



دانشکده مهندسی صنایع

گزارش تمرین اول مبانی داده‌کاوی و کاربردهای آن (۲۱۰۱۹)

استاد: دکتر مانا مس کار

دستیار آموزشی: محمدمهری منتظری هدش

اعضای گروه: صبا عبدی (۴۰۱۱۰۴۲۷۶)، آوا صدیقی (۴۰۱۱۱۰۱۵۹)

بخش ۱: شناخت انواع داده‌ها

۱. (۵ نمره) برای هر ستون، مشخص کنید که نوع داده آن چیست؟ (Nominal, Ordinal, Binary, Numeric)

CustomerID	Name	Age	Salary	Gender	Marital Status	Education	Loan Approval	Zip Code	Satisfaction Score
Nominal	Nominal	Numeric	Numeric	Binary	Nominal	Ordinal	Binary	Nominal	Ordinal

CustomerID: در این قسمت از اعداد استفاده شده است. اما نکته مهم این است که ترتیب یا معنای عددی ندارند و شناسایی برای شناسایی مشتریان است. بنابراین این attribute از نوع Nominal است.

Name: این قسمت نام مشتریان را شامل می‌شود و به همین علت از نوع Nominal است.

Age: این قسمت سن مشتریان را شامل می‌شود و به همین علت از نوع Numeric است.

Salary: این قسمت حقوق مشتریان را شامل می‌شود و به همین علت از نوع Numeric است.

Gender: این قسمت جنسیت مشتریان را نشان می‌دهد. در این ستون تنها دو مقدار male و female داریم. به همین علت از نوع binary است.

Marital Status: این قسمت وضعیت تأهل مشتریان را نشان می‌دهد و همچنین ترتیب خاصی ندارد. بنابراین از نوع Nominal است.

Education: این قسمت وضعیت تحصیلی مشتریان را نشان می‌دهد. به دلیل اینکه سطح تحصیلات به ترتیبی دنباله‌روی یکدیگر هستند، مثال پس از کارشناسی می‌توان کارشناسی ارشد را خواند، این ستون از نوع Ordinal است.

Loan Approval: این قسمت وضعیت وام درخواستی را نشان می‌دهد که تأیید یا رد شده است. به دلیل داشتن دو گزینه yes/no، این ستون از نوع binary است.

Zip Code: این قسمت، کدپستی مشتریان را نشان می‌دهد. کدپستی از اعداد تشکیل شده است اما به علت آنکه موقعیت مکانی را نشان داده و بار عددی ریاضی ندارد، آن را از نوع Nominal به شمار می‌آوریم.

Satisfaction Score: این قسمت میزان رضایت را نشان می‌دهد. به دلیل اینکه میزان رضایت از ۱ تا ۵ است و در دل خود ترتیبی جا داده است، این ستون را از نوع Ordinal به شمار می‌آوریم.

۲. (۵ نمره) بین متغیرهای Gender و Loan Approval، تفاوت در نوع داده‌ها را توضیح دهید.

متغیرهای Gender و Loan Approval هر دو از نوع binary به شمار می‌آیند اما از برخی جنبه‌ها، با یکدیگر تفاوت‌هایی دارند که به شرح آن‌ها می‌پردازیم. در ابتدا، وام می‌تواند قبول یا رد شود؛ یعنی برای ستون Loan Approval نمی‌توان حالت دیگری را تصور کرد. اما در بحث Gender می‌توان موارد متعددی را در نظر گرفت و به نوعی آن را از حالت Binary خارج کرد و به Nominal تبدیل کرد یا رویکرد دودویی را به ابعاد بالاتر برد؛ به عنوان مثال، اگر Non-binary اضافه شود، آن را با ۱۰۰ Male را با ۰۱۰ و Female را با ۰۰۱ نشان داد که همین رویکرد نیز مشکلاتی را با خود ایجاد خواهد کرد. در ادامه، ستون Gender یک ویژگی توضیحی است و یک متغیر مستقل در تحلیل‌ها به شمار می‌رود. اما در Loan Approval نتیجه یک تصمیم اعلام می‌شود و یک متغیر وابسته است که متناسب با ورودی‌ها نتیجه آن مشخص می‌شود. علاوه بر این، attribute جنسیت یک متغیر Loan Approval است. یعنی صفر و یک آن بار تحلیلی خاصی را برای ما به ارمغان نمی‌آورد. این در حالی است که در متغیر متناسب با تحلیل مدنظر، به هر یک از موارد yes و no مقدار یک را تخصیص داده و تحلیل خود را انجام می‌دهیم و

۳. (۵ نمره) آیا می‌توان از متغیر Education به عنوان داده عددی استفاده کرد؟ چرا؟

در سوال ۱ این بخش، نوع ستون Education را مشخص کردیم که Ordinal بود. این ترتیب سطوح مختلف تحصیلات را نشان می‌دهند ولی نمی‌توان آن‌ها را به داده عددی تبدیل و استفاده کرد. در ابتدا باید بدانیم که فاصله بین مقاطع تحصیل با یکدیگر یکسان نیست. یعنی لزوماً مقدار عددی Phd ۴ برابر مقطع High School نیست. بنابراین نمی‌توان آن‌ها را به عدد درستی تبدیل کرد. علاوه بر این، اگر فواصل مساوی را نیز بپذیریم، تحلیل‌های آماری بر روی این داده‌ها بی‌معنا خواهد بود. به عنوان مثال ما نمی‌دانیم میانگین ۳.۲۵ به چه معنی است و چگونه باید آن را تفسیر کنیم. بنابراین امکان تبدیل این مقادیر به داده عددی نیست. لازم به ذکر است که می‌توان متناسب با رتبه‌بندی موجود، به مقاطع تحصیلی عددی تخصیص دهیم که این عدد رنک مقاطع را نشان دهد و غیر از این کاربرد، تبدیل دیگری نمی‌توان داشت.

بخش ۲: توصیف آماری و نمایش داده‌ها

۱. (۳ نمره) برای متغیرهای `total_bill` و `tip`، مقادیر زیر را محاسبه کنید:
- میانگین، میانه، مد، دامنه، واریانس و انحراف معیار.

برای این قسمت از سوال از توابع کتابخانه `pandas` استفاده می‌کنیم. با استفاده از توابع اولیه چون `mean`, `var`, ... موارد خواسته شده را حساب می‌کنیم. حال اعداد به دست آمده برای هر ستون را تحلیل می‌کنیم.

	total_bill	tip
measures		
<code>mean</code>	19.785943	2.998361
<code>median</code>	17.795000	2.935000
<code>mode</code>	13.420000	1.820000
<code>range</code>	47.740000	4.580000
<code>variance</code>	79.252939	1.921810
<code>standard deviation</code>	8.902412	1.386293

ابتدا با ستون `total_bill` آغاز می‌کیم. میانگین این ستون برابر با ۱۹.۷۹، میانه این ستون برابر با ۱۷.۸ و مد این ستون برابر ۱۳.۴۲ است. مقادیر این سه معیار، از منظر مقدار و به ترتیب کاهشی، به شکل میانگین و میانه و مد هستند. میانگین نسبت به میانه کمی بیشتر است. به همین دلیل احتمال دارد که چند داده پرت با مقادیر بالا در این ستون وجود داشته باشند. همچنین به دلیل کمتر بودن مد از دو معیار دیگر، می‌توان گفت که این ستون چولگی مثبت دارد و داده‌ها در مقادیر کمتر تمرکز دارند. دامنه این ستون برابر با ۴۷.۷۴ است که نشان‌دهنده وجود پراکندگی زیاد در داده‌ها است. در نهایت، واریانس این ستون برابر ۷۹.۲۵ و انحراف معیار این ستون برابر ۸.۹ است. این اعداد، همانند دامنه، حاکی از پراکندگی زیادی در داده‌ها هستند؛ به نحوی که به طور میانگین داده‌ها چیزی در حدود ۹ دلار از میانگین فاصله دارند.

در نهایت با ستون `tips` ادامه می‌دهیم. میانگین این ستون برابر با ۳، میانه این ستون برابر با ۲.۹۴ و مد این ستون برابر با ۱.۸۲ است. در این اعداد تفاوت فاحشی، مثل ستون `total_bill` به چشم نمی‌خورد. میانه به میانگین نزدیک است که این نشان‌دهنده یک دست بودن داده، تا حدودی، است. همچنین مد نیز کمی از میانگین کمتر است. دامنه این ستون برابر با ۴.۵۸، واریانس این ستون برابر ۱.۹۲ و انحراف معیار این ستون برابر با ۱.۳۹ است. تمامی این اعداد حاکی از این هستند که داده‌های موجود در این ستون پراکندگی زیاد و فاحشی نداشته و در اطراف میانگین به حالت پایداری قرار گرفته‌اند.

در نهایت می‌توان نتیجه گرفت که در ستون total_bill شاهد پراکندگی زیاد و چولگی مثبت هستیم و در ستون tips داده‌ها مورد خاص و عجیبی ندارند.

۲. (۷ نمره) آیا توزیع متغیر tip دارای چولگی است؟ چگونه متوجه می‌شوید؟

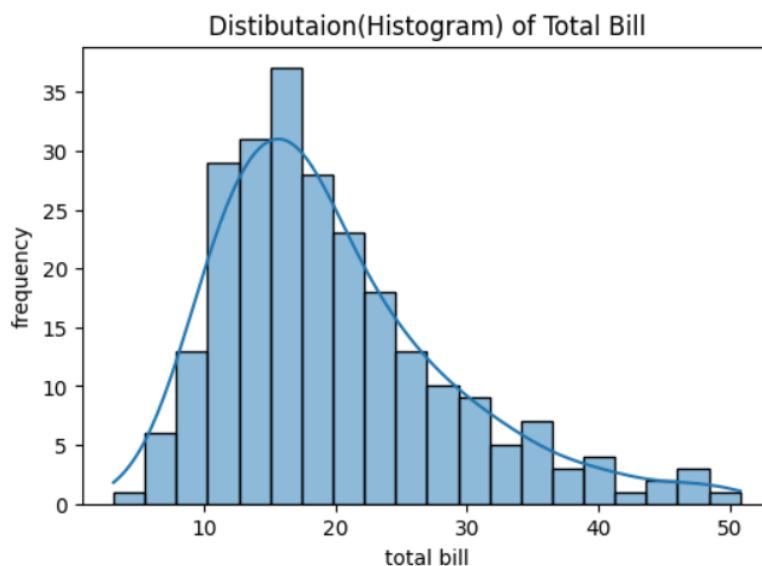
می‌دانیم چولگی زمانی اتفاق افتاده است که یکی از روابط میانگین > میانه > مد یا مد > میانگین برقرار باشد. در ستون tip حالت دوم، یعنی مد > میانگین، رخ داده است (به عکس جدول در سوال اول این بخش رجوع کنید). به دلیل اینکه این مقادیر به یکدیگر نزدیک هستند، از آزمون skewtest برای اطمینان بیشتر استفاده می‌کنیم. آماره آزمون برابر 0.29 است که از مقدار 0.05 بزرگتر است و فرض صفر، چولگی مثبت داشتن، را رد می‌کند.

```
Skewness stat: 1.0585242280120504
P-value: 0.28981651111233464
```

متناسب با دو نکته‌ای که ذکر شد، بهترین نتیجه این است که بگوییم در این ستون چولگی مثبت وجود دارد ولی بسیار ضعیف است.

۳. (۷ نمره) با استفاده از نمودار هیستوگرام، توزیع متغیر total_bill را نمایش دهید. آیا این متغیر دارای توزیع نرمال است؟

ابتدا با استفاده از کتابخانه‌های موجود، نمودار هیستوگرام را رسم می‌کنیم. از نمودار رسم شده می‌توان دریافت که از توزیع نرمال پیروی نمی‌کند.

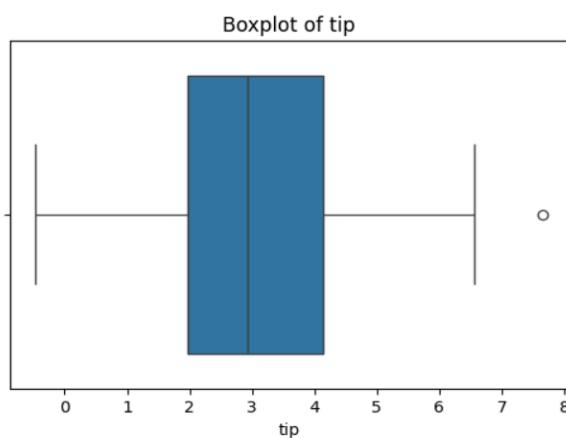


برای اطمینان و اعتبار بیشتر تحلیل خود از آزمون Shapiro استفاده می‌کنیم. این آمون برای تعیین نرمال بودن یا نبودن داده‌های کوچک تا متوسط، همانند دیتاست ما، استفاده می‌شود. p value این آزمون برابر با 3.33×10^{-5} است که از مقدار 0.05 بسیار بزرگتر است. بنابراین با قطعیت می‌توان گفت که این داده‌های این ستون از توزیع نرمال پیروی نمی‌کنند.

```
Shapiro stat: 0.9197187941346584
P-value: 3.3245391868090786e-10
```

۴. (۷ نمره) نمودار جعبه‌ای (Boxplot) متغیر tip را رسم کنید و نقاط پرت را مشخص کنید.

با استفاده از کتابخانه‌های `ggplot2`، نمودار جعبه‌ای ستون tip را رسم می‌کنیم.



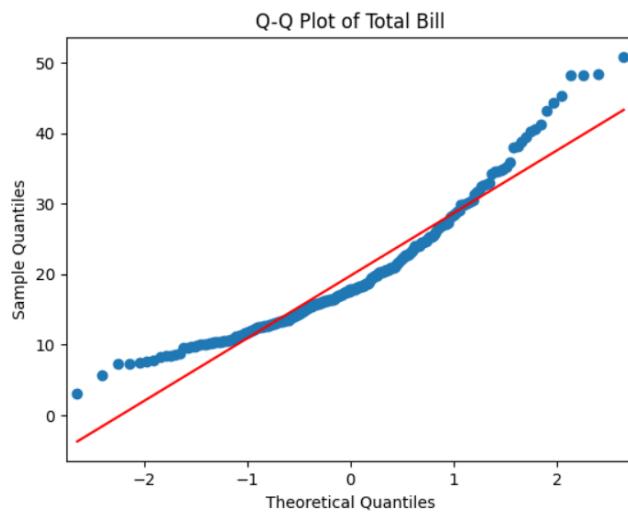
در شکل به دست آمده حدود بالا و پایین، چارک‌های اول و سوم و میانه مشخص هستند. طول جعبه نشان‌دهنده گستره میانی داده‌ها است و می‌توان بیان کرد که حجم خوبی از داده‌ها در این گستره قرار دارند. علاوه بر این چون میانه دقیقاً وسط جعبه نیست، داده‌های ما تقارن کامل ندارند. به دلیل وجود یک داده پرت در سمت راست، می‌توان گفت که داده‌ها تا حدودی چولگی مثبت ضعیفی دارند. این چولگی ضعیف را کمتر بودن میانه نسبت به میانگین نیز تأیید می‌کند (دقت کنید که این تفاوت زیاد نیست؛ بنابراین چولگی بسیار ضعیف است اگر وجود داشته باشد؛ همانگونه که در بخش‌های بالاتر توضیح داده شد).

در شکل مشهود است که تنها یک نقطه از حدود مشخص خارج است؛ این نقطه داده پرت ما محسوب می‌شود. در سلول بعدی کد با محاسبه IQR و حدود نمودار جعبه‌ای، داده پرت موردنظر را اسالیس می‌کنیم. این داده (۲۱امین داده (ایندکس گذاری از صفر است). مقدار tip در این سطر ۷.۶۵ واحد است و از میانگین، میانه و مد فاصله زیادی داشته و حتی از دامنه به دست آمده برای این ستون بیشتر است.

	total_bill	tip	sex	smoker	day	time	size
209	12.76	7.65	Female	Yes	Sat	Dinner	1

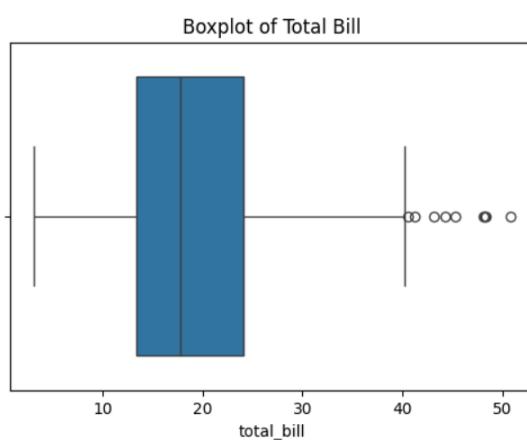
۵. (۹ نمره) نمودار Q-Q Plot برای متغیر total_bill رسم کنید و تحلیل کنید که آیا این متغیر از توزیع نرمال پیروی می‌کند؟

برای رسم Q-Q Plot موردنظر از کتابخانه statmodel استفاده می‌کنیم. لازم به ذکر است که مولفه 's line=' استفاده می‌شود که داده دیگر مورد استفاده در این نمودار از توزیع نرمال باشد. حال نمودار را تحلیل می‌کنیم. در ابتداء، نقاط رفته به خط نرمال نزدیک می‌شوند اما در ادامه از آن فاصله گرفته، دوباره به آن نزدیک شده و در انتهای مجدد به سمت چپ و دور از خط نرمال تمایل پیدا می‌کنند. این روند نشان‌دهنده این است که ستون total_bill از توزیع نرمال پیروی نمی‌کند. علاوه بر این متناسب با روند داده‌ها می‌توان بیان داشت که داده‌ها چولگی مثبت نیز دارند.



۶. (۷ نمره) با استفاده از معیار ۱.۵ برابر IQR، داده‌های پرت متغیر total_bill را شناسایی کنید.

ابتدا برای درک بهتر این ستون نمودار جعبه‌ای آن را رسم می‌کنیم.



در نمودار به دست آمده تعدادی داده پرت در سمت راست نمودار جعبه‌ای قابل مشاهده است. در سلول بعدی کد با استفاده از معیار ذکر شده در صورت سوال این داده‌های پرت را شناسایی می‌کنیم. ۹ داده به دست می‌آوریم که همگی از میانگین، میانه و دامنه بزرگتر هستند. لازم به ذکر است که این داده‌های پرت گواه بر چولگی مثبت این ستون هستند.

	total_bill	tip	sex	smoker	day	time	size
59	48.27	2.98	Male	No	Sat	Dinner	1
102	44.30	1.71	Female	Yes	Sat	Dinner	2
142	41.19	2.42	Male	No	Thur	Lunch	5
156	48.17	5.74	Male	No	Sun	Dinner	4
170	50.81	3.10	Male	Yes	Sat	Dinner	3
182	45.35	2.93	Male	Yes	Sun	Dinner	4
184	40.55	3.74	Male	Yes	Sun	Dinner	4
197	43.11	4.10	Female	Yes	Thur	Lunch	4
212	48.33	4.87	Male	No	Sat	Dinner	3

بخش ۳: مقایسه شباهت بین داده‌ها

دو فرد زیر را در نظر بگیرید:

Age	(kg) Weight	(cm) Height	PersonID
۵۸	۱۰۰.۲۳	۱۷۷	۹۶
۲۵	۱۰۲.۶۲	۱۹۵	۵

۱. (۲ نمره) فاصله اقلیدسی بین شماره ۹۶ و ۵ را به صورت دستی محاسبه کنید:

$$d(96, 5) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

$$d(96, 5) = \sqrt{(25 - 58)^2 + (102.62 - 100.23)^2 + (195 - 177)^2} = 37.6658$$

۲. (۵ نمره) همان فاصله اقلیدسی را مجدداً محاسبه کنید، اما این بار بعد از استانداردسازی داده‌ها با Z-score. آیا نتیجه تغییر کرد؟ چرا؟

PersonID	Age	Weight (kg)	Height (cm)
96	1.197358	0.737807	0.059573
5	-1.029971	0.855912	1.212602

$$d(96, 5) = \sqrt{(-1.030 - 1.197)^2 + (0.856 - 0.738)^2 + (1.213 - 0.060)^2} = 2.51$$

بله، نتیجه تغییر کرد. مقدار فاصله کاهش یافت چراکه همه مقادیر به عددی بزرگتر از ۱ تقسیم شده‌اند. می‌توان گفت اگر به طور میانگین انحراف معیار استاندارد پارامترهای مختلف بیشتر از ۱ باشد، فاصله اقلیدسی میان داده‌ها پس از استاندارد سازی کاهش خواهد یافت. حال بررسی می‌کنیم که این کمیت چه چیزی به ما می‌گوید:

$$\begin{aligned} d(X, Y) &= \sqrt{\sum_i^n (X_i - Y_i)^2}, \quad X_i, Y_i \sim N(\cdot, 1), \quad X_i - Y_i \sim N(\cdot, 2) \\ \frac{X_i - Y_i}{\sqrt{2}} &\sim N(\cdot, 1), \quad \sum_i^n 2(X_i - Y_i)^2 \sim \chi_{n-1}^2 \\ E[d(X, Y)] &= \sqrt{n} \cdot \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \approx \sqrt{2n} \end{aligned}$$

می‌بینیم که در صورت استانداردسازی داده‌ها، انتظار می‌رود فاصله اقلیدسی میان دو ردیف داده عددی در حدود $\sqrt{2n} = 2.45$ به دست بیاید که به مقدار به دست آمده نسبتاً نزدیک است و این محاسبات را تایید می‌کند.

۳. (۱۰ نمره) چه زمانی باید داده‌ها را استاندارد کنیم و چه زمانی نیاز نیست؟

۱. زمانی که میانگین ستون‌های مختلف داده‌ها اختلاف معناداری با یکدیگر دارند، نیاز است که داده‌ها را استاندارد کنیم تا معنای فاصله اقلیدسی حفظ شود. تصور کنید داده‌هایی به فرم زیر داریم:

- قد: (cm) بین ۱۵۰ تا ۲۰۰

- وزن: (kg) بین ۵۰ تا ۱۰۰

- درآمد (\$): بین ۱۰,۰۰۰ تا ۱,۰۰۰,۰۰۰

در این حالت اگر داده‌ها را استاندارد نکنیم بدون آنکه بخواهیم، وزن بیشتری برای پارامتر درآمد در نظر گرفته‌ایم! در این شرایط بهتر است که داده‌ها را استاندارد سازی کنیم. در حالتی که میانگین‌ها به یکدیگر نزدیک‌اند، اگر داده‌ها را استاندارد سازی نکنیم ممکن است به درک بهتری از فواصل دست یابیم.

۲. در استفاده از الگوریتم‌های حساس به مقیاس نیز بهتر است استانداردسازی صورت بگیرد. برای مثال در کار با الگوریتم‌هایی مانند Linear Regression، K-Means و K-Nearest Neighbors، بهتر است استاندارد سازی صورت بگیرد. چراکه این الگوریتم‌ها از فاصله اقلیدسی یا انحراف معیار استاندارد استفاده می‌کنند، پس ویژگی‌هایی با مقیاس بزرگ‌تر، مدل را تحت تأثیر قرار می‌دهند.

۳. اگر داده‌ها توزیع نرمال داشته باشند نیز استاندارد سازی مشکلی برای داده‌ها ایجاد نمی‌کند و اثر آن به نحوی است که انگار داده‌ها از یک نرمال به نرمال دیگر منتقل کرده‌ایم و فواصل همچنان در هر بعد متناسب خواهند بود اما وقتی داده‌ها توزیع نرمال ندارند استاندارد سازی بی‌معنا و غیرمجاز است چراکه از اساس آن داده‌ها از توزیع نرمال پیروی نمی‌کردند. در این حالت اعمال این تغییر به روی آن‌ها می‌تواند حتی موجب تغییر توزیع داده‌ها شود بنابر این در این موارد بهتر است استاندارد سازی انجام نشود یا لاقل ابتدا داده‌ها نرمال شوند!!

۴. اگر پارامترها دارای معنای فیزیکی مهمی باشند که برایمان اهمیت دارد و در نتیجه ابعاد فاصله را نیز معنادار می‌کند، استانداردسازی نه تنها کمکی نمی‌کند که پیچیدگی‌های کار را دوچندان خواهد کرد.

استانداردسازی؟

موقعیت	استانداردسازی؟
ویژگی‌ها مقیاس‌های متفاوتی دارند.	<input checked="" type="checkbox"/>
از الگوریتم‌های حساس به فاصله استفاده می‌کنیم.	<input checked="" type="checkbox"/>
داده‌ها تقریباً نرمال‌اند.	<input checked="" type="checkbox"/>
ویژگی‌ها در یک بازه‌ی مشخص هستند یا نرمال نیستند.	<input checked="" type="checkbox"/>
ویژگی‌ها ذاتاً دارای معنای فیزیکی مهمی هستند.	<input checked="" type="checkbox"/>
استفاده از مدل‌هایی مثل درخت تصمیم یا رندوم فارست	<input checked="" type="checkbox"/>

۴. (۳ نمره) محاسبه شباهت کسینوسی (Cosine Similarity) بین این دو فرد را به صورت دستی انجام دهید.

دو فرد زیر را در نظر بگیرید:

Type Body	Color Hair	Color Eye	PersonID
Average	Black	Blue	۹۶
Average	Black	Green	۵

با توجه به آنکه ویژگی‌های مورد بررسی در این بخش، nominal هستند، در ابتدا باید آن‌ها به نحوی به ویژگی‌های binary تبدیل کرد. برای این تبدیل روش‌های متعددی موجود است، در ادامه از روش One Hot Encoding استفاده می‌کنیم تا در نهایت یک آرایه binary برای هر یک از record‌ها تعریف کنیم.

این روش را مختصرًا شرح می‌دهیم و سپس به حل مسئله می‌پردازیم. تصور کنید که ۲ ویژگی داریم که هر یک شامل ۲ و ۴ حالت می‌باشند. آرایه‌ای که برای هر یک از داده‌ها در این سیستم تعریف می‌شود به شرح زیر خواهد بود:

اگر ویژگی اول حالت اول را داشته باشد: [۱, ۰]

اگر ویژگی اول حالت دوم را داشته باشد: [۰, ۱]

اگر ویژگی دوم حالت اول را داشته باشد: [۰, ۰, ۰, ۰]

اگر ویژگی دوم حالت دوم را داشته باشد: [۰, ۱, ۰, ۰]

اگر ویژگی دوم حالت سوم را داشته باشد: [۰, ۰, ۱, ۰]

اگر ویژگی دوم حالت چهارم را داشته باشد: [۰, ۰, ۰, ۱]

حال اگر بخواهیم آرایه شامل ویژگی اول حالت اول و ویژگی دوم حالت چهارم را تشکیل دهیم، به حالت زیر می‌رسیم:

[۱, ۰, ۰, ۰, ۰, ۰, ۱]

با همین رویکرد سطرهای ۵ و ۹۶ را به آرایه تبدیل می‌کنیم.

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \|d_2\|}$$

$$d_1 = [1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0], \quad \|d_1\| = \sqrt{3}$$

$$d_2 = [0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0], \quad \|d_2\| = \sqrt{3}$$

$$d_1 \cdot d_2 = 1 + 1 = 2$$

$$\cos = \frac{2}{3} \rightarrow \cos^{-1}\left(\frac{2}{3}\right) = 48.19^\circ$$

۵. (۵ نمره) فاصله جاکارد را برای ویژگی های رنگ چشم، رنگ مو و نوع بدن بین شماره ۹۶ و ۵ محاسبه کنید.

برای محاسبه فاصله جاکارد ابتدا باید تعداد مواردی که در هر دو سطر یکسان یا برابر ۱ هستند را حساب کنیم. برای دو سطر مورد نظر دو ویژگی یکسان داریم؛ یعنی $q = 2$. سپس باید مواردی که در هر دو سطر با یکدیگر متفاوت هستند را شناسایی کنیم، یعنی $s = 1$ و $r = 1$

		Object <i>j</i>		
		1	0	sum
Object <i>i</i>	1	<i>q</i>	<i>r</i>	<i>q+r</i>
	0	<i>s</i>	<i>t</i>	<i>s+t</i>
sum		<i>q+s</i>	<i>r+t</i>	<i>p</i>

$$sim_{jaccard}(96, 5) = \frac{q}{r + s + q} = \frac{2}{2 + 1 + 1} = \frac{2}{4} = 0.5$$

$$diss_{jaccard}(96, 5) = 1 - sim_{jaccard}(96, 5) = 1 - 0.5 = 0.5$$

لذا فاصله جاکارد این دو سطر برابر با ۰.۵ می باشد.

۱. (۲۰ نمره) بررسی کنید که آیا رابطه‌ای بین جنسیت و نوع بدن وجود دارد؟

- یک جدول توافقی (Contingency Table) بین دو متغیر Gender و Body Type بسازید.
- آزمون کای-دو (Chi-Square Test) را روی این داده‌ها انجام دهید.
- مقدار p-value را تفسیر کنید: آیا این دو متغیر وابسته هستند یا مستقل؟

ابتدا contingency table را تشکیل می‌دهیم. جدول به دست آمده به شرح زیر است.

Gender	Female	Male	Total
Body Type			
Athletic	13	12	25
Average	16	11	27
Overweight	11	12	23
Slim	11	14	25
Total	51	49	100

در ادامه لازم است که آزمون کای-دو با استفاده از داده‌های این جدول پیاده‌سازی شود. نتایج این آزمون به شرح زیر است:

```
▶ chi2_stat, p_value, dof, expected = chi2_contingency(contingency_table)
print(f"Chi-Square Statistic: {chi2_stat}")
print(f"P-Value: {p_value}")
print(f"Degrees of Freedom: {dof}")
print("Expected Frequencies Table:")
print(pd.DataFrame(expected, index=contingency_table.index, columns=contingency_table.columns))
```

```
→ Chi-Square Statistic: 1.3299361612599956
P-Value: 0.9951847612442957
Degrees of Freedom: 8
Expected Frequencies Table:
   Gender    Female   Male  Total
   Body Type
Athletic      12.75  12.25  25.0
Average       13.77  13.23  27.0
Overweight     11.73  11.27  23.0
Slim          12.75  12.25  25.0
Total         51.00  49.00 100.0
```

p-value به دست آمده برابر 0.995 است که بسیار از سطح معناداری بزرگتر است؛ لذا فرض صفر، یعنی وابسته بودن این دو متغیر رد شده و این دو متغیر مستقل از یکدیگر هستند و هیچ ارتباط معناداری میان gender و body type وجود ندارد.