FTL GIS

A Capstone Project: Urban Growth Prediction using Machine Learning

Prepared By

Mikre Getu Mihrete

mikregetuinfo@gmail.com

Ethiopia

August 2, 2025

Table of Contents

List of Tables

List of Figures

1. Objective and SDG Relevance

The accelerating pace of global urbanization necessitates advanced tools for monitoring land use change and fostering sustainable development. Unplanned urban sprawl often leads to environmental degradation, loss of natural habitats, and increased pressure on infrastructure and resources. Therefore, accurate and timely information on urban extents is critical for effective urban planning and policy-making.

This project's primary objective is to develop and evaluate a machine learning model that accurately predicts urban expansion from multi-spectral Sentinel-2 satellite imagery. By identifying and mapping urban extents, this initiative directly supports **Sustainable Development Goal 11 (SDG 11)**: Sustainable Cities and Communities. The outputs provide crucial data for informed urban planning, aiding in the management of urban sprawl and reducing the potential for slum proliferation.

Furthermore, accurate urban mapping is vital for resource management, infrastructure development, and monitoring the environmental impact of urban sprawl. The "Urban Risk Map" generated by the model offers a powerful tool for planners to identify areas with a high probability of future urbanization. This project also indirectly supports **Sustainable Development Goal 13 (SDG 13)**: Climate Action by highlighting how urbanization affects local ecosystems and green spaces, thereby guiding mitigation strategies to preserve natural landscapes.

2. Study Area Description

This project focuses on **Bilbao, Spain**, a prominent industrial city located in the Basque Country of northern Spain. The study area specifically encompasses the region surrounding the **Port of Bilbao**, where the Nervión River meets the Cantabrian Sea.

The diverse landscape of this coastal region makes it an ideal and challenging environment for evaluating a land cover classification model designed for urban growth prediction. Based on the false-color composite imagery used in this study (where near-infrared light is displayed as red), the key features of the study area include:

- **Urban and Industrial Areas:** These appear as distinct cyan, white, and light gray patches. They are heavily concentrated along the Nervión River and around the expansive Port of Bilbao, reflecting the region's significant history as a major industrial and shipping hub.
- **Vegetation:** Various shades of red highlight healthy vegetation in the hills and natural landscapes that surround the urban core and port facilities.
- **Water Bodies:** The Nervión River estuary and the larger Bay of Biscay are clearly visible as large, dark blue or black areas, emphasizing the strong coastal influence on the region's geography.
- **Bare Soil/Other:** Lighter, non-red areas mixed within the landscape likely represent bare soil or other non-vegetated surfaces.

This blend of dense urban infrastructure, a crucial international port, and notable natural landscapes provides a rich and complex environment for analyzing land use and land cover changes, particularly in the context of urban expansion.

3. Data Sources and Preprocessing

This project heavily relies on **multi-spectral raster imagery** obtained from the European Space Agency's (ESA) **Sentinel-2 satellite mission**, part of the Copernicus Programme. Sentinel-2 data is well-suited for land cover analysis due to its high spatial and temporal resolution.

3.1. Data Sources

The dataset specifically includes three key Level-1C products:

- **Training Features**: ***`S2A_MSIL1C_20220516_TrainingData.tif`*** is a 4-band image (Blue, Green, Red, Near-Infrared) captured by Sentinel-2A on **May 16, 2022**. This image serves as the primary feature set, providing the spectral information necessary to train the machine learning model.
- **Training Labels (Ground Truth)**: ***`S2A_MSIL1C_20220516_Train_GT.tif`*** is a corresponding single-band raster from the same date, representing the **ground truth** data. This file contains integer values that categorize pixels into eight distinct land cover classes (e.g., Water, Bare Soil, Dense Vegetation, Cropland, Urban/Built-up, Sparse Vegetation,

Urban, Cloud/Shadow). The model learns to associate the spectral signatures from the training features with these predefined land cover classes.

- **Test Features**: ***`S2B_MSIL1C_20220528_Test.tif`*** is another 4-band image captured by Sentinel-2B on **May 28, 2022**. This image represents unseen data and is used to evaluate the trained model's ability to predict urban expansion in a new, unclassified scenario.

3.2. Data Preprocessing

To prepare the raw Sentinel-2 raster data for the machine learning model, the following critical steps were executed using Python, primarily leveraging the ***`rasterio`*** and ***`scikit-learn`*** libraries:

- **Data Reshaping**: The initial 3D raster arrays, structured as (bands, height, width), were transformed into a 2D array format. This new structure (pixels, bands) is a prerequisite for `scikit-learn`, where each row represents an individual pixel sample and each column corresponds to a spectral band value. For the training data, this reshaping resulted in an initial array shape of **967,005 pixels by 4 bands**.
- **NoData Filtering**: Pixels identified with a ground truth value of `0` were considered `NoData` (invalid or unclassified) and systematically removed from both the feature set (X) and the corresponding labels (y). This crucial step ensured that the model trained exclusively on valid, classified pixels. After filtering, **965,589 valid pixels** remained for analysis, accounting for approximately **99.85%** of the total initial pixels.
- **Train-Validation Split**: The filtered dataset of valid pixels was then randomly divided into two subsets: a **training set** and a **validation set**. A standard 80/20 split was applied, allocating **80% of the data (772,471 pixels)** for model training and **20% (193,118 pixels)** for evaluating the model's performance on unseen data before final prediction. This split is essential for assessing the model's generalization capabilities and preventing overfitting.

4. Geospatial Analysis Summary

The geospatial analysis in this project is fundamentally built upon a **pixel-based supervised classification** framework. This workflow systematically utilizes a ground truth map to train a machine learning model, specifically a **Random Forest classifier**, to discern the intricate relationships between distinct spectral signatures (pixel values across the Blue, Green, Red, and Near-Infrared bands) and various land cover classes.

Once successfully trained, this model gains the capability to process new, unclassified satellite imagery of the same geographic area. It then assigns a predicted land cover class to every individual pixel within the new image, thereby generating a comprehensive, classified map. This automated approach significantly streamlines the process of detailed land cover mapping.

A key output of this geospatial analysis, beyond a standard classification map, is the generation of **derivative products**, notably a **probability-based "Urban Risk Map."** This unique map quantifies the likelihood of each pixel belonging to the urban class, providing a more nuanced understanding of potential urban growth rather than a simple binary classification.

The land cover classes identified and analyzed within this project, inferred from the provided ground truth data and the model's performance, are detailed as shown in Table 1.

Table 1. Land Cover Classes and Model Performance Insights

Class ID	Inferred Description	Support (Validation Pixels)	Key Characteristics & Model Performance Insights
10	Water	11645	Achieved an F1-score of 0.68, with a precision of 0.73 and recall of 0.65. Exhibited distinct Near-Infrared absorption.
20	Bare Soil	3163	Had an F1-score of 0.58, with precision 0.61 and recall 0.55. Showed spectral confusion with urban areas due to low vegetation.
30	Dense Vegetation	21537	Achieved an F1-score of 0.72, with precision 0.69 and recall 0.74. Exhibited high Near-Infrared reflectance, though 2,977 pixels were incorrectly classified as Urban/Built-up.
40	Cropland	844	A minority class that struggled significantly, with a low F1-score of 0.20 (precision 0.48, recall 0.13), indicating the model often failed to identify it correctly due to limited samples.
50	Urban / Built-up	29055	A primary target class, performing strongly with an F1-score of 0.83 (precision 0.79, recall 0.88).

60	Sparse Vegetation	3013	Had an F1-score of 0.37, with precision 0.66 and recall 0.26. Often mixed spectrally with soil and urban areas.
80	Urban	123504	A key target class and majority class, demonstrating exceptional performance with an F1-score of 0.99 (precision 0.99, recall 0.99), indicating near-perfect classification.
90	Cloud/Shadow	357	A significant minority class with very low accuracy, yielding an F1-score of 0.21 (precision 0.32, recall 0.16) due to scarcity of training samples.

The geospatial analysis culminated in two crucial visual outputs: a discrete land cover classification map, represented by the **Predicted Urban Growth map** (Figure 5), and a probabilistic **"Urban Risk Map"** (Figure 6). The latter effectively highlights areas at the highest risk of future urban development, offering more nuanced and valuable insights for urban planning than a simple classification alone.

5. Methodology and Result

For this project, a **Random Forest (RF) Classifier** was chosen as the machine learning model. This ensemble learning method is particularly well-suited for remote sensing applications due to several key advantages:

- **Robustness to Overfitting:** It's highly resistant to overfitting, especially when configured with a large number of trees (in this case, ***`n_estimators=100`***).
- **High-Dimensional Data Handling:** The Random Forest algorithm can efficiently process high-dimensional data, which is typical for multi-band satellite imagery.
- **Feature Importance and Class Probabilities:** It provides valuable metrics such as feature importance and class probabilities, which are crucial for understanding model behavior and generating nuanced outputs like the Urban Risk Map.
- **Parallelization:** The model can be parallelized (using ***`n_jobs=-1`*** in Python's scikit-learn library) to significantly speed up training.

The model was implemented using Python's ***scikit-learn*** library. Training was performed on a dataset of **772,471 pixels** from the training set, and this process was completed in **89.34 seconds**. To handle the large test raster effectively and prevent memory errors, a batch processing approach was implemented. The test image was divided and processed in **47 batches**, each containing **100,000 pixels**.

The model generated two primary outputs:

- A final classification map, saved as ***predicted_urban_growth.tif***.
- A probability map specifically for the urban class (label 50), saved as ***urban_risk_map.tif***.

5.1. Results & Evaluation

The Random Forest model's performance was rigorously evaluated using a dedicated validation set, yielding strong results:

- **Overall Accuracy:** The model achieved a notable **overall accuracy of 90.08%** on the validation set, which comprised **193,118 pixels**.

The detailed per-class performance is summarized in the classification report below, with "Class ID" corresponding to the labels in the confusion matrix (Figure 2 Random Forest Confusion Matrix). The report highlights precision, recall, and F1-score for each land cover class:

	precision	recall	f1-score	support
Water	0.73	0.65	0.68	11645
Bare Soil	0.61	0.55	0.58	3163
Dense Vegetation	0.69	0.74	0.72	21537
Cropland	0.48	0.13	0.20	844
Urban / Built-up	0.79	0.88	0.83	29055
Sparse Vegetation	0.66	0.26	0.37	3013
Urban	0.99	0.99	0.99	123504
Cloud/Shadow	0.32	0.16	0.21	357
accuracy			0.90	193118
macro avg	0.66	0.54	0.57	193118
weighted avg	0.90	0.90	0.90	193118

Figure 1. Classification Report

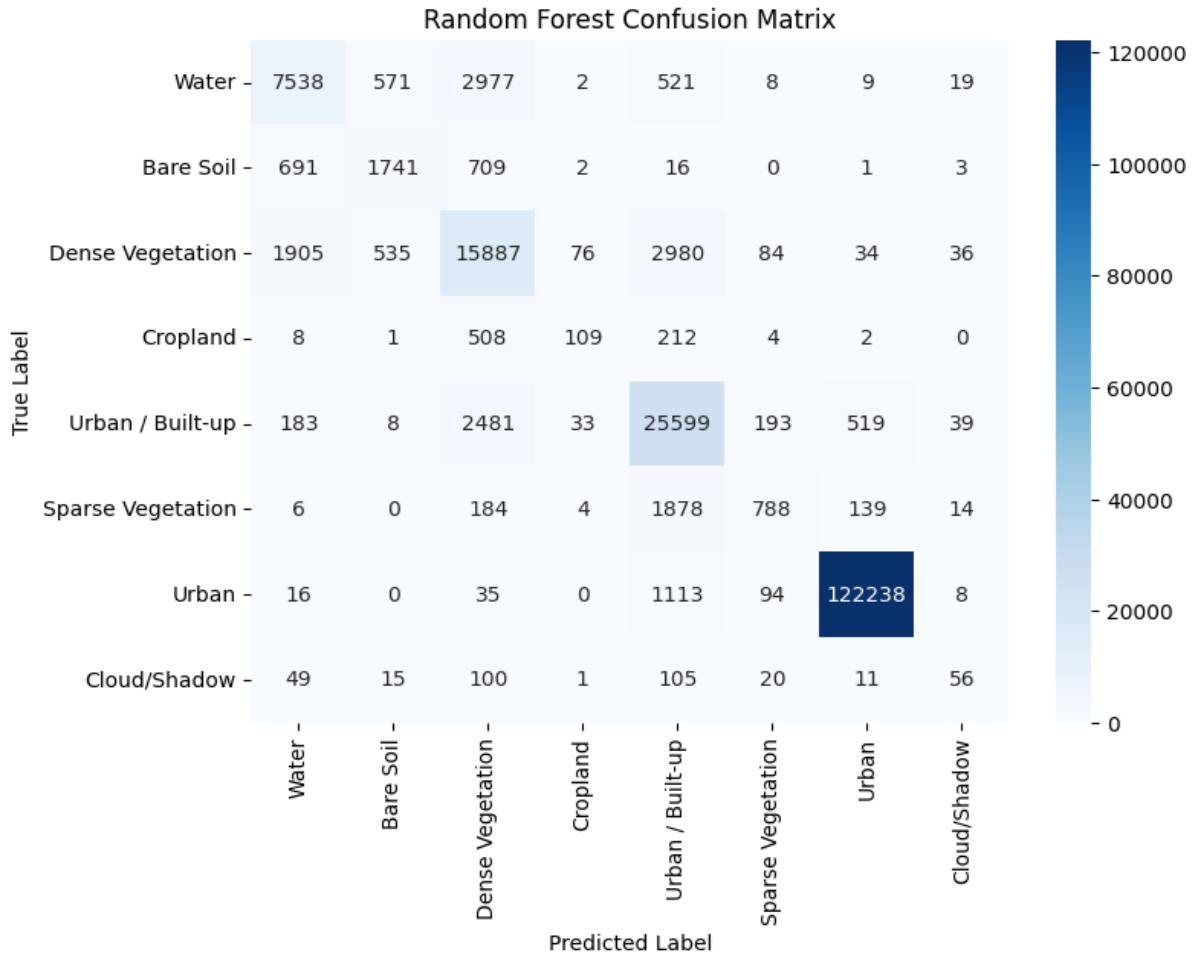

Figure 2. Confusion Matrix

Key observations from the classification report and confusion matrix (as depicted in Figure 1 and Figure 2.):

- **Excellent Performance for Urban Classes:** The model demonstrated exceptional performance for the **Urban (Class 80)** and **Urban / Built-up (Class 50)** classes, which are the primary targets of this study. Specifically, Class 80 achieved a near-perfect **F1-score of 0.99** (with 99% precision and 99% recall), **correctly classifying 122,238 pixels**. Class 50 also showed strong results with an **F1-score of 0.83** (79% precision and 88% recall), **correctly classifying 25,599 pixels**.
- **Strong Performance for Other Major Classes:** The model performed well on **Water (Class 10)**, **correctly classifying 7,538 pixels** and achieving an F1-score of 0.68, and **Dense Vegetation (Class 30)**, with **15,887 correctly classified pixels** and an F1-score of 0.72.

- **Struggles with Minority Classes:** The model notably struggled with **minority classes** such as **Cropland (Class 40)** and **Cloud/Shadow (Class 90)**. This is clearly indicated by their very low recall scores (13% for Cropland, with only **109 correctly classified pixels**, and 16% for Cloud/Shadow, with just **56 correctly classified pixels**). This suggests the model often failed to identify these classes correctly, a direct result of class imbalance within the training data.
- **Spectral Confusion:** The confusion matrix visually confirmed these findings. While strong diagonal values for classes like 50 and 80 indicate high accuracy, off-diagonal values highlight areas of spectral confusion. For instance, **2,977 pixels of Dense Vegetation (Class 30)** were incorrectly classified as **Urban / Built-up (Class 50)**, and conversely, **2,980 pixels of Urban / Built-up (Class 50)** were misclassified as **Dense Vegetation (Class 30)**. Similarly, **709 pixels of Bare Soil (Class 20)** were incorrectly identified as **Dense Vegetation (Class 30)**. These misclassifications are likely due to spectrally similar materials or mixed pixels (e.g., parks within urban areas).

6. Visualization Outputs

The project generated a comprehensive set of visual outputs, effectively summarizing the entire workflow from input data to final predictions. These visualizations are primarily organized into results_visualization.png (a 2x2 grid of maps) and a confusion_matrix.png (Figure 2), along with individual figures for clarity.

This composite image provides a side-by-side comparison of the input data, the ground truth used for training, and the model's key outputs.

A) Original Test Imagery (False-Color Composite)

Figure 3 displays the study area of Bilbao, Spain, as a false-color composite derived from Sentinel-2 data (specifically, likely using Near-Infrared, Red, and Green bands). This representation enhances the visibility of different land cover types:

- ✓ **Dense vegetation** appears prominently in **red**.
- ✓ **Urban areas** are highlighted in **cyan/white**.
- ✓ **Water bodies** are distinctly rendered in **dark blue**. This visualization provides a clear initial view of the complex interplay between urban, natural, and coastal environments

characteristic of the region. A percentile-based stretching function was applied during visualization to improve contrast by handling the wide range of pixel values effectively.

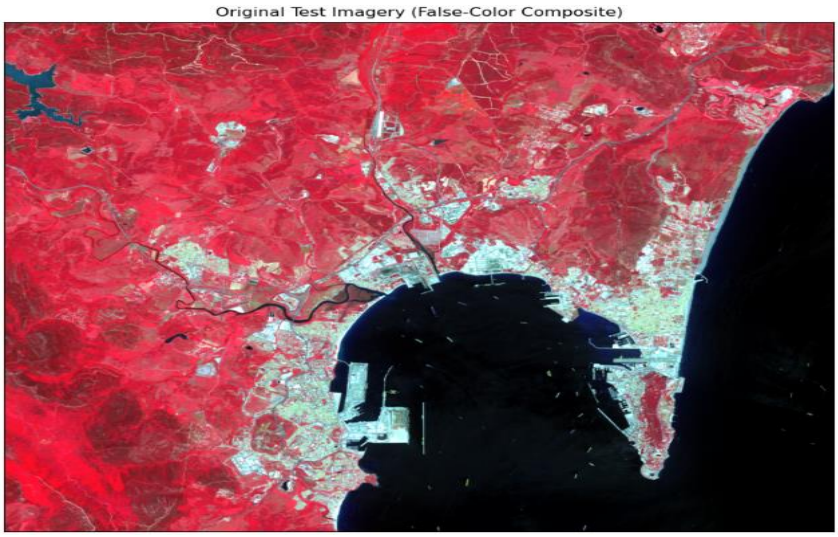

Figure 3. Original Test Imagery (False-Color Composite)

B) Training Ground Truth

Figure 4, displays the manually labeled ground truth data that was used to train the machine learning model. It serves as the reference land cover map for the study area, where each pixel is assigned a specific integer label corresponding to one of the eight distinct land cover classes (e.g., urban, water, vegetation). The accuracy and detail of this ground truth data were fundamental for the model's ability to learn and correctly classify different land cover types.

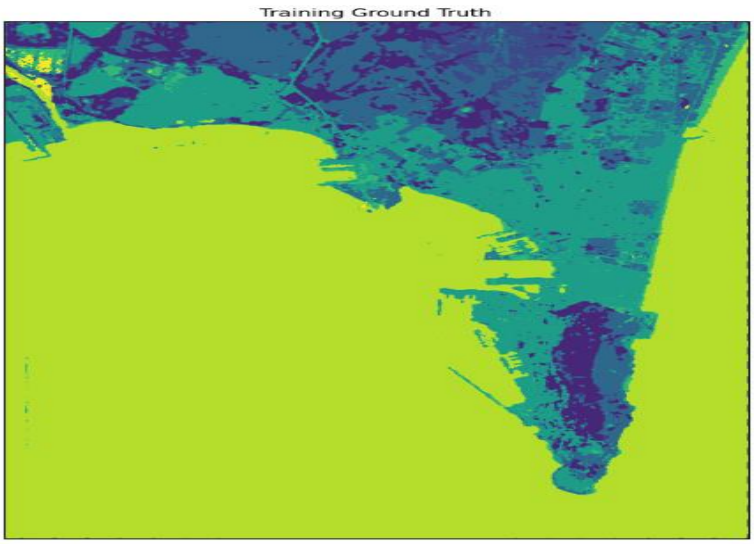

Figure 4. Training Ground Truth

C) Predicted Urban Growth

Figure 5, represents the final discrete land cover classification output generated by the trained Random Forest model on the unclassified test imagery. Each pixel is assigned a single, definitive land cover class. The map visually depicts the model's prediction of various land cover types, including a clear distinction between urban and non-urban areas, reflecting the direct result of the classification process.

Figure 5. Predicted Urban Growth

D) Urban Risk Map (Probabilities)

This is one of the most valuable outputs of the project, as shown in Figure 6. Instead of a discrete class, this map presents the probability of each pixel belonging to an urban area. A color gradient is used to represent these probabilities:

- ✓ **Dark blue/purple areas** indicate a **low probability** of urbanization (values closer to 0.0).
- ✓ **Bright yellow areas** signify a **high likelihood** of urbanization (values closer to 1.0). These high-probability yellow areas are often found at the edges of existing urban centers, identifying regions with the highest risk of future urban development. This probabilistic output offers a more nuanced and actionable tool for urban planners, as it highlights areas in transition or under pressure for future development, going beyond a simple categorical classification.

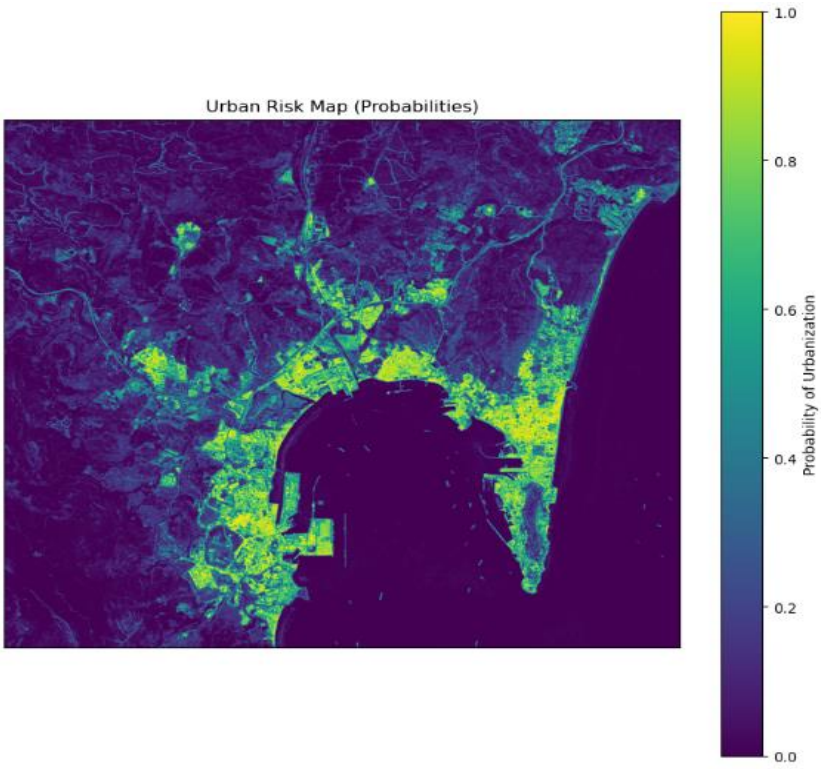

Figure 6. Urban Risk Map (Probabilities)

The heatmap, as shown in Figure 2, visually summarizes the performance of the Random Forest model on the validation set. It depicts the number of pixels for each land cover class that were correctly and incorrectly classified:

- **Correct Classifications (Diagonal):** The strong dark blue values along the main diagonal, particularly for the "Urban" (Class 80) and "Urban / Built-up" (Class 50) classes, confirm the model's high accuracy in identifying these dominant classes. For example, 122,238 pixels were correctly classified as "Urban" and 25,599 pixels as "Urban / Built-up."
- **Misclassifications (Off-Diagonal):** The off-diagonal values highlight instances of spectral confusion between classes. For example:
  - ✓ **2,977 pixels of "Dense Vegetation"** were incorrectly classified as "Urban / Built-up."
  - ✓ **2,980 pixels of "Urban / Built-up"** were incorrectly classified as "Dense Vegetation."
  - ✓ **709 pixels of "Bare Soil"** were misclassified as "Dense Vegetation."
- **Minority Class Performance:** The matrix also visually reinforces the model's struggles with minority classes like "Cropland" and "Cloud/Shadow," which show very low numbers on their respective diagonal cells (e.g., 109 and 56 correctly classified pixels, respectively) and more prominent off-diagonal values, indicating frequent misclassification.

7. Challenges and Future Work

This project encountered several challenges inherent in remote sensing and machine learning for land cover classification, which in turn led to valuable learnings.

7.1. Challenges

Severe Class Imbalance: The most significant challenge was the **severe class imbalance** within the training dataset. This disproportionate representation of different land cover types directly impacted the model's performance on minority classes.

- ✓ **Concrete Numbers:** For instance, "Cropland" (Class 40) had a support of only **844 pixels** in the validation set, and "Cloud/Shadow" (Class 90) had an even smaller support of **357 pixels**. In stark contrast, "Urban" (Class 80) accounted for a large majority with **123,504 pixels**.
- ✓ **Impact:** This imbalance led to significantly poorer performance on these underrepresented classes. The model's recall for "Cropland" was a mere **13% (only 109 correctly classified pixels)**, and for "Cloud/Shadow," it was an even lower **16% (only 56 correctly classified pixels)**. This indicates the model frequently failed to identify these classes, often misclassifying them as other, more dominant land cover types.

Spectral Confusion/Ambiguity: The model exhibited **spectral confusion** between certain land cover types due to their similar spectral signatures or the presence of mixed pixels.

- ✓ **Concrete Numbers:** The confusion matrix clearly illustrates this. For example, **2,977 pixels of "Dense Vegetation" (Class 30)** were incorrectly classified as "Urban / Built-up" (Class 50), and conversely, **2,980 pixels of "Urban / Built-up" (Class 50)** were misclassified as "Dense Vegetation" (Class 30). Similarly, **709 pixels of "Bare Soil" (Class 20)** were incorrectly identified as "Dense Vegetation" (Class 30). This confusion likely arises from areas where urban structures are interspersed with vegetation (e.g., parks, urban green spaces) or where bare soil has similar reflectance properties to certain built-up materials.
- ✓ **Additional Specifics:** The presence of "legacy industrial sites" (e.g., abandoned factories) could further skew reflectance signatures, adding to the complexity of distinguishing urban areas.

Computational Constraints (Memory Management): Processing large raster files, typical of high-resolution satellite imagery, presented a significant **memory management challenge**.

- ✓ **Impact:** This could lead to `MemoryError` during processing, requiring careful handling of the data.
- ✓ **Solution Implemented:** This challenge was successfully addressed by implementing a **batch processing workflow**. The test raster was processed in **47 smaller batches, each consisting of 100,000 pixels**, which prevented memory overload and allowed for efficient processing of the entire dataset.

Data Visualization (Initial Attempts): An initial hurdle in data visualization was encountered where attempts to display the raw satellite data resulted in a completely white image.

- ✓ **Reason:** This issue stemmed from the wide range of pixel values present in the raw imagery, causing extreme outlier values to dominate the display.
- ✓ **Solution Implemented:** This was resolved by implementing a **percentile-based stretching function**, which effectively improves contrast by ignoring these extreme outlier values and mapping the more representative data range to the displayable color scale.

7.2. Learnings

Through addressing these challenges, several key learnings emerged, which are crucial for future geospatial machine learning projects:

A. **Importance of Per-Class Metrics**

Evaluating model performance solely on overall accuracy can be misleading, especially with imbalanced datasets. A crucial learning was the importance of using **per-class metrics** such as precision, recall, and F1-score (as seen in the detailed classification report) to gain a nuanced understanding of the model's strengths and weaknesses for each individual land cover type. This allows for targeted improvements, particularly for minority classes.

B. **Value of Probabilistic Outputs**

Generating **probabilistic maps**, such as the "Urban Risk Map" (Figure 5), provides significantly more actionable and nuanced results than a simple, hard classification map.

- ✓ **Benefit:** Instead of just a binary "urban" or "non-urban" label, the probability map quantifies the likelihood of urbanization, highlighting areas in transition or at the highest risk of future development (represented by brighter yellow areas in the Urban Risk Map). This is invaluable for urban planners, enabling more informed and sustainable decision-making.

C. **Strategic Data Preprocessing and Handling**

Gained practical experience in robust data preprocessing techniques for geospatial data, including:

- ✓ **Data Reshaping:** Effectively transforming 3D raster arrays (bands, height, width) into 2D arrays (pixels, bands) for compatibility with machine learning models.
- ✓ **NoData Filtering:** Systematically removing invalid data pixels (those with a ground truth label of 0), which improved data quality and model training efficiency (retaining **965,589 valid pixels** out of 967,005).
- ✓ **Batch Processing:** The successful implementation of batch processing for large rasters is a critical learning for managing computational resources and processing efficiency in large-scale remote sensing analyses.

D. **Continuous Improvement through Feature Engineering and Advanced Models**

The project highlighted the potential for future enhancements through:

- ✓ **Feature Engineering:** Recognizing the benefit of incorporating additional spectral indices (e.g., NDVI for vegetation, NDBI for built-up areas) to provide the model with more discriminative information, which can help overcome spectral ambiguity.
- ✓ **Exploring Advanced Models:** Acknowledging that while Random Forest performed well, investigating more sophisticated models like Gradient Boosting Machines (e.g., XGBoost) or Deep Learning architectures (e.g., Convolutional Neural Networks), which are adept at learning complex spatial-spectral patterns, could yield even higher accuracy.

E. **Understanding and Addressing Class Imbalance is Key**

The consistent underperformance on minority classes underscored that **addressing class imbalance is paramount** for building a truly robust and equitable classifier across all land cover types. This learning directly informs future recommendations, such as employing techniques like SMOTE or class weighting.

8. Conclusion and Recommendations

This project successfully established and validated a robust machine learning framework for monitoring urban land cover and predicting urban growth using Sentinel-2 imagery. The implemented Random Forest model proved highly effective in the challenging, heterogeneous coastal-urban environment of Bilbao, Spain, achieving a strong overall accuracy of 90.08%. The model demonstrated particular excellence in identifying dominant land cover types such as "Urban" (Class 80, with an F1-score of 0.99) and "Urban / Built-up" (Class 50, with an F1-score of 0.83).

A significant outcome of this study is the generation of a probabilistic urban risk map, which serves as a powerful, data-driven tool. This map provides urban planners with a more nuanced understanding of potential future urbanization, highlighting areas at high risk for development (as indicated by brighter yellow areas in the visualization outputs), thereby enabling more informed and sustainable urban planning decisions than traditional discrete classification maps. This research underscores the considerable potential of integrating open-access remote sensing data with advanced machine learning techniques to support proactive and sustainable urban development initiatives.

To further enhance the model's performance and address the identified limitations, the following recommendations are put forth for future work:

- **Address Class Imbalance:** The observed poor performance on minority classes (e.g., "Cropland" and "Cloud/Shadow," with recall scores of 13% and 16% respectively) due to severe class imbalance in the training data must be systematically addressed. Future work should implement data-level techniques such as **SMOTE (Synthetic Minority Over-Sampling Technique)** to create synthetic samples for underrepresented classes, or algorithm-level methods like using **class weighting** within the model

(***`class_weight='balanced'`***) to assign greater importance to minority classes during the training process.

- **Feature Engineering:** To mitigate spectral confusion between land cover types with similar signatures (e.g., "Urban / Built-up" and "Dense Vegetation"), the feature set provided to the model should be enhanced. This can be achieved by incorporating **additional spectral indices**, such as the **Normalized Difference Vegetation Index (NDVI)** to better differentiate vegetation, and the **Normalized Difference Built-up Index (NDBI)** to specifically highlight built-up areas. These indices provide more discriminative information, potentially improving classification accuracy.
- **Explore Advanced Models:** While the Random Forest classifier performed admirably, investigating more sophisticated machine learning models could yield even higher accuracy and better capture complex spatial-spectral patterns. This includes exploring **Gradient Boosting Machines (e.g., XGBoost)**, known for their strong predictive power, or **Deep Learning architectures, particularly Convolutional Neural Networks (CNNs)**, which are inherently designed to learn hierarchical features from image data and have shown remarkable success in remote sensing image classification tasks. Additionally, incorporating external data layers like population density, road networks, or topography could further improve model robustness.
- **Temporal Analysis:** Applying the developed model to a time-series of Sentinel-2 imagery would enable the monitoring of urban growth patterns over several years. This temporal analysis could provide invaluable insights into the dynamics of urbanization, aiding in long-term urban planning and environmental impact assessments.
- **Hyperparameter Tuning:** Systematically optimizing the hyperparameters of the chosen machine learning model, such as `n_estimators` and `max_depth` for the Random Forest, through techniques like `GridSearchCV` or `RandomizedSearchCV`, could lead to further performance gains.

9. References

[1]. European Space Agency, Copernicus Sentinel-2 Level-1C Data. (2022). https://sentinels.copernicus.eu/sentinel-data-access/sentinel-products/sentinel-2-data-products/collection-1-level-1c

[2]. Python 3: https://www.python.org/

[3]. Matplotlib: https://matplotlib.org/

[4]. NumPy: https://numpy.org/

[5]. Rasterio: Gillies, S., et al. (2013-2022). *Rasterio: Geospatial raster I/O for Python programmers*. https://rasterio.readthedocs.io/

[6]. scikit-learn: Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research. https://scikit-learn.org/

[7]. Seaborn: https://seaborn.pydata.org/

[8]. Abdi, A. M. (2019). Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data. *GIScience & Remote Sensing, 57*(1), 1–20. https://doi.org/10.1080/15481603.2019.1650447

[9]. Breiman, L. (2001). Random Forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

[10]. Eisavi, V., Homayouni, S., Yazdi, A.M. et al. (2015). Land cover mapping based on random forest classification of multitemporal spectral and thermal images. *Environ Monit Assess, 187*, 291. https://doi.org/10.1007/s10661-015-4489-3

[11]. Hemashreekilari. (2023, August 11). *Understanding Random Forest*. Medium. https://medium.com/@hemashreekilari9/understanding-random-forest-a87d08416280

[12]. Liaw, A., & Wiener, M. (2002). Classification and Regression by Random Forest. *R News, 2*(3), 18-22. https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf

[13]. Maxwell, A.E., Warner, T.A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: an applied review. *Int. J. Remote Sens., 39*, 2784–2817. https://doi.org/10.1080/01431161.2018.1433343

[14]. Pal, M. (2005). Random Forest Classifier for Remote Sensing Classification. *International Journal of Remote Sensing, 26*(1), 217–222. https://doi.org/10.1080/01431160412331269698

[15]. Phan, T. N., Verena K., & Lukas W. Lehnert. (2020). Land Cover Classification using Google Earth Engine and Random Forest Classifier—The Role of Image Composition. *Remote Sensing, 12*(15), 2411. https://doi.org/10.3390/rs12152411