

تعداد سوالات: تستی: ۲۵ تشریحی: ۵

زمان آزمون (دقیقه): تستی: ۶۰ تشریحی: ۶۰

سری سوال: یک ۱

عنوان درس: (داده کاوی و انبار داده ها، داده کاوی و انبار داده ها) آموزش محور

رشته تحصیلی/کد درس: مهندسی کامپیوتر-نرم افزار ۱۱۱۵۲۵۸ - ، مدیریت فناوری اطلاعات(مدیریت منابع اطلاعاتی)۱۲۱۸۳۳۰

استفاده از ماشین حساب ساده مجاز است

۱- یک معامله که هیچ k -itemset مکرری را دربر ندارد نمی تواند هیچ ... مکرری را در بر داشته باشد بنابراین چنین معامله ای را می توان علامت زد و یا از لیست حذف نمود. به چنین روشی Transaction reduction می گویند که برای ... به کار گرفته می شود.

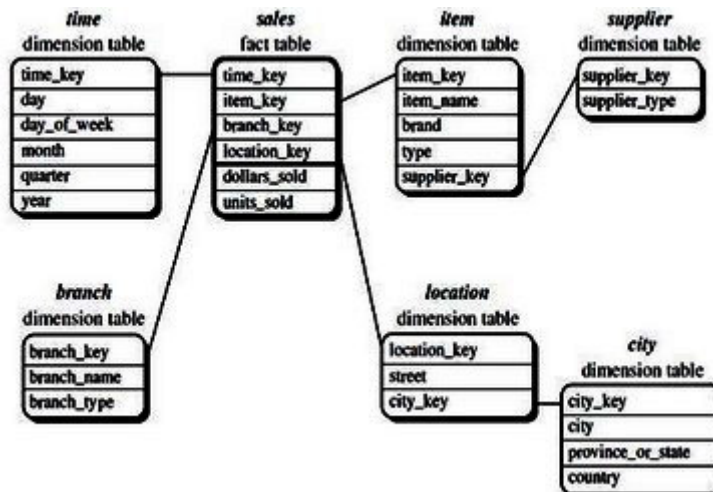
۱. k -itemset گسترش کارایی Apriori

۲. $(k+1)$ -itemset یافتن همبستگی

۳. $(k-1)$ -itemset گسترش کارایی Apriori

۴. $(k-1)$ -itemset یافتن همبستگی

۲- شکل زیر چه نوع نمایشی از دیتابیس می باشد؟



۲. Snowflake schema

۱. Star schema

۴. Data Cub

۳. Fact constellation

۳- در عملیات ... بر روی داده های چند بعدی در تکنیک OLAP، یک ... بر روی مکعب داده صورت می پذیرد.

۱. Slice and dice - عملیات از جزئیات کمتر به جزئیات بیشتر

۲. Drill-down - تجمیع

۳. Roll-up - تجمیع

۴. Drill-down - گزینشی از یک بعد

تعداد سوالات: تستی: ۲۵ تشریحی: ۵

زمان آزمون (دقیقه): تستی: ۶۰ تشریحی: ۶۰

سری سوال: ۱ یک

عنوان درس: داده کاوی و انبار داده ها، داده کاوی و انبار داده ها (آموزش محور)

رشته تحصیلی/کد درس: مهندسی کامپیوتر-نرم افزار ۱۱۱۵۲۵۸ - ، مدیریت فناوری اطلاعات(مدیریت منابع اطلاعاتی)۱۲۱۸۳۳۰

۴- در طراحی **data warehouse** دیدگاه **Top-down view** به چه معنی است؟

۱. اطلاعات اخذ شده، ذخیره شده و مدیریت شده توسط سیستمهای عملیاتی در سطوحی مجزا، و با صحتی متفاوت ارائه می گردد.
 ۲. شامل **fact tables** و **dimension tables** می باشد
 ۳. دیدگاهی از اطلاعات را به همراه دارد که از نقطه نظر آخرین کاربر (**end user**) مهم است.
 ۴. انتخاب اطلاعات مورد نیاز و مرتبط متناظر با نیازهای حال و آینده هدف مورد نظر.
- ۵- در بررسی دقیق روش نزدیکترین همسایگی وقتی تفاوت در مقیاس اندازه گیری ویژگیها وجود داشته باشد راه حل چیست؟

۱. ویژگی بامقیاس بالا اثر ویژگی بامقیاس پایین رامحو میکند.
۲. ویژگی بامقیاس پایین به صورت E در نظر گرفته میشود.
۳. مقادیر قبل از مقایسه نرمال میشوند.
۴. از میان استفاده میشود.

۶- متا دیتا باید شامل چه مورد زیر باشد؟

الف. توصیفی از ساختار **data warehouse**

ب. **Operational metadata**

ج. نگاشت از محیط عملیاتی به **data warehouse**

۱. گزینه الف ۲. گزینه ب ۳. گزینه ج ۴. گزینه الف - ب - ج

۷- تست طرحها و تعیین پارامترهای مورد نیاز مربوط به کدام مرحله از متدولوژی **crisp** است؟

۱. گام مدلسازی
۲. گام شناخت داده
۳. گام شناخت سیستم
۴. گام ارزیابی

تعداد سوالات: تستی: ۲۵ تشریحی: ۵

زمان آزمون (دقیقه): تستی: ۶۰ تشریحی: ۶۰

سری سوال: ۱ یک

عنوان درس: داده کاوی و انبار داده ها، داده کاوی و انبار داده ها (آموزش محور)

رشته تحصیلی/کد درس: مهندسی کامپیوتر-نرم افزار ۱۱۱۵۲۵۸ - ، مدیریت فناوری اطلاعات(مدیریت منابع اطلاعاتی)۱۲۱۸۳۳۰

۸- فرض کنید که بانک اطلاعاتی معاملات تنها دارای سه معامله

$\{ \langle a_1, a_2, \dots, a_{100} \rangle; \langle a_1, a_2, \dots, a_{50} \rangle; \langle a_1, a_2, \dots, a_{25} \rangle \}$ باشد و آستانه حداقل شمارش حمایت را برابر با

$min_sup=2$ بگیریم. مجموعه کالای بسته مکرر را بیان کنید؟

۱. $\{ \{a_1, a_2, \dots, a_{50}\}; 2; \{a_1, a_2, \dots, a_{25}\}; 3 \}$

۲. $\{ \{a_1, a_2, \dots, a_{100}\}; 2; \{a_1, a_2, \dots, a_{25}\}; 3 \}$

۳. $\{ \{a_1, a_2, \dots, a_{50}\}; 2; \{a_1, a_2, \dots, a_{100}\}; 1 \}$

۴. $\{ \{a_1, a_2, \dots, a_{100}\}; 1; \{a_1, a_2, \dots, a_{50}\}; 2; \{a_1, a_2, \dots, a_{25}\}; 3 \}$

۹- کدامیک از موارد زیر جزء تکنیک های مرتبط با گسترش کارایی Apriori نمی باشد.

۲. Sampling

۱. Hash-based technique

۴. Clustering

۳. Dynamic itemset counting

۱۰- کدامیک از موارد زیر جزء عملکرد داده کاوی نمی باشد؟

۱. توصیف مفهومی/کلاسی: مشخص کردن، تمایز دادن و یا هر دو

۲. کاوش الگوهای مکرر، انجمنی و همبستگی

۳. طراحی رابط کاربر

۴. آنالیز خوشه

تعداد سوالات: تستی: ۲۵ تشریحی: ۵

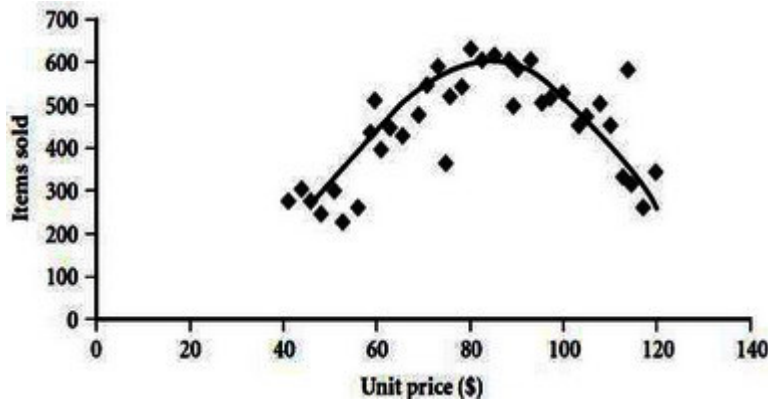
زمان آزمون (دقیقه): تستی: ۶۰ تشریحی: ۶۰

سری سوال: ۱ یک

عنوان درس: داده کاوی و انبار داده ها، داده کاوی و انبار داده ها (آموزش محور)

رشته تحصیلی/کد درس: مهندسی کامپیوتر-نرم افزار ۱۱۱۵۲۵۸ - ، مدیریت فناوری اطلاعات(مدیریت منابع اطلاعاتی)۱۲۱۸۳۳۰

۱۱- نمودار زیر چه عملیاتی را نمایش می دهد؟



۲. quantile plot

۱. رگرسیون محلی

۴. quantile-quantile plot

۳. Histograms

۱۲- ضریب همبستگی مثبت مابین دو متغیر دلالت بر ... و ضریب همبستگی صفر مابین دو متغیر دلالت بر دارد.

۱. افزایش یک متغیر در مقابل افزایش متغیر دیگر- عدم استقلال دو متغیر
۲. افزایش یک متغیر در مقابل افزایش متغیر دیگر- استقلال دو متغیر
۳. افزایش یک متغیر در مقابل کاهش متغیر دیگر-عدم استقلال دو متغیر
۴. افزایش یک متغیر در مقابل کاهش متغیر دیگر- استقلال دو متغیر

۱۳- جهت تست همبستگی بین دو متغیر برای متغیرهای ... از ضریب همبستگی و برای متغیرهای از chi-square استفاده می شود.

۱. غیر عددی- دسته ای
۲. کمی و یا عددی - غیر عددی
۳. کمی و یا عددی - دسته ای
۴. غیر دسته ای - دسته ای

۱۴- چهار ویژگی که انباره داده را از دیگر سیستم های پایگاه داده ای و فایلینگ متمایز می سازد عبار تند از:

۱. Subject-oriented, Durability, Time-variant, Isolation
۲. Subject-oriented, Isolation, Time-variant, Consistency
۳. Subject-oriented, Integrated, Time-variant, Nonvolatile
۴. Subject-oriented, Isolation, Consistency, Nonvolatile

تعداد سوالات: تستی: ۲۵ تشریحی: ۵

زمان آزمون (دقیقه): تستی: ۶۰ تشریحی: ۶۰

سری سوال: ۱ یک

عنوان درس: داده کاوی و انبار داده ها، داده کاوی و انبار داده ها (آموزش محور)

رشته تحصیلی/کد درس: مهندسی کامپیوتر-نرم افزار ۱۱۱۵۲۵۸ - ، مدیریت فناوری اطلاعات(مدیریت منابع اطلاعاتی)۱۲۱۸۳۳۰

۱۵- بخشی از فرآیند KDD است که ورودی آن از DataWarehouse و خروجی آن می باشد.

(KDD=Knowledge Discovery from Data)

۲. Pattern - Mining

۱. Knowledge - Data Mining

۴. Pattern - Data Mining

۳. Knowledge - OLAP

۱۶- روش FP-growth نسبت به روش در تولید الگوهای مکرر دارای امکان تولید بدون ایجاد کاندید می باشد و تعداد دفعات اسکن دیتابیس را می دهد.

۲. K-Mean - افزایش

۱. Apriori - کاهش

۴. K-Mean - کاهش

۳. Apriori - افزایش

۱۷- اگر $I_1 = \{T100, T400, T500, T700, T800, T900\}$ و $I_2 = \{T100, T400, T800, T900\}$ باشند آنگاه:

۱. $\text{diffset}(\{I_1, I_2\}, \{I_1\}) = \{ T100, T400, T800, T900 \}$

۲. $\text{diffset}(\{I_1, I_2\}, \{I_1\}) = \{ T500, T700 \}$

۳. $\text{diffset}(\{I_1, I_2\}, \{I_1\}) = \{ T500 \}$

۴. $\text{diffset}(\{I_1, I_2\}, \{I_1\}) = \{ T700 \}$

۱۸- بیش برآزش در کدام یک از حالات زیر می تواند رخ دهد؟

۱. وقتی که سیستم در جستجوی مدل‌های خاص است.

۲. وقتی حجم داده ها نامحدود بوده و سیستم در جستجوی مدل‌های بهینه است.

۳. وقتی الگوریتم به دنبال بهترین پارامترهای یک مدل خاص با استفاده از مجموعه محدودی داده میگردد.

۴. وقتی که سیستم در جستجوی مدل‌های خاص نیست.

تعداد سوالات: تستی: ۲۵ تشریحی: ۵

زمان آزمون (دقیقه): تستی: ۶۰ تشریحی: ۶۰

سری سوال: ۱ یک

عنوان درس: داده کاوی و انبار داده ها، داده کاوی و انبار داده ها (آموزش محور)

رشته تحصیلی/کد درس: مهندسی کامپیوتر-نرم افزار ۱۱۱۵۲۵۸ - ، مدیریت فناوری اطلاعات(مدیریت منابع اطلاعاتی)۱۲۱۸۳۳۰

۱۹- کدام مورد صحیح است؟

الف. الگوریتم **Apriori** از آن جهت از آلفگوریتم های رگرسیون و شبکه های عصبی بهتر است زیرا که منجر به ایجاد قواعد ملموس می شود.

ب. در هنگام آموزش مدلها ی دسته بندی، توزیع متوازن داده های آموزشی بین تمامی کلاسها منجر به نتیجه بهتری می شود.

ج. در هنگام آموزش مدلها ی دسته بندی، استفاده از تعداد داده های آموزشی کمتر از داده های آزمونی غالباً منجر به نتیجه بهتری می شود.

۱. گزینه الف ۲. گزینه ب ۳. گزینه ج ۴. گزینه الف و ب و ج

۲۰- در خصوص عملیات هرس در درخت تصمیم گیری، کدام گزینه صحیح می باشد؟

۱. کاهش نرخ خطا در آموزش، همواره نرخ خطا در آزمون را کاهش میدهد.

۲. افزایش نرخ خطا در آموزش، ممکن است نرخ خطا در آزمون را کاهش دهد.

۳. کاهش نرخ خطا در آموزش، ربطی به نرخ خطا در آزمون ندارد.

۴. افزایش نرخ خطا در آموزش، همواره نرخ خطا در آزمون را افزایش میدهد.

۲۱- کدامیک جزء مراحل پیش پردازش داده نمی باشد؟

۲. Data Transformation

۱. Data integration

۴. Data Reduction

۳. Classification

۲۲- کدامیک از پارامترهای زیر توصیف گر مرکزیت داده نمی باشد؟

۱. میانگین ۲. مد

۳. متوسط محدوده (midrange) ۴. انحراف معیار

۲۳- کدامیک از روش های زیر جهت برطرف کردن نویز قابل استفاده نمی باشد؟

۱. طبقه بندی ۲. رگرسیون

۳. خوشه بندی ۴. کاهش داده به عنوان نمونه گسسته سازی

۲۴- نرمالیزاسیون **Z-Score** باعث آشکارسازی می گردد ولی مقدار میانگین در این رابطه و شدیدتر از آن مقدار انحراف معیار در این رابطه به حضور این نوع داده ها حساس است.

۱. داده های ناسازگار ۲. داده های پرت

۳. داده های گسسته ۴. داده های ناسازگاری - داده های گسسته

تعداد سوالات: تستی: ۲۵ تشریحی: ۵

زمان آزمون (دقیقه): تستی: ۶۰ تشریحی: ۶۰

سری سوال: ۱ یک

عنوان درس: داده کاوی و انبار داده ها، داده کاوی و انبار داده ها (آموزش محور)

رشته تحصیلی/کد درس: مهندسی کامپیوتر-نرم افزار ۱۱۱۵۲۵۸ - ، مدیریت فناوری اطلاعات(مدیریت منابع اطلاعاتی)۱۲۱۸۳۳۰

۲۵- کدامیک از موارد زیر را نمی توان جزء تبدیل داده در نظر گرفت.

۱. ساختن شاخص جدید (Attribute construction or feature construction)
۲. تعمیم داده (Generalization)
۳. رگرسیون (Regression)
۴. هموارسازی داده (Smoothing)

سوالات تشریحی

۱.۴۰ نمره

۱- A- Data warehouse را تعریف نمایید و بیان کنید که تکنولوژی Data warehouse شامل چه مواردی است؟

B- با یک مثال مفاهیم زیر را برای یک مکعب OLAP توضیح داده و نمایش دهید.
Slice, Dice, Roll-up, Drill-Down, Pivot

۱.۴۰ نمره

۲- اگر در یک بانک اطلاعاتی داده فراموش شده وجود داشته باشد راههای حل مشکل آن را بیان نمایید؟

۱.۴۰ نمره

۳- داده زیر را با طول Bin های یکسان و برابر با سه را در نظر بگیرید. جهت حذف نویز از میانگین در طول Bin استفاده نمایید و داده نتیجه شده را بیان کنید.

۲۹ و ۲۵ و ۲۱ و ۲۰ و ۲۰ و ۱۹ و ۱۳ و ۱۲ و ۱۱ و ۱۰ و ۱۰ و ۱۰ و ۱۰ و ۵ و ۳ و ۲

تعداد سوالات: تستی: ۲۵ تشریحی: ۵

زمان آزمون (دقیقه): تستی: ۶۰ تشریحی: ۶۰

سری سوال: ۱ یک

عنوان درس: داده کاوی و انبار داده ها، داده کاوی و انبار داده ها (آموزش محور)

رشته تحصیلی/کد درس: مهندسی کامپیوتر-نرم افزار ۱۱۱۵۲۵۸ - ، مدیریت فناوری اطلاعات(مدیریت منابع اطلاعاتی)۱۲۱۸۳۳۰

۱.۴۰ نمره

۴- Answer only one of the following Questions

A- آلوگوریتم **Apriori** را بر روی اقلام زیر بکار برده و گامهای آن را شرح دهید. (حداقل پشتیبانی = ۳۰٪ و حداقل حمایت = ۶۰٪)

و برای مجموعه {M, D, K} قواعدی را که شرط حداقل حمایت را بر آورده می سازند مشخص نمایید.

Items	Tid
B, M	۱
B, D, K, E	۲
M, D, K, C	۳
B, M, D, K	۴
B, M, D, C	۵

B- با استفاده از شاخص جینی، درخت تصمیم را برای موارد زیر رسم نمایید.

سن	نوع ماشین	ریسک
۲۳	خانوادگی	زیاد
۱۷	اسپورت	زیاد
۴۳	اسپورت	زیاد
۶۸	خانوادگی	کم
۳۲	باری	کم
۲۰	خانوادگی	زیاد

۱.۴۰ نمره

۵- دیتا بیس های متداول (OLTP) on-line transaction processing و on-line analytical processing (OLAP) دارای چه تفاوت هایی می باشند؟

(OLAP) دارای چه تفاوت هایی می باشند؟