محمد عبدالهی
400422133

تمرین دوم:

در این تمرین ما به بررسی و تحلیل دیتا Apartment rental offers in Germany میپردازیم. این دیتاست شامل اطلاعات مربوط به خانه های برای اجاره در شهر نیویورک بوده و پس از انجام مراحل پاکسازی داده سعی در تخمین قیمت خانه براساس فیچرهای موجود داریم.

تسک های اصلی:

1. **در بخش اول به کارهایی همچون لود کردن دیتا و پاکسازی دیتا و پر کردن مقادیر خالی و حذف فیچرهای بدون اطلاعات و حذف داده های پرت پرداختیم. کدها و پیاده سازی ها به طور کامل در نوت بوک پیاده سازی شده و شامل کامنت برای توضیحات بیشتر میباشد.**

2. **Data visualization: در این بخش سعی داریم تا با مصورسازی داده اطلاعات کلی درباره فیچرها و فهم بهتر آنها داشته باشیم.**
   در گام اول چند مورد از توزیع های فیچرهای continuous را نمایش داده ایم و در ادامه هم سعی کرده تا تلفیقی از فیچرها و تاثیر آنها بر یکدیگر را نمایش دهیم.

3. **در این بخش با استفاده از دیتاست پاکسازی شده و در چند حالت با اعمال متدهای مختلف بر روی فیچرها سعی در مدل سازی و تخمین قیمت براساس این فیچرها داشتیم.**
   در حالت اول تمام مقادیر نال را حذف را کرده و با استفاده از ordinal encoding فیچرهای کتگوریکال را انکود کرده و سپس از طریق linear regression قیمت را تخمین زده ایم.
   در حالت دوم مقادیر نال را با مد پر کرده و بقیه پایپ لاین به صورت قبل طی میشود.

در حالت سوم باز هم از مد برای پر کردن مقادیر نال استفاده کرده و فیچرهای کتگوریکال را با one hot encoding انکود کرده ایم که باعث کاهش ارور تخمین ما شد.

در حالت چهارم پس از استفاده از مد و one hot الگوریتم pca را برای کاهش ابعاد بر روی دیتاست پیاده سازی کردیم و ارور تخمین را کاهش دادیم.

در حالت پنجم پس از استفاده از مد و one hot الگوریتم select k best features را برای کاهش ابعاد بر روی دیتاست پیاده سازی کردیم و ارور تخمین را کاهش دادیم.

در حالت ششم پس از استفاده از مد و one hot و pca با الگوریتم های linear regression, lasso regression, ridge regression, gradient boosting سعی در تخمین قیمت داشتیم که بهبود نتایج در حد قابل توجهی نبود.

در حالت هفتم برای پر کردن مقادیر null از الگوریتم iterative imputation کتابخانه sckit-learn بر روی فیچرهای continuous استفاده کردیم و بر از متد one hot بر روی فیچرهای کتگوریکال استفاده کردیم سپس برای کاهش ابعاد از pca استفاده کردیم و در نهایت از linear regression برای تخمین قیمت استفاده کردیم که در این حالت با کاهش چشمگیر مقدار ارور روبرو شدیم و بهترین نتیجه حاصل به صورت زیر شد:

```
Mean absolute error : 30.502854
Root mean squared error : 58.327246
R2_score : 0.988366
```

4. در این قسمت تابع نرمالیزشن خود را با کمک multiprocessing بر روی دیتاست پیاده کرده و زمان اجرا را گزارش کرده ایم.
5. در این بخش هم به کمک dask و pyspark سعی در پاکسازی دیتاست خود داشته ایم و به کمک dask تابع نرمالیزشن خود را بر روی دیتاست پیاده سازی کردیم و به کمک pyspark اعمال مقدماتی همچون حذف داده های نال و حذف فیچرهای بدون استفاده را داشته ایم.