"نام"

**خلاصه**

در این پروژه سعی شده است با توجه به دیتا ست داده شده قوانین تجمعی ای (Association Rules) را استخراج کرد که بتواند در یافتن ارتباطات بین محصولات (ProductName) عمل کند.

برای پاکسازی داده ها از روش های مختلفی استفاده شده که در زیر خواهیم دید

**نمای داده ها**

در این پروژه بنا بر اینکه تمرکز روی استخراج قوانین تجمعی بوده است تنها از ویژگی ProductNames استفاده شده است تا بتوان ارتباط بین محصولات را دریافت (میتوان ویژگی هایی مانند قیمت را نیز گسسته سازی کرد (Binning) و در کنار محصولات استفاده کرد اما با توجه به کابرد رایج ProductNames انجام اینکار کمی بی معنی خواهد بود



**تمیز سازی داده ها**

از آنجایی که نام محصولات تا حد تثیر گذاری با نام برند آن ها مشخص شده بررسی سطر به سطر سبد ها محصول و جایگذاری نام های برند با خود محصول ممکن نیست اما سعی شده است از کلیاتی تا جای ممکن برای تمیز سازی استفاده شد.

برای تمیز سازی داده ها از روش های زیر به ترتیب استفاده شده است:

* محصولاتی که کمتر تر از 3 حرف (2 یا کمتر) داشته باشند را حذف میکنیم با اینکار تعدادی از سطر ها نیز حذف خواهد شد. (مثال زیر سایز یک محصول را نشان میدهد)



* اعداد را در محصولات حذف میکنیم:



* محصولاتی نیز وجود دارند که هم بشکل مفرد و هم بشکل غیر مفرد نامبرده شده اند (مانند Pork و Porks) این محصلات باید هر دو با نام مفرد خود ظاهر شوند.
* محصولاتی وجود دارند که با پسوند هایی مانند سایز آن ها یا رنگ یک محصول ذکر شده اند.



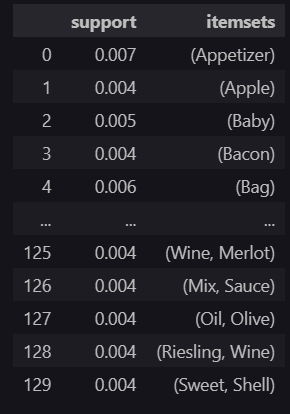
* پسوند هایی وجود دارند که با هم مترادف بوده و با یکی کردن نام آن ها به پیچیدگی کمتر مدل منجر میشود. (مانند Organic و Untrimmed).

روش های دیگری را نیز میتوان در پاکسازی داده ها بکار برد (برای مثال استفاده از Regex میتواند بسیار مفید باشد) اما بنا بر اینکه در داده های ما سبد های کالا حداکثر 3 محصول دارند (داده ها میتواند مصنوعی یا Synthetic باشد چرا که عموما در سبد های کالا تعداد محصولات بیشتر خواهد بود. که خود به نفع ما است چرا که در ابتد هدف از این نوع آنالیز داده پیدا کردن محصولاتی است که معمولا با هم خریداری میشوند)

با توجه به توضیح بالا انتظار خواهیم داشت که Support آیتم ست ها کم و آیتم ست های بزرگی نداشته باشیم. در ادامه خواهیم دید که این انتظار به وقوع میپیوندد.

**استخراج آیتم ست های معمول**

در عکس زیر تعدادی از آیتم ست های رایج همراه ساپورت آن ها آمده است

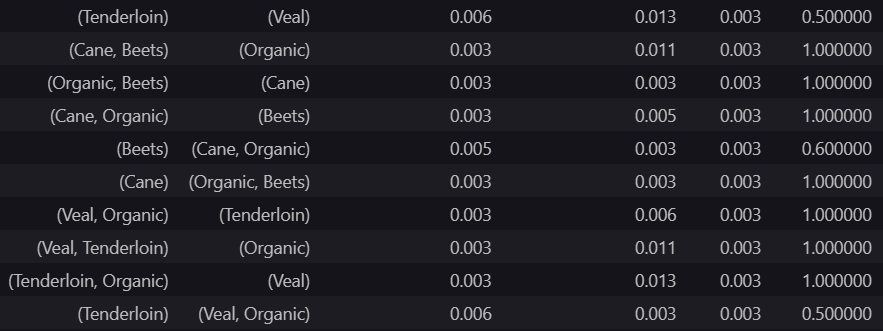
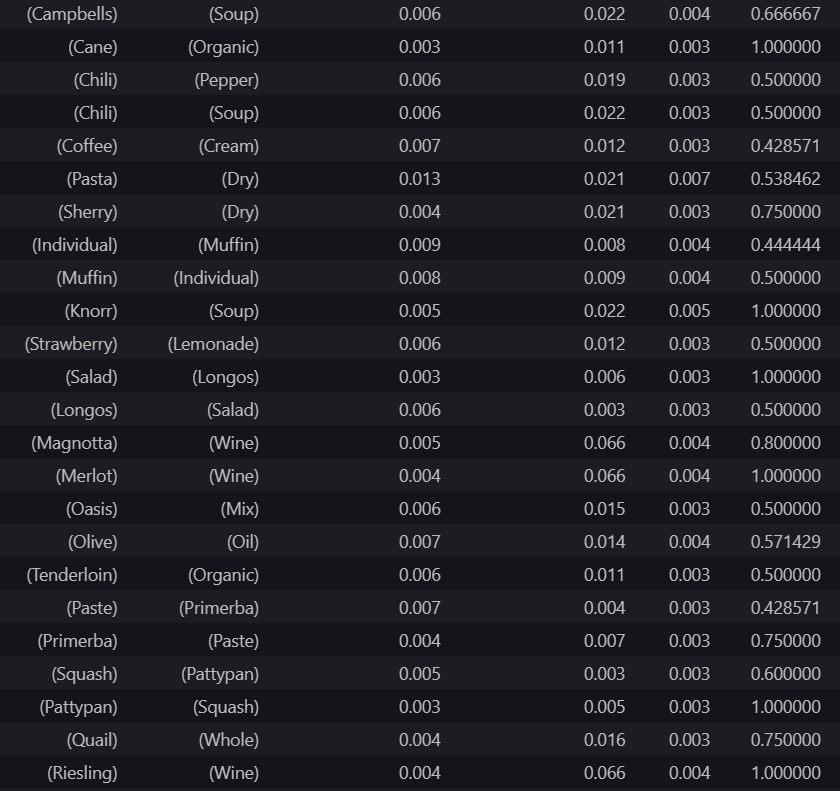
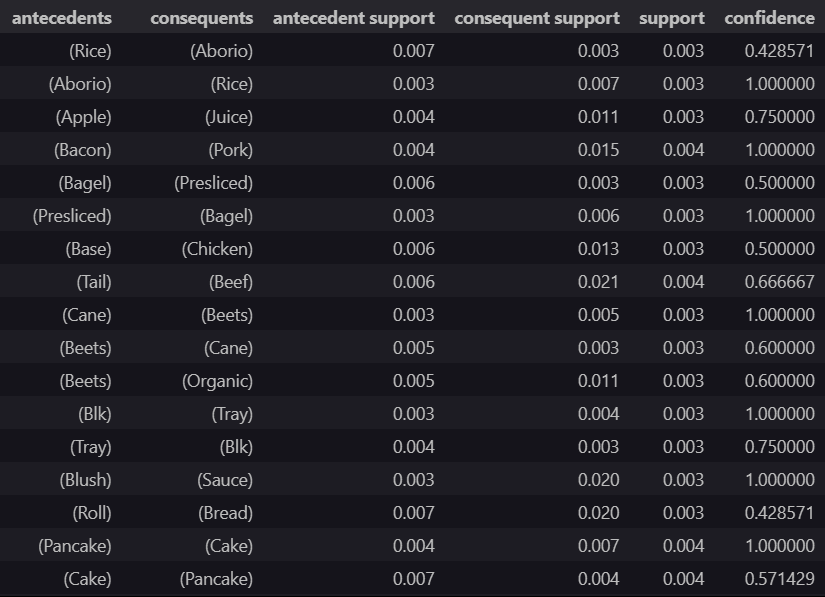


همانطور که میبینیم ایتم ست های بزرگتری از 2 آیتم دارای ساپورت نسبتا زیادی نیستند.

برای اینکه قوانین استخراجی دارای معنا باشند مجبور هستیم مینیمم ساپورت را عددی پایین مانند (0.003 یعنی حداقل 3 رخداد از 1000 سطر داده ها) قرار دهیم، البته سودی در این ساپورت پایین نهفته است که در زیر خواهیم دید.

تعدادی از محصولاتی که در زیر میبینیم نام های غریبی داشته به همین دلیل توضیح مختصری از این محصولات میبینیم:

* Aborio: نوعی از برنج (شرقی)
* Knor: نام برندی از طعم دهنده سوپ (این آیتم با محصول Soup ارتباط زیادی دارد)
* Oasis: نوعی نوشیدنی الکلی
* Merlot: نوعی نوشیدنی الکلی
* Magnotta: نوعی نوشیدنی الکلی
* Riesling: نوعی نوشیدنی الکلی
* Rose: نوعی نوشیدنی الکلی
* Organic: منظور محصولات سبزیجاتی است (این محصولات با تعدادی از انواع گوشت مانند گوشت پرندگان و گاو زیاد خریداری شده است)
* ...



اگر به میزان اعتماد (confidence) توجه کنیم میبینیم که با این حال که Support محصولات به نسبط عرف آنچه که باید در مسائل مشابه باشد کمتر است اما با درجه اطمینان بسیار خوبی میتوان گفت قوانین بسیار قابل اعتماد هستند.

**نتیجه گیری**

مهمترین نتیجه ای که از این تمرین حاصل میشود این است که برای آنکه قوانین استخراج شده نه تنها قابل اعتماد بلکه قابل استفاده باشند (عموما برای پیشنهادات خرید های اینترنتی از این نوع آنالیز داده ای استفاده میشود) باید در داده های خود سبد های خرید نسبتا بزرگی داشته باشیم در غیر این صورت درجه اعتماد بسیار بالا و Support پایین خواهد بود (البته در مورد این گونه داده ها میتوان از Support نسبی استفاده کرد)

تمیز سازی داده ها بسیار پر اهمیت بوده و احتمال رخداد اشتباه در ثبت داده ها توسط انسان بسیار رایج است.

همچنین گویا نوشیدنی های الکلی بسیار در فروشگاه های عمومی پر طرفدار هستند D: