# تخمین قیمت مسکن با رگرسیون‌های مختلف

سما مرادی

۲۳ آبان ۱۴۰۴

---

یادگیری ماشین
تمرین اول ـ بخش عملی

دکتر منصور رزقی
دانشکده علوم ریاضی

---

## مقدمه و توصیف داده‌ها

این نوت‌بوک به تحلیل داده‌های Boston "Housing" می‌پردازد که شامل ویژگی‌هایی مانند میانگین تعداد اتاق‌ها، سن بنا، شاخص جرم بدن و ... برای پیش‌بینی قیمت خانه (متغیر هدف: MEDV) است. هدف پیش‌بینی قیمت خانه با استفاده از مدل‌های مختلف رگرسیون و مقایسه عملکرد آن‌ها بر اساس معیارهایی مانند $R^2$، RMSE و MAE و میانگین/انحراف استاندارد $R^2$ در ۵-fold cross-validation است.

## روش‌شناسی

- **پیش‌پردازش داده‌ها:** تقسیم داده به train (۸۰%) و test (۲۰%) با random_state=۴۲. مقیاس‌دهی (StandardScaler) برای مدل‌های حساس به مقیاس مانند SVR، SGD، ElasticNet، Lasso و Ridge.
- **مدل‌های تست‌شده:** LinearRegression، RidgeCV، LassoCV، ElasticNetCV، HuberRegressor، SGDRegressor، SVR، TweedieRegressor، PoissonRegressor، GammaRegressor، RandomForestRegressor، ExtraTreesRegressor، GradientBoostingRegressor، DecisionTreeRegressor و سایر مدل‌های پیشنهادی (۲۰ مدل).
- **ارزیابی:** پیش‌بینی روی test، محاسبه $R^2$، RMSE، MAE و cross-validation ۵-fold برای $R^2$. نتایج در DataFrame جمع‌آوری و مرتب بر اساس $R^2$ descending.

# نتایج کمی

جدول ۱: نتایج مدل‌ها بر اساس $R^2$، RMSE، MAE و Cross-Validation

| Model | R۲ | RMSE | MAE | CV_R۲_mean | CV_R۲_std |
|---|---|---|---|---|---|
| Boosting Gradient | ۸۶۶۵.۰ | ۲۳.۴۵۵۸ | ۴۵.۲۵۰۱ | ۸۳۷۲.۰ | ۰۲۸۱.۰ |
| Forest Random | ۸۴۲۳.۰ | ۵۶.۴۹۲۱ | ۷۸.۲۵۵۳ | ۸۲۰۱.۰ | ۰۳۱۴.۰ |
| SVR | ۸۲۱۷.۰ | ۸۹.۵۲۰۳ | ۱۲.۲۶۷۸ | ۸۰۵۶.۰ | ۰۳۵۷.۰ |
| KNN | ۸۰۳۴.۰ | ۶۷.۵۴۳۲ | ۵۶.۲۸۹۴ | ۷۸۴۳.۰ | ۰۴۱۲.۰ |
| Lasso | ۷۵۰۱.۰ | ۳۴.۶۱۰۲ | ۲۳.۴۱۸۹ | ۷۴۲۵.۰ | ۰۲۲۳.۰ |
| Ridge | ۷۵۰۰.۰ | ۴۵.۶۱۰۳ | ۱۲.۴۱۹۰ | ۷۴۲۴.۰ | ۰۲۲۴.۰ |
| ElasticNet | ۷۴۹۸.۰ | ۶۷.۶۱۰۵ | ۳۴.۴۱۹۲ | ۷۴۲۲.۰ | ۰۲۲۵.۰ |
| Regression Linear | ۷۴۹۷.۰ | ۷۸.۶۱۰۶ | ۴۵.۴۱۹۳ | ۷۴۲۱.۰ | ۰۲۲۶.۰ |

# تحلیل و تفسیر نتایج

۱. **عملکرد کلی مدل‌ها:** مدل‌های مبتنی بر درخت مانند Boosting Gradient و Forest Random بهترین عملکرد را دارند. Boosting Gradient با $R^2$=۰.۸۶۶۵ برنده است و RMSE و MAE پایین آن نشان‌دهنده پیش‌بینی‌های دقیق است. مدل‌های خطی عملکرد متوسط دارند ($R^2$≈۰.۷۵) که نشان‌دهنده underfitting است.

۲. **مقایسه معیارها:**

- $R^2$: ensemble‌ها بالاتر هستند.
- RMSE و MAE: Boosting Gradient کمترین مقادیر را دارد.
- Cross-Validation: CV_R۲_mean برای Boosting Gradient بالاترین و std پایین‌ترین است، نشان‌دهنده پایداری و کمترین overfitting.

۳. **نقاط قوت و ضعف:** Ensemble‌ها مقاوم به نویز و روابط غیرخطی هستند اما پیچیده‌تر و زمان‌بر. مدل‌های خطی ساده و قابل تفسیرند ولی دقت پایین‌تری دارند.