پیش بینی روند بروز بیماری های تنفسی براساس شرایط آب و هوایی

امین محمدیان[۱]، سیدابوالقاسم میرروشندل[۲]، حمیدرضا احمدی فر[۳]، سید علی علوی[۴]، بهروز گل چای[۵]

[۱] گروه مهندسی کامپیوتر، دانشگاه گیلان ، رشت
Amin.muhammadian@gmail.com

[۲] گروه مهندسی کامپیوتر، دانشگاه گیلان ، رشت
mirroshandel@guilan.ac.ir

[۳] گروه مهندسی کامپیوتر، دانشگاه گیلان ، رشت
ahmadifar@guilan.ac.ir

[۴]دانشگاه علوم پزشکی گیلان، رشت
saalavi@gums.ac.ir

[۵]دانشگاه علوم پزشکی گیلان، رشت
behrooz.golchay@gmail.com

چکیده

حفظ و تقویت آمادگی برای پذیرش و درمان بیماران یکی از مهم ترین نگرانی های مراکز درمانی اســت. روزانه بیماران زیادی با علائم مختلف بیماری به مراکز درمانی مراجعه می نمایند. بیماران تنفسی، جامعه بزرگی از این مراجعه کنندگان را تشکیل می دهند. هدف این تحقیق بهره گیری از روش های داده کاوی در جهت ســاخت مدل پیش بینی کننده میزان مراجعین اســت. در ابتدا داده های ورودی مورد پردازش اولیه قرار گرفته و سپس الگوریتم های داده کاوی بر آنها اِعمال شده است. الگوریتم های گروه دسته بندی نظیر درخت تصمیم به عنوان الگوریتم هدف انتخاب شده اســت. میزان تاثیر عملیات های پیش پردازش نیز مورد بررســی و نتایج حاصــل از مرحله داده کاوی با داده های ورودی مختلف مورد ارزیابی قرارگرفته است. برای ارزیابی عملکرد الگوریتم ها، از روش اعتبار سنجی متقابل ۱۰ برابری استفاده شده است. الگوریتم های دسته بندی تنبل و درخت تصمیم بهترین عملکرد را از خود نشان دادند. بهره گیری از عملیات های پیش پردازش مختلف نیز موجب بهبود نتایج شده است. درنهایت می توان اینگونه بیان نمود که بهره گیری از روش های یادگیری ماشــین می تواند کمک شــایانی به پیش بینی تعداد مراجعین و در نتیجه افزایش میزان آمادگی مراکز درمانی نماید.

کلمات کلیدی

یادگیری ماشین، داده کاوی، پیش بینی، بیماری های تنفسی

۱- مقدمه

در دنیای جدید با گسترش روز افزون فناوری های اطلاعاتی و ارتباطی، روزانه حجم زیادی از داده ها گردآوری می شود. استخراج دانش موجود در این قبیل داده ها با بهره گیری از روش های مختلف آماری و غیر آماری امکان پذیر است [1]. داده کاوی به عنوان زیرمجموعه ای از علم استخراج دانش همواره مورد توجه قرار گرفته است. اهداف اصلی فرآیندهای داده کاوی را می توان در دو گروه پیش بینی و توصیف جای داد [2]. در پیش بینی داده ها، سعی بر آن

است که با بهره گیری از داده های گردآوری شده در گذشته، مقادیری تا حد امکان دقیق از متغیر هدف در آینده تخمین زده شود [3]. عموما پیش بینی ها برای بازه های زمانی مختلف در آینده مورد استفاده قرار می گیرند. در شکل (۱) می توان به طور کلی مراحل استخراج دانش را مشاهده نمود.

شکل (۱) مراحل استخراج دانش از داده های اولیه

با توجه به ماهیت داده ها در گام اول، داده ها مورد پردازش های ابتدایی قرار می گیرند تا برای عملیات های داده کاوی آماده گردند. در گام بعد فرآیندهای داده‌کاوی بر روی داده ها اِعمال می شوند که در نهایت دانش حاصل می‌شود.

امروزه افراد زیادی در سرتاسر جهان تحت تاثیر بیماری های مختلف قرار دارند. شیوع بیماری هایی نظیر آسم و همچنین بیماری های انسدادی دستگاه تنفس به طور روز افزون در حال افزایش است. این قبیل بیماری ها بسیار به تغییرات شرایط محیط پیرامون حساس هستند. هر انسان در طول روز به طور مداوم در حال تنفس هوای اطراف است. از این رو اولین بخش از بدن هر فرد که تحت تاثیر محیط پیرامون قرار می گیرد، دستگاه تنفسی او است. از سوی دیگر، با تغییر شرایط اقلیمی زمین، وضعیت آب و هوایی نیز به سرعت در حال تغییر است. علاوه بر این تغییرات، فعالیت های انسان ها نیز در حال آلوده نمودن محیط زندگی آن ها است. با در نظر گرفتن شرایط فوق می توان به این نتیجه رسید که، میزان خطر ابتلا به بیماری های تنفسی بسیار بیشتر از گذشته بوده و به طور روزافزون در حال افزایش است [6 - 4].

بنابر آنچه بیان شد، امروزه یکی از اصلی ترین دلایل مراجعه بیماران به مراکز درمانی به ویژه در شهرهای پر جمعیت، بروز علائم بیماری های تنفسی است. در برخی شهرها مراکز خاصی برای اینگونه بیماری ها اختصاص یافته است. این مراکز همواره آماده پذیرش و درمان بیماران تنفسی هستند. اما میزان مراجعه بیماران به این مراکز در روزهای مختلف متفاوت است. یکی از علل اصلی در تغییر میزان مراجعات، تغییر شرایط آب و هوایی و تاثیر آن بر بروز این نوع از بیماری ها است [7].

بنابراین با استفاده از یک مدل پیش بینی نسبتا دقیق که براساس شرایط آب و هوایی عمل می نماید، می توان آمادگی مراکز درمانی را برای پذیرش بیماران تنفسی افزایش داد. این امر موجبات افزایش کارآیی مراکز و همچنین رضایتمندی بیماران را بدنبال خواهد داشت. از این رو این تحقیق سعی دارد تا با بهره گیری از روش های داده کاوی، مدلی را برای پیش بینی تعداد بیماران مراجعه کننده به مراکز بیماری های تنفسی ارائه نماید.

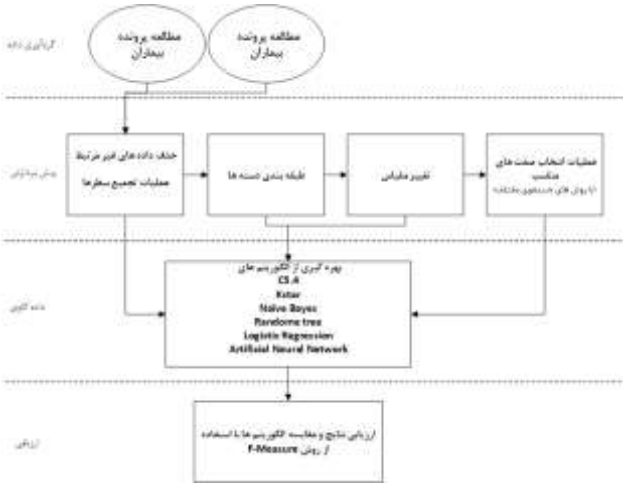
دربخش اول کلیات مورد مطالعه قرار گرفته و همچنین اهداف اصلی تحقیق بیان شده است. مروری بر مطالعات صورت پذیرفته در گذشته که شامل علوم پزشکی و غیر پزشکی نیز هست در این مرحله بیان شده است. راهکارهای پیشنهادی این تحقیق برای یافتن مدل مناسب در بخش دوم بیان شده است. بخش سوم به بیان نتایج حاصل از به کار گیری روش های پیشنهادی و همچنین تحلیل این نتایج پرداخته است. بخش چهارم نیز به جمع بندی کارهای صورت پذیرفته اختصاص یافته است.

علم داده کاوی پیوسته در حال پیشرفت و گسترش بوده است. از این رو همواره در جهت اعتلای سایر علوم مورد استفاده قرار گرفته است. علم پزشکی نیز به عنوان علمی به روز و کارآمد و به واسطه بهره گیری از ابزار آلات جدید، به عنوان یکی از تولیدکنندگان داده های عظیم شناخته می شود. پردازش این داده ها و بهره گیری از آن ها در جهت ارتقای سلامت جامعه یکی از اهداف به کار گیری داده کاوی در علم پزشکی است [8].

داده کاوی همواره در حوزه های گوناگونی نظیر انرژی و اقتصاد مورد استفاده قرار گرفته است [11 - 9]. تا کنون تحقیقات مختلفی در زمینه به‌کارگیری داده کاوی در علم پزشکی نیز انجام پذیرفته است. مطالعه ای سعی‌داشته تا با بهره گیری از روش Auto-regressive پلکانی تعداد بیماران کلیوی را پیش بینی نماید [12]. تحقیقات دیگری از الگوریتم های دسته بندی نظیر درخت تصمیم و نایو بیز برای ساخت یک مدل پیش بینی کننده میزان حملات قلبی استفاده کرده اند [14 ,13]. شبکه عصبی و رگرسیون منطقی و همچنین درخت تصمیم نیز به عنوان ابزاری برای پیش بینی احتمال وقوع سرطان استفاده شده اند [15]. در مثالی دیگر از ترکیب شبکه عصبی و سیستم فازی برای بیماری های مختلف و با تاکید بر بیماری آسم استفاده شده است [16]. در تحقیقی سعی برآن بوده است که با استفاده از یک سیستم منطق فازی به تشخیص بیماری های تنفسی کمک شود [17]. همچنین محققان از روش ترکیبی Auto-regressive با سیستم فازی نیز برای پیش بینی تعداد حملات آسم بهره گرفته اند [18]. به عنوان راهکاری دیگر ازیک سیستم فازی مبتنی بر قواعد برای تشخیص بیماری آسم استفاده شده است [19]. درنوع دیگری از تحقیقات نیز داده های پز شکی را به عنوان سری زمانی در نظر گرفته اند و از روش های داده کاوی برای پیش‌بینی سری ها بهره برده اند [21 ,20].

۲- راهکار پیشنهادی

عملیات استخراج دانش از چندین پردازش جداگانه تشکیل شده است. هریک از این پردازش ها ورودی پردازش بعدی را تامین می نماید. هدف اصلی در این مطالعه نیز به کارگیری فرآیند های داده کاوی است. از این رو سعی شده است تا در جهت افزایش دقت و کارایی مدل پیشنهادی، مراحل پردازشی به درستی پیاده سازی شود. بدین منظور در ابتدا داده‌ها مورد پردازش اولیه قرار گرفته اند. در گام بعد عملیات های مختلف داده کاوی مورد استفاده قرار می گیرند. در گام نهایی نتایج با یکدیگر مقایسه شده و بهترین روش معرفی می شود. در شکل (۲) می توان مراحل کلی فرآیند را مشاهده نمود. شایان ذکر است در این مطالعه علاوه بر تلاش برای یافتن بهترین روش داده کاوی برای ساخت مدل پیش بینی کننده، تاثیر روش های مختلف پیش‌پردازش نیز بر میزان دقت مدل نهایی مورد ارزیابی و مقایسه قرار خواهد گرفت.



شکل (۲) مراحل فرآیندهای انجام پذیرفته در این تحقیق

۱-۲- پیش پردازش

داده های گردآوری شده عموما داری نقایصی هستند که نتایج فرآیند داده‌کاوی را تحت تاثیر قرار می دهند. معمولا داده ها غیرمعتبر و ناکامل، دارای خطا و همچنین حجیم هستند [22]. داده هایی که در علوم پزشکی مورد استفاده قرار می گیرند نیز با توجه به شرایط محیطی نظیر منابع اطلاعاتی گوناگون و نحوه گردآوری به صورت غیرمتجانس و به همراه محدودیت هایی هستند [23]. از این رو در این گام، ابتدا توضیحاتی در باب مجموعه داده ی مورد استفاده بیان شده و سپس مراحل پیش پردازش صورت پذیرفته در جهت آماده سازی داده ها یک به یک شرح داده می شود. این اطلاعات از پرونده پزشکی مراجعان به بیمارستان رازی رشت طی سال های ۸۷ الی ۹۲ و همچنین مرکز هواشناسی گیلان گردآوری شده است. داده ها مربوط به ۱۴۶۲ بیمار هستند.

داده های مورد استفاده در این تحقیق از سه بخش اصلی تشکیل شده است: ۱.داده های مربوط به بیمار ۲. داده های هواشناسی ۳. زمان مراجعه. صفات موجود در داده های اولیه را می توان در جدول (۱) مشاهده نمود. صفات در سه گروه فوق دسته بندی شده اند.

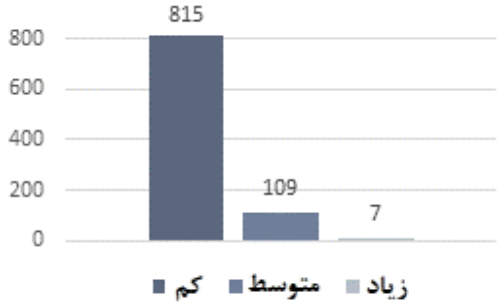
جدول (۱) صفات موجود در مجموعه داده اولیه

عنوان بخش	صفات موجود
صفات مربوط به اطلاعات بیماران	سن؛ شغل؛ جنسیت؛ مصرف دخانیات؛ تماس با آلاینده؛ نوع بیماری؛ بیماری همراه؛ نتایج آزمایشات انجام پذیرفته بعد از مراجعه به مرکز (FVC,PAO2,…)
صفات مربوط به اطلاعات هواشناسی	ساعات آفتابی روز؛ سرعت باد؛ کمینه، بیشینه و میانگین دما،رطوبت و فشار؛ میانگین کمینه و بیشینه نقطه شبنم
صفات مربوط به زمان	فصل؛ ماه؛ روز؛ ساعت؛ تاریخ

در گام اول با توجه به هدف تحقیق کلیه داده های غیر مرتبط در جهت کاهش حجم داده ها و افزایش دقت حذف می شوند. بدین منظور کلیه داده های دسته اول که صرفا اطلاعاتی در راستای شناخت بیماران و مشخصات آن‌ها بوده و بر نتیجه تحقیق تاثیر ندارند حذف می شوند. بعد از این مرحله هر سطر از داده ها صرفا حاوی اطلاعات تاریخ و هواشناسی است. با توجه به اینکه تعداد مراجعین به عنوان هدف در نظر گرفته شده است، با استفاده از عملیات تجمیع، سطرهای دارای مقادیر یکسان بایکدیگر ترکیب و تعداد آنها به صورت صفت جدیدی و با عنوان تعداد مراجعین در کنار هر سطر قرار گرفته است. بعد از مرحله فوق داده های دسته سوم که مربوط به زمان مراجعه است نیز فاقد کاربرد خاصی بوده و از این رو حذف می شوند. روش دیگری برای افزایش دقت الگوریتم های داده کاوی وجود دارد که در آن دسته بندی های مختلف که ارتباط نزدیکی با یکدیگر دارند در یک دسته مشترک گردآوری می شوند. از این رو متغیر هدف که همان تعداد مراجعین است به سه دسته مجزا تقسیم می شود. این متغیر شامل اعداد بین ۱ الی ۶ است. نحوه طبقه بندی دسته ها را می توان در جدول (۲) و همچنین نحوه توزیع داده ها را در شکل (۳) مشاهده نمود.

جدول (۲) نحوه دسته بندی متغیر تعداد مراجعین

مقادیر متغیر	دسته بندی
۱و۲	کم
۳و۴	متوسط
۵و۶	زیاد



شکل (۳) نحوه توزیع داده ها در سه کلاس

یکی از مشکلاتی که نتایج فرآیند های داده کاوی را تحت تاثیر قرار می‌دهد، تفاوت در مقیاس داده ها است [24]. از این رو در مرحله سوم از پیش پردازش داده ها، فرآیند تبدیل مقیاس داده ها صورت می پذیرد. مقیاس کلیه داده های مربوط به هواشناسی طبق معادله(۱) به بازه [۱-۰] برگردانده می شوند.

برای تغییر مقیاس هرداده X از یک مجموعه داده می توان از معادله (۱) استفاده کرد. بر این اساس هرعضو(X) به مقداری جدید(X′) در بازه صفر تا یک تبدیل می شود.

$$X' = \frac{X - X_{Min}}{X_{Max} - X_{Min}} \qquad (۱)$$

یکی دیگر از مشکلات موجود در داده کاوی، وجود صفات غیر تاثیر گذار در مجموعه داده ها ا ست. این صفات می توانند زمان انجام فرآیند داده کاوی را افزایش داده و در مقابل دقت آن را کاهش دهند. از این رو یکی از مراحل مهم در پیش پردازش داده ها، انتخاب صـفت های مناسـب از میان مجموعه صفت های موجود ا ست. فرآیند انتخاب صفت خود از دو بخش مجزا تشکیل شده ا ست. بخش اول نحوه جستجوی میان صفت ها را مشخص می نماید. بخش دوم نیز ارزیابی کیفیت صـفت های انتخابی در مرحله قبل را بر عهده دارد. در نهایت نیز مجموعه صفات مشخص شده با بیشترین میزان کیفیت به عنوان مجموعه نهایی شناخته می شود [25]. ازاین رو در مرحله چهارم از فرآیند پیش پردازش، این روش و میزان تاثیر گذاری آن بر فرآیند داده کاوی مورد بررسـی قرار گرفته اسـت. در جدول (۳) صـفت های انتخاب شـده طی فرآیند انتخاب صفت با روش های جستجوی ابتدا بهترین، ژنتیک و تصادفی و با معیار ارزیابی CFS نمایش داده شده است.

جدول (۳) صفت های انتخاب شده حاصل از فرآیند انتخاب صفت

روش جستجو	صفت های انتخاب شده
جستجوی ابتدا بهترین (دوطرفه)	سرعت باد
جستجوی ژنتیک	میانگین فشار
جستجوی تصادفی	میانگین دما؛ بیشینه دما؛ کمینه دما؛ کمینه رطوبت؛ میانگین بیشینه نقطه شبنم؛ میانگین فشار؛ کمینه فشار

در گام فرآیند داده کاوی، الگوریتم های مشخص شده بر روی نتایج هر یک از مراحل فوق اجرا می شود.

۲-۲- فرآیند داده کاوی

روش های داده کاوی مختلفی برای پیش بینی مقادیر هدف در یک مسـأله وجود دارد. در این مرحله روش دسـته بندی به عنوان مبنا در نظر گرفته شـده اسـت. به طور کلی روش های دسـته بندی را می توان در گروه های مختلف نظیر درخت ها، بیزها، تنبل ها و سـایر گروه ها تقسـیم نمود. در این تحقیق سعی بر آن بوده است تا از هر گروه نماینده ای برای داده کاوی انتخاب شود. در میان الگوریتم های دسـته بندی از درخت تصـمیم C4.5، درخت تصـمیم تصـادفی، شـبکه عصـبی، رگرسـیون منطقی و الگوریتم KStar استفاده شده است.

۳- ارزیابی نتایج

روش های گوناگونی برای ارزیابی نتایج حاصـل از عملیات داده کاوی نظیر میانگین مجذور خطای پیش بینی وجود دارد [26]. در اینجا از معیار های دقت، بازخوانی و همچنین معیار F که ترکیبی از دو معیار قبلی اسـت برای ارزیابی عملکرد الگوریتم ها استفاده شده است.

تعاریف معیارهای فوق در ادامه مورد بررسی قرار می گیرد. حاصل تقسیم تعداد دسته های درست تشخیص داده شده به تعداد کل دسته های تشخیص داده شـده را دقت می نامند. نحوه محاسـبه این معیار در معادله (۲) قابل مشاهده است.

(۱)

$$\text{Precision}=\frac{True\ Positive}{True\ Positive+False\ Positive}$$

حاصل تقسیم تعداد دسته های درست تشخیص داده شده به تعداد کل دسته های درست در مجموعه داده را بازخوانی می نامند. نحوه محاسبه این معیار نیز در معادله (۳) قابل مشاهده است.

$$\text{Recall}=\frac{True\ Positive}{True\ Positive+False\ Negative} \qquad (۲)$$

نحوه محاسبه معیار F نیز که ترکیبی از دو معیار قبلی است به صورت معادله (۴) تعیین می شود.

$$\text{F-Measure}=\frac{2*Precision*Recall}{Precision+Recall} \qquad (۳)$$

در گام اول نتایج حاصـل از اِعمال روش های داده کاوی بر داده های مرحله ی اول پیش پردازش مورد ارزیابی قرار گرفته اسـت. نتایج حاصـل در جدول (۴) قابل مشـاهده اسـت. همانطور که نشـان داده شـده، الگوریتم های گروه تنبل و درخت تصـمیم با معیار F بیشـتر از ۰,۹ بهترین عملکرد را داشته‌اند. از این میان درخت تصمیم C4.5 با معیار F ۰,۹۳ بیشترین دقت را به خود اختصاص داده است.

جدول (٤) نتایج حاصل از اجرای الگوریتم های داده کاوی بر روی داده‌های اولیه

الگوریتم	دقت	بازخوانی	معیار F
Naïve Bayes	۰,۶۳۰۳	۰,۷۴۲۲	۰,۶۸۰۱
Logistic Regression	۰,۸۸۵۲	۰,۹۱۰۷	۰,۸۹۷۲
ANN	۰,۸۸۵۵	۰,۹۱۲۶	۰,۸۹۸۳
Kstar	۰,۸۷۷۶	۰,۹۸۹۰	۰,۹۲۹۹
C4.5	۰,۸۷۶۵	۰,۹۹۲۶	**۰,۹۳۰۹**
Random Tree	۰,۸۷۸۷	۰,۹۷۳۷	۰,۹۲۳۷

در گام بعد داده ها بعد از طبقه بندی متغیر هدف در سـه طبقه کم، متوسـط و زیاد مورد ارزیابی قرار گرفته اند. نتایج حاصـل در جدول (۵) قابل مشاهده است.

جدول (۵) نتایج حاصل از داده کاوی بر روی داده های طبقه بندی شده

الگوریتم	دقت	بازخوانی	معیار F
Naïve Bayes	۰,۶۰۴۲	۰,۹۷۷۹	۰,۷۴۶۸
Logistic Regression	۰,۸۷۵۷	۰,۹۹۴۴	۰,۹۳۱۳
ANN	۰,۸۷۵۵	۰,۹۹۴۴	۰,۹۳۱۲
Kstar	۰,۸۷۵۴	۱,۰۰۰۰	**۰,۹۳۳۶**
C4.5	۰,۸۷۵۴	۱,۰۰۰۰	**۰,۹۳۳۶**
Random Tree	۰,۸۷۶۲	۰,۹۹۸۸	۰,۹۳۳۵

در این مرحله برخلاف مرحله قبل همه گروه ها به غیر از بیز با معیار F بیشتر از ۰,۹ عملکرد مناسبی از خود نشان داده اند. با این تفاوت که در این مرحله الگوریتم KStar و درخت C4.5 عملکرد تقریبا یکسانی داشته و درخت تصادفی نیز بسیار نزدیک به این دو الگوریتم عمل نموده است. اما نکته حائز اهمیت در این بخش بهبود عملکرد الگوریتم ها بعد از یک مرحله پیش پردازش است. به طوری که میانگین معیار F برای نتایج بعد از طبقه بندی با میزان ۰,۰۲۴۹ رشد همراه بوده است.

در گام های قبل، داده ها بدون تغییر مقیاس مورد داده کاوی قرار گرفتند. در این مرحله نتایج حاصل بعد از مرحله‌ی سوم پیش پردازش، یعنی تغییر مقیاس داده های هواشناسی مورد ارزیابی قرار می گیرند. نتایج حاصل از به کارگیری داده ها بعد از تغییر مقیاس در جدول (۶) ارائه شده است.

جدول (٦) نتایج حاصل از فرآیند داده کاوی بر روی داده های تغییر مقیاس داده شده

الگوریتم	دقت	بازخوانی	معیار F
Naïve Bayes	۰,۶۰۵۶	۰,۹۱۴۱	۰,۷۲۸۱
Logistic Regression	۰,۸۷۵۳	۰,۹۸۵۵	۰,۹۲۷۱
ANN	۰,۸۷۵۳	۰,۹۸۵۵	۰,۹۲۷۱
Kstar	۰,۸۷۵۳	۰,۹۹۸۸	۰,۹۳۲۹
C4.5	۰,۸۷۵۴	۱,۰۰۰۰	**۰,۹۳۳۶**
Random Tree	۰,۸۷۵۲	۰,۹۹۸۰	۰,۹۳۲۶

در این جدول نیز بهبود عملکرد نسبت به مرحله اول به چشم می خورد. اما عملکرد نسبت به مرحله قبل که ورودی داده های مرحله جدید را تامین نموده است، بدتر شده است. به طوری که میانگین معیار F نسبت به مرحله قبل ۰,۰۰۴۷ کاهش پیدا نموده است. هرچند بر الگوریتم های با بهترین عملکرد نظیر C4.5 تاثیر چندانی نداشته است.

در گام آخر نتایج حاصل از به کار گیری روش انتخاب صفات در مرحله پیش پردازش مورد ارزیابی قرار می گیرد. نتایج حاصل در جدول های (۷)، (۸) و (۹) قابل مشاهده است.

جدول (۷) نتایج حاصل با روش جستجوی ابتدا بهترین

الگوریتم	دقت	بازخوانی	معیار F
Naïve Bayes	۰,۵۹۹۷	۰,۶۳۱۴	۰,۶۱۴۴
Logistic Regression	۰,۸۸۷۳	۰,۹۰۰۹	۰,۸۹۳۸
ANN	۰,۸۸۷۳	۰,۹۰۰۶	۰,۸۹۳۶
Kstar	۰,۸۷۵۴	۱,۰۰۰۰	**۰,۹۳۳۶**
C4.5	۰,۸۷۵۴	۱,۰۰۰۰	**۰,۹۳۳۶**
Random Tree	۰,۸۷۶۸	۰,۹۳۹۴	۰,۹۰۶۸

جدول (۸) نتایج حاصل با روش جستجوی ژنتیک

الگوریتم	دقت	بازخوانی	معیار F
Naïve Bayes	۰,۶۱۰۰	۰,۷۴۶۳	۰,۶۷۰۰
Logistic Regression	۰,۸۷۵۳	۰,۹۹۷۳	۰,۹۳۲۳
ANN	۰,۸۷۵۳	۰,۹۹۷۳	۰,۹۳۲۳
Kstar	۰,۸۷۵۴	۱,۰۰۰۰	۰,۹۳۳۶
C4.5	۰,۸۷۵۴	۱,۰۰۰۰	۰,۹۳۳۶
Random Tree	۰,۸۷۵۴	۱,۰۰۰۰	۰,۹۳۳۶

جدول (۹) نتایج حاصل با روش جستجوی تصادفی

الگوریتم	دقت	بازخوانی	معیار F
Naïve Bayes	۰,۶۰۴۸	۰,۵۹۹۲	۰,۶۰۰۷
Logistic Regression	۰,۸۷۷۸	۰,۸۷۱۲	۰,۸۷۴۰
ANN	۰,۸۷۸۱	۰,۸۶۶۷	۰,۸۷۱۹
Kstar	۰,۸۷۵۱	۰,۹۹۷۱	**۰,۹۳۲۱**
C4.5	۰,۸۷۶۴	۰,۹۰۲۱	۰,۸۸۸۸
Random Tree	۰,۸۷۳۵	۰,۸۵۲۱	۰,۸۶۲۳

با توجه به جداول (۷)، (۸) و (۹) می توان به این نتیجه رسید که هیچ یک از سه عملیات انتخاب صفات مناسب نتوانسته میزان میانگین معیار F را افزایش دهد. از این رو انتخاب صفت، گزینه مناسبی برای مرحله پیش پردازش نبوده است.

در جدول (۱۰) می توان میانگین معیار F را بعد از هر مرحله پیش پردازش و همچنین میزان اختلاف آن با مرحله‌ی قبل را مشاهده نمود. شایان ذکر است مرحله چهارم که از سه زیرمرحله تشکیل شده و هر زیر مرحله به صورت مجزا از خروجی مرحله قبل بهره گرفته، به صورت جداگانه در جدول نمایش داده شده و اختلاف هریک از آن ها، با مرحله سوم درج شده است. علامت "+" نشان دهنده بهبود و علامت "–" نشان دهنده کاهش کیفیت عملیات داده کاوی است.

جدول (۱۰) میانگین و میزان تغییرات معیار F در نتایج حاصل از داده‌کاوی بر روی خروجی مراحل مختلف پیش پردازش

مرحله پیش پردازش	میانگین	تغییرات نسبت به مرحله قبل	نحوه رشد
مرحله اول	۰,۹۱۶۰	۰	
مرحله دوم	۰,۹۳۲۶	۰,۰۱۶۶	+
مرحله سوم	۰,۹۳۰۷	۰,۰۰۲۰	–
مرحله چهارم(جستجوی ابتدا بهترین)	۰,۹۱۲۳	۰,۰۱۸۴	–
مرحله چهارم(جستجوی ژنتیک)	**۰,۹۳۳۱**	۰,۰۰۲۴	+
مرحله چهارم(جستجوی تصادفی)	۰,۸۸۵۸	۰,۰۴۴۸	–

با توجه به جدول (۱۰) می توان اظهار داشت که تنها مرحله دوم پیش پردازش یعنی طبقه بندی متغیر هدف توانسته است میانگین معیار F را بهبود ببخشد. سایر مراحل پیش پردازش تاثیر منفی در میزان دقت عملیات داده کاوی را از خود نشان داده اند. اما بیان نکته ای در اینجا حائز اهمیت بسیار است. با یک نگاه کلی به جداول نتایج حاصل از عملیات های داده کاوی، می توان به این نتیجه مهم دست یافت که در تمامی آنها الگوریتم نایو بیز دارای پایین ترین میزان دقت بوده که خود تاثیر منفی بر نتایج میانگین گذارده است. از این رو در جدول (۱۱) مقادیر جدول(۱۰) بدون در نظر گرفتن الگوریتم نایو بیز نمایش داده شده است. همانطور که ملاحظه می شود، بدون در نظر گرفتن مقدار معیار F در الگوریتم نایو بیز، دقت سایر الگوریتم ها با بهبود همراه شد. به طوری که بهترین میانگین مربوط به نتایج حاصل از اجرای مرحله چهارم پیش پردازش یعنی انتخاب صفت با روش جستجوی ژنتیک بوده است. در این مرحله معیار F افزایش ۰,۰۱۷۰ نسبت به مرحله اول را تجربه نموده است.

جدول (۱۱) نتایج حاصل بدون در نظر گرفتن الگوریتم نایو بیز

مرحله پیش پردازش	میانگین	تغییرات نسبت به مرحله قبل	نحوه رشد
مرحله اول	۰,۸۷۶۷	۰	
مرحله دوم	**۰,۹۰۱۶**	۰,۰۲۴۹	+
مرحله سوم	۰,۸۹۶۸	۰,۰۰۴۷۴۹	–
مرحله چهارم(جستجوی ابتدا بهترین)	۰,۸۶۲۶	۰,۰۳۴۲۸۳	–
مرحله چهارم(جستجوی ژنتیک)	۰,۸۸۹۲	۰,۰۰۷۶۷۸	–
مرحله چهارم(جستجوی تصادفی)	۰,۸۳۸	۰,۰۵۸۶۰۵	–

نتایج جداول فوق حاصل اجرای ۵ دور هریک از الگوریتم ها است. در هر دور نیز از روش اعتبار سنجی متقابل ۱۰ برابری استفاده شده است.

٤- جمع بندی

این مطالعه روشی را برای پیش بینی احتمال بروز بیماری های تنفسی ارائه داده است. برای نیل به این منظور، از الگوریتم های داده کاوی و فرآیندهای پیش پردازش مختلف بهره گرفته شده است. داده های گردآوری شده از پرونده بیماران مراجعه کننده به مرکز پزشکی با علائم بیماری تنفسی بوده است. بخش دیگری از داده ها از مرکز هواشناسی گیلان به دست آمده است. سعی بر آن بوده است که با بهره گیری از روش های مختلف داده کاوی، بهترین مدل برای این منظور شکل گیرد. همچنین این مطالعه سعی داشته تا تاثیر مراحل مختلف پیش پردازش داده ها را بر دقت الگوریتم های داده کاوی مورد ارزیابی قرار دهد. نتایج اجرای الگوریتم ها با یکدیگر مقایسه شده و بهترین عملکرد مشخص گردیده است. بهره گیری از الگوریتم های داده کاوی بر روی داده های حاصل از مرحله چهارم پیش پردازش بهترین نتیجه را از خود نشان داد. بر ا ساس آنچه انجام پذیرفت می توان به این نتیجه ر سید که بهره گیری از روش‌های یادگیری ما شین می تواند با دقت قابل قبولی احتمال بروز بیماری های تنفسی را بر اساس شرایط آب و هوایی پیش بینی نماید.

مراجع

[1] S. Piramuthu, "*Evaluating feature selection methods for learning in data mining applications*," European journal of operational research, vol. 156, no. 2, pp. 483-494, 2004.

[2] Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A., & Kamaev, V. A. E., "*A survey of forecast error measures*," World Applied Sciences Journal, vol. 24, pp. 171-176, 2013.

[3] J. P. J. &. K. M. Han, *Data mining: concepts and techniques*, Elsevier, 2011.

[4] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P., "*From data mining to knowledge discovery in databases*," AI magazine, vol. 17, no. 3, p. 37, 1996.

[5] Hyndman, R. J., & Athanasopoulos, G., "*Forecasting, planning and goals," in Forecasting: principles and practice.*, OTexts, 2018.

[6] D'amato, G., Rottem, M., Dahl, R., Blaiss, M.S., Ridolo, E., Cecchi, L., Rosario, N., Motala, C., Ansotegui, I. and Annesi-Maesano, I., "*Climate change, migration, and allergic respiratory diseases: an update for the allergist*," World Allergy Organization Journal, vol. 4, no. 7, p. 121, 2011.

[7] D'Amato, G., Cecchi, L., D'Amato, M., & Annesi-Maesano, I. , *Climate change and respiratory diseases*, European Respiratory Society publication, 2014.

[8] Götschke, J., Mertsch, P., Bischof, M., Kneidinger, N., Matthes, S., Renner, E.D., Schultz, K., Traidl-Hoffmann, C., Duchna, H.W., Behr, J. and Schmude, J., "*Perception of climate change in patients with chronic lung disease*," PloS one, vol. 12, no. 10, p. p.e0186632., 2017.

[9] Weiland, S. K., Hüsing, A., Strachan, D. P., Rzehak, P., & Pearce, N., "*Climate and the prevalence of symptoms of asthma, allergic rhinitis, and atopic eczema in children*," Occupational and environmental medicine, vol. 61, no. 7, pp. 609-615, 2004.

[10] Bellazzi, R., & Zupan, B., "*Predictive data mining in clinical medicine: current issues and guidelines*," International journal of medical informatics, vol. 77, no. 2, pp. 81-97, 2008.

[11] Kim, K. J., "*Financial time series forecasting using support vector machines*," Neurocomputing, vol. 55, pp. 307-319, 2003.

[12] Kusiak, A., Zheng, H., & Song, Z., "*Short-term prediction of wind farm power: A data mining approach*," IEEE Transactions on energy conversion, vol. 24, no. 1, pp. 125-136, 2009.

[13] Nogales, F. J., Contreras, J., Conejo, A. J., Espínola, R., "*Forecasting next-day electricity prices by time series models*," IEEE Transactions on power systems, vol. 17, no. 2, pp. 342-348, 2002.

[14] Xue, J. L., Ma, J. Z., Louis, T. A., & Collins, A. J., "*Forecast of the number of patients with end-stage renal disease in the United States to the year 2010*," ournal of the American Society of Nephrology, vol. 12, no. 12, pp. 2753-2758, 2001.

[15] Chaurasia, V., & Pal, S., "*Data Mining Approach to Detect Heart Dieses*," International Journal of Advanced Computer Science and, vol. 2, no. 4, pp. 56-66, 2013.

[16] Soni, J., Ansari, U., Sharma, D., & Soni, S., "*Predictive data mining for medical diagnosis: An overview of heart disease prediction*," International Journal of Computer Applications, vol. 17, no. 8, pp. 43-48, 2011.

[17] Delen, D., Walker, G., & Kadam, A., "*Predicting breast cancer survivability: a comparison of three data mining methods*," Artificial intelligence in medicine, vol. 34, pp. 113-137, 2005.

[18] Patra, S., & Thakur, G. S. M., "*A proposed neuro-fuzzy model for adult asthma disease diagnosis*," Computer Science & Information Technology (CS & IT), vol. 3, pp. 191-205, 2013.

[19] Mishra, N., Singh, D., Bandil, M. K., & Sharma, P., "*Decision support system for asthma (DSSA)*," v, vol. 3, pp. 549-445, 2013.

[20] Kaku, Y., Kuramoto, K., Kobashi, S., & Hata, Y., "*Predict time series data for the number of asthmatic attacks in Himeji by Fuzzy-AR model.*," in In Emerging Trends in Engineering and Technology (ICETET), 2012 Fifth International Conference on.

[21] Zarandi, M. F., Zolnoori, M., Moin, M., & Heidarnejad, H. , " *A fuzzy rule-based expert system for diagnosing asthma*," Scientia Iranica. Transaction E, Industrial Engineering, vol. 17, no. 2, pp. 129-142, 2010.

[22] Anguera, A., Barreiro, J. M., Lara, J. A., & Lizcano, D., "*Applying data mining techniques to medical time series: an empirical case study in electroencephalography and stabilometry*," Computational and structural biotechnology journal, vol. 14, pp. 185-199, 2016.

[23] Abe, H., Yokoi, H., Ohsaki, M., & Yamaguchi, T., "*Developing an integrated time-series data mining environment for medical data mining*," in Seventh IEEE International Conference on Data Mining , 2007.

[24] Zhang, S., Zhang, C., & Yang, Q, "*Data preparation for data mining.*," Applied artificial intelligence, vol. 17, no. 5-6, pp. 375-381, 2003.

[25] Cios, K. J., & Moore, G. W., "*Uniqueness of medical data mining*," Artificial intelligence in medicine, vol. 26, no. 1-2, pp. 1-24, 2002.

[26] Tustison, B., Harris, D., & Foufoula-Georgiou, E., "*Scale issues in verification of precipitation forecasts.*," Journal of Geophysical Research: Atmospheres, vol. 106, no. D11, pp. 11775-11784., 2001.