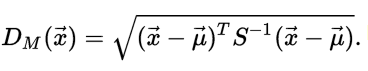
امیرحسین باوند 9331028

سوال یک : فاصله ماهالانوبیس عبارت است از نوعی اندازه گیری برای به دست آوردن فاصله یک توزیع و یک نقطه در فضای چندبعدی است . به طور کلی فاصله ماهالانوبیس به صورت زیر تعریف میشود



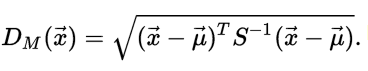
که s برابر است با ماتریس کواریانس بین متغیر های توزیع و x برابر است با برداری که میخواهیم فاصله اش را از توزیع به دست آوریم و u نیز برابر است با میانگین مقادیر هر بعد در توزیع

این فاصله از دوجهت با فاصله اقلیدسی متفاوت است .

اول این که همبستگی بین داده ها را برای محاسبه فاصله لحاظ می کند .

دوم این که به مقیاس اندازه گیری هر بعد بررسی ندارد و میتواند همه را به هم به درستی ارتباط دهد .

سوال دو : برای محاسبه این فاصله در یک فضای n بعدی نیازمند محاسبه ی ماتریس کواریانس توزیع داده شده که میخواهیم فاصله ی نقطه را از آن توزیع به دست آوریم و همچنین میانگین هر بعد از توزیع به صورت یک بردار میباشد همان طور که در بالا گفته شد فاصله ماهالانوبیس از رابطه زیر به دست می آید



که این که هر قسمت چیست در بالا توضیح داده شد .

سوال سه : نقطه صفر و دو احتمال نویز بیشتری دارد زیرا که با توجه به بیضی های شکل در بیضی های بیرونی تر قرار دارد بنابراین مستعد بیشتری برای قرارگیری به عنوان نویز دارد

سوال چهار : maximom likelihood روشی است برای تخمین زدن پارامترهای یک مدل آماری .

برای مثال میخواهیم اطلاعات قد یک جمعیت را داشته باشیم اما به دلیل محدوددیت زمانی نتوانیم قد همه را به دست آوریم تعدادی نمونه را از حمعیت انتخاب میکنیم و قد آنها را اندازه میگیریم حال با در دست داشتن اطلاعات مربوط به نمونه ای محدود از جمعیت و همچنین دانستن این که جمعیت به طور محتمل از خانوداده توزیع نرمال پیروی میکند اما نمیدانیم از چه نوع توزیعی از این خانواده استفاده میکند و مجهول است . میتوانیم با استفاده از روش maximum likelihood estimation تخمین مناسبی از واریانس و انحراف معیار داده ها به دست می آوریم . در حالت کلی روش mle در مورد یک مجموعه مشخص از داده ها عبارت است از نسبت دادن مقادیری به پارامتر های مدل که در نتیجه آن توزیعی تولید شود که بیشترین احتمال را به داده های مشاهده شده نسبت دهد . mle یک سازوکار مشخص را برای تخمین ارائه میدهد که در مورد توزیع نرمال و بسیاری توزیع های دیگر به صورت خوش تعریف عمل میکند .

سوال پنج

سوالات شش و هفت :

حال در این قسمت و قسمت بعد کد سوال خواسته شده که در فایل کدهای ارسالی در بخش p1 موجود است همچنین دقت الگوریتم برایر با 74.97 درصد می شود

بخش دوم :

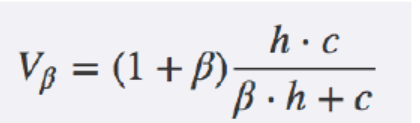
سوال یک :

معیار v-measure به عنوان یک میانگین متعادل شده بین homogeneity و completness در خوشه بندی است .

Hemogenity : شباهت داده های درون کلاستر است و زمانی ماکزیمم میشود که تعداد کلاس های متفاوت در یک کلاستر به کمترین مقدار برسد .

Completness : سعی بر این میکند که تمام داده های یک کلاس را در یک کلاستر قرار دهد.

رابطه v- measure به صورت زیر است که

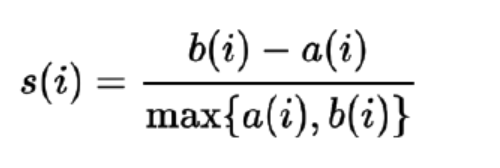


که h همان hemogenity و c همان completeness است

قسمت دو :

این معیار یک روش برای تفسیر و ارزیابی سازگاری داده های درون یک کلاستر می باشد. این مقدار بیانگر میزان شباهت یک نمونه با خوشه نظیر شده به آن در مقایسه با سایر خوشه ها می باشد. که این مقدار بین ۱و- ۱ می باشد. هرچه این مقدار بیشتر باشد بیانگر این است که نمونه به صورت بهتری به خوشه مربوط به خودش نظیر شده است. هرچه داده های بیشتری مقدار بالایی داشته باشد بیانگر عملکرد بهتر الگوریتم و خوشه بندی مناسب انجام شده است ولی اگر این مقدار برای داده های زیادی پایین باشد بیانگر این است که تعداد خوشه ها یا خیلی کم و یا خیلی زیاد است

مفدار آن بع صورت فرمول زیر است



که bi برابر میانگین کمترین فاصله نقطه i ار سایر خوشه ها است و ai برابر میانگین تفاوت داده i از سایر داده هایی که داخل خوشه یکسانی هستند قرار میدهیم .

قسمت ب :

سوال یک :

یک روشه خوشه بندی پایین به بالا می باشد که در ابتدا هر داده یک خوشه در نظر گرفته می شود و در هر مرحله

دو خوشه نزدیک به هم با یکدیگر ادغام شده و خوشه بالاتر را تشکیل می دهند.

Linkage : تفاوت خوشه ها رابه عنوان یک function از فاصله باینری داخل خوشه ها مشخص میکند

خودش دارای دو روش متداول است که یکی ward و یکی complete است

در complete بیشترین فاصله بین اجزای مجموعه رادرنظر میگیریم .

در روش ward واریانس خوشه هایی که با هم ادعام میشوند را مینیمم میکنیم.

k = 2

0.596301398889

k = 3

0.471359346878

k = 4

0.488539509655

k = 5

0.497783023538

k = 6

0.499782804877

k = 7

0.462655414494

k = 8

0.443103485153

k = 9

0.41744369051

k = 10

0.416903844508

k = 11

0.402519371478

k = 12

0.38569969115

k = 13

0.369852245418

k = 14

0.374849604182

k = 15

0.372417807523

k = 16

0.365389285535

k = 17

0.356733678602

k = 18

0.347309436664

k = 19

0.353233918169

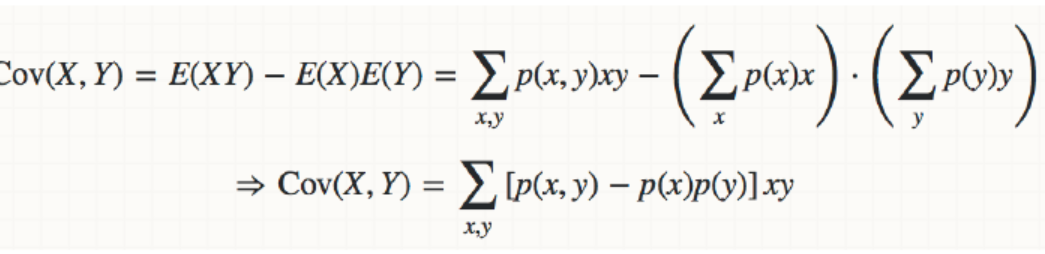
k = 20

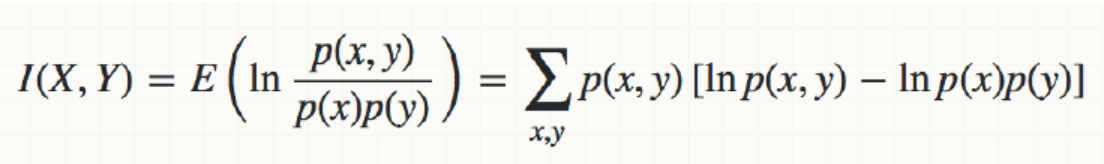
0.350551981611

سوال دو :یک مقداراست که برایtarget value برای داده های آموزش و تست انداره گیری میشود. این مقدار حاوی خظا است زیرا فیزیکی اندازه گیری میشود . مقادیر و دقت و صحبت اندازه گیری این ویژگی است که کیفیت مدل را مشخص میکند

سوال سه :

L(x,y) در واقع مقادیر بین ۰ تا مثبت بی نهایت برای ارزیابی میزان اطلاعاتی که بیت x, y در اشتراک قرار دارد را نسبت میدهد . در حالیکه normalized آن سقف معیار را محدود میکند .





هر دو رابطه بین ویژگی ها رامشخص میکنند اما nml کلی تر میباشد .

هر دو یک مقدار در یک بازه مخدود ثابت برمیگزداندد

Nml فاصله بین توزیع احتمال را محاسبه میکند اما cov فاصله خطی بین دو متغیر تصادفی را محاسبه میکند.

در محاسبه cov بین دو متغیر تصادفی ضرب آن ها را به صورت وزن دار جمع میکنیم ولی در l آن ها ضرب احتمال توام را محاسبه کرده و جمع وزن دار میکنیم .

قسمت چهار :

NMI 0.00278929661532

v measure 0.00156711296395

Ward:

NMI 0.0217726380547

v measure 0.0217578104762

در این چا استفاده از ward بهتر است زیرا v-measure را افزایش داده است

بخش سوم :

سوال یک : در داده‌کاوی، **یادگیری قانون وابستگی**  یک متد مناسب برای یافتن روابط جذاب بین متغیرهای موجود در پایگاه داده‌های بزرگ است. در واقع هدف آن این است که از داده ها و پایگاه داده اطلاعاتی استخراج کند که بتواند با استفاده از آن ها قوانینی به دست بیاورد که به احتمالی درست باشند . برای مثال در یک سوپر مارکت میگوییم اگر یک نفر شیر بخرد به احتمالی ماست هم میخرد و این اطلاعات میتواند به ما در مواردی مانند قرار دادن شیر و ماست کنار هم در یک یخچال کمک کند .

سوال دوم : یعنی این که اگر پدیده A اتفاق بیافتد احتمالا پدیده B نیز احتمالا اتفاق خواهد افتاد . البته این یک احتمال است و قطعی نیست .و این احتمال قابل محاسبه خواهد بود .

سوال سه :

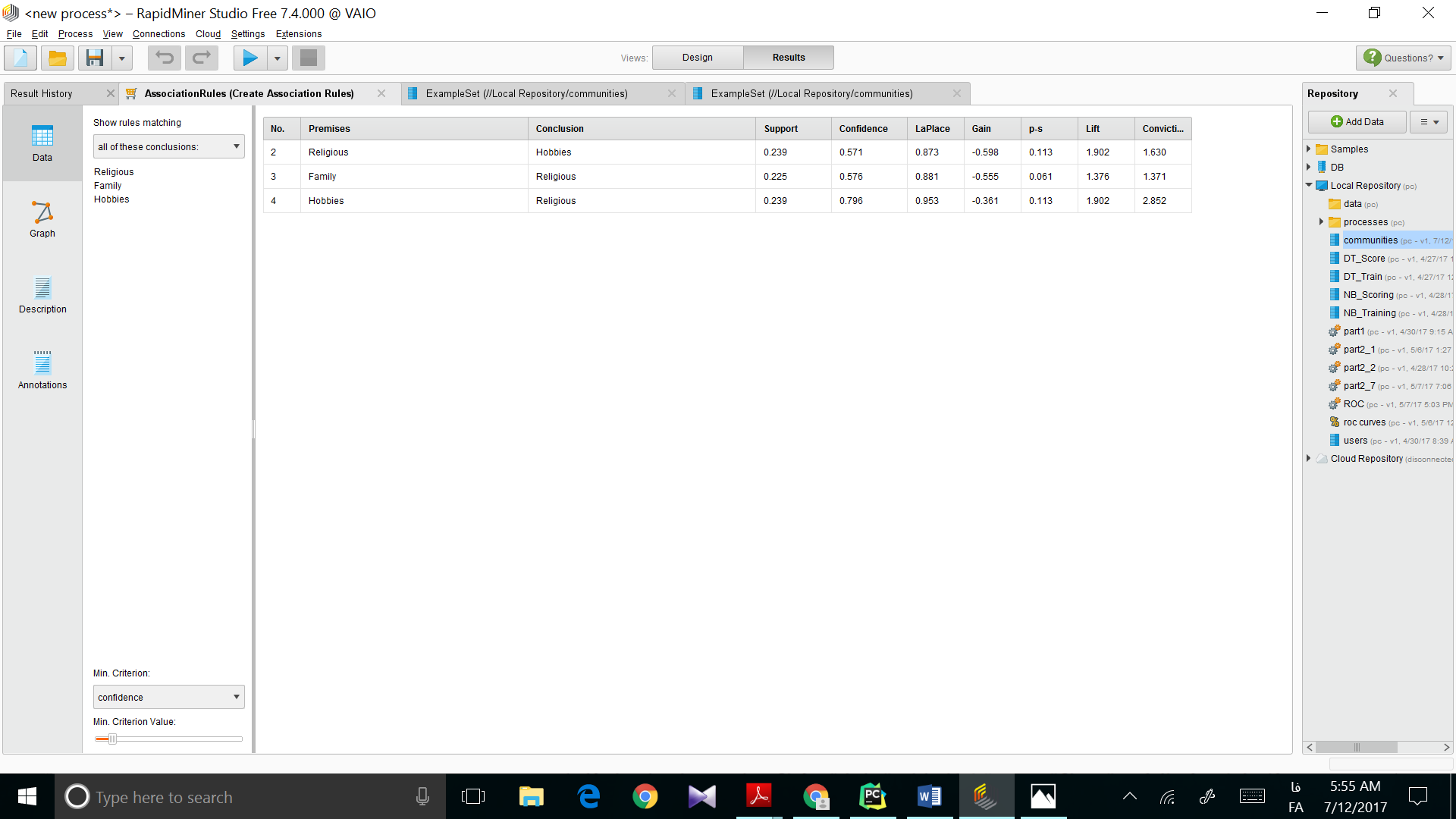
Support به معنی این است که نسبت هر تراکنش در پایگاه داده ما جقدر است در واقع مثلا در سوپر مارکت اگر تراکنش شیر و ماست خریدن را یک تراکنش در نظر بگیریم در چند درصد مواقع از میان تراکنش های تعریف شده در پایگاه داده این تراکنش خاص اجرا میشود.

Confidence به معنای این است که احتمال این که یک قانون عملی شود چند درصد است. در واقع همان احتمالی است که در سوال یک تعریف کردیم. برای مثال در یک سوپر مارکت اگر یک قانون این باشد که اگر کسی شیر بخرد احتمالا ماست هم خواهد خرید و این قانون در نیمی از مواقع درست باشد یعنی این که confidence این قانون برابر است با یک دوم .

حال سراغ مساله پیاده سازی میرویم در ابتدا سوال پرسیده شد که فرق بین binomial و binominal چیست که در واقع فرقی ندارند و احتمالا طراح سوال در جایی آن را خوانده که متعلق به همان موضوع خاص بود اما در واقعیت بین این دو تفاوتی وجود ندارد .سپس سوال پرسیده شده است که FP-GROWTH چه عملی را انجام میدهد :

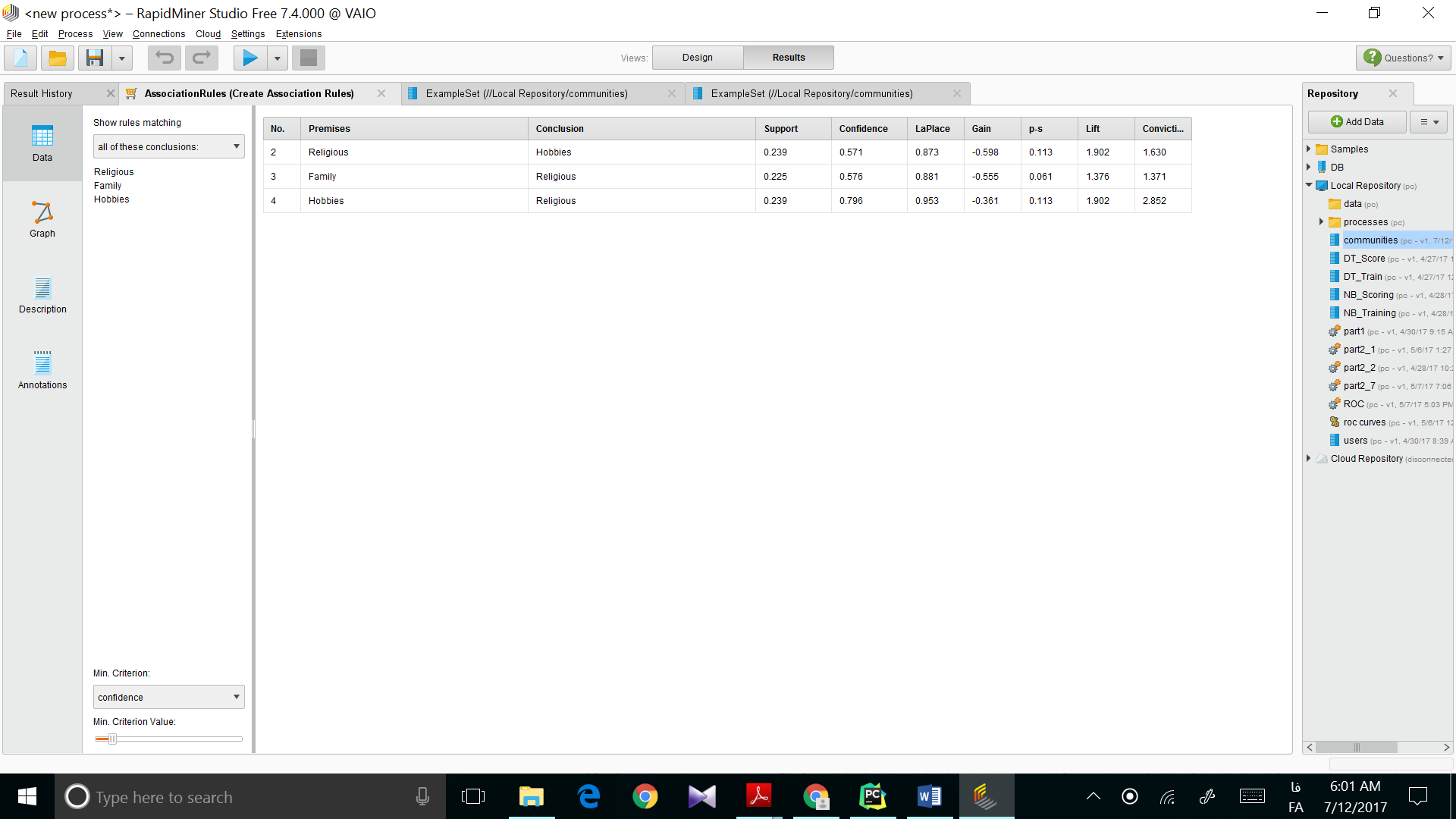
این عملگر در واقع تمام item set های مکرر را در که در دیتاست ما وجود دارد را به وسیله درخت fp استخراج میکند این الگوریتم نیاز دارد که تمام صفت های دیتاست ورودی باید binominal باشد .

ابتدا با سه دهم اجرا میکنیم و به جواب زیر می رسیم



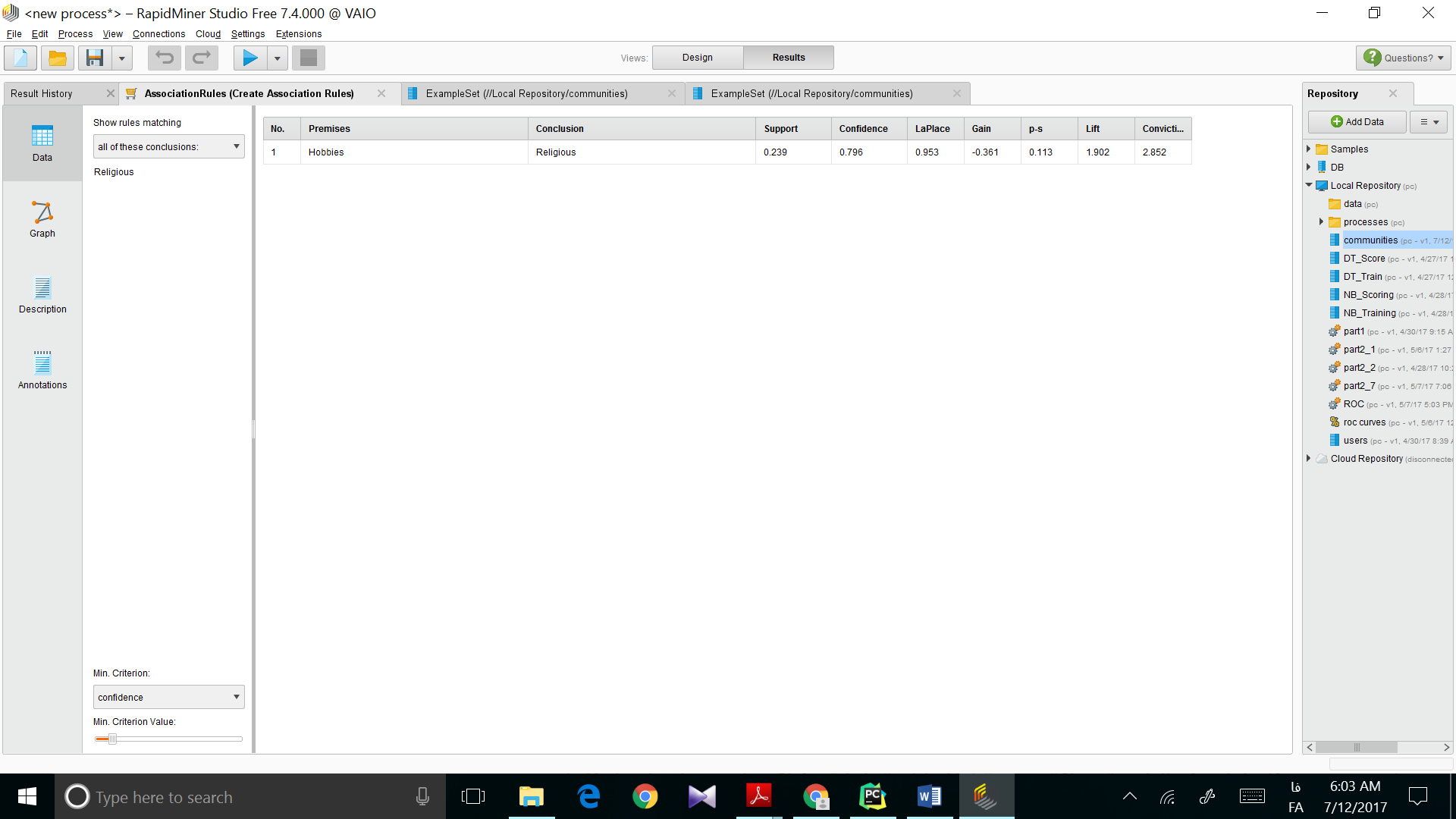
که نتایج در جدول بالا مشخص شده است .

حال نتایج را با عدد پنج دهم اجرا میکنیم و به جواب زیر می رسیم



که ملاحظه میشود تغییری در support و confidence به وجود نیامده است

اما وقتی با عدد هفت دهم اجرا میکنیم جواب متفاوت میشود و فقط یک قانون پیدا میشود



سوال چهار :

در این سوال فایل های خواسته شده را رسم میکنیم و در پوشه سوال چهار قرار میدهیم

ماژولاریتی:یک پارامتر برای ارزیابی شبکه ها و گراف ها می باشد که قدرت تقسیم شبکه به ماژول های جدا از هم را نشان می دهد. این پارامتر اغلب در روشهای بهینه سازی برای تشخیص ساختار جامعه در شبکه ها استفاده می شود