One of the applications of machine learning (ML) is predicting house prices. This process involves a series of steps that include data understanding, preprocessing, exploratory data analysis (EDA), model development and evaluation.

Workflow of the Project

In this section we will understand each step, making sure everything is crystal clear. Here are the steps that will be followed in the project:

1. Understanding the Data

The first step is to understand the dataset. This means looking at where the data comes from, what it's about and why it's useful for predicting something like house prices. Knowing what each feature (variable) represents helps in the next steps of preparing and using the data.

2. Data Preprocessing Pipeline

After understanding the data, the next step is to clean and prepare it for the model. This includes filling in missing values, dealing with data that doesn't fit the pattern (outliers) and converting non-numeric data (like categories) into a usable format. A good pipeline ensures everything is done the same way every time.

3. Exploratory Data Analysis (EDA)

4. Model Training

In this step, we teach the model to make predictions using the cleaned data. We divide the data into training and testing sets, train the model on the training set and see how well it learns to predict outcomes based on the input data.

5. Model Prediction

After training the model, we test it by making predictions with new, unseen data. The model uses what it learned to predict future outcomes, like predicting house prices based on certain features. This is where we see how well the model can generalize to new data.

6. Model Evaluation

After the model makes predictions, we check how accurate or useful the predictions are. We use metrics like accuracy or error rates to measure its performance. If the model doesn't do well, we may need to adjust it or try a different approach.

7. Saving the Model

Once the model performs well, we save it so we can use it later without retraining. Saving the model means we can load it in the future to make predictions with new data without going through the training process again. This is important for using the model in real-world situations.