EY Open Science AI & Data Challenge

Important FAQ:

**What will challenge participants do?**

The competition provides entrants with an opportunity to use AI for good and take part in helping build a more sustainable future for society and the planet.

Participants will be asked to develop machine learning and AI models using ground temperature data, satellite data, building footprints and heights, and weather data to identify location and severity of urban hot spots. Those who proceed to finals will be asked to develop a practical "business plan" that describes how the model summitted could be applied by local beneficiaries to help address the issue of urban heat islands, including the use of expanded datasets and analysis.

**How does the challenge process work?**

Once registered and the challenge begins (January 20, 2025), participants will be provided with many datasets to consider for their models. The ability to determine which datasets and parameters are the most important for model accuracy will determine the finalists. These datasets include:

* Ground datasets, temperature data collected by CAPA Strategies using ground traverses
* European Sentinel-2 Satellite – Spectral bands and indices to identify land classification and assess proximity to vegetation or water​
* NASA Landsat Satellite - Land Surface Temperature data to assess surface responses from buildings and land​
* Building Footprints and Height – Building locations can be used to assess urban density and building heights can be used to assess shading
* Local Weather Data – Local weather data (temp, humidity, wind speed and direction) can be used in the model

Data will be used to build machine learning models to forecast temperatures at micro-scales (meters) across a city. Models will be tested against known temperatures at specific locations to determine an overall accuracy (least-squares error). In addition, participants will be asked to submit a short document describing their analysis approach and address scaling such solutions to other cities, additional datasets that could improve model accuracy, socioeconomic impact, and practical application for local decision makers.

The challenge closes on March 20, 2025, at which time evaluations begin. Finalists will be announced on April 11 and be required to submit a content package by May 4. Winners will be announced on May 15, 2025.

**What is the evaluation methodology for the challenge?**

An out-of-sample validation dataset will be provided to participants. Submissions/predictions (.csv file), will be compared with the ground truth file and a R-squared (R²) score will be generated to evaluate the performance of the model.

**What is the required threshold to receive a certificate of completion?**

R-squared (R²) score of 0.8 on validation/submission dataset is required to receive a certificate of completion.

**Is the model of current state the main challenge deliverable, or is there an expectation that this is brought to a conclusion by solving the issue of heat islands?**

Finalists will be asked to prepare a potential implementation plan for their models. They will also be asked to record a short video explaining their implementation vision.

**Could you explain the methodology used to determine the scores on the leaderboard and the specific metrics that are employed?**

After constructing the model, you should generate predictions for the submission dataset and upload the outcomes to the platform. The platform will then evaluate the predicted results against the ground truth dataset and calculate the R-Squared score, which will be displayed on the leaderboard.

**Can you explain the process and criteria used to determine the winner?**

The selection of the winner begins with an initial screening where submissions are ranked according to their performance on the leaderboard. The top 10 participants are then asked to provide their code files, a document detailing their approach, and a video presentation of their solution. These submissions undergo a comprehensive evaluation based on both technical and conceptual criteria. The technical assessment focuses on the accuracy, methodology used, data preparation techniques, the model's ability to generalize, and the robustness of the results interpretation. For the conceptual aspect, the evaluation considers the innovation of the approach, the effectiveness of data storytelling, the potential real-world impact of the solution, and its scalability.

**Can I use Longitude and Latitude values as features for building the model?**

No, Participants are strictly prohibited from using Longitude and Latitude values as features in building their machine learning models. Submissions that employ longitude and latitude values as model features will be disqualified. These values should only be utilized for understanding the attributes and characteristics of the locations.

Incorporating latitude and longitude data, whether in their raw forms or through any form of manipulation (such as multiplication, embedding, or conversion to polar coordinates), as predictive features in your model is strictly prohibited. Utilizing these geographical coordinates in such a manner can compromise the adaptability of your model across diverse scenarios. Submissions that include these types of features will be considered non-compliant and will be disqualified.

**Could you please clarify the specific date for which we are required to generate predictions, as the submission dataset does not include a date-time column?**

We anticipate that participants will develop a model that is generalized and not confined to a specific date or time. The Urban Heat Island (UHI) index is a normalized measure of temperature that correlates more closely with a location than with the time of data collection. Therefore, the model should be capable of predicting the UHI Index for any given location, independent of the date or time. For instance, it is acceptable to select a single day (same day as mentioned in training dataset) to base your predictions for all the locations in the submission dataset.

**If we want to use cloud-based resources, are there designated locations or libraries to tap into?**

Please review the benchmark notebook for details regarding Python libraries. The first block of code includes several libraries including scikit-learn which are needed for machine learning execution.

**Are we allowed to use methods such as logistic regression, elastic net, random forest etc.?**

Any publicly available Python machine learning method is allowable. Participants should consider all such options to optimize their model performance and are free to employ any machine learning or deep learning algorithms and architectures to develop their models. However, the use of automated machine learning techniques is strictly prohibited.