OpenSW Architecture

**User**

**텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.**

1. Data preparation and preprocessing

\* Key preprocessing content

- Remove unused columns

- Calculate the center coordinate reference distance

- Date → Convert to normalized numerical features

- Categorical variables → One-hot encoding

- Processing missing values (reviews\_per\_month = 0)

- Location information is removed after clustering (latitude, longitude drop)

2. clustering

\* StandardScaler

\* KMeans clustering

- Determining the optimal number of clusters: Elbow Method (SSE) + Silhouette Score

- Finally set k=4

- Store cluster results in df\_scaled ['cluster']

3. Cluster prediction modeling

\* Configure training data using cluster results as labels

\* RandomForestClassifier

- Hyperparameter tuning → overfitting prevention

- train\_test\_split

- Validating performance with cross\_val\_score → Stratified K-Fold (default: cv=5)

- Model accuracy: OOB Score ≈ CV Score ≈ Test Accuracy → Not overfitting

4. Visualizaion

- PCA (cluster visualization)

- confusion matrix

- feature importance

- OOB vs Test Accuracy

5. Predicting with new data, visualizing maps

\* Input : a new csv file

\* Same preprocessing, cluster prediction

\* Visualize a map based on cluster prediction results

**Host**

**텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.**

1. Data preparation and preprocessing

\* Key preprocessing content

- Eliminating outliers

- Remove price below zero, 2000 or more

- Limited to a minimum range of 1 to 30 nights

- Log conversion: price → log\_price

2. Clustering

\* Performing KMeans clustering

\* Search for optimal k value with Elbow Method → k=4

3. Linear Regression

\* target : log\_price

\* Using Scaled Data

\* Performance Evaluation: RMSE, MAE, R² (based on log and actual price)

\* Visualization

- Prediction vs. Actual

- Residual plot

4. XGBoost Regression

\* Randomized Search CV + 3-Fold CV → parameter optimization

\* performing best\_estimator\_based prediction

\* Performance Evaluation (RMSE, MAE, R²): Check both log and actual price basis

\* Visualization of Feature Importance: Check the ranking of the most important features

\* Visualization

- Prediction vs. Actual

- Residual plot