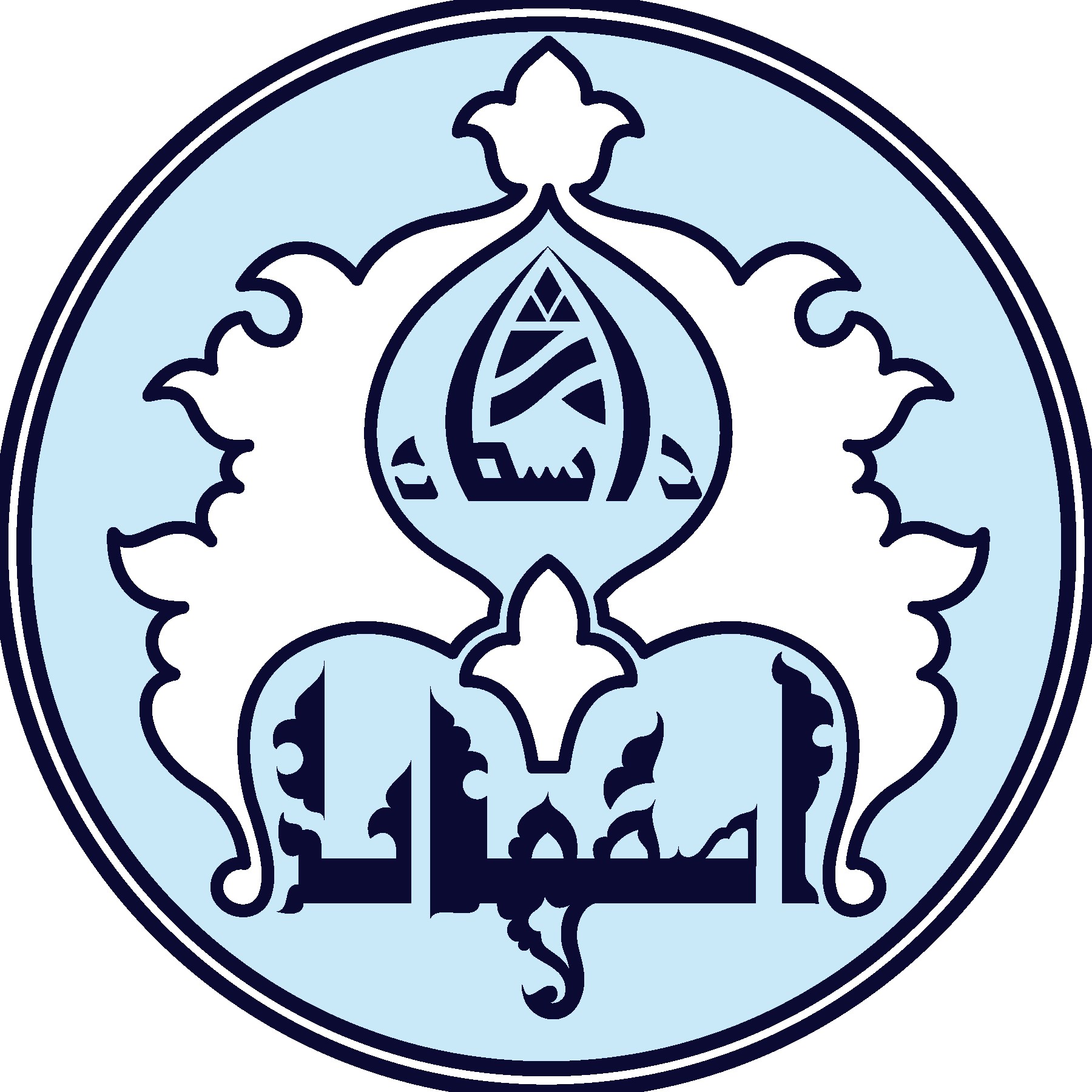
**به نام خداوند بخشنده مهربان**

****

**عنوان پروژه**پیش‌بینی قیمت منزل با استفاده از رگرسیون خطی

**عنوان درس**

یادگیری ماشین

**استاد**

دکتر الهام قصر‌الدشتی

**دستیاران آموزشی**

مهرداد قصابی

مریم صفوی

**گردآورنده**

سید حسین حسینی

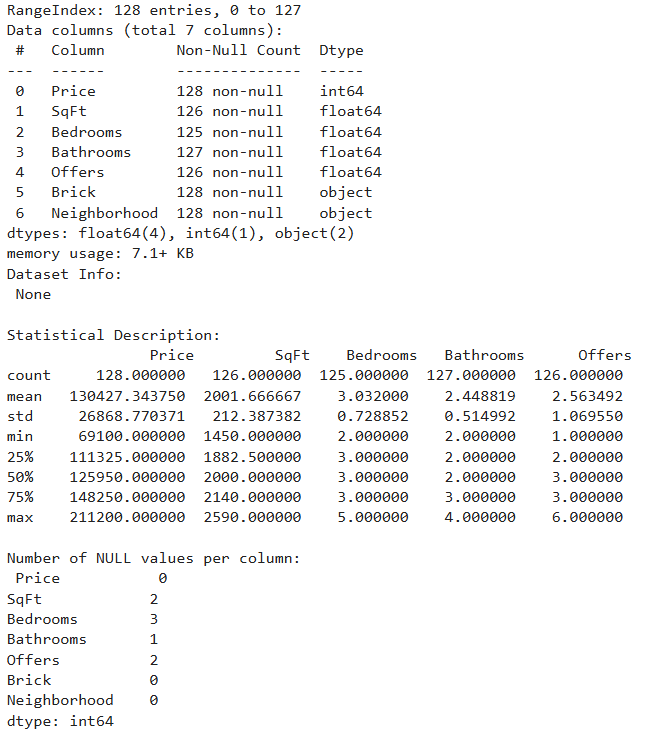
**بهار ۱۴۰۴**

**دانشکده مهندسی کامپیوتر**

**دانشگاه اصفهان**

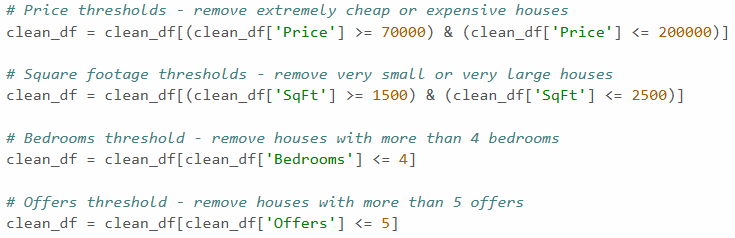
* **کتابخانه‌ها**
* Numpy : برای کار با آرایه‌ها و داده‌ها و انجام عملیات جبری مناسب
* Pandas : برای تحلیل داده‌ها و همچنین انجام عملیات مناسب بر روی دیتافریم‌ها
* Seaborn : برای مصورسازی داده‌ها و روابط بین آن ها و انجام تحلیل بر روی آن‌ها
* Matplotlib : برای مصورسازی داده‌ها و روابط بین آن ها و انجام تحلیل بر روی آن‌ها
* Sklearn : یکی از کتابخانه‌های یادگیری ماشین برای انجام عملیات‌های مرتبط با آن
* Gdown : برای دانلود دیتاست از لینک دریافتی
* **پیش‌پردازش**
* حذف داده‌های NULL

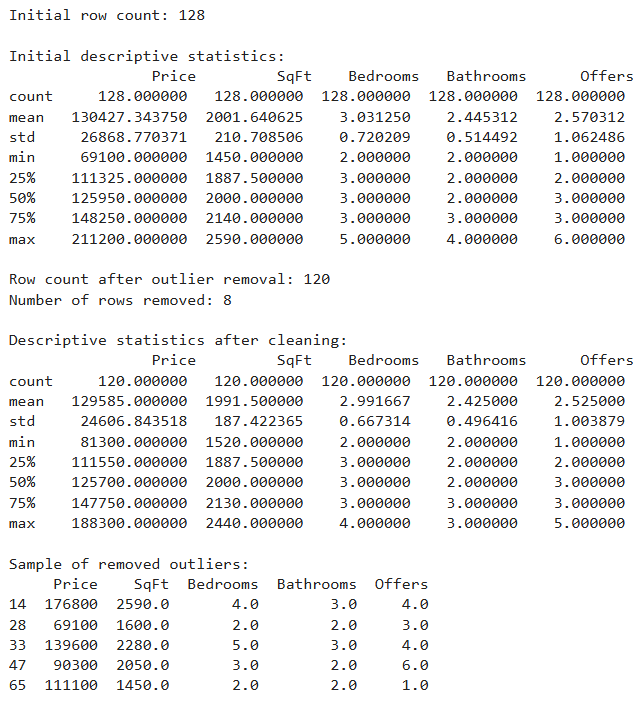
در ابتدا با توجه به اینکه داده‌های NULL باعث کاهش قدرت مدل میشوند باید در ابتدا از دیتاست حذف و با داده مناسب جایگذاری شوند، که در اینجا برای اینکه بتوانیم تعادل مناسبی بین داده‌ها برقرار کنیم از میانگین در داده‌های عددی و از مد در داده‌های غیر عددی استفاده کردیم. در جزئیاتی که در زیر مشاهده میکنید این نوع داده‌ها بسیار کم بوده و نیاز به پردازش‌های خیلی خاص و دقیقی نداشته است و همچنین این روش جز یکی از روش های مرسوم جایگذاری است.



* حذف داده‌های پرت

با توجه به اینکه داده‌های پرت باعث می‌شود که دچار بیش برازش شویم، در نتیجه نیاز است که شناسایی و حذف شوند و عموما از داده‌های آموزش حذف میشوند ولی در اینجا به علت اینکه تعداد داده‌های پرت خیلی محدود هستند و داده‌های آموزش و تست شباهت زیادی بهم دارند، این مورد در هر دو داده اتفاق می‌افتد و همچنین به علت اینکه با استفاده از روش‌هایی مثل IQR و Z\_Score بخش زیادی از داده‌ها به عنوان داده‌های پرت شناسایی شدند، در این زمینه از تحلیل دستی استفاده شده و بر اساس یکسری از متریک‌ها بخشی از داده‌ها شناسایی و حذف شدند که در زیر قابل مشاهده است. (مثلا در روش IQR به علت محدود بودن بازه عددی باعث میشد هر عددی غیر از 3 به عنوان داده پرت شناسایی شود.)





* **مصورسازی و تحلیل**

تحلیل داده‌های قیمت مسکن

 نمودار پراکندگی: ارتباط متراژ و قیمت

* توضیحات: این نمودار نشان می‌دهد که با افزایش متراژ (SqFt) قیمت نیز افزایش می‌یابد.
* نتیجه: ارتباط مثبت بین متراژ و قیمت نشان‌دهنده این است که بزرگ‌تر بودن خانه‌ها معمولاً به قیمت‌های بالاتر منجر می‌شود.

 نمودار جعبه‌ای: تعداد اتاق خواب‌ها و قیمت

* توضیحات: این نمودار توزیع قیمت‌ها را بر اساس تعداد اتاق خواب نشان می‌دهد.
* نتیجه: خانه‌های با تعداد بیشتر اتاق خواب‌ها قیمت‌های بالاتری دارند. همچنین، در خانه‌هایی با اتاق خواب کمتر، قیمت‌ها تنوع بیشتری دارند.

 نمودار جعبه‌ای: تعداد حمام‌ها و قیمت

* توضیحات: این نمودار ارتباط بین تعداد حمام‌ها و قیمت را نمایش می‌دهد.
* نتیجه: مشابه اتاق خواب، تعداد بیشتر حمام‌ها معمولاً به قیمت‌های بالاتر منجر می‌شود. قیمت‌ها برای خانه‌هایی با تعداد کم حمام، بیشتر پراکنده هستند.

 نمودار جعبه‌ای: تعداد پیشنهادات و قیمت

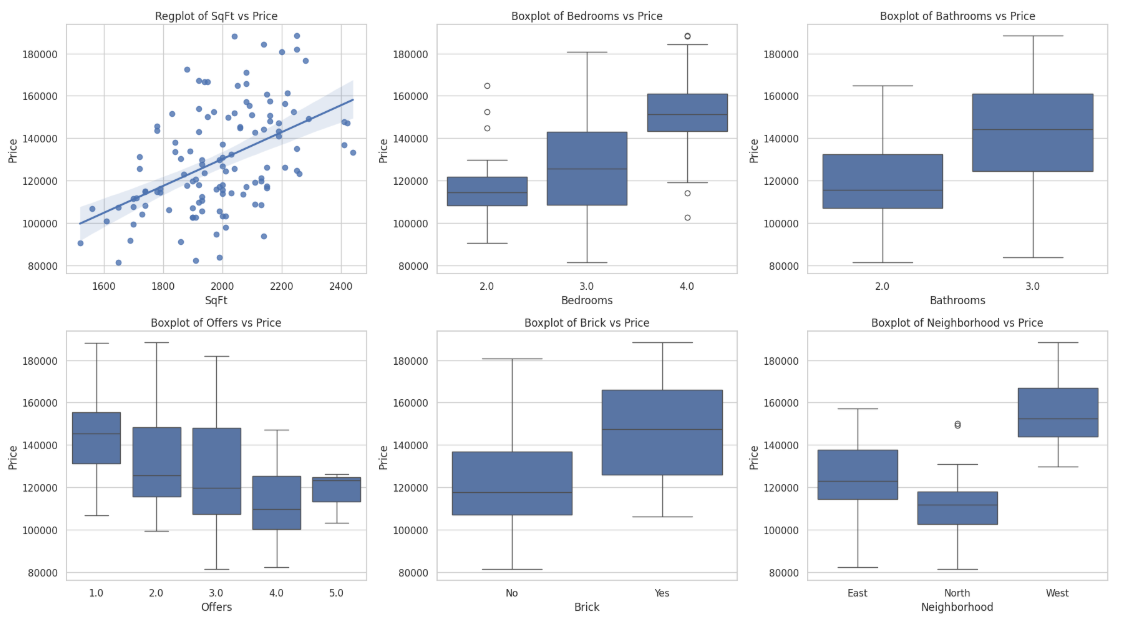
* توضیحات: این نمودار نشان می‌دهد که چگونه تعداد پیشنهادات دریافتی برای یک ملک بر قیمت تأثیر می‌گذارد.
* نتیجه: خانه‌هایی که پیشنهادات بیشتری دارند، قیمت‌های کمتری نیز دارند. این می‌تواند نشان‌دهنده تقاضای بیشتر برای آن ملک باشد و همچنین نشان دهنده سطح توان مالی افراد را نیز در بر داشته باشد.

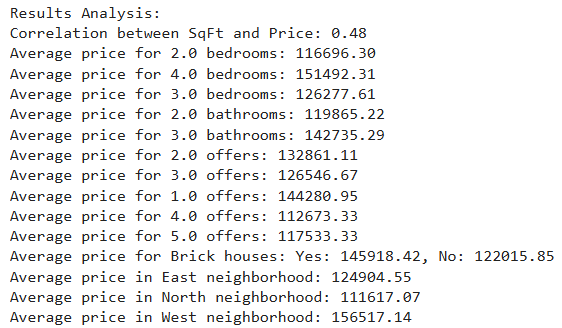
 نمودار جعبه‌ای: نوع ساخت (آجر) و قیمت

* توضیحات: این نمودار قیمت‌گذاری را بر اساس نوع ساخت، یعنی «آجر» یا «غیراجر» نشان می‌دهد.
* نتیجه: خانه‌های ساخته شده از آجر معمولاً قیمت‌های بالاتری دارند که می‌تواند به دلیل کیفیت بالاتر و جذابیت بیشتر این نوع ساخت باشد.

 نمودار جعبه‌ای: محله و قیمت

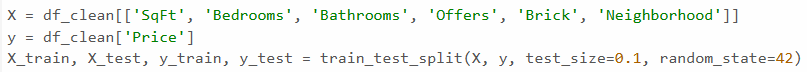
* توضیحات: این نمودار تفاوت قیمت‌ها را بر اساس محله نمایش می‌دهد.
* نتیجه: به نظر می‌رسد که محله‌های غرب (West) قیمت‌های بالاتری تولید می‌کنند در حالی که محله‌های شرق (East) قیمت‌های کمتری دارند. این می‌تواند به عوامل مختلفی از جمله امکانات اطراف و دسترسی به خدمات مرتبط باشد.





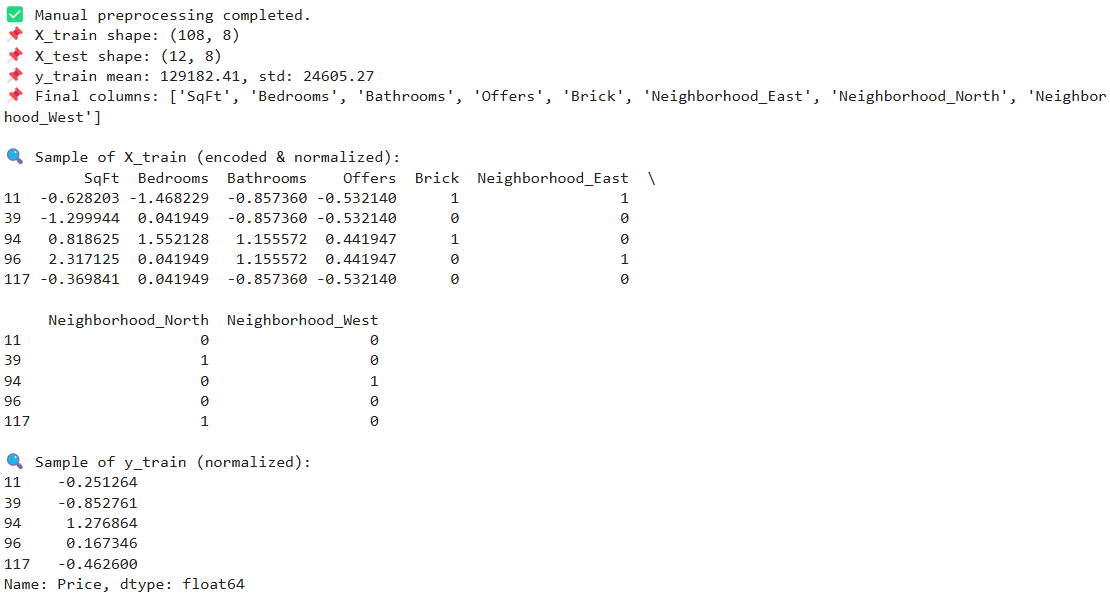
* **تقسیم بندی داده‌های آموزش و تست**

در این مرحله با توجه به اینکه نیاز است مدل را با بخشی از داده‌ها آموزش دهیم و با بخشی دیگر تست کنیم، بعد از عملیات‌های پیش پردازش که انجام شد به سراغ این تقسیم بندی به صورت 90 درصد داده‌های آموزش و 10 درصد داده‌های تست می‌رویم.



* **انکودینگ و نرمال‌سازی**

برای مدل‌هایی مانند رگرسیون خطی که به داده‌ها و ارتباط بین آن ها حساس هستند، نرمال‌سازی یک امر بسیار مهم است، از این رو ابتدا داده‌های آموزش را بر اساس میانگین 0 و انحراف معیار 1 نرمال میکنیم تا بتوانیم به درستی روابط بین داده‌ها را در آموزش مدل مدل ترسیم کنیم و همچنین بعد از نرمال‌سازی داده‌های آموزش از جزئیات آن برای نرمال سازی داده‌های تست استفاده میکنیم و بعد از آن به سراغ انکود کردن و تبدیل داده‌های دسته‌ای به عددی میرویم، که به آن انکود کردن می‌گویند. با توجه به اینکه چه در لیبل انکودینگ و چه در وان هات انکودینگ از 0 و 1 فراتر نمیرویم نیازی به نرمال‌سازی نیست.



* **پیاده‌سازی مدل و تابع هدف**

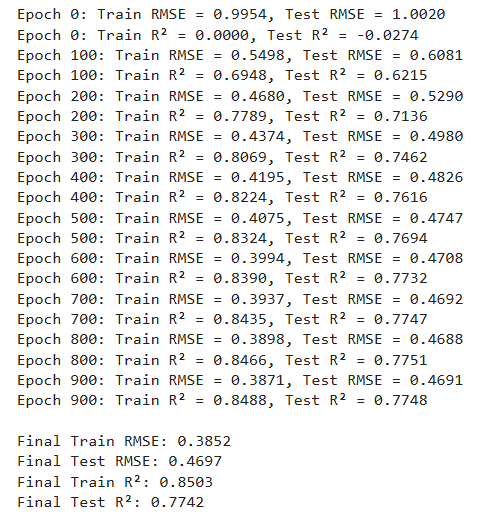
این بخش مهم‌ترین مرحله در این ساختار است و در اینجا از مدل رگرسیون خطی استفاده می‌کنیم که یکی از مدل‌های پایه در زمینه هوش مصنوعی شناخته میشود، که روش عملکرد آن با استفاده از فرمول خط هست، که در آن شیب خط و عرض از مبدا تعریف میکنیم و در اینجا این‌ها به عنوان پارامترها و یا وزن‌های مدل تلقی میشوند و شیب را نیز بایاس میخوانند.

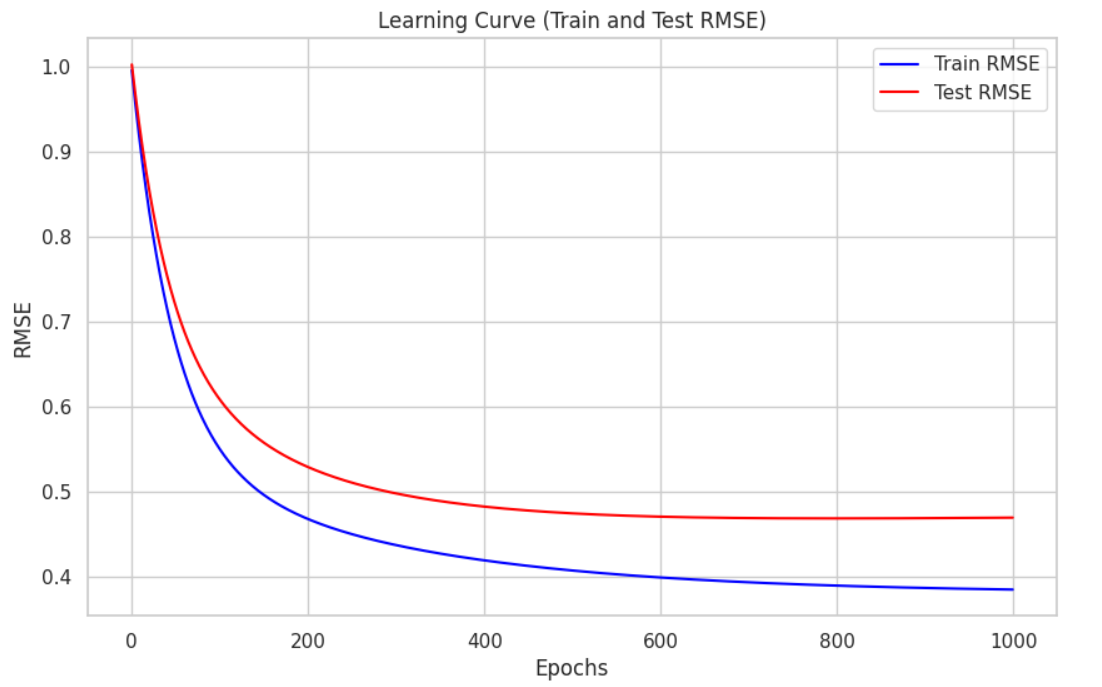
بعد از آن استفاده از تابع هدف یا لاس هست که میزان عملکرد مدل را بررسی میکند، تا بتواند آپدیت‌های مورد نیاز را بر روی وزن‌ها انجام دهد و مدل را به‌درستی آموزش دهد، که در اینجا از Root Mean Square Error استفاده میکنیم که یکی از تابع‌های لاس معروف به شمار میرود.

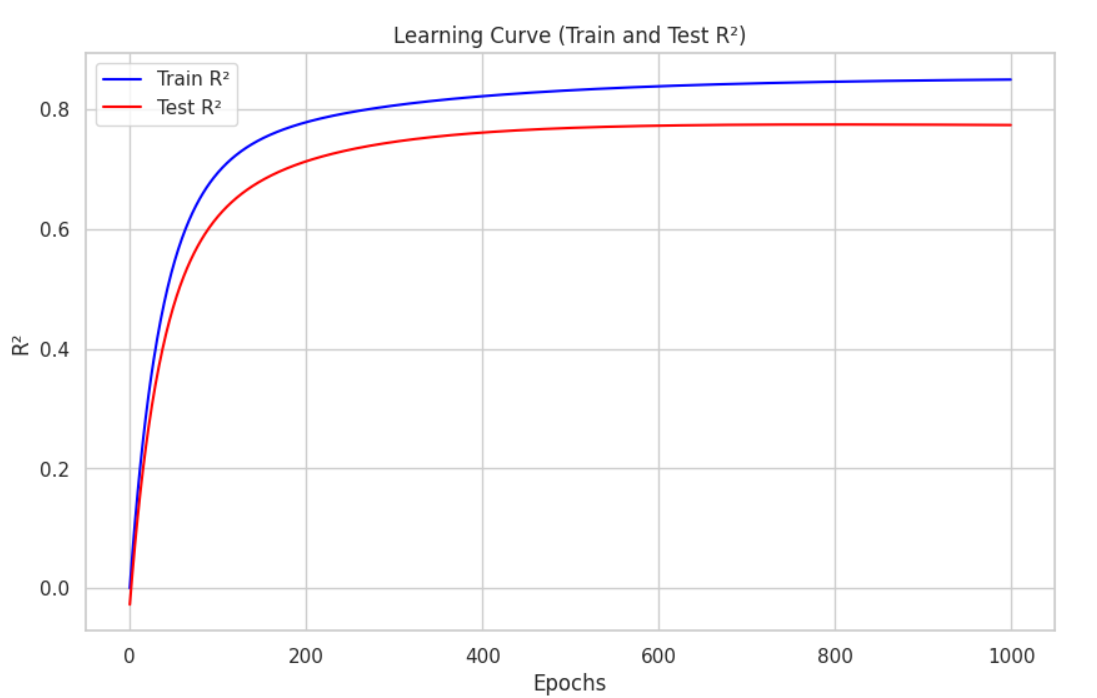
حال با استفاده از این موارد نیاز است که در هر دور وزن‌ها آپدیت شوند، که در اینجا بر اساس یکی از روش‌های مرسوم به نام گرادیان کاهشی این اتفاق می‌افتد و تا زمانی که به مقدار مناسب و بهینه همگرا نشویم این روند متوقف نخواهد شد تا اینکه تعداد دورهای مشخص شده را به پایان برسانیم.

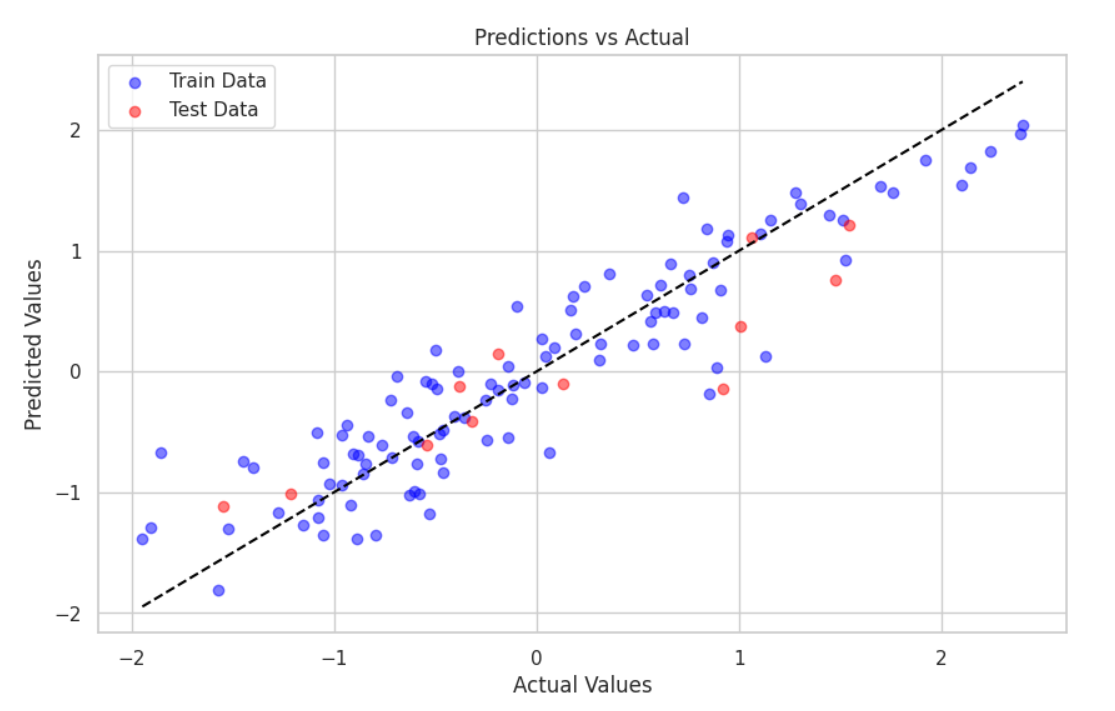
* **ارزیابی و نتیجه‌گیری**

در پایان نتایج به دست آمده را بررسی میکنیم و سعی میکنیم نرخ آموزش و تعداد دور آموزش را به درستی تعیین کنیم تا بهینه ترین خروجی را دریافت کنیم که در اینجا نرخ آموزش را 0.05 و تعداد دور آموزش را 1000 در نظر گرفته ایم تا بهینه‌ترین حالت ممکن که در آن نیز باید مقدار تفاوت خطا داده آموزش و تست هم کم باشد را شناسایی کنیم به دلیل اینکه با مشکل بیش‌برازش مواجه نشویم و همچنین برای مقایسه از R2 نیز بهروری کردیم. (نمودار‌های تحلیلی در زیر قابل مشاهده است)









**در پایان خطی که به صورت خط‌چین مشاهده می‌شود همان خط مورد نظر ما برای پیش‌بینی قیمت منازل خواهد بود.**

* **منابع**
* Ng, Andrew. Machine Learning (Coursera)
* شریفی زارچی، علی. دوره یادگیری ماشین، دانشگاه صنعتی شریف
* Géron, Aurélien. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, 2019
* OpenAI. ChatGPT

**سید حسین حسینی دولت آبادی**

**😊 موفق باشید 😉**