# 1.پیش پردازش

در فصل 4 توضیحات مجموعه داده استفاده شده ارایه خواهد شد. داده‌ی LEIE برچسب‌ها را برای هر شماره نظام پزشکی ملی نگهداری می‌کند و داده‌ی Medicare Provider Utilization and Payment تجمیعی از اطلاعات ثبت شده از داروهای تجویزی در هر نسخه یا نوع خدمت پزشکی از جانب پزشک می باشد. همچنین مقدار پرداختی در ازای نسخه یا خدمت پزشکی ذکر شده است. البته نوع داروهای به کار رفته در این مجموعه داده به جهت خلاصه سازی در دسترس نیست. بجز اطلاعات اولیه مانند نام و نام خانوادگی پزشکان تخصص پزشک نیز برای هر ردیف مشخص شده‌است. شماره نظام پزشکی‌ (NPI) برای هر پزشک یا ارائه کننده‌ی خدمت پزشکی و تخصص وی عددی ده رقمی و منحصر به فرد و تخصصش می‌باشد. شماره‌ی نظام پزشکی افراد نیز برای هر فرد یا بصورت عام ارائه کننده‌ی خدمت پزشکی ذکر شده‌است. لازم به ذکر است که داده مدعی است اطلاعات افرادی را ثبت کرده که خدمات پزشکی ارائه می‌دهند و ممکن است در این داده و برای مثال کاردمان، فیزیوتراپ، بینایی سنج و .. نیز حضور داشته باشند و از نظر فنی پزشک محسوب نشوند. لذا عبارت کلی‌تر ارائه‌دهنده‌ی خدمت پزشکی یا بصورت مختصر ارائه‌دهنده به کار می‌رود.

بعد ازا ارائه اطلاعاتی اجمالی در مورد داده حال روند مرتب‌‌سازی و آماده‌سازی داده قدم به قدم اشرح داده می‌شود.

حجم اولیه داده‌ی Medicare Provider Utilization and Payment، 2GB بوده و شامل 9،847،445 رکورد و 26 ویژگی می‌باشد. همچنین داده‌ی LEIE شامل حدود ۷۳۷۰۴ رکورد می‌باشد که این داده برچسب‌هایی را برای افراد بررسی شده درنظر گرفته و تعدادی از این برچسب‌ها بیانگر تقلب افراد و تعداد نشان از سلامت رفتار ارائه‌دهنده می‌باشد. حال نیاز است این دو داده به هم متصل و ترکیب شوند. برای ترکیب نیاز است در هر دو داده ستونی را به عنوان کلید برای انجام روند ترکیب استفاده کرد. حال از آنجا که نام و نام خانوادگی افراد می‌تواند تکراری باشد، نوشته نشده‌باشد و یا غلط املایی داشته باشد نامزد مناسبی برای انتخاب به عنوان کلید نمی‌باشد. باید اضافه کرد که یک فرد ممکن است بیش از یک تخصص داشته باشد و این کار را پیچیده‌تر نیز می‌کند. ترکیب ستون‌های تخصص و نام و نام خانوادگی نیز عبارتی طولانی و عملا نامناسب برای انتخاب به عنوان ستون کلیدی ارائه می‌دهد. به نظر می‌رسد که ستون NPI که عددی ده رقمی بوده و برای هر فرد و هر تخصص در سطح ملی منحصربه فرد است، نامزد مناسب‌تری برای این منظور می‌باشد. از طرفی به دلیل شباهت اسمی و تخصص امکان ردیابی افراد و کارآیی سیستم در صورت استفاده از نام و تخصص پایین می‌آید. لذا عملا هم بهتر بوده و هم مجبوریم که ستون NPI را به عنوان ستون کلیدی برای هر دو داده در نظر گرفته و اشخاص را با همان NPI نمایندگی کرده و نتایج را مبتنی بر NPI گزارش کنیم.

پس در هر دو داده تنها رکوردهایی درنظر گرفته‌شد که برای ستون NPI مقدار مشخص و غیرصفر (صفر در ستون NPI نشان دهنده‌ی داده‌ی از دست رفته است) دارند. به گونه‌ای که داده‌ی LEIE بعد از حذف NPIهای از دست رفته، تنها دارای 5،588 رکورد بود. روند حذف رکوردهای بدون NPI برای داده‌ی Medicare Provider Utilization and Payment نیز انجام شد و با داده‌های باقی مانده روند ادامه۸ پیدا کرد.

قبل از اینکه دو داده باهم ترکیب شوند لازم بود مقداری پیش پردازش بروی داده‌ی Medicare Provider Utilization and Payment انجام پذیرد. این داده شامل ۲۶ ستون بوده و همانطور که گفته شد دارای 9،847،445 رکورد می‌باشد. تعداد زیادی از ستون‌های این داده مقادیر اسمی دارند. از بین بین این ستون‌ها اکثرا مقادیر اسمی همچون نام و نام خانوادگی، آدرس ارائه‌ی خدمت، کدهای دارویی و سرویس و ... هستند که ارزش معناداری برای تبدیل به عددی ندارند (نام، ایالت، شهر و ...) یا تعداد حالت‌های آن بسیار زیاد (کدهای دارویی) و داده‌ی برچسب‌دار ما نسبتا کوچک است و یا مقادیر از دست رفته‌ی زیادی دارند و عملا این ستون‌ها شرایط استفاده را ازدست می‌دهند. لذا از ۲۶ ستون مذکور ۱۶ ستون حذف شده و ۱۰ ستون نگه‌داشته‌شد. از ستون‌های موجود دو ستون جنسیت ارائه دهنده (nppes\_provider\_gender) و تخصص ارائه‌دهنده (provider\_type) اسمی هستند.

در ستون جنسیت (nppes\_provider\_gender) ۴۳۱۳۱۸ رکورد وجود داشت که مقادیر جنسیت از دست رفته بود. از آنجایی که ۶۸۳۶۸۵۵ ارائه دهنده مرد و ۳۰۱۰۵۸۸ ارائه‌دهنده زن بودند، لذا مقادیر از دست رفته به صورت تصادفی و با احتمال ۶۹درصد با مقادر «M» و با احتمال ۳۱درصد با مقدار «F» جایگزین شد. درنهایت مقادیر M به یک و مقادیر F با صفر جایگزین شد تا این ستون هم مقادیر از دست رفته را تصحیح کند و هم به عددی تبدیل شود.

ستون تخصص ارائه‌دهنده (provider\_type) دارای ۹۴ تخصص مختلف که هر کدام فراوانی‌های متعددی دارند. این ستون و براساس فراوانی مرتب شده و به جای نوع تخصص با اعدادی از یک تا ۹۴ پر شده‌است. تا بدین ترتیب آخرین ستون اسمی باقی مانده به نوع عددی آن تبدیل شود. حال داده‌ای داریم با ۱۰ ستون که یک ستون متعلق به NPI بوده و بقیه‌ی ۹ ستون به عنوان ماتریس ویژگی‌ها می‌توانند مورد استفاده قرار گیرند.

بعد از آماده‌سازی های اولیه‌ی دو داده‌ی Medicare Provider Utilization and Payment و LEIE، این دو داده براساس ستون کلیدی NPI با هم ترکیب (Inner join) شدند. داده‌ی حاصل شامل ۱۱ ستون است که یک ستون برای NPI و یک ستون برای برچسب‌هاست. همچنین هر رکورد یک بردار ویژگی ۹تایی دارد که نقش مهمی را در ساخت مدل خواهد داشت. از ۵۵۸۸ رکورد ممکن ماتریس داده‌ای حاصل شد که شامل 2،993 رکورد می‌باشد. در نهایت 2،167 نمونه‌‌ی مثبت و 826 نمونه‌‌ی منفی بدست آمد. مدل برروی این داده طراحی و ارائه شده‌است.

# **2.توصیف روش**

در این بخش به نحوه‌ی طراحی و ساخت مدل می‌پردازیم و روند ارائه‌ی پیش بینی برای ارائه‌دهنده‌ی جدید تبیین می‌شود. در این بخش از جداسازی ارائه‌دهندگانی که بیش از یک تخصص دارند، پرهیز شده‌است و همچنین به دلیل امکان بروز حالاتی که یک ارائه دهنده فقط در یک تخصص تخلف داشته و در دیگر تخصص‌ها نیز تخلف و تقلب نداشته‌باشد. لذا هر رکورد با NPI آن مشخص شده و از اضافه کردن جزئیات مربوط به تخصص و اشخاص پرهخیز شده‌است.

## **2.1 تقسیم داده به دو بخش**

یک مجموعه مرجع[[1]](#footnote-1) از ارایه دهندگان تشکیل و به دو بخش مثبت و منفی تقسیم می‌شوند. به قسمی که یک گراف تشکیل شود که در یک توده از راس‌ها، ارائه‌دهنگان متقلب (مثبت) و در توده‌ی دیگر ارائه‌دهندگان غیرمتقلب (منفی) قرار گرفته‌اند. در مدل خود برای پرهیز از افزایش هزینه‌ی نگهداری ماتریس‌های شباهت، از محاسبه و درنظر گرفتن یال‌های متناظر بین ارائه‌دهندگان پرهیز شده‌است. درواقع کار مدل زمانی آغاز می‌شود که NPI جدیدی به مجموعه معرفی شود که برچسبی ندارد و فقط یک بردار ویژگی از آن در دسترس باشد و مدل ملزم به پیش‌بینی نوع مثبت یا منفی برای آن NPI باشد.

## **2.2محاسبه شباهت NPI جدید با هر کلاس**

ابتدا محاسبه‌ی شباهت با روش تبدیل کوسینوس معرفی می‌شود. این تبدیل دارای فرمول محاسباتی همانند شکل ۱-۳ است. به شکلی که x بردار ویژگی NPIi و y بیانگر بردار ویژگی NPIj می‌باشند. و ||.|| یا نرم-۲ اندازه‌ی بردار ویژگی را محاسبه می‌کند. حاصل ضرب داخلی دو بردار x و y بر حاصل‌ضرب اندازه‌ی آن‌ها مقدار کوسینوس زاویه‌ی مجاور بین دو بردار را نمایش می‌دهد که هرچه این مقدار به یک نزدیک باشد دو بردار بیشتر دو بردار شباهت کم‌تری دارند و هرچه مقدار کوسینوس کم‌تر بوده و به صفر نزدیک‌تر باشد، دو بردار به هم شباهت بیش‌تری دارند.

شکل ‏3‑1 فرمول محاسبه معیار فاصله‌ی کسینوسی

همانطور که از توضیح بر‌می آید مقدار کوسینوس در واقع فاصله‌ی دو بردار را محاسبه می‌کند. لذا مقدار شباهت بین دو بردار x و y می‌شود که به آن شباهت کوسینوسی گویند. در این پژوهش از تبدیل برای محاسبه‌ی شباهت استفاده شده‌است.

در روش ارائه‌شده، برای هر NPI جدید که برچسب آن مشخص نیست، ابتدا شباهت بردار ویژگی آن با بردار ویژگی‌های دو دسته‌ی مثبت و منفی محاسبه می‌شود. برای محاسبه‌ی شباهت از تبدیل کوسینوسی استفاده می‌شود. تا دو بردار شباهت بدست آید. یک بردار به طول ۲167 بیانگر شباهت بردار ویژگی NPI جدید با بردار ویژگی NPI های مثبت و دیگریبردار ویژگی به طول 826 می‌باشد که بیانگر شباهت NPI جدید با هریک از NPIها در دسته‌ی منفی می‌باشد.

لازم به ذکر است نگارنده اصراری بر استفاده از تبدیل کوسینوسی نداشته و می‌توان این تبدیل را با دیگر تبدیلات شناخته‌شده جایگزین نمود. دلیل اصلی استفاده از ین معیار در این پایان نامه این است که این معیار یکی از پرکاربردترین معیارها است. و همچنین باید در نظر گرفت، در عمل درصورت ورود NPI جدید بردارهای شباهت محاسبه شده و یال‌هایی از NPIهای برچسب دار به نمونه‌ی جدید رسم شده که مقدار وزن هر یال برابر با شباهت متناظر بین دو NPI می‌باشد. مقدار وزن یال‌ها پیوسته بوده و مقداری بین صفر و یک است.

## **2.3 محاسبه امتیاز بر حسب k نزدیکترین همسایگی**

در این مرحله برای اتخاذ تصمیم نیاز است میانگین شباهت‌ها در هر بردار شباهت محاسبه شود. توده‌های مثبت و منفی در صورت رسم در یک فضای ۹-بعدی می‌توانند در هر نقطه از فضا بوده و لزوما تمامی موارد مثبت یا منفی در نزدیک هم نیستند. لذا محاسبه‌ی میانگین شباهت هر بردار نمی‌تواند معیار مناسبی باشد و نیاز است که روند تصمیم امتیازدهعی مدل محلی‌تر باشد تا از درگیری با چالش چگالی نقاط در فضا دوری شود. از این ور بردارهای شباهت مثبت و منفی از بزرگ به کوچک مرتب شده و ۵ مقدار شباهت بیشینه از هر بردار انتخاب می‌شود. در این‌صورت میانگین ۵ نزدیک‌ترین همسایگی به NPI جدید محاسبه می‌شود تا دو امتیاز عضویت در گروه منفی و امتیاز عضویت در گروه مثبت محاسبه شود. استفاده از k همسایگی منجر به کاهش اثر داده پرت می‌شود زیرا تصمیم‌گیری تنها بر مبنای بیشترین شباهت نیست. لذا مدل همچون روش چند نزدیک‌ترین همسایگی[[2]](#footnote-2) (KNN) به صورت محلی عمل کرده و برای نمونه‌ی جدید تصمیم‌گیری می‌کند. تعداد همسایگی در این تحقیق ۵ درنظر گرفته‌شده است. این مقدار بهینه نشده‌است.

## **2.4 محاسبه نسبت بخت تعلق به هر کلاس**

بعد از محاسبه‌ی امتیازهای میانگین محاسبه شده که عضویت به دو گروه را نشان می‌داد، مقدار log ratio یا نسبت بخت محاسبه می شود. در نهایت از روی مقدار log ratio و فرمول مربوط به ان طبق شکل زیر است.

شکل ‏3‑3 فرمول محاسبه log ratio

به قسمی که در صورت مقدار میانگین شباهت ۵ نزدیک‌ترین همسایه از زیرگراف مثبت و در مخرج میانیگین ۵ نزدیک‌ترین همسایه از زیرگراف منفی قرار می‌گیرد.

درصورتی که NPI جدید مقدار Log ratio بالاتر و مساوی یک داشته باشد به عنوان متقلب درنظر گرفته و احتمال تعلق به زیرگراف متقلب‌ها بیش‌تر خواهد بود. در غیر این‌صورت NPI جدید منفی پیش بینی شده و به زیر گراف ها غیرمتقلب‌ها نسبت داده می‌شود.

1. Reference set [↑](#footnote-ref-1)
2. K- Nearest Neighbor [↑](#footnote-ref-2)