باسمه تعالی

تکلیف سری اول درس داده کاوی

سارا برادران (شماره دانشجویی : ۹۶۲۴۱۹۳)

سوال ۱)

CRISP-DM phases

✓ **Business/Research Understanding Phase**

فاز آشنایی با مسئله و اینکه حل مسئله تعیین روند مصرف روند دارو ها در فصل زمستان قرار است دقیقا چه مشکلی را از سازمان مربوطه حل کند :

می بایست بررسی کنیم آیا داده هایی که در اختیار است برای حل مسئله داده کاوی مور نظر کافی است یا خیر.

برای مثال آیا داده های آب و هوایی حاوی پارامتر های لازم می باشد یا خیر و یا اینکه داده های مربوط به مراجعات سالیانه به بیمارستان اصلا حاوی اطلاعاتی در باره میزان تجویز و مصرف دارو ها هست یا خیر.

✓ **Data Understanding Phase**

collect the data : داده های مربوط به آب و هوا و مراجعات بیمارستان را جمع آوری می کنیم.

Exploratory data analysis : رنج مصرف دارو در شرایط مختلف را مورد بررسی قرار می دهیم و سعی میکنیم ارتباط میان میزان مصرف دارو و شرایط آب و هوایی را به صورت ساده و اولیه کشف کنیم برای مثال ممکن است داروی مربوط به سرماخوردگی در فصول سرد تر سال همچون پاییز و زمستان مصرف بیشتری داشته باشد. این روابط ساده و اولیه را پیدا می کنیم تا در فاز های بعدی با کمک مدلینگ به شکل قوی تر مورد بررسی قرار گیرند.

✓ **Data preparation phase**

فاز آماده سازی داده می تواند شامل موارد زیر باشد :

۱- شناسایی داده های پرت : اگر میزان مصرف یک دارو در برخی داده ها بیش از حد تصور بزرگ یا کوچک باشد احتمالا داده پرت است و می بایست این مورد در نظر گرفته و بررسی شود.
۲- داده های مربوط به وضعیت آب و هوایی گوناگون می تواند دسته بندی شود و با توجه به پارامتر های آب و هوا این داده ها در چندین دسته همچون فصول سال (بهار ، تابستان ، پاییز و زمستان) قرار گیرد.
۳- انتخاب فیلد ها و پارامتر های مناسب و حذف فیلد های اضافه ای که در حل مسئله داده کاوی مربوطه کمکی نمیکند.

✓ **Modeling phase**

داده های در دسترس را cross-validate کرده و به چندین دسته طبقه بندی و تقسیم می کنیم.

به نحوی می بایست داده های در دسترس را بالانس کنیم فرضا اگر تعداد داده هایی که نمایانگر میزان مصرف یک دارو در آب و هوای گرم است بسیار بیشتر از تعداد داده های در دسترس نمایانگر میزان مصرف همان دارو در آب و هوای سرد باشد آن گاه می بایست به نحوی داده های متعادل ایجاد کنیم تا مدل به خوبی یادگیرد در غیر اینصورت احتمال اینکه الگوریتم داده کاوی به درستی عمل نکند زیاد خواهد بود.

فرمت داده های آماده سازی شده را بررسی کرده و در صورت لزوم به فاز Data preparation بازگشته و داده ها را به فرمت مناسب برای حل مسئله با یک مدل و تکنیک خاص تبدیل می کنیم.

می بایست میزان دقت مدل مورد نیاز را تعیین کنیم و براساس آن مدل مناسب را طراحی کنیم.

✓ **Evaluation phase**

وضعیت مدل ایجاد شده را به شکل کاربردی و عملی مورد بررسی قرار داده و اطمینان حاصل می کنیم که مدل طراحی شده دقیقا مسئله مد نظر را حل کرده باشد و تعیین نماید روند مصرف دارو ها در فصل زمستان نسبت به دیگر فصول سال چگونه است. به علاوه در صورتی که برخی پارامتر های مهم و تاثیرگذار در حل مسئله داده کاوی مربوطه فراموش شده باشد می بایست آن ها را به مدل اضافه کنیم.

✓ **Deployment phase**

براساس نتایج بدست آمده در داده کاوی می بایست سازمان مربوطه تصمیماتی اتخاذ کرده و یا روندی را در پی نتایج حاصل از داده کاوی طی کند. برای مثال اگر بر اساس نتایج بدست آمده مشخص شد که داروی x در فصل زمستان بیشتر مورد مصرف است چنانچه مشتری این فرآیند داده کاوی شرکت تولید کننده دارو باشد می بایست میزان تولید دارو x خود را در فصل زمستان نسبت به دیگر فصول سال افزایش دهند و یا اگر مشتری مربوطه بیمارستان است پس می بایست در فصل زمستان به میزان مورد استفاده دارو را خریداری کرده و بالعکس در فصول دیگر میزان کمتری از این دارو را از شرکت ها خریداری کند چرا که در صورتی که دارو گران قیمت باشد و بدون آنکه استفاده شود منقضی گردد بیمارستان را متوجه خسارت مالی خواهد کرد.

سوال ۲)

مثال : تخمین نمره یک درس دانشجویان برمبنای نمرات دروس پیش نیاز آن درس(estimation task)

با استفاده از روش regression می توان این task داده کاوی را حل نمود :

Goal : تخمین نمره یک درس دانشجو بر اساس نمراتی که در دروس پیش نیاز آن درس کسب کرده است.

Approach :

جمع آوری نمرات دانشجویان در دروس پیش نیاز

داده های جمع آوری شده توسط سیستم یادگیری می شود و ارتباط میان نمرات دروس کشف می گردد.

سپس سیستم تلاش می کند بر مبنای نمرات دروس قبلی یک مدل برای تخمین نمره درس فعلی پیدا کند.

مثال : دسته بندی دانشجویان دانشگاه بر اساس پارامتر هایی چون جنسیت ، رشته تحصیلی، معدل ، خوابگاهی بودن یا نبودن و مقطع تحصیلی(clustering task)

با استفاده از روش clustering می توان این task داده کاوی را حل نمود :

Goal : خوشه بندی دانشجویان دانشگاه بر اساس پارامتر هایی چون جنسیت ، رشته تحصیلی، معدل ، خوابگاهی بودن یا نبودن و مقطع تحصیلی برای دادن امکانات و خدمات دانشگاهی مناسب به هر دسته از دانشجویان.

Approach :

اطلاعات و ویژگی های مربوط به دانشجویان من جمله معدل ، جنسیت ، خوابگاهی بودن ، رشته و مقطع تحصیلی را جمع آوری می کنیم.

سپس دانشجویان مشابه را در یک دسته قرار داده و سپس کیفیت این خوشه بندی را بررسی می کنیم برای مثال آیا روند معدل و نمرات دانشجویان یک خوشه در طول زمان مشابه است یا خیر.

شاید نیاز باشد چندین خوشه در گام های بعدی ادغام یا جداسازی شوند. خوشه بندی می بایست به گونه ای باشد که دانشجویان یک دسته بیشترین میزان مشابهت را داشته و دانشجویان دسته های گوناگون بیشترین تفاوت را نسبت به هم داشته باشند.

سپس می توانیم در مراحل بعدی امکانات و خدمات مشابه را به دانشجویانی که در دسته های یکسان قرار می گیرند ارائه دهیم و یا به دسته های مشابه پیشنهادات مشابهی برای اخذ دروس اختیاری ارائه کنیم.

مثال : دسته بندی دانشجویان دانشگاه برمبنای رشته تحصیلی ، سال ورود ، مقطع تحصیلی و معدل به سه دسته ممتاز ، خوب و متوسط (classification task)

با استفاده از روش classification می توان این task داده کاوی را حل نمود :

Goal : تشخیص دسته ای که یک دانشجو بر مبنای رشته تحصیلی ، سال ورود به دانشگاه ، مقطع و معدل در آن قرار می گیرد.

Approach:

می بایست اطلاعات دانشجویان اعم از رشته تحصیلی ، سال ورود ، معدل و مقطع تحصیلی را جمع آوری نماییم.

در طی فرآیند یادگیری داده های برچسب گذاری شده به مدل داده می شود و مدل یاد میگیرد کدام یک از دانشجویان با چه اطلاعتی در کدام دسته قرار می گیرند

از مدل بدست آمده برای برچسب گذاری وضعیت تحصیلی یک دانشجوی جدید استفاده کنیم.

مثال : تشخیص افت تحصیلی یک دانشجو (classification task)

با استفاده از روش classification می توان این task داده کاوی را حل نمود :

Goal : هدف این است که بتوانیم دانشجویی که احتمالا در حال افت تحصیلی است را شناسایی کنیم و با ارائه خدماتی همچون مشاوره تحصیلی و ... به وی از این امر جلوگیری کنیم.

Approach :

اطلاعات مورد نیاز من جمله ریز نمرات درس های دانشجو در ترم های گذشته و ترم جاری را جمع اوری کرده وضعیت مشروطی یا تعداد دروس مردود شده را جمع آوری کرده و مورد بررسی قرار می دهیم.

دانشجویان را برچسب گذاری کرده وضعیت تحصیلی آن ها را نرمال یا در حال افت قرار داده و به مدل می دهیم تا بر مبنای این داده ها یادگیرد چه دانشجویانی در حال افت تحصیلی هستند.

سپس سیستم تلاش می کند مدلی کشف کند که نشان دهد تحت چه شرایطی دانشجو در حال افت تحصیلی است و درصورتی که یک دانشجوی جدید به همراه اطلاعات وی به مدل بدهیم مشخص نماید دانشجو در حال افت تحصیلی است یا خیر.

مثال : بررسی و کشف پارامتر های موثر در موفقیت تحصیلی دانشجویان اول تا سوم دانشکده مهندسی کامپیوتر (description task)

مثال : کشف ارتباط میان دروس اختیاری اخذ شده توسط دانشجویان رشته کامپیوتر (association rule task)