دسترسی در سایتِ http://jnrm.srbiau.ac.ir
سال هشتم، شماره چهلم، بهمن و اسفند ۱۴۰۱
شماره شاپا: ۲۵۸۸-۵۸۸X

دانشگاه آزاد اسلامی، واحد علوم و پژوهش‌های نوین در ریاضی

ارائه مدل پیش‌پردازش داده‌های سازمانی جهت پیش‌بینی فرآیندهای کسب و کار بیمه‌های اجتماعی

مهرداد فدائی پله‌شاهی[۱]، سهراب کردرستمی[۲*]، امیرحسین رفاهی شیخانی[۳]، مرضیه فریدی ماسوله[۴]، سهیل شکری[۵]

[۱و۲و۳و۵] گروه ریاضی و علوم کامپیوتر، دانشکده علوم پایه، واحد لاهیجان، دانشگاه آزاد اسلامی، لاهیجان، ایران.
[۴] دانشکده کامپیوتر و فناوری اطلاعات، دانشگاه احرار، رشت، ایران

تاریخ ارسال مقاله: ۱۴۰۰/۰۶/۰۷ تاریخ پذیرش مقاله: ۱۴۰۱/۰۳/۱۱

چکیده

در این مقاله، یک روش جدید پیش‌پردازش داده‌ها در زمینه پیش‌بینی فرآیندهای کسب و کار با استفاده از شبکه عصبی بازگشتی، زنجیره مارکوف و یادگیری عمیق بازگشتی ارائه شده است. هدف، بدست آوردن داده‌های با کیفیت و استخراج اطلاعات مهمترین متغیرهای دخیل در فرآیند از کارافتادگی از سازمان تامین اجتماعی است. برای دستیابی به این هدف، روش، شامل کاهش تعداد ویژگی و نرمال‌سازی داده‌ها در مقایسه با ویژگی‌های اولیه می‌باشد. روش پیشنهادی با داده‌های حقیقی سازمان تامین اجتماعی؛ پیاده‌سازی شده و در مدل، پیش‌بینی اعمال گردیده و نتایج نشان می‌دهد که استفاده از روش ارائه شده، میزان استفاده از منابع حافظه‌ای را به میزان بسیار جزئی افزایش داده ولی میزان زمان استفاده از CPU، نسبت به روش‌های مورد مقایسه کاهش چشمگیری داشته و مضاعف بر آن، میزان دقت و کارایی را نیز به میزان قابل توجه‌ای افزایش می‌دهد.

واژه‌های کلیدی: پیش‌پردازش، پیش‌بینی، کسب و کار بیمه‌های اجتماعی، شبکه عصبی بازگشتی، داده کاوی.

* عهده‌دار مکاتبات: Email: sohrabkordrostami@gmail.com

۱- مقدمه

پیش‌بینی نتایج در رویدادهای یک فرآیند کسب و کار، کاربردهای جالب زیادی دارد. برای مثال زمانی که افراد به مرور کردن صفحات مختلف وب سایت مربوط به زمینه‌های مرتبط فرآیند دلخواه، تمایل پیدا می‌کنند، پیش‌بینی موفقیت‌آمیز مرحله بعدی باعث می‌شود موارد مرتبط را قبل از انتخاب، آماده نمایند و جذابیت بیشتری را برای مشتریان ایجاد کنند.

وب، مجموعه‌ای از اسناد متنی نوشته‌شده در HTML به نام صفحات وب است. این صفحات وب از طریق URLها به یکدیگر متصل می‌شوند. وب سه جنبه دارد: محتوای وب، ساختار وب و کاربرد وب. بنابراین، کشف دانش درباره وب عمدتا به سه دسته تقسیم می‌شود. اول محتوا کاوی وب که هدف آن یافتن اطلاعات مفید در محتوای صفحات وب مانند متن، تصاویر و تگ‌های HTML است. دوم ساختار کاوی وب است و هدف آن کشف اطلاعات از ساختار هایپرلینک وب بوده و در نهایت سومین جنبه، کاربرد کاوی وب است و هدف آن کشف دانش در الگوهای استفاده کاربران وب است.

وظایف چندگانه پیش گویانه در مدل سازی فرآیند کسب و کار وجود دارد. فرض می‌کنیم که در نقطه پیش‌بینی، یک نمونه فرآیند جزئی اجرا می‌شود و می‌خواهیم ویژگی‌های آینده، به عنوان مثال، ادامه یک نمونه فرآیند یا زمان باقیمانده یا مجموعه‌ای از منابع مورد نیاز برای تکمیل نمونه، را پیش بینی کنیم. تحت شبکه‌های عصبی با حافظه تقویت شده[۱]، بسیاری از این وظایف پیش‌بینی را می‌توان به پیش‌بینی دنباله تبدیل کرد، یعنی دنباله‌ای از نمادهای گسسته تولید می‌کنیم. به عنوان مثال، فرآیند ادامه، یک مورد طبیعی است، که در آن هر نماد یک رویداد است [۱].

پیش‌بینی و برآورد در کسب و کار هر سازمان، دارای اهمیت بسیاری بوده و نتایج پیش‌بینی‌های دقیق، شامل بهره‌وری بالاتر، صرفه‌جویی بیشتر در هزینه‌ها، افزایش کمیت و کیفیت سود و همچنین ارائه خدمات بهتر به ذینفعان خواهد بود.

تکنیک‌های فرآیندکاوی، استخراج اطلاعات مفید از ثبت رویدادها و اطلاعات بر اساس تاریخ فرآیندهای کسب و کار را ممکن می‌سازد. این اطلاعات به بهبود فرآیندها کمک می‌کنند و معمولا پس از اتمام فرآیند استخراج می‌شوند. با این حال، علاقه به استفاده از فرآیندکاوی برای اجرای فرآیندهای در حال اجرا رو به افزایش است[۲].

پیش‌بینی دقیق دنباله‌ی رویدادها به ارائه خدمات بهتر به مشتریان و جلوگیری از خطرات کمک می‌کند. در تحقیقات اخیر تلاش شده تا به مسئله یادگیری تاثیر رویدادهای گذشته بر رویدادهای آینده با استفاده از روش های یادگیری عمیق پرداخته شود. چنین کارهایی اغلب دنباله رویدادهای گذشته را به عنوان ورودی و تغییرات رویدادها را مدل می‌کنند و تعداد کمی از آنها به تاثیر ویژگی‌های رویداد، مربوط می‌شود [۳-۸].

الگوریتم‌های فرآیندکاوی متداول [۳،۹-۱۱] قادر به استخراج دانش از ثبت رویدادها هستند که به ایجاد یک مدل فرآیند و ایجاد رویداد بعد از آن کمک می‌کند. اگر چه روش های استخراج فرآیند در تحلیل اطلاعات ورودی به سیستم‌های اطلاعاتی موثر بوده‌اند اما در پیش بینی‌ها محدودیت‌هایی وجود دارد:

۱- تعداد زیادی از رویدادها و یا روابط پیچیده بین رویدادها قابل کنترل نیستند.
۲- دقت نسبی به میزان قابل توجهی، بسته به نوع داده ورودی متفاوت است.

تحولات اخیر در هوش مصنوعی، فرصت‌های جدیدی را در صنایع بیمه به منظور ایجاد راه حل‌های مناسب و خدمات مبتنی بر دانش جدید

[۱] Memory-Augmented Neural Networks

مشتریان و اجرای عملیات‌های پیشرفته و عملکردهای تجاری، ارائه کرده است. با این حال، داده‌های مربوط به بیمه، ناهمگن بوده و توزیع کلاس نامتعادل با فراوانی پایین و ابعاد بالا، چهار چالش عمده را در زمینه یادگیری در کسب و کار دنیای واقعی، ایجاد کرده است [۱۲]. در [۱۲]، بر روی چارچوب یادگیری موازی مقرون به صرفه جهت افزایش عملیات‌های بیمه با رویکرد یادگیری عمیق که نیازمند پیش‌پردازش نیست، تمرکز شده است که شامل یک شبکه عصبی موازی منسجم و جدید بوده و داده‌های همگن واقعی را ارائه می‌کند.

آماده‌سازی داده‌ها، مهم‌ترین و زمان‌برترین بخش در پروژه‌های مدیریت فرآیندها می باشد. در پیش‌بینی فرآیندهای کسب و کار بیمه‌های اجتماعی با چالش تعداد زیاد ویژگی، توزیع نامتوازن، نرخ پایین تخصیص داده‌ها، مواجه بوده لذا یک روش پیش‌پردازش مطلوب برای انتخاب ویژگی‌های بهینه که منجر به پیش‌بینی با دقت بالاتر گردند، ارزشمند خواهد بود.

در تحقیقات گذشته، کاهش تعداد ویژگی و نرمال‌سازی داده‌ها قابل مشاهده بوده و لیکن، چالش ویژگی‌های فراوان، توزیع نامتوازن، نرخ پایین تخصیص در داده‌های ناهمگن از فرآیندهای نیمه‌ساخت‌یافته از سازمان تامین اجتماعی وجود داشته که در مباحث پیش‌بینی فرآیندهای کسب و کار تامین اجتماعی از اهمیت بالایی برخوردار بوده و روش ارائه شده در این مقاله در جهت رفع این چالش و امکان پیش‌بینی با دقت بالاتر ارائه می‌گردد.

قلمرو موضوعی این پژوهش، عملیات پیش‌پردازش بر روی داده‌های بیمه شدگان با درخواست ازکارافتادگی در سازمان تامین اجتماعی می‌باشد و قلمرو مکانی سازمان تامین اجتماعی می‌باشد.

مراحل این روش پیشنهادی از طریق مدل پیش‌بینی با ترکیب شبکه عصبی بازگشتی و زنجیره مارکوف و یادگیری عمیق بازگشتی تایید می‌گردد و علاوه بر این به ما امکان می‌دهد اهمیت ویژگی‌هایی را که برای انتخاب مجموعه‌ای از مناسب‌ترین ویژگی‌ها استفاده می‌شود، به دست آوریم. در واقع با توجه به کاهش چشمگیر تعداد ویژگی و افزایش دقت در پیش بینی مدل، به این ارزیابی دست خواهیم یافت. ادامه مقاله به شرح زیر سازماندهی می‌شود:

بخش ۲ به پیشینه پژوهش پرداخته، در بخش ۳ تشریح مساله عنوان گردیده و در بخش ۴ روش کامل پیشنهادی برای پیش‌پردازش داده‌ها را ارائه می‌دهد. در بخش ۵ یافته‌های پژوهش و بخش ۶ به بحث و نتیجه گیری در مورد یافته‌ها پرداخته و نتایج با روش‌های معتبر مقایسه گردیده و چالش‌های باز در این زمینه را شناسایی می‌کند.

۲- پیشینه پژوهش

در سنوات گذشته و طی دهه‌های اخیر، محققان با روش‌های مختلف از جمله سری‌های زمانی، شبکه‌های عصبی بازگشتی و حافظه کوچک ماندگار، مسائل پیش‌بینی فعالیت بعدی، زمان فعالیت بعدی و ادامه فرآیند در پردازش داده‌های متنی و داده‌های پویا را تحلیل نموده‌اند که به بررسی موردی از عناوین ذکر شده می‌پردازیم.

در دهه ۱۹۶۰ میلادی و ماقبل آن به ایجاد پایگاه‌های داده و جمع‌آوری داده‌ها پرداخته‌شده و از دهه ۱۹۷۰ تا اواسط ۱۹۸۰، سیستم‌های مدیریت پایگاه داده شکل گرفته است.

از اواسط ۱۹۸۰ تا به حال، سیستم‌های پایگاه داده پیشرفته در حال بکارگیری بوده و از اواخر۱۹۸۰ میلادی تاکنون به صورت همزمان بر روی این پایگاه داده‌ها، تحلیل پیشرفته داده که در واقع شامل داده کاوی می‌باشد، در حال انجام است.

تکنیک‌های مختلف داده کاوی برای استخراج الگوها و در نتیجه، دانش از انواع مختلف پایگاه داده‌ها و

جمع‌آوری داده‌ها، یک فعالیت ضروری و با اهمیت در فرآیندکاوی می‌باشد [۱۳].

پاک‌سازی داده‌ها، تجمیع داده‌ها، کاهش داده‌ها و تبدیل داده‌ها از مراحل پیش‌پردازش داده‌ها در بحث داده کاوی بوده که با انواع داده‌های ذیل قابل پیاده‌سازی می‌باشد:

- انواع داده ساده: داده‌های موجود در مخزن داده‌ها و داده‌های معاملاتی و مبادلاتی
- انواع داده‌های پیچیده: صدا، تصویر، داده‌های جریانی و داده‌های درختی

در مقاله [۱۴] بهینه‌سازی کلونی موچه‌ها بررسی‌شده و چارچوب مبتنی بر بهینه‌سازی ازدحام ذرات برای طبقه‌بندی داده‌ها و پیش‌پردازش در محیط کلان داده که از وزن‌ها بر اساس اندازه، محتوا و کلمات کلیدی اختصاص داده‌شده، پیشنهاد گردیده است [۱۴].

در مقاله [۱۵]، مروری بر افزایش پیش‌پردازش داده‌ها، یک دسته‌بندی به روزرسانی شده در رایانش ابری تحت چارچوب کلان داده‌ها ارائه‌شده و با تکنیک‌های پیش‌پردازش مانند انتخاب ویژگی، داده‌های ناقص، یادگیری نامتعادل، کاهش نمونه و همچنین حداکثر اندازه پشتیبانی‌شده، بررسی خانواده‌های مختلف داده‌ها پوشش داده‌شده و مسائل کلیدی در پیش‌پردازش داده‌های بزرگ برجسته شده‌اند[۱۵].

جان آنتونیوکورتس و همکاران در سال ۲۰۲۰ یک مدل پیش‌پردازش در زمینه پیش‌بینی سری زمانی صنعتی ارائه داده‌اند که از میانگین مرکزی جهت رفع سرو صدا و هموارسازی سری‌های زمانی استفاده‌شده و انتخاب مهمترین ویژگی‌ها براساس اهمیت استفاده الگوریتم‌های رگرسیون شکل گرفته و هریک از مراحل روش از طریق آزمایش با سه رگرسیون پیشرفته تایید شده است[۱۶].

اگرچه در مطالعات مختلف، پیش‌پردازش داده‌ها را بخشی از فرآیند اکتشاف داده‌ها می‌دانند اما این مطالعات، فاقد تحلیل مناسب در خصوص نیاز به فن آوری‌ها و راه حل‌های جداگانه برای پیش‌پردازش می‌باشند [۱۷].

در [۱۷] به بررسی اثرات پیش‌پردازش از طریق پردازش داده‌های متنی مدل‌های جمله می‌پردازد. برای این منظور، آزمایش برای ارزیابی ترکیبات انواع پیش‌پردازش داده‌های معمولی انجام‌شده است. بعلاوه، آثار دو تکنیک جدید در مورد دقت مدل تجزیه و تحلیل پیش‌پردازش اصطلاحات فنی متشکل از کلمات مرکب و تعیین ترتیب یادگیری بر اساس پیچیدگی داده‌ها وجود دارد.

در زمانی که چالش‌ها یا مشکلات به درستی شناسایی و به صورت مناسب مرتب شوند، داده کاوی بسیار موفقیت‌آمیز خواهد بود [۱۳].

سیستم‌های مدیریت روند کسب و کار سنتی بر جریان کارهای اجرایی کامل ساختارمند به طور جامع متمرکز هستند. اخیرا این تمرکز به سمت رسیدگی به فرآیندهای پوشش وسیع‌تری از رفتار، حرکت کرده است. در حقیقت، فرآیندهای کسب و کار حاوی طیفی از فرآیندهای سنتی ساختارمند هستند به همراه جریان‌های غیرساختارمند تک منظوره که توسط انسان هدایت می‌شوند. سیستم‌های فرآیند کسب و کار سنتی، در یک طرف این طیف، به یک مدل فرآیند نیاز دارند که بتواند به طور کامل از پیش تعریف شود و معمولا شامل محدودیت‌هایی مانند جریان کنترل شدید است. به عنوان یک حرکت به سوی فرآیندهای کمتر شدت‌یافته و سخت در این طیف، رجوع به مراحل به عنوان نیمه‌ساختارمند متداول است. فرآیندهای نیمه ساختارمند به طور گسترده در صنایعی از قبیل دولتی، بیمه، بانکداری و مراقبت‌های بهداشتی بوجود می‌آیند [۱۸]. فرآیند درخواست خسارت از بیمه خودرو، مدیریت نسخه‌های دارو و ارائه خدمات مناسب به بیمار در بیمارستان، چند نمونه از چنین فرآیندهایی هستند. این فرآیندها از نوع سنتی

فرایندهای ساختارمند و متوالی از پیش تعیین‌شده خارج می‌شوند، زیرا چرخه عمر آنها به طور کامل توسط یک مدل فرآیند به طور کامل هدایت نمی‌شود. به طور معمول مقدار زیادی از اطلاعات مربوط به فرآیندهای نیمه ساختارمند موردگرا وجود دارد و از منابع داده متفرقه می‌آید. کسانی که روی یک مورد کار می‌کنند، به عنوان کارشناسان پرونده نیز شناخته می‌شوند، که مستقل قضاوت می‌کنند و در عین حال در انتخاب مجموعه و دنباله‌ای از مراحل فرآیند برای رسیدگی به یک نمونه پرونده در محتویات سند موجود و اطلاعات مربوط به آن پرونده تابع دستورالعمل‌های شرکت هستند. کارشناسان پرونده می‌توانند به موازات، چندین وظیفه را انجام دهند و یک یا چند وظیفه را تکرار کنند. حجم داده‌های موجود ممکن است در هر مرحله در طی بررسی یک پرونده تغییر کند و مقادیر داده‌های جدید می‌تواند بر تصمیم‌گیری کارشناسان پرونده مربوطه در نحوه روند کار تاثیر بگذارد. بنابراین، مجموعه‌ای از وظایف و دستورات اجرایی آن‌ها در یک نمونه پروسه کسب و کار نیمه ساختارمند، یک دلیل پیش فرض نیست [۱۹].

در محیطی که روند پرونده‌ها به تجزیه و تحلیل حجم زیاد داده‌های پویا و مهلت اجرا به تصمیم‌گیری سریع نیاز دارد، اداره کردن پرونده‌ها چالش برانگیز و مستعد خطا است حتی برای کارشناسان پرونده که تخصص بالایی دارند. حجم زیادی از داده‌های مربوط به یک نمونه پرونده ممکن است حتی برای یک کارشناس پرونده برای ایجاد یک تصویر منسجم از آن پرونده مشکل باشد. همه این عوامل باعث می‌شود مدیران پرونده‌ها در شناسایی موقعیت‌های بحرانی که در آن مداخله مستقیم، لازم است و اجرای سیاست‌ها در طول زمان اجرا شود، با سختی مواجه شوند. با این وجود نتایج تصمیمات گذشته می‌تواند برای تصمیم‌گیری بهتر در آینده استفاده شود [۱۹].

تجزیه و تحلیل فرآیند، شامل یک لایه پیچیده تجزیه و تحلیل داده‌ها بر اساس مفهوم سنتی فرآیند کاوی است [۱۹]. در مقایسه با فرآیند کاوی، تجزیه و تحلیل فرآیند، مسئله اصلی‌تری را برای استفاده از داده‌های تولیدشده یا مرتبط با اجرای فرآیند برای بدست آوردن دیدگاه‌های عملی در مورد فرآیندهای کسب و کار نشان می‌دهد. تجزیه و تحلیل فرآیند، طیف وسیعی از داده‌ها را استفاده می‌کند و اگر چه تنها به پردازش ثبت و درج‌ها محدود نمی‌شود و اطلاعات مربوط به رویداد، ثبت و درج‌های تامین، اطلاعات مربوط به تصمیم گیری و محتوای فرآیند و سوالات پرس و جوها را نیز پوشش می‌دهد بلکه تعدادی از کاربردهای دنیای واقعی و خاص مربوط به تجزیه و تحلیل، مانند بهینه‌سازی منابع و اولویت‌بندی نمونه را هم به نتیجه می‌رساند. سایر عملیات تجزیه و تحلیل فرآیند عبارتند از پیش‌بینی فعالیت بعدی، زمان فعالیت بعدی، ادامه فرآیند و مدت زمانی که تا تکمیل یا لغو فرآیند طول می‌کشد [۱].

پیشرفت‌های اخیر در معماری شبکه های عصبی و الگوریتم‌های یادگیری باعث محبوب‌شدن روش‌های یادگیری عمیق شده است. روش‌های یادگیری عمیق به ویژه در کشف ساختار پیچیده و نمایش‌های قوی و مقادیر زیاد داده‌های خام بسیار مفید است و بنابراین به طور قابل توجهی نیاز به ویژگی‌های ساخت با کمک دست که معمولا در هنگام استفاده از تکنیک‌های سنتی یادگیری ماشین مورد نیاز است را کاهش می‌دهد. شبکه‌های عصبی بازگشتی، به ویژه حافظه کوچک ماندگار، سبب وقوع پیشرفت غیرمنتظره در حل توالی پیچیده وظایف مدل‌سازی در حوزه‌های مختلف مانند درک تصویر، شناخت گفتار و پردازش زبان طبیعی شده است [۲۱,۶].

به طور مشابه، حافظه کوچک ماندگار می‌تواند به طور مداوم تکنیک‌های کلاسیک را برای تعدادی از

وظایف تجزیه و تحلیل فرآیند مانند پیش‌بینی فعالیت بعدی، زمان برای فعالیت بعدی و غیره بهتر اجرا کند [۲۳,۲۲]. بیشتر تکنیک‌های کلاسیک برای تکنیک‌های پیش بینی در آثار تحقیقی تحلیل پیشگویانه فرآیند با هدف حل مسائل مربوط به وظیفه هدف پیشنهاد شده است. در اینجا ما تعدادی خروجی خواهیم داشت که این خروجی‌ها در مجموع Target یا هدف ما را تشکیل می‌دهند. یادگیری عمیق در مسائل خارج از چهارچوب‌های خاص نیز کاربرد دارد. تکنیک‌های یادگیری عمیق نسبت به تکنیک‌های کلاسیک مزیت بیشتری دارند، زیرا بدون نیاز به مهندسی صحیح ویژگی یا تنظیم پیکربندی به وظایف مختلف، قابل تعمیم‌اند. علاوه بر این، روش‌های مذکور، تقویت برای نویز را نشان می‌دهند و مقیاس عملکرد را به عنوان ورودی داده‌های بزرگتر بیان می‌کنند [۲۴]. شبیه‌سازی این بررسی با تعداد محدودی داده در یک بازه زمانی کوتاه انجام گرفته، حال اینکه با داده‌های بسیار زیاد در طول یک سال و یا حتی در بازه‌های زمانی بیشتر نیز قابل بررسی می‌باشد.

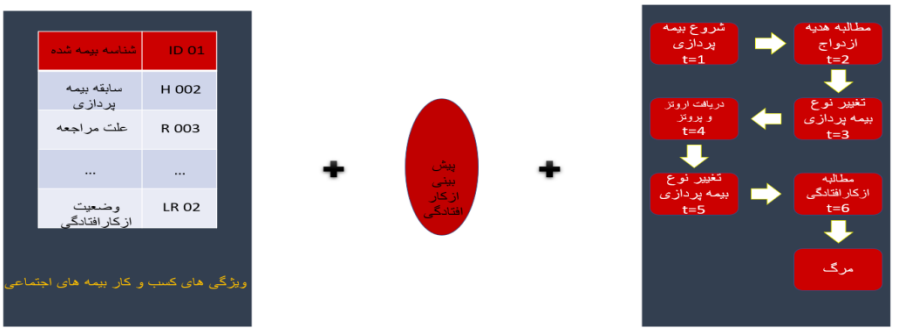
فرآیند کاوی شامل بهبود فرآیندها و مدل‌های فرآیند نیز می‌شود. یکی از فرآیندهای بهینه‌سازی می‌تواند کاهش هزینه‌ها باشد. با این حال بهینه‌سازی یک اصطلاح بسیار گسترده است. فرآیند را می‌توان به لحاظ هزینه‌ها، زمان، منابع مورد استفاده و سایر عوامل بهینه کرد. با تجزیه و تحلیل فرآیند، نقص در سیستم را می‌توان یافت و بهبود دارد. یک مثال از چنین پیشرفتی، استخدام افراد بیشتری است تا فعالیت خاصی، سریعتر انجام شود. مدل‌های فرآیندها نیز می‌توانند با تغییر یک مدل فرآیند برای توصیف واقعی‌تر، فرآیند واقعی را بهبود دهند [۲۵].

مقاله [۲۶] رویکرد جدید برای پیش‌بینی سیستم‌های بیمه‌ای ارائه می‌دهد. با استفاده از ترکیب روش یادگیری عمیق، به طور خاص شبکه عصبی بازگشتی و زنجیره مارکوف برای مسئله پیش‌بینی نتایج در یک فرآیند بیمه‌ای بکار گرفته شده است. همچنین روش پیشنهادی با داده‌های حقیقی سازمان تامین اجتماعی شبیه‌سازی شده است. استفاده از روش ارائه‌شده، میزان استفاده از منابع حافظه‌ای را نسبت به روش مارکوف تا حدی افزایش داده؛ ولی میزان زمان استفاده از CPU نسبت به دو روش مارکوف و شبکه عصبی بازگشتی، کاهش چشمگیری داشته و مضاعف بر آن، میزان دقت و کارایی نیز در روش ارائه‌شده ارتقاء داده شده است.

از دستاوردهای مدل‌های پیش‌پردازش بیان‌شده، می‌توان کاهش تعداد ویژگی و نرمال‌سازی داده‌ها در شرایط مدل‌های ارائه‌شده (تعداد پایین ویژگی و توزیع متوازن) را عنوان کرد ولیکن با توجه به ویژگی‌های فراوان، توزیع نامتوازن و نرخ پایین تخصیص در داده‌های ناهمگن سازمان تامین اجتماعی موجب بروز افزایش خطای محاسباتی می‌گردد که مدل پیش پردازش ارائه‌شده در این مقاله، در جهت رفع چالش‌های مذکور و اخذ نتایج همراه با بهبود دقت و کارایی هم زمان با کنترل محدودیت‌های بیان‌شده در زمان کاهش ویژگی‌ها خواهد بود.

۳- تشریح مساله

در این بخش، تشریح هدف پژوهش و تجزیه و تحلیل داده‌های سازمان تامین اجتماعی برای مدل پیش بینی فرآیندهای کسب و کاری ارائه شده است. داده‌های بیمه های اجتماعی می‌تواند به دو جزء تقسیم شود که در شکل ۱ نشان داده شده است.



شکل (۱): مثالی از مجموعه داده‌های بیمه‌های اجتماعی

(کد ID شناسه بیمه شده، H مقدار سابقه بیمه پردازی به روز، R علت مراجعه و LR وضعیت ازکارافتادگی می‌باشد.)

ویژگی‌های کسب و کار نیمه ساخت یافته در طول زمان ثابت بوده و خدمات مربوطه بابت درخواست‌هایی که در طول زمان چرخه عمر بیمه شده ثبت می‌شود و در صورت احراز شرایط، اعطا می‌گردد.

در این مقاله به بررسی این چالش، انتخاب ویژگی‌های موثر و مناسب در پیش‌بینی مربوطه می‌پردازیم که یک روش پیش‌پردازش برای به دست آوردن داده‌های با کیفیت بالا را ارائه می‌دهد. این روش شامل موارد زیر است:

- کاهش تعداد ویژگی‌های داده
- نرمال‌سازی داده‌ها

بخش ۳-۱ منبع داده، نحوه بدست آوردن و تقسیم بندی داده‌ها را توضیح می‌دهد. در بخش ۳-۲ یک تجزیه و تحلیل داده با محوریت داده‌های با کیفیت پایین، مقادیر گمشده و نویز ارائه می‌شود.

۳-۱. توضیحات داده‌های سازمان تامین اجتماعی

با توجه به گستردگی داده‌ها و جهت دریافت پاسخ پیش بینی به صورت بهینه، تعداد تصادفی ۸۴۶ داده از بیمه‌شدگان سازمان تامین اجتماعی که در سال ۱۳۹۸ هجری شمسی درخواست ازکارافتادگی داده‌اند و جهت تشخیص به کمیسیون پزشکی مربوطه معرفی شده‌اند از معاونت درمان سازمان مربوطه اخذ گردیده است. این داده‌ها که نمونه‌ای از آن در شکل ۲ نشان داده شده، شامل داده‌های رشته‌ای و عددی به شرح ذیل می‌باشد:

- داده‌های رشته‌ای: وضعیت ازکارافتادگی، علت مراجعه، جنسیت، نوع بیمه و نتایج کمیسیون‌های قبلی
- داده‌های عددی: ردیف، شناسه بیمه شده، کد جنسیت، سن، سابقه بیمه پردازی به روز، تاریخ شروع طول درمان و تاریخ اتمام طول درمان

	نتایج کمیسیونهای قبلی	تا تاریخ	از تاریخ	سابقه بیمه پردازی به روز	نوع بیمه	سن	جنسیت	کد جنسیت	وضعیت از کار افتادگی	علت مراجعه
2	طول درمان			4834	اجباري كارگران ساختماني	51	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
3	از کار افتادگي کلي نمي باشد			4834	اجباري كارگران ساختماني	51	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
4	از کار افتادگي کلي نمي باشد	31/03/1398	12/02/1398	6207	اجباري	51	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
5	از کار افتادگي کلي نمي باشد	31/03/1398	12/12/1397	6207	اجباري	51	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
6	از کار افتادگي کلي نمي باشد	31/04/1398	01/04/1398	6207	اجباري	51	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
7	نقص عضو مقطوع	30/09/1385	27/09/1385	6756	همكار سازماني	48	آقا		01 از کار افتاده کلي نمي باشد حادثه ناشي از کار	
8	از کار افتاده جزئي مي باشد	30/09/1385	27/09/1385	6756	همكار سازماني	48	آقا		01 از کار افتاده کلي نمي باشد حادثه ناشي از کار	
9	از کار افتادگي کلي نمي باشد			5571	اجباري	42	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
10	از کار افتادگي کلي نمي باشد			4057	راننده‌گان درون شهري	42	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
11	از کار افتادگي کلي نمي باشد			4710	اختياري	60	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
12	از کار افتادگي کلي نمي باشد			4710	اختياري	60	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
13	رأي بدوي مورد تأييد مي باشد			4710	اختياري	60	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
14	از کار افتادگي کلي نمي باشد			4710	اختياري	60	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
15	از کار افتادگي کلي نمي باشد			2617	حرف و مشاغل آزاد	50	آقا		01 از کار افتاده کلي نمي باشد بيماري و حادثه ناشي از کار	
16	از کار افتاده جزئي مي باشد			2474	اجباري	38	آقا		01 از کار افتاده کلي نمي باشد حادثه ناشي از کار	
17	از کار افتاده جزئي مي باشد			2474	اجباري	38	آقا		01 از کار افتاده کلي نمي باشد حادثه ناشي از کار	
18	از کار افتاده جزئي مي باشد			2474	اجباري	38	آقا		01 از کار افتاده کلي نمي باشد حادثه ناشي از کار	
19	از کار افتاده جزئي مي باشد	12/01/1398	13/10/1397	5000	اجباري	57	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
20	از کار افتادگي کلي نمي باشد			4000	اجباري	56	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
21	طول درمان			4000	اجباري	56	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
22	طول درمان			4000	اجباري	56	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
23	از کار افتادگي کلي نمي باشد			8406	اجباري	44	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
24	از کار افتادگي کلي نمي باشد			8406	اجباري	44	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
25	از کار افتادگي کلي نمي باشد				اجباري	54	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
26				2473	اجباري	41	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
27	از کار افتادگي کلي نمي باشد			9326	حرف و مشاغل آزاد	59	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
28				8518	راننده‌گان حمل بار بين شهري	52	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
29	از کار افتادگي کلي نمي باشد			5475	راننده‌گان حمل مسافر بين شهري	65	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	
30	از کار افتادگي کلي نمي باشد			5475	راننده‌گان حمل مسافر بين شهري	65	آقا		01 از کار افتاده کلي نمي باشد بيماري غير ناشي از کار	

شکل (۲): نمونه‌ای از مجموعه داده‌های خام

۳-۲. تجزیه و تحلیل داده‌ها

داده‌های موجود در مجموعه داده پژوهش شامل اطلاعات عمومی افراد بیمه‌شده شامل شناسه، جنسیت، کد جنسیت و سن می‌باشد و اطلاعات مربوط به سوابق بیمه‌شده، مشخصات بیماری و دیگر داده‌های تخصصی وابسته به سازمان تامین اجتماعی بوده و نهایتا یک ستون نتیجه کمیسیون پزشکی و تعیین تکلیف ازکارافتادگی متقاضیان که صرفا به صورت صفر یا یک بوده و هدف نهایی پیش بینی نتایج می‌باشد.

نکته قابل ملاحظه در این بخش، انتخاب حداقل ویژگی‌های مهم و اثرگذار از ویژگی‌های متعدد در داده‌های نیمه‌ساخت‌یافته سازمان تامین اجتماعی می‌باشد که هم زمان با کنترل هزینه‌های محاسباتی، امکان پیش بینی با دقت بالاتر را نیز فراهم نماید.

۴- روش پیشنهادی

با توجه به اینکه داده نامناسب، یکی از چالش‌های پیش رو در پیش‌بینی فرآیندهاست لذا در این بخش، روش پیشنهادی پیش‌پردازش داده‌ها جهت مناسب سازی داده‌ها برای دستیابی به داده با کیفیت مناسب، خواهیم پرداخت. برخی از مهم‌ترین مواردی که طی فرآیند پیش‌پردازش داده‌ها باید به آن‌ها پرداخته شود، به شرح ذیل می‌باشد:

- داده‌های ناموجود (Missing Data)
- داده‌های پرت (Outliers)
- نرمال‌سازی داده‌ها (Normalization)

در روش پیشنهادی، ۲ مرحله را پیش رو خواهیم داشت:

- کاهش تعداد ویژگی‌های داده
- نرمال‌سازی داده‌ها

۴-۱. کاهش تعداد ویژگی‌های داده

کاهش بعد عمودی در این روش دارای دو قسمت اساسی است که از هم مستقل بوده و همزمان در نرم افزار weka قابل پیاده‌سازی می‌باشند:

- روش جستجو[۱]: روش جستجو مشخص می‌کند چه تعداد داده به‌عنوان زیرمجموعه یک ویژگی شناخته شود. همچنین در روش جستجو، نوع جستجوی یک عامل در بین زیرمجموعه‌های ایجاد شده نیز مشخص می‌گردد.
- مقایسه‌گر[۲]: نوع مقایسه و پیدا کردن شباهت‌ها و تفاوت‌های میان مقادیر صفت‌ها توسط مقایسه‌گر مشخص می‌گردد. برای مثال، یک مقایسه‌گر ممکن است، تعداد تکرار را به‌عنوان شباهت در نظر بگیرد، یا مقایسه‌گر دیگر، فواصل بین داده‌های یک صفت را به‌عنوان عامل اصلی شباهت میان مقادیر، محسوب نماید.
- در روش پیشنهادی برای بخش جستجو از الگوریتم BESTFIRST و برای بخش مقایسه ویژگی از الگوریتم CFSSUBSETEVAL استفاده خواهیم کرد.
- الگوریتم BESTFIRST جزء الگوریتم‌های جستجوی مبتنی بر درخت[۳] است. جهت کاهش تعداد ویژگی‌های داده‌ها که باعث افزایش سرعت نهایی می‌شود، الگوریتم ۱ به شرح زیر بیان می‌شود.

(شروع الگوریتم۱): این الگوریتم در روش پیشنهادی، هر زیرمجموعه‌ای از داده‌ها را به عنوان یک گره از درخت در نظر می‌گیرد. گره‌ای که کمترین هزینه را داشته باشد، بسط داده می‌شود.

- این کار با در نظر گرفتن همه ویژگی‌ها به عنوان زیرمجموعه بهینه آغاز می‌گردد و پس از تقسیم و رسیدن به زیرمجموعه‌ای که دارای یک یا چند ویژگی است پایان می‌یابد. این عمل می‌تواند در جهت عکس نیز انجام گیرد یعنی با خالی کردن گره حاوی زیرمجموعه

---

[۱] Search method

[۲] Attribute evaluator

[۳] Tree– Search

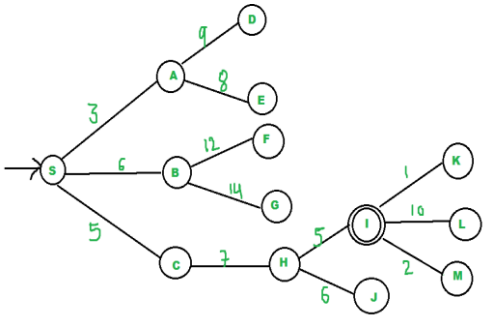
بهینه، هر ویژگی به‌عنوان یک زیرمجموعه در نظر گرفته شده و هزینه آن محاسبه شود، سپس ترکیب‌های ۲ تایی از صفات مورد بررسی قرار گرفته و پس از آن ترکیب‌های ۳ تایی و ... تا زمانی این عمل ادامه پیدا خواهد کرد که به مجموعه کل صفات برسیم. درنهایت از بین زیرمجموعه‌های انتخاب‌شده، زیرمجموعه‌ای که داری کمترین هزینه باشد به‌عنوان مجموعه بهینه انتخاب خواهد شد. (پایان الگوریتم۱)

در ادامه بحث کاهش تعداد ویژگی داده‌ها، برای پیدا کردن شباهت‌ها و تفاوت‌های میان مقادیر صفت‌ها از کل داده‌ها، الگوریتم دو به صورت زیر بیان می‌شود.

(شروع الگوریتم۲): الگوریتم CFSSUBSETEVAL به مقادیری که دارای کمترین درجه همبستگی با سایر مقادیر درگیر در پژوهش هستند، کمترین ارزش را اعطا می‌کند. در حالت کلی این الگوریتم با جستجو در میان ویژگی‌ها، ویژگی‌هایی را که دارای بیشترین مقادیر داده تکراری هستند را شناسایی نموده و به آن‌ها ارزش کمتری می‌بخشد. هر مقدار تکراری، از ارزش محاسباتی یک ویژگی می‌کاهد. این الگوریتم به همراه الگوریتم BESTFIRST به کار گرفته می‌شود و می‌تواند برای هر مقدار تکراری یک واحد به هزینه یک گره اضافه نماید، در نهایت هزینه هر گره با توجه به هزینه هر عضو از زیرمجموعه ویژگی‌های آن گره و تعداد اعضای آن گره محاسبه خواهد شد. (پایان الگوریتم۲)

اگر بخواهیم نحوه عمل این دو الگوریتم با هم را مورد بررسی قرار دهیم می‌توانیم بگوییم که CFSSUBSETEVAL مسئول ارزش‌دهی هر ویژگی است و BESTFIRST بهترین زیرمجموعه از ویژگی‌ها را با توجه به ارزش اعطا شده، انتخاب می‌کند و حاصل کار این دو الگوریتم، انتخاب بهینه از تعداد ویژگی‌های اصلی و کلیدی داده‌ها (علت مراجعه، کد جنسیت، سابقه بیمه پردازی و نتایج کمیسیون‌های قبلی از شکل ۴ و با اجرای بخش ۴-۱ و ۴-۲ از روش پیشنهادی) می‌باشد.

اگر شکل ۳ را به ترتیب ویژگی‌های ارزش‌گذاری‌شده توسط CFSSUBSETEVAL با این قاعده که ارزش مورد نظر با توجه به تعداد داده‌های یکتا در هر ویژگی تعیین‌شده و یال آن ویژگی نشان‌دهنده میزان ارزش آن ویژگی است، می‌توان با استفاده از روش BESTFIRST به شرح زیر به، با ارزش‌ترین زیر مجموعه از ویژگی‌ها دست‌یافت.



شکل (۳): مجموعه‌ای از ویژگی‌های مورد بررسی با استفاده از CFSSUBSETEVAL و BESTFIRST

ابتدا از گره با میزان s شروع نموده و درخت مورد بررسی را با یالی که دارای بیشترین ارزش یعنی ۶ است، ادامه می‌دهیم، تا به B برسیم. اگر B یک برگ باشد که الگوریتم پایان پیدا می‌کند، در غیر این صورت دوباره یالی با بیشترین ارزش انتخاب می‌گردد. این کار تا زمانی که به یک گره برگ

برسیم ادامه پیدا می‌کند. در این مثال مسیر B به G نیز دارای ارزش ۱۴ است. در نهایت مسیر پیموده شده S به G به عنوان کاندید با ارزش‌ترین مسیر و ویژگی‌های S,B,G به عنوان با ارزش‌ترین ویژگی‌ها انتخاب می‌گردند.

هر چند می‌توان دید که بهترین مجموعه از ویژگی‌های S,C,H,I,L با ارزش‌ترین مجموعه ویژگی است، اما این روش می‌تواند در زمان اندک و با پیچیدگی نسبتا پایین بهترین نتایج را رقم بزند که نمونه‌ای از داده‌های پیش‌پردازش شده در شکل ۴ نشان داده شده است.

منطقاً نیز این محاسبه قابل دفاع است چرا که ویژگی‌هایی با مقادیر تکراری قاعدتاً در نتیجه‌گیری کلی تأثیرگذار نخواهند بود. برای مثال وقتی همه نمونه‌های یک مجموعه داده از نظر ویژگی نوع وسیله نقلیه، وسیله نقلیه سبک هستند، نمی‌توان تأثیری را برای ویژگی نوع وسیله نقلیه در نتیجه‌گیری کلی در نظر گرفت.

۴-۲. نرمال‌سازی داده‌ها

داده‌های موجود در مجموعه داده‌ها مورد استفاده دارای سه نوع مقدار به شرح زیر است:

- عددی
- رشته‌ای
- مقادیر منطقی شامل دو مقدار ۰ و ۱

استفاده از این مقادیر در رابطه محاسبات عددی مورد استفاده در این پژوهش عملاً غیرممکن است. نخستین عملی که برای ایجاد یک داده قابل استفاده باید انجام داد تبدیل تمامی مقادیر موجود به مقادیر عددی است.

برای تبدیل داده‌های رشته‌ای به عددی از رابطه زیر بهره می‌بریم:

$$X_i = \frac{t_i}{T_{all}} \qquad (۱)$$

که در آن X_i مقدار عددی که قرار است جایگزین مقدار رشته‌ای شود، t_i تعداد تکرار آن رشته در مجموعه داده‌ها و T_{all} تعداد کل داده‌های موجود برای آن ویژگی است. بر این اساس مقدار X_i همواره عددی بین ۰ تا ۱ خواهد بود.

بعد از تبدیل همه رشته‌ها (وضعیت از کارافتادگی، علت مراجعه، جنسیت، نوع بیمه و نتایج کمیسیون‌های قبلی از شکل ۲) به عدد برای از بین بردن تأثیر داده‌های خیلی بزرگ در محاسبات باید فواصل بازه‌های عددی به میزان ثابتی برای همه ویژگی‌ها تبدیل گردد. به همین دلیل از رابطه (۲) استفاده خواهد شد.

$$X_i = \frac{D_i - Min}{Max - Min} \qquad (۲)$$

این رابطه داده‌ها را در بازه‌ای بین ۰ و ۱ نگاشت می‌کند. که در آن X_i مقدار عددی نرمال شده، D_i مقدار اصلی داده در دیتاست است. Min و Max به ترتیب بیشترین و کمترین داده موجود برای آن ویژگی در دیتاست می‌باشند.

Purchased	Gender	Estimated	Age	id	1
0	0	19000	19	1	2
0	0	20000	35	2	3
0	1	43000	26	3	4
0	1	57000	27	4	5
0	0	76000	19	5	6
0	0	58000	27	6	7
0	1	84000	27	7	8
1	1	150000	32	8	9
0	0	33000	25	9	10
0	1	65000	35	10	11
0	1	80000	26	11	12
0	1	52000	26	12	13
0	0	86000	20	13	14
0	0	18000	32	14	15
0	0	82000	18	15	16
0	0	80000	29	16	17
1	0	25000	47	17	18
1	0	26000	45	18	19
1	0	28000	46	19	20
1	1	29000	48	20	21
1	0	22000	45	21	22
1	1	49000	47	22	23
1	0	41000	48	23	24
1	1	22000	45	24	25
1	0	23000	46	25	26
1	0	20000	47	26	27
1	0	28000	49	27	28
1	1	30000	47	28	29

شکل (۴): نمونه‌ای از مجموعه داده‌ها پس از پیش‌پردازش

مهم‌ترین فاکتور برای مقایسه دو سیستم پیش بینی (مدل‌های شبیه‌سازی با استفاده از روش‌های هوش مصنوعی مطرح‌شده در [۲۶])، میزان کارایی[۱] (P_r) یک سیستم پیش بینی است. کارایی یک سیستم پیش بینی تحت تأثیر دو عامل است. هزینه اجرا و دقت. این دو متغیر خود دارای بخش‌های متعددی هستند اما در این پژوهش آن‌ها را به صورت زیر محدود نموده‌ایم:

- ✓ هزینه اجرا[۲] (C_i): که خود شامل:
  - هزینه دستیابی به منابع[۳] (C_s) از جمله:
    - هزینه دستیابی به منابع داده[۴] (C_{sd})
    - هزینه دستیابی به منابع پردازشی (C_{sp})[۵]
  - هزینه اجرای روش[۶] (C_r): که شامل موارد زیر است:
    - هزینه‌های زمانی[۷] (C_{rt})
    - هزینه پردازشی[۸] (C_{rp})
    - هزینه حافظه[۹] (C_{rm})
- ✓ دقت[۱۰] (A_c): در واقع دقت را میزان تطابق دنیای واقعی با مقدار حاصل شده در پیش بینی در نظر خواهیم گرفت.

[۱] Performance
[۲] Implementation cost
[۳] Cost of access to resources
[۴] Cost of acquiring data sources
[۵] Cost of access to processing resources
[۶] Cost of running
[۷] Time costs
[۸] Processing costs
[۹] Memory cost
[۱۰] Accuracy

برای به دست آوردن رابطه‌ای خطی میان متغیرهای تأثیرگذار بر کارایی روش پیش بینی باید به بررسی تک تک هر یک از متغیرها پرداخت، اما به صورت کلی می‌توان در یک رابطه ساده اظهار کرد که کارایی یک روش پیش بینی با میزان دقت رابطه مستقیم و با میزان هزینه رابطه عکس دارد. این مسئله را می‌توان در رابطه (۳) مشاهده نمود.

$$P_r \propto \frac{A_c}{C_i} \qquad (۳)$$

برای تبدیل این رابطه به یک رابطه تساوی با قابلیت تخمین درست، نیازمند ضرایبی هستیم که بتواند شرحی درست از میزان تأثیرگذاری هریک از متغیرهای مورد بررسی را ارائه دهد. این ضرایب را به ترتیب w_1 و w_2 می‌نامیم. بنابراین رابطه (۳) به رابطه (۴) تبدیل می‌گردد:

$$P_r = \frac{w_1 A_c}{w_1 C_i} \qquad (۴)$$

مقدار w_1 و w_2 وابسته به نوع مسئله پیش بینی است و باید از تاریخچه روش مورد بررسی و سنجش میزان همبستگی میان دو متغیر A_c و C_i با متغیر P_r به دست آید.

با بررسی جزئی C_i می‌توان یک رابطه تناسب بین اجزای سازنده‌ی آن یافت. بدیهی است که رابطه میان C_s و C_r با C_i یک رابطه مستقیم است. پس به این ترتیب می‌توان گفت:

$$C_i \propto (C_s + C_r) \qquad (۵)$$

همین طور در مورد C_s و C_r نیز می‌توان روابط (۶) و (۷) را بیان نمود.

$$C_s \propto (C_{sd} + C_{sp}) \qquad (۶)$$

$$C_r \propto (C_{rp} + C_{rt} + C_{rm}) \qquad (۷)$$

با همان منطق ذکر شده برای w_1 و w_2 می‌توان روابطِ (۵)، (۶)، (۷) را به روابط تساوی تبدیل نمود. به این ترتیب خواهیم داشت:

$$C_i = w_3 C_s + w_4 C_2 \qquad (۸)$$

$$C_s = w_5 C_{sd} + w_6 C_{sp} \qquad (۹)$$

$$C_r = w_7 C_{rp} + w_8 C_{rt} + w_9 C_{rm} \qquad (۱۰)$$

دقت اما مقوله‌ای است که با بررسی میزان تطابق یک پیش بینی با میزان رخداد آن در واقعیت سنجیده خواهد شد. اگر یک پدیده مانند X دارای n ویژگی قابل سنجش باشد، آن پدیده به صورت مجموعه‌ای از مقادیر قابل سنجش‌اش به صورت زیر تعریف خواهد شد.

$$X = (x_1, x_2, ..., x_n) \qquad (۱۱)$$

به ترتیب مقادیر x_i میزان ویژگی i ام برای پدیده X است. میزان دقت پیش بینی پدیده مورد نظر با استفاده از رابطه زیر سنجیده خواهد شد.

(۱۲)

$$A_c = \frac{1}{\sqrt{(x_1 - f_1)^2 + (x_2 - f_2)^2 + ... + (x_n - f_n)^r}}$$

در این رابطه f_i نشان دهنده میزان پیش بینی برای ویژگی i ام پدیده X خواهد بود. بدیهی است هدف نهایی این پژوهش افزایش مقدار P_r است.

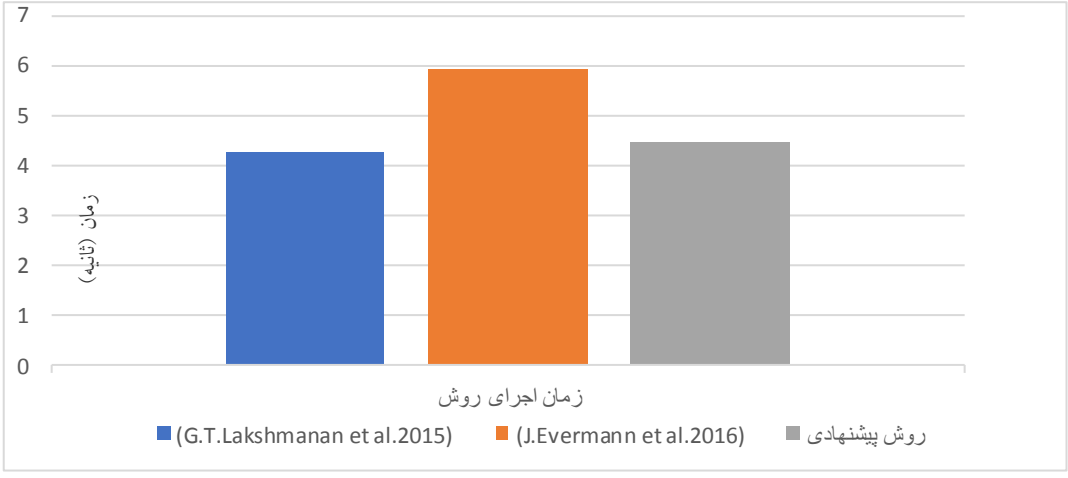
۵- یافته‌های پژوهش

با تعداد تصادفی ۸۴۶ داده از بیمه شدگان سازمان تامین اجتماعی که در سال ۱۳۹۸ درخواست ازکارافتادگی داده‌اند و جهت تشخیص به کمیسیون پزشکی مربوطه معرفی شده‌اند و انتخاب نام ستون‌ها به (reason, sex , sex_r , age , type, history , exp_s , exp_f , last_result , id , id_num , state) که اکثر مقادیر به صورت رشته‌ای بوده و تبدیل داده‌های رشته‌ای به عددی

در حالی‌که تغییری در داده‌های عددی نداشته باشیم با یک مجموعه داده کاملا عددی مواجه خواهیم شد.

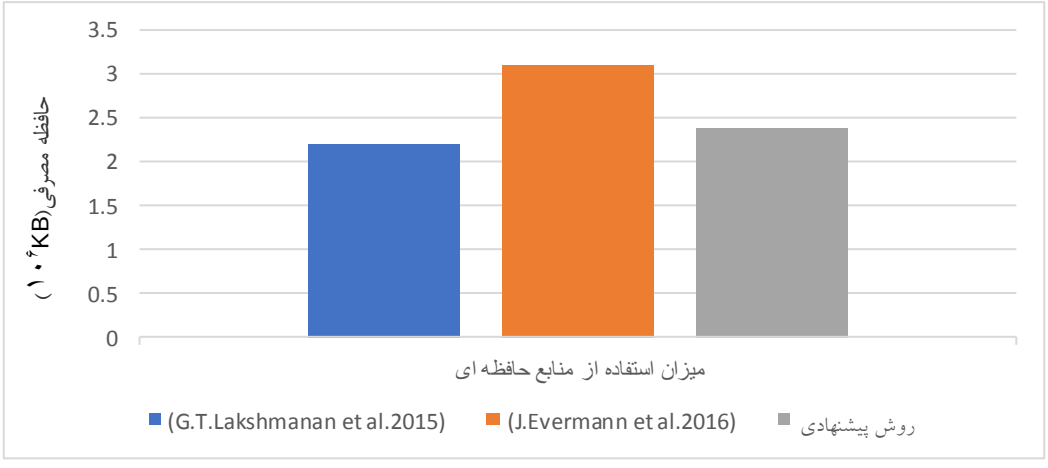
تعداد یازده ستون ویژگی که شامل اطلاعات عمومی افراد بیمه شده و هفت ستون از اطلاعات مربوط به سوابق بیمه‌شده و مشخصات بیماری، با استفاده از نرم افزار Weka و الگوریتم‌های Cfssubseteval و Bestfirst تبدیل یافته و نهایتا به چهار ویژگی reason , sex_r , history و Last_result در ستون های ۳، ۵، ۸ و ۱۱ کاهش یافته و پس از نرمال‌سازی به بازه [0,1] نگاشت می‌گردند.

پس از اجرای شبیه‌سازی روش ارائه‌شده در بخش ۴ و مطرح‌شده در [۲۶] با نرم افزار متلب و مقایسه آن با دو نمونه از روش‌های معتبر علمی [۵] و [۱۹] با داده‌های تعریف شده، میزان زمان مصرفی هریک از روش‌های مورد مطالعه در شکل ۵ نشان داده شده است.



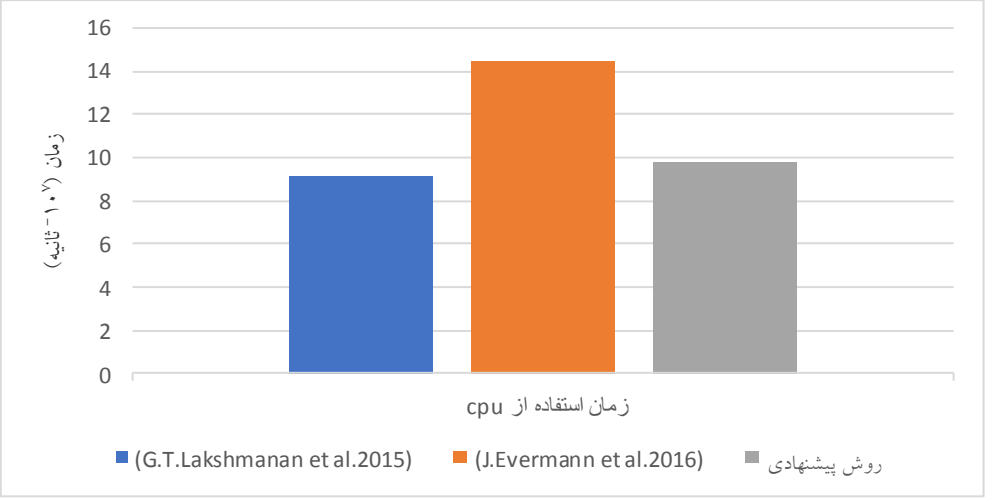
شکل (۵): میزان زمان مصرفی در ۳ روش مورد مقایسه

میزان استفاده از منابع حافظه‌ای در روش پیشنهادی مقاله حاضر و مقایسه آن با دو نمونه از روش‌های مذکور [۲۶] با همان داده‌های سازمان تامین اجتماعی در شکل ۶ به نمایش گذاشته شده است.



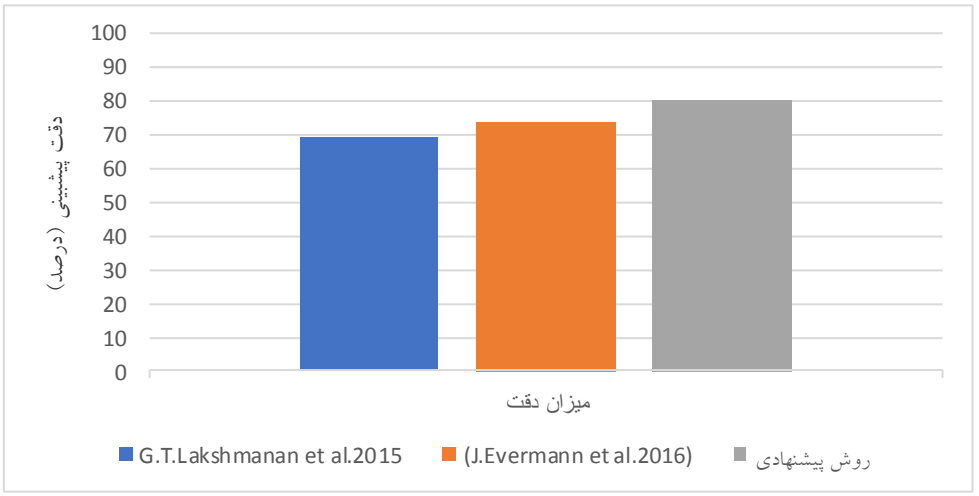
شکل (۶): میزان استفاده از منابع حافظه‌ای (RAM)

پس از اجرای شبیه سازی روش پیشنهادی ارائه شده در بخش ۴ و مقایسه آن با دو نمونه از روش‌های فوق با داده‌های تعریف شده، میزان زمان استفاده از CPU هر یک از روش‌های مورد مطالعه در شکل ۷ نشان داده شده است.



شکل (۷): زمان استفاده از CPU

میزان دقت کسب شده از روش پیشنهادی مقاله و مقایسه آن با دو نمونه از روش های صدرالذکر و با داده‌های سازمان تامین اجتماعی پس از اجرای شبیه سازی در شکل ۸ به نمایش گذاشته شده است. استفاده از روش پیشنهادی پیش‌پردازش داده‌ها، بیش از ۳۷ درصد از ویژگی‌ها را کاهش داده که در حجم زیاد داده‌ها در علم داده کاوی از اهمیت فوق العاده‌ای برخوردار بوده است. ضمنا با بررسی شکل‌های ۵، ۶ و۷ به این نتیجه می‌رسیم که روش پیشنهادی باکمی افزایش منابع حافظه‌ای، کاهش زیادی در میزان زمان استفاده از CPU نسبت به دو روش مورد مقایسه و همچنین افزایش دقت و کارایی طبق شکل ۸ که از نقاط قوت می‌باشد، داشته است.



شکل (۸): میزان دقت در روش‌های مورد مطالعه

۵-۱. چارچوب پیشنهادی معماری سازمانی

با توجه به نتایج بدست آمده از روش ارائه‌شده پیش‌پردازش و استفاده در مدل‌های پیش‌بینی و اخذ نتایج با دقت بالاتر به همراه کاهش منابع هزینه‌ای در بخش داده‌های از کارافتادگی، می توان واحدهای اجرایی سازمان تامین اجتماعی را از نتایج این پژوهش برخوردار نموده تا در بررسی ویژگی‌های موثر در فرآیند ار کارافتادگی بکارگرفته شده و همچنین از مدل ارائه شده در سایر فرآیندهای سازمان مذکور (در بخش بیمه‌ای و درمان) و بهره برداری مدیران سطح میانی از نتایج این تحقیق استفاده نمود.

۶- بحث و نتیجه‌گیری

استفاده از نتایج پیش‌پردازش ارائه شده در این مقاله، طبق شکل‌های ۵ تا ۸، دستاورد جدیدی در مدل‌های پیش بینی فرآیندهای کسب و کار بیمه‌های اجتماعی و در جهت رفع چالش‌های موجود شامل ویژگی‌های فراوان، داده‌های ناهمگن، توزیع نامتوازن و نرخ پایین تخصیص داده‌ها ارائه می‌دهد.

در این روش پیش‌پردازش، برای بخش جستجو از الگوریتم BESTFIRST و برای بخش مقایسه ویژگی از الگوریتم CFSSUBSETEVAL استفاده‌شده و با در نظر گرفتن همه ویژگی‌ها به عنوان زیرمجموعه بهینه آغاز می‌گردد. در این پیش‌پردازش، مقادیر رشته‌ای به عددی تبدیل‌شده و همچنین تعداد ویژگی کاهش‌یافته و نهایتا داده‌های باقیمانده به روش نرمال سازی به بازه [۰،۱] نگاشت شده است.

برای نشان دادن عملکرد هر روش، مقایسه آن با روش‌های قبلی، امری متداول بوده بنابراین نتایج حاصل از این تحقیق، با دو روش معتبر [۵] و [۱۹] مقایسه گردیده است. استفاده از روش ارائه شده، میزان ۳۷ درصد از ویژگی‌ها را کاهش داده که در تعداد بالای داده‌ها در علم داده کاوی از اهمیت بالایی برخوردار بوده و در عین حال با افزایش نسبی منابع حافظه‌ای، کاهش چشمگیری در میزان زمان استفاده از CPU نسبت به روش‌های مورد مقایسه و همچنین افزایش میزان دقت و کارایی داشته است.

موارد قابل استفاده از نتایج این پژوهش در موارد کاربردی با داده‌های مختلف و با محدودیت نتایج صفر یا یک می‌باشد. با توجه به اینکه مقوله پیش‌بینی در امر سیاست مانند انتخاب رییس جمهور از میان کاندیداتورهای مختلف، در امر اقتصاد مانند پیش بینی رشد اقتصادی و در امر درمان مانند پیش بینی بهبودی یا فوت بیماران مبتلا به کووید ۱۹ و با علائم مختلف، بسیار مهم بوده و دارای پیامدهای گرانبهایی می‌تواند باشد لذا این مقاله با افزایش دقت همراه با کاهش نسبی منابع مصرفی، سهم اندکی در پیش بینی مسائل روزمره و حیاتی دنیای امروز خواهد داشت.

برای تحقیقات آینده می‌توان ایجاد مدلی که بتواند عمل پیش‌پردازش را به روی داده‌های با نتایجی حتی غیر از صفر و یک نیز انجام دهد، پیشنهاد کرد. همچنین می‌توان بر روی روش‌هایی که ویژگی‌های بیشتری را پوشش دهد، کار کرده و نتایج را مقایسه نموده و بهره برداری از مدل ارائه‌شده در داده‌های مختلف سازمان تامین اجتماعی و کسب نتایج کاربردی پیشنهاد می‌گردد.

سپاسگزاری

ما از حمایت مالی و معنوی واحد آموزش و پژوهش اداره کل منابع انسانی سازمان تامین اجتماعی تشکر می‌کنیم.

فهرست منابع

[۱] A. Khan, H. Le, K. Do, T. Tran, A. Ghose, H. Dam, R. Sindhgatta. Memory-augmented neural networks for predictive process analytics: arXiv preprint arXiv: ۱۸۰۲. ۰۰۹۳۸ (۲۰۱۸).

[۲] A. E. Márquez-Chamorro, M. Resinas, A. Ruiz-Corts.Predictive monitoring of business processes: a survey. IEEE Transactions on Services Computing ۹۹:۱-۱ (۲۰۱۷).

[۳] L. Lin, L. Wen, J. Wang. Mm-pred: A deep predictive model for multi-attribute event sequence. In Proceedings of the ۲۰۱۹ SIAM international conference on data mining: Society for Industrial and Applied Mathematics. ۱۱۸-۱۲۶ (۲۰۱۹).

[٤] J.-P. Briot, G. Hadjeres, F. Pachet. Deep learning techniques for music generation-a survey. arXiv preprint arXiv: ۱۷۰۹,۰۱٦۲. (۲۰۱۹)

[٥] J. Evermann, J.-R. Rehse, P. Fettke. A deep learning approach for predicting process behavior at runtime: in International Conference on Business Process Management, Springer ۳۲۷-۳۳۸ (۲۰۱٦).

[٦] Y. LeCun, Y. Bengio, G. Hinton. Deep learning: Nature ٥۲۱: ٤۳٦–٤٤٤ (۲۰۱٥).

[۷] T. A. Gibson, J. A. Henderson, J. Wiles. Predicting temporal sequences using an event-based spiking neural network incorporating learnable delays ۳۲۱۳-۳۲۲۰ (۲۰۱٤).

[۸] F. Daniel, K. Barkaoui, S. Dustdar,eds. Process Mining Manifesto: in Business Process Management Workshops. IEEE Task Force on Process Mining. vol. ۹۹ of Lecture Notes in Business Information Processing, Springer-Verlag, Berlin ۱٦۹-۱۹٤ (۲۰۱۲).

[۹] J. Kumar, A. K. Singh. Workload prediction in cloud using arti_cial neural network and adaptive di_erential evolution. Future Generation Computer Systems ۸۱: ٤۱-٥۲ (۲۰۱۸).

[۱۰] N. Tax, I. Verenich, M. La Rosa, M. Dumas. Predictive business process monitoring with lstm neural networks: In International Conference on Advanced Information Systems Engineering, ٤۷۷–٤۹۲ (۲۰۱۷).

[۱۱] W. M. P. van der Aalst, M.H. Schonenberg, M. Song. Time prediction based on process mining. Inf. Syst ۳٦(۲): ٤٥۰–٤۷٥ (۲۰۱۱).

[۱۲] X. Jiang, S. Pan, G. Long, F. Xiong, J. Jiang, C. Zhang. Cost-sensitive parallel learning framework for insurance intelligence operation. Transactions on Industrial Electronics ۱–۱۱ (۲۰۱۸).

[۱۳] M.A. Jassim, S.N. Abdulwahid. Data Mining preparation: Process, Techniques and Major Issues in Data Analysis, IOP Conf. Ser.: Mater. Sci. Eng. ۱۰۹۰ ۰۱۲۰٥۳ (۲۰۲۱).

[۱٤] A.K. Dubey, A. Kumar, R . Agrawal. An efficient ACO- PSO- based framework for data classification and preprocessing in big data, Evolutionary Intelligence, part of Springer Nature ۲۰۲۰ https://doi. org/۱۰. ۱۰۰۷/s۱۲۰٦٥-۰۲۰-۰۰٤۷۷-۷ (۲۰۲۰)

[۱٥] S. García, S. Ramírez-Gallego, J. Luengo, J.M. Benítez, F.,Herrera. Big data preprocessing: methods and prospects, Big Data Analytics, DOI ۱۰,۱۱۸٦/s٤۱۰٤٤-۰۱٦-۰۰۱٤-۰ (۲۰۱٦).

[۱٦] J. A. Cortés-Ibáñez, S. González, J. J. Valle-Alonso, J. Luengo, S. García, F. Herrera. Preprocessing methodology for time series: an industrial world

monitoring with lstm neural networks: in International Conference on Advanced Information Systems Engineering. Springer ٤٧٧-٤٩٢ (٢٠١٧).

[٢٣] N. Navarin, B. Vincenzi, M. Polato, A. Sperduti. Lstm networks for data-aware remaining time prediction of business process instances: arXiv preprint arXiv: ١٧١١,٠٣٨٢٢ (٢٠١٧).

[٢٤] J. Evermann, J. Rehse, P. Fettke. Predicting process behaviour using deep learning: Decision Support Systems. (٢٠١٧).

[٢٥] P.H. Klees, M. leoni A. W. Veenstra. The Application of Comparative Process Mining in Logistics.Student theses are made available in the TU/e repository upon obtaining the required degree. (٢٠١٧).

[٢٦] M. Fadaei Pellehshahi, S. Kordrostami, A.H. Refahi Sheikhani, M. Faridi Masouleh. Predicting business processes of the social insurance using recurrent neural network and Markov chain: Journal of Modelling in Management, Vol. ahead-of- print No. ahead- of- print. (٢٠٢١).

application case study. Information Sciences, ٥١٤, ٣٨٥-٤٠١ (٢٠٢٠).

[١٧] H.woo, J. kim, W. Lee. Validation of Text Data Preprocessing Using a Neural Network Model: Mathematical Problems in Engineering. (٢٠٢٠).

[١٨] G. report.Critical Capabilities for Composite Content Management Applications (٢٠١٠).

[١٩] G. T. Lakshmanan, D. Shamsi, Y. N. Doganata, M. Unuvar, R. Khalaf. A markov predictionmodel for data-driven semi-structured business processes: Knowledge and Information Systems ٤٢: ٩٧-١٢٦ (٢٠١٥).

[٢٠] W.V. Aalst, A. Adriansyah, A.K. Medeiros, F. Arcieri, T. Baier, T. Blickle, J. Chandra Bose, P.Van Den Brand, R. Brandtjen, J. Buijs. Process mining manifesto: In International Conference on Business Process Management ١٦٩–١٩٤ (٢٠١١).

[٢١] J. Schmidhuber. Deep learning in neural networks: An overview. Neural Networks ٦١: ٨٥–١١٧ (٢٠١٥).

[٢٢] N. Tax, I. Verenich, M. La Rosa, M. Dumas. Predictive business process