«به نام خدا»

تکلیف اول - مرضیه علیدادی – 9631983

1.

- فاز اول: business understanding phase :

در این فاز باید با مسئله آشنا شویم. باید درک متقابلی از طرف ما، به عنوان دانشمند داده، و مشتری به وجود آید. در وهله ی اول، باید هدف انجام این کار را بررسی کنیم. اینکه پیشبینی روند مصرف دارو، با چه هدفی قرار است صورت گیرد؟! و از چه جنبه هایی باید بررسی شود؟! آیا مسائل آب و هوایی مدنظر است؟! آیا قرار است مقایسه نسبت به سال های مختلف صورت گیرد؟! آیا مقایسه بیمارستان ها مدنظر است؟! آیا بررسی روند مصرف دارو، به صورت فردی برای مراجعه کنندگان مدنظر است؟! آیا تعداد دفعات مصرف دارو فقط اهمیت دارد، یا مقدار مصرف آن نیز مدنظر است؟! و ... . هدف های مربوط به حوزه ی سلامت و داروی این پروژه، هدف های مربوط به بحث های مدیریتی، و هدف های اصلی حوزه ی بیزنسی مشتری که موجب موفقیت کار می شوند، را تعیین می کنیم. هدف های مربوط به حوزه ی داده کاوی و شاخص های مهم رسیدن به نتیجه ی موفقیت آمیز را تعریف می کنیم. به دنبال این شفاف سازی ها، باید برنامه ریزی مناسب برای رسیدن به آن ها صورت گیرد. در وهله ی دیگر، باید اهداف مشخص شده، برای استفاده در تحلیل های داده کاوی، فرموله شوند. و باید بررسی کنیم که آیا مطابق هدف هایمان، data های لازم را در اختیار داریم یا خیر. باید بررسی شود که آیا برای داروها، میزان مصرف آن ها، فصل مصرف آن ها، وضعیت آب و هوایی، و بیمارستان محل ثبت آن ها ثبت شده است یا خیر. باید بررسی شود که داده ها ناقص نباشند و پاسخگوی اهداف مسئله باشند.

- فاز دوم: data understanding phase :

در این فاز باید به فهم data ای که در اختیار ما قرار داده اند، بپردازیم. پروتکل ها و مدل های مرسوم در حوزه ی سلامت و دارو را بررسی و مشخص می کنیم. سپس داده های مشتری را در یافت می کنیم. و به توصیف داده های در دسترس می پردازیم. داده های مورد نیاز برای تحلیل مان را، از بین داده های در دسترس استخراج می کنیم. مثلا مشخصات مراجعه کنندگان، دارو ها، مقدار مصرف آن ها، بیمارستان محل ثبت آن ها و فصل مصرف آن ها و به طور کلی اطلاعاتی که پاسخگوی هدف های مشخص شده در فاز قبل است، را استخراج می کنیم. در نهایت، کیفیت داده های موجود در

dataset را تایید می کنیم تا به فاز بعد برویم.

- فاز سوم: data preparation phase :

باید data را آماده سازی کنیم، تا برای اجرای الگوریتم های داده کاوی روی آن ها، مناسب شوند.

1- داده های پرت را باید بررسی کنیم. برای مثال اگر مقدار مصرف دارو برای یک فرد در یک زمان خاص، منفی ثبت شده باشد، یا از مقداری که به صورت معمول و منطقی در پایگاه داریم، بیشتر باشد، داده ی پرت محسوب می شود. و باید آن را بررسی کنیم که آیا درست است یا خیر. یا مثلا اگر وضعیت آب و هوایی در یکی از داده های ثبت شده، با بقیه نامتناسب باید، یا با وضعیت آب و هوایی آن فصل متناقض و عجیب باشد، داده ی پرت محسوب می شود. و باید آن را بررسی کنیم که آیا درست است یا خیر.

2- داده هایی که در یک ستون از یک نوع هستند، باید در یک رنج و یک فرمت باشند. اگر اینگونه نبود، باید آن ها را استاندارد سازی کنیم. برای مثال، اگر در یکی از داده های یک ستون خاص، مقدار مصرف دارو به واحد "سی سی" نوشته شده باشد، ولی در بقیه بر حسب "دوز" باشد، باید در همه یکسان شود. یا اگر وضعیت آب و هوایی در فیلد هایی بر حسب "درجه ی سانتی گراد" نوشته شده است، در بعضی بر حسب "فارنهایت" نوشته شده است و در بقیه به صورت "برفی، آفتابی، بارانی و ..." نوشته شده باشد، باید یک روند ثابت برای همه در نظر گرفته شود و همه را به یک صورت ثبت کنیم.

3- بعضی از داده ها ممکن است categorical باشند. باید بررسی کنیم که دسته ها به درستی ثبت شده باشند و از لحاظ منطقی با هم تناقض نداشته باشند. ممکن متوجه شویم که دسته بندی بهتری وجود دارد برای یک نوع داده، یا اینکه باید بعضی دسته ها را باید با هم ادغام کنیم، یا دسته ها را به دسته های ریز تری تفکیک کنیم. برای مثال ممکن است زمان مراجعات برحسب ماه ثبت شده باشد، و ما متناسب با هدف تحلیل مان، تصمیم بگیریم دسته بندیِ فیلدِ زمانِ مراجعه را، برحسب فصل انجام دهیم؛ چون می خواهیم دارو های فصل زمستان را تحلیل کنیم.

4- ممکن است لازم باشد عمل binning صورت بگیرد و متغیر های عددی، به بازه هایی تفکیک شوند. مثلا ممکن است لازم باشد سن مراجعه کنندگان را به رده ی سنی تبدیل کنیم.

5- ممکن است لازم باشد فیلدی برای ایندکس اضافه کنیم؛ تا با استفاده از آن بتوانیم رکورد ها را با شناسه های مختلفی که وجود دارد، به یکدیگر لینک کنیم و از آن استنتاج بهتری کنیم.

- فاز چهارم: modeling phase :

در این فاز مدلسازی را انجام می دهیم. با تکنیک های مختلف داده کاوی، یادگیری ماشین، خوشه

بندی و ... مدلسازی را انجام می دهیم. از بین الگوریتم های مختلف، بسته به اینکه تحلیل ما به کدام سمت باید برود، یک روش را انتخاب می کنیم و مدل را تشکیل می دهیم و سعی می کنیم train را انجام دهیم. باید مطمئن شـــویم که مدل ما از baseline بهتر عمل می کند. ســـپس با تحلیل های مناسب، مدلمان را fine-tune می کنیم، تا سود بیشتری از الگوریتم حاصل شود.
ما اینجا در مسئله مان، قصد داریم روند مصرف دارو را پیشبینی کنیم. می توانیم از prediction بهره بگیریم و بر اساس تحلیل سوابق، پیشبینی را انجام دهیم. در کنار آن، قبل از prediction می توانیم از classification و تحلیل های این چنینی بهره بگیریم و با توجه به هدفی که در تحلیل مان داریم، جزئیات مورد نیاز و تاثیرگذار بر پیشبینی را نیز تعیین کنیم. برای مثال می توانیم بر اساس وضعیت آب و هوایی، classification انجام دهیم و روند دارو های مصـــرف شـــده در هر دســـته را به تفکیک وضعیف آب و هوایی مشاهده و تحلیل کنیم، تا بر اساس آن ها پیشبینی صورت گیرد. یا مثلا می توانیم بر اساس نوع دارو، classification انجام دهیم و همان روند قبل را، این بار به تفکیک نوع دارو انجام دهیم، تا بر اساس آن ها پیشبینی صورت گیرد. این دسته اقدام ها، بستگی به جزئیات هدف مسئله مان دارد؛ که در فاز یک مشخص می شود.

- فاز پنجم: evaluation phase :

در این فاز باید مدلی که در فاز قبلی ایجاد کرده ایم را ارزیابی کنیم.
ما به هر ترتیبی می توانیم مدلی را برای مسـئله مان تهیه کرده باشـیم؛ اما کار ما زمانی ارزش دارد، که براســـاس الگوریتم ها و رویکرد های اســـتاندارد ارزیابی، خود را اثبات کند و عملکرد قابل قبولی داشته باشد.
به علاوه، باید بررسی کنیم که مدل ما، با تعریفی که برای مسئله در فاز اول داشتیم، همخوانی داشته باشد، و همان مسئله را حل کرده باشد. مثلا اگر هدف مشتری، پیشبینی مصرف دارو ها، به تفکیک نوع آن ها، در فاز اول برآورد شـــده بود؛ اینجا باید بررســـی کنیم، که همین تحلیل، به درســـتی و بدون سوبرداشت، صورت گرفته باشد.
همچنین باید نرخ خطا ها را اندازه گیری کنیم و تاثیر هر خطا روی تحلیل مان را برآورد کنیم. و در نظر داشته باشیم که خطاهای با خسارت زیاد و غیرقابل جبرانی رخ ندهد؛ و روند هایی را پیش بگیریم که خطاهای کم ارزش تری را ناشـی شـوند. مثلا فرض می کنیم که مشـتری ما هدفش از این تحلیل، این باشـد که می خواهد قبل از فصـل زمسـتان، دارو های مهم را به اندازه ی نیاز، تهیه و خریداری کند، تا در زمستان در اختیار مراجعه کنندگان قرار دهد. در این صورت، اگر ما میزان مصرف را اشتباهاً خیلی کمتر تحلیل کرده باشیم، دراین صورت، ممکن است داروهای تهیه شده توسط مشتری، خیلی زودتر به پایان برسـد و بیمار خاصـی که به آن دارو نیاز حتمی دارد، نتواند دارو را تهیه کند. یا از طرفی ممکن

است دارو ها را با درصد خطای بالایی، بیشتر از آنچه می شد، پیشبینی کرده باشیم. در این صورت، ممکن اســت با توجه به تاریخ مصــرف دارو ها، تعداد بالایی از آن ها بلااســتفاده بماند و با توجه به تاریخ مصرف آن ها، دیگر قابل استفاده نباشد. بنابراین، ما باید با توجه به ماهیت دارو ها، و بررسی خطاهای احتمالی، بهترین تحلیل، با حداقل خسارت را داشته باشیم. و بررسی کنیم که مثلا با توجه به ماهیت یک دارو و اهمیت آن، تعداد مورد نیاز مصرف آن را دست بالا درنظر بگیریم یا دست پایین؛ تا کمترین ضرر ناشی شود.
باید مدل هایی که داریم را ارزیابی کنیم و مدلی از بین مدل هایمان، که بهترین عملکرد را از خود نشان داد، به مرحله ی بعدی ببریم.

[تا این مرحله، در هر کدام از فاز ها، اگر ببینیم نیاز است؛ به فاز قبلی برمی‌گردیم و اصلاح می‌کنیم.]

- فاز ششم: deployment phase :

در این فاز، مدل را deploy می کنیم. قرار است بر اساس کار داده کاوی، مشتری کاری را انجام دهد و بر اساس نتایج داده کاوی، روال هایی را در شرکتش عوض کند. در این مرحله هم باید در کنار مشتری بمانیم و آثار پیاده سازی را در شرکت مشتری ببینیم. در این مرحله، نیاز است مسائلی را با محیط انطباق داد. بنابراین داده کاوی در هر مسئله ای، یک فرآیند در جریان و ادامه دار است.

.2

- مسئله اول: Association : بررسی شباهت میان مهارت ها و استعداد های مورد نیاز برای گذراندن دو درس پایگاه داده و ساختمان داده

- هدف: بررســی شــباهت میان مهارت ها و اســتعداد های مورد نیاز برای گذراندن دو درس پایگاه داده و ساختمان داده، با بررسی و تحلیل ارتباط میان رنج نمره ی دو درس پایگاه داده و ساختمان داده در بین دانشجویان
- گام ها: اطلاعات مربوط به دانشــجویان و نمرات دو درس پایگاه داده و ســاختمان داده ی آن ها را در یک دیتاست در نظر می گیریم.
  برای هر یک از دانشــجویان، بررســی می کنیم که آیا رنج نمره ی این دو درس، برای آن ها معمولا یکسان است یا خیر.
  اگر نتیجه ی این بررســی ها این بود که، غالباً رنج این دو نمره برای هر فرد یکســان اســت(یعنی مثلا معمولا افرادی که نمره ی پایگاه داده ی آن ها بالا است، نمره ی ساختمان داده ی آن ها نیز،

به همان نسبت بالا است. و بالعکس.)، متوجه می شویم که میان مهارت های مورد نیاز برای این درس ها، ارتباط مستقیم وجود دارد و احتمالا مهارت ها و استعداد هایی در یک راستا طلب می کنند.

- مسئله دوم: Estimation : تخمین معدل هر دانشجوی رشته ی کامپیوتر تا پایان تحصیل در دانشگاه
  - هدف: تخمین معدل کل هر دانشجوی در حال تحصیل رشته ی کامپیوتر، تا پایان تحصیل در دانشگاه، با تحلیل نمرات دروس گذرانده ی او تاکنون
  - گام ها: اطلاعات دانشجویان رشته ی کامپیوتر و نمرات آن ها در درس های مختلف گذرانده شان را از دیتاست مدنظر قرار می دهیم.

    دانشجویانی که قبلا فارغ التحصیل شده اند را در نظر می گیریم. در دیتاستی، برای این دسته از دانشجویان که قبلا فارغ التحصیل شده اند، معدل کل شان را در کنار نمرات درس های مختلف گذرانده شان قرار می دهیم.

    سپس با استفاده از این اطلاعات مربوط به این دسته از دانشجویان که فارغ التحصیل شده اند، به مدل یاد می دهیم که افراد با چه ریز نمراتی در چه دروسی، چه معدل کلی داشته اند.

    بنابراین، سیستمِ آموزش داده شده ی ما، حالا می تواند با استفاده از داده هایی که از ریز نمرات دانشجویان در حال تحصیل، در دسترس دارد، برای آن ها یک معدل کل متناسب تخمین بزند.

- مسئله سوم: Classification : قرار دادن دانشجویان رشته ی کامپیوتر، در دسته های مختلف گرایش
  - هدف: قرار دادن دانشجویان رشته ی کامپیوتر در دسته هایی به تفکیک گرایش، برحسب سوابق تحصیلی و نمرات آن ها در دروس مختلف
  - گام ها: اطلاعات دانشجویان رشته ی کامپیوتر و سوابق تحصیلی آن ها و نمرات آن ها در درس های مختلف گذرانده شان را از دیتاست مدنظر قرار می دهیم.

    دانشجویانی که قبلا انتخاب گرایش کرده اند را در نظر می گیریم. در دیتاستی، برای این دسته از دانشجویان که گرایش خود را تعیین کرده اند، گرایش را در کنار سوابق آن ها قرار می دهیم.

    سپس با استفاده از این اطلاعات مربوط به این دسته از دانشجویان که گرایش خود را تعیین کرده اند، به مدل یاد می دهیم که افراد با چه سوابق تحصیلی و چه ریز نمراتی، چه گرایشی را انتخاب کرده اند.

    بنابراین، سیستمِ آموزش داده شده ی ما، حالا می تواند با استفاده از داده هایی که از سوابق دانشجویان انتخاب گرایش نکرده، در دسترس دارد، آن ها را در دسته های مختلفِ گرایشیِ موجود، دسته بندی کند.

- مسئله چهارم: Prediction : پیشبینی نفرات برتر مشغول تحصیل در ترم پنجم رشته ی کامپیوتر، از نظر معدل، در همان نیم سال جاری

- هدف: پیشبینی نفرات برتر مشغول تحصیل در ترم پنجم رشته ی کامپیوتر، از نظر معدل، در همان نیم سال جاری ، با بررسی سوابق تحصیلی آن ها
- گام ها: اطلاعات همه ی دانشجویان رشته ی کامپیوتر و سوابق تحصیلی آن ها و نمرات آن ها در درس های مختلف گذرانده شان تا ترم پنجم را از دیتاست مدنظر قرار می دهیم.

دانشجویان در حال تحصیل در ترم ششم یا بالاتر یا فارغ التحصیل رشته ی کامپیوتر را در نظر می گیریم. از میان این دانشجویان، آن دانشجو هایی که در ترم پنجم خود، جز نفرات برتر بوده اند را جدا می کنیم؛ و اطلاعات مربوط به سوابق تحصیلی آن ها تا ترم پنجم شان را در دیتاستی قرار می دهیم.

سپس با استفاده از این اطلاعات مربوط به این دسته از دانشجویان که قبلا در ترم پنجم نظیر خود، جز نفرات برتر بوده اند، به مدل یاد می دهیم که افراد با چه سوابق تحصیلی و چه ریز نمراتی، در ترم پنجم جز نفرات برتر قرار گرفته اند.

بنابراین، سیستمِ آموزش داده شده ی ما، حالا می تواند با استفاده از داده هایی که از سوابق تحصیلی و ریز نمرات دانشجویان در حال تحصیل در ترم پنجم رشته ی کامپیوتر، در دسترس دارد، پیشبینی کند که کدام دانشجویان از میان آن ها، نفرات برتر این ترم، از لحاظ معدل خواهند بود.