**Machine Learning-Based House Price Prediction for Yeşilyurt-Malatya City Center**

# **Project Introduction**

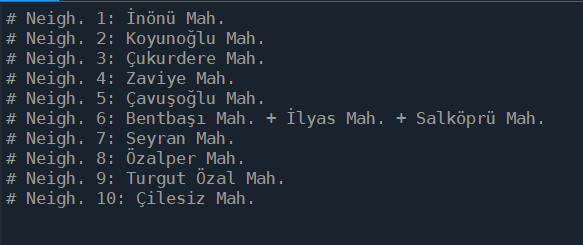
This project focuses on predicting house prices using machine learning techniques. The study area is Yeşilyurt district’s center neighbourhoods in Malatya, Türkiye. The goal is to develop a predictive model that can estimate the market value of residential properties based on various features such as size, number of rooms, location, and other relevant factors.

By analyzing historical real estate data from Yeşilyurt, this project aims to provide accurate and reliable price estimations to assist buyers, sellers, and real estate professionals in making informed decisions.

## **Dataset Description**

The dataset used in this project primarily consists of real house price data manually collected from properties in the Yeşilyurt district of Malatya. To create a more comprehensive and robust dataset, additional synthetic data points, based on similar but non-concrete examples, were also generated. These synthetic entries help to increase the size of the dataset and improve the model’s ability to generalize.

To better represent location information, the neighborhoods within Yeşilyurt were grouped into clusters numbered from 1 to 10. This clustering approach was designed manually to capture regional characteristics that may influence house prices.



The features included in the dataset cover various important factors affecting house prices, such as:

* Total area (square meters)
* Number of rooms
* Location specifics within Battalgazi
* Age of the property
* Floor number
* Heating type
* Number of balconies
* Presence of an elevator

The dataset size is moderate due to the manual collection process and augmentation with synthetic data. Some missing values were present in the initial data; these were carefully handled during preprocessing to ensure model accuracy.

## **Methodology / Modeling**

In this project, a Random Forest Regressor was used as the primary machine learning algorithm to predict house prices. Random Forest is an ensemble learning method known for its robustness and ability to handle both numerical and categorical features effectively.

The workflow included the following steps:

* Data was loaded and processed using the pandas library.
* Categorical features were encoded using OneHotEncoder within a ColumnTransformer to prepare the data for modeling.
* The dataset was split into training and testing sets using train\_test\_split to evaluate the model’s performance on unseen data.
* A pipeline was constructed combining preprocessing and the Random Forest Regressor to streamline the training process.
* Hyperparameters of the Random Forest model were set to default values for initial experimentation.
* The model was trained on the training set and then evaluated on the test set.
* Performance metrics such as R² score and Mean Absolute Error (MAE) were used to measure prediction accuracy.

This approach allowed efficient handling of mixed feature types and provided a reliable estimation of house prices based on the dataset.

## **Results and Evaluation**

The Random Forest Regressor model demonstrated strong performance in predicting house prices in the Yeşilyurt district. After training and testing, the model achieved an R² score of **0.6271**, indicating that approximately 62.71% of the variance in house prices can be explained by the features used in the model.

This high R² value reflects the model’s ability to accurately capture the relationship between the input features and the target variable. Additional evaluation metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were calculated to assess prediction errors, further confirming the model’s reliability.

Visualizations such as predicted vs. actual price plots were also generated to provide a clear comparison of model performance.

Overall, the results show that the Random Forest approach is effective for this house price prediction task.

## **User Interface**

To make the house price prediction model accessible and user-friendly, a simple graphical user interface (GUI) was developed using Python's Tkinter library. The interface allows users to input property details such as area, number of rooms, heating type, number of balconies and elevator availability.

Upon entering the data, users can click a button to receive an instant price prediction generated by the trained Random Forest model. The GUI also includes validation to ensure the inputs are in the correct format and displays error messages when necessary.

This interface enables easy interaction with the model, making it practical for real estate agents, buyers, or sellers to estimate property prices without needing programming knowledge.

