فصلنامه رویکردهای نوین در تحقیقات علوم پایه، فنی و مهندسی
سال ششم، شماره ۲۱، تابستان ۱۴۰۲
ISSN: ۲٦٤٥-٥٤٣٩
www.jonarbset.ir

تشخیص بیماری سرطان ریه با بکارگیری الگوریتم طبقه بندی داده کاوی

سهیلا رحیمی[۱]

[۱] گروه کامپیوتر، واحد بیضا، دانشگاه آزاد اسلامی، بیضا، ایران.

چکیده

سرطان ریه یکی از بیماریهای خطرناک است که باعث مرگ گسترده افراد در سراسر جهان می شود. تشخیص زودرس سرطان ریه تنها روش ممکن برای بهبود شانس بیمار برای بقا است. بنابراین وجود تکنیک هایی برای تشخیص به روز تومورهای سرطانی در مراحل اولیه ضروری می باشد. با استفاده از تکنیک های داده کاوی می توان پیش بینی زود هنگام سرطان ریه را انجام داد. از اینرو در این تحقیق یک مدل برای تشخیص زودهنگام سرطان ریه با استفاده از الگوریتم های دسته بند متا موجود در داده کاوی ارائه شده که با بکارگیری آن می توان خیلی سریع و با دقت بسیار بالا افراد مبتلا به سرطان ریه را شناسایی و تشخیص داد. مدل پیشنهادی پژوهش حاضر که با استفاده از الگوریتم های ترکیبی دسته بند متا پیاده سازی شده است، که مطابق با نتایج شبیه سازی از دقت بسیار بالایی برابر با ۹۷/۱۴٪ برخوردار می باشد. این میزان دقت با استفاده از ترکیب دو الگوریتم Logit Boost از دسته بند متا و Decision Stump از دسته بند trees حاصل گردید.

واژه‌های کلیدی: بیماری سرطان ریه، تکنیک داده کاوی، انتخاب ویژگی، دقت تشخیص

۱. مقدمه

بر اساس آمار سازمان بهداشت جهانی (WHO) در سراسر دنیا، سرطان ریه شایع‌ترین سرطان از لحاظ شیوع و مرگ و میر به‌شمار می‌رود. در سال ۲۰۰۸، ۱/۶۱ میلیون مورد جدید، و ۱/۳۸ میلیون مرگ و میر ناشی از سرطان ریه گزارش شد-۱] [۳. بالاترین نرخ‌ها در اروپا و آمریکای شمالی است. جمعیتی که احتمال ایجاد سرطان ریه در آن‌ها بیشتر است افراد بالای ۵۰ سال هستند که سابقه کشیدن سیگار دارند. بر خلاف میزان مرگ و میر در مردان، که از بیش از ۲۰ سال پیش در حال کاهش است، نرخ مرگ و میر ناشی از سرطان ریه در زنان در طول دهه‌های گذشته افزایش یافته‌است، و اخیراً در حال تثبیت است. در آمریکا، خطر مادام‌العمر ابتلا به سرطان ریه در مردان ۸٪ و در زنان ۶٪ است. بنابراین از آنجا که اگر این بیماری دیر هنگام تشخیص داده شود طول عمر و در نتیجه شانس بقای شخص کاهش می یابد. امروزه در ایران و بسیاری از کشور های دنیا کیفیت و امید به زندگی در انواع مختلف سرطان بهبود یافته است[۵-۴] ، اما‌همچنان بسیاری از این بیماران در مرحله ای تشخیص داده می شوند که به علت گسترش بیماری نمی توان برای آنها کار زیادی انجام داد. به عنوان مثال در اسکاتلند به علت تشخیص دیر هنگام و گسترش بیماری، تنها ۱۵ ٪ بیماران دارای سرطان ریه از نوع سلول های غیر کوچک یا NSCLC در زمان تشخیص قابل جراحی شدن بوده اند.

تشخیص اولیه سرطان ریه موجب افزایش شانس زنده بودن افراد می شود ، که شکست منجر به مرگ می شود. داده کاوی یک روش قدرتمند برای کمک به افراد در بهداشت ، علم و مهندسی است. از یک روش یادگیری برای درک الگوهای داده استفاده می کند[۶]. این تکنیک ها در حال استخراج اطلاعات مخفی از پایگاه داده های بزرگ هستند که به یافتن روابط و الگوهایی از داده ها کمک می کند. طبقه بندی ترکیبی برای طبقه بندی مجموعه داده های سرطان ریه مورد استفاده قرار می گیرد زیرا دقت بسیار بالاتری نسبت به سایر طبقه بندی کننده ها دارد[۷].

در کلینیک های پزشکی بیماران برای انجام یکسری آزمایشات ، متحمل هزینه های هنگفت و زیادی می شوند و حتی بیماران ممکن است دچار آسیب های روحی و روانی زیادی شوند اما دیگر نجات جان افراد امکان پذیر نیست زیرا که اکثر سرطانها در مراحل آخر و خیلی دیر تشخیص داده می شوند بنابراین در این تحقیق نشخیص سرطان ریه با استفاده از الگوریتم های ترکیبی(ترکیب های دوتایی) دسته بند متا صورت پذیرفته، که هدف از انجام این پژوهش ایجاد یک مدل تشخیص زودهنگام سرطان ریه می باشد بطوریکه این مدل دارای بالاترین میزان دقت و سرعت باشد. اهداف کلان تحقیق حاضر عبارتند از:

۱- با استفاده از تکنیک های داده کاوی می توان دقت مدل های تشخیص سرطان ریه را افزایش داد.

۲- ترکیب دسته بندهای متا می تواند به تشخیص دقیق تر سرطان ریه کمک کند.

پس از بدست آوردن داده های هدف، ممکن است کیفیت لازم برای شروع و انجام مراحل داده کاوی نداشته باشند. بنابراین لازم است که یکسری از عملیات های پیش روی داده ها اعمال گردد. در ابتدا لازم است که با استفاده از فیلتر SMOTE[۱] ، تعداد نمونه های موجود در هر دو کلاس "بیماران و افراد سرطانی" و " بیماران غیر سرطانی" متعادل شوند زیرا که این کار در کیفیت و نتیجه کار تاثیر به سزایی دارد.

در مرحله اول، ترکیبی از تمام الگوریتم های دسته بند متا به همراه مهمترین و معروف ترین الگوریتم های مربوط به دسته بندهای دیگر موجود در نرم افزار وکا بر روی داده های موجود در دیتاست اعمال و مدل مورد نظر با بالاترین دقت بدست آمده است.

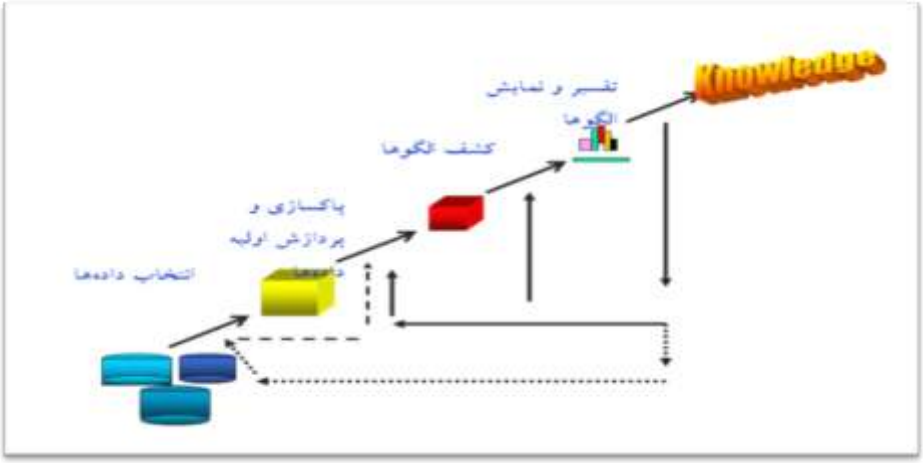
سپس در مرحله دوم، ابتدا مهمترین ویژگی های موجود در دیتاست که بیشترین تاثیر در تشخیص سرطان ریه داشته اند توسط Selected Attribute انتخاب و مجددا ترکیبی از تمام الگوریتم های دسته بند متا به همراه مهمترین و معروف ترین الگوریتم های مربوط به دسته بندهای دیگر موجود در نرم افزار وکا بر روی داده های موجود در دیتاست اعمال و مدل مورد

[۱] Synthetic Minority Oversampling Technique(SMOTE)

نظر با بالاترین دقت و سرعت بدست آمده است. بطوریکه بتوان با کاهش بعد و حجم محاسباتی داده ها علاوه بر دقت ، سرعت مدل تشخیصی را افزایش داد.

۲. داده کاوی

داده کاوی عبارت از اقتباس یا استخراج دانش از مجموعه ی بسیار حجیم داده ها به عبارتی دیگر داده کاوی فرایندی است که با استفاده از تکنیکهای هوشمند، دانش را از مجموعه ای از داده ها استخراج می کند[۸]. مفاهیم مهم در داده کاوی شامل خوشه بندی و دسته بندی می باشد. شکل (۱) فرآیند داده کاوی را نشان می دهد.



شکل ۱ – شمای کلی فرایند داده کاوی

۳. نرم افزار وکا

تا به امروز نرم افزارهای تجاری و آموزشی فراوانی برای داده کاوی در حوزه های مختلف داده ها به دنیای علم و فناوری عرضه شده‌اند. هریک از آنها با توجه به نوع اصلی داده هایی که مورد کاوش قرار می‌دهند، روی الگوریتم های خاصی متمرکز شده‌اند. مقایسه دقیق و علمی این ابزارها باید از جنبه های متفاوت و متعددی مانند تنوع انواع و فرمت داده های ورودی، حجم ممکن برای پردازش داده ها، الگوریتم های پیاده سازی شده، روش های ارزیابی نتایج، روش های مصور سازی[۲]، روش های پیش پردازش[۳]داده ها، واسط های کاربر پسند[۴] ، پلت فرم[۵]های سازگار برای اجرا، قیمت و در دسترس بودن نرم افزار صورت گیرد. از آن میان، نرم افزار وکا با داشتن امکانات بسیار گسترده، امکان مقایسه خروجی روش های مختلف با هم، راهنمای خوب، واسط گرافیگی کارآ، سازگاری با سایر برنامه‌های ویندوزی و از همه مهمتر وجود کتابی بسیار جامع و مرتبط با آن معرفی می‌شود[۹].

میزکار[۶]وکا، مجموعه‌ای سازمان‌یافته از الگوریتم های یادگیری ماشین ، هنر و ابزارهای پیش‌پردازش اطلاعات است. این نرم-افزار به گونه‌ای طراحی شده است که می‌توان به سرعت، روش‌های موجود را به صورت انعطاف‌پذیری روی مجموعه‌های جدید

---

[۲] Visualization

[۳] Preprocessing

[۴] User friendly

[۵] Platform

[۶] Workbench

داده، آزمایش نمود. این نرم‌افزار، پشتیبانی‌های ارزشمندی را برای کل فرآیند داده کاوی های تجربی فراهم می‌کند. این پشتیبانی‌ها، آماده سازی داده‌های ورودی، ارزیابی آماری چارچوب‌های یادگیری و نمایش گرافیکی داده‌های ورودی و نتایج یادگیری را در بر می‌گیرند. همچنین، هماهنگ با دامنه وسیع الگوریتم‌های یادگیری، این نرم‌افزار شامل ابزارهای متنوع پیش پردازش داده‌هاست. این جعبه ابزار متنوع و جامع، از طریق یک واسط متداول در دسترس است، به نحوی که کاربر می‌تواند روش‌های متفاوت را در آن با یکدیگر مقایسه کند و روش‌هایی را که برای مسایل مدنظر مناسب‌تر هستند، تشخیص دهد.

فرنک و همکاران [۱۰] بیان کردند که نرم‌افزار وکا در دانشگاه Waikato واقع در نیوزلند توسعه یافته است و اسم آن از عبارت Waikato " Environment for knowledge Analysis " استخراج گشته است. همچنین Weka ، نام پرنده‌ای با طبیعت جستجوگر است که پرواز نمی‌کند و در نیوزلند، یافت می‌شود. این سیستم به زبان جاوا نوشته شده و بر اساس لیسانس عمومی و فراگیر GNU[۷] انتشار یافته است. نرم افزار وکا تقریباً روی هر پلت فرمی اجرا می‌شود و نیز تحت سیستم عامل‌های لینوکس، ویندوز، مکینتاش، و حتی روی یک منشی دیجیتالی شخصی ، آزمایش شده است.

این نرم‌افزار، یک واسط همگون برای بسیاری از الگوریتم‌های یادگیری متفاوت، فراهم کرده است که از طریق آن روش‌های پیش پردازش، پس از پردازش[۸]و ارزیابی نتایج طرح های یادگیری روی همه مجموعه های داده های موجود، قابل اعمال است . همچنین این نرم‌افزار شامل مجموعه متنوعی از ابزارهای تبدیل مجموعه‌های داده‌ها، همانند الگوریتم‌های گسسته سازی[۹]می-باشد. در این محیط می‌توان یک مجموعه داده را پیش پردازش کرد، آن را به یک طرح یادگیری وارد نمود و دسته‌بندی حاصله و کارآیی‌اش را مورد تحلیل قرار داد.

در اغلب برنامه‌های کاربردی داده کاوی، یادگیری ماشینی، بخش کوچکی از سیستم نرم‌افزاری نسبتاً بزرگی را شامل می-شود. در صورتی که نوشتن برنامه کاربردی داده کاوی مد نظر باشد، می‌توان با برنامه‌نویسی اندکی به برنامه‌های وکا از داخل کد شخصی دسترسی داشت. همچنین اگر پیدا کردن مهارت در الگوریتم‌های یادگیری ماشینی مدنظر باشد، اجرای الگوریتم-های شخصی بدون درگیر شدن با جزییات دست و پا گیر مثل خواندن اطلاعات از یک فایل، اجرای الگوریتم‌های فیلترینگ یا تهیه کد برای ارزیابی نتایج یکی از خواسته‌ها می‌باشد که وکا دارای همه این مزیت‌ها است. برای استفاده کامل از این ویژگی، باید با ساختارهای پایه‌ای داده‌ها آشنا شد [۱۰].

۱.۳. الگوریتم های یادگیری

مفهوم الگوریتم در داده‌کاوی و یا در یادگیری ماشینی به مجموعه‌ای از استنتاج‌ها و محاسبات اطلاق می‌شود که مدلی از داده‌ها را ارائه می‌نماید. به منظور ایجاد مدل، در ابتدا الگوریتم به آنالیز داده‌های ارائه شده می‌پردازد تا انواع خاصی از الگوها یا روندها را جستجو نماید. سپس از نتایج این آنالیز به دفعات استفاده می‌کند تا به پارامترهای مطلوب برای ایجاد مدل داده‌کاوی دست یابد .در مرحله‌ی بعد این پارامترها جهت استخراج الگوهای عملیاتی و فرآیندهای آماری دقیق در تمامی مجموعه داده به کار گرفته می‌شوند.

- **الگوریتم های یادگیری دسته بند Bayes**

دسته بند مبتنی بر رابطه نظریه بیز از یک چهارچوب احتمالی برای حل مسائل دسته بندی استفاده می‌کند. از جمله الگوریتم های این دسته بند می توان به Naïve Bayes و Naïve Bayes Simple و ... اشاره کرد.

---

[۷] General Public License

[۸] Postprocessing

[۹] Discretization

- **الگوریتم های یادگیری دسته بند Trees**

درخت تصمیم ابزاری است که در آن نمونه ها به نحوی دسته بندی می شوند که از ریشه به سمت پایین درخت رشد کرده و در نهایت به گره برگ می رسند. هر گره داخلی نشان دهنده یک ویژگی بوده و برگهای این درخت نشان دهنده یک کلاس یا مجموعه ای از جوابها هستند. این الگوریتم برای داده های ورودی با ابعاد و حجم بالا مناسب و نسبت به نویز مقاوم است و همچنین برای تقریب توابع گسسته به کار می رود ، اما کارایی این الگوریتم نسبت به دیتاست های مختلف، متفاوت می باشد.
از جمله الگوریتم های این دسته می بتوان به Decision stump که برای استفاده توسط روش‌هایboosting طراحی شده است، اشاره کرد. این الگوریتم برای مجموعه‌های داده عددی یا رده‌ای، درخت تصمیم‌گیری یک سطحی می‌سازد. این الگوریتم، با مقادیر از دست رفته، به صورت مقادیر مجزا برخورد کرده و شاخه سومی از درخت توسعه می‌دهد[۱۱].

- **الگوریتم های یادگیری دسته بند Functions**

این الگوریتم برای برآورد مقادیر گسسته بر اساس مجموعه‌ای از متغیر(های) وابسته استفاده می‌شود. به بیان ساده‌تر این الگوریتم احتمال رخداد یک رویداد را بر حسب گنجاندن داده‌ها در یک تابع لجستیکی پیش‌بینی می‌کند. از این‌رو به نام رگرسیون لاجیت نیز شناخته می‌شود. از آنجا که این الگوریتم، احتمال را پیش‌بینی می‌کند، مقادیر خروجی آن بین ۰ تا ۱ می باشند. اگر چه رگرسیون لجستیک یک مدل بسیار قدرتمند محسوب می شود، لیکن بستگی به تجربه مدل ساز در ارتباط با داده ها و آنالیز داده ها دارد و مدل ساز باید بر اساس تجربه ، پاسخ صحیح را در رابطه با متغیرها مشخص کند.

- **الگوریتم های یادگیری دسته بند Rules**

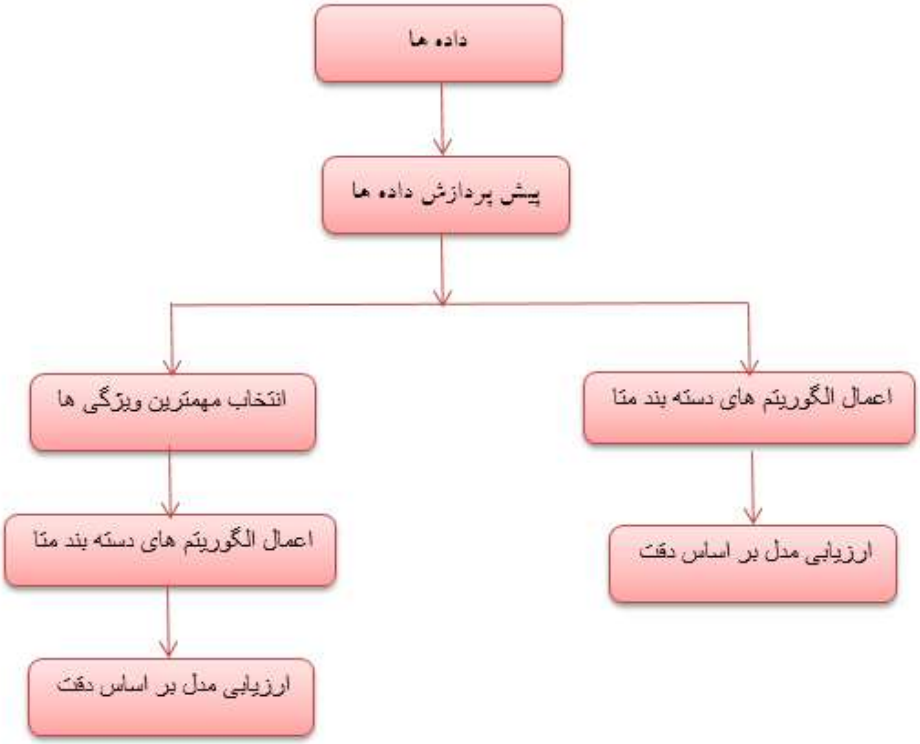
این دسته بندها دانش خروجی خود را به صورت یک مجموعه از قوانین «اگر-آنگاه» نشان می‌دهند. هر قانون یک بخش شرایط (LHS: Left Hand Side) و یک بخش نتیجه (RHS: Right Hand Side) دارد. بدیهی است اگر تمام شرایط مربوط به بخش مقدم یک قانون درباره یک رکورد خاص درست تعبیر شود، آن قانون آن رکورد را پوشش می‌دهد. دو معیار Accuracy و Coverage برای هر قانون قابل محاسبه است که هر چه میزان این دو معیار برای یک قانون بیشتر باشد، آن قانون؛ قانونی با ارزش‌تر محسوب می‌شود.

- **الگوریتم های یادگیری دسته بند Lazy**

الگوریتم های یادگیری دسته بند Lazy نمونه‌های آموزشی را ذخیره می‌کنند و تا زمان رده بندی هیچ کار واقعی انجام نمی‌دهند.

۴. اجرای روش پیشنهادی

در این بخش به روش اجرای تحقیق و ایجاد مدل تشخیص سرطان ریه با استفاده از الگوریتم های دسته بند متا، بر اساس دیتاستی که شامل ۱۶ ویژگی و ۳۱۰ نمونه می باشد، خواهیم پرداخت و تمام مراحل را بطور کامل شرح خواهیم داد. نمای کلی روش پیشنهادی این پژوهش در شکل (۲) نشان داده شده است.



شکل ۲ – روند پیاده سازی مراحل روش پیشنهادی

۱.۴. داده های تحقیق

در این تحقیق، از داده های بین المللی که حاوی اطلاعات بیماران سرطان ریه می باشد، استفاده گردید. این دیتاست شامل ۳۱۰ نمونه و ۱۶ ویژگی است. ویژگی های موجود در این دیتاست در جدول (۱) نمایش داده شده است.

جدول ۱- ویژگی های موجود در دیتاست

ردیف	نام ویژگی	ردیف	نام ویژگی
۱	Gender	۹	Allergy
۲	Age	۱۰	Wheezing
۳	Smoking	۱۱	Alcohol Consuming
۴	Yellow Fingers	۱۲	Coughing
۵	Anxiety	۱۳	Shortness of Breath
۶	Peer Pressure	۱۴	Swallowing Difficulty
۷	Chronic Disease	۱۵	Chest Pain
۸	Fatigue	۱۶	Lung Cancer

ویژگی Lung Cancer به عنوان فیلد کلاس برای تشخیص اینکه فرد دچار بیماری سرطان ریه می باشد یا خیر، در نظر گرفته شده است.

۲.۴. انتخاب مهم ترین ویژگی های بیماران مبتلا به سرطان ریه

کاهش بعد یکی از مهمترین و چالش برانگیزترین فعالیت‌ها در حوزه یادگیری ماشین و تشخیص الگو می باشد و بخش قابل توجهی از تحقیقات این دو حوزه، به روش‌های کاهش بعد و انتخاب ویژگی اختصاص می یابد. درواقع کاهش بعد، یکی از اساسی‌ترین روش های رایج، در حوزه های مرتبط با داده‌ها است و پروسه ای لازم، در پیش‌پردازش پروسه‌های مرتبط با الگوریتم های داده کاوی می باشد.

الگوریتم‌های کاهش بعد، با کاهش تعداد متغیرهای مجموعه داده‌ها و یا انتقال داده‌ها به فضایی با متغیرهای کمتر، کار پژوهش بر روی داده‌های معمول آزمایشگاهی را آسان و آنالیز داده‌های بسیار حجیم را امکان‌پذیر می‌کند. متغیر هایی که برای داده‌ها اندازه‌گیری می‌شوند، بعد و یا همان ویژگی های داده خوانده می‌شوند. تعداد بسیار زیاد ویژگی‌ها، علی‌رغم به وجود آوردن فرصت‌ها و اطلاعات جدید، باعث افزایش بار محاسباتی در تحلیل و آنالیز داده‌ها و کمبود فضا حین ذخیره‌سازی آنان می شوند و کارایی سیستم‌های آنالیز داده‌ها را دشوار می سازد. از این رو کاهش بعد، اولین قدم در حل این گونه مسائل شناخته می شود.

ویژگی هایی که دارای اطلاعات تفکیک کننده[۱] بیشتری بوده و اشیا را بهتر توصیف می کنند مطلوب سیستم‌ها می باشند. از طرفی هرچه تعداد ویژگی‌ها کمتر باشد، پیچیدگی سیستم نیز کمتر می شود. در واقع سیستم‌های تشخیص الگو و یادگیری ماشین با ویژگی‌های کمتر و توصیف کنندگی بیشتر مطلوب‌تر می باشند. هر چند که تعداد ویژگی کمتر و توصیف کنندگی بیشتر تا حدی در تناقض می باشند و باید موازنه ای بین آن ها به وجود آید. در مواردی نیز داده ها دارای ویژگی‌های ناقص[۱]، نامربوط[۲] و یا زائد[۳] می باشند. ویژگی‌های ناقص حاوی اطلاعات اندکی هستند و استفاده از آنان در سیستم نه تنها سودی ندارد بلکه با افزایش بعد، حجم داده‌ها را بسیار بیشتر کرده و محاسبات را پیچیده می کنند.

بنابراین انتخاب ویژگی به فرآیندی گفته می‌شود که منجر به کاهش بعد مجموعه داده اصلی می‌شود. مجموعه ابعاد انتخاب شده باید شامل اطلاعات کافی و قابل اعتماد که نشان دهنده مجموعه اصلی هستند، باشد. پس از انجام کاهش بعد، ابعاد اضافی و بدرد نخور حذف می‌شوند و ابعادی که به بهترین شکل ماهیت داده را حفظ می‌کنند در مجموعه داده باقی می مانند. در بسیاری از تحقیق ها تلاش می‌شود تا با کاهش بعد، دقت الگوریتم ها و همچنین سرعت آنها بهبود پیدا کند.

در طرح حاضر نیز پس از ورود و بارگذاری داده ها در نرم افزار وکا، در ابتدا با استفاده از Selected Attributes ، مهمترین ویژگی ها استخراج و انتخاب می شوند که این ویژگی ها در جدول (۲) نمایش داده شده است.

جدول ۲- انتخاب مهم ترین ویژگی های بیماران مبتلا به سرطان ریه در این پژوهش

ردیف	نام ویژگی
۱	Allergy
۲	Wheezing

Discriminative

Sparse

Irrelevant

Redundant

Alcohol Consuming	۳
Coughing	۴
Swallowing Difficulty	۵

سپس همانگونه که در بخش های قبل بیان شد ترکیبی از الگوریتم های مختلف دسته بند متا و دیگر دسته بندهای موجود در نرم افزار وکا بر روی داده ها اعمال و نتایج حاصل از آنها با هم مقایسه گردید.

۵. نتایج شبیه سازی

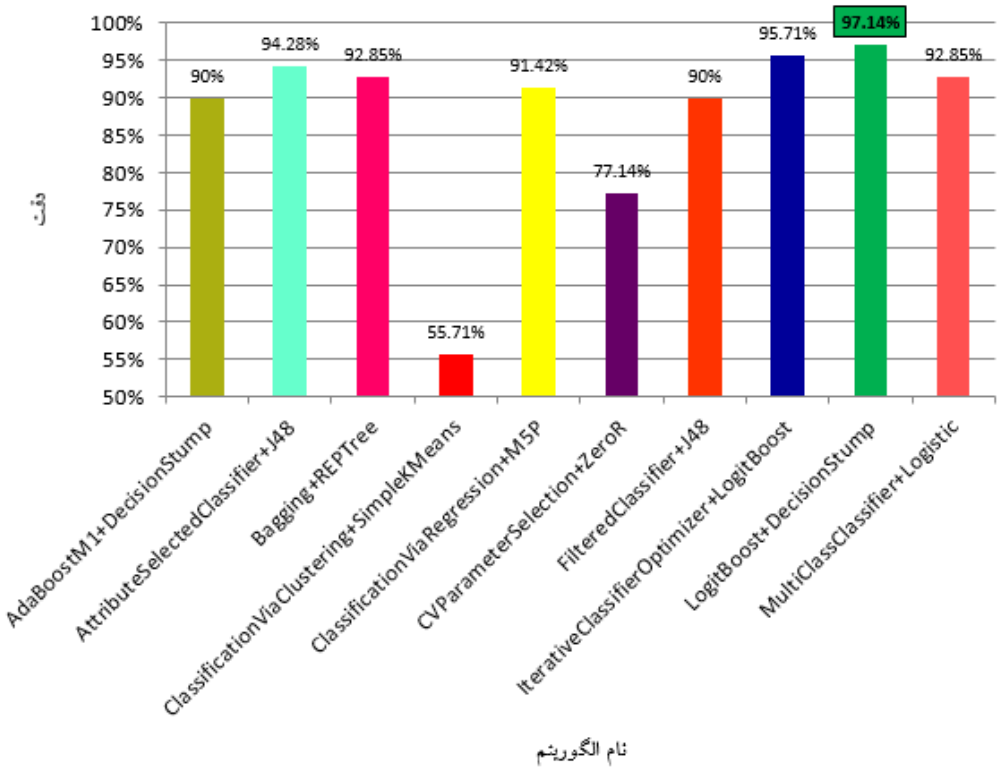
یکی از اصلی ترین اهداف این پژوهش تشخیص زودهنگام سرطان ریه با استفاده از الگوریتم های ترکیبی دسته بند متا می باشد. ترکیب های مختلفی از تمام الگوریتم های دسته بند متا و الگوریتم های دسته بندهای دیگر موجود در نرم افزار وکا در نظر گرفته شده است و در نهایت آن الگوریتم هایی که بالاترین میزان دقت را داشته اند انتخاب و در اینجا ذکر گردیده است. نتایج حاصل از اعمال طبقه بندهای مختلف متا بر روی کل داده های موجود در دیتاست در جدول (۳) نشان داده شده است.

جدول ۳- نتایج حاصل از اعمال طبقه بندهای مختلف متا بر روی کل داده های موجود در دیتاست

ردیف	نام الگوریتم	دقت(درصد)
۱	AdaBoostM1+DecisionStump	٪۹۰
۲	AttributeSelectedClassifier+J48	٪۹۴,۲۸
۳	Bagging+REPTree	٪۹۲,۸۵
۴	ClassificationViaClustering+SimpleKMeans	٪۵۵,۷۱
۵	ClassificationViaRegression+M5P	٪۹۱,۴۲
۶	CVParameterSelection+ZeroR	٪۷۷,۱۴
۷	FilteredClassifier+J48	٪۹۰
۸	IterativeClassifierOptimizer+LogitBoost	٪۹۵,۷۱
۹	LogitBoost+DecisionStump	٪۹۷,۱۴
۱۰	MultiClassClassifier+Logistic	٪۹۲,۸۵
۱۱	MultiClassClassifierUpdateable+SGD	٪۹۴,۲۸
۱۲	MultiScheme+ZeroR	٪۷۷,۱۴
۱۳	OrdinalClassClassifier+J48	٪۹۴,۲۸
۱۴	RandomCommittee+RandomTree	٪۹۲,۸۵
۱۵	RandomizableFilteredClassifier+IBK	٪۸۷,۱۴
۱۶	RandomSubSpace+REPTree	٪۹۱,۴۲
۱۷	Stacking+ZeroR	٪۷۷,۱۴
۱۸	ThresholdSelector+Logistic	٪۹۴,۲۸

۱۹	Vote+ZeroR	٪۷۷,۱۴
۲۰	WeightedInstancesHandlerWrapper+ZeroR	٪۷۷,۱۴

با توجه به شکل (۳) اکثر الگوریتم های دسته بند متا از دقت بسیار خوبی(دقت بالای ۹۰٪) برخوردار می باشند ولی نتایج حاصل نشان می دهد که ترکیب دو الگوریتم LogitBoost از دسته بند متا و DecisionStump از دسته بند trees دارای بالاترین میزان دقت و برابر با ۹۷/۱۴٪ می باشد.



شکل ۳ – نتایج حاصل از اعمال طبقه بندهای مختلف متا بر روی کل داده های موجود در دیتاست

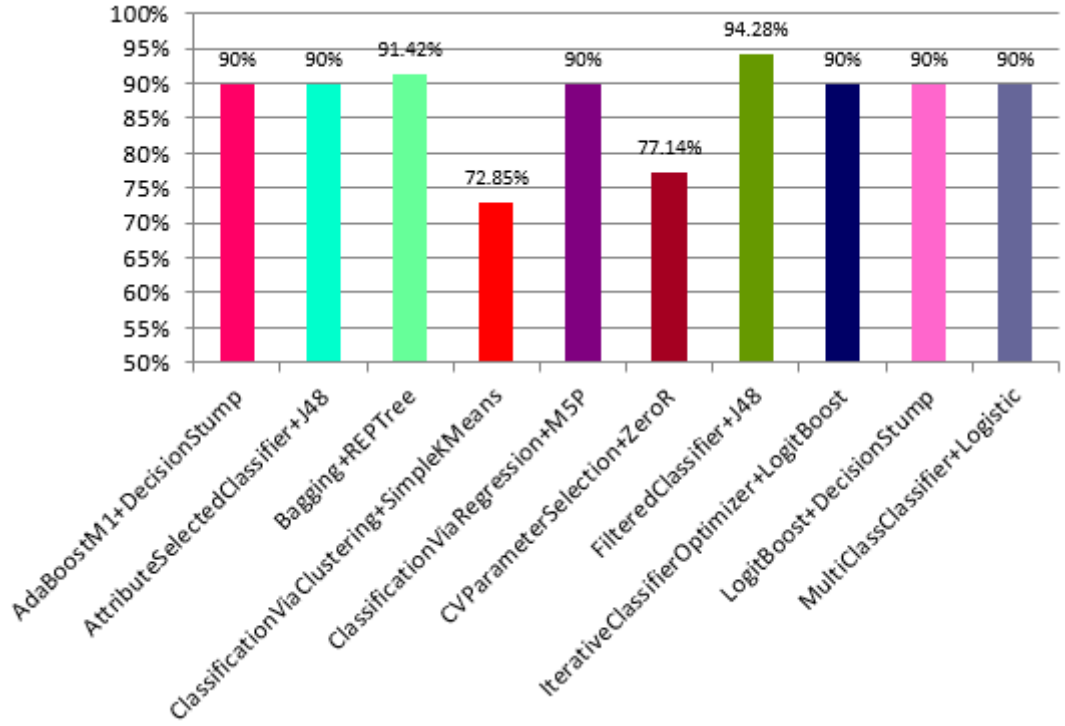
پس از اعمال فیلتر AttributeSelection بر روی داده هایی که شامل اطلاعات سرطان ریه می باشد تعداد ویژگی ها از ۱۶ عدد به ۵ فیلد کاهش یافته است که این ویژگی ها عبارتند از: allergy ، wheezing ، alcohol_consuming ، coughing و swallowing_difficulty . بعد از انتخاب مهمترین ویژگی ها، تمام الگوریتم های ترکیبی طبقه بند متا بر روی داده ها اعمال گردید که نتایج آن در جدول (۴) نشان داده شده است.

جدول ۴- نتایج حاصل از اعمال طبقه بندهای مختلف متا بر روی داده های موجود در دیتاست بعد از انتخاب مهمترین ویژگی ها

ردیف	نام الگوریتم	دقت(درصد)
۱	AdaBoostM1+DecisionStump	٪۹۰
۲	AttributeSelectedClassifier+J48	٪۹۰
۳	Bagging+REPTree	٪۹۱,۴۲
۴	ClassificationViaClustering+SimpleKMeans	٪۷۲,۸۵

۵	ClassificationViaRegression+M5P	٪۹۰
۶	CVParameterSelection+ZeroR	٪۷۷,۱۴
۷	FilteredClassifier+J48	٪۹۴,۲۸
۸	IterativeClassifierOptimizer+LogitBoost	٪۹۰
۹	LogitBoost+DecisionStump	٪۹۰
۱۰	MultiClassClassifier+Logistic	٪۹۰
۱۱	MultiClassClassifierUpdateable+SGD	٪۹۱,۴۲
۱۲	MultiScheme+ZeroR	٪۷۷,۱۴
۱۳	OrdinalClassClassifier+J48	٪۹۰
۱۴	RandomCommittee+RandomTree	٪۹۵,۷۱
۱۵	RandomizableFilteredClassifier+IBK	٪۹۴,۲۸
۱۶	RandomSubSpace+REPTree	٪۹۰
۱۷	Stacking+ZeroR	٪۷۷,۱۴
۱۸	ThresholdSelector+Logistic	٪۹۱,۴۲
۱۹	Vote+ZeroR	٪۷۷,۱۴
۲۰	WeightedInstancesHandlerWrapper+ZeroR	٪۷۷,۱۴

پس از انتخاب مهمترین ویژگی های موجود در دیتاست همانگونه که اطلاعات شکل (۴) نشان می دهد اکثر الگوریتم های دسته بند متا باز هم از دقت بسیار خوبی(دقت بالای ٪۹۰) برخوردار می باشند که نتایج حاصل نشان می دهد که ترکیب دو الگوریتم RandomCommittee از دسته بند متا و RandomTree از دسته بند trees دارای بالاترین میزان دقت و برابر با ٪۹۵,۷۱ می باشد. همانگونه که نتایج نشان می دهد حتی با کاهش ابعاد و پیچیدگی مسئله، الگوریتم های دسته بند متا همچنان از دقت بالایی برخوردار هستند بطوریکه حتی در مواردی الگوریتم های ترکیبی عملکرد بهتری نسبت به وقتی که تمام ویژگی های مسئله موجود بودند، داشته اند مانند الگوریتم های ترکیبی ClassificationViaClustering و SimpleKMeans ، FilteredClassifier و J48 ، RandomizableFilteredClassifier و IBK (که با رنگ آبی نشان داده شده است.)



شکل ۴ – نتایج حاصل از اعمال طبقه بندهای مختلف متا بر روی داده های موجود در دیتاست بعد از انتخاب مهمترین ویژگی ها

۶. نتیجه گیری

پژوهشی که توسط ناصر و ابوناصر در [۱۲] انجام شد هدف تشخیص سرطان ریه با استفاده از الگوریتم شبکه های عصبی مصنوعی بود. برای انجام این پژوهش از دیتاستی شامل ۱۶ ویژگی و ۳۱۰ نمونه استفاده گردید بطوریکه هم شامل داده های سرطانی و هم داده های غیر سرطانی می باشد. پس از اعمال الگوریتم شبکه عصبی مصنوعی بر روی این داده ها دقتی برابر با ۹۶/۶۷ ٪ بدست آمد. اما در پژوهش حاضر با استفاده از اعمال الگوریتم های ترکیبی دسته بند متا بر روی داده های موجود در دیتاست مورد بحث به دقتی برابر با ۹۷,۱۴٪ دست یافتیم که نشان از کارایی و دقت بالاتر مدل پیشنهادی پژوهش حاضر دارد. باید توجه داشت که برای یک مقایسه عادلانه باید تمام شرایط از جمله داده های هر دو پژوهش با هم یکسان باشند به همین دلیل برای انجام پژوهش حاضر نیز از همان داده هایی استفاده گردید که در پژوهش ناصر و ابوناصر[۱۲] استفاده شده بود که این دیتاست شامل ۱۶ ویژگی و ۳۱۰ نمونه بوده است بطوریکه هم شامل داده های سرطانی و هم داده های غیر سرطانی می باشد. در پژوهش حاضر پس از انتخاب داده های مورد نظر و اعمال الگوریتم های ترکیبی مختلف از دسته بند متا، دقت های متفاوتی بدست آمده است که بالاترین دقت یعنی ۹۷,۱۴٪ مربوط به ترکیب دو الگوریتم LogitBoost از دسته بند متا و DecisionStump از دسته بند trees می باشد.

مراجع

۱.Agrawal, A., Misra, S., Narayanan, R., Polepeddi, L., & Choudhary, A. (2011) ,“ A lung cancer outcome calculator using ensemble data mining on SEER data,” In Proceedings of the Tenth International Workshop on Data Mining in Bioinformatics (5), pp. 1-9, ACM.

۲.Agrawal, A., Misra, S., Narayanan, R., Polepeddi, L., & Choudhary, A. (2012) ,“ Lung cancer survival prediction using ensemble data mining on SEER data,” Scientific Programming, **20**(1), pp. 29-42.

۳.Ahmed, Kawsar, Emram, A.A., Jesmin, T., Mukti, R.F., Rahman, M.Z., Ahmed, F., (2013),“Early detection of lung cancer risk using data mining,” Asian Pacific Journal of Cancer Prevention, **14**(1), pp: 595-598.

٤.Ahmed, S. R. A., Al Barazanchi, I., Mhana, A., & Abdulshaheed, H. R. (2019) ,“ Lung cancer classification using data mining and supervised learning algorithms on multi-dimensional data set,” Periodicals of Engineering and Natural Sciences, **7**(2), pp.438-447.

٥.Barash, O., Peled, N., Tisch, U., Bunn, P. A., Hirsch, F. R., & Haick, H. (2012) ,“ Classification of lung cancer histology by gold nanoparticle sensors. Nanomedicine: Nanotechnology,” Biology and Medicine, **8**(5), pp. 580-589.

٦.Yadav, A. K., Tomar, D., & Agarwal, S. (2013) ,“ Clustering of lung cancer data using Foggy K-means,” In Recent Trends in Information Technology (ICRTIT), 2013 International Conference on IEEE, pp. 13-18.

۷.Singh, N., Singh, P., & Kaur, R. (2019) ,“ Design and Development a Hybrid Classifier to Improve Lung Cancer Diagnosis, Journal of the Gujarat Research Society, **21**(15).

۸.Han, J., Kamber, M., Pei, J. (2011) ,“ Data mining: concepts and techniques,” third edition. Morgan Kaufmann.

۹.Mikut, R., Reischl, M. (2011) ,“ Data mining tools. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,” **1**(5), pp. 431-443.

۱۰. Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., & Trigg, L. (2009) ,“ Weka-a machine learning workbench for data mining. In Data mining and knowledge discovery handbook,” (pp. 1269-1277). Springer, Boston, MA.

۱۱. Kamel, H., Abdulah, D., & Al-Tuwaijari, J. M. (2019, June) ,“ Cancer Classification Using Gaussian Naive Bayes Algorithm,” In 2019 International Engineering Conference (IEC), pp. 165-170.

۱۲. Nasser, I. M., Abu-Naser, S. S. (2019) ,“ Lung Cancer Detection Using Artificial Neural Network,” International Journal of Engineering and Information Systems (IJEAIS), **3**(3), pp. 17-23.