فرهاد امان ۹۹۳۱۰۰۶

۱-
دیتاست: این داده‌ها شامل اطلاعاتی مانند مواد موثره، دوزها، نحوه‌ی تولید، تاریخ انقضاء، هزینه تولید، فروش‌ها، بازخوردها از مشتریان و نظرات متخصصان پزشکی می‌شوند.

مساله: بهبود کارایی داروها و پیش‌بینی فروش بهتر

شرکت قصد دارد با استفاده از داده‌های خود، کارایی داروهای خود را ارزیابی کرده و بهبودهای لازم را اعمال کند. این شامل بهینه‌سازی فرمولاسیون داروها، تغییر در دوزها، و یا تغییر در روش‌های تولید می‌شود. همچنین، شرکت می‌خواهد بتواند فروش داروها را پیش‌بینی کرده و استراتژی‌های بازاریابی بهتری ایجاد کند.

با استفاده از داده‌های جمع‌آوری شده از مواد موثره، دوزها، تاریخ‌های انقضاء، هزینه تولید، فروش‌های گذشته، و نظرات مشتریان، می‌توان الگوریتم‌های داده کاوی را برای پیش‌بینی فروش، ارزیابی کارایی داروها و پیشنهاد بهبودها به کار برد. این کارها می‌تواند به شرکت کمک کند تا بهبودهای لازم در محصولات خود ایجاد کرده و در نهایت فروش و سودآوری شرکت را افزایش دهد

گام ۱: تعریف مساله و اهداف
در این مرحله، ابتدا باید با تیم بیزینس شرکت تعامل کرد تا مساله‌ی داده کاوی مشخص شود. این مساله می‌تواند شامل پیش‌بینی، دسته‌بندی، یافتن الگوها، یا هر کاربرد دیگری از داده کاوی باشد. همچنین، اهداف کلی مشخص می‌شوند تا بفهمیم چه اطلاعاتی از داده‌ها باید استخراج شود.

گام ۲: جمع‌آوری داده‌ها
در این مرحله، داده‌های مورد نیاز جمع‌آوری می‌شوند. این داده‌ها ممکن است از منابع مختلفی مانند پایگاه داده‌ها، فایل‌های متنی، یا حتی از اینترنت آمده باشند. باید دقت شود که داده‌های جمع‌آوری شده کامل و مرتبط با مساله مورد نظر باشند.

گام ۳: پیش‌پردازش داده‌ها
پس از جمع‌آوری داده‌ها، نیاز است تا داده‌ها پیش‌پردازش شوند. این شامل حذف داده‌های ناقص، تبدیل داده‌های متنی به اعداد، حذف داده‌های تکراری و نویزهای داده می‌شود. همچنین، می‌توانید ویژگی‌های مهم را انتخاب کنید و داده‌ها را مقیاس دهید.

گام ۴: انتخاب و اجرای الگوریتم‌های داده کاوی
بعد از پیش‌پردازش، الگوریتم‌های داده کاوی انتخاب می‌شوند و روی داده‌ها اجرا می‌شوند. این الگوریتم‌ها ممکن است شامل خوشه‌بندی، یادگیری ماشین، یا هر الگوریتم دیگری باشند که بر اساس نیازهای مساله انتخاب می‌شوند.

گام ۵: ارزیابی مدل‌ها
پس از اجرای الگوریتم‌ها، نیاز است تا مدل‌های حاصله ارزیابی شوند. این ارزیابی ممکن است با استفاده از معیارهایی مانند دقت، صحت، فراخوانی، و یا معیارهای دیگر انجام شود تا بفهمیم مدل‌ها به چه اندازه عملکرد خوبی دارند.

گام ۶: تفسیر نتایج و گزارش‌گیری
در این گام، نتایج حاصله تفسیر می‌شوند و به کارفرما گزارش داده می‌شود. این گزارش باید شامل توضیحاتی در مورد مساله، روش‌های استفاده شده، نتایج حاصله و نتیجه‌گیری‌های به دست آمده باشد. همچنین، پیشنهاداتی برای اقدامات آینده نیز در این گزارش قرار می‌گیرند.

گام ۷: استقرار مدل
اگر مدل‌ها و نتایج کاوش داده مورد تایید قرار گرفتند، مدل‌ها در محیط تولیدی یا سیستم مرتبط استقرار می‌یابند. این مرحله شامل اجرای مداوم مدل در داده‌های واقعی و مانیتورینگ عملکرد مدل‌هاست.

۲-
Noise : نویز به اختلاف‌ها و تغییرات تصادفی در داده‌های اندازه‌گیری اشاره دارد که ناشی از خطاها و عدم دقت در فرآیند اندازه‌گیری می‌باشند. نویز می‌تواند شامل انحراف و تغییر مقدار و یا افزودن اشیاء بی‌ارتباط به داده‌ها باشد.

Outlier : داده‌ای که به نوعی ویژگی‌های متفاوتی از اکثر دیگر داده‌ها در مجموعه داده دارد. به عبارت دیگر داده‌ای که به طریقی نمایانگر ویژگی‌های منحصر به فرد یا نادری است که از بقیه داده‌ها متمایز می‌شود.

۱) در برخی موارد مطالعه داده‌های پرت برای ما سودمند است. برای مثال:
کشف تقلب: در اطلاعات مالی و بانکداری، داده‌های پرت ممکن است نمایانگر تراکنش‌های مالی ناهنجار و تقلبی باشند. برای مثال، اگر یک تراکنش مالی بسیار بزرگ یا غیرعادی وجود داشته باشد، ممکن است به عنوان یک داده‌پرت معتبر در شناسایی تقلب استفاده شود.

شناسایی نفوذ در شبکه: در حوزه امنیت شبکه، داده‌های پرت ممکن است نمایانگر فعالیت‌های نفوذی به شبکه باشند. برنامه‌ها و الگوریتم‌های خاصی در شبکه‌های کامپیوتری به دنبال داده‌های پرتی می‌گردند که نمایانگر حملات یا فعالیت‌های مشابه‌اند.

۲) مدل‌هایی مانند رگرسیون خطی به داده‌های پرت حساس‌اند و در آموزششان تداخل ایجاد می‌شود و دقت مدل کاهش می‌یابد.
نویزها نیز در آموزش شدن مدل تاثیر منفی می‌توانند داشته باشند و منجر به تحلیل و آنالیز غلط داده‌ها شوند. برای مثال ممکن است در یک مساله، سیگنال‌های نویز دار نیاز به یک پیش پردازش برای کاهش نویز داشته باشند و این بار پردازشی و محاسباتی دارد.

۳-
۱)
Nominal: کد ملی، شماره پرسنلی کارکنان این داده‌ها در اکثر مواقع صرفا برای تشخیص متفاوت بودن دو داده استفاده می‌شوند و برای یکتایی داده به کار می‌روند و روی آنها تحلیلی صورت نمی‌گیرد.
Ordinal: کیفیت و مدرک تحصیلی این مقادیر برای مقایسه و ترتیب‌دهی می‌توانند به کار روند.
Interval: تاریخ‌های تقویم، دمای هوا به سانتی‌گراد و فارنهایت. این مقادیر عددی اند و قابلیت جمع و تفریق دارند ولی ضرب و تقسیم را ندارند چون صفر حقیقی ندارند.
Ratio: دمای هوا به کلوین، سن. این مقادیر صفر حقیقی دارند و قابلیت جمع، تفریق، ضرب و تقسیم نیز دارند.

۲) برای شماره دانشجویی مثلا گرفتن غذا از سلف طوری که به هر دانشجو یک غذا برسد و فقط کسانی که رزرو کرده‌اند غذا دریافت کنند.
برای جنسیت نیز می‌توان مثلا برای سیستم هوشمند گرفتن برنامه غذایی نمی‌توان این فاکتور را حذف کرد زیرا زن و مرد وزن و قد و به طور کلی ویژگی‌های فیزیکی متفاوتی دارند.

۳)
میانگین : Ratio , Interval
میانه : Ordinal (البته برای Interval و Ratio نیز می‌توان تعیین کرد)
مد : Nominal

۴-
۱) One-hot encoding: این روش برای نمایش دادن متغیرهای Categorical به شکل عددی است. هر مقدار ممکن برای متغیر مورد نظر را به عنوان یک ستون جدید به ویژگی‌های داده اضافه می‌کنیم. در صورتی که داده آن مقدار را داشته باشد در ستون مورد نظر 1 و در بقیه ستون‌ها 0 قرار می‌دهیم.

Label encoding: این روش نیز برای نمایش دادن متغیر های Categorical به شکل عددی است. به ازای هر مقدار ممکن برای متغیر Categorical یک عدد تعیین می‌کنیم و به جای مقادیر غیر عددی دسته بندی استفاده می‌کنیم.

۲) هر کتگوری منحصر به فرد در فیچرهای Color و Shape به یک ستون جدید تبدیل می‌شوند. در نتیجه حداقل به ۹ ستون نیاز است.

ID	Blue	Red	Green	Size	Weight	Triangle	Round	Square
1	1	0	0	10	0.5	1	0	0
2	0	1	0	8	0.3	1	0	0
3	1	0	0	8	0.7	0	1	0
4	0	0	1	12	0.3	0	0	1

۳) از آنجایی که فیچرهای Color ,Shape از نوع Ordinal نبوده و برتری میان کلاس‌های آن‌ها وجود ندارد در این جدول روش One Hot مناسب‌تر است زیرا در روش Label Encoding مدل آموزش دیده می‌تواند به غلط این رابطه را یاد بگیرد که ارزش کلاس Square نسبت به Triangle بیشتر است زیرا مقدار نسبت داده شده بیشتری دارد.
در حالت کلی روش One Hot Encoding وقتی داده‌های Categorical زیادی در اختیار داریم و هرکدام شامل تعدادی زیادی کلاس منحصر به فرد است، مناسب نمی‌باشد زیرا این روش می‌تواند ابعاد جدول را بسیار بزرگ کند و در نتیجه آموزش مدل را کندتر و پیچیده‌تر می‌کند.

۵- تمام بردارهای مجموعه داده ما روی خطی در راستای v هستند. بنابراین باید بردار یکه آن را محاسبه کنیم.

$$u = \frac{v}{|v|_2}$$

$$v.u = |v|_2 = 17.46$$

$$2v.u = 34.92$$

$$-2v.u = -34.92$$

-۶

$$cov(\beta_1, \beta_2) = cov(\beta_1, \overline{y} - \beta_1 \overline{x}) = cov(\beta_1, \overline{y}) - cov(\beta_1, \beta_1 \overline{x})$$

همچنین می‌دانیم که $\overline{y}$ عددی ثابت است در نتیجه:

$$\beta_1 = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^n (x_i - \overline{x})^2} = \frac{\sum_i^n (x_i - \overline{x}) y_i - \sum_i^n (x_i - \overline{x})\overline{y}}{\sum_i^n (x_i - \overline{x})^2} = \frac{\sum_i^n (x_i - \overline{x}) y_i}{\sum_i^n (x_i - \overline{x})^2}$$

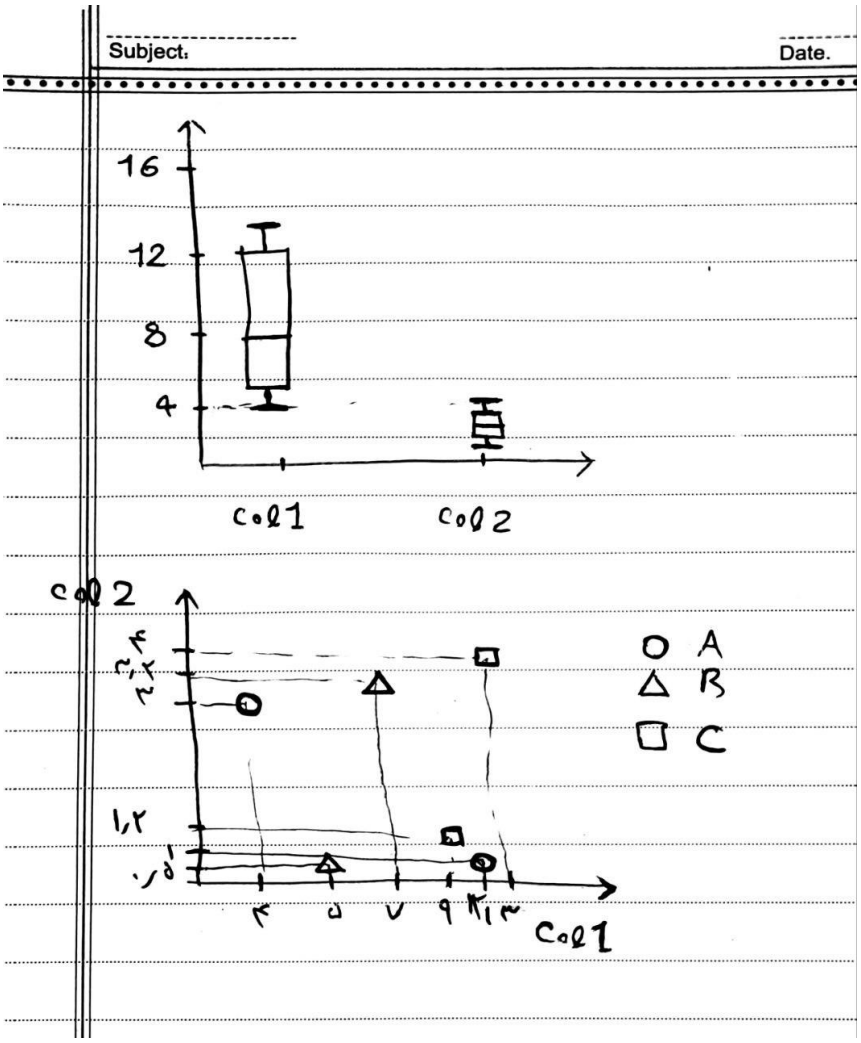
$$\Rightarrow - \; x \; \frac{\sum_i^n var(y_i)}{\sum_i^n (x_i - \overline{x})^2} = cov(\beta_1, \; \beta_2)$$

اگر دو متغیر مستقل باشند، به این معنی که هیچ ارتباط خطی بین آن‌ها وجود نداشته باشد، ضریب همبستگی (covariance) بین آن‌ها برابر با صفر خواهد بود. این بدان معناست که تغییرات در یکی از متغیرها هیچ تأثیری بر تغییرات متغیر دیگر ندارد و بالعکس. به عبارت دیگر، هیچ رابطه خطی بین این دو متغیر وجود ندارد و هیچ‌گونه پیش‌بینی یا تبادل اطلاعاتی بین آن‌ها امکان‌پذیر نیست.

-۷

نمودار جعبه‌ای: یکی از روش‌های متداول برای نمایش توزیع داده‌ها است. این نوع نمودار به شما اجازه می‌دهد تا اطلاعاتی در مورد میانه، کوچکترین و بزرگترین مقادیر، وجود نقاط پرت و انحراف معیار داده‌ها بدست آورید. در اینجا باید داده‌ها را مرتب کنید، سپس کوچکترین مقدار، پنجاه درصدی (میانه)، و بزرگترین مقدار را نشان دهید. همچنین، نقاط پرت نیز نشان داده می‌شوند و این نمودار به شما ایده‌ای از توزیع داده‌ها می‌دهد.

ماتریس پراکندگی: ماتریس پراکندگی یک روش برای نمایش روابط بین دو یا بیشتر از ویژگی‌های یک مجموعه داده است. این نوع نمودار اجازه می‌دهد تا هر ویژگی در محورهای مختصات قرار گیرد و نقاط داده‌ها بر اساس این ویژگی‌ها روی نمودارها نشان داده شوند. این روش به شما امکان می‌دهد تا الگوها و روابط میان داده‌ها را بررسی کنید.



-۸

$$X = \{x_1, x_2, \dots, x_n\} \Rightarrow \frac{\sum_i^n x_i}{n} = 0$$

$$\frac{\sum_p (x_i . x_i + x_p . x_p - 2x_i . x_p)}{n} = \frac{n(x_i . x_i) + (x_1 . x_1 + \dots + x_n . x_n)}{n}$$

$$(x_i . x_i) + \frac{(x_1 . x_1 + \dots + x_n . x_n)}{n}$$

$$(x_j . x_j) + \frac{(x_1 . x_1 + \dots + x_n . x_n)}{n}$$

$$\frac{\sum_j n(x_j \,.\, x_j) + (x_1 \,.\, x_1 + \ldots + x_n \,.\, x_n)}{n^2} = \frac{2(x_1 \,.\, x_1 + \ldots + x_n \,.\, x_n)}{n}$$

$$(x_i \,.\, x_i) + \frac{(x_1 \,.\, x_1 + \ldots + x_n \,.\, x_n)}{n} + (x_j \,.\, x_j) + \frac{(x_1 \,.\, x_1 + \ldots + x_n \,.\, x_n)}{n} - \frac{2(x_1 \,.\, x_1 + \ldots + x_n \,.\, x_n)}{n}$$

$$x_i \,.\, x_i + x_j \,.\, x_j = |x_i|^2 + |x_j|^2$$

۹-

۱) می‌دانیم که F = 0 است و T = 1

$$SMC = \frac{f_{00} + f_{11}}{f_{tot}} = \frac{1}{3}$$

$$JC = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = 0$$

۲) طبق داده‌های سطر سوم و چهارم می‌دانیم که $R_3 = (1,\ 1,\ 0)\ ,\ R_4 = (1,\ 1,\ 1)$

سپس آن‌ها را نرمال می‌کنیم.

$$R_3 = (0.5,\ 0.5,\ 0)\ ,\ R_4 = (\frac{1}{3},\ \frac{1}{3},\ \frac{1}{3})$$

$$|R_3| = \sqrt{2}\ ,\ |R_4| = \sqrt{3}$$

در نهایت با توجه به فرمول هرکدام را محاسبه می‌کنیم.

$$cosine(R_3,\ R_4) = \frac{(1 * 1) + (1 * 1) + (1 * 0)}{\sqrt{2} \times \sqrt{3}} = \frac{\sqrt{6}}{3}$$

$$D_B = \ -\ ln(\frac{1}{\sqrt{6}} + \frac{1}{\sqrt{6}} + 0) = \ -\ ln(\frac{2}{\sqrt{6}}) = 0.2$$

۳)

$$\overline{c_1} = \frac{3}{4}\ ,\ \overline{c_2} = \frac{1}{2}$$

$$cov(c_1,\ c_2) = \frac{1}{6} \ \Rightarrow\ correlation(c_1,\ c_2) = \frac{\frac{1}{6}}{\frac{1}{2} * \frac{1}{\sqrt{3}}} = \frac{\sqrt{3}}{3}$$

(۴

$$H(x) = -\sum_{x \epsilon X} P(x) \,.\, lg(P(x)) \,,\; p = 0.5$$

$$-\, 2\, lg(\tfrac{1}{2}) = 2\, lg(2) = 2$$

(۵

$$P(x) = \int_{y=0}^{\infty} e^{-(x+y)} dy = e^{-x} \int_{y=0}^{\infty} e^{-y} dy = e^{-x}(-\, e^{-y})|^{\infty} = -\, e^{-x}$$

$$P(y) = -\, e^{-y}$$

$$I(x,\, y) = \int\int P(x,\, y)\, log(\tfrac{P(x,y)}{P(x)P(y)}) dx\, dy = \int\int e^{-(x+y)}\, log(\tfrac{e^{-(x+y)}}{e^{-x}\, e^{-y}})$$

$\int\int e^{-(x+y)}\, log(1) = 0 \Rightarrow x, y$ مستقل