Project Idea: House Price Prediction with External Features

Base Dataset

- Use the **Boston Housing dataset** (or a Kaggle housing dataset).
  Columns: rooms, crime rate, property tax, etc.

Extra Data via Web Scraping

- Scrape **neighborhood info** online (example: average income, population density, crime index, school ratings).
- Merge it with base dataset (by location or zip code).

🔹 **Workflow (End-to-End)**

1. Problem Definition

- Goal: Predict house prices based on features.
- Business use: Real estate agencies, buyers, sellers.

---

2. Data Collection

- Base dataset: Kaggle / sklearn's housing dataset.
- Extra dataset: Web scrape from sources like:
  - Zillow / RealEstate websites
  - Government data portals (population, crime rates)

📌 Example (Python scraping with BeautifulSoup):

```
import requests
from bs4 import BeautifulSoup

url = "https://www.numbeo.com/crime/country_result.jsp?country=United+States"
response = requests.get(url)
soup = BeautifulSoup(response.text, 'html.parser')

# Example: extract city crime rates
table = soup.find("table", {"class": "table_indices"})
rows = table.find_all("tr")
for row in rows[:5]:
```

```
cols = [c.text.strip() for c in row.find_all("td")]
print(cols)
```

---

3. Data Understanding

- Check base dataset shape, missing values, distributions.
- Merge with scraped dataset (joining on city/zipcode).

---

4. Data Preprocessing

- Handle missing values (median/mean imputation, external lookup).
- Handle outliers (Z-score, IQR method).
- Encode categorical data (city names → one-hot encoding).
- Scale numerical data (StandardScaler/MinMaxScaler).

---

5. Feature Engineering

- Create new features:
    - "Rooms per household"
    - "CrimeRate x Income" interaction
    - "Proximity to school score"

---

6. Splitting Data

- Train/validation/test (70/15/15).

---

7. Model Training

- Train multiple models: Linear Regression, Random Forest, XGBoost.

---

8. Model Evaluation

- Metrics: RMSE, MAE, R^2.
- Compare models with cross-validation.

---

9. Hyperparameter Tuning

- GridSearchCV / RandomizedSearchCV.

---

10. Monitoring

- Add logging for predictions.
- Update model as new property data arrives.

🔹 **Why this Project is Special**

✓ Uses **real dataset**.
✓ Includes **data scraping** to enrich dataset (very rare in student projects).
✓ Full **data preprocessing pipeline**.
✓ Proper **ML workflow** (train/test, tuning, evaluation).