Machine Learning Model Documentation

# 1. Overview

This project involves building and evaluating multiple regression models to predict rental prices based on various features. Several machine learning algorithms were implemented, including Linear Regression, Ridge, Lasso, ElasticNet, KNN, Random Forest, and Decision Tree Regression. The models were evaluated using R² and Mean Squared Error (MSE) metrics with K-Fold Cross Validation.

# 2. Dataset Details

The dataset consists of features representing rental property characteristics and the target variable representing rent price. Data preprocessing steps such as scaling, encoding categorical variables, and feature selection were performed before model training.

# 3. Model Development

The following regression models were trained and compared using a consistent evaluation strategy:

• Linear Regression  
• Lasso Regression  
• Ridge Regression  
• ElasticNet Regression  
• KNN Regression  
• Random Forest Regression  
• Decision Tree Regression

# 4. Cross-Validation Strategy

Each model was evaluated using 5-Fold Cross Validation (KFold) with data shuffled randomly (random\_state=2025). The evaluation metrics computed for each fold include R² (coefficient of determination) and MSE (Mean Squared Error). The mean and standard deviation across folds were used to assess model stability and generalization.

# 5. Model Comparison

Linear Regression - R²: 0.414, MSE: 765815  
Lasso Regression - R²: 0.481, MSE: 693229  
Ridge Regression - R²: 0.457, MSE: 717482  
ElasticNet Regression - R²: 0.485, MSE: 686536  
KNN Regression - R²: 0.348, MSE: 802094  
Random Forest Regression - R²: 0.606, MSE: 541277  
Decision Tree Regression - R²: 0.454, MSE: 696680

# 6. Model Tuning

The Random Forest model achieved the highest R² score (0.606) and was further tuned by adjusting the 'n\_estimators' parameter, which defines the number of trees in the forest. The tuning results are shown below:

n\_estimators | R² Mean | R² Std | MSE Mean  
-------------|----------|----------|-----------  
50 | 0.606 | 0.175 | 541277  
100 | 0.612 | 0.175 | 534691  
200 | 0.615 | 0.176 | 531715  
300 | 0.615 | 0.176 | 531436  
400 | 0.615 | 0.177 | 531565  
500 | 0.616 | 0.177 | 530465

# 7. Key Parameter Explanation

The parameter 'n\_jobs=-1' was used to parallelize the training process by utilizing all available CPU cores. This helps speed up model training, especially when using ensemble methods like Random Forest.

# 8. Conclusion

Among all the models tested, the Random Forest Regressor performed best with an R² of approximately 0.616 and the lowest MSE. Further optimization and feature engineering may help improve accuracy beyond 0.65 R².