پروژه برنامه نویسی

ساغر جانکی

۴۰۱۲۳۳۵۳

استاد قاسمی نیا

تابستان۱۴۰۴

موضوعی که من برای پروژه خودم انتخاب کردم درباره آزمایش خون و پارامترهاشه. البته من خیلی توی اینترنت گشتم تا داده پیدا کنم و آخر سر هم تونستم توی سایتkaggleداده پیدا کنم.

داده هایی که از این سایت پیدا کردم داده های بالینی و خونیه که شامل:

BMI,Chol,TG,HDL,LDL,Cr,BUN,Diagnosis

این داده ها به ترتیب شاخص توده بدن، کلسترول، تری گلیسرید، کلسترول خوب، کلسترول بد، کراتینین خون، اوره خون و بیمار بودن یا نبودن رو نشون میدن.

برای مورد آخری اگر کد صفر بود یعنی اون فرد سالمه و اگر ۱ بود یعنی بیماره.

من با استفاده از این داده ها میتونم ریسک ابتلا به بیماری قلبی و دیابت و عملکرد کلیه ها رو بررسی کنم.

چجوری؟

مثلا برای ریسک بیماری مزمن کلیه به کرآتینین خون، اوره خون، شاخص توده وسن نیاز دارم.

کرآتینین ماده زائد و محصول جانبی‌ایه که از متابولیسم عضلات به دست میاد. وقتی سطح این ماده داخل خون بالا میره نشون میده که کلیه‌ها دچار نارسایی هستند. برای همین سنجش کرآتینین برای بررسی عملکرد کلیه ها ضروریه. در ضمن این ماده خیلی سمیه.

(کرآتینین رو نباید با کراتین که یه پروتئینیه که داخل مو هست اشتباه بگیریم!)

خب حالا اوره چجوری به دست میاد؟ قبل از هر چیزی بگم که یه ماده ای هست به اسم آمونیاک. که آمونیاک خیلی خیلی سمیه به طوری که حتی یه ذرش توی خون تجمع پیدا کنه فوری باعث مرگ میشه. خب حالا راه حل چیه؟

کبد یه چرخه‌ای داره به اسم چرخه اوره. توی این چرخه آمونیاک رو به اوره تبدیل میکنه. میزان سمی بودن اوره خیلی کم تر از آمونیاکه. بعدش اوره میره و وارد خون میشه تا دفع بشه. این آمونیاک از متابولیسم پروتئین ها به دست میاد.

هم کرآتینین و هم اوره باید از طریق کلیه ها دفع بشن.

برای ریسک دیابت باید شاخص توده بدنی، نسبت تری گلیسرید به اچ دی ال(من سعی میکنم کم تر انگلیسی تایپ کنم چون تنظیمات ورد بهم میخوره) و سن برای ما مهمه. چونکه بی ام آی و سن بالا خطر ابتلا به دیابت نوع دوم رو افزایش میده.

و نسبت تری گلیسرید به اچ دی ال نشانه مقاومت به انسولین است. هر چقدر بالاتر باشه ریسک دیابت هم بالاتره.

در آخر برای بیماری قلبی ما به کلسترول، ال دی ال و اچ دی ال، تری گلیسرید، شاخص توده، سن و جنسیت نیاز داریم.

اول از همه بگم که ال دی ال رو بهش میگن کلسترول بد و در مقابل به اچ دی ال میگن کلسترول خوب. چونکه ال دی ال میاد توی رگ ها رسوب میکنه ولی اچ دی ال میاد از توی رگ برش میداره و مانع از تجمع توی رگ ها میشه.

خب حالا پروژه ای که من میخوام انجام بدم تحلیل ریسک این بیماری ها با استفاده از داده هاییه که دانلود کردم.

من برای اینکه بتونم کد بزنم از یه پروژه ای که توی گیت هاب بود استفاده کردم و از چت جی پی تی هم کمک گرفتم تا بتونم کد ها رو متناسب با داده های خودم بزنم.

واقعا واقعا سخت بود حتی با حضور چت جی پی تی. چون چند روز من فقط این کدها رو میزدم و همش ارور میداد. حتی بعضی وقتا نمیدونستم برای چی ارور میده. برای همین مجبور بودم ارورهایی که میده رو از چت جی پی تی بپرسم و اونا رو رفع کنم. بعضیاشون حتی بعد از اینکه از چت جی پی تی پرسیدم و بازم ارور میداد.

توی پروژم چندتا شاخص رو محاسبه کردم:شاخص سلامت کلیه ها و محاسبه نسبت چربی ها.این نسبت ها توی پزشکی خیلی مهم هستند. و داخل پروژه شاخص های ابتلا به این بیماری ها رو حساب کردم و گفتم که مثلا دیابت توی خانم ها و آقایان احتمال ابتلا بهش چقدره. در آخر هم محاسبه کردم که مثلا تعداد آقایونی که احتمال ریسک قلبی بالا دارند چقدره.

خب بریم سراغ خود پروژه. من همونطور که قبلا گفتم داده های مربوط به پروژم رو از سایتkaggleگرفتم. لینکش هم این پایین میذارم.

https://www.kaggle.com/datasets/simaanjali/diabetes-classification-dataset

قبلا هم گفتم چه پارامترهایی داخلش هست و به چه درد ما میخوره پس دیگه نمیگم بریم سراغ کدهای پروژه.

خب من کد اولیه رو از یه پروژه توی گیت هاب برداشتم. پروژش درباره داده های بالینی توی همون سایتی که خودم گفتمه. با این داده ها میاد بیماری مزمن کلیه رو پیش بینی میکنه. مراحلش اینجوریه که پیش پردازش داده ها، پاک سازی، آموزش مدل های یادگیری ماشین مثل درخت تصمیم گیری و رگرسیون لجستیک داره ولی من نمیخواستم که تمام این کارا رو انجام بدم و فقط میخواستم که با استفاده از همون چیزایی که سر کلاس یاد گرفته بودم انجامش بدم برای همین مجبور شدم که اینجا از چت جی پی تی کمک بگیرم.

یکی دو روز اول هر چی کد میزدم همش ارور میداد. ولی بالاخره روز سوم تونستم یه کدی بزنم که ارور نده و همش ران بشه اما بعدش دوباره مشکل داشتم.

توی خروجی باید برای هر بیماری مثلا کاردیو قلب جنسیت رو مشخص میکرد و تعداد افرادی که درصد بالا، متوسط و پائین برای این مشکل داشتند رو مشخص میکرد ولی این کار رو انجام نداده بود و برای کل بیمارا حساب کرده بود و جنسیت رو همunknown

زده بود. از چت جی پی تی که پرسیدم چرا ارور میده گفت ممکنه از این باشه که بعضی ستون‌ها توی ستون جنسیت جاشون خالی باشه حالا یا گم شده باشه یا نذاشته باشن. که من رفتم چک کردم و دیدم که همش مشخص شده.

وسط این چیزی که دارم میگم یه موضوعی یادم اومد. چندتا پروژه که داخل گیت هاب دیدم، فهمیدم که باید از تابع()fillna استفاده کنیم که از کتابخونهpandaهست. این تابع

میاد مقادیر گمشده رو پر میکنه. من نیازی بهش نداشتم چون وقتی که چک کردم هیچ مقدار گمشده یا خالی ندیدم.

خب برین سراغ موضوع خودمون. یه ایراد دیگه هم که گفت ممکنه این باشه که داخل ستون داده‌ها حروف بزرگ و کوچیک باشه یا اولش با یه فاصله شروع شده باشه یا غیر از اف و ام (منظورش زن و مرده) توی داده ها یه چیز دیگه نوشته که من بازم هر دوتای اینا رو چک کردم و مشکلی نبود. بعدش رفتم یه کدی زدم تا ببینم کلا چند تا از داده‌ها رو به صورت ناشناخته نشون میده. کدش رو این پائین میذارم.

```
print(df['Gender'].value_counts(dropna=False))
```

و دیدم تمام داده ها رو به صورت ناشناخته نشون میده. چت جی پی تی گفتش که مشکل از پاکسازی یا تبدیل اولیه داده‌هاست نه خود داده.

بعدش گفتم شاید من درست ننوشتم ستون ها رو. گفتم یه کدی بزنم که اسم ستون‌ها رو بهم بگه من ببینم شاید درست شد. کد زیر رو زدم:

```
print(df.columns.tolist())
```

ستون ها رو هم بهم نشون داد و دیدم که درست زدم. چت جی پی تی هم همش بهم میگفت که مشکل از اینه که تو اسم ستون رو درست ننوشتی ولی خب من درست نوشته بودم بعد گفتش که باید داده‌ها رو پاکسازی کنم که کدی که بهم گفته بود زدم و بازم همون نتیجه بود. بعدش یه چیز دیگه بهم گفت که اونو زدم کلا درصد حساب نکرد مشکل قبلی هم برطرف نشد. آخرش مجبور شدم از یکی کمک بگیرم و دیگه درست شد. تمام این کارا حدود یک و نیم روز وقت منو گرفت ولی آخرش درست شد البته فکر میکردم چت جی پی تی کمک میکنه که نکرد.

آها اینم بگم که من موقعی که میخواستم پروژه رو شروع کنم یه صفحه پایتون که درست کردم یه پوشه هم درست کردم و فایل داده ها که به صورتCSVبود رو داخلش کپی

پیست کردم. برای اینکه یه کاری کنم تا داده ها رو بخونه هم باز مشکل داشتم. چون از روی پروژه اصلی هم نتونستم انجامش بدم. ولی بالاخره تونستم انجامش بدم.

خب بعدش میخواستم که یهsearch engineدرست کنم که مثلا بهش بگم زن بیمار

بالای ۵۰ سال و کسایی که این ویژگی رو دارن برام پیدا کنه. برای نوشتن این قسمت به مشکل خاصی برنخوردم.

یه مشکل دیگه ای هم که باهاش مواجه بودم این بود که وقتی کد رو ران میکردم ارور

FutureWarning بهم نشون میداد و من اصلا نمی دونستم که چی هست.

بعدش توی اینترنت سرچ کردم که ببینم چیه. یه سایت پیدا کردم که ننوشته بود چیه ولی گفته بود که یه هشدار کوچیکه و اصلا نیازی به نگرانی نیست میتونید نادیدش بگیریم تا نمایش داده نشه. یه کدی هم داده بود که بزنیم تا نادیده بگیرتش. این پائین لینک سایته رو میذارم:

https://stackoverflow.com/questions/64426905/how-to-resolve-future-warning-errors

توی پروژه یه قسمت دیگه هم هست که من برای اینکه ریسک هر کدوم از بیماری ها رو محاسبه کنم به هر شاخص یه عددی دادم. بهش میگم وزن یعنی دارم مشخص می کنم کدوم شاخص توی محاسبه اون ریسک مهم تره. مثلا برای محاسبه برای کلیه ها، کرآتینین و اوره خیلی مهم‌تر از شاخص توده بدنی و سنه. برای کرآتینین ۴۵/.، اوره ۰/۳۵ ، شاخص توده بدنی ۰/۱۵ و برای سن که خیلی کم مهمه ۰/۰۵. گذاشتم. پس وقتی که دارم امتیاز نهایی

رو حساب میکنم اگه کرآتینین یک فرد بیشتر باشه تاثیرش روی نمره ریسک خیلی بیشتر از سن اون بیماره.

یه سوال اینجا پیش میاد که برای خودمم پیش اومد. اینکه این عددها رو بر چه اساسی میذاریم یعنی میشه هر عددی گذاشت یا باید از روی یه چیز خاص بگیم؟

جوابش اینه که این وزن ها به صورت فرضی انتخاب شدن یعنی خود این اعداد توی مقاله پزشکی یا مدل علمی خاصی نیومده. من از یه مقاله که مربوط به پیش بینیCKDبود کمک گرفتم. توی این مقاله هم این اعداد نبود. این مقاله گفته بود که کدوم شاخص مهم‌تره.

اینم لینک مقاله:

https://bjgp.org/content/bjgp/62/597/e243.full.pdf

برای اینکه این اعداد رو بذاریم، نمیتونیم هر عددی رو بذاریم. بهتره که مجموع اعداد ۱ بشن و نکته دیگه اینه که این اعداد باید منطقی باشن.

برای دو بیماری بعدی هم به همین صورت عدد گذاری کردم. برای دیابت به ترتیب چربی بدن، شاخص توده بدنی، نسبت تری گلیسرید بهHDLو سن مهمه. توی کاردیو هم به ترتیب نسبتLDL به HDL، نسبت تری گلیسرید به HDL، کلسترول و سن مهمه.

سوال بعدی باز برای خودم هم پیش اومد موقع زدن کد باهاش درگیر بودم. اینکه حالا من درصد و همه چیز رو حساب کردم تا چه درصدی رو می‌تونم توی گروهlow,mediumو highقرار بدم. وقتی میخوایم داده‌ها رو به ۳ قسمت تقسیم کنیم باید شکستگی ها رو جایی بذاریم که حدود یک سوم پائین، یک سوم وسط و یک سوم بالا باشن. به زبان آماری می‌شه صدک۳۳ و صدک۶۶. چرا ۳۳/. و ۶۶/. رو من انتخاب کردم چون باید سه بخش مساوی ایجاد بشه. یعنی اگر ۴ تا بخش بود می‌شد: ۰/۲۵، ۰/۵ و ۰/۷۵

پس اینطوری میشه که هر کسی که امتیازش تا ۳۳٪ باشه توی گروه ریسک کم قرار میگیره، هرکسی هم که بین ۳۳٪ تا ۶۶٪ باشه توی گروه متوسط و کسی هم که بالاتر از ۶۶٪ باشه توی گروه ریسک زیاد قرار میگیره.

این قسمت رو من خودم نمیدونستم موقع کد زدن و از چت جی پی تی پرسیدم.

خب حالا ممکنه این سوال پیش بیاد که توی یکی از ستون ها مشخص کرده که فرد بیمار یا سالمه چه نیازی به این همه کد که زدی هست فقط یهsearch engineانجام میدادی و دیگه نیازی به بقیش نبود. در جواب باید بگم که من چند تا مورد رو محاسبه کردم. اولین موردش شاخص ریسکه که نشون میده شدت یا احتمال خطر چقدر بالاست. مورد بعدی اینه که من دسته بندی کردم. همه بیمارها رو بر اساس بیمال تو سه تا گروه قرار دادم:

Low, medium, high

اینا گروه های ریسک مختلف هستند که میتونه برای افرادی که سالم هستند مفید باشه یعنی توی جدول زده طرف سالمه ولی وقتی که این شاخص‌ها رو نگاه میکنه میبینه که در معرض خطر اون بیماری قرار داره و ریسک ابتلا به اون بیماری رو داره.

این کا رو با استفاده از search engineکه ساختم میتونیم انجام بدیم. اینجوری که مثلا بهش میگیم مرد سالم با ریسک قلبی بالا و همه اونایی که سالمن ولی شاخص ریسکشون توی دستهhigh یا حتی mediumقرار میگیره رو به ما نشون میده.

اینم لینک پروژه گیت هابی که ازش استفاده کردم ولی همونطور که گفتم پروژش خیلی سنگین بود و کارایی انجام داده بود که من نمیخواستم برای همین از چت جی پی تی کمک گرفتم در کل باید صادقانه بگم که همش رو خودم ننوشتم.

https://github.com/salsabiltanjim/Chronic-Kidney-Disease-Prediction

برای قسمتsearch engineهم از یه پروژه پزشکی بود الهام گرفتم و با کمک چت جی پی تی و یه سری اطلاعاتی که خودم داشتم انجامش دادم.

https://github.com/Gedichter/medical_record_search_engine

آخر سر هم بگم که خیلی موضوع عوض کردم؛ کلی دیتا از دیتاست های مختلف دانلود کردم؛ سایت های مختلف رو نگاه کردم؛ چند تا پروژه توی گیت هاب دیدم و پوشه های مختلف داخلشو دیدم تا فهمیدم کد اصلی پروژه کجاست و در نهایت به این موضوع و دیتا و کدها رسیدم.