 یبلیبلبلا

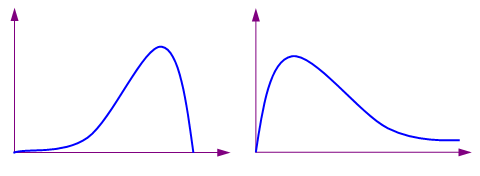
نام و نام خانوادگي: وزارت علوم و تحقیقات و فناوری عنوان امتحان: داده کاوی شماره دانشجويي: تاريخ امتحان: 23/8/97

نام استاد: ملكوتي مدت امتحان: 120دقیقه

1. مراحل فرایند کشف دانش از مجموعه پایگاه داده های متفرق را بصورت شکل ترسیم کنید.
2. بنویسید موارد زیر در کدام دسته از انواع صفات قرار می گیرند:

* نام کالا
* وزن
* رنگ
* درجه کیفیت (متوسط، خوب، عالی)
* موجود بودن
* حداکثر درجه حرارت محیط برای نگهداری (سانتی گراد)

1. در نمودارهای زیر که نشان دهنده توزیع دو گروه مختلف از داده هاست محل تقریبی میانگین، میانه و مد را مشخص کنید:



1. برای داده های زیر boxplot ترسیم کنید. آیا در این داده ها، داده پرت وجود دارد؟

**1, 2, 3, 5, 6, 9, 10, 11, 12, 13, 28**

1. فرض کنید چهار صفت A و B و C و D ، صفات دودویی نامتقارن کالا باشند. فاصله دو کالای داده شده در جدول زیر را محاسبه کنید:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| نام کالا | A | B | C | D |
| کالای 1 | Y | N | Y | Y |
| کالای 2 | N | N | Y | Y |

1. ماتریس فاصله برای سه دانشجوی زیر در درس ریاضی را محاسبه کنید. فرض کنید نمرات میان ترم در محدوده 9 تا 18 بوده است. حداقل و حداکثر تعداد تمرین های تحویل شده از طرف دانشجویان 2 و 10 می باشد. پروژه اختیاری بوده و فقط تحویل یا عدم تحویل آن ثبت می شود. محدوده معدل کل دانشجو توسط یکی از حروف A,B,C,D نشان داده می شود.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | رشته | نمره میان ترم | تعداد تمرین تحویل شده | تحویل پروژه | معدل کل |
| دانشجوی 1 | مهندسی کامپیوتر | 12 | 4 | Y | C |
| دانشجوی 2 | مهندسی برق | 15 | 2 | Y | B |
| دانشجوی 3 | مهندسی کامپیوتر | 18 | 8 | N | A |

1. در مرحله پاکسازی داده برای حل مشکل داده های مفقود(Null) چه راهکارهایی وجود دارد؟
2. در یک نظرسنجی بین یک گروه از دانشجویان نتایج زیر مشاهده شده است:

|  |  |  |  |
| --- | --- | --- | --- |
|  | مرد | زن | جمع |
| موافق | 50 | 80 | 130 |
| مخالف | 150 | 20 | 170 |
| جمع | 200 | 100 | 300 |

* مقادیر مورد انتظار را برای هر یک از سلول ها بدست بیاورید.
* با فرض اینکه برای این جدول حدوث با درجه آزادی یک و سطح معنی داری 0.001 حداقل مقدار کای دو برای وابستگی صفات 10.828 باشد، وابستگی صفات رای و جنسیت را محاسبه و بررسی کنید.

1. یک بنگاه تجاری که کالاهایی را از تولید کنندگان خریده و به فروشندگان جزء می فروشد، قصد دارد مکعب داده هایی حول دو موضوع میزان خرید و میزان فروش (هر دو به ریال) تولید کند. ابعاد مرتبط با موضوع خرید تولید کننده، زمان، کالا و ابعاد مرتبط با موضوع فروش زمان، کالا و خریدار (فروشگاه های خرده فروشی) می باشد. شمای مناسب برای این انبار داده را طراحی کنید.

موفق باشید.

cos(*d1, d2*) = (*d1* • *d2*) /||*d1*|| ||*d2*||

