Islands

1. Project description

The 2025 EY Open Science AI & Data Challenge focuses on the Urban Heat Island (UHI) effect, a phenomenon where urban areas experience significantly higher temperatures than surrounding rural regions due to dense infrastructure, limited vegetation, and human activities. In some cases, temperature differences can exceed 10°C, leading to health risks, increased energy consumption, and environmental degradation.

Vulnerable populations, including young children, older adults, outdoor workers, and low-income communities, are particularly at risk of heat-related illnesses. Prolonged exposure to extreme heat can lead to increased cases of heatstroke, respiratory diseases, and cardiovascular conditions, placing additional strain on healthcare systems.

This challenge aims to develop a machine learning model capable of predicting UHI hotspots in urban environments. Beyond identifying high-risk areas, the model will analyze the key contributing factors, such as building density, land cover, green spaces, and water bodies, to provide actionable insights for urban planners and policymakers.

This project aims to contribute to the development of cooler, more sustainable, and climate-resilient cities, for example by:

- Targeted Mitigation Strategies: By pinpointing the most affected areas, city officials can prioritize interventions such as planting more trees, implementing reflective surfaces, and optimizing green infrastructure to reduce heat retention.
- Improved Public Health: Early identification of heat-prone zones allows for better emergency planning, cooling center placements, and public health advisories, reducing heat-related illnesses and fatalities.
- Energy Efficiency: Understanding UHI dynamics can inform smart urban planning, reducing excessive energy demand for air conditioning and lowering greenhouse gas emissions.

2. Data sources

a. European Sentainel-2 optical satellite data
b. NASA Landsat optical satellite data
c. Elevation data
d. Building footprint data
e. Detailed local weather data
f. Water data

3. Satellite data

Satellite data is typically stored in multiple spectral bands, each capturing different wavelengths of light (e.g., visible, infrared, thermal). These bands provide crucial information about land cover, temperature, and vegetation.

There are two common methods to extract band values from satellite data:

- Using API Calls: Retrieve band values directly from satellite datasets via APIs, such as the planetary_computer.

- Using GeoTIFF Images: Create and download a GeoTIFF image containing the desired bands and extract the band values locally. The GeoTIFF image can represent any desired time period (single date or time series mosaic) and include any number of spectral bands.

You can select any of these approaches as per their convenience. Since our dataset is large, the API method can be time-consuming and resource-intensive. Therefore, we have opted for the second method and extracted the values for selected band. Please refer to the following sample notebook for details about the creation of the GeoTIFF image.

- Landsat_LST-2.ipynb for the landsat-8-c2-l2 satellite
- Landsat_LSToriginal.ipynb for the landsat-c2-l2 satellite
- Sentinel2_GeoTIFF-2.ipynb for the sentinel-2-l2a satellite

We then extracted the different satellite image band values for locations specified in the training dataset. It transforms coordinates, retrieves band values from the different GeoTIFF file, and organizes the extracted features into a DataFrame for further analysis and training the machine learning model.

Notes about the GeoTIFF file format

A GeoTIFF is an image file that, unlike a standard image format (e.g., JPEG or PNG), contains geospatial metadata. This means each pixel in the image is tied to real-world geographic coordinates (latitude, longitude).

Unlike a JPEG, which only stores visual information, a GeoTIFF links the image to a specific location on Earth. This makes it useful for mapping, geographic analysis, and machine learning models that require spatial awareness. This format is widely used in GIS (Geographic Information Systems), remote sensing, and geospatial machine learning applications.

For example, for a satellite image of NYC:

- A JPEG would show buildings, roads, and parks but wouldn't tell you where they are in the real world.
- A GeoTIFF would include the exact latitude and longitude of every pixel, enabling precise location-based analysis.

4. How to deal with KML file for Building footprint

A KML (Keyhole Markup Language) file is a text-based format used to store geographic data. It defines points, lines, shapes, and other features on a map using latitude and longitude coordinates. Unlike a JPEG or even a GeoTIFF, which store images, a KML file describes geographic locations, paths, and regions in a structured way. It's widely used in Google Earth, Google Maps, and GIS applications to display location-based data.

For example, if you want to mark landmarks in NYC:

- A JPEG could show a map but wouldn't contain coordinate data.
- A GeoTIFF could store an image with geospatial metadata.
- A KML file, however, would list specific locations (e.g., the Empire State Building) with precise latitude and longitude, allowing interactive mapping.

For the KML file provided we calculated the building density for the region of NYC and saved it into a GeoTIFF file to then map the values to the training dataset.

5. ML model and results

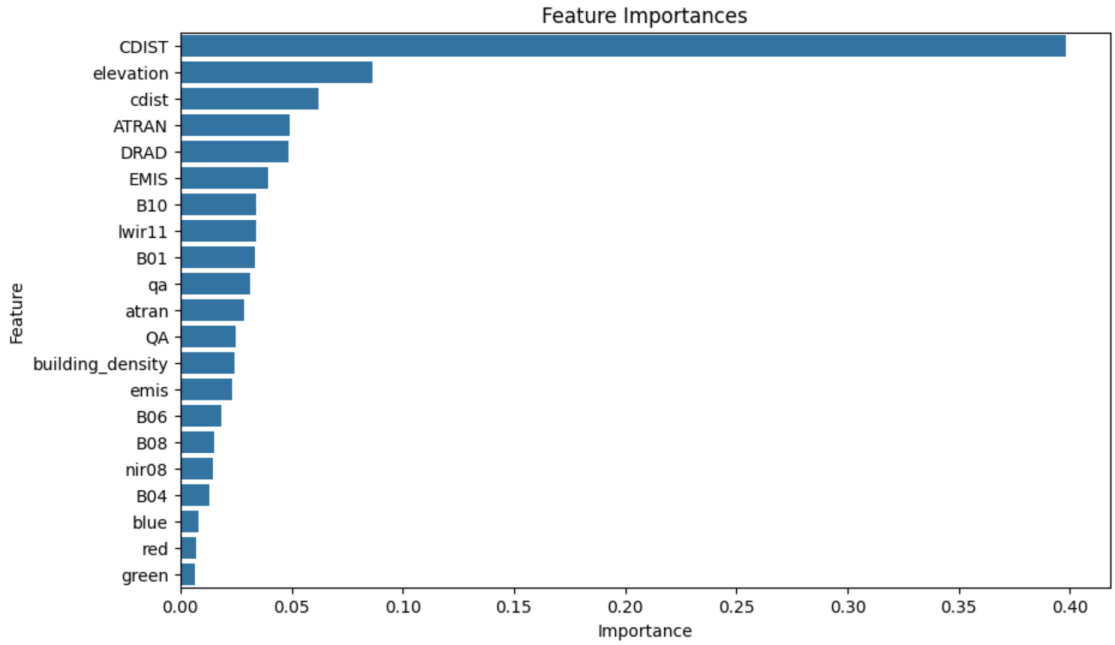
Among different ML model, it turned out that the Random Forest regression model gave the best results. We trained 300 decision trees trained on the dataset containing predictor features from the different data sources.

Model Performance (metric: R^2 Score)

- Training Data: 0.9925
- Testing Data: 0.9601
- EY Validation set: 0.9603

You can see that our model effectively captures the relationship between the selected features and the UHI index, exhibiting generalization capabilities.

Here are the different feature importance for the predictor variables:



You can see that the most important features are:

- **CDIST (Cloud Distance)** – The most important feature. This likely represents the effect of clouds on surface temperature, as areas with fewer clouds tend to have higher land surface temperatures.

- **Elevation** – Higher elevation areas generally experience lower temperatures, while lower elevation regions (urban areas) retain more heat.

- **ATRAN (Atmospheric Transmittance)** – Related to how much radiation passes through the atmosphere, impacting land surface temperature measurements.

- **DRAD (Downwelling Radiation)** – Measures incoming radiation, which directly affects surface heating and thus the UHI effect.

- **EMIS (Emissivity)** – Represents surface material properties; urban areas with concrete and asphalt tend to have higher emissivity, contributing to UHI.

6. Areas of improvements

Here are some takeaways and ways of improvements for this project:

- Instead of using satellite data from just one day that coincides with ground-based data collection, we could explore methods to enhance data quality. For example, selecting a date with minimal cloud cover or creating a median mosaic from multiple images in a time series could provide more reliable data. Additionally, we could develop a systematic method to select the best files for each satellite.
- While we conducted a lot of trial and error in selecting different bands, a better understanding of the UHI effectcould help us make more informed decisions. With deeper knowledge, we could improve our selection of both the satellite source and the specific bands or band combinations that are most effective for predicting the UHI index. Additionally, for future work, we could implement a systematic method to select the best bands for UHI prediction across different satellites, rather than relying solely on experimentation.
- We faced challenges integrating the weather dataset, as it only contained data from two locations in the NY region. Since the challenge prohibited using coordinates (longitude and latitude) for extrapolation, we couldn't extend this data spatially. Finding an alternative method for regional weather data integration could be a valuable improvement