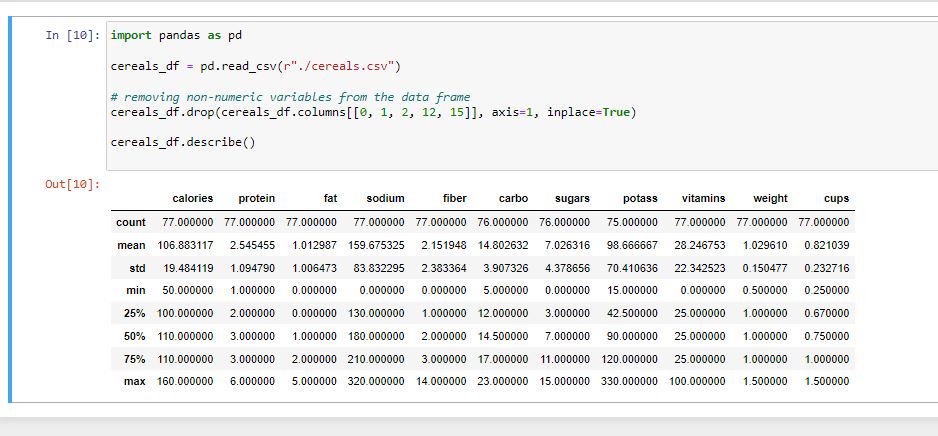
**تکلیف دوم درس مبانی داده کاوی**

**امید رئیسی (۹۶۲۱۱۶۰۰۱۵)**

**۱)**

**الف) متغییر‌های name, mfr, type اسمی هستند و به ترتیب نام غلات، تولید کننده آن‌ها و نوع سرد و یا گرم بودن را بیان می‌کنند. متغییر‌های shelf, rating ترتیبی هستند زیرا shelf ارتفاع قفسه غلات را بیان می‌کند که با کیفیت و رضایت کاربران از غلات رابطه مستقیم دارد و متغییر rating هم امتیاز کاربران به غلات را نشان می‌دهد. سایر متغییر‌ها کمّی و عددی هستند.**

**ب)**

****

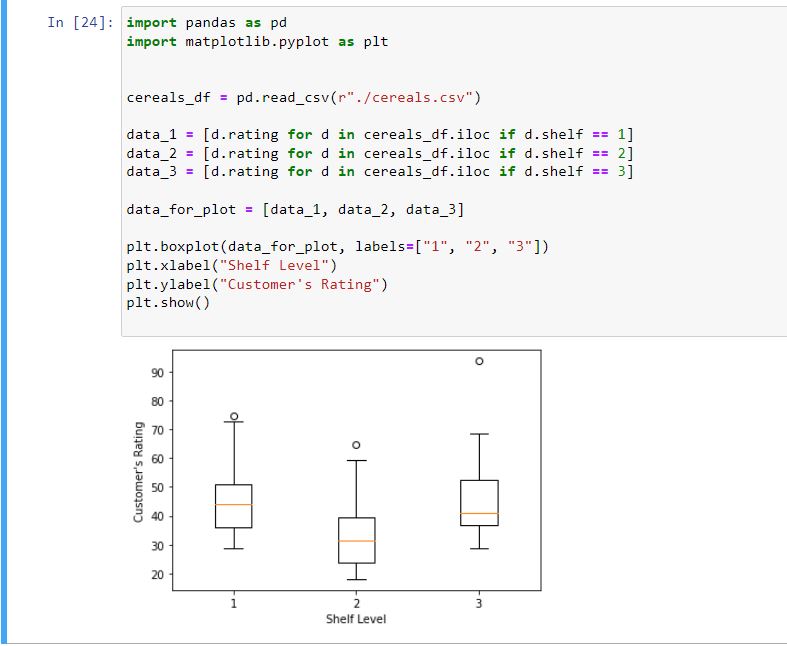
**ج)**

1. **قند، پتاسیم و کربوهیدرات دارای بیشترین تغییر پذیری هستند.**
2. **متغییر‌های پتاسیم، فیبر و چربی دارای چولگی می‌باشند. (در نمودار هیستوگرام این متغییر‌ها نقطه اوج وتراکم مقادیر نمودار از میانگین متغییر فاصله دارد.)**
3. **متغییر max در جدول قسمت ب کرانه متغییر‌ها را نشان می‌دهد که برای مثال قند و پتاسیم به ترتیب ۱۵ و ۳۳۰ می‌باشد.**

****

**د) نمودار زیر نشان می‌دهد که تعداد غلات از نوع گرم خیلی کم است و با این تعداد کم نمی‌توان مقایسه‌ای بین دو نوع سرد و گرم انجام داد.**

**ه) با توجه به نمودار زیر سطح ارتفاع قفسه از زمین در امتیاز دهی کاربران مؤثر است و بهتر است برای بررسی بیشتر داده‌های قفسه شماره ۲ صرف نظر شوند.**

****

**و)**

1. **زوج متغییر های [(potas, fiber), (suger, rating), (calories, rating)] دارای بیشترین همبستگی هستند.**
2. **متغییر‌های که دارای همبستگی بالا می‌باشند باید یکی از آن‌ها حذف شود این کار اصولا با روش‌های مختلف انجام می‌شود که اساس کار آن‌ها سعی و خطا و تاثیر متغییر‌های جدید بر مدل می‌باشد. در این حالت که با متغییر‌های عددی سروکار داریم روش PCA روش مناسبی برای این کار می‌باشد.**
3. **در صورت نرمال‌سازی داده‌‌ها متغییر‌ها دارای میانگین، واریانس و در نتیجه انحراف معیار جدید می‌شوند و تغییر این شاخص ها در مقادیر ماتریس همبستگی تأثیر زیادی می‌گذارد و ممکن است همبستگی‌های جدید به وجود آمده و همبستگی‌های قبلی از بین بروند.**

****

**ز) مؤلفه اصلی اول که در روش PCA کاربرد دارد در واقع خطی است که بیشترین تغییر پذیری را درنمودار پراکنشی دو متغییر با همبستگی بالا را شامل می‌شود.در این نمودار که به هدف بررسی امکان حذف متغییر با همبستگی بالا رسم می‌شود این خط دارای بیشترین واریانس است و همچنین مجموع فواصل نقاط مختلف نمودار از این خط دارای کمترین مقدار ممکن است.**

**2)**

|  |  |  |  |
| --- | --- | --- | --- |
| **رده پیش‌بینی شده** | |  | |
| **غیر‌کلاهبردار** | **کلاهبردار** |
| **32** | **30** | **کلاهبردار** | **رده واقعی** |
| **920** | **58** | **غیر‌کلاهبردار** |

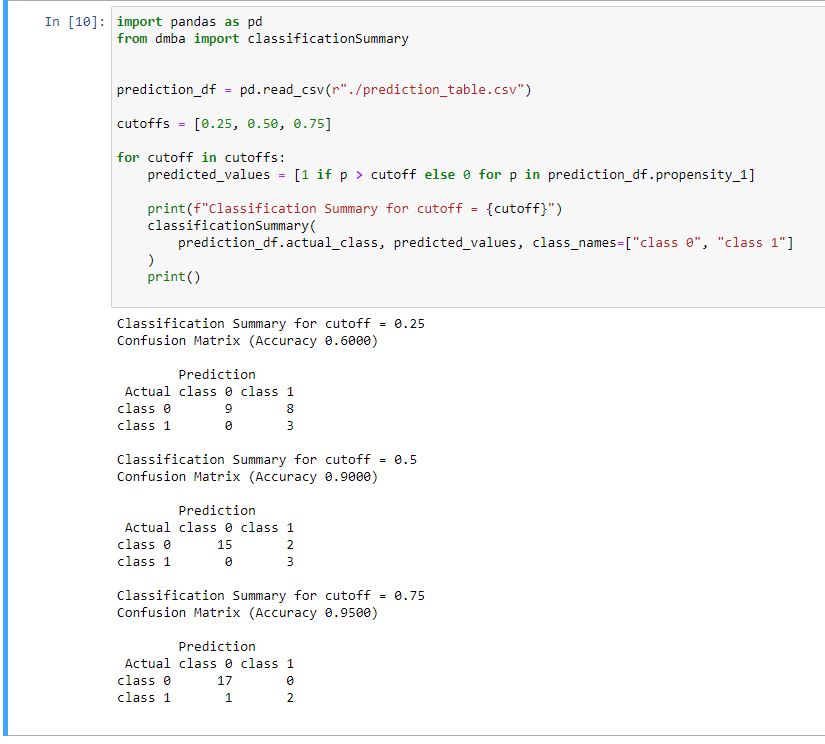
**نرخ خطای کل از رابطه زیر بدست می‌آید.**

**که و تعداد پیش‌بینی‌های اشتباه را نشان می‌دهند.**

**این رابطه نشان می‌دهد که این مدل در تقریبا ۹٪ داده‌ها مرتکب اشتباه می شود.**

**3)**

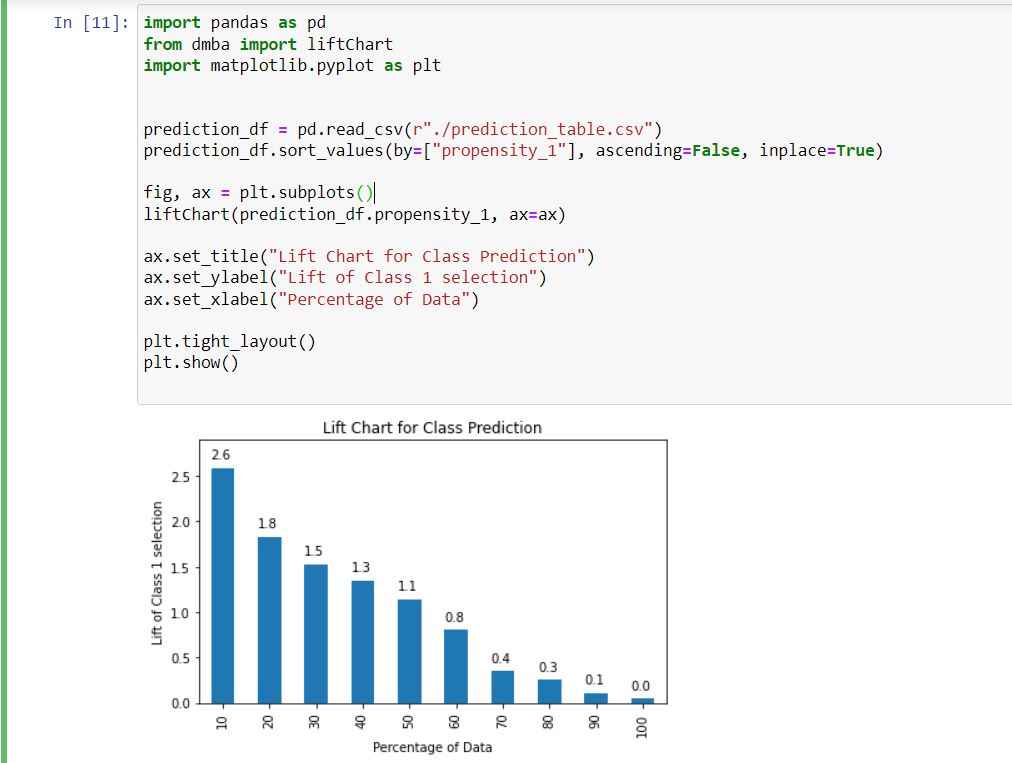
**الف)**

****

**نرخ خطا از رابطه بدست می‌آید و همچنین حساسیت (sensitivity) و وضوح (specifity) از روابط زیر بدست می‌آیند: ( در اینجا پیشبینی کلاس 1 از پیشبینی کلاس 0 مهمتر است.)**

|  |  |
| --- | --- |
| *Cutoff = 0.25* |  |
|  |
|  |
| *Cutoff = 0.50* |  |
|  |
|  |
| *Cutoff = 0.75* |  |
|  |
|  |

*ب)*

**