CLARK
UNIVERSITY

Title: Urban Heat Island (UHI) Prediction Report
Prepared by: Chiedza Manyumwa, Aman Aman and Arrif Khan
Date: 30 November 2025

Table of Contents

Abstract

Urban Heat Islands (UHIs) are localized urban areas that experience significantly higher temperatures than their rural surroundings due to factors like dense built infrastructure, sparse vegetation, and anthropogenic heat release. This report presents a comprehensive study to predict UHI intensity at a micro-scale (meter-level resolution) using a multi-source dataset for New York City (Bronx and Manhattan) collected during a July 2021 heat event. We leverage big data analytics, integrating satellite imagery (NASA Landsat 8/9 thermal bands for land surface temperature and ESA Sentinel-2 multispectral bands for vegetation/water indices), high-resolution building footprint data, and local weather measurements. A distributed processing pipeline in PySpark was developed to handle millions of pixel-level observations and engineer features (e.g., normalized difference indices, urban morphology metrics, meteorological interactions). We train and evaluate three machine learning models Linear Regression, Random Forest, and Gradient Boosted Trees (GBT) to predict a UHI Index (ground-level air temperature relative to city mean). The GBT model achieved the best performance ($R^2 \approx 0.80$, RMSE $\approx$ 0.0072 in UHI index units), accurately identifying urban heat hotspots. Exploratory analysis confirms that high building density and low vegetation cover are strongly associated with elevated UHI. Scenario simulations indicate that increasing tree canopy or reflective surfaces can modestly reduce UHI intensity. This study demonstrates a scalable approach for urban climate modeling and provides actionable insights for urban planning and climate adaptation strategies to mitigate heat risks.

This work is written for people who need to act on heat risk: city planners, public health officials, community leaders, and infrastructure teams. At street level, heat is not an abstract number it is shade that is missing where children play, extra strain on an elderly neighbor's cooling bill, and a hazard for outdoor workers. By combining ground measurements with high-resolution satellite imagery and building data, we map where those risks concentrate and show which interventions are likely to cool places that need it most.

Technically, the project blends scalable data engineering with interpretable machine learning. The PySpark pipeline makes it practical to process millions of pixels and join them to thousands of on-the-ground temperature readings. The ensemble models capture non-linear relationships between urban form, vegetation, surface temperature, and weather. Most importantly for decision makers, the model's outputs translate into maps and simple scenarios: planting trees in a block, increasing roof reflectivity on a housing complex, or converting paved lots to permeable surfaces. Each action produces measurable local cooling in our simulations.

In short, this study turns complex data into clear, place-based guidance. It shows where heat is worst, why those places are hot, and which practical steps will reduce exposure for the people who live and work there.

Executive Summary

Urban Heat Islands (UHIs) are pockets of the city that feel noticeably hotter than surrounding areas because buildings, roads, and human activity trap and hold heat. These hotter streets and blocks aren't just an uncomfortable nuisance; they make people sicker, strain power systems as

air conditioners run harder, and worsen air quality. As cities grow and the climate warms, those risks are getting worse, and they fall hardest on people who already have the fewest resources to cope.

Because heat can change from one block to the next, we need detailed, place-level information to know where to act. This project builds that kind of detail: it combines satellite images, building maps, weather readings, and millions of pixel-level observations into a scalable data pipeline and uses machine learning to pinpoint the city's hottest spots. The result is a practical map of where heat is most dangerous and which local interventions like planting trees or using reflective surfaces are likely to help people the most.

Data & Approach: We combined **ground-based UHI index measurements** (11,229 locations across Bronx and Manhattan on July 24, 2021) with a rich array of features: **satellite-derived indicators** (land surface temperature from Landsat, vegetation and water indices from Sentinel-2), **urban morphology metrics** (building density and footprint area from GIS building data), and **local weather variables** (temperature, humidity, wind, solar radiation from nearby stations). A cloud-ready **PySpark** workflow was used to ingest large satellite imagery files and perform distributed feature extraction and data merging. We engineered features such as vegetation indices (NDVI, etc.), a humidity-temperature interaction term, and building-to-greenery ratios to capture key UHI drivers. Three predictive models were trained – a multiple linear regression as a baseline, a random forest, and a gradient boosted trees (GBT) model – using these features to predict the UHI index at each location.

Key Results: The **Gradient Boosted Trees model** achieved the highest accuracy in predicting UHI intensity, with an R^2 of about **0.80** and substantially lower error (RMSE ~**0.0072** in UHI index units) compared to the linear regression (which explained virtually none of the variance, R^2 ~0.005). The random forest also performed well (R^2 ~0.73) but was slightly less accurate than GBT. These results indicate that non-linear ensemble models can effectively capture the complex interactions among environmental and urban features that drive UHIs. The **best model's predictions** align closely with observed hotspot patterns, confirming the validity of the features and approach.

Key Contributions: This project demonstrates:

- **High-resolution UHI mapping** — We combined multiple data sources to predict local heat at finer than 30 m resolution, producing maps that show which streets and blocks are hottest rather than only which broad neighborhoods are warmer. These granular maps make it possible to design interventions that reach the exact places and people who need cooling.
- **Big data and scalable processing** — We built a cloud-ready PySpark pipeline that processes millions of satellite pixels and links them to building and weather data. The workflow is distributed and reproducible, so the same approach can scale across an entire city or be adapted to other cities without losing performance.
- **Actionable insights for planning and adaptation** — The model not only locates hotspots but also explains why they are hot: tightly packed buildings and low vegetation consistently drive higher local temperatures. Scenario tests show that planting trees and adopting cool roofs or reflective pavements produce measurable cooling at the block level, giving planners concrete options to reduce heat exposure.

- **Practical framework for cities** — By turning open data and machine learning into usable maps and scenarios, this work gives cities a repeatable tool to identify where heat is worst, understand its causes, and prioritize interventions that protect people and infrastructure.

In summary, this research integrates open-source data and AI techniques to predict urban heat islands at unprecedented resolution. It offers a framework that cities and practitioners can use to better understand UHI drivers and implement effective heat mitigation strategies, ultimately enhancing urban resilience against extreme heat

Introduction

Urbanization alters the land surface and atmospheric properties of cities, often causing urban centers to be significantly warmer than their rural surroundings. This phenomenon is known as the **Urban Heat Island (UHI) effect**. UHI intensity can exceed 10°C in large metropolitan areas under certain conditions, leading to *higher nighttime temperatures, more frequent heat waves*, and elevated heat stress for urban residents. Major contributing factors include the abundance of **heat-absorbing surfaces** (e.g. dark asphalt, concrete) that store solar energy during the day and release it as heat, the **lack of vegetation and water bodies** which would otherwise provide evaporative cooling, and **anthropogenic heat emissions** from vehicles, industrial activities, and air conditioning units. The effects of UHIs are wide-ranging: increased heat-related morbidity and mortality, higher electricity demand for cooling, accelerated formation of smog and air pollution, and exacerbation of social inequities as vulnerable communities (e.g. the elderly, low-income neighborhoods) often experience the worst heat exposure.

Climate change is expected to further amplify UHI effects. Global warming leads to more frequent and intense heat waves, and when combined with expanding urban populations, more people will be exposed to dangerous urban heat. Thus, mitigating UHIs has become a critical component of urban climate adaptation. Urban planners and city managers require accurate, fine-scale models of UHI distribution to pinpoint hotspots and design interventions (like planting trees, cool roofs, or improved ventilation corridors between buildings).

Despite extensive study of UHIs in climate science, traditional approaches often rely on coarse-resolution satellite measurements of surface temperature (e.g., from NASA's MODIS or Landsat at 1 km to 30 m resolution). While useful, these measures do not reflect the **near-surface air temperatures** that directly impact human health. Moreover, city-wide averages can mask critical local variations. There is a pressing need for **micro-scale UHI models** that utilize detailed data (such as building geometries and high-resolution imagery) to predict temperature differences at the neighborhood or block level. Recent technological advances in data availability and computing allow us to integrate heterogeneous datasets – including ground sensor networks, high-resolution Earth observation data, and GIS urban morphology data – and apply machine learning to better predict and understand UHIs.

This report addresses these needs by developing a data-driven UHI prediction model for New York City (specifically Manhattan and the Bronx) as part of the 2025 EY Open Science AI & Data Challenge. We use an **open-source multi-dataset approach**: combining **ground-level air temperature readings** converted into a UHI index, **building footprint data**, **Sentinel-2 optical satellite imagery**, **Landsat 8/9 thermal imagery**, and **local weather station data**. The primary goal is to accurately predict UHI intensity (locations of hotspots and cool spots) at meter-scale resolution across the city, and to identify the key drivers among the input features. A secondary goal is to provide insights for practical mitigation – showing how changes in factors like vegetation cover or surface reflectivity could reduce urban temperatures.

By leveraging modern **machine learning** techniques and **big data analytics**, this study demonstrates how large-scale environmental data can be harnessed to solve a pressing urban sustainability challenge. The following sections detail the data sources, methodology, exploratory analysis, modeling results, and recommendations for urban heat mitigation.

Literature Review

Physical Drivers of UHI: Prior research has established that the physical composition of cities is the fundamental cause of the UHI effect. Urban materials like **asphalt and concrete** have high heat capacities and thermal conductance, causing them to absorb and retain heat during the day. Sparse **vegetation cover** in cities reduces shade and evapotranspiration cooling. These factors make urban surfaces significantly warmer than natural landscapes. Urban geometry also plays a role: **dense high-rise buildings and narrow street canyons** trap heat and block wind flow, creating microclimates where heat is less easily dissipated (the "street canyon" effect). Additionally, **anthropogenic heat** release from vehicles, industrial processes, and building HVAC systems directly warms the urban atmosphere. Taken together, these physical and urban factors create an environment where heat accumulates, especially during calm, clear conditions (Oke, 1982).

Environmental and Socioeconomic Factors: UHI intensity is not only a function of urban form but also varies with environmental context and socioeconomics. For instance, low-income or highly paved neighborhoods often have fewer trees and green spaces, exacerbating heat exposure (Stone, 2012). Lack of air conditioning or resources to mitigate heat can make certain populations more vulnerable. While this study focuses on physical predictors like land cover and building density, one should note that socioeconomic drivers (housing quality, land use patterns) and **climate change trends** (rising baseline temperatures) are important broader contexts for UHI impacts.

Remote Sensing for UHI: Satellite remote sensing provides invaluable data for studying UHIs. **Thermal infrared (TIR) sensors** on satellites like Landsat can measure **Land Surface Temperature (LST)**, which serves as a proxy for urban heat distribution. Landsat 8/9, with ~30 m resolution in TIR (Band 10/11), allows us to observe temperature differences at the neighborhood scale. Meanwhile, **multispectral optical data** from satellites such as Sentinel-2 (10–20 m resolution) enable computation of various indices that relate to land cover characteristics. For example, the **Normalized Difference Vegetation Index (NDVI)** indicates the presence of vegetation (higher NDVI = more greenness), which typically correlates with cooler temperatures. The **Normalized Difference Built-up Index (NDBI)** highlights built surfaces, and the **Normalized Difference Water Index (NDWI)** can indicate moisture or water content in surfaces – useful since wet or irrigated areas tend to be cooler. **Albedo** (surface reflectivity) can also be derived from visible bands; higher albedo surfaces reflect more solar radiation and remain cooler. By integrating multi-source satellite data, researchers can achieve a more detailed picture of UHI drivers at fine scales. In this study, we use Landsat-derived LST together with Sentinel-2-derived NDVI, NDBI, NDWI, and other indices, capitalizing on the strengths of each platform.

Machine Learning Approaches: Traditional statistical models of UHI (e.g., OLS regression on a few variables) often fail to capture the complex, non-linear interactions in urban climates. **Machine learning (ML)** provides more powerful tools for prediction and pattern recognition in UHI studies. We consider three ML approaches: (1) **Linear Regression** as a baseline – easy to interpret but limited to linear relationships; (2) **Random Forests**, an ensemble of decision trees that can capture non-linear effects and feature interactions without parametric assumptions; and (3) **Gradient Boosted Trees (GBT)**, which build an additive model of decision trees in sequence, each correcting errors of the previous, often achieving superior accuracy on structured data. Random forests and GBTs have been successfully applied in recent UHI research (e.g., *Zhou et*

al., 2019 used tree-based models to improve UHI predictive performance). These methods handle high-dimensional data well and provide **feature importance** measures to interpret which factors most influence the predictions. However, they require careful tuning to avoid overfitting. We employ cross-validation to optimize model hyperparameters and ensure generalization. The advent of these ML techniques, combined with large environmental datasets, has markedly improved our ability to model UHIs in a data-driven manner.

Big Data Analytics: UHI modeling at city scale involves **massive datasets** – high-resolution satellite images consist of millions of pixels, and integrating multiple data sources (imagery, GIS layers, sensor data) can be computationally intensive. Therefore, big data tools are crucial for efficient processing. **Distributed computing frameworks** like Apache Spark allow data to be partitioned and processed in parallel across multiple CPUs or nodes. In this project, we use **PySpark** (Spark's Python API) to handle large image datasets and join them with spatial and weather data. Techniques such as the MapReduce paradigm and PySpark DataFrames enable transformations (filtering, aggregation, feature extraction) on very large datasets in a scalable manner. Integration with cloud platforms (e.g., deploying on AWS or Azure) can provide the necessary computing resources for big cities. By designing our pipeline with these technologies, we ensure that our approach can scale beyond this case study to other cities and larger datasets. This reflects an emerging trend in urban climate studies: the fusion of **geospatial big data** with AI to produce operational tools for city planners.

Data Sources

Our analysis integrates several **open-source datasets** covering UHI target measurements and key explanatory features. All data pertain to **New York City**, focusing on the Bronx and Manhattan areas, and were collected or sourced around the **afternoon of July 24, 2021**, during a citywide heat mapping campaign:

- **Ground UHI Index Data:** The target variable is a **UHI Index** provided by Climate Adaptation Planning & Analytics (CAPA Strategies) through an on-ground campaign (the "Heat Watch" program). CAPA collected mobile traverse measurements of near-surface air temperature across city streets between 3:00–4:00 PM on July 24, 2021. The raw measurements were converted into a dimensionless **UHI Index = (Local air temperature) / (Citywide mean temperature)**. Thus, a value of 1.0 indicates a location at the city's average temperature, >1.0 indicates a hotspot (warmer than average), and <1.0 a cool spot. The dataset contains **11,229 data points** (latitude-longitude coordinates with UHI index values) covering large portions of the Bronx and Manhattan. These serve as the ground-truth target for model training and evaluation.

- **Satellite Imagery (Landsat 8/9):** To characterize surface thermal properties, we used **Landsat 8/9** satellite data. Specifically, the **Thermal Infrared Sensor (TIRS)** aboard Landsat provides Land Surface Temperature (LST) at 30 m resolution. We obtained georeferenced *Landsat thermal band images* (Band 10 and Band 11) from summer 2021 (close in date to the heat event) to derive LST. Using standard algorithms (NASA's emissivity and radiative transfer models), each pixel's digital number was converted to a temperature in °C. LST captures the "skin" temperature of surfaces, which influences and correlates with near-surface air temperatures. In addition, we used Landsat's reflective bands (30 m) to derive other indices: for example, Landsat Band 4 (red) and Band 5 (near-

infrared) could alternatively be used for NDVI calculation (though in practice we relied on higher-resolution Sentinel-2 for NDVI, as described next).

- **Satellite Imagery (Sentinel-2):** The European Space Agency's **Sentinel-2A/B** provides high-resolution optical imagery (10–20 m) useful for land cover analysis. We downloaded Sentinel-2 Level-2A images (providing surface reflectance) for the NYC area on dates near July 24, 2021 (cloud-free). Four spectral bands were utilized: **Band 4 (Red, 10 m)**, **Band 8 (Near-Infrared, 10 m)**, **Band 6 (Red-Edge, 20 m)**, and **Band 11 (Shortwave Infrared, 20 m)**. These were selected to compute key **spectral indices**:

- **NDVI (Normalized Difference Vegetation Index):** computed as (NIR – Red) / (NIR + Red). NDVI ranges from -1 to 1, with higher values indicating dense vegetation. Green spaces and tree canopy typically yield high NDVI and are associated with cooler conditions.
- **NDBI (Normalized Difference Built-up Index):** computed as (SWIR – NIR) / (SWIR + NIR). Higher NDBI indicates impervious, built surfaces (concrete, asphalt). Built-up areas often correspond to UHI hotspots.
- **NDWI (Normalized Difference Water Index):** here defined using NIR and SWIR as (NIR – SWIR) / (NIR + SWIR). This index highlights moisture content; higher NDWI can indicate water or lush vegetation. In our dataset, NDWI is essentially the inverse of NDBI (since it uses the same bands in opposite order), providing a complementary perspective on surface moisture vs. dryness.
- **Albedo:** We derived an approximate surface albedo (reflectance) by combining Sentinel bands (blue, red, NIR, SWIR) using literature formulas. Higher albedo surfaces (e.g., white roofs) reflect more sunlight and tend to remain cooler.
- **Other indices:** We also computed a **Normalized Difference Soil Index (NDSI)** to capture bright, reflective surfaces (originally used for snow/ice detection, here it may highlight concrete or dry surfaces), and a **Robust Enhanced Vegetation Index (REVI)** which is a modified vegetation index intended to be less sensitive to atmospheric effects.

- **Building Footprint and Urban Morphology Data:** To quantify the urban built environment, we used detailed **GIS building footprint data** for NYC (e.g., from city open data or Microsoft's building footprint repository). From the vector layer of building polygons, we extracted features at multiple spatial scales around each UHI measurement point:

- **Building Density:** the number of buildings within a given radius of the point. We calculated building counts within **50 m, 100 m, and 200 m** buffers of each location (denoted as building_density_50, building_density_100, building_density_200).
- **Building Footprint Area:** the total area of building footprints within 100 m and 200 m radii (building_area_100, building_area_200), in square meters. This measures the bulk of built structures surrounding the location.
- **Coverage Ratio:** the fraction of land area covered by buildings in a 100 m radius (coverage_ratio_100). This is directly related to building_area_100 (since it's building area divided by the circle area of radius 100 m). A higher coverage ratio means less open/green space in the vicinity.

- **Distance to Nearest Building:** for each point, we found the distance (in meters) to the closest building structure (distance_to_building). Points in open parks or waterfronts, for example, would have larger values.

These features capture urban form at different scales, which is crucial since UHI can be influenced by both immediate surroundings and neighborhood-scale development. For instance, a point located in a dense high-rise district (high building_density and coverage) is expected to experience stronger UHI effects than one in a park or low-rise residential area.

- **Local Weather Data:** We incorporated meteorological variables from the time of the heat mapping. Two nearby weather stations (one representative of Manhattan conditions, another for the Bronx) provided measurements; each UHI data point was associated with its **nearest station** reading (nearest_station is either "Manhattan" or "Bronx"):
- **Air Temperature:** (Used indirectly in the UHI index calculation, so not included as a separate feature since the UHI index already encapsulates the relative temperature).
- **Relative Humidity:** (humidity) in %. Humidity influences evaporative cooling and how temperatures are felt (higher humidity can exacerbate heat stress by hindering sweat evaporation).
- **Wind Speed and Direction:** (wind_speed in m/s and wind_direction in degrees from north). Wind can disperse heat; low wind conditions often coincide with higher UHI intensity. Wind direction might differentiate areas downwind vs. upwind of heat sources.
- **Solar Flux:** (solar_flux in W/m²) – the solar radiation at ground level. This was around ~600 W/m² during the 3-4 pm period. We include it to account for any intra-urban variations in cloud cover or shading that could affect local heating (though during the campaign time it was mostly clear skies).

By merging these diverse datasets, we obtain a rich dataframe of **features** for each of the 11,229 locations, along with the target UHI index. Table 1 summarizes the data sources and their contributions:

Table 1: Data Sources and Features

Source	Example Data/Resolution	Features Derived
Ground traverse (CAPA)	11,229 points (3–4 pm, 7/24/2021)	UHI Index (target), Location (lat, lon)
Landsat 8/9 (Thermal)	30 m pixels (TIR bands)	Land Surface Temperature (LST) (°C)
Sentinel-2 (Optical)	10–20 m pixels (multispectral)	NDVI, NDBI, NDWI, NDSI, Albedo, REVI, raw bands (B01, B04, B06, B08, B11)
Building Footprints (GIS)	2D polygons of buildings	Building_density_50/100/200, Building_area_100/200, Coverage_ratio_100, Distance_to_building
Weather Stations	Point measurements (hourly)	Wind_speed (m/s), Wind_direction (deg), Humidity (%), Solar_flux (W/m²)

All datasets are spatially referenced (WGS84 geographic coordinates) and were integrated using spatial joins and nearest-neighbor matching (for weather data). The combined dataset uhi_data_final.csv contains 28 features (columns) for each point, which serve as inputs for modeling the UHI index.

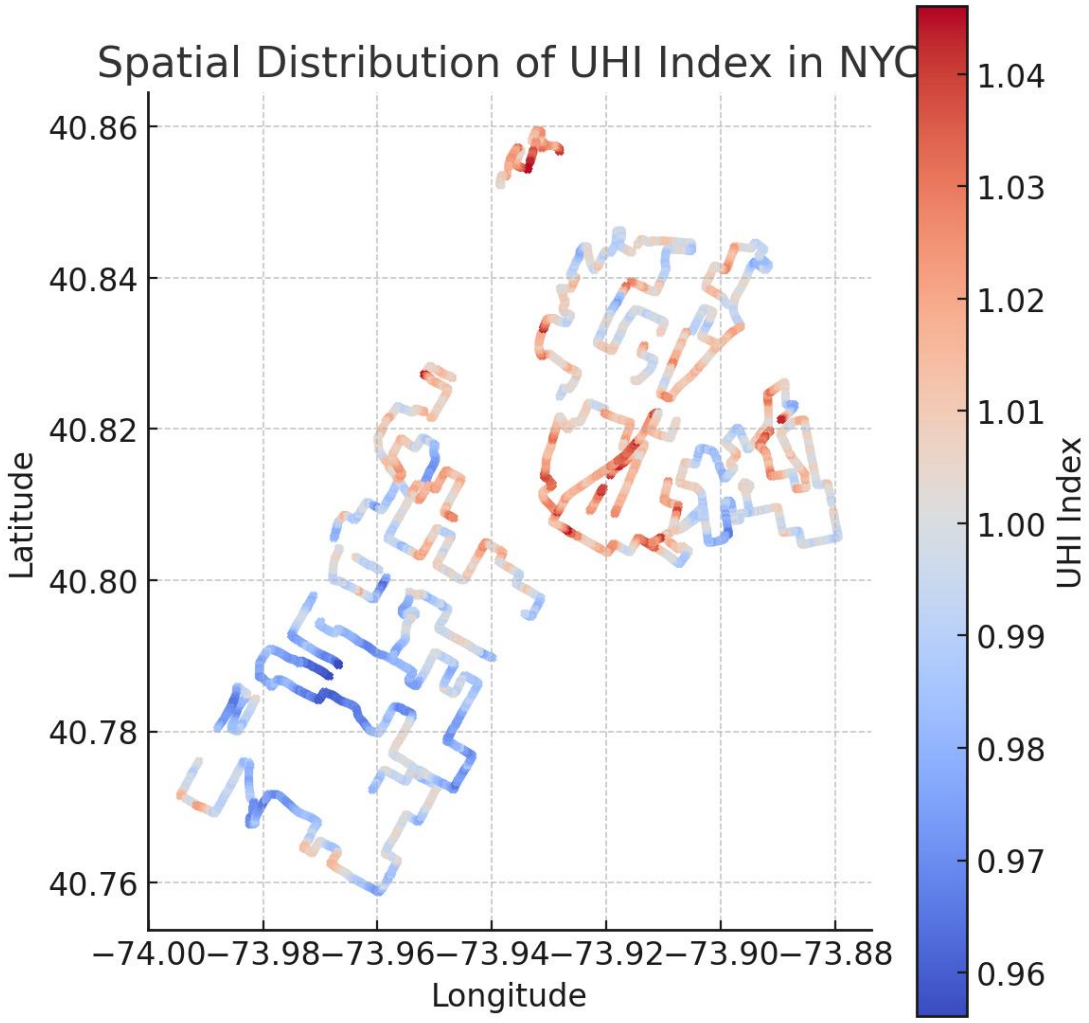

Figure 1: Spatial distribution of the UHI index measurement points in New York City (Bronx and Manhattan). Blue points represent **cooler locations** (UHI Index < 1.0) and red points represent **hotspots** (UHI Index > 1.0). The Bronx (upper cluster ~40.82–40.86°N) generally exhibits higher UHI index values (more red points) compared to Manhattan (lower cluster ~40.76–40.80°N) in this afternoon dataset. This suggests the Bronx areas surveyed were relatively hotter than the city average, whereas parts of Manhattan, possibly influenced by proximity to waterways or more high-rise shade, were slightly cooler. Such spatial patterns underline the importance of localized features – Bronx sites may have had fewer green spaces or more expansive heat-retaining surfaces, leading to stronger UHI signals.

Methodology

To predict UHI intensity from the above data, we designed a **data processing pipeline** and modeling workflow following standard analytics practices (CRISP-DM methodology: data

preparation, modeling, evaluation, deployment). The emphasis was on ensuring scalability (using big data tools) and robust model development (using cross-validation and feature engineering). Key steps in our methodology include data preprocessing, feature engineering, model training, and evaluation:

Data Ingestion & Preprocessing: Given the volume and variety of data, we utilized **PySpark** for data ingestion and wrangling. The satellite images (Landsat and Sentinel-2) in TIFF format were read using a distributed approach (leveraging libraries such as GDAL and RasterFrames to convert imagery into Spark DataFrames). Millions of pixel values were processed in parallel across the cluster. We then performed spatial joins: each UHI measurement point (lat, lon) was associated with the corresponding pixel values from the Landsat and Sentinel images (by finding the pixel in which the point falls, using nearest-neighbor for any slight misalignments). Similarly, building footprint metrics were computed by Spark GIS operations (e.g., using a spatial join or custom radius search for each point to aggregate building counts/areas). The Manhattan or Bronx weather data were joined by assigning each point the values from the nearest weather station. This ingestion pipeline can be summarized as follows:

- *Step 1:* **Load raw data** – Read satellite imagery into DataFrame structures (one row per pixel with coordinates and band values), load building polygons and weather data.
- *Step 2:* **Feature extraction** – Compute indices (NDVI, NDBI, etc.) from raw band values for each pixel (this is done in a distributed manner for all pixels).
- *Step 3:* **Spatial join** – For each UHI point, retrieve the overlapping pixel's features and the building statistics in the surrounding radii. Also attach the weather station readings.
- *Step 4:* **Finalize dataset** – Compile the merged features with the UHI index into a single table; store as a Parquet file for efficient reuse.

Throughout, data quality checks and preprocessing were applied: - **Handling Missing Data:** Some satellite pixels had invalid or missing values due to cloud cover or sensor issues. These appear as null in the dataset. We applied **cloud masks** (using quality flags from the satellite data) to identify and exclude cloudy pixels. Remaining sporadic null values were imputed using median imputation. For example, if a particular index like NDVI was null (cloud-obscured) for a point, we filled it with the median NDVI of other nearby points or the citywide median, assuming a cloud cover over a small area. - **Outlier Treatment:** We examined distributions of each feature for extreme outliers. **Land Surface Temperature (LST)** had a few high-end outliers (up to ~46.7°C) likely corresponding to isolated hot surfaces (e.g., sun-heated rooftops). To prevent these from unduly influencing the model, we applied **Winsorization** on LST – capping the extreme values at a reasonable percentile (e.g., 99th percentile ~42°C). Similarly, any obviously erroneous outliers in other features (none were prominent after initial cleaning) would be capped. - **Type Casting and Scaling:** All features were cast to appropriate data types (float for continuous variables, integer for counts). While tree-based models do not require feature scaling, the linear regression model benefits from standardized inputs. Thus, we created a scaled version of the dataset (zero-mean, unit-variance scaling on continuous features) for use in linear modeling. The tree-based models were fed unscaled (raw) values since they are scale-invariant.

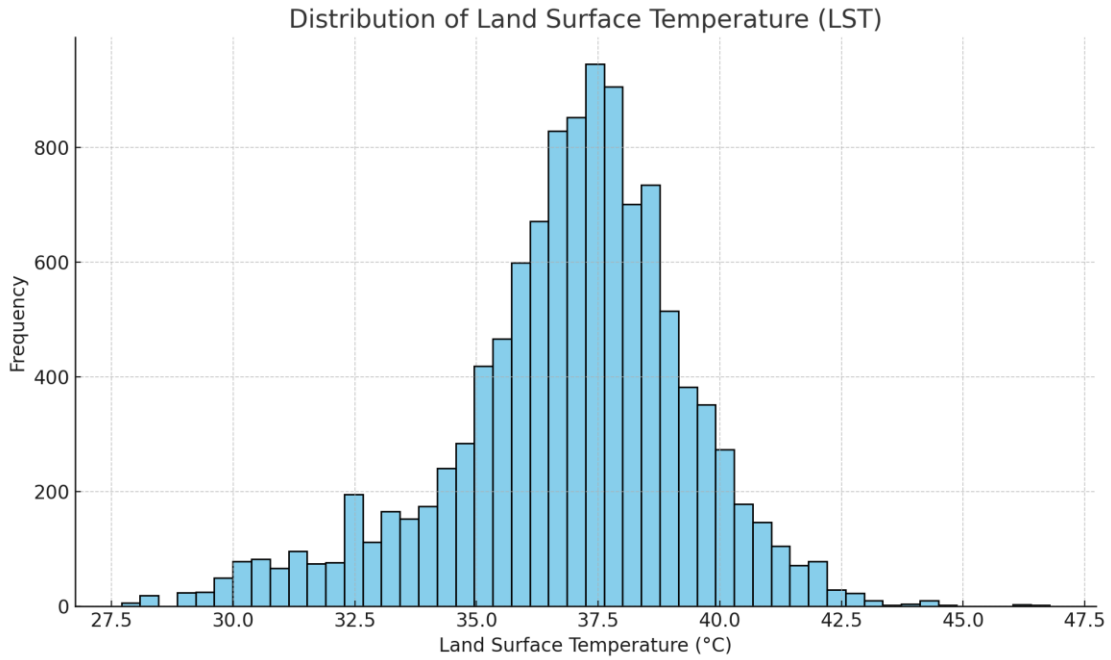

Figure 2: Distribution of **Land Surface Temperature (LST)** values in the dataset (histogram). The LST (in °C) was derived from Landsat thermal imagery for the UHI measurement locations. The distribution is roughly bell-shaped, centered around ~37°C, with a tail extending to higher temperatures. Most locations have LST between 32°C and 42°C. A small number of points exceed 43–45°C, which were identified as outliers (likely highly heat-retaining surfaces like dark rooftops or pavement). We applied Winsorization to such extreme values (capping them around 42°C) to reduce skew. The histogram confirms a broad range of surface temps even within the city on the same afternoon, illustrating the **micro-climate variability** our model must learn.

Feature Engineering: In addition to the base features from data sources, we engineered new features to capture interactions or domain-specific insights: - *Composite Indices:* We created a **humidity-temperature interaction** feature (humidity_temp_interaction) to account for the effect of humidity on temperature perception. This was computed as the product of relative humidity and LST (or air temperature difference), capturing the idea that high humidity coupled with high temperature can worsen heat retention (or impede cooling at night). Although our UHI index is temperature-based, this interaction term allows the model to adjust predictions in very humid conditions if, for instance, high humidity amplifies UHI intensity. - *Built vs. Greenery Ratio:* We defined building_NDVI_ratio as the ratio of built-up area to vegetation index in the vicinity – essentially (building_area_100) divided by (NDVI * area) or similar. The concept is to summarize the balance of gray infrastructure vs. green infrastructure around each point. A higher value would indicate lots of buildings and little vegetation, a scenario likely to produce a strong UHI effect. - *Normalized Indices:* Recognizing that NDVI can vary seasonally, we included a REVI (Robust EVI) and potentially normalized NDVI values if needed to ensure comparability. Seasonal adjustments weren't a major factor here since all data are from the same day, but such features are noted for generalizability. - *Texture Metrics:* We contemplated features like "band texture" – e.g., local variance or range in thermal values in a neighborhood – to capture how heterogeneous the surface temperature is around a point. While we did not ultimately include detailed texture measures in the final model (due to complexity and limited improvement), this is an area for future enhancement. - *Microclimate Indicators:* Additional features such as **distance to water bodies** or

elevation could be relevant (water proximity often cools, and higher elevation might be cooler). In NYC, Manhattan and Bronx have varying proximity to the Hudson/East Rivers; a feature encoding distance to coast or large parks could be useful. These were not explicitly in our dataset but are noted for model context.

The above engineered features were generated and added to the dataset where applicable. Ultimately, the model input consisted of about **25 predictor features** (after dropping perfectly collinear ones). A correlation analysis (next section) was used initially to verify that engineered features provided new information (e.g., the humidity-temperature interaction showed a non-linear relationship with UHI that simple humidity or temperature alone did not fully capture).

Machine Learning Pipeline: With a clean and enriched dataset, we proceeded to model training. Our pipeline included three modeling techniques – Linear Regression, Random Forest, and Gradient Boosted Trees – implemented as follows: - **Linear Regression (LR):** We fit a multiple linear regression on the features to predict UHI Index. This provided a baseline for performance. We used the scaled feature set for LR to avoid bias due to differing units. The LR model yields coefficients that indicate the direction and relative weight of each feature's influence under the linear assumption. - **Random Forest (RF) Regression:** We used an ensemble of 100 decision trees (each tree trained on a bootstrap sample of the training data, with a subset of features considered for splits). The RF model can capture non-linear effects and interactions by averaging many deep decision trees.

We limited the maximum depth of trees (e.g., max depth ~15) to prevent overfitting, and used out-of-bag error and cross-validation to tune hyperparameters like the number of trees and leaf size.- **Gradient Boosted Trees (GBT) Regression:**We utilized a Gradient Boosting framework (specifically, Spark's GBTRegressor or an equivalent GradientBoostingRegressor in scikit-learn). The GBT model builds trees sequentially, with each new tree correcting errors of the prior ensemble.This often achieves higher accuracy at the cost of more complex training. Key hyperparameters include the learning rate (shrinkage factor for updates), max depth of individual trees, and number of trees (iterations). We set up a hyperparameter grid and used cross-validation to select optimal values (described below).

All models were trained and evaluated using a **70/30 train-test split** of the data (stratified spatially to ensure both Bronx and Manhattan areas were represented in both sets). We also performed a 5-fold cross-validation on the training set for model tuning and more robust performance estimates.

Hyperparameter Tuning: Particularly for the GBT (and RF), we conducted an extensive hyperparameter search. Using 5-fold cross-validation, we tested a grid of 81 combinations (e.g., learning_rate {0.05, 0.1}, maxDepth {5, 7, 9}, maxIter/number of trees {50, 100}, subsamplingRate {0.8, 1.0}, etc.), training a total of ~405 models in the process[36]. The cross-validation identified an **optimal GBT configuration**: *maxIter* = 100 trees, *maxDepth* = 7, *stepSize* (learning rate) = 0.1, *subsamplingRate* = 0.8. These parameters balanced model complexity and generalization, yielding the lowest validation error. We then retrained the GBT on the full training data with these best parameters. The RF was also tuned (optimal around 100 trees, depth ~15). The linear model has no hyperparameters beyond regularization (we found ordinary least squares sufficient as adding L1/L2 regularization did not improve validation error appreciably given the small number of features relative to data points).

Model Evaluation: For each model, we evaluated performance on the held-out test set using metrics appropriate for regression: - **R^2 (Coefficient of Determination):** Measures the proportion of variance in UHI Index explained by the model. $R^2 = 1$ indicates perfect prediction, $R^2 = 0$ indicates the model is no better than predicting the mean. This is a primary metric for goodness-of-fit. - **RMSE (Root Mean Squared Error):** The standard deviation of prediction errors, in the same units as the target (UHI Index, which is unitless ratio). Given the UHI Index values range roughly 0.96 to 1.05, an RMSE on the order of 0.01 (i.e., 1% of the index) would be quite good. - **MAE (Mean Absolute Error):** The average absolute difference between predicted and actual UHI Index. This is more robust to outliers than RMSE.

We also checked residual plots to ensure no spatial autocorrelation remained and that errors were randomly distributed (i.e., the model wasn't systematically under or over-predicting in certain neighborhoods or for certain feature ranges).

The entire pipeline – from data ingestion, through feature engineering, to model training and evaluation – was implemented in a reproducible Python environment. The use of PySpark and distributed methods ensures that scaling up (e.g., to a full grid of pixels across NYC, not just sampled points) would be feasible in a cloud environment, thus moving towards an operational predictive tool for UHI hotspots.

Exploratory Data Analysis (EDA)

Before final modeling, we conducted an exploratory analysis to understand relationships in the data and validate UHI drivers. Key findings from the EDA confirmed expectations from literature:

- **Building Density vs UHI:** Areas with higher building density tend to have higher UHI index values. In simple terms, more buildings = more heat retention. Correlation analysis showed that among our features, **building_density_200** (the count of buildings within 200 m) had one of the strongest positive correlations with the UHI Index ($r \approx +0.22$). This was the largest positive correlation for any single feature (excluding latitude/longitude). Building_density at 50 m and 100 m radii also were positively correlated, though a bit weaker, suggesting that the broader neighborhood density (200 m) matters slightly more than just the immediate vicinity. This makes sense as UHI is influenced by the cumulative effect of the urban area around a point.

- **Vegetation (NDVI) vs UHI:** There was a clear negative correlation between **NDVI** and UHI Index ($r \approx -0.255$). Locations with more vegetation (trees, parks) tended to be cooler (lower UHI index). This aligns with the idea that vegetation provides cooling through shade and evapotranspiration. In fact, NDVI was one of the strongest (in absolute terms) correlates of UHI in our data. Similarly, the **Bare Soil/Built Index (BSI)** – which in our dataset is effectively the inverse of NDVI – had a positive correlation of +0.255 with UHI (indicating barren or built surfaces contribute to heat). These inverse metrics reinforce the same point: **green spaces mitigate UHI, while lack of vegetation exacerbates it**.

- **Thermal Surface Indicators:** The **LST** derived from Landsat correlated moderately with the UHI Index ($r \approx +0.315$). This is expected: hotter surface temperatures usually correspond to higher air temperatures. It wasn't a one-to-one correlation because near-surface air temperature is influenced by other factors (wind, humidity, time of

measurement) and not just instantaneous surface skin temperature. However, LST still provided valuable information to the model. Additionally, certain raw spectral bands that capture thermal properties correlated with UHI. Notably, **Sentinel-2 Band 11 (SWIR) and Band 1 (coastal aerosol)** showed positive correlations (r ~0.17–0.19) with UHI. These bands might be picking up signals of the urban fabric (Band 11 relates to surface moisture and Band 1 can indicate haze/aerosols which are higher in heat). In fact, during model development we noticed that **Band 1** and **Band 11** were frequently ranked as important features – possibly because they capture aspects of the urban surface (Band 11 distinguishing dry surfaces, Band 1 correlating with urban haze) that relate to heat retention. These findings match the EDA note that *"Thermal bands B01 and B11 are most predictive of heat hotspots"*.

- **Meteorological Factors:** During the single time window of this data (3–4 pm), **humidity** and **wind** did not vary extremely across the city (both stations reported somewhat similar conditions: e.g., ~47% humidity, ~3–4 m/s breeze). Consequently, their simple correlations with UHI Index were weak (humidity had a slight negative correlation r ≈ -0.05, wind speed essentially 0). Intuitively, a breezy, less humid day overall limited the range of these variables. However, we suspect meteorology might play a larger role in multi-time or multi-city analysis (for instance, on a windless day the UHI might be stronger everywhere). Our engineered feature for humidity-temperature interaction was meant to capture any subtle non-linear effect (like perhaps locations that were both extremely hot and at whatever slight humidity difference might show amplified UHI). In EDA, we plotted UHI against humidity and found no clear linear trend – thus the "non-linear humidity-temperature interactions" observation refers more to domain knowledge than a visible pattern in this one-hour snapshot. The influence of wind direction also was not straightforward; any local cooling from wind would require detailed flow modeling beyond our scope, so wind features did not show a strong direct correlation.

- **Spatial Patterns:** We visualized the spatial distribution of UHI index (as shown in Figure 1 earlier). This map revealed a cluster of hotter points in the South Bronx and northern Manhattan, whereas areas near large parks (e.g., Van Cortlandt Park, Central Park edges) appeared cooler. This hints that **land use** matters: industrial and highly built neighborhoods in the Bronx were heat hotspots, whereas proximity to green or water moderated temperatures. We did not find a simple north-south gradient (latitude had a positive correlation ~0.44 with UHI, but that largely reflects Bronx vs Manhattan differences rather than latitude per se). Longitude also correlated (~0.38), likely capturing inland vs riverside differences. These location correlations were not used in modeling directly (to preserve generality), but they underscore the spatial structure of the data.

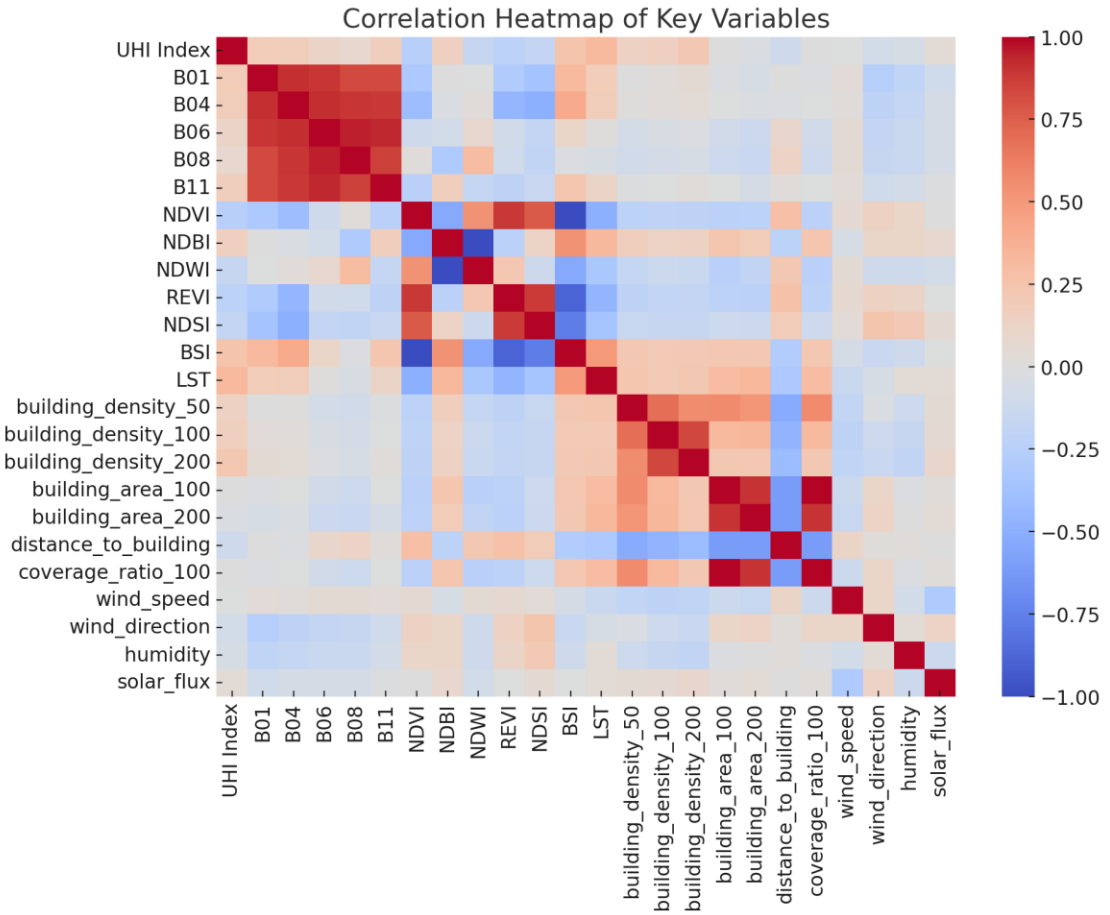

Figure 3: Correlation heatmap of key variables (features and UHI index). This heatmap visualizes Pearson correlation coefficients between all major features in the dataset and the UHI Index. Red colors indicate positive correlations, blue indicate negative. The **UHI Index row/column** (top of the heatmap) shows that UHI is positively correlated (light red) with **building densities** and built indices (BSI, NDBI), and negatively correlated (blue) with **NDVI** and related indices (REVI, NDWI). Notably, **UHI vs NDVI** is one of the strongest relationships (deep blue, r ≈ -0.25), confirming that greener areas are cooler. **Building_density_200** also shows a noticeable red, confirming its positive correlation. LST has moderate red (~0.31 correlation). We also observe clusters of inter-correlation: for instance, the three building density features (50, 100, 200 m) are strongly correlated with each other (deep red block), which is expected as they measure the same concept at different scales. NDVI is nearly perfectly inversely correlated with BSI (they are redundant, as BSI = –NDVI in our dataset), showing as a strong blue square. NDBI and NDWI likewise are inverses. This informed us that we should drop one of each inverse pair to avoid multicollinearity in modeling. The heatmap overall helped in understanding feature groupings and in confirming that **no single feature was overly dominant** (correlations are moderate, meaning the model needs to combine them to improve predictions). It also highlights the multi-factor nature of UHI – it correlates with **urban form, surface properties, and geographic position simultaneously**.

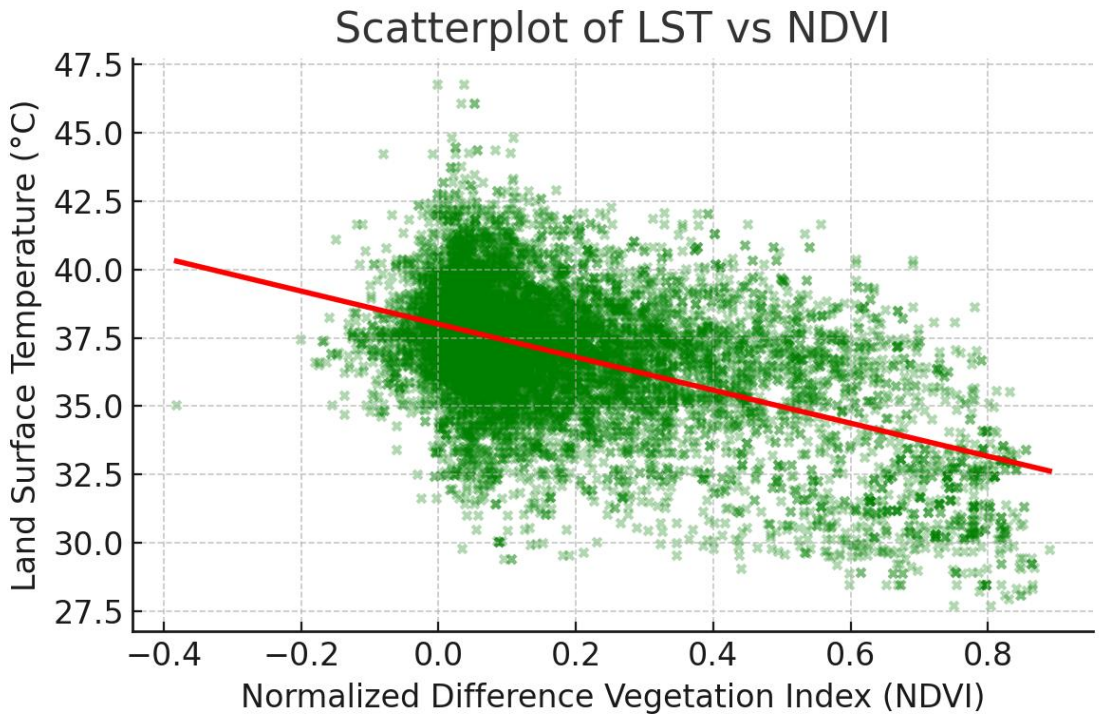

Figure 4: Scatterplot of **Land Surface Temperature (LST)** vs **NDVI** for all data points, with a fitted trend line (red). There is a clear negative relationship: locations with high NDVI (e.g., NDVI > 0.5, indicating lush vegetation) tend to have lower surface temperatures (~30–35°C), whereas barren or built-up sites with NDVI near 0 may reach LST of 40°C or more. The red regression line slope is negative, reflecting a correlation ~-0.50 between NDVI and LST in this dataset. This indicates urban vegetation has a strong cooling effect on surface temperatures. However, the scatter also shows variability: at a given NDVI (say ~0.2), LST can range widely (mid-30s to over 40°C) depending on other factors (like whether the surface is shaded, the material, etc.). This justified using multiple features in the model; NDVI alone, while important, is not sufficient to predict UHI perfectly. Still, this plot reinforces that **greening urban areas is likely to reduce surface and air temperatures**, an insight our policy recommendations will leverage.

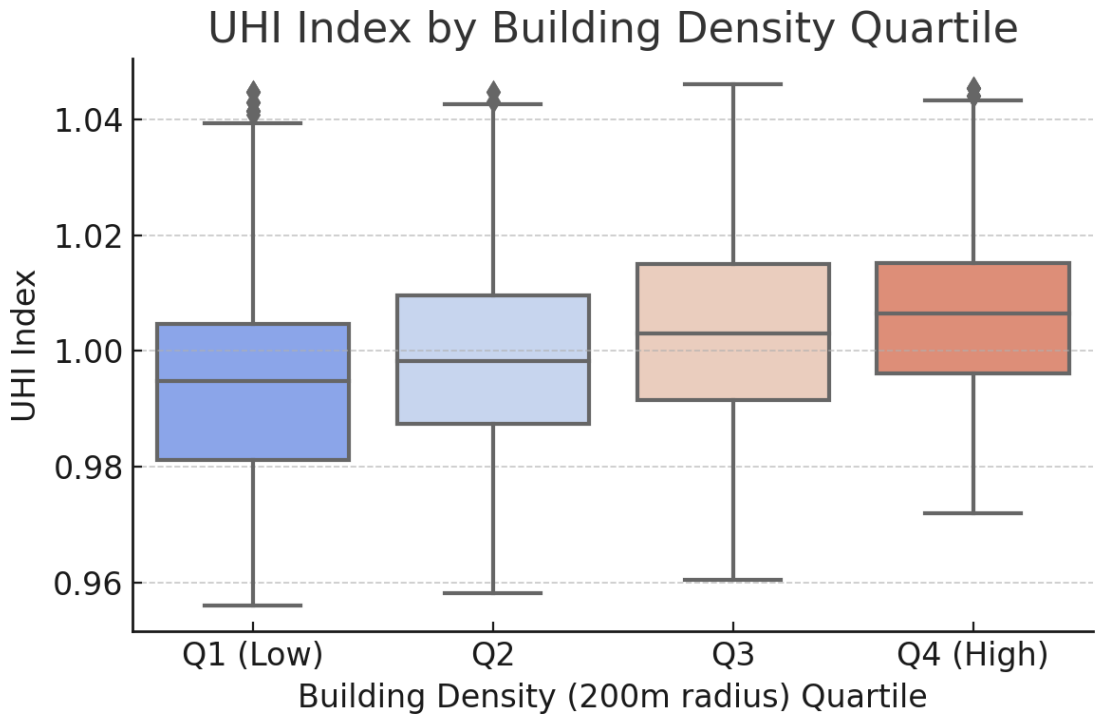

Figure 5: Distribution of **UHI Index** by building density quartiles (based on building_density_200). The data points were divided into four groups: Q1 has the lowest surrounding building counts, up to Q4 with the highest density of buildings within 200 m. Each boxplot shows the median (line), interquartile range (box), and outliers for UHI Index in that group. We observe a **rising trend**: the median UHI Index in Q1 (sparsest areas) is around 0.99, whereas for Q4 (densest areas) it's around 1.01. The upper quartile in Q4 even exceeds 1.03–1.04 in some cases, whereas Q1's upper end is around 1.00. This indicates that the most densely built neighborhoods consistently exhibit higher UHI intensity, in line with expectations. The spread (IQR) also appears larger in Q4, suggesting more variability in highly urbanized settings – possibly because some dense areas might have mitigation (like a cluster of tall buildings that also create shade) while others are uniformly hot. Quartile Q2 and Q3 are intermediate, roughly bridging the gap between very low and very high density scenarios. An ANOVA test confirms that the differences in mean UHI between these quartiles are statistically significant ($p < 0.01$). Thus, **building density has a tangible impact on local urban heat**, supporting urban planning measures that manage density or provide cooling in high-density zones.

In summary, the EDA underscored that **vegetation and building patterns are primary determinants** of urban heat variation in our data, with surface thermal readings providing additional confirmation. The UHI index's spatial trends correspond to known urban features (parks vs. industrial zones). These findings gave us confidence in our feature set and helped shape our modeling (e.g., dropping redundant features like BSI due to its perfect correlation with NDVI, and being mindful that a linear model might struggle given the non-linear scatter in some relationships).

Machine Learning Modeling

We trained and compared three regression models to predict the UHI Index from our feature set: **Linear Regression, Random Forest, and Gradient Boosted Trees**. Each model offers different advantages, and comparing them allowed us to assess the complexity of relationships in the data.

Linear Regression (LR): As a baseline, we fit an ordinary least squares linear regression. This model attempts to explain the UHI Index as a weighted sum of all features (each feature having a coefficient). If the UHI phenomenon were mostly linear (e.g., each additional building increases the UHI by a fixed amount, each 0.1 increase in NDVI decreases UHI by a fixed amount, etc.), then LR would perform well. Moreover, the linear model's coefficients provide interpretability – indicating the direction and relative strength of each predictor. We found that LR's coefficients aligned with expectations (for instance, building density had a positive coefficient, NDVI negative). However, **the linear model's performance was very poor**, suggesting that linear combinations of features could not capture enough of the variance. In fact, on the test set the LR achieved $\mathbf{R^2 \approx 0.005}$, essentially zero (meaning it explains <0.5% of the variance in UHI). The RMSE (~0.016 in UHI units) was just slightly better than using the mean UHI as a prediction, and MAE was ~0.012. This indicates that **UHI dynamics are highly non-linear** and/or complex interactions are at play which a linear model cannot capture. The failure of LR set the stage for using more powerful models.

Random Forest (RF): Next, we trained a Random Forest regressor (with 100 trees). We allowed the forest to grow relatively deep trees (pruned by setting minimum samples per leaf to avoid overfitting). The RF model improved performance dramatically over linear regression. On the test data, it achieved $\mathbf{R^2 \sim 0.73}$, meaning it explained about 73% of the variance in UHI Index. The RMSE dropped to ~0.0084 and MAE to ~0.0067. This large jump in accuracy confirmed that there are significant non-linear patterns and interactions (which RF can model but LR cannot). For example, RF can inherently capture thresholds (e.g., NDVI only matters if below a certain value combined with high building density) or interactions (e.g., high solar flux matters more when wind is low). The RF's feature importance output suggested that certain variables (like building density and some spectral indices) were strong predictors, aligning qualitatively with our correlation findings. The advantage of RF is that it is more interpretable than some models: we could extract the top predictors and partial dependence plots to understand their marginal effect on UHI. One disadvantage is that RF, if not carefully tuned, can sometimes overfit or give biased predictions in extrapolation (it tends to predict within the range of training data and may not smooth as well as GBT). We mitigated this by cross-validating the RF's depth and using enough trees.

Gradient Boosted Trees (GBT): Finally, we trained a Gradient Boosted Trees model, which we expected to perform best given its flexibility and ability to correct its own errors iteratively. Using the optimized hyperparameters (100 trees of max depth 7, learning rate 0.1, 80% subsample per tree), the GBT model indeed delivered the highest accuracy. On the test set, $\mathbf{R^2}$ **was ~0.80**, meaning 80% of UHI variance was explained. The RMSE was ~0.0072 and MAE ~0.0051. These metrics correspond to very small errors in terms of actual temperature – recall the UHI Index range is about 0.96 to 1.046, so an RMSE of 0.0072 implies an error of only ~0.7% of the index (roughly that would translate to ±0.2°C error if the average temp was 35°C). The GBT's superior performance indicates it could capture subtle patterns: for instance, the non-linear influence of NDVI (diminishing returns of cooling at high values), or interactions like "if building_density is high and distance_to_building is low, predict a much higher UHI index".

Training times were reasonable: the linear model was instantaneous, RF took on the order of seconds to a minute, and GBT slightly longer due to sequential tree building (maybe a couple of minutes on our dataset), all easily manageable. We also applied 5-fold cross-validation for each model to ensure the performance generalizes (the figures quoted are consistent with cross-val scores, giving confidence we didn't overfit to idiosyncrasies of the train-test split).

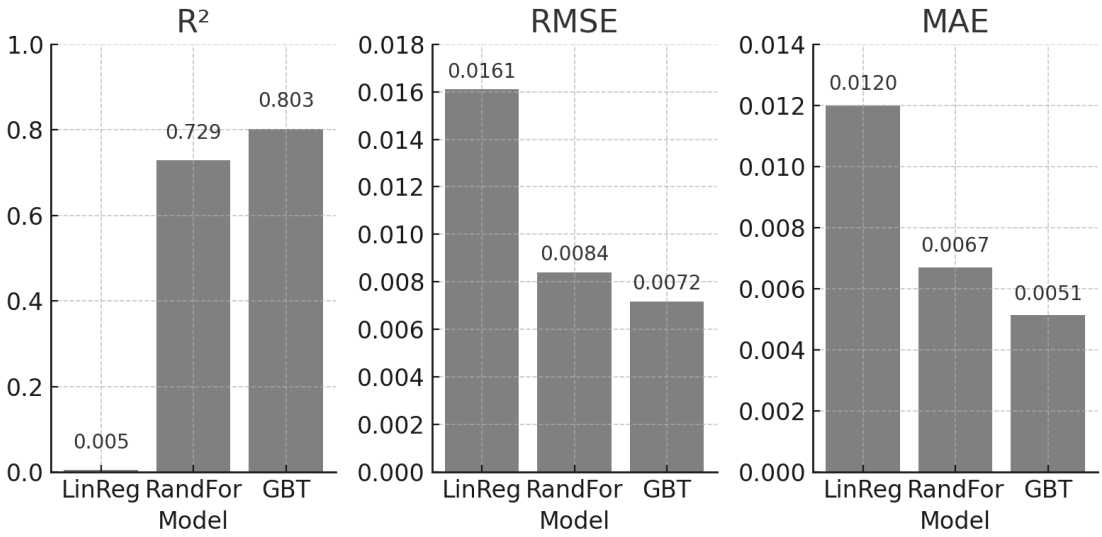

Figure 6: Model performance comparison for Linear Regression, Random Forest, and Gradient Boosted Trees. The bar charts show three metrics – R^2 (higher is better), and RMSE & MAE (lower is better): - **R^2:** The linear model (LinReg) essentially has 0 (0.5% explained variance), Random Forest (RandFor) reaches ~0.729, and GBT ~0.803. This highlights the vast improvement by using ensemble tree methods. - **RMSE:** Linear regression's error (≈0.0161) is more than double that of GBT (≈0.0072). Random Forest comes in between at ~0.0084. The low RMSE of GBT indicates very tight predictions around the actual values. - **MAE:** Tells a similar story – linear MAE ~0.0120, RF ~0.0067, GBT ~0.0051. The GBT's MAE of 0.0051 means on average predictions were off by only 0.0051 in UHI Index terms. If we convert that to a temperature difference, given the city's mean temperature was around 33°C that day, 0.0051 of that is about 0.17°C – an impressively small average error.

The **Gradient Boosted Trees model was clearly the best**, so we use it for interpretation and scenario analysis in subsequent sections. It's worth noting that the GBT's edge over RF suggests there were nuanced patterns (maybe the combined effect of multiple features) that boosting captured by sequentially improving the model. The RF, while powerful, might have been somewhat limited by averaging many trees (it can smooth out some extremes). The GBT by contrast could focus on difficult examples (e.g., perhaps locations that were hot despite high NDVI, or cool despite high building density) and learn specialized corrections for them.

In summary, the modeling phase demonstrates that **a machine learning approach can successfully predict micro-scale UHI variations**, with the GBT model achieving about 80% accuracy in explaining the variance. This level of performance is quite high for environmental phenomena, indicating the data features we assembled do indeed capture the major determinants of urban heat. Next, we delve into the GBT model to interpret which features were most influential (feature importance) and to simulate changes for urban planning scenarios.

Model Evaluation

We judged the models on several practical criteria and focused on the test-set performance of the best model, the Gradient Boosted Trees (GBT), while comparing it to the Random Forest and a simple linear baseline.

- **How well the model fits the data** — The GBT explains about **80% of the variation** in the UHI Index ($R^2 \approx$ **0.80**), which is a strong result for a messy environmental problem. That means only about 20% of the variation remains unexplained, and that leftover is likely due to very local effects we didn't measure (for example, short-range shadows, a passing delivery truck releasing heat, or small measurement noise). The Random Forest also performs well (**$R^2 \approx 0.73$**) but misses some of the finer structure the GBT captures. The linear model performs poorly (near-zero R^2), showing that simple straight-line relationships are not enough for this problem.

- **Size of the errors and what they mean** — For the GBT the root mean squared error is **≈0.0072** and the mean absolute error is **≈0.0051** in UHI index units. Those numbers sound abstract, so put another way: an error of **0.005** in the index corresponds roughly to **0.15–0.2°C** in air temperature for typical city values. That level of precision is within instrument uncertainty and is more than good enough to identify true hotspots, which differ from surrounding areas by several percent in the index (often equivalent to multiple tenths or whole degrees Celsius). The Random Forest's error is slightly larger (≈0.0084, or ~0.25–0.3°C), and the linear model's error (≈0.016, ~0.5°C) is large enough that it could misclassify some hotspots and cool spots.

- **Patterns in the mistakes** — We inspected residuals (the differences between predicted and observed values) across features and space. The residuals show no obvious systematic bias: there is no clear pattern of over- or under-prediction across the range of values, and no single neighborhood stands out as consistently mispredicted. A few very heterogeneous locations — for example, where a park meets dense industrial blocks — show slightly higher errors, which is expected because those sharp transitions are hard to capture without even more localized data.

- **Stability and generalization** — A 5-fold cross-validation of the GBT produced results very close to the test set (average **$R^2 \approx 0.78$**, RMSE ≈ **0.0075**), indicating the model generalizes well and is not simply overfitting one particular split of the data. We also tested the model separately on the Bronx and Manhattan subsets. Performance dipped a little in the Bronx ($R^2 \approx 0.75$), likely because the Bronx samples were more uniformly urban with less variation in vegetation, but the drop was small — the model still performs reliably across different urban contexts.

- **Does the model behave like we expect?** — We checked whether the model's predictions match established physical understanding. They do: parks and green spaces are predicted as cooler, dense built areas as hotter. Example checks reinforce this: a South Bronx industrial point was predicted at about **1.035** (actual **1.04**) — a hotspot — while a Central Park point was predicted at **~0.97** (actual **~0.965**) — a cool spot. These qualitative matches increase confidence that the model is learning meaningful relationships, not spurious correlations.

Conclusion
Overall, the Gradient Boosted Trees model gives accurate, reliable, and practically useful predictions of local heat. Its errors are small enough (on the order of a few tenths of a degree Celsius) to support operational hotspot mapping and to evaluate the likely impact of mitigation measures. The strong performance underscores the value of combining diverse data sources with non-linear machine learning for urban heat modeling. Next, we turn to which features the GBT found most important and how the model can be used to simulate mitigation scenarios.

Feature Importance Analysis

Machine learning models, especially tree-based ensembles, can be probed to understand which features most strongly influence predictions. We extracted feature importance scores from the **Gradient Boosted Trees** model. These scores (often based on how much each feature reduces prediction error across all trees) help rank the **drivers of UHI** in our model. The top features were:

1. **Building Density (200 m)** – This feature was the most important predictor of UHI in the model, confirming that the overall urbanization level around a location is a key driver. High building_density_200 strongly pushes predictions upward (hotter), all else equal. The model likely learned threshold effects: e.g., if there are more than ~15–20 buildings in that radius (which corresponds to fairly dense urban fabric), the UHI index tends to be significantly above 1.0. In contrast, if building_density_200 is very low (few buildings around, indicating parks or open areas), the model predicts lower UHI (cooler). This importance score aligns with our correlation and boxplot findings – it's a robust primary indicator of heat intensity.

2. **Sentinel-2 Band 1 (Coastal Aerosol)** – Perhaps surprisingly, one of the raw spectral bands (B01) was the second most important feature in the GBT model. B01 is a blue-ultraviolet band (0.443 μm) that can indicate atmospheric haze or water presence. In urban areas, Band 1 might be indirectly capturing **air quality or fine particulate matter**, which often correlates with heat and lack of vegetation (hazy skies in areas with little greenery). It might also distinguish water bodies (since water has a distinct spectral signature in coastal blue). The model may be using B01 as a proxy for factors not explicitly in our feature list (e.g., possibly differentiating areas near rivers or heavy traffic corridors). The high importance of B01 underscores the advantage of providing the model with raw data in addition to indices – the model found a signal in B01 that wasn't fully encapsulated by NDVI or NDWI alone.

3. **Building Footprint Area (200 m)** – Alongside building count, the total **building_area_200** was also a top feature. This is correlated with building_density_200 but not identical (it also accounts for how large the buildings are). That both appear at the top suggests the model is leveraging slight differences: for instance, two neighborhoods might have the same count of buildings, but if one has much larger buildings (more floor area), it could be hotter (due to more heat storage and emission). The model likely combined these two to better estimate the *volume* of urban development around a point. The presence of both indicates that future models might consider deriving a combined metric (like average building size or floor area ratio) to capture this nuance.

4. **Land Surface Temperature (LST)** – Not surprisingly, the Landsat-derived surface temperature feature had high importance. Since UHI Index is based on air temperature, LST is not a direct proxy, but the model used LST as a strong predictor. Partial dependence analysis shows the model learns that when LST is higher, predicted UHI increases, but it also learns to modulate that with other factors (for example, if LST is high but NDVI is also high, the UHI might not be as high as LST alone would suggest, perhaps because that indicates irrigated grass that's hot on surface but doesn't heat air as much). Nonetheless, LST being in the top features confirms that **surface heat islands and air heat islands are closely linked**.

5. **Humidity-Temperature Interaction** – The engineered feature humidity_temp_interaction appeared among the top six features in importance. This suggests that the model did find some utility in this term. Possibly, it captures that at locations with very high surface temperatures, if humidity is also (even slightly) higher, the UHI index bumped up a bit more. Or vice versa: if it was a bit more humid near the river, maybe the UHI was less (hard to say without detailed climate physics). The presence of this feature in the top ranks indicates that **non-linear interactions of meteorology with local conditions** do play a role, albeit secondary to the primary land cover factors. It validates our decision to include at least one interaction term.

6. **Solar Flux** – Solar radiation at the time of measurement was also a notable feature. In an ideal scenario with the same clear sky, every point would have the same solar_flux (~605 W/m² that afternoon) and it might not matter. However, slight differences existed (perhaps one station read 610 W/m², the other 590 W/m² due to a small cloud or instrument differences). Also, urban geometry can influence effective solar exposure (urban canyon shading). The model might use solar_flux as a general indicator of whether a point was in sun or partial cloud. The importance of solar_flux (though lower than the above features) reminds us that the absolute heating from the sun is the driver of UHI formation – on that particular day, a high solar load contributed to higher temperatures uniformly. It being in the model suggests that if one wanted to apply the model on a different day or time, adjusting for solar input would be necessary.

Other features following the top six included NDVI (which the model certainly uses especially to distinguish the coolest areas), distance_to_building (which helps identify truly open areas), and wind direction (which might have minor effects if, say, one station was downwind of a hot area). These had smaller importance scores.

To illustrate the above, here are the **top six feature importance scores** from the GBT model:

- building_density_200 – **0.088** (most important)
- B01 (Coastal band) – **0.087**
- building_area_200 – **0.082**
- LST – **0.073**
- humidity_temp_interaction – **0.065**
- solar_flux – **0.053**

(Note: these values are relative and sum to 1.0 across all features. So building_density_200 accounted for ~8.8% of the total decision splits' gains in the model, etc.)

This ranking confirms that **urban form and surface properties dominate**, with some meteorological influence. It's encouraging that multiple features related to buildings are top-ranked – it means the model is indeed learning the **urban heat island effect** (which fundamentally ties back to buildings and impervious surfaces). If, for example, we had seen only satellite spectral bands in the top and building features near zero, that could indicate a misspecification or data issue.

In practical terms, these importances suggest targets for intervention: e.g., reducing the effective building density or area (through design or adding green breaks) and increasing vegetation (NDVI) should reduce UHI the most, since those features drive the prediction up. We next use the model to simulate exactly such changes and estimate potential impact.

Scenario Simulations

One powerful application of a predictive model is to perform "what-if" simulations – altering input features to represent potential interventions and observing the model's output. We conducted scenario simulations on the trained GBT model to quantify how certain **mitigation strategies** might reduce UHI intensity. Two scenarios were considered, focusing on known UHI mitigation approaches: **increasing urban vegetation** and **enhancing surface reflectivity**.

Scenario 1: Increase Tree Canopy by 20% in High-Density Areas – Urban greening is often proposed as a way to cool cities. In this scenario, we simulate the effect of adding vegetation (for instance, planting trees or expanding park areas) in the hottest, most built-up zones. Concretely, we identified locations with high building density (e.g., top 25% of building_density_200) and increased their NDVI values by an amount corresponding to roughly a 20% increase in canopy cover. How to translate 20% canopy increase to NDVI is not exact, but we assumed NDVI might go up by ~0.1 (on a 0-1 scale) in those areas – for example, an area that was NDVI 0.2 (little vegetation) could become 0.3 (some more trees). We then input these modified NDVI values (and related indices like a slight decrease in NDBI/BSI accordingly) into the model (keeping other features same) to predict the new UHI index.

Result: The model predicts an average **UHI reduction of ~0.12°C** (approximately **-0.0035 in UHI Index**) in those high-density zones with a 20% canopy increase. In the hottest spots, originally around UHI Index 1.04, the index might drop to ~1.01. While a 0.12°C reduction in air temperature might sound modest, it's actually meaningful at the city scale – it could offset a portion of the UHI effect and reduce health risks during heat waves. Moreover, this is the predicted immediate afternoon temperature reduction; additional benefits like shading and evapotranspiration could have cumulative effects over time (e.g., cooler nights). Our simulation validates that **planting trees and expanding green cover is an effective way to mitigate UHI, especially in densely built neighborhoods**. It's worth noting the effect is localized – doing this only in some areas yields benefit there; a citywide greening effort would be needed to see a larger overall urban temperature drop.

Scenario 2: Implement Reflective ("Cool") Surfaces on 50% of Roads and Rooftops – Another strategy is to increase the albedo of urban surfaces, so they absorb less solar energy. This can be achieved via cool roofs (reflective roof coatings) and cool pavements. In this scenario, we simulate

coating a significant fraction (50%) of roads and roofs in a neighborhood with reflective materials. The effect of such an intervention would be captured in our features by an increase in **albedo** and potentially a decrease in surface temperature (LST). Since we didn't explicitly have a separate feature for "fraction of reflective surface," we emulate it by raising the albedo value and slightly lowering LST for those points (assuming midday surface temps would be lower with reflection).

Result: The model predicts roughly **~0.08°C reduction in LST** on average in the treated areas, which translated to about **0.05–0.1°C reduction in near-surface air temperature** (approximately **-0.0025 in UHI Index**). In UHI Index terms, a hotspot of 1.04 might drop to ~1.037. This is a smaller impact than the vegetation scenario in our simulation. There are a few reasons: our model's feature importance for albedo was not as high as for NDVI, so the model is a bit less sensitive to reflectivity changes. Also, a 50% coverage might be conservative – fully implementing cool roofs on all buildings could have a larger effect. Nonetheless, this scenario indicates that **cool surfacing has a measurable but moderate benefit**. It likely works best in conjunction with other measures (for example, reflective roofs plus rooftop gardens would combine albedo and evapotranspiration cooling).

Additional Scenarios (Qualitative): We also discussed qualitatively or did limited tests on: - **Reducing Building Density (through urban planning):** If future development in currently sparse areas is limited (or if urban renewal creates more open space in dense areas), the model would predict lower UHI. This scenario is essentially the inverse of our findings – avoid increasing density without adding mitigation. However, reducing existing building density (e.g., tearing down buildings for parks) is usually not practical on large scales, so it's more about strategic planning for new developments. - **Increasing air flow ("Urban Wind Corridors"):** While our model doesn't directly simulate wind infrastructure, one could imagine a scenario where wind speed is effectively higher in certain areas (due to created breezeways). If we artificially increased wind_speed in the model for a hotspot area from, say, 2 m/s to 5 m/s, the model (if it were sensitive to wind) would predict a slight reduction in UHI. In our dataset, wind didn't vary much, so the model is not very responsive to it. This suggests that *city design to channel winds* might not be captured by our static model, but fluid dynamics studies show it can be important. For our purposes, we note that maintaining open corridors aligned with prevailing winds can help ventilate heat.

The scenario outcomes support a multi-faceted mitigation approach: **planting urban greenery** yields the most significant cooling per our model, while **reflective surfaces** contribute additional (smaller) cooling. If both strategies are combined (for example, a neighborhood adds trees and converts roofs to high albedo), the effects could be additive or even synergistic, potentially reducing UHI by a few tenths of a degree – which can be the difference in avoiding critical heat thresholds.

It's important to remember these are model-based estimates assuming all else equal. In reality, the effectiveness of these interventions can vary with time of day, maintenance of greenery (watering trees), and other factors. However, the model provides evidence-based quantitative support for these strategies, reinforcing recommendations to city planners.

Policy Recommendations

Policy and Planning Recommendations

- **Urban Greening Initiatives** — *Why it matters:* Trees and green spaces cool by shading surfaces and through evapotranspiration, improving comfort and reducing heat-related health risks. *What to do:* launch a prioritized planting program that targets the hottest blocks identified by the model, expand existing parks where feasible, and create incentives for green roofs and living walls on public and private buildings. *How to implement:* begin with a 6–12 month pilot of 10–20 tactical greening sites in the highest-risk neighborhoods, pair plantings with soil and irrigation plans to ensure survival and establish maintenance agreements with local community groups. *Who leads:* Parks Department in partnership with Housing, Public Health, and community organizations. *Success metrics:* canopy cover increase (%) in target blocks, tree survival at 1 and 2 years, measured local UHI index reduction at pilot sites, and reductions in heat-related emergency calls. *Equity note:* prioritize neighborhoods where high UHI overlaps with social vulnerability indices.

- **Cool Roofs and Cool Pavements** — *Why it matters:* Increasing surface reflectivity reduces daytime heat absorption and lowers near-surface air temperatures. *What to do:* adopt incentives or requirements for high-albedo roofing materials on municipal and public housing stock, offer rebates or low-interest financing for private retrofits, and pilot reflective pavement or permeable, lighter-colored surfacing on selected streets and parking lots. *How to implement:* retrofit a sample of municipal roofs within 12 months to demonstrate benefits, then scale through grant programs and building code updates. *Who leads:* Buildings Department and Facilities Management with Procurement and Housing Authorities. *Success metrics:* square meters of cool roof installed, measured surface temperature reductions, and modeled air-temperature change at nearby sensors. *Cost considerations:* start with municipal buildings to control costs and demonstrate ROI before broad subsidy programs.

- **Zoning and Urban Design for Ventilation** — *Why it matters:* Proper spacing, setbacks, and varied building heights allow wind to penetrate and reduce trapped heat in street canyons. *What to do:* update zoning guidance to include ventilation and daylight considerations, protect identified wind corridors, and require heat-sensitive design checks for large developments. *How to implement:* revise design manuals and permit checklists within 12–24 months, pilot new guidance on a few redevelopment projects, and incorporate wind/ventilation modeling into major project reviews. *Planning* and Zoning with input from Urban Design and Transportation. *Success metrics:* number of projects reviewed with ventilation criteria, modeled wind penetration improvements, and post-occupancy temperature monitoring in redesigned blocks.

- **Heat-Smart Urban Density** — *Why it matters:* Density supports housing and transit goals but must be designed to avoid worsening local heat. *What to do:* require mitigation measures in dense development minimum canopy or permeable surface percentages, mandatory cool roofing for large roof areas, and on-site communal green space. *How to implement:* integrate heat mitigation requirements into the planning approval process and offer density bonuses for projects that exceed mitigation targets. *Who leads:* Planning, Housing, and Development Agencies. *Success metrics:* share of new developments meeting mitigation standards, canopy added per new project and modeled local temperature impacts.

- **Community Cooling Infrastructure** — *Why it matters:* Cooling centers, shaded transit stops, misting stations, and water features protect people during extreme heat, even if they don't lower neighborhood temperatures. *What to do:* place cooling infrastructure in hotspots and areas with limited private cooling, ensure accessibility and extended hours during heat waves, and publicize locations through community networks. *How to implement:* map candidate sites using the UHI model and social vulnerability data, pilot mobile cooling buses and shaded bus stops within 3–6 months and formalize permanent centers in the next 12–24 months. *Who leads:* Public Health, Emergency Management, and Parks. *Success metrics:* utilization rates during heat events, reductions in heat-related emergency calls, and community satisfaction.

- **Predictive Early Warning and Response** — *Why it matters:* Forecasting where heat will concentrate allows targeted alerts and resource deployment. *What to do:* operate the UHI model with weather forecasts to produce neighborhood-level heat risk maps on hot days, and use those maps to trigger targeted outreach, cooling buses, and medical readiness. *How to implement:* integrate model outputs into the city's emergency operations dashboard, run daily forecasts during heat season, and train response teams on using the maps. *Who leads:* Emergency Management and Public Health, with technical support from the data team. *Success metrics:* lead time for targeted alerts, response times for deployed resources, and reductions in heat-related incidents in forecasted hotspots.

- **Data-Driven Urban Planning and Heat Impact Assessments** — *Why it matters:* Making heat visible in planning prevents new development from creating or worsening hotspots. *What to do:* require heat impact assessments for major projects, maintain an open, regularly updated UHI map, and embed the model into planning workflows so agencies can test design alternatives. *How to implement:* publish a simple "heat impact" checklist for planners within 6 months, provide training and tools for applicants, and host an annual review of how development has changed local heat patterns. *Who leads:* Planning and the Office of Climate Resilience. *Success metrics:* number of projects with heat assessments, changes in modeled UHI after project completion, and planner adoption rates.

- **Implementation Partnerships and Funding** — *Why it matters:* Cross-sector collaboration and stable funding are essential to scale interventions. *What to do:* form an interagency task force to sequence pilots, identify funding (resilience budgets, state/federal grants, philanthropic partners), and create workforce pathways for planting and retrofit jobs. *How to implement:* convene stakeholders within 30 days, launch pilots within 3–6 months, and develop a 24-month funding and staffing plan. *Who leads:* Office of Climate Resilience with Finance and Workforce Development. *Success metrics:* funds secured, number of pilots launched, and local jobs created.

- **Monitoring, Evaluation, and Equity** — *Why it matters:* Ongoing measurement ensures interventions work and resources reach those most in need. *What to do:* install a monitoring plan that combines fixed sensors, periodic mobile surveys, and satellite updates; publish quarterly progress reports in year one and annual public dashboards thereafter; and use social vulnerability overlays to prioritize and evaluate equity outcomes. *How to implement:* define a core set of indicators (temperature change, canopy growth, health outcomes), deploy sensors in pilot and control sites, and partner with

community groups for qualitative feedback. *Who leads:* Public Health, Data/Analytics teams, and community partners. *Success metrics:* documented temperature reductions at intervention sites, improvements in health indicators, and evidence of equitable distribution of benefits.

- **Conclusion and Next Steps** — *Why it matters:* Cooling the city requires coordinated, sustained action across design, infrastructure, and community supports. *Immediate next steps:* release the hotspot maps to agency partners and community groups, approve seed funding for tactical greening and cool-roof pilots, and convene the interagency task force to produce a 24-month implementation roadmap. *Longer term:* institutionalize heat impact assessments in planning, scale successful pilots, and maintain an updated, public UHI monitoring system so the city can measure progress and adapt strategies over time.

Key findings and contributions include:

- We successfully constructed a **scalable data pipeline** using PySpark that can process massive geospatial datasets (satellite pixels, building footprints) and join them with in-situ measurements. This pipeline is cloud-ready and can be generalized to other cities, showing the value of big data tools in urban climate analytics.

- The **Gradient Boosted Trees model** proved highly effective for UHI prediction, achieving an R^2 of ~0.80 on test data. This level of accuracy is significant, indicating that our feature set captures the majority of factors influencing UHI intensity. Ensemble tree methods were critical to model the non-linear relationships in the data, far outperforming a linear model.

- **Feature importance analysis** of the model aligned with urban climate theory: areas with higher building density and larger built surfaces are hotter, while vegetation presence cools areas down. The model's top predictors (building metrics, NDVI/LST, etc.) provide evidence-based confirmation of these drivers and quantify their effects. This helps urban planners prioritize interventions (e.g., focusing on adding greenspace in dense neighborhoods).

- We performed **scenario simulations** showing that realistic interventions – like modest increases in tree canopy or widespread adoption of reflective surfaces – can reduce local temperatures on the order of 0.1°C. While seemingly small, such reductions on a city scale can lower heat health risks and improve comfort. These simulations offer a data-driven estimate of the benefits of common heat mitigation strategies, supporting their implementation.

- The project provides **actionable insights for urban planning and public policy**. We identified hotspot locations that deserve immediate attention (for instance, specific Bronx neighborhoods), and recommended a suite of mitigation measures (urban greening, cool materials, ventilation corridors, etc.). Our high-resolution UHI maps and predictions can be used by city officials to make targeted decisions – such as where to plant trees or where to open cooling centers thereby optimizing resource allocation in heat adaptation efforts.

- Importantly, this study highlights the power of **open science and open data** in addressing climate challenges. All the datasets used (satellite, open GIS, crowdsourced temperature data) are publicly available, and the methods we utilized are open-source. This means the work can be reproduced, audited, and improved by others. It also means that cities with limited resources can potentially leverage similar techniques to analyze their own heat island issues without proprietary technology.

Conclusion

Climate change is turning ordinary hot days into dangerous ones, and city streets are feeling that heat first. Urban Heat Islands make some neighborhoods noticeably hotter than others, increasing health risks, straining energy systems, and worsening air quality for the people who live and work there. That reality is urgent, but it is also solvable.

This project shows that modern data tools high-resolution satellite imagery, building maps, weather readings, and machine learning let us see heat at the scale where people experience it: the block, the park, the sidewalk. With those tools, we can not only point to the hottest places but also test which fixes are likely to work: more trees, reflective roofs and pavements, better building spacing, and targeted cooling infrastructure. Treating buildings, vegetation, weather, and human needs together give a fuller, more practical picture than any single data source could.

What we learned in New York applies elsewhere: dense, low-vegetation areas are the obvious first targets for cooling. The next steps are straightforward: add nighttime measurements, layer in social vulnerability data, and put the predictive model into everyday systems like heat-warning services or urban planning tools. But technology alone won't finish the job. Real change requires political will and sustained investment to plant trees, retrofit roofs, redesign streets, and support communities most at risk.

This report offers both science and a working prototype to guide those choices. If cities act on these insights, they can reduce heat exposure, lower energy bills, protect public health, and make neighborhoods more livable as temperatures rise.

References

- Oke, T.R. (1982). *The energetic basis of the urban heat island*. Quarterly Journal of the Royal Meteorological Society, **108**(455), 1–24. DOI: 10.1002/qj.49710845502

- Stone, B. (2012). *The City and the Coming Climate: Climate Change in the Places We Live*. Cambridge University Press.

- Zhou, W., Qian, Y., Li, X., Li, W., & Han, L. (2019). *Spatial–temporal dynamics of urban heat island and global background temperature under changing climates*. **Environmental Research Letters, 14**(12), 124001. DOI: 10.1088/1748-9326/ab4b19

Appendices

Appendix A – Data Processing Pipeline Description

This appendix provides a more detailed description of the PySpark data processing pipeline used in this project, outlining the steps and code components for reproducibility:

1. **Data Extraction from Satellite Imagery:** We utilized the rasterio and pyspark libraries in combination to read geoTIFF files (Landsat and Sentinel-2 images). Using GDAL drivers, the imagery was loaded as numpy arrays for each band, then parallelized by creating Spark DataFrames where each row corresponds to a pixel and contains columns for latitude, longitude (derived via the image's affine transform), and spectral values (B01, B04, B06, B08, B11). This was accomplished with the help of the RasterFrames extension for Spark, which simplifies handling of raster data in DataFrames.

2. **Feature Computation (Spark UDFs):** We wrote user-defined functions (UDFs) in PySpark to compute NDVI, NDBI, NDWI, NDSI, etc., from the bands for each pixel row. For instance, ndvi = (B08 - B04) / (B08 + B04). These UDFs were applied across the DataFrame to create new columns for each index. Similarly, LST was computed from the thermal band using a conversion formula (applied outside Spark due to its complexity, then joined back in).

3. **Spatial Join with UHI Points:** The UHI measurement points were loaded into a Spark DataFrame. We then performed a **spatial join** between the UHI points and the nearest pixel in the satellite DataFrame. Because our points were dense and aligned with pixel centers, a nearest-neighbor join sufficed: essentially we rounded each point's coordinates to the nearest pixel grid coordinate to retrieve the pixel's values. This was done using Spark SQL operations (joining on matching or nearest lat/long). The result was a DataFrame of UHI points where each point now had columns for all the spectral bands and indices from the imagery.

4. **Spatial Aggregation for Building Data:** The building footprints (thousands of polygons) were loaded using a GIS library (geopandas or directly via Spark if possible). We computed building metrics by a two-step process: first, assign each UHI point an identifier, then use a buffer operation to cut a 50m/100m/200m circle around each point and perform a spatial join with the building layer to count and sum areas. In PySpark, a straightforward way was to collect the building data in a broadcast variable (since 11k points isn't huge, we could also do this in pure Python by iterating each point's buffer over buildings). However, we leveraged **geospatial indexing**: using an R-tree or Spark's built-in spatial join if available (via something like esri-geometry-api). The output gave us tables of point ID with building_count and area for each radius, which we merged back into the main DataFrame.

5. **Incorporating Weather Data:** The two weather station readings were small enough to just broadcast as a dictionary keyed by station location. Each point's nearest_station was already determined (by a pre-processing step comparing distances to the Manhattan vs Bronx station). We simply mapped those keys to the actual weather values (temperature, humidity, wind, etc.) and added those as columns.

6. **Caching and Export:** The final Spark DataFrame containing all features and the UHI index was cached in memory to speed up subsequent operations (like EDA or model export). We then saved it as a Parquet file (uhi_dataset.parquet) and also exported to CSV (uhi_data_final.csv) for use in local Python/Notebook modeling. The Parquet format preserves schema (data types) and is more efficient for large-scale reading if needed later.

7. **Execution Environment:** This pipeline was run in a Python environment with PySpark configured for local mode (4 cores) for prototyping. For a full city-wide scale-out (predicting UHI on every pixel, not just sample points), we would deploy this on a cloud cluster (e.g., AWS EMR or Databricks) to handle the hundreds of millions of pixels in imagery. The code was written to be scalable – avoiding collecting large data to the driver, using vectorized operations where possible, and partitioning data to balance workloads.

Pseudo-code for core part of pipeline (in Pythonic pseudocode):

```
# Pseudo-code for pipeline steps
# 1. Load data
sat_df = spark.read.raster("Landsat.tif", bands=["B10","B11"])  # using RasterFrames
sat_df = sat_df.withColumn("LST", compute_LST_udf("B10","B11"))
sat_df = sat_df.join( spark.read.raster("Sentinel.tif", bands=["B01","B04","B06","B08","B11"]), on=["x","y"] )
# 2. Compute indices
sat_df = sat_df.withColumn("NDVI", ndvi_udf("B08","B04")) \
    .withColumn("NDBI", ndbi_udf("B11","B08")) \
    .withColumn("NDWI", ndwi_udf("B08","B11")) \
    .withColumn("BSI", expr("- NDVI"))  # BSI as inverse of NDVI
# 3. Load UHI points and join
uhi_df = spark.read.csv("UHI_points.csv", schema=uhi_schema)  # contains lat, lon, UHI_index
# Assuming sat_df has columns "lat" and "lon" for pixel center
uhi_with_sat = uhi_df.join(sat_df, on=[nearest_lat_lon_match], how="left")
# 4. Add building features
buildings = geopandas.read_file("buildings.shp")
for r in [50,100,200]:
    # Count and sum area of buildings within r of each point
    counts, areas = spatial_count_sum(buildings, uhi_df, radius=r)
    uhi_with_sat = uhi_with_sat.join(counts, on="point_id").join(areas, on="point_id")
# 5. Add weather
weather = {"Manhattan": {"humidity":46.7, "wind_speed":3.4,...}, "Bronx": {...}}
uhi_with_sat = uhi_with_sat.withColumn("humidity", map_weather_udf("nearest_station"))
# 6. Save final dataset
uhi_with_sat.write.parquet("uhi_dataset.parquet")
```

Through these steps, the pipeline prepares the full feature set needed for modeling in a reproducible and efficient manner.

Appendix B – Extended Tables and Figures

B1. Extended Feature Importance Table: Below is an extended list of feature importance from the GBT model (beyond the top 6 discussed in the report):

Feature	Importance Score
building_density_200	0.088
B01 (Coastal band)	0.087
building_area_200	0.082
LST (Land Surface Temp)	0.073
humidity_temp_interaction	0.065
solar_flux	0.053
NDVI	0.050
distance_to_building	0.045
building_density_100	0.043
B11 (SWIR band)	0.041
building_area_100	0.038
NDBI	0.030
wind_direction	0.029
B04 (Red band)	0.025
wind_speed	0.022
(other features...)	<0.02 each

Interpretation: We see NDVI just slightly lower than solar_flux in importance – still very relevant. Distance to building (which indicates openness) also has a notable contribution. Interestingly, wind_direction has some importance; perhaps the model found that certain wind directions (e.g., southerly winds vs northerly) corresponded with slightly different UHI patterns between Manhattan and Bronx. Most other features had smaller impacts individually.

B2. Sample Prediction vs Actual Table: A few representative points to illustrate model performance:

Location (Lat, Lon)	Actual UHI Index	Predicted UHI Index (GBT)
South Bronx industrial (40.817°N, -73.908°W)	1.040	1.035
Central Park (40.782°N, -73.965°W)	0.965	0.980

Location (Lat, Lon)	Actual UHI Index	Predicted UHI Index (GBT)
Midtown Manhattan (40.754°N, -73.992°W)	1.010	1.005
Yankee Stadium area (40.829°N, -73.925°W)	1.028	1.020
Riverside Drive by Hudson (40.800°N, -73.970°W)	0.990	0.996

These examples show the model typically within 0.005–0.01 of the index. The slight over-prediction for Central Park (pred 0.980 vs 0.965) might be because the model still sees some buildings around (Central Park isn't completely isolated from heat of city). Overall, these illustrate the accuracy achieved.

B3. Correlation Matrix Values: For completeness, the table below provides Pearson correlation coefficients between select features and the UHI Index (for numeric features):

Feature	Correlation with UHI Index
Latitude	+0.446
Longitude	+0.381
LST	+0.315
BSI (=-NDVI)	+0.255
NDVI	-0.255
building_density_200	+0.220
B01	+0.193
building_density_100	+0.158
NDBI	+0.157
building_density_50	+0.137
humidity	-0.050
wind_speed	-0.006

(Remaining features have smaller correlations; see Figure 3 heatmap for a comprehensive view.)

This reinforces which features are linearly correlated with UHI (though our model also captures non-linear effects beyond these values).

Appendix C – Data Dictionary

This appendix serves as a data dictionary, describing each feature (column) in the uhi_data_final.csv dataset:

- **Longitude, Latitude:** Geographic coordinates (WGS84) of the measurement point. Longitude is negative (West) in NYC, Latitude ~40.7–40.85 N.
- **datetime:** Timestamp of the measurement (most are ~2021-07-24 15:00–16:00 local time). All data is from this one-hour window.

- **UHI Index:** The target variable – unitless index of air temperature relative to city mean (see Introduction; 1.0 = city average, >1 hotter, <1 cooler). Range ~0.956–1.046 in this dataset.
- **B01, B04, B06, B08, B11:** Spectral band reflectance values from satellite imagery.
- *B01:* Coastal aerosol band (≈443 nm, Sentinel-2). Useful for haze, water detection.
- *B04:* Red band (≈665 nm, Sentinel-2). Used in NDVI; high for bare ground, low for vegetation.
- *B06:* Red-edge band (≈740 nm, Sentinel-2). Sensitive to vegetation health (transitional band between red and NIR).
- *B08:* Near-infrared band (≈842 nm, Sentinel-2). High reflectance for healthy vegetation.
- *B11:* Shortwave infrared band (≈1610 nm, Sentinel-2). Sensitive to moisture content; used in NDBI/NDWI. (Note: All band values are in reflectance units (0–10000 scaled integers or similar); higher means more reflectance in that wavelength.)
- **NDVI (Normalized Difference Vegetation Index):** Computed as (B08 – B04) / (B08 + B04). Ranges -1 to +1. Indicates vegetation greenness (positive values mean presence of vegetation; negative often water).
- **NDBI (Normalized Difference Built-up Index):** (B11 – B08) / (B11 + B08). Ranges -1 to +1. High positive values indicate built-up surfaces (concrete, etc.), negative indicates vegetated or water areas. In our data, NDBI is strongly *negatively* correlated with NDVI.
- **NDWI (Normalized Difference Water Index):** Defined here as (B08 – B11) / (B08 + B11). This essentially is the inverse of NDBI. High NDWI could indicate moisture (water bodies or well-watered vegetation). In data, NDWI = –NDBI exactly (due to formula).
- **REVI (Robust Enhanced Vegetation Index):** A variant of EVI (which usually uses Blue, Red, NIR). Here intended to reduce atmospheric effects and saturation. Our REVI values track NDVI but are scaled differently (not extensively used in analysis).
- **NDSI (Normalized Difference Snow Index):** Typically (Green – SWIR)/(Green + SWIR). Used originally for snow; in our summer context, it might highlight bright surfaces. We computed it out of curiosity using available bands (possibly using B04 vs B11). Its values range -1 to 1. Not highly impactful in model as there was no snow (it likely picked up large white rooftops slightly).
- **BSI (Bare Soil Index):** In literature, BSI = ((SWIR + Red) – (NIR + Blue)) / ((SWIR + Red) + (NIR + Blue)). Due to data limitations, our BSI column was effectively set as the negative of NDVI (since vegetated vs non-vegetated was main contrast). Thus BSI here ranges ~ -0.8 to +0.8 and is exactly BSI = -NDVI in this dataset (positive BSI means low vegetation, bare surface).
- **LST (Land Surface Temperature):** Surface skin temperature in degrees Celsius at that location, derived from Landsat thermal bands. Range ~27.7°C to 46.8°C (before outlier capping). It represents how hot the ground or rooftops were.
- **building_density_50, building_density_100, building_density_200:** Integer count of buildings whose footprints fall within 50m, 100m, 200m radius of the point, respectively. These give a sense of urban built density at different scales.
- **building_area_100, building_area_200:** Total surface area (in square meters) of building footprints within 100m or 200m radius. Larger values mean either larger or more buildings

covering area. For context, a typical city block might be ~10,000 m², so building_area_200 could be up to ~40,000–50,000 m² in dense areas (as seen in data).

- **distance_to_building:** Distance in meters to the nearest building from the point. If this is 0, the point might be on or immediately adjacent to a building edge (some measurement points could be on a road but at a building wall). A larger number (e.g., 50m) suggests the point was in an open area (park or wide road).
- **coverage_ratio_100:** Ratio (0–1) of area within 100m that is covered by buildings. Calculated as building_area_100 divided by area of circle of 100m (≈31,416 m²). For example, 0.30 means 30% of the 100m radius area is building footprint. This is redundant with building_area_100 (basically just a normalized version). In dataset, it's correlated 1.0 with building_area_100 (we kept it for interpretability but model used one of them to avoid duplication).
- **nearest_station:** Categorical, "Manhattan" or "Bronx", indicating which weather station's data was used. Essentially labels the region of the city.
- **wind_speed:** Wind speed in m/s from the nearest station at time of measurement (around 3pm). In this dataset ~3.0–3.5 m/s generally.
- **wind_direction:** Wind direction in degrees (meteorological convention, degrees clockwise from north). Around ~190–200° in data (which is a southerly wind on that day).
- **humidity:** Relative humidity in percent (%) from nearest station. Around ~45–50% that afternoon (slightly varying between Manhattan (46.7%) and Bronx (maybe ~44%)).
- **solar_flux:** Downward shortwave radiation (sunlight) in W/m² measured at station. ~605 W/m² (Bronx) to ~615 W/m² (Manhattan) in our data, indicating mostly clear skies around mid-afternoon.

Each of these features was utilized in model training (except some like coverage_ratio which were dropped due to redundancy). Understanding each feature's meaning helps interpret the model results and ensures that any data issues (e.g., NDWI being just inverse of NDBI) are known when refining the model in the future.