House Price Prediction using AI: Basic Loading and Preprocessing

Introduction

The goal of this project is to develop a machine learning model that can predict house prices based on various features such as square footage, number of bedrooms, location, etc. In this document, we'll focus on the initial steps of loading the dataset and preprocessing it for training.

Tools and Libraries Used

- Python (programming language)

- Pandas (data manipulation library)

- NumPy (numerical computation library)

- Scikit-learn (machine learning library)

- Jupyter Notebook (for code execution)

Step 1: Dataset Loading

1.1. Data Source

The dataset used for this project is sourced from [provide the dataset source]. It contains information about various houses including features like square footage, number of bedrooms, location, etc., along with their corresponding prices.

1.2. Loading the Dataset

python

import pandas as pd

Load the dataset

df = pd.read\_csv('house\_data.csv')

```

Step 2: Exploratory Data Analysis (EDA)

EDA involves exploring the dataset to understand its structure, features, and relationships.

2.1. Data Overview

python

Display first few rows of the dataset

df.head()

```

2.2. Data Summary

python

Get summary statistics

df.describe()

```

2.3. Data Visualization

Use plots and charts to visualize relationships between features and the target variable (house prices).

Step 3: Data Preprocessing

3.1. Handling Missing Values

```python

# Check for missing values

df.isnull().sum()

If there are missing values, decide whether to impute or drop them

Example: df = df.dropna() # Drop rows with missing values

```

3.2. Feature Selection

Decide which features are relevant for the prediction task and drop any unnecessary columns.

3.3. Encoding Categorical Variables

If the dataset contains categorical variables, they need to be encoded for model training.

```python

# Example for one-hot encoding

df = pd.get\_dummies(df, columns=['location'], drop\_first=True)

```

3.4. Splitting Data

```python

from sklearn.model\_selection import train\_test\_split

# Split the dataset into training and testing sets

X = df.drop('price', axis=1) # Features

y = df['price'] # Target variable

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

```

Conclusion

At this stage, we have successfully loaded the dataset, performed basic exploratory data analysis, and preprocessed the data for training a machine learning model. The next steps would involve selecting a model, training, evaluating, and fine-tuning it for optimal performance in house price prediction.

Please note that this is a basic guide, and depending on the specific characteristics of your dataset and project objectives, further preprocessing steps or techniques may be required.