Airbnb Price Prediction Report

1. Introduction

This project aims to predict Airbnb listing prices using machine learning. The dataset contains information such as amenities, room type, distance to the city center, number of bathrooms, and bedrooms.
The goal is to build a model that can estimate the price of a listing based on these features.

2. Objective

- To clean and preprocess the Airbnb dataset
- To build different machine learning models and compare their performance.
- To find the best model and tune it for better accuracy.
- To understand which features affect the price using SHAP analysis

3. Steps Taken

3.1 Data Preprocessing

- Loaded the dataset
- Cleaned missing values
- Extracted important features such as:
    - Amenities
    - Number of bathrooms
    - Number of bedrooms
    - Room type
    - Distance from city center
- Converted categorical data to numeric format

3.2 Model Building

Four models were trained:

1. Random Forest
2. XGBoost
3. Gradient Boosting
4. Extra Trees

Each model was evaluated using RMSE, MSE and R^2

3.3 Model Results

Model	RMSE	MAE	R^2
RandomForest	1.739943+16	8.958992e+15	-0.126549
XGBoost	1.742538e+16	8.489380e+15	-0.129912
GradientBoosting	1.745533e+16	8.196741e+15	-0.133799
ExtraTrees	1.837953e+16	1.015334e+16	-0.257039

3.4 Hyperparameter Tuning

RandomForest was tuned using RandomizedSearchCV

Best Parameters found:

- n_estimators: 400
- max_depth: 20
- min_samples_leaf: 1

Tuned RMSE:

1.7395742507536234E+16

3.5 SHAP Explanation

A SHAP summary plot was used to understand feature impact.

Simple interpretation:

- Amenities like AC, heating, dryer, washer, kitchen increase price.
- Features like room type = shared room reduce price
- Features with higher SHAP values have more influence on predictions

4. Conclusion

- RandomForest performed the best among all tested models.
- The dataset likely has very large or log transformed price values, leading to extremely high predictions
- SHAP analysis helped identify important features influencing price.
- The cleaned dataset and model can now be used for dashboards or further development.