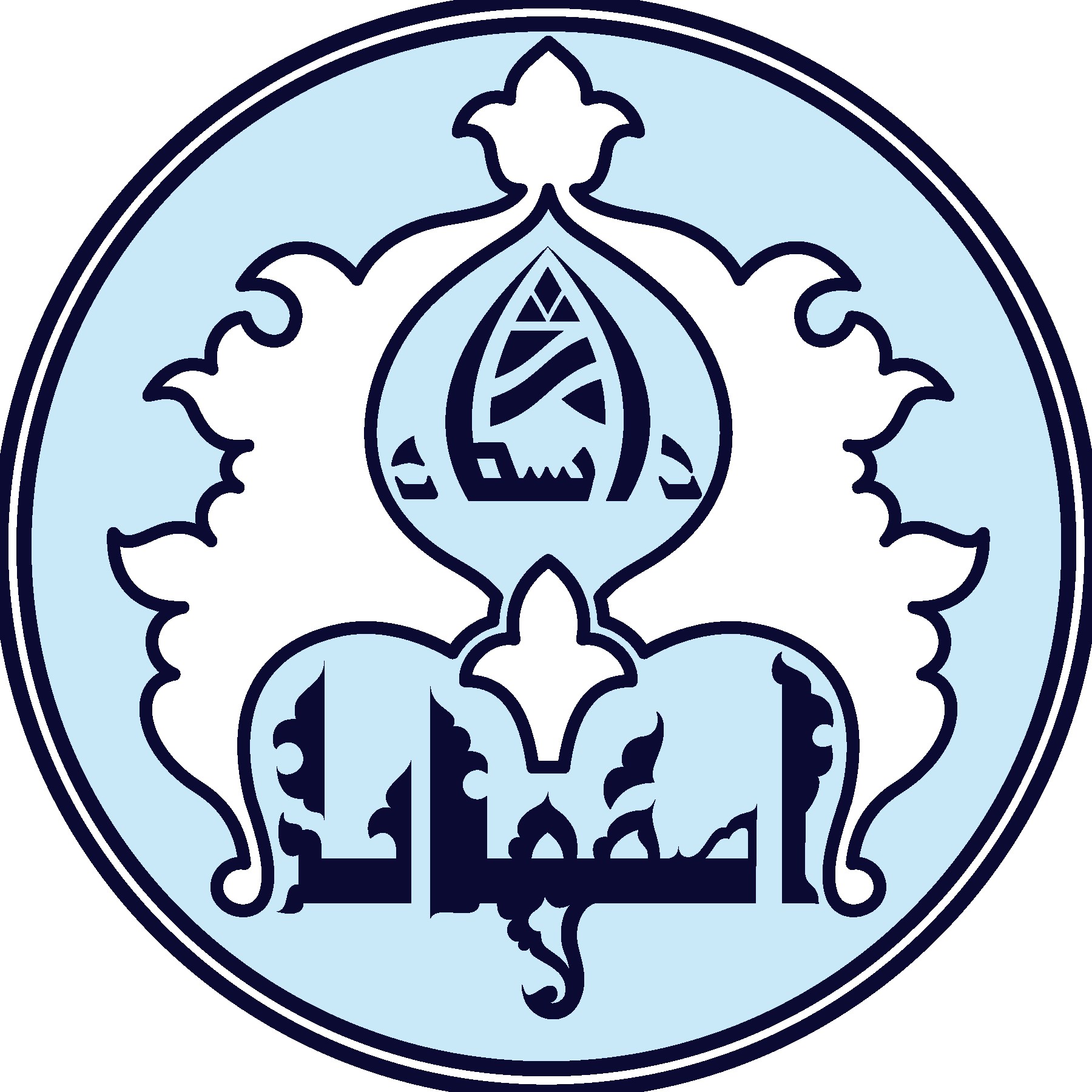
**به نام خداوند بخشنده مهربان**

****

**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

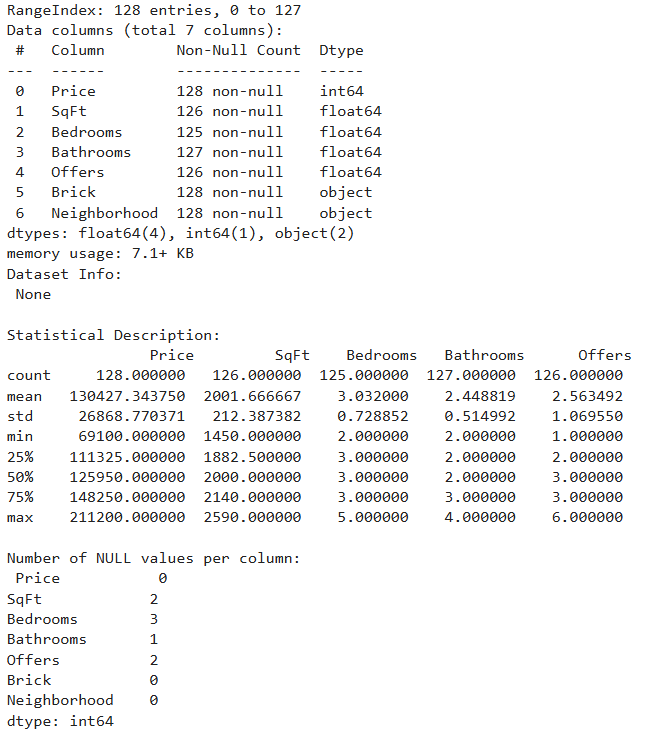
**بهار ۱۴۰۴**

**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.)

