تکلیف اول درس مبانی داده کاوی

امید رئیسی (۹۶۲۱۱۶۰۰۱۵)

۱)

1. **یادگیری با ناظر**
2. **یادگیری بدون ناظر**
3. **یادگیری با ناظر**
4. **یادگیری بدون ناظر**
5. **یادگیری با ناظر**
6. **یادگیری با ناظر**
7. **یادگیری با ناظر**
8. **یادگیری بدون ناظر**

۲) **داده‌های آموزشی داده‌های برچسب داری هستند که در الگوریتم‌های با ناظر برای یادگیری الگوریتم بکار گرفته می‌شوند و در واقع روابط میان متغییر‌های پیشگو(ورودی) و خروجی را به الگوریتم نشان می‌دهند اما در مقابل داده‌های اعتبار‌سنجی داده‌های برچسب داری هستند که پس از اتمام یادگیری الگوریتم برای ارزیابی آن استفاده می‌شوند.**

۳**) با بررسی اولیه داده‌ها متوجه می‌شویم که کلاس هدف به صورت نامتوازن می‌باشد به صورتی که ۷۸٪ از داده‌ها از طبقه هدف 0 و مابقی از طبقه هدف 1 می‌باشند. ابتدا باید داده‌ها را در یک جدول مرتب سازی کنیم و پس از بررسی اینکه آیا متغییر بدون مقداری وجود دارد یا نه (بررسی missing value ها) باید برای داده‌های آموزشی خود به صورتی نمونه برداری کنیم که تعادل میان هر دو طبقه هدف برقرار باشد.پس از متعادل کردن داده‌ها حدود 80٪ آن‌ها را به عنوان داده آموزشی و مابقی را به عنوان داده‌های اعتبارسنجی انتخاب می‌کنیم.**

**حال برای مدل خود با استفاده از کتابخانه‌های پایتون و توابع رگرسیون خطی تأثیر هر متغییر ورودی را بر متغییر هدف پیدا کرده و مدل خود را با داده‌های اعتبارسنجی آزمایش می‌کنیم.(ممکن است در مراحل مدل سازی با مصور سازی داده‌ها و استفاده از خلاصه سازی داده‌ها مانند ماتریس همبستگی از روش های کاهش ابعاد نیز استفاده کنیم.)**

**۴) تعداد رکورد‌ها برابر با 1000 و تعداد متغیر‌ها برابر با 50 می‌باشد پس در کل 50000 مقدار داریم که 5٪ آن برابر با 2500 می باشد:()**

**اگر بدترین حالت را در نظر بگیریم به صورتی که هر رکورد به صورت میانگین به اندازه 2.5 مقدار نداشته باشد آنگاه همه رکورد‌ها حذف می‌شوند.()**

**اگر بهترین حالت را در نظر بگیریم به صورتی که رکورد‌ها همه متغییر‌هایشان بی مقدار باشند آنگاه باید 50 رکورد را حذف کنیم.()**

۵)

الف)**متغییر‌های Color و Fuel\_Type متغیر‌های طبقه‌ای هستند که به ترتیب دارای 10 و 3 کلاس می باشند.**

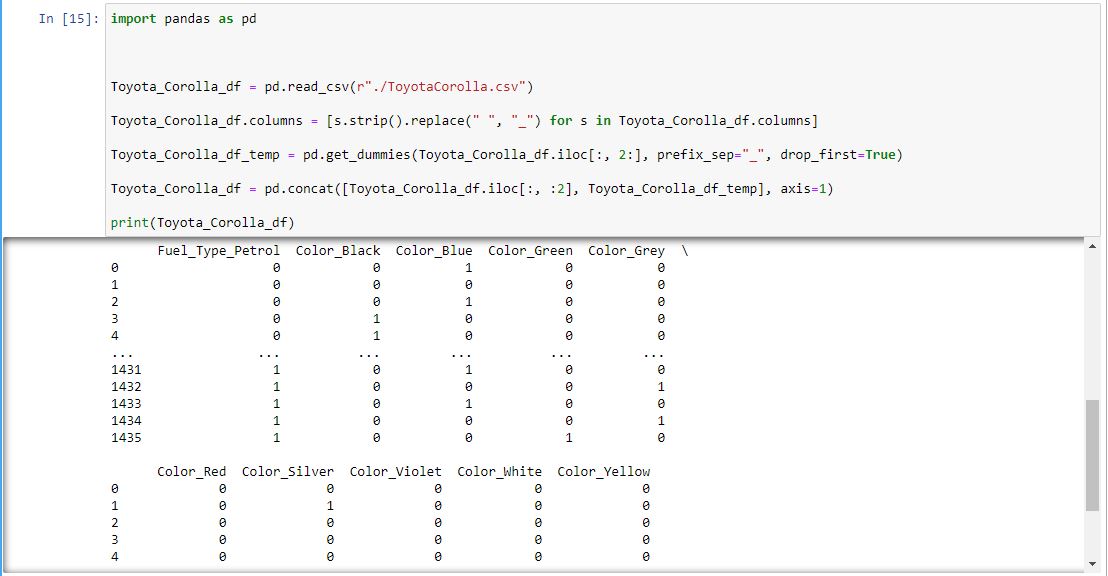
**Color = ['Beige', 'Black', 'Blue', 'Green', 'Grey', 'Red', 'Silver', 'Violet', 'White', 'Yellow']**

**Fuel\_Type = ['CNG', 'Petrol', 'Diesel']**

**برای مثال اگر متغییر Color را در نظر بگیریم می‌توانیم این 10 کلاس را در 9 متغییر دودویی که هر کدام نماینده یکی از این رنگ‌ها می‌باشد ذخیره کنیم (به جز رنگ Beige) بدین ترتیب که اگر متغییر مربوط به رنگ داری مقدار 1 بود بدین ترتیب رنگ ماشین برابر با آن رنگ است (0 بودن تمامی متغییر‌ها نشان می‌دهد که رنگ ماشین Beige است.)**

**به همین ترتیب برای متغییر Fuel\_Type اگر تمام متغییر‌های جایگزین 0 باشند نوع سوخت مصرفی ماشین CNG می‌باشد.**

**این جایگزینی متغییر‌ها را (تبدیل متغییر‌های طبقه‌ای به چند متغییر باینری و عددی را که به آن‌ها Dummy Variables می‌گوییم) در پایتون با استفاده از کتابخانه pandas و متود get\_dummies انجام می‌شود.**

****

ب)**برای آماده سازی داده‌ها ابتدا باید بررسی کنیم که آیا missing value داریم یا نه اما چون داده‌ها کامل هستند مستقیم سراغ افراز داده‌ها می‌رویم.**

**افراز داده‌ها را با استفاده از متود train\_test\_split از کتابخانه scikit-learn در پایتون انجام می‌دهیم.**

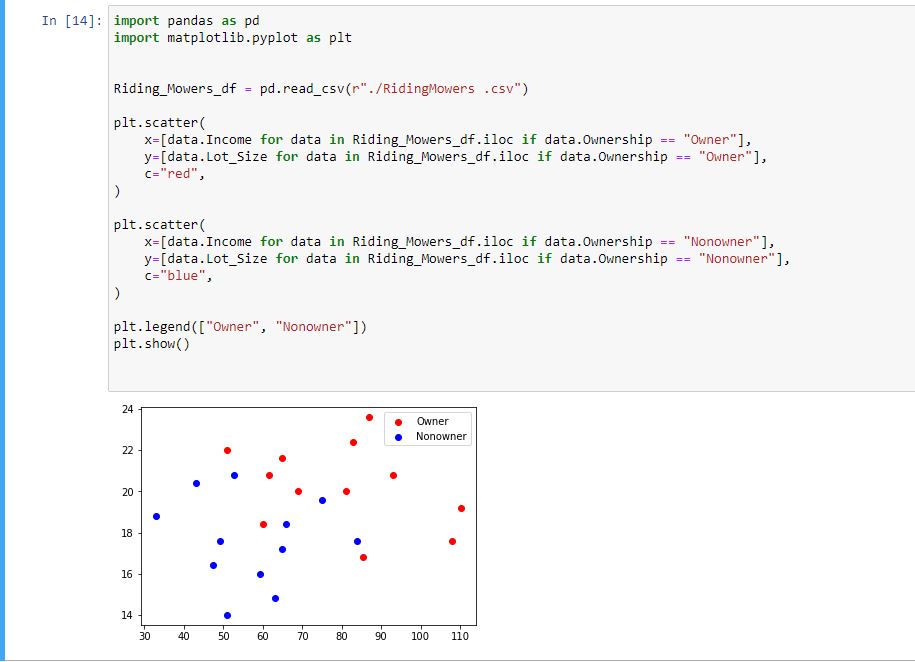
****

**training\_data (داده‌های آموزشی) : داده‌هایی که برای آموزش الگوریتم‌های خود استفاده می‌کنیم تا آنها را تبدیل به مدل کنیم.**

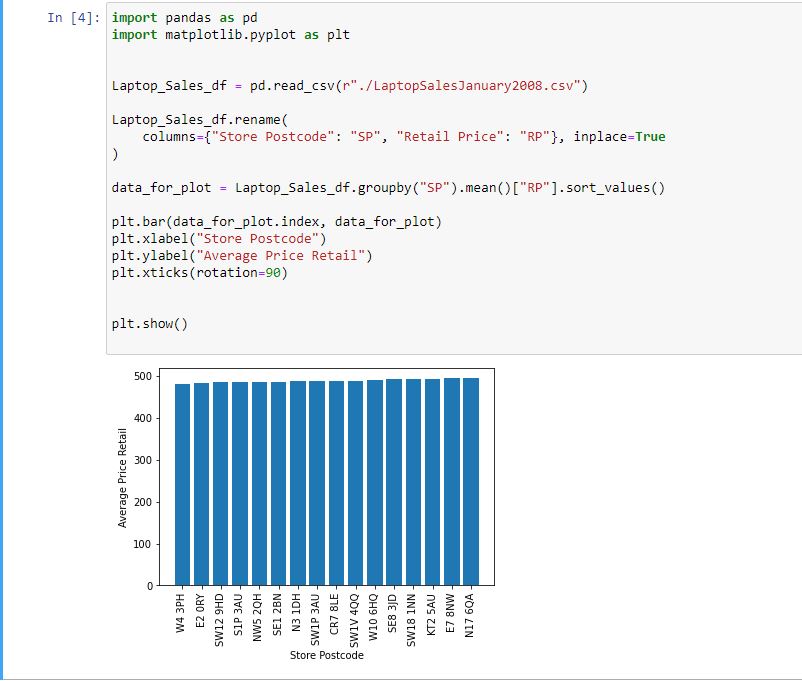
**test\_data (داده‌های اعتبارسنجی) : داده‌هایی که برای مقایسه مدل‌های بدست آمده استفاده می‌کنیم تا بهترین مدل را پیدا کنیم.**

**evaluation\_data (داده‌های ارزشیابی) : داده‌هایی که برای ارزیابی مدل منتخب بر روی داده‌های جدید استفاده می‌شوند تا از عدم رخداد بیش‌ برازش بر روی داده‌های test\_data مطمئن شویم.**

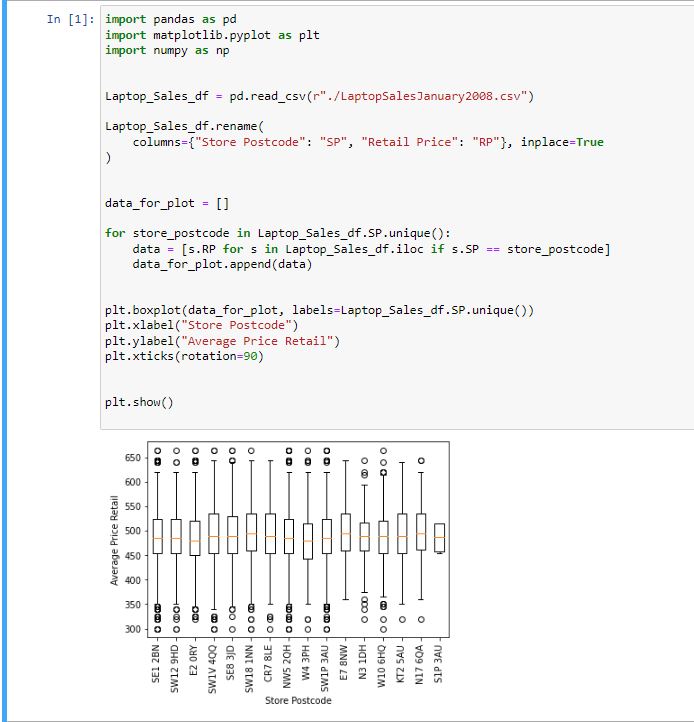
**۶) برای رسم نمودار از کتابخانه matplotlib استفاده می‌کنیم.**

****

**۷) الف) همانطور که از شکل زیر مشخص است بیشترین میانگین خرده فروشی مربوط به شعبه N17 6QA به ارزش 494 دلار و کمترین میانگین خرده فروشی مربوط به شعبه W4 3PH به ارزش481 دلار می‌باشد.**

****

**ب)** **در مقایسه N17 6QA (بالاترین) با W4 3PH (پایین‌ترین) میانه ها مشابه هستند، اما W4 3PH محدوده قیمت‌های بزرگ‌تری دارد و خارج از چارک‌های 1 و 4 پرت‌تر است.**

****