Big Data Fraud detection using multiple medicare data sources

# چکیده

# مقدمه

# مجموعه داده

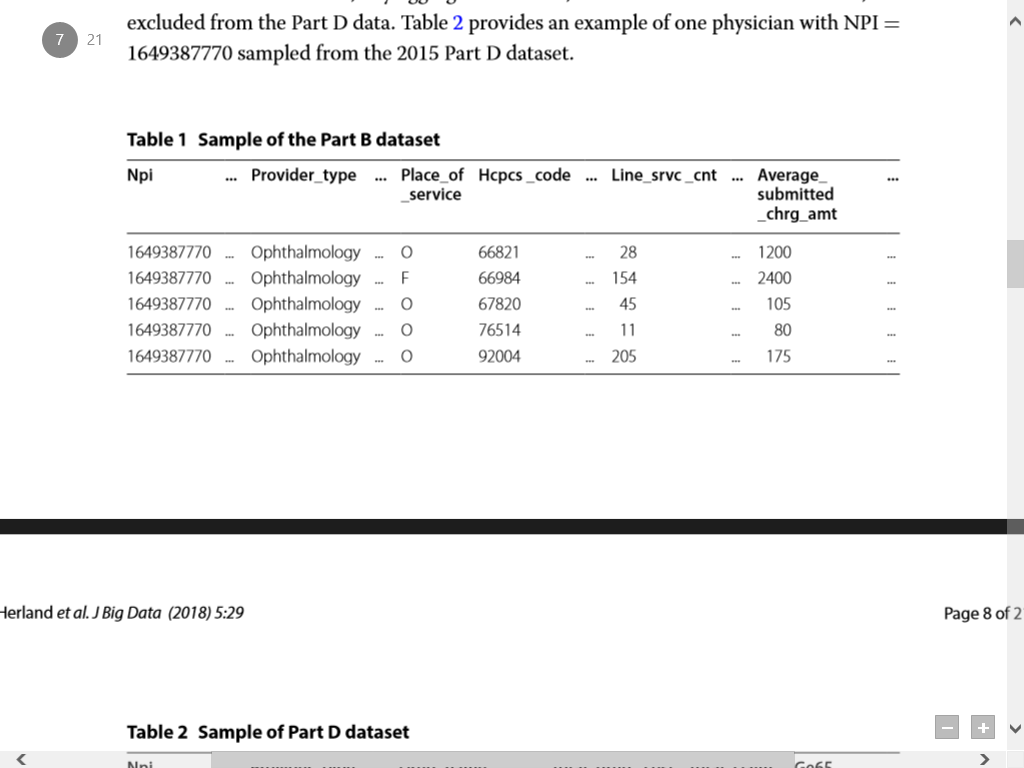
در این بخش ما مجموعه داده ی CMS را که استفاده نمودیم، توصیف میکنیم. بعلاوه روش پردازش داده ی استفاده شده برای ایجاد هر مجموعه داده، شامل پردازش، نگاشت برچسب تقلب بین مجموعه های داده مراقبت سلامت و LEIE، و one-hot encoding برای متغییرهای طبقه ای بحث شده است. اطلاعات در هر مجموعه داده بر اساس داده ی مطالبات ناظر CMS برای ذینفعان مراقبت سلامت ثبت نام شده در برنامه رایگان پرداخت برای خدمات[[1]](#footnote-2) می باشد. توجه داشته باشید که این داده اطلاعات هر مطالبه ی ثبت شده در برنامه ی سود مراقبت سلامت [[2]](#footnote-3) را به حساب نمی آورد. از آنجایی که رکوردهای CMS ،اطلاعات مطالبات بعد از پرداخت صورت گرفته است، ما فرض می کنیم داده ی مراقبت سلامت در حال حاضر پاکسازی شده و درست است. توجه داشته باشید که NPI در گام داده کاوی استفاده نشده است، بلکه برای تجمیع و شناسایی استفاده شده است. بعلاوه برای هر مجموعه داده ، یک متغییر سال که برای تجمیع و شناسایی نیز استفاده شده است را هم اضافه می کنیم.

# توصیف مجموعه داده ی مراقبت سلامت

### Part B

مجموعه داده ی بخش B اطلاعات مطالبات برای هر روند(نسخه) که یک پزشک در یک سال انجام داده را فراهم میکند. اخیرا این مجموعه داده در وبسایت CMS از سال 2012 تا 2015 در دسترس است. پزشکان با استفاده از NPI[[3]](#footnote-4) یکتا شناسایی شده اند، در حالیکه روندها بر اساس کد [[4]](#footnote-5)HCPCS برچسب زده می شوند.سایر اطلاعات مطالبات شامل میانگین هزینه و پرداخت ها، تعداد روندهای انجام شده و تخصص پزشکی(که بعنوان نوع ارایه دهنده نیز شناخته میشود) میباشد.CMS تصمیم گرفت که داده ی بخش B را تجمیع کند با:

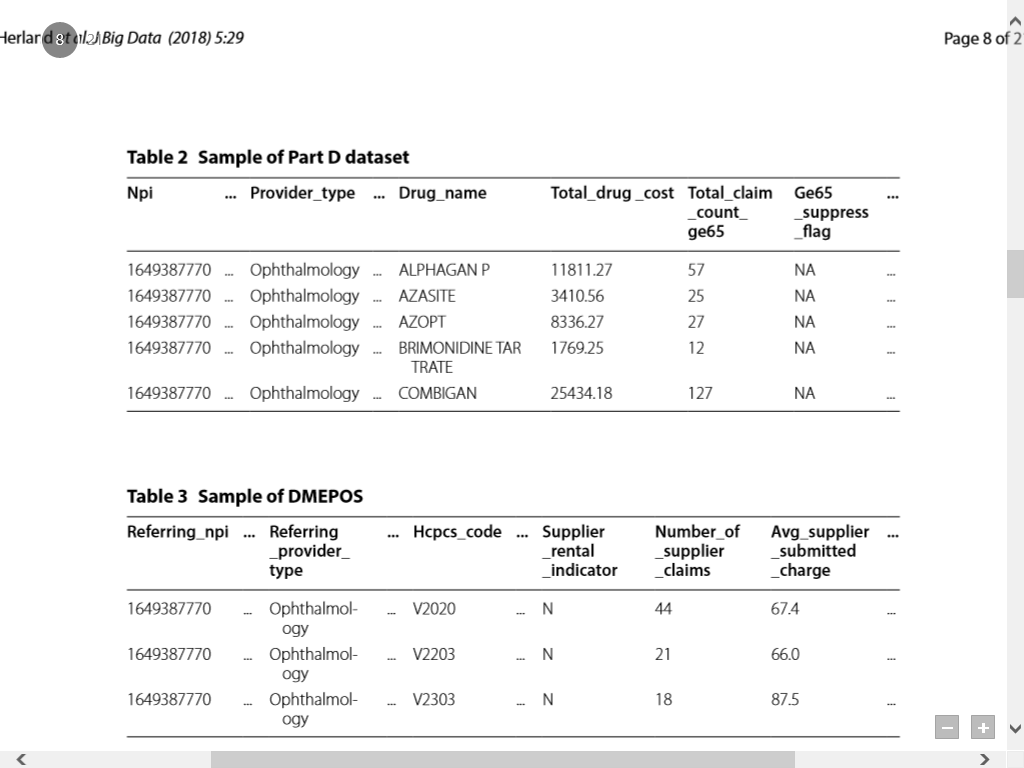
1)NPI مربوط به ارایه دهنده ی خدمات 2)کد HCPCS برای روند یا سرویس انجام شده 3) محل خدماتی که یک تسهیلات است([[5]](#footnote-6)F) یا غیرتسهیلات(O[[6]](#footnote-7))، به ترتیب مانند یک بیمارستان یا یک مطب. هر سطر در مجموعه داده شامل NPI یک پزشک، نوع ارایه دهنده،کد HCPCSتقسیم شده بر اساس محل خدمات همراه با اطلاعات خاص منطبق بر این تقسیم بندی(مانند تعداد مطالبات)، و سایر ویژگی های تغییر نیافتنی(مانند جنسیت). ما دریافتیم که در عمل جراحی، پزشکانی هستند که چه در بیمارستان یا در مطب شان روندهای مشابهی انجام میدهند، همچنین تعداد کمی از پزشکان که تحت چند نوع ارایه دهنده(تخصص) مانند متخصص داخلی و متخصص قلب عمل انجام میدهند. بنابراین برای هر پزشک تعداد زیادی سطر مانند ترکیب های یکتایی از NPI، نوع ارایه دهنده، کد HCPCS، و محل خدمات وجود دارد و بنابراین داده ی بخش B میتواند برای فراهم نمودن اطلاعات سطح روند در نظر گرفته شود. جدول 1 مثالی از یک پزشک با NPI=1649387770 نمونه گیری شده از مجموعه داده ی بخش B سال 2015 میباشد.



### Part D

مجموعه داده ی بخش D اطلاعات مربوط به تجویز داروهایی است که تحت برنامه تجویز دارو بخش D مراقبت سلامت در یک سال تهیه میشود. اخیرا این داده در وبسایتCMS از سال 2013 تا 2015 در دسترس است. پزشکان با استفاده از NPI یکتایشان در داده شناسایی میشوند در حالیکه هر دارو توسط برند و نام عمومی شان برچسب زده میشود. سایر اطلاعات شامل میانگین هزینه و پرداخت ها ، متغییرهای توصیف کننده ی مقدار داروی تجویزشده و تخصص پزشکی میباشد.CMS تصمیم گرفت که داده ی بخش D را تجمیع کند با:

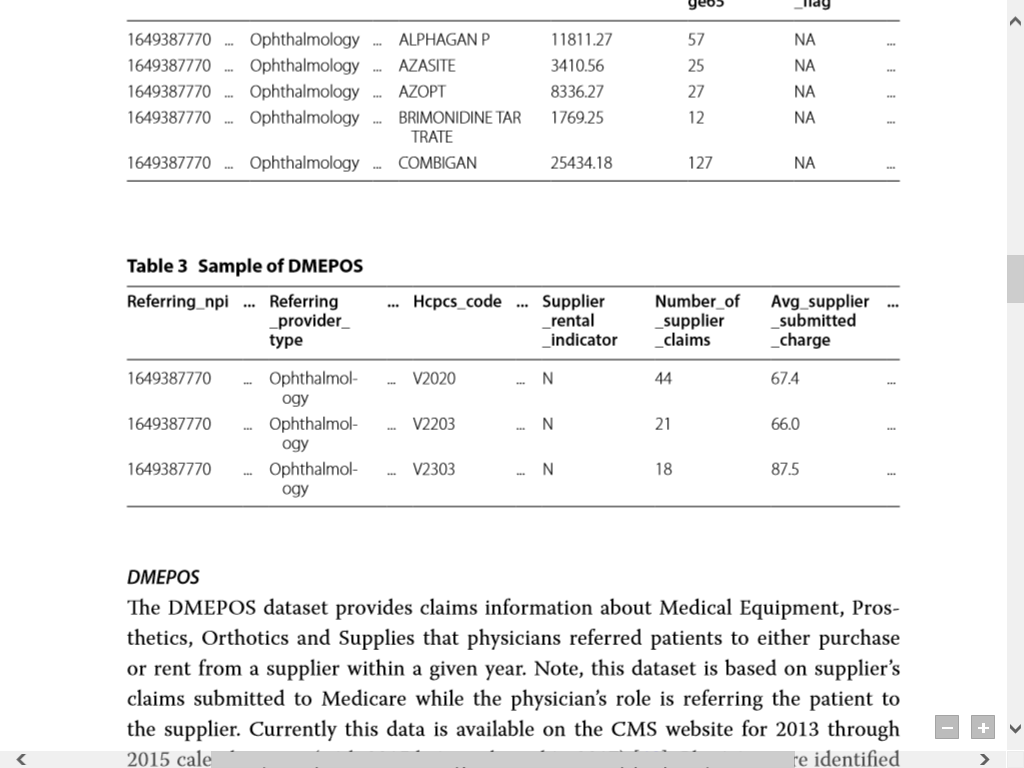
1)NPI مربوط به تجویزکننده 2) نام دارو(نام برند در مورد داروهای بازرگانی) و نام عمومی. هر سطر در مجموعه داده ی بخش D ،NPI مربوط به پزشک،نوع ارایه دهنده و نام دارو به همراه اطلاعات خاص منطبق بر این تقسیم بندی، را لیست می کند. مانند بخش B ما دریافتیم که تعدادی از پزشکان تحت چند تخصص مانند متخصص داخلی و قلب عمل میکنند. بنابراین برای هر پزشک تعداد زیادی سطر مانند ترکیب های یکتایی از NPI، نوع ارایه دهنده، نام دارو، نام عمومی وجود دارد و در نتیجه داده ی بخش D میتواند برای فراهم نمودن اطلاعات سطح روند درنظر گرفته شود. برای محافظت از حریم خصوصی ذینفعان مراقبت سلامت، هر رکورد تجمیع شده که از 10 مطالبه یا کمتر مشتق شده است، از داده ی بخش D خارج شده است. جدول2 مثالی از یک پزشک با NPI=1649387770 نمونه گیری شده از مجموعه داده ی بخش D سال 2015 می باشد.



### DMEPOS

مجموعه داده DMEPOS اطلاعات مطالبات درباره ی تجهیزات پزشکی،پروتز،ارتوتیک و تجهیزاتی که پزشکان به بیماران ارجاع میدهند که بخرند یا از یک تهیه کننده اجاره کنند در یک سال.توجه داشته باشید که این مجموعه داده بر اساس مطالبات فراهم کنندگان ارسال شده به بیمه پزشکی است، در حالیکه نقش پزشک ارجاع بیمار به فراهم کننده است. اخیرا این داده در وبسایت CMS از سال 2013 تا 2015 در دسترس است. پزشکان با استفاده از NPI یکتایشان در داده شناسایی شده اند ، درحالیکه محصولات بر اساس کد HCPCS برچسب زنی میشوند. سایر اطلاعا شامل میانگین هزینه و پرداخت ها، تعداد خدمات/محصولات اجاره داده شده یا فروخته شده و تخصص پزشکی میباشد.CMS تصمیم گرفت که داده ی بخش B را تجمیع کند با :

1)NPI مربوط به ارایه دهنده ی خدمات 2)کد HCPCS برای روند یا خدمات انجام شده توسط فراهم کننده DMEPOS و3) شاخص اجاره (مقدار’y’ یا ‘N’) مشتق شده از مطالبات فراهم کننده DMEPOS. هر سطر، NPI یک پزشک، نوع ارایه دهنده، یک کد HCPCS تقسیم شده به اجاره یا غیراجاره همراه با اطلاعات خاص منطبق بر این تقسیم بندی (مانند تعداد مطالبات فراهم کننده) و سایر ویژگی های غیرقابل تغییر(مانند جنسیت) را فراهم میکند. ما دریافتیم که برخی پزشکان برای تجهیزات DMEPOS یکسان، یا کد HCPCS یکسان، ارجاعات را بعنوان اجاره و غیراجاره قرار میدهند. همچنین برخی پزشکانی که اعمال جراحی تحت چند تخصص مثلا داخلی و قلب انجام میدادند. بنابراین برای هر پزشک، سطرهای زیادی مانند ترکیب NPI، نوع ارایه دهنده،کدhcpcs و وضعیت اجاره وجود دارد و بنابراین داده ی DMEPOS همچنین میتواند برای فراهم نمودن اطلاعات سطح روند در نظر گرفته شود. جدول 3 مثالی از یک پزشک با NPI=1649387770 از مجموعه داده ی بخش B سال 2015 است.

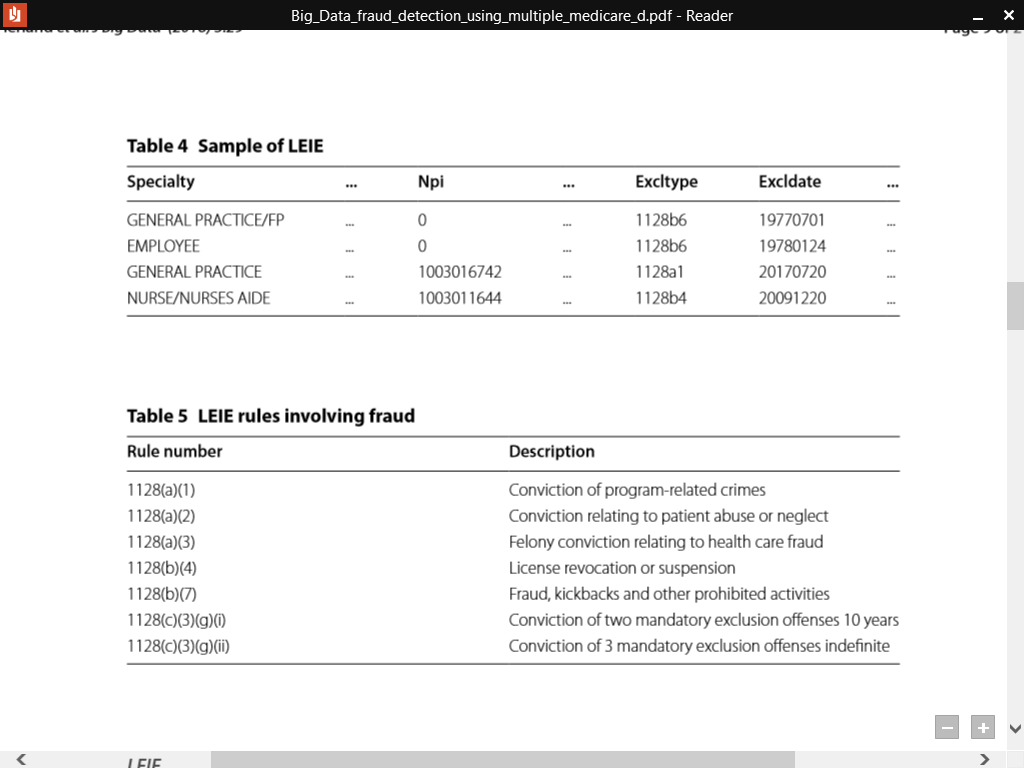


### LEIE

برای دستیابی صحیح به کارایی تشخیص تقلب، همانگونه که در دنیای واقعی عمل میکند، ما به یک منبع داده نیاز داریم که شامل پزشکانی باشد که مرتکب تقلب در دنیای واقعی شده باشند.بنابراین لیستی از اشخاص و موجودیت های اخراج شده را به کار میگیریم که شامل اطلاعات زیر است:

دلیل اخراج، تاریخ اخراج، تاریخ بازگردانی/ابطال برای همه ی پزشکان نامناسب شناخته شده برای عمل پزشکی و بنابراین اخراج از عمل در امریکا برای یک بازه زمانی داده شده.

این مجموعه داده ایجاد شده و ماهیانه توسط اداره بازرسی عمومی (OIG[[7]](#footnote-8))مطابق با بخش 1128و 1156 قانون امنیت اجتماعی نگه داری میشود. OIG اختیار محرومیت افراد از برنامه های مراقبت سلامت فدرالی مانند بیمه پزشکی را داراست. متاسفانه LEIE فراگیر نیست و 38% ارایه دهندگان با محکومیت تقلب به جراحی ادامه میدهند و 21% با وجود محکومیت از عمل جراحی تعلیق نشده اند.بعلاوه مجموعه داده LEIE فقط شامل مقادیر NPI برای درصد کمی از پزشکان و موجودیت هاست. مثالی از چهار پزشک مختلف و اینکه چگونه در LEIE به تصویر کشیده شده اند در جدول 4 نشان داده شده است، که هر پزشک بدون NPI مقدار 0 را دارد.LEIE در سطح ارایه دهنده تجمیع شده و اطلاعات خاصی با توجه به روندها، داروها یا تجهیزات مرتبط با فعالیت های کلاهبردارانه ندارد. دسته های مختلفی از محرومیت/ اخراج بر اساس شدت گناه وجود دارد که توسط شماره قوانین توصیف شده اند. ما از همه ی محرومیت ها استفاده نمیکنیم، بلکه ارایه دهندگان محروم شده را با قوانین انتخاب شده ی نشان دهنده ی ارتکاب تقلب فیلتر میکنیم. جدول 5 این قوانین را که منطبق بر محرومیت ارایه دهندگان کلاهبردار است و طول محرومیت اجباری را میدهد.ما تعیین کرده ایم که هر رفتاری که قبل از پایان تاریخ محرومیت اجباری یا در طول آن است، تشکیل دهنده ی تقلب است.



## 

## پردازش داده

برای هر مجموعه داده، ما اطلاعات را برای همه ی سال های در دسترس ترکیب کردیم. برای تحقیق ما مجموعه داده ی بخش B از سال 2012 تا 2015 در دسترس بود، درحالیکه بخش D و DMEPOS برای سال های 2013 تا 2015 در دسترس بودند. برای بخش B و DMEPOS گام اول حذف همه ویژگی هایی که در سال های در دسترس وجود نداشتند بود. مجموعه داده ی بخش D در تمام سال های در دسترس ویژگی های یکسان داشت. برای بخش B متغییرهای انحراف از معیار را از سال های 2012 و 2013 حذف کردیم و متغییرهای پرداخت را از سال های 2014و 2015 استانداردسازی نمودیم، زیرا در سال های دیگر در دسترس نبودند. برای DMEPOS یک متغییر انحراف از معیار را از سال 2014 و2015 حذف نمودیم زیرا در 2013 در دسترس نبوده است.برای همه ی سه مجموعه داده، ما همه ی نمونه هایی که هردوی مقادیر NPI وHCPCS/نام دارو از دست رفته، یا NPI نامعتبر داشتند را حذف کردیم. برای بخش B ما همه ی نمونه ها با کدهای HCPCS ارجاع شده به نسخه ها را فیلتر کردیم.این کدهای مرتبط با نسخه، روندهای پزشکی واقعی نیسند. اما در عوض برای خدمات خاص لیست شده در فایل متوسط قیمت فروش داروی بخش B بیمه هستند.نگه داشتن این نمونه ها نتایج را گل الود میکند، مانند ویژگی line-srvc-cnt در اینجا نشان دهنده ی وزن یا مقدار دارو است، به جای تعیین ساده ی تعداد روندها. در این تحقیق ما فقط به ویژگی های خاصی از هر مجموعه داده علاقمندیم ، تا پایه ای محکم برای ازمایشات و تحلیل هایمان فراهم آوریم. برای مجموعه داده ی بخش B هشت ویژگی را نگه داشتیم در حالیکه 22 ویژگی را حذف کردیم. برای مجموعه داده بخش D هفت ویژگی را نگه داشتیم و 14 ویژگی را حذف کردیم. برای مجموعه داده DMEPOS نه ویژگی را نگه داشتیم و 19 ویژگی را حذف کردیم. اطلاعات حذف شده اطلاعات خاصی درمورد مطالبات، داروهای تجویزشده یا مرجوعی ارایه نمیکرد بلکه شامل اطلاعات مربوط به ارایه دهنده مانند محل و نام هستند و همچنین متغییرهای اضافی مانند توصیف متنی که میتواند با استفاده از متغییرهایی که شامل روند یا کد دارو هستند نشان داده شود.برای بخش D ما همچنین متغییرهایی که تعداد و اطلاعات پرداخت برای بیماران بالای 65 سال را در نظر نگرفتیم زیرا این اطلاعات در متغییرهای نگه داشته شده وجود دارند. در اینجا متغییر تعداد مطالبات(total\_claim\_count) شامل تعداد همه ی سنین و در برگیرنده ی سن بالای 65 میباشد.جدول 6و7و8 جزییات ویژگی هایی که از مجموعه داده انتخاب کردیم، شامل یک توصیف و نوع ویژگی(عددی یا کیفی) همراه با ویژگی های محرومیت(برچسب تقلب) مشتق شده از LEIE میباشد. مراحل پردازش داده برای بخشB، بخشD و DMEPOS یکسان است. همه ی سه مجموعه داده ی تغییر ناپذیر در اصل در سطح HCPCS یا روند هستند، به این معنی که با NPI وHCPCS/دارو تجمیع شده اند. برای مواجه شدن با نیازهای نگاشت برچسب تقلب با استفاده از LEIE ، ما هر مجموعه داده را مجددا میزان میکنیم، تجمیع می کنیم به سطح ارایه دهنده که همه ی اطلاعات از روی NPIگروه بندی و تجمیع شود. برای بخش B فرایند تجمیع شامل گروه بندی تاریخ با استفاده از NPI ، نوع ارایه دهنده، جنسیت و سال، تجمیع با HCPCS و محل خدمات است. بخشD توسط NPI، نوع ارایه دهنده و سال گروه بندی شد و تجمیع با داروها.DMEPOS توسط NPI و نوع ارایه دهنده، جنسیت و سال گروه بندی و تجمیع با HCPCS و وضعیت اجاره است. برای مجموعه داده بخش D و DMEPOS ، تعداد ذینفعان اگر کمتر از 11 باشد به 0 کاهش می یابد و در پاسخ مقدار 5 را نسبت دادیم، به توصیه CMS . در تلاشی برای گذشتن از ازدست رفتن اطلاعات به دلیل تجمیع این مجموعه داده ها، شش ویژگی عددی تولید کردیم برای هر ویژگی عددی انتخاب شده در زیربخش قبلی برای هر مجموعه داده . بنابراین برای هر مقدار عددی، هر سال، در هر مجموعه داده، متغییر عددی اصلی را با مجموع میانگین، مجموع، میانه، انحراف معیار، بیشینه و کمینه مقادیر جایگزین کردیم، شش ویژگی برای هر ویژگی عددی تولید شد. ویژگی های حاصل همگی کامل هستند به جز انحراف معیار که شامل مقادیر NA است. مقادیرNA وقتی تولید میشوند که یک پزشک یک دارو/کد HCPCS را یک بار در سال انجام میدهد یا تجویز می کند. بنابراین انحراف معیار جمعیت برای یک نمونه واحد 0 ست و بنابراین همه ی مقادیرNA را با صفر نشان دادیم که این نمونه ی تنها هیچ تغییری در آن سال خاص ندارد. دو ویژگی دیگر ویژگی های کیفی هستند:نوع ارایه دهنده و جنسیت.(بخش D شامل متغییر جنسیت نیست.)

مجموعه داده ی ترکیبی

مجموعه داده ترکیبی بعد از پردازش مجموعه داده ی بخشB، بخش DوDMEPOS ایجاد شد، شامل همه ی ویژگی ها از هریک همرا با برچسب تقلب مشتق شده ازLEIE است. فرایند ترکیب شامل یک عملیات الحاق روی NPI، نوع ارایه دهنده و سال است. به دلیل اینکه متغییر جنسیت در مجموعه داده ی بخش D وجود ندارد، این متغییر را در شروط عملیات الحاق شامل نکردیم و از برچسب های جنسیت از بخش B استفاده کردیم درحالیکه برچسب های جنسیت جمع اوری شده از مجموعه داده DMEPOS را پس از الحاق حذف کردیم. در ترکیب این مجموعه داده ها ما به پزشکانی که در همه ی سه بخش مراقبت سلامت شرکت داشتند محدود شدیم. این مجموعه داده ی ترکیب شده پایه ی فراگیرتر و بزرگتری از ویژگی ها برای به کارگیری الگوریتم های داده کاوی برای تشخیص رفتارهای کلاهبردارانه دارد.

## برچسب زنی **تقلب**

## **One-hot encoding**

# **روش ها**

## **یادگیرنده**

برای اجرا و ارزیابی مدل ها، به دلیل حجم زیاد مجموعه داده از اسپارک بر روی یک خوشه ی Yarn هدوپ استفاده کردیم. ما از سه مدل کلاس بندی قابل دسترس در کتابخانه یادگیری ماشین آپاچی اسپارک2.3.0 استفاده نمودیم. رگرسیون منطقی، درخت های Gradient Boosted و جنگل تصادفی. در این بخش ما به اختصار هر یادگیرنده را توصیف میکنیم و به تغییرات تنظیمات که از تنظیمات پیش فرض متفاوت است توجه داشته باشید.

رگرسیون منطقی احتمالات را برای اینکه یک متغییر کیفی وابسته به چه کلاسی تعلق دارد پیش بینی میکند، با استفاده از مجموعه ای از متغییرهای مستقل که یک تابع منطقی را به کار میگیرند. رگرسیون منطقی از یک تابع سیگموید برای تولید مقادیری که میتواند بعنوان احتمالات کلاس ها تفسیر شود، استفاده میکند. رگرسیون منطقی شبیه رگرسیون خطی است اما از یک کلاس فرضیه برای پیش بینی کلاس عضویت استفاده میکند. ماتریس مرز برای تطبیق شکل داده (تعداد کلاس ها و ویژگی ها) تنظیم شد بنابراین الگوریتم تعداد کلاس ها و ویژگی های مجموعه داده را میداند. اندازه بردار مرز برای رگرسیون دوجمله ای برابر 1 است، و برای کلاس بندی دوتایی آستانه ای تنظیم نمیشود. جنگل تصادفی یک روش یادگیری گروهی[[8]](#footnote-9) است که تعداد زیادی درخت تولید میکند. مقدار کلاس که اغلب میان این درختان ظاهر میشود، کلاس پیش بینی شده بعنوان خروجی از مدل است. مانند یک روش یادگیری گروهی، جنگل تصادفی مجموعی از درختان پیش بینی کننده ی مختلف است. هر درخت در جنگل وابسته است به مقدار دیکته شده توسط یک بردار تصادفی که بطور مستقل نمونه گیری شده و در جایی که هر درخت در جنگل بطور برابری توزیع شده است. جنگل تصادفی گروهی تصادفی بودن را به فرایند یادگیری اضافه میکند که میتواند تناسب بیش از حد [[9]](#footnote-10) را کمینه کند و در برابر داده های نامتوازن مقاوم است. ما هر یادگیرنده ی جنگل تصادفی را با 100 درخت ساختیم زیرا گروه تحقیقاتی ما متوجه شد که استفاده از درخت بیشتر سودی ندارد. پارامتری که ID درختان را برای هر نمونه نگه میدارد،True تنظیم شد و پارامتر بیشینه حافظه 1024 مگابایت تنظیم شد تا زمان آموزش کمینه شود. تنظیماتی که تعداد ویژگی ها را اداره میکند، برای اینکه تقسیمات را در هر نود درخت در نظر بگیرد، یک سوم تنظیم شد زیرا این تنظیم بر اساس تحقیقات اولیه نتیجه بهتری فراهم می اورد. پارامتر بیشینه bin، که بیشترین تعداد bin ها برای گسسته کردن ویژگی های پیوسته است، به مقدار 2 تنظیم شد زیرا ما دیگر ویژگی کیفی نداریم چون با استفاده از one-hot encoding تبدیل شدند. درخت gradient boosted (GBT)یکی دیگر از درختان تصمیم گروهی است. برخلاف جنگل تصمیم، GBT هر درخت تصمیم را در یک زمان یک بار آموزش میدهد تا خسارت تعیین شده توسط تابع loss الگوریتم را کاهش دهد. در هر چرخه، گروه فعلی برای پیش بینی کلاس برای هر نمونه در داده ی آموزشی استفاده میشود. مقادیر پیش بینی شده با مقادیر واقعی ارزیابی میشوند که به الگوریتم اجازه ی pinpoint و اصلاح نمونه های به اشتباه برچسب خورده ی قبلی را میدهد. پارامتری که ID نودها برای هر نمونه را نگه میدارد،true تنظیم شد و پارامتر بیشینه حافظه به 1024 مگابایت تنظیم شد تا زمان آموزش را کمینه کند.

## **معیار کارایی**

## Cross-validation

## Significance testing

# بحث و نتایج

# نتیجه گیری

1. Fee-For-Service program [↑](#footnote-ref-2)
2. Medicare Advantage program [↑](#footnote-ref-3)
3. National Provider Identifier [↑](#footnote-ref-4)
4. Health Common Procedure Coding System [↑](#footnote-ref-5)
5. Facility [↑](#footnote-ref-6)
6. Non-facility [↑](#footnote-ref-7)
7. Office of inspector general [↑](#footnote-ref-8)
8. Ensemble learning [↑](#footnote-ref-9)
9. overfitting [↑](#footnote-ref-10)