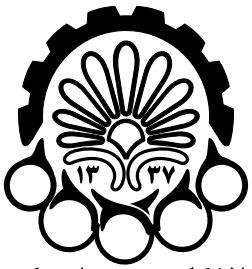

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

به نام خدا

سری اول تمارین درس داده کاوی

استاد درس:

دکتر مریم امیر مزلقانی

نیم سال اول ۱۴۰۲-۱۴۰۳

راه ارتباطی:

Aut.DataMining.Fall@gmail.com

توضیحات:

۱. این تمرین شامل دو بخش عملی و تئوری هست و پاسخ به هر دو بخش الزامی است.

۲. تمرین عملی در قالب یک نوت بوک آماده شده است و دیتای لازم برای این تمرین در پوشه *Practical* موجود است.

۳. در نوت بوک تمرین عملی حتما هر خواسته را در بلوک مربوط به خودش انجام دهید.

٤. حین حل تمرین عملی شما نیازمند به ساخت ۴ فایل جدید هستید (۱ فایل csv و ۳ فایل png)، به نحوه نامگذاری این فایل ها دقت کنید.

٥. **ملاک اصلی انجام تمارین عملی، گزارش است** و ارسال کد بدون pdf گزارش فاقد ارزش است. لذا برای این بخش یک فایل گزارش تهیه کنید و در آن برای هر بخش از تمرین عملی، توضیحات مربوط به آن را ذکر کنید.

٦. خوانا و مرتب بودن پاسخ های شما در نمره‌تان تاثیر مثبت خواهد داشت.

۷. مطابق قوانین دانشگاه هر نوع کپی برداری و اشتراک کار دانشجویان غیرمجاز بوده و برخورد خواهد شد. استفاده از کدها و توضیحات اینترنت به منظور یادگیری **الزاما با ذکر منبع** بلامانع است.

۸. فایل های ایجاد شده در تمرین عملی + نوت بوک + گزارش + پاسخ تمارین تئوری را به صورت زیپ در آورده و با فرمت **StudentID_DM01.zip** در سامانه کورسز آپلود نمایید.

نکته مهم:

در طول ترم شما **مجموعاً ۷ روز** مجاز به ارسال تمارین خود با تاخیر هستید.

مثال: در تمرین اول پاسخ خود را با ۱ ساعت تاخیر ارسال می‌کنید، در این صورت شما ۶ روز دیگر امکان ارسال با تاخیر دارید. (۱ دقیقه تاخیر با ۲۳ ساعت تاخیر تفاوتی ندارد و یک روز از مهلت های شما کسر می‌شود.)

مدیریت این ۷ روز با شماست و پس از اتمام این مهلت **به ازای هر روز تاخیر ۲۰٪** از نمره شما کسر خواهد شد.

بخش تئوری:

سوال ۱.

فرض کنید شرکتی در حوزه تولید دارو شما را به عنوان یک متخصص در زمینه داده کاوی به خدمت گرفته است. تمام مراحل مورد نیاز در این مسیر را از ابتدا تا انتها نامبرده و به صورت مختصر توضیح دهید. (دیتاست و مسئله مورد نظر را به اختیار مشخص کنید)

سوال ۲.

Noise و Outlier را تعریف کرده و سپس به سوالات زیر پاسخ داده و توضیح دهید:

۱) پیدا کردن outlier چه سودی دارد؟ مسئله مرتبط با داده کاوی پیدا کنید که پیدا کردن outlier ها هدف اصلی آن باشد.
۲) درباره معایب Noise و outlier در داده به طور مختصر توضیح دهید.

سوال ۳.

به سوالات زیر پاسخ دهید:

۱) برای هر نوع داده دو مثال بیان کرده و دلیل خود را به اختصار توضیح دهید. (Nominal, Ordinal, Interval, Ratio)
۲) ویژگی های "شماره دانشجویی " و "جنسیت" را در نظر بگیرید. برای هر کدام مسئله ای مطرح کنید که این ویژگی ها در آن تاثیرگذار بوده و نباید حذف شوند.
۳) هرکدام از مفاهیم میانه، مد و میانگین برای کدام یک از داده های اسمی، ترتیبی، بازه ای و نرخی قابل تعریف است؟

سوال ۴.

با توجه به دیتاست ارائه شده به سوالات پاسخ دهید.

ID	Color	Size	Weight	Shape
1	Blue	10	0.5	Triangle
2	Red	8	0.3	Triangle
3	Blue	8	0.7	Round
4	Green	12	0.3	Square

۱) روش های one-hot encoding و label encoding را شرح دهید.
۲) هر کدام از روش های نامبرده شده را بر روی دیتاست پیاده کنید. در روش one-hot encoding حداقل تعداد ستون مورد نیاز چند تاست؟
۳) کدام روش مناسب تر است؟ آیا این نتیجه گیری در هر دیتاستی برقرار است؟

سوال ۵.

بردار $v = \begin{bmatrix} 10 \\ 3 \\ -14 \end{bmatrix}$ را در نظر بگیرید. برای مجموعه داده $\{v, -v, 2v, -2v\}$ با استفاده از روش PCA و تحلیل آن تصویر یک بعدی با بیشترین واریانس و خطی که داده ها روی آن تصویر شده است را بیابید (نیازی به حساب کردن ماتریس کوواریانس نیست)

سوال ۶.

با در نظر گرفتن دو رابطه زیر برای مقادیر β_0, β_1 ، مقدار کوواریانس مربوط به این دو پارامتر را حساب کنید. در چه حالتی این دو پارامتر مستقل می شوند؟(y_i و x_iپارامتر یا متغیر تصادفی هستند و $\bar{x}$ و $\bar{y}$ میانگین این متغیر های تصادفی هستند)

$$\beta_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \beta_2 = \bar{y} - \beta_1 \bar{x}$$

سوال ۷.

به دیتاست به سوالات زیر پاسخ دهید.

id	Col1	Col2	Class
1	12	1	A
2	5	0.5	B
3	9	1.2	C
4	13	4	C
5	4	3	A
6	7	3.2	B

۱) Box plot مربوط به هر یک از ویژگی هارا رسم کنید و درباره کاربرد این نمودار به صورت مختصر توضیح دهید.

۲) scatter matrix مربوط به دیتاست ارائه شده را رسم کنید و درباره کاربرد این نمودار به صورت مختصر توضیح دهید.

سوال ۸.

mean centering عمل کم کردن میانگین یک متغیر از تمام مشاهدات روی آن متغیر در مجموعه داده است به طوری که میانگین جدید متغیر صفر باشد. ثابت کنید برای یک مجموعه داده که این عمل روی آن انجام شده است با نقاط $\overline{X}_1 \dots \overline{X}_n$ عبارت زیر صادق است.

$$||\overline{X_i}||^2 + ||\overline{X_j}||^2 = \frac{\sum_{p=1}^{n}||\overline{X_i} - \overline{X_p}||^2}{n} + \frac{\sum_{q=1}^{n}||\overline{X_j} - \overline{X_q}||^2}{n} - \frac{\sum_{p=1}^{n}\sum_{q=1}^{n}||\overline{X_p} - \overline{X_q}||^2}{n^2}$$

سوال ۹.

با توجه به داده های زیر به سوالات پاسخ دهید:

C1	C2	C3	Class
F	F	F	A
T	F	T	B
T	T	F	B
T	T	T	A

۱) Simple Matching Coefficient و Jaccard Coefficient را برای داده های سطر اول و دوم بدست آورید
۲) Cosine و Bhattacharya را برای داده های موجود در سطر سوم و چهارم بدست آورید.
۳) correlation ستون های C1 و C2 را محاسبه کنید.
۴) آنتروپی را با توجه به برچسب های موجود بدست آورید.
۵) Pdfزیر را در نظر بگیرید. و برای متغیر تصادفی xوy مقدار $I(x,y)$ را محاسبه کنید.

$$p(x,y) = \begin{cases} e^{-(x+y)}, & x > 0, y < 0 \\ 0 & o.w \end{cases}$$

بخش عملی:

در این تمرین با اطلاعات بیش از ۴۰هزار شهر در قالب یک دیتافریم مواجه می‌شوید و با هدف تمرین در کار با داده و مفاهیم ابتدایی داده کاوی روی این دیتافریم تغییراتی ایجاد خواهید کرد.

برای انجام بخش عملی تمرین، نوت بوک موجود در فایل Practical را باز نمایید و طبق خواسته های هر بخش کد مربوط را در بلوک مختص به همان بخش بنویسید و اجرا کنید (از ذخیره بودن نوت بوک خود پیش از زیپ کردن پاسخ هایتان مطمئن شوید)

* در صورت وجود هرگونه ابهام میتوانید از طریق ایمیل این درس با تدریسیاران در ارتباط باشید.