Sprawl Monitoring in Sri Lanka

Machine Learning-Based Environmental Forecasting Dashboard
for Sri Lankan Districts

Submitted By:
PVNN Jayalath
2019/ICTS/176

Machine Learning Project
TICT4253 Intelligent Systems (Practical)

Department of Information and Communication Technology
Faculty of Technological Studies
University of Vavuniya, Sri Lanka
2026

Contents

0.1 Problem Statement

The uncontrolled horizontal expansion of urban centers in Sri Lanka, commonly referred to as "Urban Sprawl," has transitioned from a developmental milestone to a multi-faceted environmental crisis. As the boundary between rural and urban landscapes blurs, the lack of a structured, data-driven monitoring framework has led to several critical issues:

0.1.1 The Intensification of Urban Heat Islands (UHI)

The conversion of pervious, natural landscapes into impervious surfaces—such as asphalt roads and concrete high-rises—has significantly altered the thermal properties of Sri Lankan districts. These man-made materials possess high thermal bulk and low albedo, absorbing solar radiation during the day and re-emitting it as heat at night. This results in:

- **Elevated Land Surface Temperatures (LST):** A measurable spike in ground-level heat that far exceeds surrounding rural baselines.
- **Energy Inefficiency:** Increased ambient temperatures lead to a surge in demand for artificial cooling, creating a feedback loop of high energy consumption and further heat emission.
- **Health Risks:** Vulnerable urban populations face increased risks of heat-related illnesses due to localized micro-climate warming.

0.1.2 Ecological Fragmentation and Biomass Depletion

Urban sprawl directly encroaches upon critical ecosystems, including wetlands and forest covers. This degradation is quantifiable through the **Normalized Difference Vegetation Index (NDVI)**. The consequences include:

- **Loss of Carbon Sinks:** The removal of urban greenery diminishes the environment's capacity to sequester carbon dioxide, accelerating the impacts of localized climate change.
- **Hydrological Imbalance:** Reduced vegetation leads to lower evapotranspiration rates and increased surface runoff. In regions like the Western Province, this has directly correlated with the rising frequency of flash floods during monsoonal periods.

0.1.3 The Predictive Knowledge Gap

Perhaps the most significant challenge is the **Information Asymmetry** between urban planners and environmental scientists. While raw satellite imagery is abundant, there is a distinct lack of tools that can:

1. **Shift from Reactive to Proactive:** Most existing reports document environmental damage after it has occurred, rather than forecasting future risks.
2. **Provide District-Level Granularity:** Planners lack localized, ML-driven simulations that can answer questions such as, *"What will the vegetation density of Kandy be in 2040 if current sprawl rates persist?"*
3. **Translate Data into Action:** Complex indices like NDBI (Built-up Index) and NDVI are often inaccessible to non-technical stakeholders without intuitive visualization tools.

0.1.4 Demographic Pressure and Land Degradation

The inevitable increase in population density creates a relentless demand for residential and commercial infrastructure. This pressure forces development into ecologically sensitive peripheries. Without an intelligent system to monitor the correlation between **Population Growth** and **Environmental Decay**, sustainable development remains a theoretical concept rather than a practical reality.

1. Methodology

The methodology adopts a systematic data science approach, integrating remote sensing data with machine learning to model urban dynamics. The workflow is divided into four critical phases: Data Acquisition, Preprocessing, Exploratory Analysis, and Feature Engineering.

1.1 Data Sourcing and Parameter Extraction

The study utilizes multi-spectral imagery acquired via the **Google Earth Engine (GEE)** cloud platform, which allows for large-scale geospatial analysis without local computational constraints.

- **Satellite Sensors:** Primary data was sourced from the **Landsat 8-9 OLI/TIRS** (Operational Land Imager and Thermal Infrared Sensor) collections.

- **Land Surface Temperature (LST):** Extracted using the thermal bands (Band 10 and 11). A Cloud-Masking algorithm was applied to ensure the data represents clear-sky conditions, providing an accurate thermal profile of Sri Lankan districts.

- **Normalized Difference Vegetation Index (NDVI):** Calculated to quantify the health and density of vegetation cover:

$$NDVI = \frac{NIR(Band5) - Red(Band4)}{NIR(Band5) + Red(Band4)} \tag{1.1}$$

- **Normalized Difference Built-up Index (NDBI):** Developed to automate the identification of urban footprints. It leverages the high reflectance of man-made structures in the Short-Wave Infrared (SWIR) spectrum.

- **Socio-Demographic Integration:** District-wise population density data was merged with geospatial indices to create a hybrid dataset that considers both physical and human drivers of sprawl.

1.2 Data Preprocessing and Standardization

Raw satellite-derived data often contains noise or inconsistent scales. To stabilize the model training, the following steps were implemented:

- **Mean Imputation for Missing Values:** Temporal gaps in the population and temperature series were filled using mean imputation, ensuring a continuous time-series for the Random Forest algorithm.

- **Categorical Encoding:** Since machine learning models require numerical input, district names were transformed using **Label Encoding**.

- **Feature Normalization:** Environmental variables like LST (measured in Celsius) and NDVI (ranging from -1 to 1) have vastly different scales. **StandardScaler** was applied to transform the features into a standard normal distribution:

$$z = \frac{x - \mu}{\sigma} \tag{1.2}$$

Where μ is the mean and σ is the standard deviation.

1.3 Exploratory Data Analysis (EDA)

The EDA phase was conducted to validate the ”Urban Heat Island” hypothesis within the Sri Lankan context.

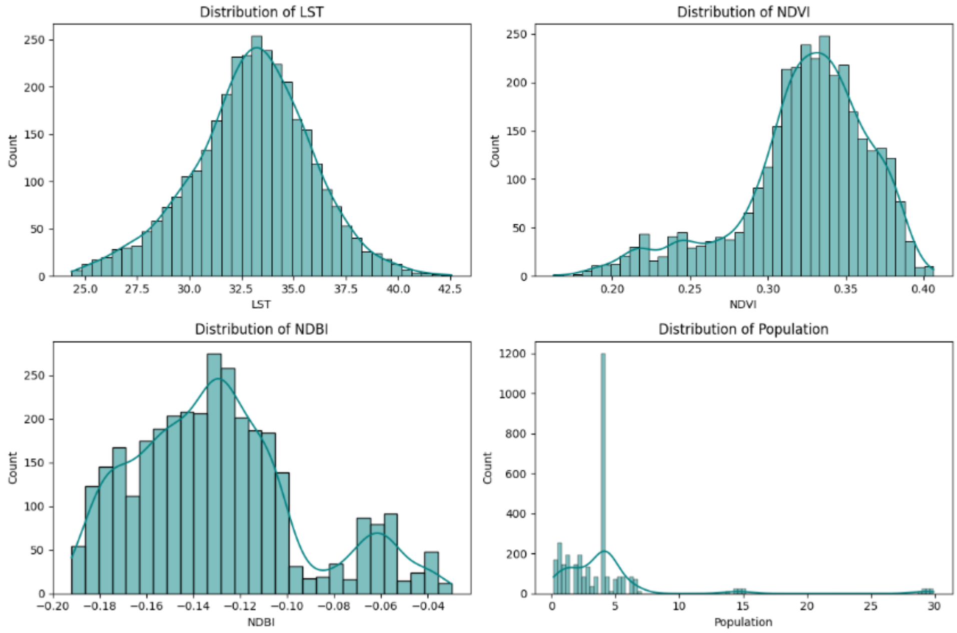


Figure 1.1: Pearson Correlation Heatmap showing relationships between LST, NDVI, and NDBI.

The statistical analysis in Figure 1.1 confirms a **strong negative correlation** between NDVI and LST. This indicates that districts with higher vegetation loss experience significantly higher ground temperatures. Conversely, the positive correlation between NDBI and LST highlights the role of concrete surfaces in heat retention.

1.4 Feature Selection and Data Leakage Prevention

A critical step in our methodology was the removal of **NDBI** from the final classification feature set.

- **The Reason:** Since the "Risk Category" (Target Variable) was initially defined based on built-up thresholds, keeping NDBI as a feature would lead to *Data Leakage*, causing the model to memorize the target instead of learning the complex relationship between temperature, population, and vegetation.
- **The Result:** This strategy ensures the model's predictive realism when applied to unseen future data where built-up indices might be unknown.

2. Model Development

The core of the ”Sprawl Monitoring System” is built upon an ensemble learning framework. By using a multi-stage modeling approach, the system first forecasts future environmental states and then assesses the risk levels associated with those states.

2.1 The Random Forest (RF) Algorithm

The **Random Forest** algorithm was selected as the primary engine for both regression and classification tasks. Unlike single decision trees which are prone to overfitting, Random Forest operates by constructing a multitude of decision trees during the training phase.

The strengths of using Random Forest in an environmental context include:

- **Non-linear Relationship Mapping:** Environmental data like LST and NDVI rarely follow a linear path; RF can capture complex, non-linear interactions between these features.
- **Resilience to Outliers:** Satellite data can occasionally contain anomalies; the ensemble nature of RF averages out these errors.
- **Feature Importance:** It allows us to identify which factor (e.g., Year vs. Population) has the highest impact on environmental degradation.

2.2 Dual-Stage Modeling Architecture

The system architecture is divided into two specialized branches:

2.2.1 Random Forest Regressor: Temporal Forecasting

The Regressor is designed to perform multi-output regression. It takes the *District* and the *Target Year* as inputs and maps them to multiple environmental outputs:

$$f(District, Year) \rightarrow \{LST, NDVI, Population\} \tag{2.1}$$

By training on historical data from Google Earth Engine, this model learns the unique ”environmental trajectory” of each Sri Lankan district, allowing it to project these values up to the year 2050.

2.2.2 Random Forest Classifier: Probabilistic Risk Assessment

Instead of providing a binary ”Urban” or ”Rural” label, the classifier is utilized to generate a **Urbanization Probability Score**. This is a crucial distinction for urban planning.

- **Beyond Binary Classification:** A simple 'Urban' label does not show the intensity of development. By extracting the probability of the 'Urban' class, the system provides a gradient of risk.
- **The Output Metric:** The output is rendered as a percentage (%). For example, a result of **85% Urbanization Probability** indicates a high-intensity built-up risk, whereas **15%** indicates a stable, high-vegetation rural area.
- **Logic of the Score:** This score is calculated based on the consensus of the underlying 100 decision trees. If 85 out of 100 trees classify the predicted environment as 'Urban' based on its heat and vegetation loss, the resulting risk score is 85%.

2.3 Model Training and Hyperparameters

The models were implemented using the `Scikit-Learn` library in Python. The following hyperparameter configuration was used to ensure an optimal balance between bias and variance:

- **n_estimators=100:** The number of trees in the forest.
- **random_state=42:** To ensure reproducibility of the results.
- **Train-Test Split:** The dataset was partitioned into 80% training data and 20% unseen testing data to validate the performance.

2.4 Mathematical Objective Function

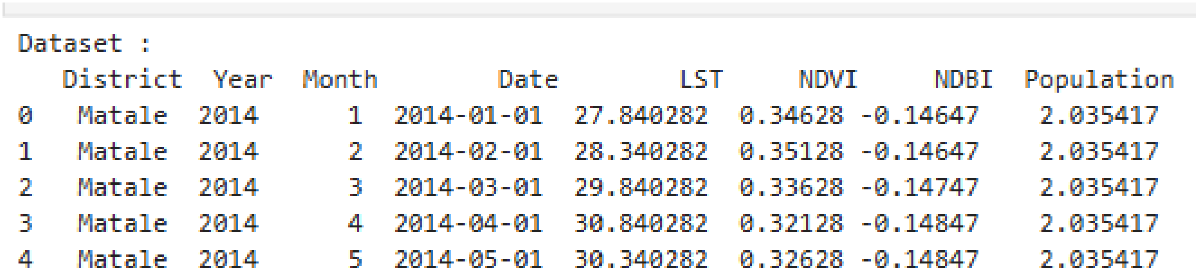
The regressor minimizes the **Mean Squared Error (MSE)**, ensuring that the predicted temperatures and vegetation indices are as close to reality as possible:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.2}$$

Where y_i is the actual satellite observation and $\hat{y}_i$ is the value predicted by the model.

3. Results and Evaluation

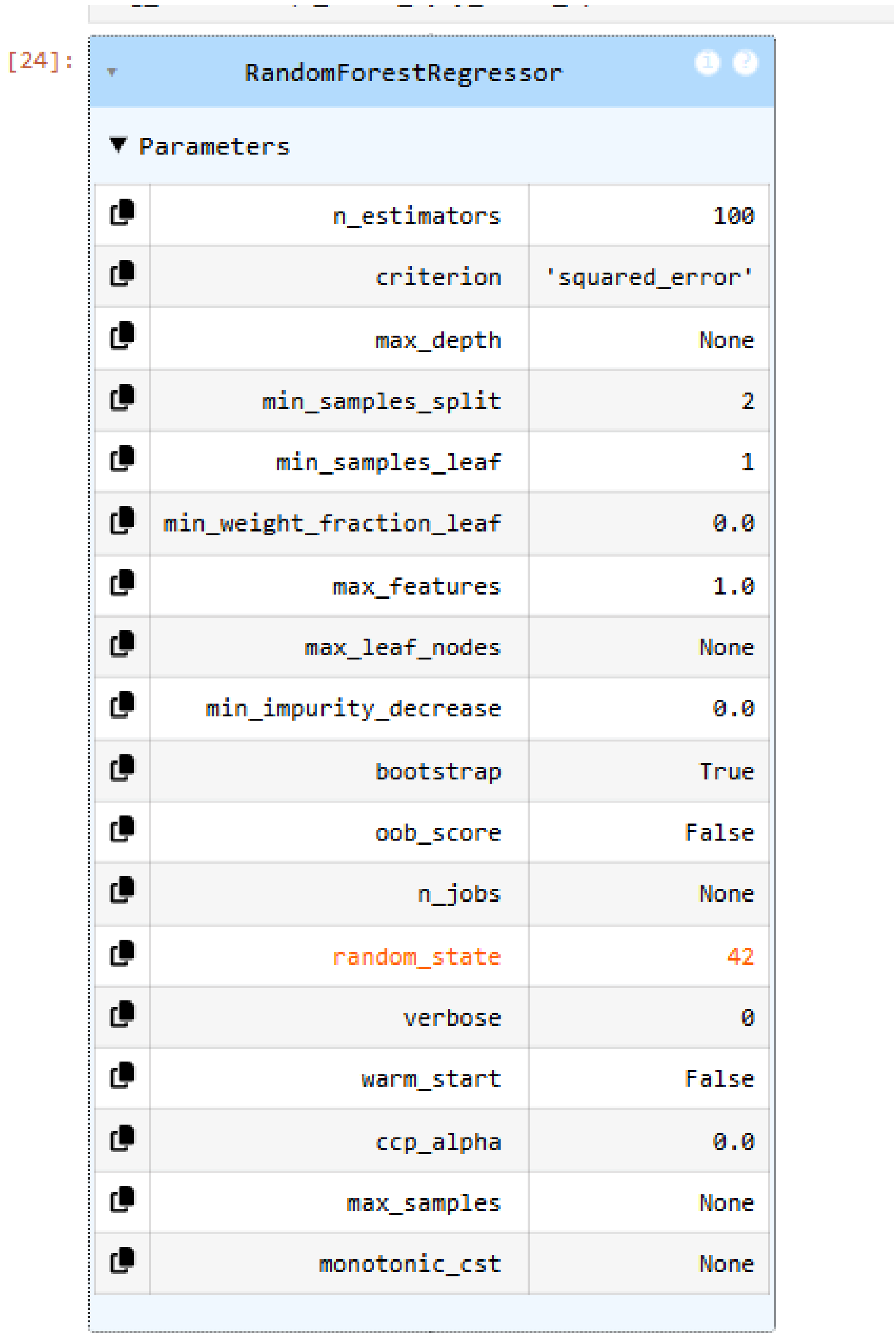
3.1 Data Overview



```
Dataset :
   District  Year  Month        Date        LST     NDVI     NDBI  Population
0    Matale  2014      1  2014-01-01  27.840282  0.34628 -0.14647    2.035417
1    Matale  2014      2  2014-02-01  28.340282  0.35128 -0.14647    2.035417
2    Matale  2014      3  2014-03-01  29.840282  0.33628 -0.14747    2.035417
3    Matale  2014      4  2014-04-01  30.840282  0.32128 -0.14847    2.035417
4    Matale  2014      5  2014-05-01  30.340282  0.32628 -0.14847    2.035417
```

Figure 3.1: Snapshot of the processed dataset.

3.2 Model Architecture Visualization



[24]:

RandomForestRegressor

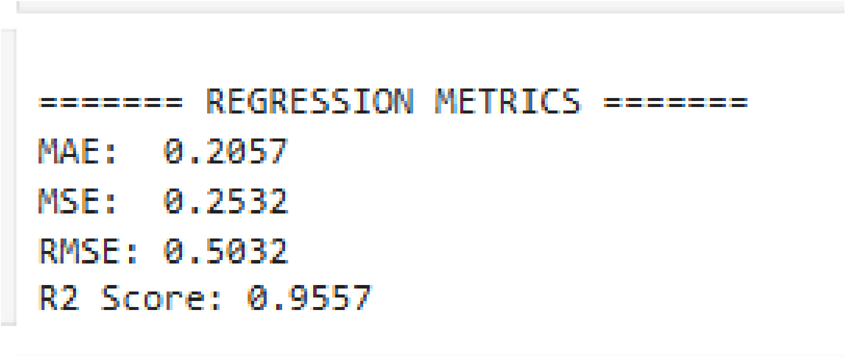
▼ Parameters

n_estimators	100
criterion	'squared_error'
max_depth	None
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_features	1.0
max_leaf_nodes	None
min_impurity_decrease	0.0
bootstrap	True
oob_score	False
n_jobs	None
random_state	42
verbose	0
warm_start	False
ccp_alpha	0.0
max_samples	None
monotonic_cst	None

Figure 3.2: Random Forest Regressor Architecture.

3.3 Quantitative Evaluation

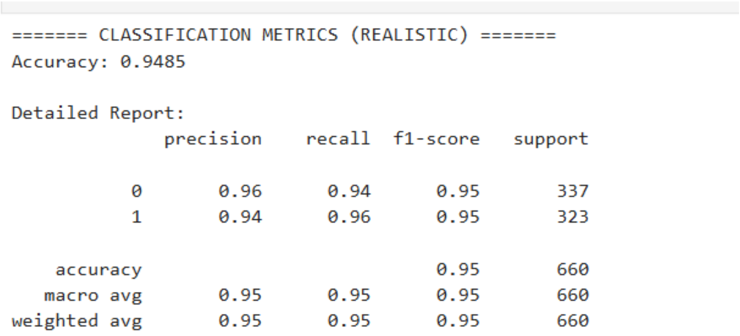
3.3.1 Regression Performance



```
======= REGRESSION METRICS =======
MAE:  0.2057
MSE:  0.2532
RMSE: 0.5032
R2 Score: 0.9557
```

Figure 3.3: Regression Evaluation Metrics.

3.3.2 Classification Performance



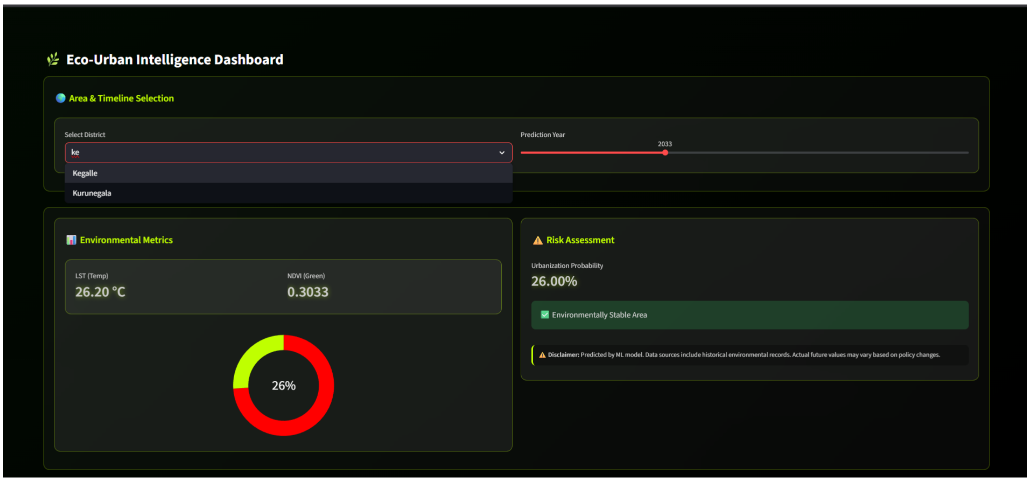
```
======= CLASSIFICATION METRICS (REALISTIC) =======
Accuracy: 0.9485

Detailed Report:
              precision    recall  f1-score   support

           0       0.96      0.94      0.95       337
           1       0.94      0.96      0.95       323

    accuracy                           0.95       660
   macro avg       0.95      0.95      0.95       660
weighted avg       0.95      0.95      0.95       660
```

Figure 3.4: Classification Performance Metrics.

4. Discussion and Dashboard Implementation

4.1 The Forecasting Dashboard

A web-based dashboard was developed using the **Streamlit** framework.



Eco-Urban Intelligence Dashboard

Area & Timeline Selection

Select District: ke

Kegalle

Kurunegala

Prediction Year: 2033

Environmental Metrics

LST (Temp): 26.20 °C

NDVI (Green): 0.3033

Risk Assessment

Urbanization Probability: 26.00%

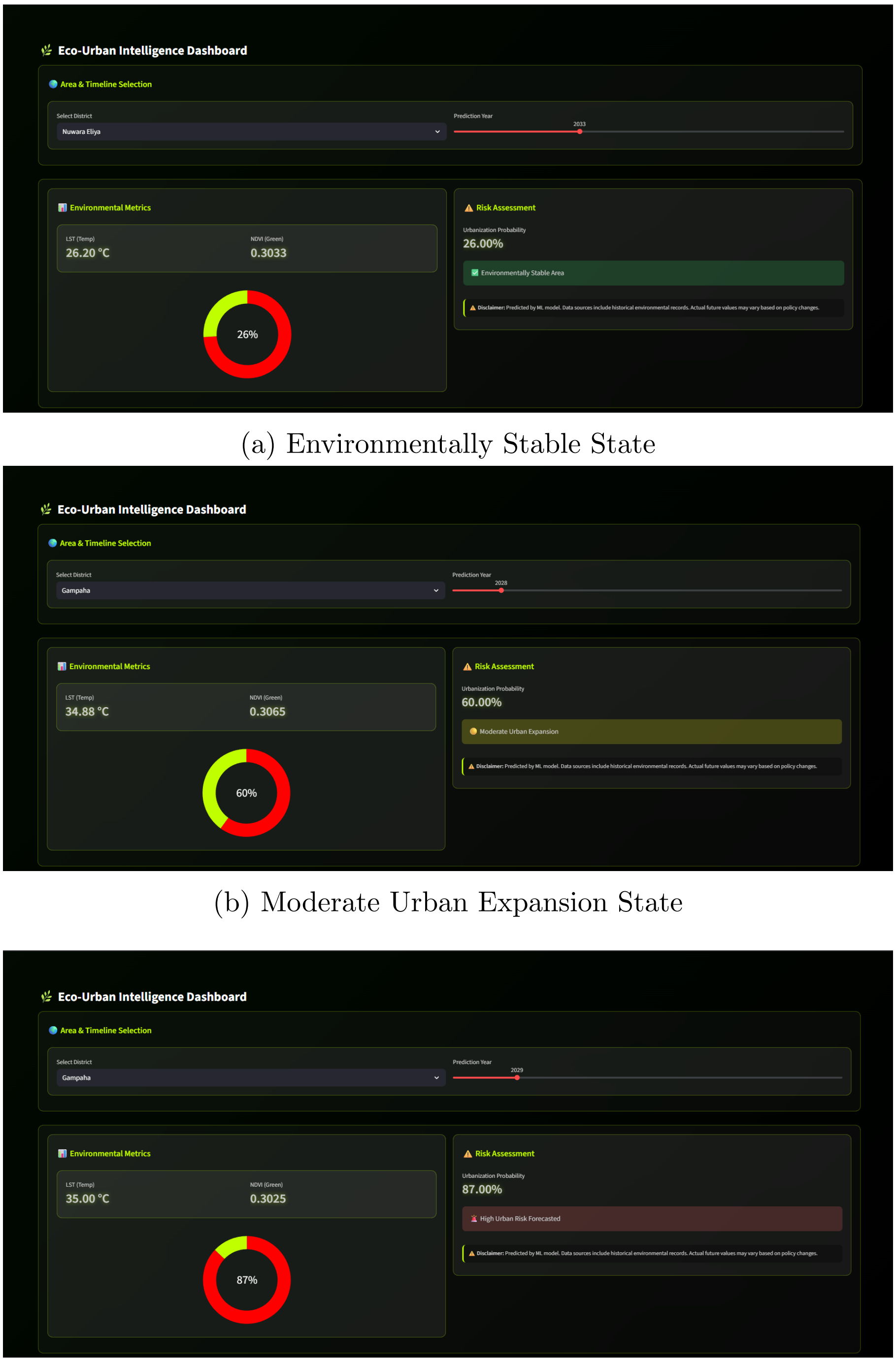
☑ Environmentally Stable Area

Disclaimer: Predicted by ML model. Data sources include historical environmental records. Actual future values may vary based on policy changes.

Figure 4.1: The Primary Eco-Urban Intelligence Dashboard Interface.

4.1.1 Risk Assessment States

The dashboard updates visual indicators for three states:



Eco-Urban Intelligence Dashboard

Area & Timeline Selection

Select District: Nuwara Eliya

Prediction Year: 2033

Environmental Metrics

LST (Temp): 26.20 °C

NDVI (Green): 0.3033

26%

Risk Assessment

Urbanization Probability: 26.00%

✓ Environmentally Stable Area

⚠ **Disclaimer:** Predicted by ML model. Data sources include historical environmental records. Actual future values may vary based on policy changes.

(a) Environmentally Stable State

Eco-Urban Intelligence Dashboard

Area & Timeline Selection

Select District: Gampaha

Prediction Year: 2028

Environmental Metrics

LST (Temp): 34.88 °C

NDVI (Green): 0.3065

60%

Risk Assessment

Urbanization Probability: 60.00%

● Moderate Urban Expansion

⚠ **Disclaimer:** Predicted by ML model. Data sources include historical environmental records. Actual future values may vary based on policy changes.

(b) Moderate Urban Expansion State

Eco-Urban Intelligence Dashboard

Area & Timeline Selection

Select District: Gampaha

Prediction Year: 2029

Environmental Metrics

LST (Temp): 35.00 °C

NDVI (Green): 0.3025

87%

Risk Assessment

Urbanization Probability: 87.00%

High Urban Risk Forecasted

⚠ **Disclaimer:** Predicted by ML model. Data sources include historical environmental records. Actual future values may vary based on policy changes.

(c) High Urban Risk Forecasted State

Figure 4.2: Comparison of Predictive Risk States.

5. Conclusion

5.1 Summary

The ”Sprawl Monitoring in Sri Lanka” project successfully demonstrated the efficacy of using ensemble learning techniques to monitor and forecast environmental degradation at a district level. By integrating multi-temporal satellite data with demographic trends, the system provides a robust framework for urban planning.

The technical evaluation of the system yielded the following conclusions:

- **High Forecasting Precision:** The Random Forest Regressor achieved an R^2 score of **0.9557**, indicating that the model can account for over 95% of the variance in environmental factors like LST and NDVI.
- **Reliable Risk Assessment:** The classification model proved highly effective in identifying urbanization risks, providing a probabilistic score that reflects the intensity of development rather than a simple categorical label.
- **Actionable Insights:** The developed Streamlit dashboard successfully bridges the gap between complex machine learning outputs and practical decision-making, allowing stakeholders to visualize future environmental scenarios up to the year 2050.

Bibliography

[1] Gorelick, N., et al. (2017). *Google Earth Engine: Planetary-scale geospatial analysis.*

[2] Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python.*

[3] Breiman, L. (2001). *Random Forests.* Machine Learning.