سیستم های اطلاعاتی هوشمند
نشریه علمی دانشگاه آزاد اسلامی
مجلد۱،شماره ۴
صفحات۱-۹
مقاله پژوهشی

تحلیل عوامل اثرگذار بر «ابتلاء به بیماری کووید-۱۹»، با استفاده از داده‌های بالینی بیماران، با روش‌های داده‌کاوی

آرش بحیرائی[۱] و منصور اسماعیل‌پور[۲*]

*نویسنده مسئول، دریافت: ۱۴۰۰/۱۱/۳، بازنگری: ۱۴۰۱/۹/۲۴، پذیرش: ۱۴۰۱/۱۰/۸

[۱] دانشجوی دکتری مدیریت فناوری اطلاعات، واحد همدان، دانشگاه آزاد اسلامی، همدان، ایران bahiraei@gmail.com
[۲] دانشیار گروه کامپیوتر، دانشکده فنی مهندسی، واحد همدان، دانشگاه آزاد اسلامی، همدان، ایران esmaeilpour@iauh.ac.ir

چکیده

مقدمه و هدف: ابتلاء فراگیر به ویروس کووید-۱۹ و شیوع سریع آن در سال‌های اخیر، به یکی از معضلات کادر درمانی در جهان تبدیل شده است. هنگامی که این بیماران به مراکز درمانی وارد و مراحل اولیه تشخیص ویروس را طی می‌کنند، تفکیک بهینه و قاعده‌مند آنان از سایر بیماران، می‌تواند کمک شایانی به کادر درمان کند. از سوی دیگر تشخیص سریع‌تر بیمار مشکوک به کووید- ۱۹، در جلوگیری از شیوع بیشتر آن، موثر خواهد بود. در این پژوهش قصد داریم با استفاده از داده‌های تصاویر پزشکی مبتلایان، و روش‌های داده‌کاوی، به تعیین مشخصه‌های اثرگذار در تشخیص و تفکیک مبتلایان به کووید-۱۹ از سایر بیماران، در مراکز درمانی بپردازیم. همچنین به استخراج قواعد تصمیم‌گیری، جهت بهینه سازی فرآیند تصمیم‌گیری کادر درمان، خصوصا" در تفکیک بهتر مبتلایان از سایر بیماران، پرداخته شده است.

طراحی/روش/رویکردها: در این پژوهش از داده‌کاوی با روش «رافست»، «شبکه عصبی مصنوعی» و «درخت تصمیم» جهت استخراج قواعد تصمیم‌گیری استفاده کرده‌ایم. به این منظور، ابتدا داده‌ها پاکسازی شده و مشخصه‌های غیرمرتبط حذف گردید، سپس با استفاده از نرم افزارهای اکسل، رُزتا، و وِکا، اقدام به گسسته‌سازی و تقسیم داده‌ها به آزمون و آموزش کرده و پس از کاهش مشخصه‌ها، به استخراج قواعد تصمیم‌گیری و تحلیل آن، اقدام شد.

یافته‌ها: در این پژوهش مجموعا" شامل ۹۵۰ داده تصویر رادیوگرافی بیماران است که از این تعداد ۳۱۱ سطر یا ۳۲/۷ ٪ مربوط به زنان و ۵۵۹ سطر یا ۵۸/۸ ٪ مربوط به مردان است. با توجه به توابع جانسون و ژنتیک و درخت تصمیم، مشخصه‌های «سن»، «تعداد روزهای تهیه تصویر پس از شروع علائم و یا بستری شدن»، «وضعیت آزمایش RT_PCR» و «نوع تصویر پزشکی»، و بر اساس تابع هولتس بعلاوه مشخصه «درخواست اکسیژن»، اثرگذاری بیشتری در تشخیص و تفکیک بیماران دارند. با استفاده از روش «رافست» بوسیله توابع جانسون و ژنتیک و هولتس، دقت قوانین مدل هریک ۸۳٪ شد که بوسیله تابع جانسون ۴۶۵ قانون و توسط تابع ژنتیک تعداد ۳۳۱۶ قانون و توسط تابع هولتس تعداد ۶۲ قانون استخراج شد. با استفاده از روش «درخت تصمیم» و الگوریتم J48 آن، دقت مدل برابر۸۲٪ و تعداد قوانین برابر ۹ قانون است. همچنین در روش «شبکه عصبی مصنوعی» با الگوریتم چندلایه پرسپرون، دقت مدل تقریبا" برابر۹۷٪ می‌باشد از بقیه روش‌ها بالاتر است. به طور متوسط ۹/۰۸ روز پس از شروع علائم و با بستری شدن از بیماران تصاویر پزشکی تهیه شده است. در نهایت پنج قانون ترکیبی حاصل از اجرای این روش‌ها تبیین گردید.

کلمات کلیدی: کرونا، Covid-19، جانمایی بیمار، درخت تصمیم، راف، شبکه عصبی، Montreal.

۱- مقدمه

حمد خدای را می‌گوییم که توفیق و فرصتی عطا فرمود[۱] تا در این مجال، به واکاوی و تحلیل عوامل اثرگذار در ابتلاء به بیماری کووید-۱۹، توجه شود.

ویروس کرونا در پایان دسامبر ۲۰۱۹ در شهر ووهان[۲] چین ظاهر شد و در ۱۲ مارس ۲۰۲۰ پس از انتشار به بسیاری از کشورها، به عنوان یک اپیدمی توسط سازمان بهداشت جهانی[۳] اعلام شد (سعیدی‌فر، ۱۳۹۹). تا قبل از سال ۲۰۰۲ تصور می‌شد کرونا ویروس‌ها، مشکل حادی ایجاد نمی‌کنند و برای اولین بار با شیوع سارس[۴] در کشور چین و ۲۹ کشور دیگر که باعث ابتلاء بیش از ۸۰۰۰ نفر و مرگ حدود ۱۰ درصد از مبتلایان شد، محققان دریافتند که کروناویروس‌ها عامل ایجاد بیماری شدیدتر از سرماخوردگی هم می‌شوند (طاهری، ۱۳۹۹).

[۱] سوره مبارکه هود، آیه ۸۸ .
[۲] Wuhan
[۳] World Health Organization
[۴] SARS

کووید-۱۹ یک کرونا ویروس، بعد از سندرم شدید تنفسی حاد یا سارس و سندرم تنفسی خاورمیانه یا مرس، متعلق به گروه بتا است، که شیوع بالایی داشته و شرایط مخاطره آمیز بسیاری را به وجود آورده است (چِن و همکاران، ۲۰۲۰).

با توجه به وجود داده‌های معتبر درباره ویروس کرونا، نیاز تبدیل این داده‌ها، به اطلاعات موثر برای تصمیم‌گیری مدیران در حوزه‌های مرتبط، از جمله حوزه پزشکی، مشاهده می‌شود. یکی از راه‌های موثر جهت این تبدیل، استفاده از روش‌های داده‌کاوی است که توجه زیادی را درعلوم مختلف از جمله پزشکی به خود جلب نموده است. داده‌کاوی در حوزه سلامت کاربردهای زیادی مانند تشخیص عوامل موثر در ابتلاء به یک بیماری، یافتن قوانین و الگوهایی جهت تشخیص دقیق‌تر بیماری‌ها، گروه‌بندی بیماران جهت سازماندهی و مدیریت هوشمند بیماران و بیمارستان، و کاربردهای دیگر، اشاره کرد.

در این مقاله سعی شده، با توجه به ضرورت تشخیص دقیق‌تر کرونا، با استفاده از داده‌های (مشخصه‌های) بالینی بیماران و روش‌های داده‌کاوی، به تحلیل عوامل موثر بر ابتلاء به کرونا ویروس (کووید-۱۹) و استخراج قوانینی مبتنی بر اطلاعات بالینی افراد با هدف «تشخیص سریع‌تر» و «پیشگیری دقیق‌تر» بیماری، بپزداریم.

۲- طرح مسئله

مسئله این است که با توجه به اطلاعات بالینی و تصاویر پزشکی بیماران که برچسب کلاس آن‌ها مشخص است، قصد داریم با استفاده از داده‌های تصاویر پزشکی مبتلایان، و روش داده‌کاوی، به تعیین مشخصه‌های اثرگذار در تشخیص مبتلایان به کووید-۱۹، هنگام طی مراحل تشخیص و درمان، در مراکز پزشکی-درمانی، بپردازیم. همچنین بتوانیم این مشخصه‌ها، را معین کرده و میزان اهمیت آن را برای متخصصان و پزشکان مرتبط با این بیماری، تبیین کنیم.

در گام دوم تحقیق نیز، به استخراج قواعد تصمیم‌گیری، جهت بهینه سازی فرآیند تصمیم‌گیری کادر درمان و مدیران مراکز درمانی، خصوصا" در جانمایی بهتر مبتلایان، اقدام نماییم و به متخصصین، پیشنهادات مناسبی در این راستا، ارائه کنیم.

۳- مبانی نظری

در سال ۱۹۵۶ در کنفرانس بنیاد راکفلر۲ برگزاری کنفرانسی را بر عهده داشت که چشم اندازش به قرار زیر بود: امکان استفاده از کامپیوترها و شبیه‌سازی در هر زمینه از یادگیری و سایر حوزه‌های هوش مصنوعی. در همین کنفرانس بود که اصطلاح هوش مصنوعی مورد استفاده عمومی قرار گرفت. به طور کلی هوش مصنوعی را به این صورت می‌توان تعریف کرد: فرآیندهای کامپیوتری که سعی دارند فرآیندهای تفکر انسان را تقلید نمایند که این فرآیندها با فعالیت‌های که نیاز به استفاده از هوش دارند در ارتباط هستند. الگوریتم‌های یادگیری مانند رافست[۵] و فناوری‌های مرتبط به آن‌ها به عنوان بخش‌هایی از هوش مصنوعی می‌باشند. اخیرا" افزایش فعالیت‌های پژوهشی در زمینه هوش مصنوعی که رشد روزافزونی داشته است بیانگر این مطلب است که مدل‌های هوش مصنوعی مانند رافست از نظر قابلیت پیشگویی و طبقه‌بندی الگو، قوی می‌باشند (ابریشمی، ۱۳۸۷).

با توجه به پیشرفت روز افزون تکنولوژی اطلاعات، داده کاوی در واقع کشف دانش از داده‌های پایگاه‌های اطلاعاتی علمی است که برای تصمیم‌گیری‌های دقیق‌تر، بسیار کاربرد دارد. روش‌های داده‌کاوی جهت یافتن الگوهای مناسب در تشخیص پزشکی و درمان، استفاده می‌شود. روش‌های متفاوتی در داده‌کاوی برای استخراج اطلاعات وجود دارد که می‌توانند برای پیش‌بینی مورد استفاده قرار بگیرند. از آن جمله می‌توان به شبکه‌های‌عصبی، درخت تصمیم و رگرسیون لجستیک اشاره کرد (استافورد و ویجی‌وارانی، ۱۹۸۴ و ۲۰۱۳).

امروزه شبکه های عصبی مصنوعی به عنوان یکی از روش‌های کلاسیک در حل بسیاری از مسائل مورد استفاده قرار می‌گیرد که بسته به وضعیت اطلاعات داده‌ای و آموزش‌پذیری شبکه، نتایج مطلوبی را ارائه می‌کند (اسماعیل‌پور و همکار، ۱۳۸۶).

در سال‌های اخیر سیستم‌های تشخیص بیماری گسترش یافته که در شرایط مختلف می‌توان از آن‌ها بهره برد. با توجه به گسترش دسترسی به پایگاه‌های داده حوزه درمان، روش‌های داده‌کاوی همچون درخت تصمیم و شبکه‌های عصبی مورد توجه قرار گرفته‌اند (لینگاریج، ۲۰۱۵).

تحلیل خوشه‌ای نیز یک روش دیگر است که در آن هیچ فرضی در مورد تعداد گروه‌ها یا ساختمان آن‌ها در نظر گرفته نمی‌شود. خوشه‌بندی تنها بر اساس مشابهت‌ها یا فواصل (عدم شباهت‌ها) انجام می‌شود (بحیرائی و همکاران، ۱۳۹۳).

یادگیری ماشین در مسائلی مانند آنالیز داده و داده‌کاوی کاربردهایی گسترده پیدا نموده و امروزه داده‌های بدون برچسب به راحتی جمع‌آوری می‌شوند اما راهی برای استفاده از آن‌ها نیست زیرا یادگیری بانظارت، از داده‌های آموزشی که برچسب کلاس آن‌ها مشخص است، برای تعیین برچسب کلاس، برای داده‌های تست، استفاده می‌کنند (افتخاری و عارفیان، ۱۳۹۲).

«درخت تصمیم» یکی از مشهورترین و قدیمی‌ترین روش‌های ساخت مدل رده‌بندی می‌باشد. در الگوریتم‌های رده‌بندی مبتنی بر درخت تصمیم، دانش خروجی به صورت درختی از حالات مختلف مقادیر و ویژگی‌ها ارائه می‌شود. نمایش دانش به شکل درخت باعث شده است که رده‌های مبتنی بر درخت تصمیم کاملا" قابل تفسیر باشند (حقیقت و همکاران،۱۳۹۱).

البته کارآیی فرآیند داده‌کاوی با میزان دقت و کیفیت داده‌ها ارتباط مستقیم دارد. مرحله پیش‌پردازش داده، امکان تبدیل داده به شکل مناسب را توسط الگوریتم داده‌کاوی مفروض، فراهم می‌کند (اسماعیل‌پور و همکاران، ۱۳۹۹).

۴- پیشینه تحقیق

گروه کرونا ویروس‌ها در انسان، پستانداران و پرندگان از نظر ژنوتایپی و سرولوژی با چهار نوع آلفا، بتا، گاما و دلتا، شناسایی شده‌اند که در انسان، توسط جنس آلفا و بتا ایجاد بیماری می‌کند نوع جدید این ویروس به نام کووید-۱۹ درسال ۲۰۱۹ برای اولین بار از شهر ووُهان در چین شایع شد، و پس از آن در ده‌ها کشور از سراسر جهان را آلوده کرده و باعث ایجاد اپیدمی گسترده در سطح جهان شده است (چِن و همکاران، ۲۰۲۰).

از سوی دیگر، یکی از روش‌های مناسب برای پیشبینی و تشخیص در حوزه‌های پزشکی رویکرد دادهکاوی می‌باشد. داده‌کاوی روشی داده مدار و بر مبنای یادگیری و کشف الگوی پنهان در میان داده‌های حقیقی می‌باشد که از این الگو جهت پیشبینی، استفاده می‌کند (لاکشمی، ۲۰۱۳).

بسکابادی و همکاران دریافتند که نتایج مدل رگرسیون و طبقه‌بندی درختی نشان می‌دهد که از متغیرهای کمی به ترتیب اهمیت سن، زمان بستری تا نتیجه، فاصله شروع علائم تا نتیجه آزمایش و فاصله بستری تا نتیجه آزمایش و همین‌طور از متغیرهای کیفی، جنسیت، در نتیجه درمان بیماران موثر می‌باشند (بسکابادی و همکاران، ۱۳۹۹).

علیزاده‌فرد و صفاری‌نیا در خصوص عوامل موثر بر کرونا در پژوهش خود به این نتیجه رسیدند که اضطراب بیماری کرونا (به صورت منفی) و همبستگی اجتماعی ناشی از بیماری کرونا (به صورت مثبت) با سلامت روان همبستگی دارد. همچنین مشخص شد که اضطراب و همبستگی اجتماعی ناشی از بیماری کرونا، به ترتیب ۲۴ و ۴۸ درصد از تغییرات سلامت روان را پیش بینی می‌کنند (علیزاده‌فرد و صفاری‌نیا، ۱۳۹۸).

---

[۵] ROUGH SET

باتیننی و همکاران در پژوهش خود چهار نوع داده آلودگی به کووید-۱۹ روزانه را برای پیش بینی ۶۰ روزه کل عفونت‌ها (در ایالات متحده آمریکا، برزیل، هند و روسیه) در نظر گرفتند. برای انجام این کار، یک مدل یادگیری ماشین (ML) به نام Fb-Prophet را به کارگرفتند و نتایج تأیید کرد که تعداد کل موارد تأیید شده در چهار کشور تا پایان ژوئیه، درست بوده و پیش‌بینی‌ها انجام شده است. به عبارت دیگر نتایج نشان داد که در اواخر سپتامبر، شیوع تخمین زده شده، به ترتیب در ۷,۵۶ ، ۴,۶۵ ، ۳,۰۱ و ۱,۲۲ میلیون مورد در ایالات متحده آمریکا، برزیل، هند و روسیه برآورد شد. هم چنین دریافتند که برخی از دست کم گرفتن‌ها و بیش از حد ارزیابی کردن‌ها در اطلاعات بیماران روزانه، به چشم می‌خورد (باتیننی و همکاران، ۲۰۲۰).

آپادهایا و همکاران در مطالعه خود نتیجه گرفتند که عوامل موثر بر مرگ و میر ناشی از کووید-۱۹ در یک مدل تجربی نشان داده که نرخ مرگ و میر در هر میلیون، متغیری وابسته است، و امید به زندگی در بدو تولد ، تراکم پزشک، تحصیلات، چاقی، نسبت جمعیت بالای ۶۵ سال، شهرنشینی (تراکم جمعیت) و درآمد سرانه، با نرخ مرگ و میر مرتبط است. او و همکارانش از داده‌های ۱۸۴ کشور برای تخمین کمّی رگرسیون استفاده شد. همچنین نتایج این تحقیق نشان می‌دهد که چاقی، نسبت جمعیت بالای ۶۵ سال و شهرنشینی، تأثیر مثبت و معناداری بر مرگ و میر ناشی از کووید-۱۹ دارد. جای تعجب نیست، اما پایین بودن درآمد سرانه، تأثیر منفی و معناداری بر میزان مرگ ناشی از کووید-۱۹ دارد. (آپادهایا و همکاران، ۲۰۲۰).

پونو و همکاران در پژوهش خود با روش محاسبه CFR[۶] (نرخ مرگ و میر) با استفاده از مدل رگرسیون خطی ساده دریافتند که از ۲۱ مه ۲۰۲۰، سنگاپور، اندونزی و فیلیپین که سه کشور برتر SEA[۷] بودند، دارای بیشترین سابقه آلودگی به کووید-۱۹ (عفونت) هستند. در همان زمان، برونئی یک کشته داشت، در حالی که کامبوج، لائوس، تیمور-لسته و ویتنام هیچ کشته‌ای نداشته‌اند. اندونزی و فیلیپین به ترتیب با بیشترین CFR بین ۶/۶۶٪ و ۶/۵۹٪، و در تجزیه و تحلیل رگرسیون R2 با مقدار ۹۷/۹۵٪ و ۹۹/۴۳٪، را داشتند. سنگاپور با وجود عفونت زیاد، کمترین میزانCFR یعنی ۰/۰۶۸٪ را داشت (پونو و همکاران، ۲۰۲۱).

کومار داس و همکاران در تحقیق خود یک مدل غربالگری اتوماتیک کووید-۱۹ برای شناسایی بیماران مبتلا به این بیماری با استفاده از تصاویر اشعه‌ایکس قفسه سینه آن‌ها طراحی کردند. این مدل تصاویر را در سه دسته طبقه بندی می‌کند: کووید-۱۹ مثبت، سایر عفونت‌های ذات‌الریه، و بدون عفونت. سه طرح یادگیری مانند CNN، VGG-16 و ResNet-50 به طور جداگانه برای یادگیری مدل استفاده شده است. از یک مجموعه داده استاندارد رادیوگرافی کووید-۱۹ از مخزن Kaggle برای بدست آوردن تصاویر اشعه ایکس قفسه سینه نیز استفاده شده است. عملکرد مدل با هر سه طرح یادگیری ارزیابی شده است و نشان می‌دهد عملکرد VGG-16 در مقایسه با CNN و ResNet-50 عملکرد بهتری است. مدل با VGG-16 مقدار Accuracy برابر ۹۷/۶۷٪، مقدار Precision برابر ۹۶/۶۵٪، و مقدار Recall برابر ۹۶/۵۴٪ و مقدار F1-score برابر ۹۶/۵۹٪ را ارائه می‌دهد. ارزیابی عملکرد این مدل همچنین نشان می‌دهد که مدل محققان، برای نشان دادن کووید-۱۹ از دو مدل موجود بهتر عمل می‌کند (کومار داس و همکاران، ۲۰۲۱).

کانگ و همکاران در تحقیق خود در یافتند که باید یک مدل پیش بینی اصلاح شده کووید-۱۹ شدید در افراد آلوده به Sars-Cov-2 ایجاد کنند. آن‌ها مدل پیش‌بینی را برای بیماران شدید کووید-۱۹ بر اساس تاریخ بالینی از مرکز تومور بیمارستان یونیون وابسته به کالج پزشکی تونجی ، در چین ایجاد کردند. در مجموع ۱۵۱ مورد از ۲۶ ژانویه تا ۲۰ مارس ۲۰۲۰ وارد شد. سپس ۵ مرحله برای پیش‌بینی و ارزیابی مدل دنبال کردند: پیش‌پردازش داده‌ها، تقسیم داده‌ها، انتخاب ویژگی‌ها، ساخت مدل ، جلوگیری از انطباق بیش از حد[۸] و همچنین ارزیابی و ترکیب با الگوریتم‌های شبکه عصبی مصنوعی را انجام دادند. سپس آن‌ها نتایج را در ۵ مرحله پردازش کردند. در انتخاب ویژگی‌ها، ALB[۹] همبستگی منفی، قوی نشان داد (r=0.771, P<0.001) در حالی که GLB[۱۰] (r=0.661, P<0.001) و BUN[۱۱] (r=0.714, P<0.001) همبستگی مثبت و شدیدی با شدت کووید-۱۹ را نشان دادند. از تِنسورفلو[۱۲] که یک کتابخانه متن باز یادگیری ماشین است، هم برای ایجاد یک مدل شبکه عصبی استفاده شد. این مدل با سطح زیر منحنی ۰/۹۵۳ (۰/۹۸۲-۰/۸۸۹)، عملکرد پیش‌بینی خوبی را به دست آورد. نتایج بدست آمده نشان داد که مدل محققان، عملکرد برجسته‌ای را در پیشبینی دارد. عوامل GLB و BUN ممکن است دو عامل خطرناک برای کووید-۱۹ شدید باشند (کانگ و همکاران، ۲۰۲۱).

۵- روش تحقیق

یکی از راه‌های مهم در یافتن مشخصه‌های اثرگذار در ابتلا به کووید-۱۹، استفاده از روش‌های داده‌کاوی است که در این پژوهش از برخی روش‌های آن استفاده شده است. لازم به ذکر است که جهت انجام فرآیند داده‌کاوی می‌بایست این مراحل انجام شود: ۱.جمع آوری داده‌ها ۲.پیش پردازش داده‌ها ۳. کاربرد الگوریتم‌های داده‌کاوی ۴.عملیات پس‌پردازش (گارسیا و همکاران، ۲۰۰۷).

برای تهیه داده‌های اولیه، از داده‌های یک تحقیق در حوزه یادگیری عمیق، استفاده شده است. مجموعه داده‌های سی‌تی‌اسکن افراد کرونا-مثبت که ما برای این پژوهش استفاده کردیم، توسط جوزف کوهن ، فوق‌دکتری دانشگاه مونترال، و هماکارانش گردآوری شده‌است. آقای کوهن، مجموعه‌ای از داده‌های CT قفسه سینه که شامل کووید-۱۹ هم هست را برای پژوهش خودشان جهت تشخیص بیماری کرونا با استفاده از یادگیری عمیق و از طریق تصاویر قفسه سینه، جمع‌آوری کردند (کوهن و همکاران، ۲۰۲۰).

ما در این پژوهش تنها مشخصه‌هایی از دیتاست که مربوط به داده‌های بالینی بیماران هستند را استفاده کرده‌ایم و از فایل تصاویر و مشخصه‌های تصاویر موجود در دیتاست استفاده نکرده‌ایم و مشخصه‌های مربوط به آن‌ها را از دیتاست حذف کردیم. علاوه بر داده‌های کووید-۱۹، داده‌های مربوط به سایر بیماری‌های ریوی از جمله سارس، ARDS و ... نیز در این دیتاست وجود دارند. نکته حائز اهمیت در این دیتاست این است که عملیات پاکسازی داده‌ها و یکپارچه سازی توسط گروه آقای کوهن و همکاران انجام شده است. یعنی روی دقت داده‌ها تمرکز شده است و این محققان معتقدند که داده‌های حاشیه‌نویسی شده و دقیق، اولین قدم برای توسعه هر ابزار تشخیصی یا مدیریتی است، این داده‌های تصویری، مرتبط با ویژگی‌های بیمارستانی، در یک دیتاست عمومی که برای یادگیری ماشین طراحی شده است، قرار دارد و امکان توسعه موازی و سریع ابزارهای تشخیص و مدیریت و اعتبارسنجی مدل‌ها را فراهم می‌کند. به علاوه، از این داده‌ها می‌توان برای کارهای مختلف تحقیقاتی استفاده کرد (کوهن و همکاران، ۲۰۲۰). این موضوع اشاره‌ای بر روایی و پایایی داده‌های مطالعه شده در این گروه می‌باشد.

مشخصه‌های بالینی در دیتاست آقای کوهن و همکاران در جدول ۱ معرفی شده است:

[۶] Case Fatality Rates
[۷] Southeast Asian (SEA) countries
[۸] Overfitting
[۹] Albumin
[۱۰] Globulin
[۱۱] Blood Urea Nitrogen
[۱۲] TensorFlow

جدول ۱- شرح مشخصه‌های دیتاست (کوهن و همکاران، ۲۰۲۰).

مشخصه (Attribute)	شرح هر مشخصه
Offset	هر تصویر، چند روز پس از شروع علائم یا بستری شدن، تهیه شده. اگر در گزارش " after a few days" ثبت شده باشد، مقدار ۵ روز فرض شده است.
Sex	مرد (M)، زن (F)، یا «خالی».
Age	سن بیمار بر اساس «سال».
Finding	نوع مشکل تنفسی تشخیص داده شده.
RT_PCR_positive	بله (Y)، خیر (N)، یا «خالی» در صورت عدم گزارش یا ارائه نشدن گزارش.
Survival	اگر بیمار از این بیماری جان سالم به در برد و زنده ماند: بله (Y) در غیراین صورت: نه (N).
Intubated	بله (Y)، اگر بیمار در طول مدت این بیماری، در هر نقطه‌ای از بدنش، لوله گذاری (یا تهویه) داشته باشد. و اگر نداشته باشد: خیر (N). در صورت ناشناخته بودن یا بی اطلاعی: «خالی».
Intubated_present	بله (Y)، اگر بیمار (در زمان نمونه برداری) هر نقطه‌ای از بدنش، لوله گذاری (یا تهویه) دارد. و اگر نداشته باشد: خیر (N). در صورت ناشناخته بودن یا بی اطلاعی: «خالی».
went_icu	بله (Y)، اگر بیمار در طول مدت این بیماری، در ICU (واحد مراقبت ویژه) یا CCU (واحد مراقبت ویژه) بستری بود، و در غیر این صورت: خیر (N). اگر این وضعیت مشخص نبود: «خالی».
in_icu	بله (Y)، اگر بیمار (در زمان نمونه برداری) در ICU (واحد مراقبت ویژه) یا CCU (واحد مراقبت ویژه) بستری بوده، و در غیر این صورت: خیر (N). اگر این وضعیت مشخص نبود: «خالی».
needed_supplemental_O2	بله (Y)، اگر بیمار در طول مدت این بیماری، به اکسیژن مکمل نیاز داشته باشد، و در غیر این صورت: خیر (N). اگر این وضعیت مشخص نبود: «خالی».
Extubated	بله (Y)، اگر بیمار با موفقیت مرخص شده، و در غیر این صورت: خیر (N). اگر این وضعیت مشخص نبود: «خالی».
Temperature	دمای بدن بیمار به سانتیگراد، در زمان تصویربرداری.
View	رادیوگرافی قفسه سینه از پشت (PA)، رادیوگرافی قفسه سینه از جلو (AP)، رادیوگرافی قفسه سینه خوابیده به پشت (AP Supine)، رادیوگرافی قفسه سینه خوابیده به پهلو (L)، برای تصویربرداری اشعه ایکس. برای سی تی اسکن Axial یا Coronal.
Modality	چگونگی عکس برداری: CT یا X-ray یا سایر موارد.
Date	سال تصویربرداری.
url	آدرس اینترنتی مقاله یا وب سایتی که تصویر از آنجا گرفته شده است.
Class	برچسب کلاس. بله (Y)، اگرتشخیص بیمار کرونا قطعی است، و در غیر این صورت: خیر (N).

در جدول ۱ مشخصه‌های بالینی بیماران ذکر شده است که ستون Finding، نتیجه تشخیص برخی دیگر از انواع بیماری‌های تنفسی را نیز شامل می‌شود، که ما در این پژوهش تنها روی تشخیص قطعی کووید- ۱۹ تمرکز داریم، لذا ستون جدیدی به نام Class تشکیل داده شد که برچسب کلاس ما محسوب می‌شود و هر جا که مقدار ستون Finding مقدار «کووید-۱۹» را نشان می‌داد، مشخصه کلاس را برابر Y و هر جا تشخیصی غیر از کووید-۱۹ داده شده بود، مشخصه کلاس را برابر N قرار دادیم و نهایتا ستون Finding که تمامی نتایج آن به ستون Class منتقل شده بود را حذف کردیم. نهایتا" داده‌ها را در دو قالب .csv و xls. ذخیره کرده‌ایم.

لازم به ذکر است که داده‌ها از پراکندگی جغرافیایی مختلفی برخوردارند و سطرهای آن از مراکز درمانی کشورهای مختلفی از جمله ایران، پاکستان، چین، تایوان، کره جنوبی، ایتالیا، استرالیا، ویتنام، کانادا، اسپانیا، انگلستان، آلمان، پرتغال، ژاپن، مالت، و امریکا نمونه‌گیری و جمع‌آوری شده است، که نمونه داده‌های برخی از کشورها بیشتر، و برخی کمتر است.

حال می‌خواهیم از روش‌های «رافست» (توابع ژنتیک[۱۳] ، هولتس[۱۴] و جانسون[۱۵]) «درخت تصمیم»[۱۶] و «شبکه عصبی مصنوعی»[۱۷] برای عملیات داده‌کاوی استفاده کنیم و نتایج و قواعد استخراج شده آن‌ها را تبیین و تحلیل نمائیم. لذا نرم‌افزارهای رزتا[۱۸] نسخه ۴۱. ۴ . ۱ و وکا[۱۹] نسخه ۵. ۸ . ۳ و اکسل نسخه ۲۰۱۳ به کار برده شده است.

۱-۵- استفاده از الگوریتم «رافست» جهت تحلیل داده‌ها

تئوری «رافست» در سال ۱۹۸۰ میلادی توسط پوولاک[۲۰] به عنوان یک روش ریاضی مناسب برای انجام این نوع تحلیل داده‌ها، ارائه شد. این تکنیک ، که مکمل روش‌های آماری استنباط است، بینش جدیدی از خصوصیات داده‌ها را نشان می‌دهد. همانند تئوری‌های آماری، تمرکز اصلی این تکنیک بر روی بررسی روابط ساختاری داده‌ها به جای توزیع احتمال است. روش‌های مبتنی بر «رافست» خصوصاً برای استدلال در موارد مبهم در داده، ابزاری مطمئن است. برخی از کاربردهای آزمایش شده و شناخته شده برای این تئوری، عبارتند از: «یک الگوریتم کنترل برای کنترل فرآیند» که اِمرُزِک[۲۱] آن را مطرح کرد، «ایجاد رابطه‌ای کسب دانش برای سیستم‌های خبره با استفاده از تکنیک‌های یادگیری ماشین مبتنی بر رافست» که وانگ[۲۲] و زیارکو[۲۳] به آن اشاره کردند، «تجزیه و تحلیل و کاهش جدول تصمیم‌گیری» که زیارکو و پوولاک آن را طرح کردند، «تشخیص پزشکی» که پوولاک و اسلوینسکی[۲۴] آن را مطرح کردند (زیارکو، ۱۹۹۱).

اسماعیل‌پور و همکاران بیان کرده‌اند که یکی از موارد مهم در استفاده از الگوریتم‌های داده‌کاوی، نحوه ارزیابی الگوریتم و قوانین استخراج شده از آن است. از جمله این راه‌ها، تشکیل «ماتریس درهم‌ریختگی[۲۵]» است. ماتریس درهم‌ریختگی

[۱۳] Genetic Algorithm
[۱۴] Holts
[۱۵] Johnson Algorithm
[۱۶] Decision Tree
[۱۷] Artificial Neural Network
[۱۸] Rosetta
[۱۹] WEKA
[۲۰] Pawlak
[۲۱] Mrozek
[۲۲] Wong
[۲۳] Ziarko
[۲۴] Slowinski
[۲۵] Confusion Matrix

چگونگی عملکرد الگوریتم‌های دسته‌بندی را با توجه به مجموعه داده ورودی، و به تفکیک انواع دسته‌های مسئله دسته‌بندی، مطابق جدول ۲ نشان می‌دهد. ماتریس مفاهیم TN، FP، FN و TP به شرح ذیل است:

جدول ۲- جدول مشخصات ماتریس درهم‌ریختگی (اسماعیل‌پور و همکاران، ۱۳۹۸)

		برچسب پیش‌بینی شده	
		دسته مثبت	دسته منفی
برچسب شناخته شده	دسته مثبت	TP	FN
	دسته منفی	FP	TN

TP: این مقدار بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم دسته‌بندی نیز دسته آن‌ها به درستی مثبت تشخیص داده است.

TN: این مقدار بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها منفی بوده و الگوریتم دسته‌بندی نیز دسته آن‌ها به درستی منفی تشخیص داده است.

FP: این مقدار بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها منفی بوده و الگوریتم دسته‌بندی، دسته آن‌ها را به اشتباه، مثبت تشخیص داده است.

FN: این مقدار بیانگر تعداد رکوردهایی است که دسته واقعی آن‌ها مثبت بوده و الگوریتم دسته‌بندی، دسته آن‌ها را به اشتباه منفی، تشخیص داده است.

همانگونه که مشخص است دو مقدار TP، TN مهمترین مقادیری هستند که در یک مسئله دودسته‌ای باید بیشینه شوند. از آنجاکه هر دو مقدار TP، TN در صورت کسر قرارگرفته‌اند بنابراین می‌توان عنوان کرد که در رابطه ۱ به تمام دسته‌های موجود در مسئله دسته‌بندی توجه شده است. به همین دلیل «معیار دقت دسته‌بندی»[۲۶]، مشهورترین و عمومی‌ترین معیار محاسبه کارآیی الگوریتم‌های دسته‌بندی است. «معیار خطای دسته‌بندی»[۲۷] نیز دقیقاً برعکس «معیار دقت دسته‌بندی» است. کمترین مقدار آن برابر صفر (بهترین کارآیی) و بیشترین مقدار آن برابر یک (ضعیف‌ترین کارایی) است (صنیعی آباده و همکاران، ۱۳۹۴). رابطه ۱ معیار کارایی الگوریتم‌های دسته‌بندی را نشان می‌دهد:

$$CA = \frac{TN+TP}{TN+FN+TP+FP} \qquad (۱)$$

اکنون می‌خواهیم با استفاده از سه الگوریتم ژنتیک، هولتس و جانسون در روش «رافست»، و نرم‌افزار رُزتا، قوانین معتبر را استخراج کنیم (و البته از قوانین کم اعتبار، صرف نظر نماییم). بنابراین ابتدا داده‌هایی که مرحله پیش‌پردازش روی آن‌ها انجام شده را در نرم‌افزار رُزتا بارگذاری می‌کنیم. سپس داده‌ها را با روش انتروپی ، گسسته می‌کنیم. در مرحله بعد، چون تعداد سطرهای ما ۹۵۰ سطر است، داده‌ها را به این صورت به دو قسمت تقسیم می‌کنیم: بخش داده‌های آزمون (۲۰٪ معادل ۱۹۰ سطر) که برای آزمودن مدل و ارزیابی آن استفاده می‌شوند، و داده‌های آموزش (۸۰٪ معادل ۷۶۰ سطر) که برای ساخت مدل استفاده می‌شوند. در ادامه می‌بایست توسط نرم‌افزار رُزتا، روی داده‌های آموزش، کاهش‌دهنده را اعمال کنیم.

در این مرحله جهت اجرای الگوریتم «رافست»، به ترتیب هر یک از توابع را اجرا می‌کنیم و نتایج و قوانین استخراج شده را مورد ارزیابی قرار می‌دهیم:

الف) تابع جانسون، ابتدا مشخصه‌ها را بررسی و مشخصه‌هایی که بیشترین تأثیر را در ابتلا به کووید-۱۹ دارند را دخالت می‌دهد و سپس قوانین مربوطه را استخراج می‌کند. حاصل اجرای تابع جانسون در شکل ۱ نمایش داده شده است:

		Predicted		
		Y	N	
Actual	Y	122	2	0.983871
	N	30	36	0.545455
		0.802632	0.947368	0.831579
ROC	Class	Y		
	Area	0.953262		
	Std. error	0.014423		
	Thr. (0, 1)	0.956		
	Thr. acc.	0.9		

شکل -۱ اجرای تابع جانسون روی داده‌های بالینی و تصاویر پزشکی بیماران مشکوک یا مبتلا به کووید-۱۹

همانطور که در شکل ۱ مشاهده می‌شود، مقدار TP برابر ۱۲۲ و TN برابر ۳۶ و مقدار FP برابر ۳۰ و FN برابر ۲ می‌باشد و نهایتا" دقت مدل ۰/۸۳۱۵۷۹ یعنی ۸۳٪ است که دقت قابل قبولی به شمار می‌آید. همچنین از اجرای تابع جانسون، ۴۶۵ قانون استخراج شد که فهرست شدند و از میان آن‌ها، مهم ترین قوانینی که احتمال مثبت بودن ابتلاء به کووید-۱۹ را تایید می‌کنند بدین شرح است:

قانون۱: بیماران مرد، که سنشان را مشخص نکرده‌اند و RT_PCR آنان شفاف نیست، و بعد از شروع سال ۲۰۲۰ مراجعه کرده‌اند، هم چنین ۴ تا ۶۰ روز پس از شروع علائم، از آن‌ها تصویر (AP Supine) تهیه شده، و به آی سی یو منتقل شده باشند، احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایر بیماران جدا شوند.

قانون۲: بیماران زن، که سنشان را مشخص نکرده‌اند و RT_PCR آنان شفاف نیست، و بعد از شروع سال ۲۰۲۰ مراجعه کرده‌اند، هم چنین ۴ تا ۶۰ روز پس از شروع علائم، از آن‌ها تصویر (AP Supine) تهیه شده، و به آی سی یو منتقل شده باشند، احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایر بیماران جدا شوند.

قانون۳: بیمارانی که جنسیت و سنشان ثبت نشده، اما RT_PCR آنان مثبت است، و بعد از شروع سال ۲۰۲۰ مراجعه کرده‌اند، هم چنین معلوم نیست چند روز پس از شروع علائم، از آن‌ها تصویر (AP) تهیه شده، و وضعیت انتقالشان به آی سی یو معلوم نیست، احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایر بیماران جدا شوند.

قانون۴: بیماران مرد، که سنشان بین ۶۵ تا ۸۱ سال است و RT_PCR آنان مثبت است، و بعد از شروع سال ۲۰۲۰ مراجعه کرده‌اند، هم چنین ۴ تا ۶۰ روز پس از شروع علائم، از آن‌ها تصویر (PA) تهیه شده، و وضعیت انتقالشان به به آی سی یو مشخص نشده باشد، احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایر بیماران جدا شوند.

قانون۵: بیماران مرد، که سنشان ثبت نشده است و RT_PCR آنان شفاف نیست، و بعد از شروع سال ۲۰۲۰ مراجعه کرده‌اند، هم چنین ۱ تا ۴ روز پس از شروع علائم، از آن‌ها تصویر (AP Supine) تهیه شده، و به آی سی یو منتقل شده باشند، احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایر بیماران جدا شوند.

قانون۶: بیماران مرد، که سنشان بین ۶۵ تا ۸۱ سال است و RT_PCR آنان مثبت است، و بعد از شروع سال ۲۰۲۰ مراجعه کرده‌اند، هم چنین ۴ تا ۶۰ روز پس از شروع علائم، از آن‌ها تصویر (AP) تهیه شده، و به آی سی یو منتقل شده باشند، احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایر بیماران جدا شوند.

[۲۶] Classification Accuracy

[۲۷] Recall

ب) در تابع ژنتیک تمام مشخصه‌ها و عوامل مؤثر بر ابتلا به کووید-۱۹ تأثیر داده می‌شوند. حاصل اجرای تابع ژنتیک همانند نتایج تابع جانسون در شکل ۱ است و نهایتا" دقت مدل ۰/۸۳۱۵۷۹ یعنی ۸۳٪ است که دقت قابل قبولی به شمار می‌آید. همچنین از اجرای تابع ژنتیک، ۳۳۱۶ قانون استخراج شد که فهرست شدند و از میان آن‌ها، قوانین مهم که احتمال مثبت بودن ابتلاء به کووید-۱۹ را تایید می‌کنند بدین شرح است:

قانون۱: بیمارانی که سنشان را مشخص نکرده‌اند و RT_PCR آنان شفاف نیست، و بعد از شروع سال ۲۰۲۰ مراجعه کرده‌اند، هم چنین ۴ تا ۶۰ روز پس از شروع علائم، از آن‌ها تصویر (AP Supine) تهیه شده، و به آی سی یو منتقل شده باشند و مشخص نیست که اکسیژن کرده‌اند یا خیر، و زنده ماندند، احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایرین جدا شوند.

قانون۲: بیماران مرد که سنشان را مشخص نکرده‌اند و RT_PCR آنان شفاف نیست، و بعد از شروع سال ۲۰۲۰ مراجعه کرده‌اند، هم چنین ۴ تا ۶۰ روز پس از شروع علائم، از آن‌ها تصویر (AP Supine) تهیه شده، و به آی سی یو منتقل شده باشند، احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایرین جدا شوند.

قانون۳: بیمارانی که سنشان را مشخص نکرده‌اند و RT_PCR آنان شفاف نیست، و بعد از شروع سال ۲۰۲۰ مراجعه کرده‌اند، هم چنین ۴ تا ۶۰ روز پس از شروع علائم، از آن‌ها تصویر (AP Supine) تهیه شده، و در بخش آی سی یو نباشند، و نحوه دریافت اکسیژن آنان نامشخص است و نهایتا" زنده ماندند، هنوز احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایرین جدا شوند.

قانون۴: بیماران مردی که سنشان را مشخص نکرده‌اند و RT_PCR آنان شفاف نیست، و بعد از شروع سال ۲۰۲۰ مراجعه کرده‌اند، هم چنین ۴ تا ۶۰ روز پس از شروع علائم، از آن‌ها تصویر (AP Supine) تهیه شده، و وضعیت قراردادن لوله در بدنشان در طول مدت حضور مشخص نیست، و نهایتا" زنده ماندند، هنوز احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایرین جدا شوند.

قانون۵: بیماران مردی که سنشان را مشخص نکرده‌اند و RT_PCR آنان شفاف نیست، و بعد از شروع سال ۲۰۲۰ مراجعه کرده‌اند، هم چنین ۴ تا ۶۰ روز پس از شروع علائم، از آن‌ها تصویر (AP Supine) تهیه شده، و وضعیت درخواست اکسیژن آن‌ها مشخص نیست، و نهایتا" زنده ماندند، هنوز احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایرین جدا شوند.

قانون۶: بیماران مردی که سنشان را مشخص نکرده‌اند و RT_PCR آنان شفاف نیست، و بعد از شروع سال ۲۰۲۰ مراجعه کرده‌اند، هم چنین ۴ تا ۶۰ روز پس از شروع علائم، از آن‌ها تصویر (AP Supine) تهیه شده، و وضعیت لوله‌گذاری در بدنشان در زمان بستری مشخص نیست، و نهایتا" زنده ماندند، هنوز احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایرین جدا شوند.

قانون۷: بیماران مردی که سنشان بین ۶۵ تا ۸۱ سال است و RT_PCR آنان مثبت است، و بعد از شروع سال ۲۰۲۰ مراجعه کرده‌اند، هم چنین ۴ تا ۶۰ روز پس از شروع علائم، از آن‌ها تصویر (AP) تهیه شده، و وضعیت درخواست اکسیژن آن‌ها مشخص نیست، و معلوم نیست که زنده باشند، هنوز احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایرین جدا شوند.

ج) در تابع هولتس، با دقت بسیار بالا به بررسی عوامل مؤثر بر ابتلا به کووید-۱۹ پرداخته و عوامل را یک به یک مورد بررسی قرار داده و تعداد تکرارهای موجود در عوامل مؤثر بر ابتلا به کووید-۱۹ را بررسی کرده و قوانین و میزان اطمینان و خطای آن‌ها را استخراج می‌کند. خلاصه اجرای تابع هولتس نیز همانند شکل ۱ می‌باشد. نهایتا" دقت مدل همانند دو تابع ژنتیک و جانسون، برابر ۰/۸۳۱۵۷۹ یعنی ۸۳٪ است که دقت قابل قبولی به شمار می‌آید. همچنین از اجرای تابع هولتس، ۶۲ قانون استخراج شد که فهرست شدند و از میان آن‌ها، قوانین مهم که احتمال مثبت بودن ابتلاء به کووید-۱۹ را تایید می‌کنند بدین شرح است:

قانون۱: اگر نوع تصویری که تهیه شده است از نوع سی تی اسکن کرونل است، احتمال ابتلاء آنان به کووید-۱۹ زیاد است و بیمار باید از سایرین جدا شوند.

قانون۲: اگر بیمار در طی فرآیند درمان درخواست اکسیژن کرده است، احتمال ابتلاء آنان به کووید-۱۹ وجود دارد.

قانون۳: اگر تصویر ۷۴ روز یا بیشتر، بعد از شروع علائم یا بستری شدن، تهیه شده است، احتمال ابتلاء آنان به کووید-۱۹ کم است و بیمار احتمالا مبتلا به کووید-۱۹ نیست.

قانون۴: اگر سن بیمار بین ۸۱ تا ۸۹ سال است، احتمال ابتلاء آنان به کووید-۱۹ زیاد است و بیمار باید از سایرین جدا شود.

قانون۵: اگر بیمار با موفقیت ترخیص نشده است و مدتی در بیمارستان مانده، احتمال ابتلاء آنان به کووید-۱۹ زیاد است و بیمار باید از سایرین جدا شود.

قانون۶: اگر سن بیمار بین ۸۹ تا ۹۲ سال است، احتمال ابتلاء او به کووید-۱۹ کم است.

۲-۵- استفاده از «درخت تصمیم» جهت تحلیل داده‌ها

یکی از روش‌های متداول دسته‌بندی، درخت تصمیم است. ساختار درخت تصمیم در یادگیری ماشین، یک مدل پیشبینی کننده است که حقایق مشاهده شده در مورد یک پدیده را به استنتاج‌هایی در مورد مقدار هدف آن پدیده، نگاشت می‌کند. یادگیری درخت تصمیم یکی از رایجترین روش‌های داده‌کاوی است که به دلیل سادگی و کارآمدی‌اش، علی‌رغم مشکلاتی از جمله امکان وجود مشخصه دارای نویز و یا مشخصه فاقد مقدار، به شکل گسترده‌ای در مسائل مختلف، استفاده می‌شود (ویتن و همکاران، ۲۰۱۶ و هان و همکاران، ۲۰۱۱).

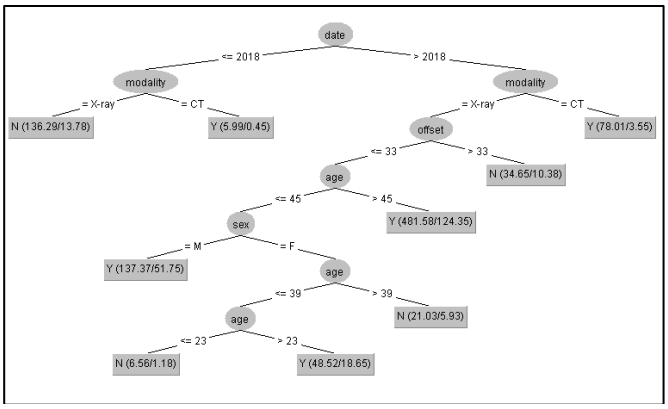
به عنوان مثال، اُولانو و همکاران از درخت تصمیم در تشخیص بیماری پارکینسون استفاده کردند. تشخیص بیماری پارکینسون در مراحل ابتدایی بیماری کاری دشوار است. آنها تلاش کردند که با استفاده از درخت تصمیم و استخراج ویژگیهای ظاهری و آزمایشگاهی به تشخیص این بیماری بپردازند. نتایج آزمایشها نشان از دقت بالای الگوریتم آن‌ها در تشخیص این بیماری دارد (اُولانو و همکاران، ۲۰۰۱).

تعداد سطرهای داده‌های پژوهش ما ۹۵۰ سطر است، در این حالت نیز داده‌ها را به دو قسمت تقسیم می‌کنیم: بخش داده‌های آزمون (۲۰٪ معادل ۱۹۰ سطر)، و داده‌های آموزش (۸۰٪ معادل ۷۶۰ سطر). که پس از بارگذاری داده‌ها در نرم‌افزار «وکا»، نتیجه حاصل از اجرای درخت تصمیم با الگوریتم J48 توسط این نرم‌افزار، در شکل ۲ نمایش داده شده است:

```
=== Summary ===

Correctly Classified Instances         143               75.2632 %
Incorrectly Classified Instances        47               24.7368 %
Kappa statistic                          0.3552
Mean absolute error                      0.3756
Root mean squared error                  0.4106
Relative absolute error                 80.1314 %
Root relative squared error             85.6308 %
Total Number of Instances              190

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.701    0.724      1.000    0.840      0.465    0.820     0.839     Y
               0.299    0.000    1.000      0.299    0.460      0.465    0.820     0.722     N
Weighted Avg.  0.753    0.454    0.821      0.753    0.706      0.465    0.820     0.797

=== Confusion Matrix ===

   a   b   <-- classified as
 123   0 |   a = Y
  47  20 |   b = N
```

شکل ۲- نتیجه اجرای درخت تصمیم با الگوریتم J48 روی داده‌های بالینی و تصاویر پزشکی بیماران مشکوک یا مبتلا به کووید-۱۹

همانطور که در شکل ۲ مشاهده می‌شود، مقدار TP برابر ۱۲۳ و TN برابر ۲۰ و مقدار FN برابر ۰ (صفر) و FP برابر ۴۷ می‌باشد و نهایتا" دقت مدل ۰/۸۲۱ یعنی ۸۲٪ است که دقت این مدل نیز قابل قبول است. همچنین جهت استخراج قوانین به نمایش درخت تصمیم می‌پردازیم که در شکل ۳ و توسط نرم‌افزار «وکا»، ترسیم شده است:



شکل ۳- نتیجه اجرای درخت تصمیم با الگوریتم J48 روی داده‌های بالینی و تصاویر پزشکی بیماران مشکوک یا مبتلا به کووید-۱۹

از جمله مهمترین قوانین استخراج شده حاصل از اجرای درخت تصمیم بر روی داده‌ها عبارتند از:

قانون۱: بیمارانی که قبل از سال ۲۰۱۸ (زمانی که شیوع کم بوده) مراجعه کرده‌اند و نوع تصویر پزشکی آن‌ها اشعه ایکس است، احتمال ابتلاء آنان به کووید-۱۹ کم است و نیازی نیست از سایرین جدا شوند.

قانون۲: بیمارانی که قبل از سال ۲۰۱۸ مراجعه کرده‌اند و نوع تصویر پزشکی آن‌ها سی تی اسکن است، احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایرین جدا شوند.

قانون۳: بیمارانی که بعد از سال ۲۰۱۸ مراجعه کرده‌اند و نوع تصویر پزشکی آن‌ها سی تی اسکن است، احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایرین جدا شوند.

قانون۴: بیمارانی که بعد از سال ۲۰۱۸ مراجعه کرده‌اند و نوع تصویر پزشکی آن‌ها اشعه-ایکس است اما بیش از ۳۳ روز پس از شروع علائم، از آن‌ها تصویر تهیه شده است، احتمال ابتلاء آنان به کووید-۱۹ کم است و نیازی نیست از سایرین جدا شوند.

قانون۵: بیمارانی با سن بیشتر از ۴۵ سال، که بعد از سال ۲۰۱۸ مراجعه کرده‌اند و نوع تصویر پزشکی آن‌ها اشعه-ایکس است و کمتر از ۳۳ روز پس از شروع علائم، از آن‌ها تصویر تهیه شده است، احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایرین جدا شوند.

قانون۶: بیماران مرد با سن کمتر از ۴۵ سال، که بعد از سال ۲۰۱۸ مراجعه کرده‌اند و نوع تصویر پزشکی آن‌ها اشعه-ایکس است و کمتر از ۳۳ روز پس از شروع علائم، از آن‌ها تصویر تهیه شده است، احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایرین جدا شوند.

۳-۵- استفاده از «شبکه عصبی مصنوعی» جهت تحلیل داده‌ها

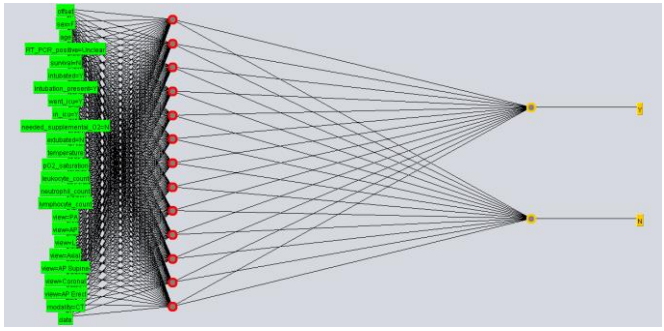
مطالعات نشان می‌دهد شبکه‌های عصبی قادر به تقریب هر تابع عملیاتی می‌باشند. در میان شبکه‌های عصبی مختلف، «شبکه عصبی پرسپترون چندلایه[۲۸]»، با ساختار پس انتشار خطا و «تابع انتقال سیگموئید[۲۹]» در لایه‌های مخفی و تابع انتقال خطی، در لایه خروجی، در ارتباط با کاربردهای رگرسیون، از کارایی مناسبی برخوردار است (خسروانیان و آیت، ۱۳۹۷).

همان‌گونه که قبلا مطرح شد، تعداد ۹۵۰ سطر داده پژوهش را به دو قسمت تقسیم می‌کنیم: بخش داده‌های آزمون (۲۰٪ معادل ۱۹۰ سطر)، و داده‌های آموزش (۸۰٪ معادل ۷۶۰ سطر). که پس از بارگذاری داده‌ها در نرم‌افزار «وکا»، نتیجه اجرای شبکه عصبی مصنوعی با الگوریتم چندلایه پرسپرون، در شکل ۴ نمایش داده شده است:

```
=== Summary ===

Correctly Classified Instances         184               96.8421 %
Incorrectly Classified Instances         6                3.1579 %
Kappa statistic                          0.9299
Mean absolute error                      0.0441
Root mean squared error                  0.1783
Relative absolute error                  9.415  %
Root relative squared error             37.1779 %
Total Number of Instances              190

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.992    0.075    0.961      0.992    0.976      0.931    0.969     0.979     Y
                 0.925    0.008    0.984      0.925    0.954      0.931    0.969     0.963     N
Weighted Avg.    0.968    0.051    0.969      0.968    0.968      0.931    0.969     0.973

=== Confusion Matrix ===

   a   b   <-- classified as
 122   1 |   a = Y
   5  62 |   b = N
```

شکل ۴- اجرای شبکه عصبی مصنوعی با الگوریتم چندلایه پرسپرون، روی داده‌های بالینی و تصاویر پزشکی بیماران مشکوک یا مبتلا به کووید-۱۹

با توجه به شکل ۴ مشاهده می‌کنیم که مقدار TP برابر ۱۲۲ و TN برابر ۶۲ و مقدار FN برابر ۱ و FP برابر ۵ می‌باشد و نهایتا" دقت مدل ۰/۹۶۹ یعنی تقریبا ۹۷٪ است که دقت این مدل نیز بسیار خوب و بهتر از مدل‌های قبل است. تصویر شبکه عصبی مصنوعی حاصل شده توسط نرم‌افزار «وکا»، در شکل ۵ نمایش داده شده است. در این مرحله به استخراج قوانین می‌پردازیم:



شکل ۵- نتیجه اجرای شبکه عصبی مصنوعی با الگوریتم پرسپرون، روی داده‌های بالینی و تصاویر پزشکی بیماران مشکوک یا مبتلا به کووید-۱۹

برخی نتایج حاصل از اجرای شبکه عصبی مصنوعی با الگوریتم چندلایه پرسپرون، بر روی داده‌ها این را بیان می‌کند که از بین ۱۹۰ داده آزمون، با در نظر گرفتن مقدار TP می‌توان دریافت که تعداد ۱۲۲ بیمار که مبتلا به کووید-۱۹ بوده‌اند به درستی تشخیص داده شده است. همچنین با توجه به مقدار TN می‌توان گفت از بین ۱۹۰ داده آزمون، ۶۲ مورد که مبتلا به کووید-۱۹ نبوده‌اند نیز به درستی تشخیص داده شده است. که این موضوع بیانگر دقت مدل است. همینطور مقدار FP نشان می‌دهد که از بین ۱۹۰ داده آزمون، تنها ۵ مورد منفی مبتلا به کووید-۱۹ را به اشتباه مثبت پیش‌بینی کرده است و مقدار FN بیانگر این است که از بین ۱۹۰ داده آزمون، تنها یک نمونه مثبت ابتلا به کووید-۱۹ را به اشتباه منفی تعیین کرده است.

[۲۸] Multi Layer Perceptron

[۲۹] Sigmoid function

۶- یافته‌ها

در این پژوهش از دیتاست جهانی کوهن و همکاران که در سال ۲۰۲۰ میلادی تهیه شده بود، استفاده کردیم. جامعه آماری مورد بررسی در آن مجموعا" شامل ۹۵۰ داده تصویر رادیوگرافی بیماران است که از این تعداد ۳۱۱ سطر یا ۳۲/۷ ٪ مربوط به زنان و ۵۵۹ سطر یا ۵۸/۸ ٪ مربوط به مردان است. با توجه به توابع جانسون، ژنتیک و درخت تصمیم، مشخصه «سن»، «تعداد روزهای تهیه تصویر پس از شروع علائم و یا بستری شدن»، «وضعیت آزمایش RT_PCR» و «نوع تصویر پزشکی»، و بر اساس تابع هولتس بعلاوه مشخصه «درخواست اکسیژن»، اثرگذاری بیشتری در جهت تشخیص و تفکیک بیماران دارند.

از لحاظ محدوده سنی نمونه‌ها، حداقل سن ۱۸ سال و حداکثر ۹۴ سال است که بیشترین نمونه‌ها در سنین ۵۰ و ۷۰ سال هستند. درصدهای فراوانی گروه‌های سنی به ترتیب، گروه سنی ۱۸ تا ۳۰ سال ۱۳/۷ ٪، گروه سنی ۳۱ تا ۴۰ سال ۱۲/۳ ٪، گروه سنی ۴۱ تا ۵۰ سال ۱۹/۲ ٪، گروه سنی ۵۱ تا ۶۰ سال ۱۷/۸ ٪، گروه سنی ۶۱ تا ۷۰ سال ۱۸/۲ ٪، گروه سنی ۷۱ تا ۸۰ سال ۱۵/۴ ٪، و گروه سنی ۸۱ سال به بالا ۳/۲ ٪ می‌باشند. از نظر میزان تهیه تصاویر پزشکی و رادیوگرافی بیماران، به طور متوسط ۹/۰۸ روز پس از شروع علائم و یا بستری شدن از بیماران تصاویر پزشکی تهیه شده است.

با در نظر گرفتن قوانین استخراج شده از توابع جانسون و ژنتیک و هولتس در اجرای تئوری «رافست» و همچنین قوانین مستخرج از درخت تصمیم، با ترکیب این قوانین به قوانین شامل‌تر و معتبرتری دست خواهیم یافت:

قانون۱ (ترکیب ۵ ژنتیک، ۱ هولتس، ۴ جانسون، ۲ درخت تصمیم): در زمانی که شیوع بیشتر است (بعد از سال ۲۰۲۰)، بیماران مرد که سنشان بین ۶۵ تا ۸۹ سال است یا سنشان را مشخص نکرده‌اند (به احتمال بد حالی)، احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایرین جدا شوند.

قانون۲ (ترکیب ۱ ژنتیک، ۱ هولتس، ۱ جانسون، ۱ درخت تصمیم): در زمانی که شیوع بیشتر است (بعد از سال ۲۰۲۰)، بیماران مرد که سنشان بالای ۴۵ سال است و کمتر از ۳۳ روز پس از شروع علائم، از آن‌ها تصویر تهیه شده است، احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایرین جدا شوند.

قانون۳ (ترکیب ۱ هولتس ، ۱ درخت تصمیم): در زمانی که شیوع بیشتر است (بعد از سال ۲۰۲۰)، بیماران مرد که نوع تصویر پزشکی آن‌ها اشعه-ایکس است اما بیش از ۷۴ روز پس از شروع علائم، از آن‌ها تصویر تهیه شده است، احتمال ابتلاء آنان به کووید-۱۹ کم است و نیازی نیست از سایرین جدا شوند.

قانون۴ (ترکیب ۶ ژنتیک، ۳ جانسون): در زمانی که شیوع بیشتر است (بعد از سال ۲۰۲۰)، بیمارانی (صرف نظر از جنسیتشان)، وضعیت آزمایش RT_PCR آن‌ها شفاف نیست یا مثبت است، احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایرین جدا شوند.

قانون۵ (ترکیب ۴ ژنتیک، ۱ هولتس): در زمانی که شیوع بیشتر است (بعد از سال ۲۰۲۰)، بیمارانی که سنشان مشخص نیست و هم چنین ۴ تا ۶۰ روز پس از شروع علائم، از آن‌ها تصویر (AP یا AP Supine) تهیه شده در طی فرآیند درمان درخواست اکسیژن کرده‌اند، احتمال ابتلاء آنان به کووید-۱۹ زیاد است و باید از سایرین جدا شوند.

۷- بحث و نتیجه‌گیری

در مجموع در این پژوهش که با هدف تعیین مشخصه‌های مهم و همچنین قوانین تفکیک بیماران مبتلا به کووید-۱۹ از سایر بیماران انجام شد، روش‌های داده‌کاوی «رافست»، «درخت تصمیم» و «شبکه عصبی» مورد بررسی قرار گرفتند که در روش «رافست»، با توجه به توابع جانسون و ژنتیک، و همچنین در روش «درخت تصمیم»، مشخصه‌های «سن»، «تعداد روزهای تهیه تصویر پس از شروع علائم و یا بستری شدن»، «وضعیت آزمایش RT_PCR» و «نوع تصویر پزشکی»، و بر اساس تابع هولتس بعلاوه مشخصه «درخواست اکسیژن»، اثرگذاری بیشتری در قوانین استخراج شده جهت تشخیص و تفکیک بیماران دارند. با استفاده از روش «رافست» بوسیله توابع جانسون و ژنتیک و هولتس، دقت قوانین مدل هریک ۸۳٪ شد که بوسیله تابع جانسون ۴۶۵ قانون و توسط تابع ژنتیک تعداد ۳۳۱۶ قانون و توسط تابع هولتس تعداد ۶۲ قانون استخراج شد. با استفاده از روش «درخت تصمیم» و الگوریتم J48 آن، دقت مدل برابر۸۲٪ و تعداد قوانین برابر ۹ قانون است. همچنین در روش «شبکه عصبی مصنوعی» با الگوریتم چندلایه پرسپرون، دقت مدل تقریبا" برابر۹۷٪ می‌باشد از بقیه روش‌ها بالاتر است. این نتایج نشان داد که روش و راستای پژوهش با تحقیقات استافورد (۱۹۸۴)، ویجی‌وارانی (۲۰۱۳)، لینگاریج (۲۰۱۵)، افتخاری و عارفیان (۱۳۹۲)، حقیقت و همکاران (۱۳۹۱)، هم راستا بوده و از سوی دیگر در حوزه تشخیص مشخصه‌ها یا تحقیقات بسکابادی و همکاران(۱۳۹۹)، آپادهایا و همکاران (۲۰۲۰)، و کومار داس و همکاران (۲۰۲۱)، به طور مستقیم و همچنین با تحقیقات علیزاده‌فرد و صفاری‌نیا (۱۳۹۸)، به طور غیر مستقیم سازگاری دارد.

۸- محدودیت‌ها

دراین پژوهش با محدودیت استفاده از داده‌های داخل کشور مواجه بودیم. با استفاده از تحلیل داده‌های داخل کشور و با ارائه این خدمات به سامانه‌های درمانی در کشور، می‌توان مدیریت علمی و بهینه هزینه‌ها و همچنین جلوگیری از شیوع عوارض ناشناخته بیماری و تبعات اجتماعی را شاهد بود.

۹- مراجع

[۱] ابریشمی، حمید، *مبانی اقتصاد سنجی*. تهران: نورعلم، ۱۳۸۷.

[۲] اسماعیل پور، منصور و بهلولی، علی و اسلامبولچی، علیرضا، بررسی عوامل استقرار مدیریت دانش در آموزش و پرورش با استفاده از تکنیک های داده کاوی (مورد مطالعه: اداره آموزش و پرورش شهرستان دورود)، *مطالعات دانش شناسی*، سال ششم، شماره ۲۲ ، بهار ۹۹ ، ص ۱ تا ۲۴، ۱۳۹۹.

[۳] اسماعیل پور، منصور و نومی گلزار، الناز، راهکاری جهت تشخیص عیب در گیربکس ماشین آلات با استفاده از شبکه های عصبی مصنوعی، *اولین همایش ملی مهندسی برق، کامپیوتر و فناوری اطلاعات*، همدان، ۱۳۸۶.

[۴] افتخاری، مهدی و عارفیان، فاطمه، روش جدیدK نزدیکترین همسایه فازی و ناهموار برای طبقه‌بندی نیمه نظارتی، *همایش ملی مهندسی کامپیوتر و توسعه پایدار با محوریت شبکه‌های کامپیوتری*، ۲۵-۱۵، ۱۳۹۲.

[۵] بحیرایی، آرش، سهیلی نیا، حسین، فیلی، حمیدرضا، طاهری، سعید، مقایسه الگوریتمهای خوشه‌بندی سلسله‌مراتبی و غیرسلسله‌مراتبی با رویکرد حل مساله، *اولین کنفرانس بین المللی مدیریت، حسابداری و اقتصاد*، شیراز، ۱۳۹۳.

[۶] بسکابادی، مصطفی؛ دوست‌پرست، مهدی، مدل بندی و داده کاوی داده های جهانی بیماران ویروس کووید ۱۹. *مجله طب اورژانس ایران*، دوره۷، شماره۱، مقاله ۴۰.

[۷] خسروانیان آسیه، آیت سعید. یک سیستم هوشمند پزشکیار مبتنی بر شبکه عصبی مصنوعی در تشخیص بیماری دیابت. *مجله دیابت و متابولیسم ایران*. ۱۳۹۷; ۱۸ (۲) :۷۹-۷۱، ۱۳۹۹.

[۸] حقیقت منفرد، جلال؛ علی نژاد، محمود و متقالچی، سارا، مقایسه مدل شبکه‌های عصبی با مدل باکس جنکینز در پیشبینی شاخص کل قیمت سهام بورس اوراق بهادار تهران، *فصلنامه مهندسی مالی و مدیریت اوراق بهادار*، شماره ۱۱، ۱۳۹۱.

[۹] صانعی‌فر، متین، سعیدی، پرویز، مقایسه شبکه های پیچیده بازارهای بورس سهام و متغیرهای اقتصادی در دوران قبل و بعد از شیوع ویروس کرونا (کووید- ۱۹)، *فصلنامه تحقیقات مدلسازی اقتصادی*، شماره ۴۰ ، تابستان۹۹، ص ۱۲۳، ۱۳۹۹.

[۱۰] صنیعی آباده، محمد، محمودی، سینا و طاهر پور، محدثه.. *داده کاوی کاربردی*، جلد ۱. تهران: نیاز دانش، ۱۳۹۴.

[۱۱] طاهری، سارا، مروری بر بیماری کروناویروس (کووید- ۱۹) و آنچه دربارهٔ آن شناخته شده است، *تصویر سلامت*، شماره ۱۱، ص ۸۷-۹۳، ۱۳۹۹.

[۱۲] علیزاده‌فرد، سوسن، صفاری‌نیا، مجید، پیش بینی سلامت روان بر اساس اضطراب و همبستگی اجتماعی ناشی از بیماری کرونا، *پژوهش های روان شناسی اجتماعی*، زمستان ۱۳۹۸، شماره ۳۶، ص ۱۲۹ -۱۴۱، ۱۳۹۸.

[13] Battineni Gopi, Chintalapudi Nalini, Amenta Francesco, "Forecasting of COVID-19 epidemic size in four high hitting nations (USA, Brazil, India and Russia) by Fb-Prophet machine learning model". *Applied Computing and Informatics*, 2020.

[14] Chen Y, Liu Q, Guo D., "Emerging coronaviruses:genome structure, replication, and pathogenesis". *Medical Virology*, 2020.

[15] Cohen Joseph Paul, Morrison Paul, Dao Lan, Roth Karsten, Duong Tim Q., "COVID-19 Image Data Collection: Prospective Predictions are the Future", *Journal of Machine Learning for Biomedical Imaging (MELBA)*, 2020.

[16] García, E., Romero, C., Ventura, S., & Calders, T.,"Drawbacks and solutions of applying association rule mining in learning management systems". *Paper presented at the Proceedings of the International Workshop on Applying Data Mining in e-Learning (ADML 2007)*, Crete, Greece, 2007.

[17] Han J, Kamber M, Pei J., "Data Mining: Concepts and Techniques. 3th ed.", *San Francisco: Morgan Kaufmann*, 2011.

[18] Kang Jianhong, Chen Ting, Luo Honghe, Li Lijian, Jiming Yang Mia, "Machine learning predictive model for severe COVID-19*", Infection, Genetics and Evolution*, 104737, 2021.

[19] KumarDas Ayan, Kalam Sidra, Kumar Chiranjeev, Sinha Ditipriya, "TLCoV- An automated Covid-19 screening model using Transfer Learning from chest X-ray images", *Chaos, Solitons & Fractals*, Volume 144, 2021.

[20] Lakshmi KV, Padmavathamma M., "Modeling an Expert System for Diagnosis of Gestational Diabetes Mellitus Based On Risk Factors", *Journal of Computer Engineering*, 8(3):29-32, 2013.

[21] Lingaraj H, Devadass R, Gopi V, Palanisamy K., "Prediction of diabetes mellitus using data mining techniques: a review", *Journal of Bioinformatics & Cheminformatics*,1(1):1-3, 2015.

[22] Olanow CW, Watts RL, Koller WC., "An algorithm (decision tree) for the management of Parkinson's disease" (2001): treatment

[23] Puno George R., Puno Rena Christina C., Maghuyop Ida V., "COVID19-case fatality rates across Southeast Asian countries (SEA): a preliminary estimate using a simple linear regression model", *Journal of Health Research*, 2021.

[24] Romero, C., Ventura, S., & García, E., "Data mining in course management systems: Moodle case study and tutorial", *Computers & Education*, 51(1), 368-384, 2008.

[25] Stafford GC, Kelley PE, Syka JEP, Reynolds WE, Todd JF., "Recent improvements in and analytical applications of advanced ion trap technology", *International Journal of Mass Spectrometry and Ion Processes*, 60(1):85-98.9, 1984.

[26] Upadhyaya Ashish, Koirala Sushant, Ressler Rand, Upadhyaya Kamal, "Factors affecting COVID-19 mortality: an exploratory study", *Journal of Health Research*, 2020.

[27] Vijiyarani S, Sudha S., "Disease prediction in data mining technique– a survey", *International Journal of Computer Applications & Information Technology*, 2(1):17-21, 2013.

[28] Witten IH, Frank E, Hall MA, Pal CJ., "Data Mining: Practical Machine L
e
a
r
n
i
n
g

T
o
o
l
s

a
n
d

T
e

[29] Ziarko, W., "The discovery, analysis, and representation of data dependencies in databases", Knowledge discovery in database, *The MIT Press*, 1991.