1. 연구목적

본 연구는 주택 가격 데이터(Housing Price Data)를 활용하여 머신러닝 기반 가격 예측 모델을 구축하고, 서로 다른 회귀 모델 간 예측 성능을 비교·분석하는 것을 목적으로 함. 이를 통해 주택 가격에 영향을 미치는 주요 요인을 확인하고, 머신러닝 모델의 실용 가능성을 검증하고자 함

```
df = pd.read_csv('/Housing_Price_Data.csv')

df.head()
```

	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
0	13300000	7420	4	2	3	yes	no	no	no	yes	2	yes	furnished
1	12250000	8960	4	4	4	yes	no	no	no	yes	3	no	furnished
2	12250000	9960	3	2	2	yes	no	yes	no	no	2	yes	semi-furnished
3	12215000	7500	4	2	2	yes	no	yes	no	yes	3	yes	furnished
4	11410000	7420	4	1	2	yes	yes	yes	no	yes	2	no	furnished

2. 데이터 개요

본 연구에 사용된 데이터는 총 545개의 주택 샘플과 13개의 변수로 구성됨

타겟 변수는 price(주택 가격)이며, 입력 변수로는 area, bedrooms, bathrooms, stories, parking과 함께 mainroad, guestroom, basement, hotwaterheating, airconditioning, prefarea, furnishingstatus 등의 범주형 변수가 포함됨

기초 통계 분석 결과, 주택 가격은 최소 약 175만, 최대 약 1,330만 수준의 분포를 보이며, 주택 면적(area)과 주차 공간(parking) 등의 변수는 비교적 넓은 분산을 가지는 것으로 나타남

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 545 entries, 0 to 544
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   price             545 non-null    int64
 1   area              545 non-null    int64
 2   bedrooms          545 non-null    int64
 3   bathrooms         545 non-null    int64
 4   stories           545 non-null    int64
 5   mainroad          545 non-null    object
 6   guestroom         545 non-null    object
 7   basement          545 non-null    object
 8   hotwaterheating   545 non-null    object
 9   airconditioning   545 non-null    object
 10  parking           545 non-null    int64
 11  prefarea          545 non-null    object
 12  furnishingstatus  545 non-null    object
dtypes: int64(6), object(7)
memory usage: 55.5+ KB
```

3. 데이터 전처리 및 변수 설정

결측치가 존재하는 행은 제거(dropna) 처리함

타겟 변수(y)는 price로 설정하였으며, 입력 변수(X)는 price를 제외한 모든 변수로 구성함

범주형 변수(mainroad, guestroom, basement, hotwaterheating, airconditioning, prefarea, furnishingstatus)는 원-핫 인코딩(pd.get_dummies, drop_first=True)을 적용하여 수치형 변수로 변환

최종적으로 X는 13개 변수로 구성되었으며, y는 545개의 가격 데이터로 구성함. 학습 데이터와 검증 데이터는 8:2 비율로 분리하였고, 재현성을 확보하기 위해 random_state=42로 고정

4. 머신러닝 모델 구성

본 연구에서는 두 가지 회귀 모델을 활용함.

첫째, 선형회귀(Linear Regression) 모델을 기준 모델(Baseline)로 설정함. 이는 변수와 가격 간의 선형적 관계를 가정하는 가장 기본적인 회귀 모델임.

둘째, 랜덤포레스트 회귀(Random Forest Regressor) 모델을 적용함. 이는 다수의 의사결정나무를 결합한 앙상블 모델로, 비선형 관계를 효과적으로 반영할 수 있는 특징을 가짐.

5. 학습 결과 및 성능 비교

선형회귀 모델의 성능은 다음과 같음.

R^2: 약 0.653

RMSE: 약 1,324,508

랜덤포레스트 회귀 모델의 성능은 다음과 같음.

R^2: 약 0.619

RMSE: 약 1,391,619

두 모델 모두 주택 가격의 전반적인 경향은 일정 수준 이상 설명하는 것으로 나타났으며, 선형회귀 모델이 본 데이터셋에서는 랜덤포레스트보다 다소 높은 설명력을 보임.

#7-1 모델 1: 선형회귀 학습 및 평가

```
[26]
✓ 0초
lr = LinearRegression()
lr.fit(X_train, y_train)

# 예측
y_pred_lr = lr.predict(X_test)

# 평가 지표 계산
r2_lr = r2_score(y_test, y_pred_lr)
rmse_lr = np.sqrt(mean_squared_error(y_test, y_pred_lr))

print("=== Linear Regression 성능 ===")
print("R²   :", r2_lr)
print("RMSE:", rmse_lr)
```

```
=== Linear Regression 성능 ===
R²   : 0.6529242642153184
RMSE: 1324506.9600914386
```

#7-2 모델 2: 랜덤포레스트 회귀 학습 및 평가

```
[27]
✓ 0초
rf = RandomForestRegressor(
    n_estimators=300,   # 트리 개수
    max_depth=None,     # 트리 깊이 제한 없음(과제용 기본)
    random_state=42
)

rf.fit(X_train, y_train)

# 예측
y_pred_rf = rf.predict(X_test)

# 평가 지표 계산
r2_rf = r2_score(y_test, y_pred_rf)
rmse_rf = np.sqrt(mean_squared_error(y_test, y_pred_rf))

print("=== Random Forest 성능 ===")
print("R²   :", r2_rf)
print("RMSE:", rmse_rf)
```

```
=== Random Forest 성능 ===
R²   : 0.6168607933155569
RMSE: 1391619.2131186817
```

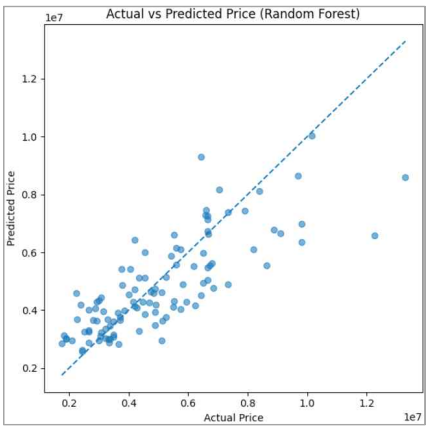
6. 시각화 결과 해석

6.1 실제값 vs 예측값 산점도

랜덤포레스트 회귀 모델의 실제 주택 가격과 예측 가격을 비교한 결과, 데이터 포인트가 대각선 주변에 비교적 밀집되어 분포함. 이는 모델이 주택 가격의 전반적인 추세를 일정 수준 이상 학습하였음을 의미함. 다만 일부 고가 주택 영역에서는 예측 오차가 확대되는 경향도 함께 관찰됨.

#9-1 시각화 1: 실제값 vs 예측값

```
plt.figure(figsize=(6, 6))
plt.scatter(y_test, y_pred_rf, alpha=0.6)
plt.xlabel("Actual Price")
plt.ylabel("Predicted Price")
plt.title("Actual vs Predicted Price (Random Forest)")
# y = x 기준선 추가
min_val = min(y_test.min(), y_pred_rf.min())
max_val = max(y_test.max(), y_pred_rf.max())
plt.plot([min_val, max_val], [min_val, max_val], linestyle='--')
plt.tight_layout()
plt.show()
```



6.2 Feature Importance 분석

변수 중요도 분석 결과, area(주거 면적)가 가장 높은 중요도를 보였으며, 그 다음으로 bathrooms, airconditioning, parking, stories 순으로 가격 예측에 중요한 영향을 미치는 것으로 나타남.

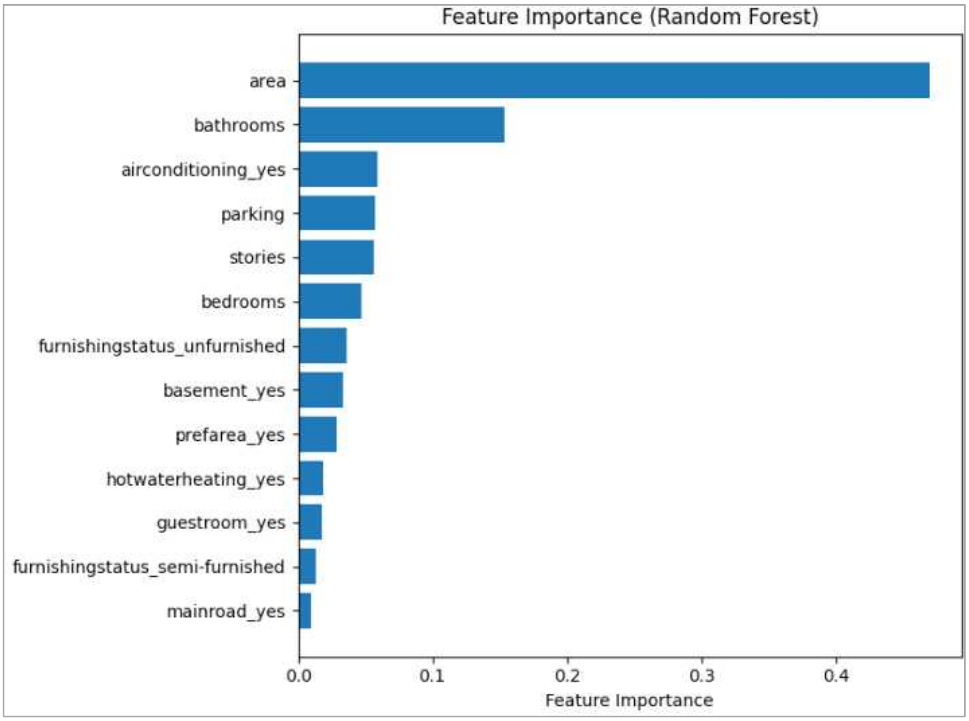
이는 주거 면적과 위생·편의시설, 주차 공간 등의 요소가 주택 가격 형성에 핵심적인 요인으로 작용함을 의미함.

#9-2 시각화 2: Feature Importance

```
importances = rf.feature_importances_
feature_names = X.columns

# 중요도 순으로 정렬
indices = np.argsort(importances)

plt.figure(figsize=(8, 6))
plt.barh(range(len(indices)), importances[indices])
plt.yticks(range(len(indices)), feature_names[indices])
plt.xlabel("Feature Importance")
plt.title("Feature Importance (Random Forest)")
plt.tight_layout()
plt.show()
```



7. 결론

본 연구에서는 주택 가격 데이터를 기반으로 선형회귀 모델과 랜덤포레스트 회귀 모델을 적용하여 예측 성능을 비교 분석함.

두 모델 모두 일정 수준 이상의 예측 성능을 확보하였으며, 본 데이터셋에서는 선형회귀 모델이 상대적으로 더 안정적인 설명력을 보임.

또한 변수 중요도 분석 결과, 주택 면적, 욕실 수, 냉방 시설, 주차 공간 등의 변수가 가격에 큰 영향을 미치는 핵심 요인임을 확인함.

이는 실제 부동산 시장의 가격 형성 논리와도 일치하는 결과로 판단됨.

향후에는 하이퍼파라미터 튜닝, 교차 검증, 추가 데이터 확보 등을 통해 모델의 예측 정확도를 더욱 향상시킬 수 있을 것으로 기대됨.