

سوالات استخدای : داده کاوی

۱- فاصله مانها تان و نمونه $X1=(1,2)$ و $X2=(3,5)$ برابر است با:

۱. ۵ ۲. ۳ ۳. ۶۱ ۴. ۲

۲- کدام گزینه فرایند کشف دانش را نشان می دهد؟

۱. پالایش داده ها، انتخاب داده ها، داده کاوی، ارائه دانش
۲. پالایش داده ها، یکپارچه سازی داده ها، انتخاب داده ها، تبدیل داده ها، داده کاوی، ارزیابی الگو، ارائه دانش
۳. پالایش داده ها، یکپارچه سازی داده ها، انتخاب داده ها، داده کاوی، کاهش الگوهای زائد، ارائه دانش
۴. جمع آوری، انتخاب داده ها، داده کاوی، پیش پردازش داده ها، ارائه دانش

۳- کدام گزینه صحیح است؟

۱. در داده های با چولگی منفی، مقدار میانگین کمتر از مقدار میانه است.
۲. مقدار میانگین همیشه بین میانه و مد است.
۳. میانه مقداری یکتا است.
۴. IQR به تفاضل چارک اول و دم اشاره دارد.

۴- نمودار جعبه ای به ترتیب چه مقادیر را نشان می دهد؟

۱. چارک اول، چارک دوم، چارک سوم
۲. میانه، میانگین، مد
۳. کمترین ، میانه، میانگین، مد، بزرگترین
۴. کمترین ، چارک اول، میانه، چارک سوم، بزرگترین

۵- با توجه به جداول رابطه زیر، عدم تشابه دو بیمار برابر است با:

	Fever	Cough	Tes-1	Test-2	Test-3	Test-4
Jack	Y	N	P	N	N	N
Jim	Y	N	N	N	N	N

۱. ۰/۳۳ ۲. ۰/۶۷ ۳. ۰/۷۵ ۴. ۱

۶- در خصوص مقادیر ناموجود، کدام استراتژی های استفاده می شود؟

- الف) چشم پوشی از تاپل
ب) استفاده از یک مقدار ثابت سراسری
ج) استفاده از شاخص میانگین
د) استفاده از مقدار واریانس

۱. همه موارد ۲. الف ، ب، ج، د ۳. الف، ج، د، ه ۴. ج، د، ه

۷- با استفاده از روش هموار سازی با مقدار کرانی بسته...

۱. ابتدا کرانها مشخص و سپس مقادیر بسته به نزدیکترین مقدار کران جایگزین می شود.

۲. هر مقدار در بسته با مقدار میانگین بسته جایگزین می شود.

۳. کوچکترین و بزرگترین مقدار مشخص و هر بسته با نزدیکترین مقدار به کوچکترین یا بزرگترین مقدار جایگزین می شود.

۴. هر مقدار در بسته با مقدار میانه بسته جایگزین می شود.

۸- اگر مقادیر یک صفت خاصه ی A در بازه ی ۹۸۶- تا ۹۱۷ باشد، مقدار ۹۰۰ با استفاده از مقیاس گذاری دهدی به چه مقداری تبدیل می شود؟

۱. ۰/۹ ۲. ۱۷ ۳. ۹۰ ۴. ۰/۰۹

۹- کدام طراحی پایگاه داده در OLAP استفاده نمی شود؟

۱. مبتنی بر ER ۲. ستاره ای ۳. برفگونه (دانه برفی) ۴. موضوع گرا

۱۰- داده هایی در مورد داده هستند.

۱. انبار داده ۲. متا دیتا ۳. دیتا مارت ۴. مکعب داده ها

۱۱- شامل یک جدول بزرگ مرکزی و مجموعه ای از جداول که هر یک به یک بعد اشاره می کنند.

۱. شمای ستاره ای ۲. شمای برفگونه ۳. شمای صور فلکی ۴. دیتا مارت

۱۲- داده ها را به صورت مفصل تر و جزئی تر نشان می دهد؟

۱. roll-up ۲. drill-down ۳. Slice ۴. Pivot

۱۳- یک مجموعه اقلام X در مجموعه داده های D است، چنانچه هیچ یک از مجموعه اقلام بزرگتر از X مانند Y، دارای پشتیبان برابر X نباشد.

۱. مکرر ۲. کاندید ۳. بسته ۴. ماکسیمال

۱۴- با توجه به جدول مقابل، مقدار اطمینان قانون مقابل برابر است با:

$\{I1, I2\} \Rightarrow I5$

TID	List of item-ID
۱	I1, I2, I5
۲	I2, I4
۳	I2, I3
۴	I1, I2, I4
۵	I1, I3
۶	I2, I3
۷	I1, I3
۸	I1, I2, I3, I5
۹	I1, I2, I3

۱. ۳۳%

۲. ۱۰۰%

۳. ۲۹%

۴. ۵۰%

۱۵- با توجه به جدول مقابل، مقدار پشتیبان قانون مقابل برابر است با:

$\{I1, I5\} \Rightarrow I2$

TID	List of item-ID
۱	I1, I2, I5
۲	I2, I4
۳	I2, I3
۴	I1, I2, I4
۵	I1, I3
۶	I2, I3
۷	I1, I3
۸	I1, I2, I3, I5
۹	I1, I2, I3

۱. ۴

۲. ۷

۳. ۲

۴. ۶

۱۶- روش مشکل یافتن الگوهای مکرر طولانی را به جستجوی الگوهای کوتاه در یک پایگاه داده کوچکتر تبدیل می کند.

۱. Fp-growth ۲. Apriori ۳. Eclat ۴. Hash Table

۱۷- قلم داده دوتایی {I3, I2} در قالب عموده داده ها، با توجه به تراکنشهای مقابل به چه صورت است؟

TID	List of item-ID
۱	I1, I2, I5
۲	I2, I4
۳	I2, I3
۴	I1, I2, I4
۵	I1, I3
۶	I2, I3
۷	I1, I3
۸	I1, I2, I3, I5
۹	I1, I2, I3

۱. {۳, ۶, ۸, ۹} ۲. {۴} ۳. {۸, ۱} ۴. {۹, ۸, ۴, ۱}

۱۸- مقدار lift قانون زیر با توجه به جدول حدوث ارائه شده ، برابر است با:

$$buys(X, "game") \Rightarrow buys(X, "Videos")$$

	game	<u>game</u>
video	۴۰۰۰	۳۵۰۰
<u>video</u>	۲۰۰۰	۵۰۰

۱. ۰/۴ ۲. ۰/۸۹ ۳. ۰/۶ ۴. ۰/۷۵

۱۹- محدودیتهای شامل تعیین آستانه برای سنجه های جالب بودن نظیر پشتیبان، اطمینان و همبستگی می باشد.

۱. مربوط به نوع دانش ۲. مربوط به بعدیا سطح ۳. مربوط به داده ها ۴. مربوط به جالب بودن

۲۰- در کدام تکنیک برچسب کلاس ها مشخص نیست؟

۱. خوشه بندی ۲. دسته بندی بیز ۳. بردار ماشین ۴. نزدیکترین همسایه

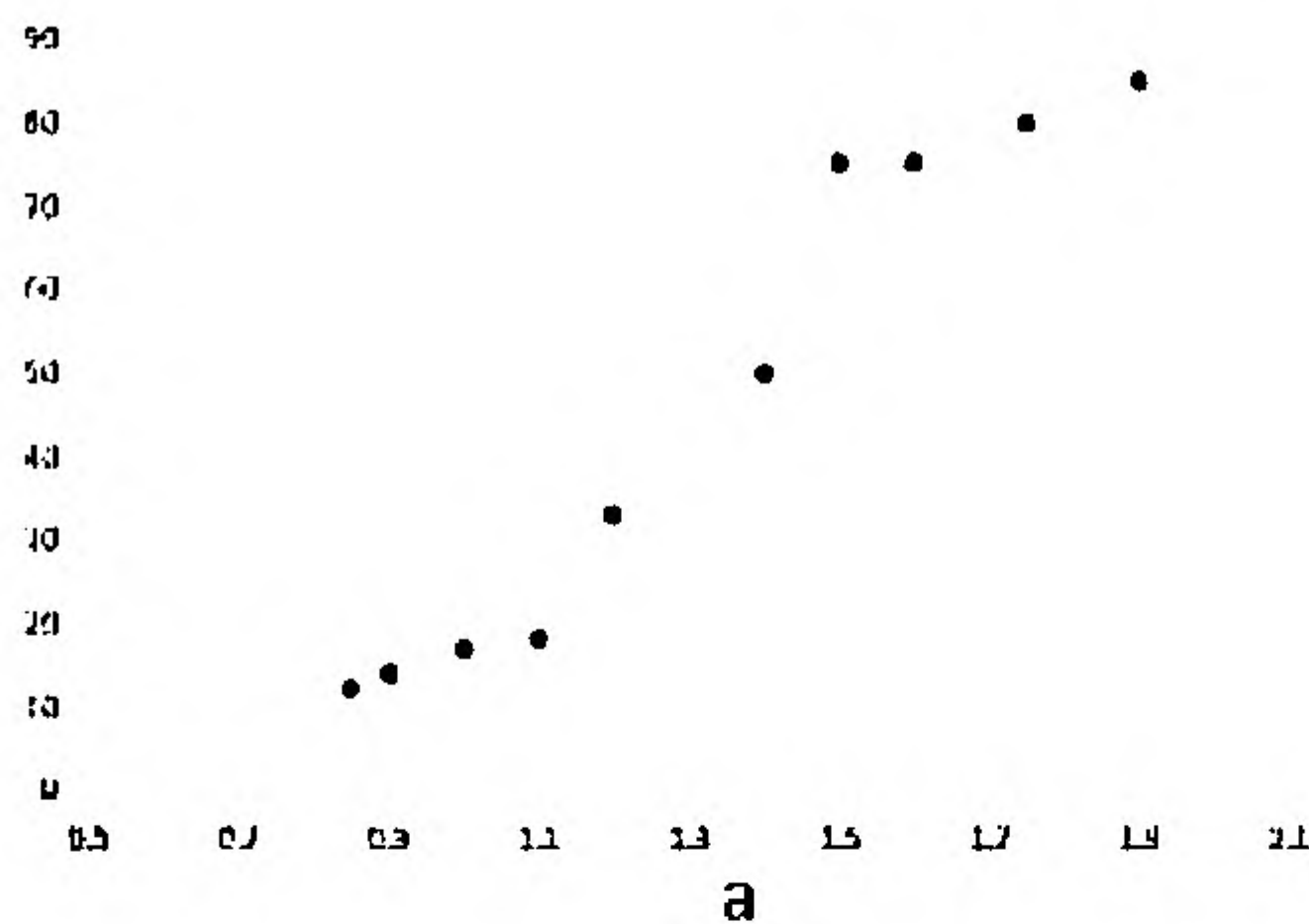
۲۱- صفت خاصه ای "شغل" چه نوعی است؟

۱. نسبی Ratio ۲. بازه ای Interval ۳. ترتیبی Ordinal ۴. اسمی Nominal

۲۲- چارک سوم به چه اشاره می کند؟

۱. نقطه ای که ۲۵ درصد از داده های ابتدایی را جدا می کند.
۲. نقطه ای که ۵۰ درصد از داده های ابتدایی را جدا می کند.
۳. نقطه ای که ۳۰ درصد از داده های ابتدایی را جدا می کند.
۴. نقطه ای که ۷۵ درصد از داده های ابتدایی را جدا می کند.

۲۳- نمودار مقابل چه چیزی را بین مقادیر دو متغیر نشان میدهد؟



۱. همبستگی مثبت بین X و Y

۲. همبستگی منفی بین دو متغیر

۳. عدم همبستگی

۴. همبستگی مثبت تا نقطه a و همبستگی منفی از a به بعد

۲۴- در جدول زیر با توجه به اینکه مقادیر test-2 به صورت ترتیبی است، فاصله اقلیدسی دو رکورد شماره ۲ و ۴ برابر است با:

Test-2	ID
Excellent	۱
Fair	۲
good	۳
Excellent	۴

۱. ۱

۲. ۰/۵

۳. ۰/۲۵

۴. ۰

۲۵- فرض کنید در یک دیتاست، افراد زیر ۱۰ سال، در دیتاست کم باشند. اگر بخواهیم در نمونه گیری حتما از افراد زیر ۱۰ سال هم در نمونه باشد، کدام روش نمونه گیری مناسب است؟

۱. نمونه گیری تصادفی ساده بدون جایگزینی

۲. نمونه گیری تصادفی ساده با جایگزینی

۳. نمونه گیری خوشه ای

۴. نمونه گیری طبقه ای

سوالات تشریحی

۱- با توجه به جدول حدوث زیر، مقدار کای اسکوئر را به دست آورید.

	male	female	total
fiction	۲۵۰	۲۰۰	۴۵۰
non-fiction	۵۰	۱۰۰۰	۱۰۵۰
total	۳۰۰	۱۲۰۰	۱۵۰۰

۱/۵۰ نمره

۲- با استفاده از الگوریتم Apriori مجموعه اقلام دوتایی مکرر را در دیتاست مقابل پیدا کنید؟ ۱/۵۰ نمره

TID	List of item-ID
۱	I1, I2, I5
۲	I2, I4
۳	I2, I3
۴	I1, I2, I4
۵	I1, I3
۶	I2, I3
۷	I1, I3
۸	I1, I2, I3, I5
۹	I1, I2, I3

۳- در خصوص قوانین انجمنی چند سطحی، چه روشهایی جهت پشتیبان برای سطوح تعریف می شود؟ ۱/۵۰ نمره

۴- نرمال سازی min-max را با ذکر فرمول و مثال توضیح دهید؟ ۱/۵۰ نمره

شماره سوال	پاسخ صحيح
1	الف
2	ب
3	الف
4	د
5	ب
6	ب
7	الف
8	الف
9	الف
10	ب
11	الف
12	ب
13	ج
14	د
15	ج
16	الف
17	الف
18	ب
19	د
20	الف
21	د
22	د
23	الف
24	الف
25	د

سوالات تشریحی

۱- ۵۰۷/۹۳

۲- فصل ۶

۳- فصل ۷

۴- فصل ۲

۱- کدام یک از مراحل فرایند کسب دانش جزو مراحل پیش پردازش داده ها محسوب نمی شود؟

۱. تبدیل داده ها
۲. یکپارچه سازی داده ها
۳. انتخاب داده ها
۴. ارزیابی الگو

۲- کدام یک از گزینه های زیر در دسته بندی داده ها مورد استفاده قرار نمی گیرد؟

۱. درخت تصمیم
۲. شبکه عصبی
۳. خوشه بندی
۴. قوانین IF-Then

۳- کلاه برداری در استفاده از کارت های اعتباری نمونه ای از مسائلی است که عموماً کدام یک از روش های زیر به تحلیل و کشف آن می پردازد؟

۱. تحلیل داده های پرت
۲. رگرسیون چند جمله ای
۳. رگرسیون غیرخطی
۴. تحلیل سبد خرید

۴- کدام یک از موارد زیر جزو چالش های مطرح شده از سوی موتورهای جستجو برای داده کاوی نیست؟

۱. کنترل مقدار حجیم داده ها توسط موتورهای جستجو
۲. بهنگام سازی افزایشی
۳. روبرو بودن موتورهای جستجو با پرسش هایی که به دفعات زیاد توسط هر کاربر پرسیده می شود
۴. مشارکت هزاران ابرکامپیوتر برای کاوش در میان داده ها

۵- در کدام یک از انواع مجموعه داده ها مقادیر میانه، میانگین و مد با هم منطبق بوده و دارای مقادیر یکسانی هستند؟

۱. در مجموعه داده های متقارن
۲. در مجموعه داده ای با چولگی مثبت
۳. در مجموعه داده های با چولگی منفی
۴. در مجموعه داده های نامتقارن

۶- کدام یک از سنجه های شاخص مرکزی زیر برای داده ای اریب (نامتقارن) سنجه بهتری محسوب می شود؟

۱. مد
۲. میانه
۳. میانگین
۴. انحراف معیار

۷- در کدام یک از حالات زیر در یک نمودار پراکنشی، داده ها همبستگی دارند؟

۱. نقاط بی هیچ نظمی در فضای دوبعدی پخش شده باشند.
۲. نقاط در حول و حوش خطی موازی محور افقی قرار داشته باشند.
۳. نقاط در حول و حوش خطی به صورت قطری در فضای دوبعدی قرار داشته باشند.
۴. نقاط در حول و حوش خطی موازی محور عمودی قرار داشته باشند.

۸- این تکنیک مصورسازی می تواند بالغ بر ۱۸ بعد را نشان داده و نمونه نامتقارن آن در جهت تعمیم تکنیک اولیه قادر به نمایش ۳۶ بعد می باشد.

۱. صورتک های چرنف
۲. تکنیک قطعه بندی دایره
۳. تکنیک دنیاهاى تودرتو
۴. ابر برچسب

۹- کدام یک از گزینه های زیر در مورد خصوصیات ریاضی فاصله اقلیدسی و فاصله مانهاتان صحیح نیست؟

۱. برقرار بودن نامعادله مثلثی بین فواصل
۲. نامتقارن بودن تابع فاصله
۳. هویت غیرقابل تشخیص
۴. غیرمنفی بودن فاصله

۱۰- از کدام یک از روش های زیر برای تشخیص همبستگی بین داده های عددی استفاده می شود؟

۱. آزمون خی دو (چی اسکوار تست)
۲. تحلیل کوواریانس
۳. تحلیل مولفه ای اصلی
۴. تبدیل موجک

۱۱- هار دو و دبوخی چهار مثال هایی از کدام یک از روش های کاهش داده محسوب می شوند؟

۱. نمونه برداری
۲. تحلیل مولفه های اصلی
۳. تبدیل موجک
۴. هیستوگرام

۱۲- در کدام یک از روش های نمونه برداری اگر ۱۰۰ داده داشته باشیم که در ده بسته ده تایی قرار داده شده باشند به عنوان مثال قادر به انتخاب سه بسته کامل از بین آنها خواهیم بود؟

۱. نمونه برداری خوشه ای
۲. نمونه برداری طبقه ای
۳. نمونه برداری تصادفی ساده با جایگزینی
۴. نمونه برداری تصادفی ساده بدون جایگزینی

۱۳- کدام یک از انواع نمودارهای زیر در طراحی یک سیستم OLAP مورد استفاده قرار نمی گیرد؟

۱. نمودارهای ER
۲. نمودارهای ستاره ای
۳. نمودارهای برفگونه
۴. نمودارهای صور فلکی

۱۴- در کدام یک از نمودارهای زیر از چند جدول حقیقت (Fact Table) استفاده می شود؟

۱. صورفلکی
۲. ستاره ای
۳. برفگونه
۴. ER

۱۵- کدام یک از توابع زیر را می توان یک تابع جبری محسوب نمود؟

۱. mode()
۲. count()
۳. sum()
۴. avg()

۱۶- کدام یک از عملگرهای OLAP زیر یک عملگر دیداری و بصری سازی است؟

۱. pivot
۲. dice
۳. slice
۴. roll-up

۱۷- کدام یک از گزینه های زیر در مورد الگوریتم star-cubing صحیح نیست؟

۱. این الگوریتم در محاسبه مکعب کوه یخی به ترتیب ابعاد حساس است
۲. در زمانی که با داده های اریب سروکار داشته باشد سریع تر از BUC عمل می کند
۳. افزایش حداقل پشتیبان باعث افزایش سرعت این الگوریتم می شود
۴. در محاسبه مکعب کامل در صورتی که داده ها متراکم باشند عملکرد الگوریتم قابل مقایسه با MultiWay است

۱۸- در کاوش اکتشافی و مبتنی بر استثنا فضای مکعب، کدام یک از سنجه های زیر درجه شگفتی را برای هر مسیر drill down از سلول نشان داده و در شناسایی آنومالی ها استفاده می شود؟

۱. SelfExp ۲. PathExp ۳. InExp ۴. OutExp

۱۹- کدام یک از رویکردهای زیر برای محاسبه مکعب کامل یک رویکرد معمولی در MOLAP بوده و از آدرس دهی مستقیم آرایه استفاده می کند؟

۱. MultiWay ۲. Star-Cubing ۳. Shell Fragment ۴. BUC

۲۰- کدام یک از گزینه های زیر در خصوص الگوریتم FP-Growth صحیح نیست؟

۱. این روش برای الگوهای مکرر کوتاه مقیاس پذیر است و برای یافتن الگوهای مکرر بلند، موثر نیست.
۲. این روش سریع تر از روش آپریوری عمل می کند.
۳. در پایگاه های داده بزرگ برای اعمال این الگوریتم لازم است که پایگاه داده در ابتدا افراز شود.
۴. در این روش، یافتن الگوهای مکرر طولانی به جستجوی الگوهای کوتاه تر در یک پایگاه داده کوچکتر تبدیل می شود.

۲۱- فرمتی که هر رکورد شامل یک قلم داده ای بوده و با مجموعه از شناسه های تراکنش ها همراه می شود اساس کار کدام یک از الگوریتم های زیر است؟

۱. Eclat ۲. FP-Growth ۳. Apriori ۴. BUC

۲۲- کدام یک از سنجه های زیر دارای خصوصیت null-variance نیست؟

۱. Lift ۲. Max-Confidence ۳. All-Confidence ۴. Kulczynski

۲۳- در میان سنجه های مورد استفاده در کاوش مشارکت ها استفاده از کدام یک از سنجه های زیر به همراه نسبت عدم توازن بهترین پاسخ را تولید می کند؟

۱. Lift ۲. Kulczynski ۳. All-Confidence ۴. Max-Confidence

۲۴- کدام یک از الگوهای زیر جزو الگوهای پایه در زمینه کاوش الگو محسوب نمی شود؟

۱. الگوهای مکرر
۲. قوانین انجمنی
۳. الگوهای بسته یا ماکسیمال
۴. الگوهای پیوسته مبتنی بر گسسته سازی

۲۵- گراف های جهت دار، گراف های بدون جهت، شبکه ها، درخت ها و توالی ها در کاوش چه نوع الگوهایی بررسی می شوند؟

۱. کاوش الگوهای ساختاری
۲. کاوش الگوهای تک بعدی
۳. کاوش الگوهای چندبعدی
۴. کاوش الگوهای تک سطحی

سوالات تشریحی

- ۱- نمودار جعبه ای و روش تشخیص داده های پرت با استفاده از این نمودار را مختصرا شرح دهید
۱.۲۰ نمره
- ۲- دو روش از میان روش های نرمال سازی داده ها به منظور تبدیل داده ها را نام برده و با مثال شرح دهید.
۱.۲۰ نمره
- ۳- سیستم های پردازش تراکنشی برخط (OLTP) و سیستم های پردازش تحلیلی برخط (OLAP) را از چهارمنظر با یکدیگر مقایسه کنید.
۱.۲۰ نمره
- ۴- روش های بهبود الگوریتم آپریوری را نام برده و دو مورد را شرح دهید.
۱.۲۰ نمره
- ۵- روش BUC برای محاسبه مکعب های کوه یخی از شبه مکعب نوک به سمت پایین را به اختصار شرح دهید.
۱.۲۰ نمره

شماره سوال	پاسخ صحيح
1	د
2	ج
3	الف
4	ج
5	الف
6	ب
7	ج
8	الف
9	ب
10	ب
11	ج
12	الف
13	الف
14	الف
15	د
16	الف
17	ج
18	ب
19	الف
20	الف
21	الف
22	الف
23	ب
24	د
25	الف

سوالات تشریحی

۱- فصل 2

۲- فصل 3

۳- فصل 4

۴- فصل 6

۵- فصل 5

۱- کدام یک جز مراحل کشف دانش نمی باشد؟

۱. پاکسازی داده ۲. تجمیع داده ۳. ارزیابی الگو ۴. تقسیم داده

۲- در جدول مقابل $Confidence(I1 \Rightarrow I2, I5)$ برابر است با:

TID	Items
1	I1, I2, I5
2	I2, I4
3	I2, I3
4	I1, I2, I3
5	I2, I4
6	I2, I3
7	I1, I3
8	I1, I2, I3, I5
9	I1, I2, I3, I4, I5

۱. 0.33 ۲. 0.6 ۳. 0.4 ۴. 0.5

۳- از scatter plot برای چه کاری استفاده می شود؟

۱. بصری سازی ۲. پاکسازی داده ۳. تجمیع داده ۴. تقسیم داده

۴- از کدام روش جهت تخمین مشابه بودن اسناد (Document) استفاده می شود؟

۱. Cosine ۲. Edclidean ۳. Minkowski ۴. Supermum

۵- کدام گزینه روش برخورد با missing value نمی باشد؟

۱. نادیده گرفتن ۲. پرکردن به صورت دستی ۳. استفاده از میانگین داده ۴. استفاده از جذر داده

۶- data about data مربوط به کدام گزینه می باشد؟

۱. Clean Data ۲. Meta data ۳. Hyper data ۴. Data warehouse

۷- کدام گزینه جهت کاهش ابعاد داده کاربرد دارد

۱. PCA ۲. k-mean ۳. cosine ۴. dwt

۸- در داده های نرمال:

۱. 95 درصد داده ها در بازه "میانگین منهای اندازه انحراف معیار" و "میانگین بعلاوه انحراف معیار" قرار دارند
۲. 68 درصد داده ها در بازه "میانگین منهای اندازه انحراف معیار" و "میانگین بعلاوه انحراف معیار" قرار دارند
۳. 99.7 درصد داده ها در بازه "میانگین منهای اندازه انحراف معیار" و "میانگین بعلاوه انحراف معیار" قرار دارند
۴. 100 درصد داده ها در بازه "میانگین منهای اندازه انحراف معیار" و "میانگین بعلاوه انحراف معیار" قرار دارند

۹- کدام گزینه جز استراتژی Transformation یا انتقال در آماده سازی داده نمی باشد؟

۱. Smoothing
۲. Aggreagation
۳. Normalization
۴. Anomaly Detection

۱۰- کدام گزینه جز روش های نرمال سازی داده نمی باشد؟

۱. Std
۲. z-score
۳. min-max
۴. decimal scaling

۱۱- کدام روش بین ER و مدل کاربردی بانک اطلاعاتی ارتباط برقرار می کند؟

۱. OLAP
۲. OLTP
۳. MOLAP
۴. ROLAP

۱۲- منبع Meta data شامل کدام موارد نمی باشد؟

۱. توضیح درباره ساختار ware house
۲. الگوریتم های خلاصه سازی
۳. داده های مرتبط با بازدهی سیستم
۴. سیستم عامل مورد استفاده

۱۳- اجازه می دهد که داده ها را مدل کرده و در چندین بعد نمایش داده شود.

۱. data warehouse
۲. data cube
۳. meta data
۴. olap

۱۴- از اعمال رایج olap کدام گزینه نمی باشد؟

۱. Roll-up
۲. Min-max normal
۳. drill-down
۴. Pivot

۱۵- در داده های 10,12,13,15,17,25,30 مقدار Five Number Summer برابر است با:

۱. 10,12,15,25,30
۲. 10,13,15,17,25
۳. 10,13,17,25,30
۴. 12,15,25,30,30

۱۶- اگر کمترین و بیشترین مقادیر تعداد از اعداد به ترتیب 5 و 20 باشد و بخواهیم این اعداد را تبدیل به اعدادی در بازه 0 تا 10 نماییم، عدد 6 در تبدیل برابر خواهد بود با:

۱. 0.66
۲. 1.84
۳. 0.8
۴. 1.2

۱۷- در داده های 10,12,13,15,17,25,30 مقدار IQR برابر است با:

۱. 25 ۲. 12 ۳. 15 ۴. 13

۱۸- کدام الگوریتم جهت کشف الگوهای پرتکرار در مجموعه داده ها می باشد

۱. k-mean ۲. apriori ۳. knn ۴. two step

۱۹- در چولگی مثبت، کدام مقدار از همه کمتر است؟

۱. میانگین ۲. میانه ۳. چارک اول ۴. مد

۲۰- نمونه گیری از بخشی از داده ها را می نامند

۱. partitioning ۲. sampling ۳. selection ۴. metadata

۲۱- کدام گزینه قابلیت اطمینان الگو را در قوانین انجمنی بیان می کند

۱. support ۲. confidence ۳. Lift ۴. RUC

۲۲- فرمول زیر مربوط به کدام گزینه می باشد؟

$$\frac{P(A \cap B)}{P(A) \cdot P(B)}$$

۱. Support ۲. Confidence ۳. RUC ۴. Lift

۲۳- کدام الگوریتم جهت کشف الگوهای پرتکرار بدون کاندید می باشد؟

۱. Fp-growth ۲. k-mean ۳. apriori ۴. Knn

۲۴- بیشتر در فیلتر کردن نویز و پاکسازی داده استفاده می شود؟

۱. frequent mining ۲. pattern mining ۳. classification ۴. clustering

۲۵- می خواهیم میزان تمایل افراد به محصول A و B را بررسی نماییم. با توجه به جدول آماری زیر، مقدار کای اسکوئر برابر است با:

	B+	B-
A+	5	15
A-	9	10

۴. 1.758

۳. 3.12

۲. 1.69

۱. 2.19

سوالات تشریحی

- ۱- مراحل کشف دانش را با استفاده از داده کاوی همراه با ترسیم شکل توضیح دهید
۱۰۲۰ نمره
- ۲- انواع ویژگی های داده ها را با ذکر مثال نام ببرید
۱۰۲۰ نمره
- ۳- مراحل و روشهای پاکسازی داده ها (cleaning data) را توضیح دهید؟
۱۰۲۰ نمره
- ۴- تفاوت بین operational database system و data warehouse چیست؟
۱۰۲۰ نمره
- ۵- مراحل الگوریتم Apriori را تشریح کنید؟
۱۰۲۰ نمره

نمبر سوال	ياسخ صحيح
1	د
2	ب
3	الف
4	الف
5	د
6	ب
7	الف
8	ب
9	د
10	الف
11	ب
12	د
13	ب
14	ب
15	الف
16	الف
17	د
18	ب
19	د
20	ب
21	ب
22	د
23	الف
24	ب
25	الف

سوالات تشریحی

۱.۲۰ نمره

۱- فصل 1 صفحه ۷

۱.۲۰ نمره

۲- فصل ۲ صفحه 40 الی 43

۱.۲۰ نمره

۳- فصل ۳ صفحه 88 الی ۹۱

۱.۲۰ نمره

۴- فصل ۴ صفحه 128

۱.۲۰ نمره

۵- فصل ۶ صفحه 248

۱- کدام گزینه از فرآیند کشف دانش (Knowledge Discovery) نمی باشد؟

۱. یکپارچگی داده ها ۲. پالایش داده ها ۳. داده کاوی ۴. نگاشت داده

۲- پایگاه داده رابطه ای از مجموعه ای تشکیل شده است که اسامی یکتایی دارند.

۱. جداول ۲. رابطه ها ۳. رکوردها ۴. اطلاعات

۳- کدام گزینه تعریف کاملتری از انبار داده (Data Warehouse) را ارائه میدهد؟

۱. مخزنی برای نگهدای دانش در یک مکان می باشد.
۲. مخزنی که از اطلاعات جمع آوری شده از پایگاه داده های توزیع شده برای تولید دانش نگهداری می کند.
۳. مخزنی از اطلاعاتی است که از چندین منبع جمع آوری شده اند و تحت شمای یکدستی و معمولا در یک سایت نگهداری و ذخیره میشوند.
۴. محلی برای نگهداری منطقی رابطه های یک پایگاه داده می باشد.

۴- مجموعه اشیاء را شکل می دهند و یک شی یک را نشان می دهد.

۱. مجموعه داده (Data Set)، کلاس (Class)
۲. مجموعه داده ها (Data Set)، موجودیت (Entity)
۳. کلاس (Class)، موجودیت (Entity)
۴. موجودیت (Entity)، کلاس (Class)

۵- مقادیر یک صفت خاصه ی اسمی نمادها یا اسامی است که هر مقدار آن نیز به طبقه، کد یا حالتی اشاره دارد.

۱. اشیاء ۲. کلاس ۳. موجودیت ۴. مخزن

۶- کدام گزینه جزو انواع صفت خاصه نمی باشد؟

۱. اسمی (Nominal) ۲. دودویی (Binary) ۳. عددی (Numeric) ۴. رسمی (Official)

۷- آمار توصیفی پایه، اساس تحلیل برای محسوب می شود.

۱. پیش پردازش آماری ۲. پیش پردازش داده ها ۳. پیش پردازش اشیاء ۴. پیش پردازش رابطه ها

۸- تکنیک مصور سازی داده (Data Visualization) برای چه نوع داده هایی مورد استفاده قرار میگیرد؟

۱. داده های رابطه ای چند بعدی ۲. داده های دودویی
۳. داده های منطقی ۴. داده های ساختار دهی شده

۹- کدام گزینه بیان کننده تعریف داده کاوی (Data Mining) می باشد.

۱. فرآیند کشف الگوهای مورد توجه کاربر از داده های حجیم می باشد.

۲. فرآیند پردازش داده ها می باشد.

۳. فرآیند تشخیص انواع مختلف داده ها می باشد.

۴. فرآیند است بازگشتی برای شناسایی اطلاعات موجود در جداول پایگاه داده

۱۰- کدام گزینه از اجزای توصیف کیفیت داده ها نمی باشد؟

۱. کامل بودن

۲. صحت و درستی

۳. انعطاف پذیری

۴. قابلیت تفسیر

۱۱- یکپارچه ساز داده (Data Integration) چه وظیفه ای بر عهده دارد؟

۱. ترکیب داده ها از چندین منبع برای فرم دهی منسجم مخزن داده ای

۲. تفسیر داده های جمع آوری شده

۳. تبدیل داده ها به یک شکل مناسب

۴. اصلاح داده های ناقص

۱۲- روش های تبدیل داده ها (Data Transformation) چه کاربردی دارند؟

۱. داده ها را با یکدیگر ترکیب می کند.

۲. داده ها را به شکل های مناسبی برای کاوش تبدیل می کنند .

۳. داده ها را تفکیک می کند.

۴. داده ها را استخراج می کند.

۱۳- متا داده ی انبار داده ها، داده هایی هستند که انبار داده را تعریف می کنند.

۱. جداول

۲. اشیاء

۳. خصیصه ها

۴. پایگاه داده

۱۴- هسته ی مرکزی مدل چند بعدی می باشد.

۱. مکعب داده ها

۲. اشیاء

۳. تحلیل تفصیلی

۴. جداول

۱۵- پردازش تحلیلی برخط (Online analytical processing) از استفاده می کند.

۱. مدل مفهومی آماری

۲. مدل خصیصه

۳. مدل داده های چند بعدی

۴. مدل توصیفی

۱۶- سرور **HOLAP** کدام سرورها را با یکدیگر ترکیب می کند؟

۱. MOLAP, OLAP

۲. MOLAP, OLAP

۳. OLAP, POLAP

۴. OLAP, POLAP

۱۷- مکعب داده ها از شبکه ای از تشکیل شده است.

۱. ستون ها ۲. شبه مکعب ها ۳. جداول ۴. پایگاه داده

۱۸- کدام گزینه از روش های محاسبات مکعب داده (Cube Computation) می باشد.

۱. چند راهه ۲. توصیفی ۳. آماری ۴. ترکیبی

۱۹- و از تکنیک های پردازش پرسش های پیچیده در فناوری مکعب می باشند.

۱. مکعب های نمونه گزین - مکعب های رتبه بندی
۲. توصیف مکعب - مکعب های رتبه بندی
۳. ترکیب پردازش مکعبی - مکعب های نمونه گزین
۴. ترکیب پردازش مکعبی - توصیف مکعب

۲۰- کدام گزینه سه روش تحلیل داده های چند بعدی را بیان می کند.

۱. مکعب های پیش گو، مکعب های چندگانه، کشف مبتنی بر استثنا
۲. مکعب توصیفی، مکعب های پیش گو، مکعب چند گانه
۳. مکعب توصیفی، مکعب پیش گو، کشف مبتنی بر استثنا
۴. مکعب توصیفی، مکعب چندگانه، کشف مبتنی بر استثنا

۲۱- کدام گزینه بعنوان یک طبقه بندی الگوریتمهای استخراج اقلام مکرر شناخته نمیشود؟

۱. الگوریتم های شبیه Apriori ۲. الگوریتم های مبتنی بر رشد الگو
۳. الگوریتم هایی که از قالب عمودی داده ها استفاده می کنند ۴. الگوریتم های عددی

۲۲- کدام گزینه تعریف الگوهای منفی می باشد؟

۱. الگوهایی هستند که در مولفه های آنها رفتار همبسته ی منفی مشاهده می شود.
۲. الگوهای نادری هستند که به ندرت به وقوع می پیوندند.
۳. الگوهای تکرارشونده ای هستند که با رفتارهای منفی را اصلاح می کنند.
۴. الگوهای تکرارشونده ای هستند که استثنائات را بررسی می کنند.

۲۳- انواع محدودیت های راهبردهای داده کاوی مبتنی بر محدودیت عبارتند از:

۱. هرس الگو و هرس داده ها
۲. هرس الگو و هرس اطلاعات
۳. هرس اطلاعات و هرس دانش
۴. هرس داده و هرس دانش

۲۴- برای کاهش تعداد الگوهای برگردانده شده توسط فرآیند کاوش می توان از یا استفاده کرد.

۱. الگوهای فشرده - الگوهای تصادفی
۲. الگوهای فشرده - الگوهای تقریبی
۳. الگوهای تصادفی - الگوهای تقریبی
۴. الگوهای تصادفی - الگوهای بازگشتی

۲۵- کاهش داده ها هنگامی بی کاست خوانده می شود که بتوانیم

۱. اطلاعات مورد نظر را از جداول استخراج کنیم.
۲. دانش مدنظر استخراج کنیم.
۳. داده ها را فشرده کنیم.
۴. داده های اصلی را بدون از دست دادن اطلاعات از داده های فشرده بازسازی کنیم.

سوالات تشریحی

- ۱- کاوش متدولوژی های جدید داده کاوی باید شامل چه موضوعاتی باشد. ۱.۲۰ نمره
- ۲- تعاملات کاربر در پردازش داده کاوی را نام ببرید و یکی از آنها را تشریح کنید. ۱.۲۰ نمره
- ۳- رویکردهای متفاوت در طراحی انبار داده را نام برده و آنها را توضیح دهید. ۱.۲۰ نمره
- ۴- روش استنتاج صفت گرا را تشریح کنید. ۱.۲۰ نمره
- ۵- پالایش داده ها چیست و روش های اصلی آن را نام ببرید. ۱.۲۰ نمره

نمبر رد سواب	ياسخ صحيح
1	د
2	الف
3	ج
4	ب
5	الف
6	د
7	ب
8	الف
9	الف
10	ج
11	الف
12	ب
13	ب
14	الف
15	ج
16	الف
17	ب
18	الف
19	الف
20	الف
21	د
22	الف
23	الف
24	ب
25	د

سوالات تشریحی

۱.۲۰ نمره

۱- فصل ۱ صفحه ۳۷ منبع

۱.۲۰ نمره

۲- فصل ۱ صفحه ۳۸

۱.۲۰ نمره

۳- فصل ۴ صفحه ۱۵۰ - ۱۵۱

۱.۲۰ نمره

۴- فصل ۴ صفحه ۱۶۵

۱.۲۰ نمره

۵- فصل ۳ صفحه ۹۱-۹۲-۹۳-۹۴