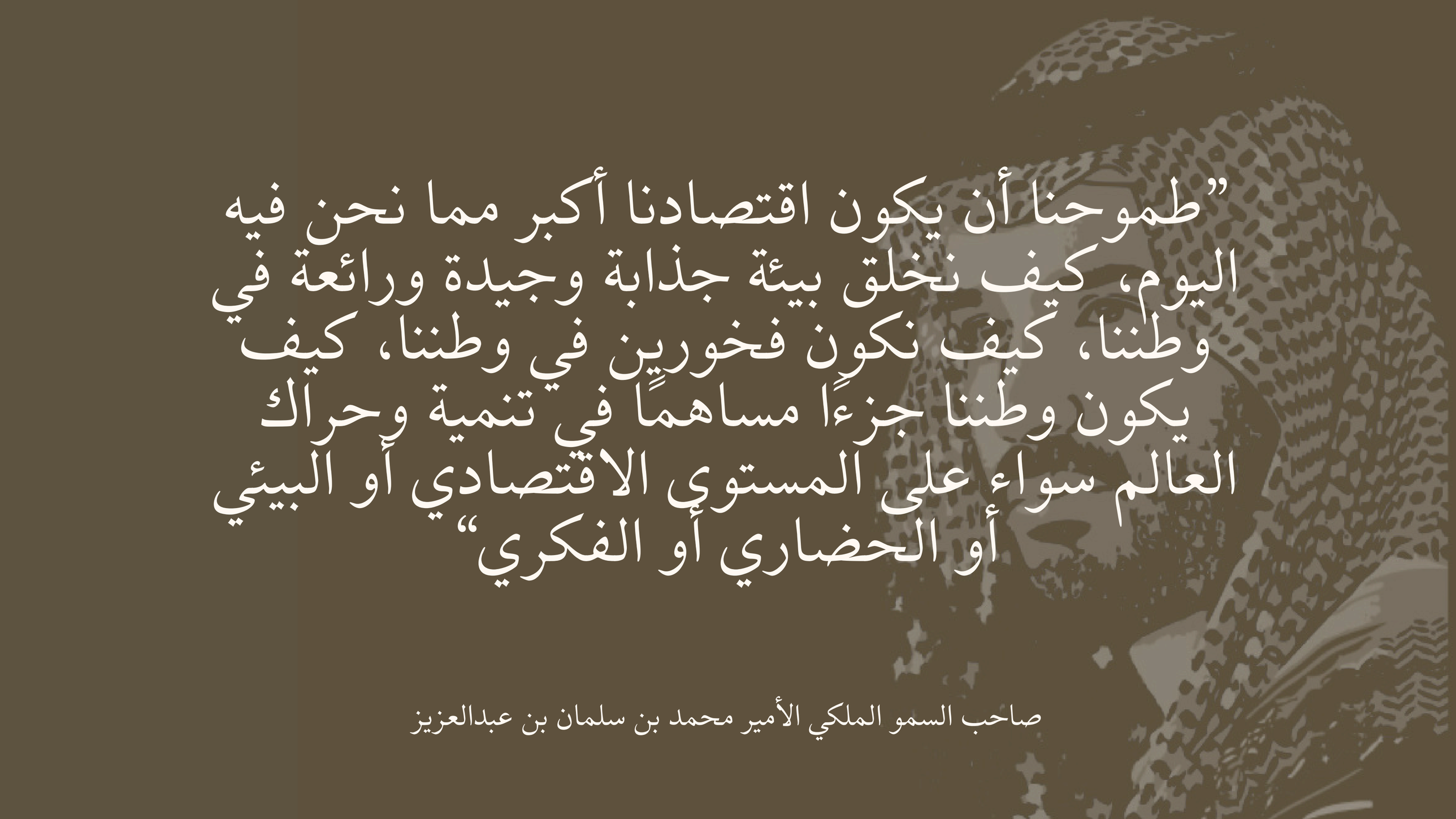
جامعة الملك سعود
King Saud University

Afaq آفاق

IT362 - Data Science

UNVEILING SAUDI ARABIA'S HIDDEN GEMS THROUGH DATA



”طموحنا أن يكون اقتصادنا أكبر مما نحن فيه اليوم، كيف نخلق بيئة جذابة وجيدة ورائعة في وطننا، كيف نكون فخورين في وطننا، كيف يكون وطننا جزءًا مساهمًا في تنمية وحراك العالم سواء على المستوى الاقتصادي أو البيئي أو الحضاري أو الفكري“

صاحب السمو الملكي الأمير محمد بن سلمان بن عبدالعزيز

Analytical Journey

From Data Collection to Model Evaluation

1. Data Collection

- Retrieved 199 public photos using the Flickr API.
- Extracted metadata including: title, views, favorites, comments, date taken/uploaded.
- Collected comment texts for sentiment analysis.

2. Feature Extraction

- OpenAI CLIP: Identified visual elements and landscape types.
- OpenCV: Determined time of day (day/night) from image brightness.
- Hugging Face Transformers: Performed sentiment analysis on photo comments.
- Created new columns: sentiment score, visual elements, region, time of day, and season.

3. Data Transformation

- Structured all raw data into a unified dataset .
- Removed irrelevant/missing fields.
- Normalized numeric values for analysis.
- Applied one-hot encoding to categorical features.

TABUK, SAUDI ARABIA

Analytical Journey

From Data Collection to Model Evaluation

4. Exploratory Data Analysis

- Analyzed engagement by region, season, and time
- Studied how landscape types and visual elements relate to interaction
- Used correlation, chi-square tests, and visualizations to explore patterns
- Cleaned data and noted biases

5. Modeling

- Defined a Popularity Score using a custom formula combining views, favorites, comments, and sentiment.
- Trained three regression models:
  - Linear Regression (baseline)
  - Random Forest Regressor
  - Support Vector Regressor (SVR)

6. Model Evaluation

- Used metrics: MAE, MSE, RMSE, R^2 to evaluate performance.
- Random Forest achieved the best accuracy ($R^2 = 0.93$).
- Compared actual vs. predicted scores to validate model reliability.

TABUK, SAUDI ARABIA

Algorithms Selected for Model Building

Linear Regression

- Chosen as a baseline model to evaluate general trends in popularity score
- Simple and interpretable, useful for understanding basic relationships between features like comments and favorites

Random Forest Regressor

- Selected for its ability to handle non-linear patterns in visual and contextual features
- Performed best in capturing complex interactions between variables such as visual elements, region, and time

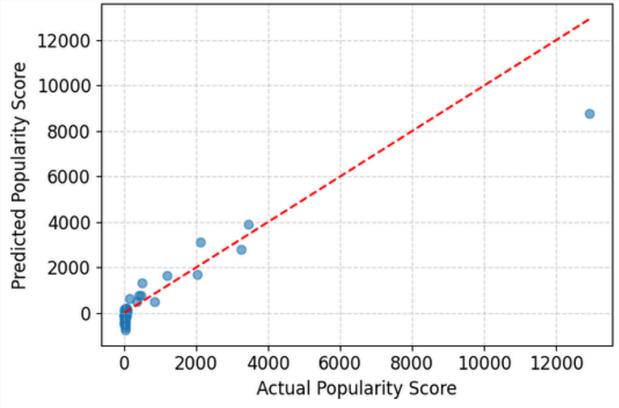
Support Vector Regressor

- Included to test margin-based modeling on our high-dimensional data
- Aimed to explore subtle correlations, especially from encoded visual features and timestamp values

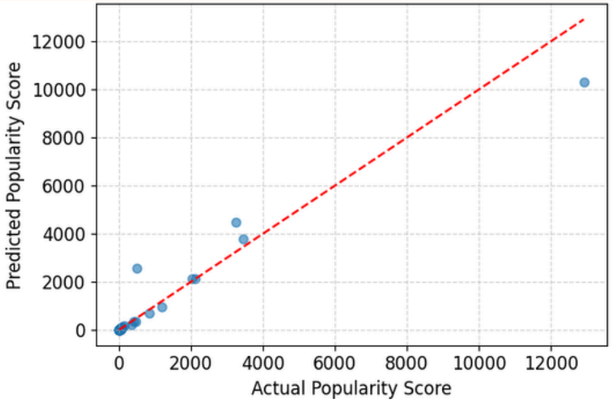
RUB' AL KHALI, SAUDI ARABIA

Algorithms Selected for Model Building

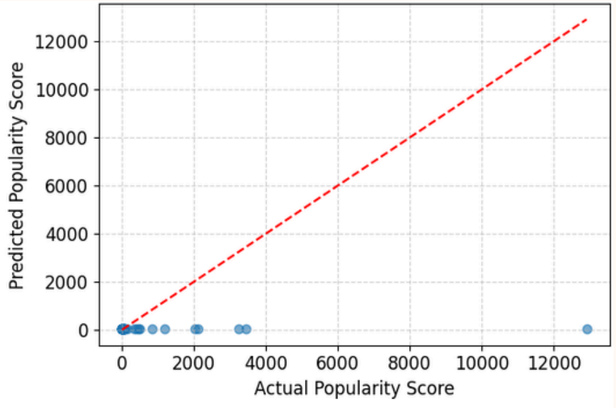
Linear Regression



Random Forest Regressor



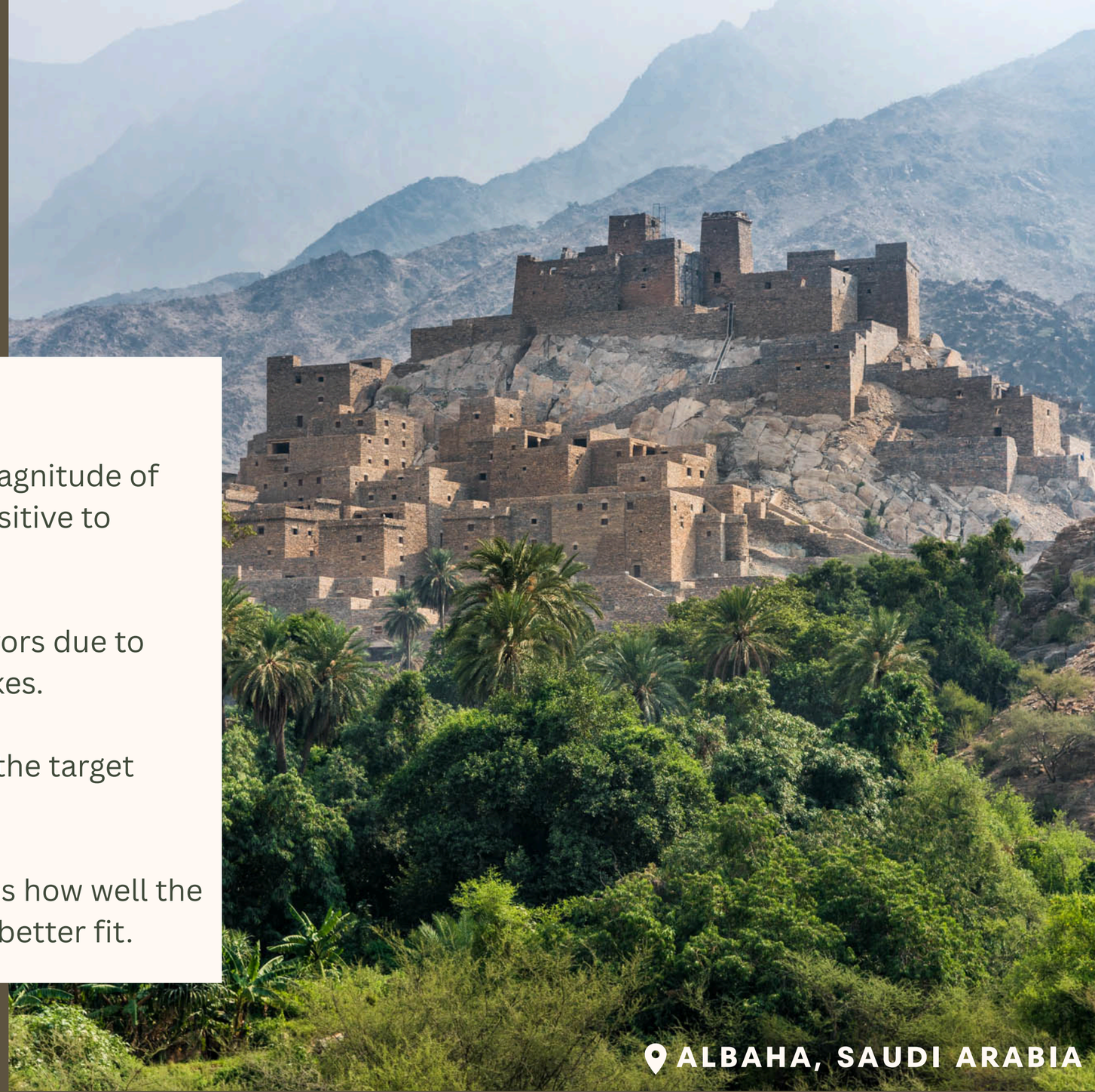
Support Vector Regressor



Model Evaluation Metrics

Why These Metrics?

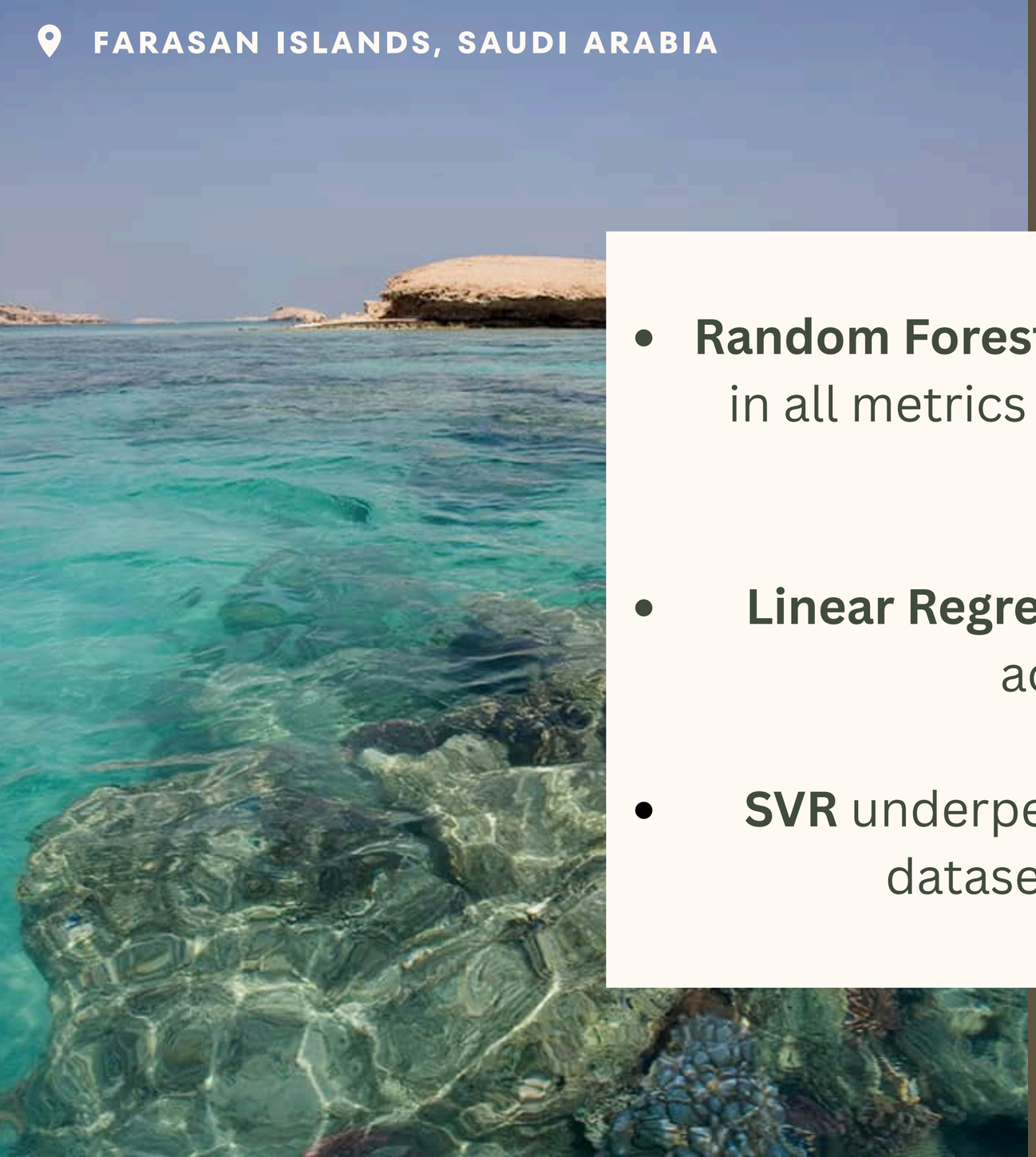
- **MAE (Mean Absolute Error)**: Measures average magnitude of errors in predictions — interpretable and less sensitive to outliers.
- **MSE (Mean Squared Error)**: Emphasizes larger errors due to squaring — useful for penalizing significant mistakes.
- **RMSE (Root Mean Squared Error)**: Same units as the target variable, easier to interpret than MSE.
- **R^2 Score (Coefficient of Determination)**: Measures how well the model explains variance — closer to 1 indicates a better fit.



Performance Comparison

Model	MAE	MSE	RMSE	R^2
Linear Regression	406.92	571,949	756.27	0.877
Random Forest	185.06	331,761	575.99	0.928
Support Vector Regressor	715.15	5,070,887	2,251.91	-0.094

ALQASSIM, SAUDI ARABIA



What These Results Reveal

- **Random Forest** significantly outperforms the others in all metrics — lowest error and highest R^2 → best choice.
- **Linear Regression** performs decently but not as accurate — good baseline.
- **SVR** underperforms — likely not suitable for this dataset (high error and negative R^2).

technical challenges

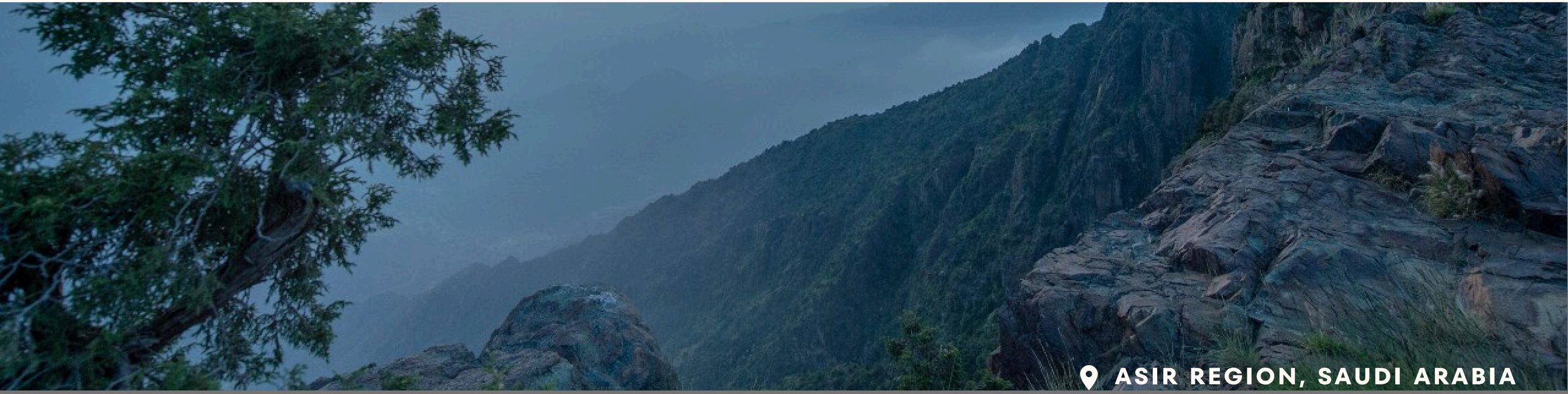
Limited metadata availability and quality from the Flickr API caused incomplete or inaccurate location information.
→ Solution: We manually categorized the photos into regions to ensure geographic accuracy.

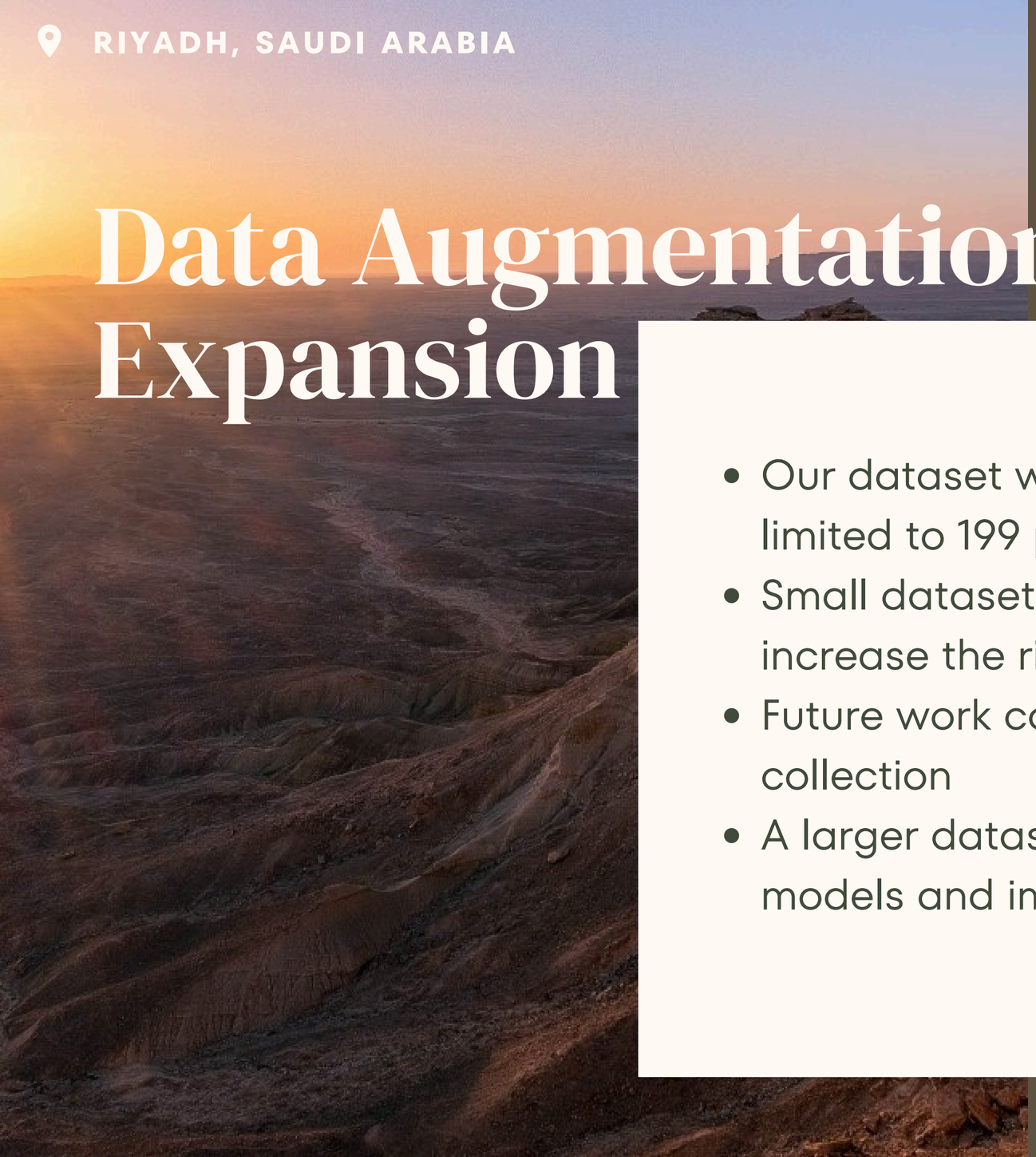
Processing high-resolution images using CLIP and OpenCV required significant computational resources and time.
→ Solution: We optimized the processing workflow and started image analysis early.

One photo was missing during processing, reducing the dataset from 200 to 199 images.
→ Solution: We continued the project without reprocessing the entire dataset.

How the models could be improved?

- **Data Augmentation or Expansion**
- **Time-Based Modelling**
- **More Advanced Models**





Data Augmentation or Expansion

- Our dataset was manually selected and limited to 199 photos
- Small datasets can limit model accuracy and increase the risk of overfitting
- Future work could use automated image collection
- A larger dataset allows for training deeper models and improves generalization

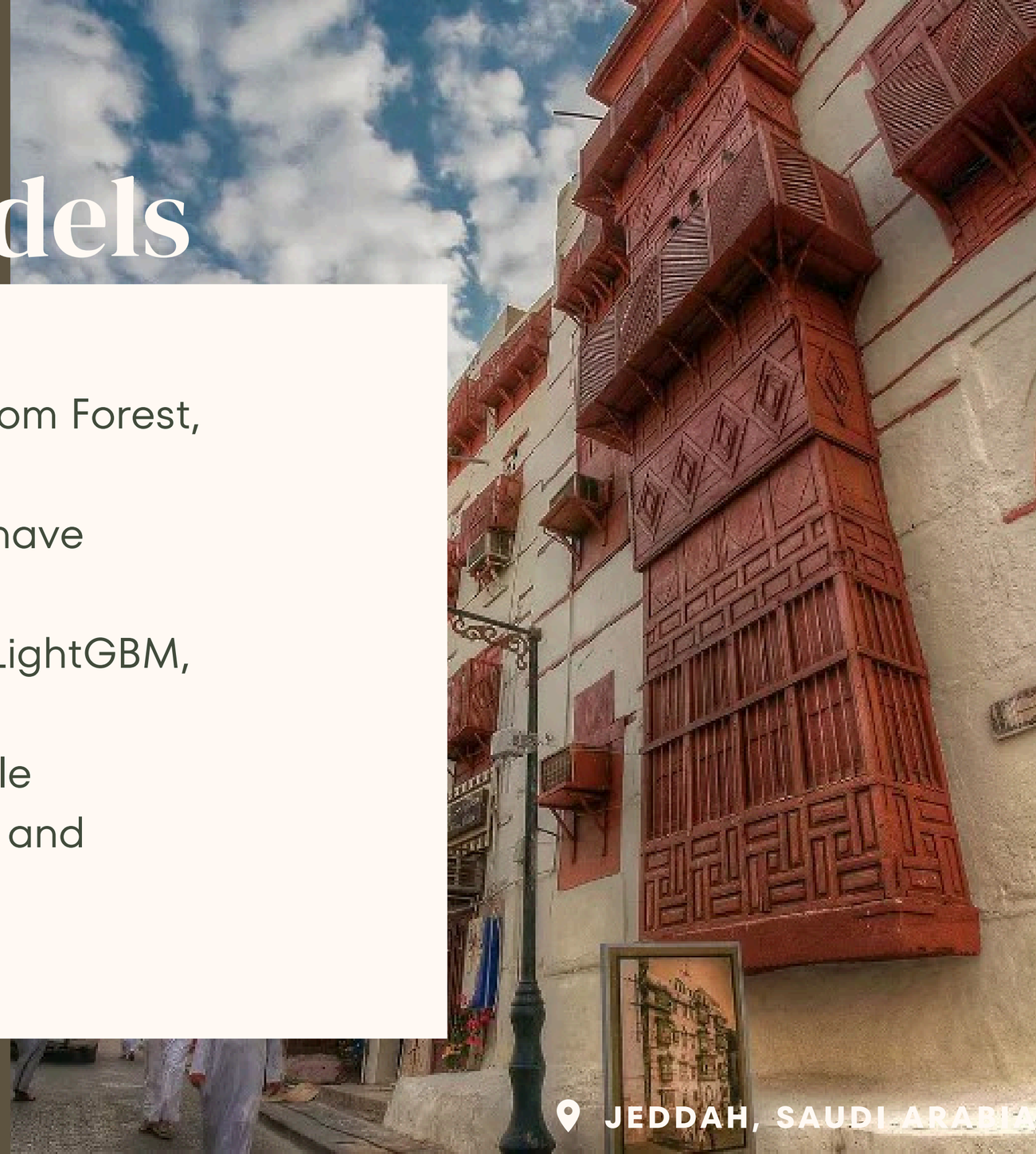
Time-Based Modeling

- We included timestamps but didn’t use them in temporal models
- Models like LSTM or ARIMA could track engagement trends over time
- This would help make seasonal or time-aware recommendations
- Useful for understanding how popularity changes by season or year

AL ULA, SAUDI ARABIA

More Advanced Models

- We tested Linear Regression, Random Forest, and SVR
- These models performed well, but have limitations with complex data
- Future work can explore XGBoost, LightGBM, or deep learning regressors
- Advanced models can better handle structured + image-based features and improve accuracy



STUDENTS

Lamees Alghamdi
444201177

Maha Alruwais
444200749

Ghadeer Alnuwaysir
444200420

Rana Albridi
444201094

Norah Almadhi
444200890