دانشگاه آزاد اسلامی

واحد علوم و تحقیقات

دانشکده مدیریت و اقتصاد، گروه مدیریت کسب و کار

پایان‌نامه برای دریافت درجه کارشناسی ارشد در رشته مدیریت کسب و کار (M.A)

گرایش: سیستم‌های اطلاعاتی و فناوری اطلاعات

عنوان:

ارائه روشی برای سازماندهی محتوا در فرایند مدیریت دانش سازمانی با رویکرد شبکه‌های عصبی مصنوعی

استاد راهنما:

دکتر محمدرضا موحدی صفت

نگارش:

عباس اسمعیلی

زمستان ۱۴۰۱


بسم الله الرحمن الرحيم

معاونت پژوهش و فن آوری

به نام خدا

منشور اخلاق پژوهش

بایاری از خداوند سبحان و اعتقاد به این که عالم محضر خداست و همواره ناظر بر اعمال انسان و به منظور پاس داشت مقام بلند دانش و پژوهش و نظر به اهمیت جایگاه دانشگاه در اعتلای فرهنگ و تمدن بشری، ما دانشجویان و اعضاء هیات علمی واحدهای دانشگاه آزاد اسلامی متعهد می گردیم اصول زیر را در انجام فعالیت های پژوهشی مد نظر قرار داده و از آن تخطی نکنیم:

۱- اصل حقیقت جویی: تلاش در راستای پی جویی حقیقت و وفاداری به آن و دوری از هرگونه پنهان سازی حقیقت.

۲- اصل رعایت حقوق: التزام به رعایت کامل حقوق پژوهشگران و پژوهیدگان (انسان، حیوان و نبات) و سایر صاحبان حق.

۳- اصل مالکیت مادی و معنوی: تعهد به رعایت کامل حقوق مادی و معنوی دانشگاه و کلیه همکاران پژوهش.

۴- اصل منافع ملی: تعهد به رعایت مصالح ملی و در نظر داشتن پیشبرد و توسعه کشور در کلیه مراحل پژوهش.

۵- اصل رعایت انصاف و امانت: تعهد به اجتناب از هرگونه جانب داری غیر علمی و حفاظت از اموال، تجهیزات و منابع در اختیار.

۶- اصل رازداری: تعهد به صیانت از اسرار و اطلاعات محرمانه افراد، سازمان ها و کشور و کلیه افراد و نهادهای مرتبط با تحقیق.

۷- اصل احترام: تعهد به رعایت حریم ها و حرمت ها در انجام تحقیقات و رعایت جانب نقد و خودداری از هرگونه حرمت شکنی.

۸- اصل ترویج: تعهد به رواج دانش و اشاعه نتایج تحقیقات و انتقال آن به همکاران علمی و دانشجویان به غیر از مواردی که منع قانونی دارد.

۹- اصل برائت: التزام به برائت جویی از هرگونه رفتار غیر حرفه ای و اعلام موضع نسبت به کسانی که حوزه علم و پژوهش را به شائبه های غیر علمی می آلایند.

تعهدنامه اصالت رساله یا پایان‌نامه

اینجانب عباس اسمعیلی دانش‌آموخته مقطع کارشناسی ارشد ناپیوسته در رشته مدیریت کسب و کار که در تاریخ ..... از پایان‌نامه خود تحت عنوان " ارائه روشی برای سازماندهی محتوا در فرایند مدیریت دانش سازمانی با رویکرد شبکه‌های عصبی مصنوعی" با کسب نمره ..... و درجه ..... دفاع نموده‌ام بدینوسیله متعهد می‌شوم:

۱) این پایان‌نامه حاصل تحقیق و پژوهش انجام شده توسط اینجانب بوده و در مواردی که از دستاوردهای علمی و پژوهشی دیگران (اعم از پایان‌نامه، کتاب، مقاله و...) استفاده نموده‌ام، مطابق ضوابط و رویه موجود، نام منبع مورد استفاده و سایر مشخصات آنرا در فهرست مربوطه ذکر و درج کرده‌ام.

۲) این پایان‌نامه قبلاً برای دریافت هیچ مدرک تحصیلی (هم‌سطح، پایین‌تر یا بالاتر) در سایر دانشگاهها و مؤسسات آموزش عالی ارائه نشده است.

۳) چنانچه بعد از فراغت از تحصیل، قصد استفاده و هر گونه بهره‌برداری اعم از چاپ کتاب، ثبت اختراع و... از این پایان‌نامه داشته باشم، از حوزه معاونت پژوهشی واحد مجوزهای مربوطه را اخذ نمایم.

٤) چنانچه در هر مقطع زمانی خلاف موارد فوق ثابت شود، عواقب ناشی از آن را می‌پذیرم و دانشگاه آزاد اسلامی واحد علوم و تحقیقات مجاز است با اینجانب مطابق ضوابط و مقررات رفتار نموده و در صورت ابطال مدرک تحصیلی‌ام هیچگونه ادعایی نخواهم داشت.

عباس اسمعیلی

تاریخ و امضاء

دانشگاه آزاد اسلامی

واحد علوم و تحقیقات

دانشکده مدیریت و اقتصاد، گروه مدیریت کسب و کار

پایان‌نامه برای دریافت درجه کارشناسی ارشد در رشته مدیریت کسب و کار (M.A)

گرایش: سیستم‌های اطلاعاتی و فناوری اطلاعات

عنوان:

ارائه روشی برای سازماندهی محتوا در فرایند مدیریت دانش سازمانی با رویکرد شبکه‌های عصبی مصنوعی

استاد راهنما:

دکتر محمدرضا موحدی صفت

نگارش:

عباس اسمعیلی

زمستان ۱۴۰۱

سپاسگزاری

عرض تشکر ویژه از جناب دکتر محمدرضا موحدی صفت، استاد راهنمای دانا و فهمیده‌ای که با گوشزد کردن نکات کلیدی، مرا در مسیر نگارش پایان‌نامه راهنمایی کردند.

همچنین از تمامی اساتید دانشکده‌ی مدیریت و اقتصاد دانشگاه آزاد واحد علوم و تحقیقات، از جمله مدیر گروه رشته‌ی مدیریت کسب و کار، جناب دکتر حسین وظیفه دوست، سپاس گزارم. این عزیزان همواره کمک و همراه بنده بودند و دانش خویش را بی هیچ مضایقه‌ای، در اختیار من قرار دادند. ضمناً از جناب دکتر ابراهیم نظری فرخی که در انتخاب موضوع پایان‌نامه به من یاری رساندند، متشکرم.

در نگارش این پایان‌نامه از منابع علمی و کتابخانه‌های نرم‌افزاری بسیاری بهره گرفته شده که بدون وجود آن‌ها، نگارش این پایان‌نامه غیرممکن بود. محصولات حوزه‌ی علوم داده و یادگیری عمیق شرکت گوگل، بستر لازم برای انجام این پایان‌نامه را فراهم کردند. این محصولات حاصل زحمات کارکنان این شرکت، به‌خصوص تیم تحقیقاتی آن است.

در انتها، از تمامی عزیزانی که مرا در نگارش این پایان‌نامه یاری کردند، به ویژه افرادی که از ذکر نام ایشان غفلت شد، ممنون و متشکرم.

فهرست مطالب

عنوان **شماره صفحه**

فصل سوم: روش اجرای تحقیق

فصل چهارم: تجزیه و تحلیل داده‌ها و یافته‌ها

فصل پنجم: نتیجه‌گیری و پیشنهادات

فهرست جداول

عنوان **شماره صفحه**

فهرست نمودارها

عنوان **شماره صفحه**

فهرست شکل‌ها

عنوان **شماره صفحه**

چکیده

امروزه جهان با سرعت زیادی در حال تغییر است. سازمان‌ها همواره با فرصت‌ها و تهدیدهای جدیدی مواجه می‌شوند. توانایی رویارویی با تغییرات و استفاده‌ی حداکثری از فرصت‌های حاصله، یک ضرورت اساسی برای سازمان‌های کوچک و بزرگ به شمار می‌رود. در اینجاست که نقش اطلاعات و دانش سازمانی پر رنگ‌تر می‌شود. محتویات موجود در سیستم مدیریت دانش، ملزومات تصمیم‌گیری را در سطوح مختلف مدیریتی فراهم می‌سازند. لذا سازماندهی مناسب و بهینه‌ی محتویات ورودی به سیستم مدیریت دانش، زمینه‌ساز اتخاذ تصمیمات بهتر شده و سازمان‌ها را قادر به مواجهه با شرایط محیطی متنوع می‌سازد. علوم داده، به‌خصوص مبحث نوین یادگیری عمیق، تحولی عظیم در عرصه‌ی سازماندهی داده‌ها و اطلاعات ایجاد کرده است. بهره جستن از زمینه‌های ناب یادگیری عمیق، نه تنها یک مزیت، بلکه یک ضرورت رقابتی برای سازمان‌ها است. در پژوهش حاضر، با اتکا بر جدیدترین رویکردهای مبحث یادگیری عمیق و شبکه‌های عصبی مصنوعی، روشی ارائه شده که با کارایی و اثربخشی بی‌نظیر، امکان سازماندهی محتویات ورودی به سیستم مدیریت دانش را ایجاد می‌کند. عملکرد این روش با سایر روش‌های مطرح حوزه‌ی یادگیری ماشین مقایسه شده و مزیت‌ها و معایب آن مورد بررسی قرار گرفته است.

روش ارائه شده، با دستیابی به دقت طبقه‌بندی ٪۹٤ در محتویات انگلیسی و ٪۹۱ برای محتویات فارسی‌زبان، عملکرد بهتری از روش‌های جایگزین داشته است. همچنین میانگین سرعت طبقه‌بندی برای هر پاراگراف انگلیسی و فارسی، به ترتیب ۰.۳ و ۰.۲ میلی‌ثانیه بوده است که نشان از کارایی فوق‌العاده‌ی این روش دارد.

کلیدواژه‌ها: سازماندهی، سیستم مدیریت دانش، هوش مصنوعی، یادگیری ماشین، یادگیری عمیق، شبکه‌های عصبی مصنوعی، کتابخانه‌ی نرم‌افزار تنسورفلو

فصل اول: کلیات تحقیق

- مقدمه
- بیان مسئله
- اهمیت و ضرورت انجام تحقیق
- جنبه‌ی جدید بودن و نوآوری در تحقیق
- اهداف تحقیق
- سؤالات تحقیق
- فرضیه‌های تحقیق
- ساختار تحقیق

۱-۱- مقدمه

در دنیای پر تکاپوی امروز، دانش سازمانی، یکی از کلیدی‌ترین سرمایه‌های سازمان برای بقا و دستیابی به اهداف است. در سیستم مدیریت دانش، با ذخیره‌سازی مناسب اطلاعات و محتویات ورودی، دسترسی به این منابع ارزشمند امکان‌پذیر می‌شود.

سازماندهی محتوا در سیستم مدیریت دانش، منجر به سهولت دستیابی به دانش سازمانی شده و به این ترتیب مزیت‌های بی‌شماری را برای آن ایجاد می‌کند. درنتیجه، به کارگیری از فناوری‌های نوین در بهبود فرایند سازماندهی محتویات ورودی به سیستم مدیریت دانش اهمیت بالایی دارد.

هوش مصنوعی مفهومی است که پنجره‌ای جدید به روی انسان برای دستیابی هر چه بهتر به اهدافش باز کرده است. یکی از زیرشاخه‌های مهم هوش مصنوعی، مقوله‌ی یادگیری ماشین است. یادگیری عمیق، نوع پیشرفته‌ای از یادگیری ماشین بوده که در آن کشف ارتباطات پیچیده میان داده‌ها مدنظر است. شبکه‌های عصبی مصنوعی، رویکردی نسبتا جدید در زمینه‌ی یادگیری عمیق است که پتانسیل زیادی برای سازماندهی کارا و اثربخش‌تر محتویات، ایجاد کرده است.

تلفیق شبکه‌های عصبی با الگوریتم‌های پردازش زبان طبیعی در فرایند سازماندهی محتویات، امکان دستیابی به نتایجی بسیار دقیق را فراهم می‌آورد. دستیابی به این میزان دقت، با بهره‌گیری از الگوریتم‌های سازماندهی سنتی غیر ممکن است.

هوش مصنوعی، هرچند پتانسیل زیادی را برای انسان در نیل به اهدافش ایجاد کرده است، اما به کارگیری مناسب آن پیچیدگی‌های فراوانی دارد. یکی از چالش‌های اساسی در به کارگیری مناسب هوش مصنوعی، دستیابی به نتیجه بهتر بدون احتیاج به صرف منابع سخت افزاری زیاد می‌باشد. باید در نظر داشت علاوه بر دقت (اثربخشی)، کارایی نیز متغیری مهم، در مسیر دستیابی به هدف است. به عبارت دیگر نقطه‌ی ایده‌آل، نقطه‌ای است که با کمترین میزان صرف منابع، بهترین نتایج حاصل شود.

هدف در این تحقیق ارائه‌ی روشی کاربردی و باکیفیت است که محتویات را در فرایند مدیریت دانش، مبتنی بر شبکه‌های عصبی و الگوریتم‌های پردازش زبان طبیعی، سازماندهی کند. اثربخشی و کارایی، متغیرهای اصلی

مدنظر در دستیابی به هدف هستند. لذا یکی از چالش‌های در این تحقیق، بهره بردن از روش‌های مناسب برای افزایش مقدار این متغیرها است.

۱-۲- بیان مسئله

فرایند سازماندهی محتوا، شامل دسته‌بندی یا طبقه‌بندی محتویات به دسته‌های مختلف (از پیش تعریف شده)، با اتکا بر ارتباط و شباهت‌های موجود در ویژگی‌ها یا صفات ذاتی عناصر سازنده‌ی آن‌ها است (سرکار[۱] ۲۰۱۹، ۲۷۵-۲۷۷).

در دنیای پر تغییر امروز، اطلاعات، منبعی مهم برای تمامی سازمان‌ها به حساب می‌آید (بینون-دیویس[۲] ۲۰۲۰، ۲). سازماندهی محتوا دستیابی و تجزیه و تحلیل محتویات را تسهیل کرده و بهبود می‌بخشد. این کار مزیت‌های فراوانی را برای سازمان‌ها به همراه خواهد داشت. از جمله این مزیت‌ها، کمک به بهبود تصمیمات مدیریتی با فراهم آوردن امکان دسترسی به اطلاعات ساختارمند، افزایش امنیت اطلاعات با ایجاد این قابلیت که فقط اطلاعات با اهمیت رمزگذاری شوند و صرفه‌جویی در منابع سخت‌افزاری درعین افزایش سرعت با ایجاد بهبود در عمل نمایه‌سازی هستند.

علاوه بر آنچه ذکر شد، سازماندهی محتویات، منجر به ذخیره‌ی مناسب‌تر آن‌ها در قسمت‌های مورد انتظار می‌شود. این عمل در شرایط بحران، بهبود به‌سزایی در دستیابی به اطلاعات حیاتی ایجاد کرده و به این ترتیب مدیریت ریسک را بهبود می‌بخشد.

از طرف دیگر، یادگیری عمیق شاخه‌ای نسبتا جدیدی از هوش مصنوعی بوده که در تلاش برای نزدیک‌تر کردن یادگیری ماشین به هدف اصلی آن یعنی «هوش مصنوعی» است (دی[۳]، بهاردواج[۴] و وی[۵] ۲۰۱۸، ۸). شبکه‌هایی که مبتنی بر یادگیری عمیق آموزش داده می‌شوند شبکه‌های عصبی مصنوعی (یا به طور ساده‌تر شبکه‌های مصنوعی) نامیده می‌شوند (راسل و نورویگ ۲۰۲۱، ۸۰۱). این شبکه‌ها، همانطور که از اسمشان معلوم است، مکانیزم یادگیری در موجودات زیستی را که مبتنی بر اعصاب است، شبیه‌سازی می‌کنند (آگاروال[۶] ۲۰۱۸، ۱).

---

[۱] Sarkar
[۲] Beynon-Davies
[۳] Di
[۴] Bhardwaj
[۵] Wei
[۶] Aggarwal

برخلاف روش‌های سنتی یادگیری ماشین که عمل یادگیری در آن‌ها سطحی بوده و متکی به برخی پیش‌فرض‌ها و دانش‌های قبلی از نوع ورودی‌ها می‌باشند، در مفهوم یادگیری عمیق با بهره‌گیری از شبکه‌های عصبی مصنوعی، عمل یادگیری بسیار عمیق‌تر انجام می‌شود (دی، بهاردواج و وی ۲۰۱۸، ١٦). درنتیجه‌ی این امر دقت و سرعت شناسایی الگوها و پیش‌بینی نتایج افزایش محسوسی پیدا می‌کند. در عین حال شبکه‌های عصبی برای استفاده در زمینه‌های مختلف قابل استفاده هستند (دی، بهاردواج و وی ۲۰۱۸، ١٦–١٨). از آنجا که شبکه‌های عصبی حاوی ویژگی‌هایی از جمله رویکرد دنباله‌ای، انعطاف‌پذیری در شرایط نقص داده، قابلیت یادگیری بدون نظارت و یادگیری در سطوح مختلف هستند؛ در زمینه‌ی پردازش زبان طبیعی (که اساس کار برای سازماندهی محتوا است)، قابلیت‌های منحصر به فردی ایجاد می‌کنند (دی، بهاردواج و وی ۲۰۱۸، ۱۲۰).

مفهوم یادگیری عمیق فرصت‌های زیادی را در زمینه‌ی سازماندهی محتوا ایجاد کرده است. با این وجود به دلیل نوظهور بودن این مفهوم که به سرعت در حال پیشرفت است، همچنان فقدان بهره‌گیری از تکنولوژی‌های جدید این عرصه در فرایند مدیریت دانش سازمانی در تحقیقات فعلی و سازمان‌ها مشاهده می‌شود.

نظرسنجی‌ای که توسط شرکت مشاوره و تحقیقات فناوری گارتنر[۱] منتشر شده نشان داد که تنها ٪۳۷ سازمان‌های مورد مطالعه شکلی از هوش مصنوعی را به کار گرفته‌اند (کاستلو[۲]، ۲۰۱۹). غفلت از به کارگیری روشی مناسب برای سازماندهی محتویات ورودی به سیستم مدیریت دانش، مدیریت و دسترسی اطلاعات ذخیره شده را برای سازمان‌ها بسیار دشوارتر می‌کند. درنتیجه، نه تنها هزینه‌ی نگهداری اطلاعات افزایش یافته بلکه با افزایش حجم اطلاعات، که امری اجتناب ناپذیر است، دسترسی کارمندان به این سرمایه‌های ارزشمند مشکل و در برخی موارد فراخوانی اطلاعات مدنظر ناممکن می‌شود. کاستی‌ای که با توجه به نتایج نظرسنجی قید شده در میان انبوهی از سازمان‌ها مشاهده می‌شود، آن‌ها را از دستیابی به فواید بی‌شمار استفاده از تکنولوژی‌های مبتنی بر هوش مصنوعی و به‌خصوص مزیت‌های ایجاد شده در عرصه‌ی سازماندهی اطلاعات بی‌بهره گذاشته و توانایی این شرکت‌ها را برای ادامه‌ی بقا در محیط پرتلاطم فعلی کاهش می‌دهد.

با توجه به مطالب بیان شده، منظور اصلی این تحقیق بهره‌گیری از مفهوم یادگیری عمیق (مشخصا شبکه‌های عصبی مصنوعی) و پردازش زبان طبیعی در راستای ارائه‌ی روشی برای سازماندهی محتوایات ورودی به سیستم مدیریت دانش سازمانی است.

اثربخشی و **کارایی** روش ارائه شده، متغیرهای اصلی در این تحقیق هستند. میزان اثربخشی و کارایی روش ارائه شده در سازماندهی محتوا و رویکردهایی که بهره‌گیری از آن‌ها این میزان را بهبود می‌بخشند، جنبه‌های

---

[۱] Gartner, Inc

[۲] Costello

مجهول و مبهم این تحقیق هستند. برای افزایش مقدار این متغیرها از تکنولوژی‌های جدید عرصه‌ی یادگیری عمیق استفاده و رویکردهای نوآورانه‌ای ارائه خواهد شد.

۱-۳- اهمیت و ضرورت انجام تحقیق

اطلاعات منبعی مهم در همه‌ی سازمان‌ها هستند (بینون-دیویس ۲۰۲۰، ۲). درنتیجه عملکرد مناسب سیستم‌های مدیریت دانش در سازمان‌ها ضروری است. علی‌رغم اهمیت موضوع، جستجوهای محقق برای یافتن تحقیقات انجام شده در زمینه‌ی بهره‌گیری از شبکه‌های عصبی در سیستم‌های مدیریت دانش، حاکی از کمبود تحقیقات انجام شده در این زمینه است.

از طرفی دیگر، در مقاله‌ی مروری‌ای که کوچ[۱]، کورت[۲] و آکبییک[۳] (۲۰۱۹، ۱) با عنوان «خلاصه‌ای از حوزه مدیریت دانش: تاریخچه ده ساله مجله مدیریت دانش» انجام داده‌اند، بررسی ۶۳۷ مقاله‌ی منتشر شده نشان داد که «مطالعه موردی» یکی از متداول‌ترین کلمات کلیدی در بین کلیدواژه‌های مقالات مورد بررسی است. محققان این مقاله نتیجه گرفتند که لازم است مقالات بیشتری به بررسی مدیریت دانش به عنوان یک زمینه‌ی علمی (نه یک مد زودگذر)، بپردازند؛ چرا که مدیریت دانش هنوز زمینه‌ی زیادی برای پیشرفت دارد.

روش ارائه شده در این تحقیق مبتنی بر نسخه‌ی دوم کتابخانه‌ی تنسورفلو است که تکنولوژی‌های نوین و قدرتمندی در عرصه‌ی یادگیری عمیق و پردازش زبان در آن موجود می‌باشند. لذا روشی که ارائه خواهد شد می‌تواند بهبود قابل توجهی در فرایند مدیریت دانش سازمانی ایجاد کرده و به این ترتیب باعث صرفه‌جویی شگرفی در مصرف منابع سازمانی شود.

از جمله مزایای روش ارائه شده بررسی جملات به صورت کلمه به کلمه و کلمه در جمله است که علاوه بر افزایش سرعت و دقت شبکه‌ی عصبی در انجام کار سازماندهی، امکان استفاده از آن در شرایط متفاوت را فراهم می‌آورد. همچنین اتکای روش ارائه شده بر رویکردهای جدیدی است که عمل الگویابی کلمات در جمله‌ها را بهبود بخشیده و منجر به افزایش دقت بالقوه‌ی روش ارائه شده، در دستیابی به اهدافش می‌شوند.

درنتیجه‌ی مطالب ذکر شده اهمیت اصلی روش ارائه شده نسبت به سایر روش‌ها، فراهم آوردن دقت و سرعت بیشتر در فرایند سازماندهی محتویات است. به طوری که در شرایط عدم وجود روش مدنظر، نه تنها

---

[۱] Koç
[۲] Kurt
[۳] Akbıyık

دقت سازماندهی محتوا کمتر بوده؛ بلکه کار سازماندهی محتوا سخت‌تر و نیازمند به منابع سخت‌افزاری قوی‌تر است.

همچنین، بهبود حاصله از به‌کارگیری روش ارائه شده، منجر به توانمندسازی سازمان‌ها در تجزیه و تحلیل بهتر کلان‌داده‌ها می‌شود. از آنجا که تعامل مناسب کلان‌داده‌ها و مدیریت دانش سازمانی، می‌تواند باعث اثربخش و کاراتر شدن فرایند اجرای پروژه‌ها شود (اکامبارم و دیگران ۲۰۱۸، ۱)، درنتیجه فراهم آوردن قابلیت‌های بیشتر در این زمینه بسیار پر اهمیت می‌باشد.

در ضمن برخلاف تحقیقات پیشین، بخش مهمی از روش ارائه شده به صورت متن‌باز در اختیار عموم قرار خواهد گرفت که قابلیت استفاده از نتایج این تحقیق را به حداکثر میزان ممکن، فراهم خواهد کرد.

۱-٤- جنبه‌ی جدید بودن و نوآوری در تحقیق

به طور کلی می‌توان، جنبه‌های جدید این تحقیق را در سه مورد بیان کرد:

۱. یکی از مهم‌ترین جنبه‌های جدید مورد بررسی، هدف اصلی تحقیق یعنی سازماندهی محتوا در فرایند مدیریت دانش سازمانی است. علی‌رغم فرصت‌هایی که یادگیری عمیق در زمینه‌ی مدیریت دانش سازمانی فراهم کرده، اما در تحقیقات پیشین کمتر به این موضوع پرداخته شده است.

۲. رویکردها و فناوری‌های مدنظر برای استفاده در این تحقیق، فناوری‌هایی هستند که با توجه به جدید بودنشان کمتر به آن‌ها پرداخته شده است. روشی که ارائه خواهد شد، مبتنی بر نسخه‌ی دوم کتابخانه‌ی تنسورفلو که تکنیک‌های بسیار جدید و قدرتمندی هستند، می‌باشد. بهره‌گیری از این تکنیک‌ها در رسیدن هرچه بیشتر به هدف مدنظر کمک به‌سزایی خواهد کرد.

۳. برخلاف تحقیق‌های قبلی، بخش مهمی از روشی مدنظر در این تحقیق، به صورت متن‌باز ارائه خواهد شد. لذا امکان استفاده، مشاهده‌ی جزئیات روش و انجام آزمایشات مختلف روی آن برای عموم فراهم خواهد شد. درنتیجه‌ی این امر، نتایج حاصله از این تحقیق در موارد زیادی کاربرد خواهند داشت.

۱-۵- اهداف تحقیق

منظور از انجام این تحقیق، ارائه روشی اثربخش، کارا و باکیفیت است که به سازمان‌ها در جهت بهبود سازماندهی محتویات در فرایند مدیریت دانش سازمانی کمک رساند. همچنین، در این تحقیق، مفهوم شبکه‌های مصنوعی بررسی شده است تا مخاطب دید کلی از این مفهوم پرکاربرد پیدا کند.

۱-۵-۱- هدف اصلی

۱. ارائه‌ی روشی کاربردی و باکیفیت، مبتنی بر شبکه‌های عصبی برای سازماندهی محتوا در فرایند مدیریت دانش سازمانی:

در این تحقیق تلاش بر این است که روش ارائه شده، قابلیت اجرایی شدن در سازمان‌ها را داشته باشد یا به آن‌ها در فرایند توسعه‌ی سیستمی مناسب برای سازماندهی محتویات در فرایند مدیریت دانش کمک رساند. در ضمن، در راستای ارائه‌ی روشی باکیفیت و انعطاف‌پذیر کوشش خواهد شد تا در شرایط مختلف، پاسخگوی نیاز سازمان‌ها باشد.

۱-۵-۲- اهداف فرعی

۲. افزایش اثربخشی در سازماندهی محتوا در فرآیند مدیریت دانش سازمانی مبتنی بر شبکه‌های عصبی:

دقت روش ارائه شده در سازماندهی صحیح محتویات، اهمیت بالایی دارد. لذا یکی از عوامل مؤثر مدنظر در ارائه‌ی این روش، دستیابی به دقتی قابل قبول است.

۳. افزایش کارایی در سازماندهی محتوا در فرآیند مدیریت دانش سازمانی مبتنی بر شبکه‌های عصبی:

میزان منابع سخت‌افزاری مورد نیاز برای اجرای روش مدنظر، دیگر عامل مهم است. تلاش بر این است که از رویکردها و الگوریتم کارآمد، برای دستیابی به بازدهی مناسب، استفاده شود.

٤. دستیابی به روشی مبتنی بر تکنولوژی‌های نوین که نسبت به رویکردهای سنتی، سازماندهی محتوا را دقیق و بهینه‌تر انجام دهند:

ما به دنبال ارائه‌ی روشی بهینه‌تر از روش‌های سنتی هستیم. لذا مدل ارائه شده مبتنی بر جدیدترین فناوری‌های حوزه‌ی یادگیری عمیق خواهد بود. همچنین روش‌های سنتی نیز به طور اجمالی بررسی خواهند شد تا از مقایسه آن‌ها با روش ارائه شده، میزان کامیابی در دستیابی به این هدف، مشخص شود.

۱-٦- سؤالات تحقیق

مبتنی بر اهداف اشاره شده، سؤالات اصلی و فرعی این تحقیق به شرح زیر است:

۱-٦-۱- سؤال اصلی

۱. چگونه می‌توان با بهره‌گیری از تکنیک‌های یادگیری عمیق ماشین، روشی کاربردی و باکیفیت در راستای سازماندهی محتوا در فرایند مدیریت دانش سازمانی ارائه داد؟

۱-٦-۲- سؤالات فرعی

۲. چه روش‌هایی برای افزایش اثربخشی در سازماندهی محتوا در فرآیند مدیریت دانش سازمانی مبتنی بر شبکه‌های عصبی وجود دارد؟

۳. چه روش‌هایی برای افزایش کارایی در سازماندهی محتوا در فرآیند مدیریت دانش سازمانی مبتنی بر شبکه‌های عصبی وجود دارد؟

٤. چگونه می‌توان با بهره بردن از فناوری‌های نوین، سازماندهی محتوا را نسبت به رویکردهای سنتی بهینه‌تر کرد؟

۱-۷- فرضیه‌های تحقیق

فرضیه‌های این تحقیق در دو دسته‌ی اصلی و فرعی، به شرح زیر هستند:

۱-۷-۱- فرضیه‌ی اصلی

۱. با بهره‌گیری از تکنیک‌های یادگیری عمیق ماشین، می‌توان روشی کاربردی و باکیفیت در راستای سازماندهی محتوا در فرایند مدیریت دانش سازمانی ارائه داد.

۱-۷-۲- فرضیه‌های فرعی

۲. میزان اثربخشی (دقت) مدل ارائه شده در طبقه‌بندی داده‌ها، بیش از ٪۹۰ خواهد بود.

۳. زمان مورد نیاز برای بررسی و سازماندهی یک پاراگراف توسط مدل ارائه شده، با داشتن ۱۰ هزار رکورد در پایگاه داده کمتر از ۰.۱ ثانیه خواهد بود (کارایی مدل قابل قبول خواهد بود).

٤. روش حاصل از به کارگیری فناوری‌های نوین، نسبت به رویکردهای سنتی مطرح، سازماندهی محتوا را دقیق‌تر انجام می‌دهد.

۱-۸- ساختار تحقیق

این تحقیق در پنج فصل تدوین شده است:

فصل اول به بررسی کلیات تحقیق اختصاص یافته است. در این فصل، نخست مسئله و ابعاد آن به طور کامل تبیین شده است. سپس به اهمیت انجام تحقیق پرداخته شده و جنبه‌های نوآورانه‌ی این تحقیق بررسی شده‌اند. در ادامه اهداف، سؤالات و فرضیات این تحقیق در دو دسته‌ی اصلی و فرعی آورده شده‌اند.

فصل دوم به مرور ادبی و بررسی پیشینه‌ی تحقیق اختصاص پیدا کرده است. در این فصل، ابتدا به مبانی نظری تحقیق پرداخته شده تا مخاطب را با مفاهیم اساسی تحقیق آشنا کنیم. سپس تحقیقات جدید پیرامون موضوع بررسی شده‌اند. هدف در فصل دوم، ایجاد درکی کلی درباره‌ی مفاهیم مدیریت دانش سازمانی، هوش مصنوعی، یادگیری ماشین، یادگیری عمیق و شبکه‌های عصبی مصنوعی بوده است. در ضمن، با بررسی مقالات به‌روز، کاربرد شبکه‌های عصبی مصنوعی را در مقوله‌ی سازماندهی و طبقه‌بندی تبیین شده است.

در فصل سوم، به بررسی روش اجرای تحقیق پرداخته شده است. در این فصل اقدامات انجام شده‌ی در راستای دستیابی به یک مدل کاربردی، باکیفیت و مقیاس پذیر، ذکر شده است. روش اجرای این تحقیق مبتنی بر هفت فاز متدلوژی علم داده انجام شده است که در این فصل مرحله به مرحله توضیح داده شده است. در این فصل ابتدا داده‌های ورودی آماده شده‌اند و سپس نه مدل یادگیری ماشین پیاده‌سازی می‌شوند. در ادامه روی مجموعه داده‌های آموزش، عمل یادگیری مدل‌ها انجام و آن‌ها برای مرحله‌ی آزمایش آماده می‌شوند. همچنین معیارهای اندازه‌گیری عملکرد مدل‌های پیاده‌سازی شده، در این فصل معرفی خواهند شد.

در فصل چهارم، ابتدا داده‌های ورودی بررسی خواهند شد. سپس با اتکا بر معیارهای معرفی شده در فصل سوم، یافته‌های حاصل از اندازه‌گیری عملکرد مدل‌های معرفی شده بیان شده است. در این فصل عملکرد مدل‌ها به طور کلی و به تفکیک هر کلاس، مورد بررسی قرار می‌گیرند. بررسی انجام شده در خصوص مدل‌های جایگزین اجمالی و در خصوص مدل اصلی مفصل خواهد بود.

در فصل پنجم، ابتدا در نتایج حاصله بحث خواهد شد. سپس از اقدامات انجام شده و مطالب بیان شده، نتیجه‌گیری می‌شود. در نهایت، به بیان محدودیت‌های تحقیق و پیشنهادهای کاربردی و پژوهشی پرداخته خواهد شد.

فصل دوم: مروری بر ادبیات تحقیق و پیشینه تحقیق

- مقدمه
- تعاریف، اصول و مبانی نظری
- بررسی مقالات اخیر مرتبط با موضوع

۲-۱- مقدمه

در این فصل ابتدا به تعاریف، اصول و مبانی نظری، بیان خواهد شد و سپس مقالات اخیر مرتبط با موضوع را در دو بخش فارسی و لاتین، بررسی می‌گردد. هدف اصلی در این فصل، آشناسازی مخاطب با مفاهیم اساسی تحقیق است تا زمینه‌های نظری لازم برای درک موضوع و اقدامات انجام شده‌ی آتی، در ایشان ایجاد شود.

در قسمت مبانی نظری، در وهله‌ی اول، دانش و مدیریت آن معرفی شده‌اند. سپس مفهوم سازماندهی در زمینه‌ی مدیریت و علوم داده به طور مجزا بررسی شده است. همچنین، مفاهیم رگرسیون، طبقه‌بندی و خوشه‌بندی به عنوان مفاهیم کلیدی سازماندهی مبتنی بر علوم داده، توضیح داده شده‌اند.

در ادامه مفاهیم هوش مصنوعی، پردازش زبان طبیعی و یادگیری ماشین، بیان شده‌اند. در قسمت یادگیری ماشین، انواع آن که شامل یادگیری نظارت شده، یادگیری بدون نظارت و یادگیری تقویتی است، مورد بررسی قرار گرفته است. در ادامه مفاهیم یادگیری عمیق و شبکه‌های عصبی مصنوعی تبیین شده‌اند. در این فصل، مجموعاً ده مدل یادگیری ماشین که سه مورد از آن در حیطه‌ی یادگیری عمیق قرار می‌گیرند، معرفی شده است.

لازم به ذکر است که مفاهیم نظری موضوع بسیار گسترده می‌باشند و باز شدن موضوعات به حدی انجام شده تا در حیطه‌ی مدنظر در این تحقیق، بگنجد. توصیه می‌شود برای کسب اطلاعات بیشتر به منابع قید شده، که اکثراً کتب مرجع معتبر می‌باشند، ارجاع شود.

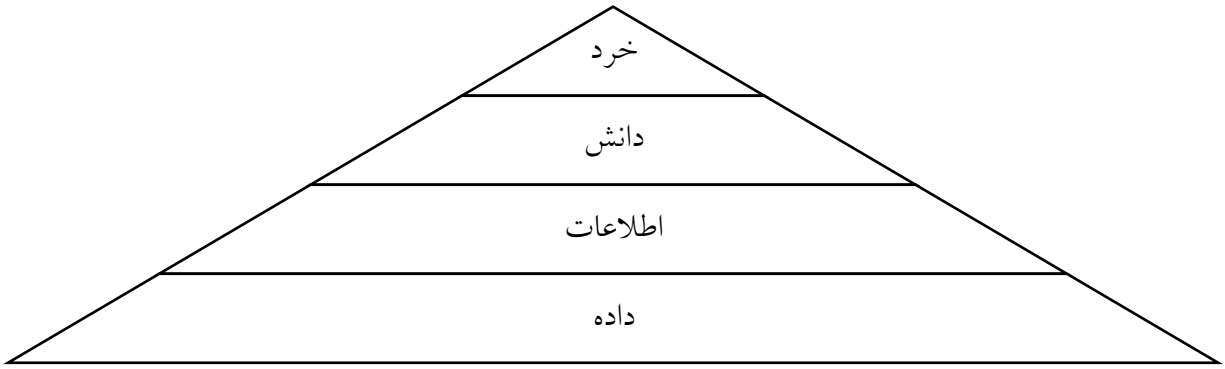
۲-۲- تعاریف، اصول و مبانی نظری

در این بخش از فصل دوم، کلیتی راجع به مفاهیمی اساسی تحقیق بیان شده است. با تبیین تعاریف مفهومی و عملیاتی تلاش شده است تا در مخاطب درکی کلی از این مفاهیم ایجاد شود.

۲-۲-۱- دانش[۱]

دانش در لغت، به معنای درک یا اطلاعات درباره‌ی موضوعی است که از طریق تجربه یا مطالعه، توسط یک فرد یا افراد، بدست می‌آید (فرهنگ لغت الکترونیکی کمبریج، واژه‌ی دانش).

طبق بیان حسین[۲] و ارمینه[۳] (۲۰۲۱، ۳)، دانش در عمل، استفاده از اطلاعات ساخته شده بر بستر داده‌ها، برای رسیدن به هدف است. دانش از طریق تجربه و تخصص کارکنان سازمان بدست می‌آید. در هرم فرایند تبدیل داده به خرد[٤]، دانش مرحله سوم است. این هرم در شکل ۲–۱ نمایش داده شده است.



(شکل ۲–۱) هرم تبدیل داده به خرد، موسوم به هرم DIKW (حسین و ارمینه ۲۰۲۱، ۳).

۲-۲-۱-۱- مدیریت دانش[۵]

فرآیندی است که از طریق آن می‌توان دانش سازمانی را به شکل ارزش افزوده ایجاد کرد و در اختیار کسانی قرار داد که در داخل سازمان برای دستیابی بهتر به اهداف سازمانی به آن نیاز دارند (حسین و ارمینه ۲۰۲۱، ۱۳). دانشی که ثبت می‌شود و به شکل سند (دانش صریح) در می‌آید، تنها درصد کمی (حدود ٪۳۰) از کل دانش موجود در سازمان را تشکیل می‌دهد؛ در حالی که بخش باقیمانده در ذهن کارکنان و کارشناسان سازمان (دانش ضمنی) جا می‌گیرد. مدیریت دانش، فرآیندهای مختلف دانش شامل کسب، ایجاد، سازماندهی، کاربرد و استفاده مجدد از دانش را ترویج می‌کند (حسین و ارمینه ۲۰۲۱، ۱۳).

---

[۱] Knowledge
[۲] Husain
[۳] Ermine
[٤] Wisdom
[۵] Knowledge Management

۲-۲-۱-۱-۱- سیستم مدیریت دانش

به بیان (حسین و ارمینه ۲۰۲۱، ۱۳)، سیستم مدیریت دانش هر نوع سیستم مبتنی بر فناوری اطلاعات است که دانش را برای بهبود درک، همکاری و همسویی فرآیندها، ذخیره و بازیابی می‌کند. سیستم‌های مدیریت دانش علاوه بر استفاده‌ی داخلی در سازمان‌ها یا تیم‌ها، می‌توانند در جهت متمرکز کردن پایگاه دانش سازمان برای کاربران یا مشتریانشان نیز استفاده شوند.

۲-۲-۲- سازماندهی در مدیریت[۱]

سازماندهی یکی از اصلی‌ترین مفاهیمی است که در این تحقیق به آن پرداخته شده است. ابتدا تعریفی مفهومی و عملیاتی از این مفهوم کلیدی ارائه می‌کنیم:

سازماندهی، در لغت، به معنای ترتیبی دادن برای اتفاق افتادن نتیجه‌ی مدنظر است (فرهنگ لغت الکترونیکی کمبریج، واژه‌ی سازماندهی).

پنج وظیفه‌ی اصلی مدیران شامل برنامه‌ریزی، سازماندهی، به‌کارگماردن، رهبری و کنترل می‌شود. سازماندهی، هماهنگی منابع سازمانی با یکدیگر و چینش مناسب آن‌ها، برای دستیابی به هدف است. هنگامی که مدیران اهدافی را تعیین کرده و برای دستیابی به آنها برنامه‌ریزی کردند، باید منابع سازمانی را به‌گونه‌ای طراحی و توسعه دهند که امکان اجرای موفقیت‌آمیز برنامه‌های تعیین شده را فراهم آورد (قوش[۲] ۲۰۲۱، ۲).

سازماندهی در مدیریت، شامل توسعه ساختار سازمانی و تخصیص منابع انسانی برای اطمینان از تحقق اهداف است. ساختار سازمان، چارچوبی را برای هماهنگ‌سازی اقدامات افراد فراهم می‌کند. این ساختار معمولاً با یک نمودار سازمانی نشان داده می‌شود که یک نمایش گرافیکی از زنجیره فرماندهی در یک سازمان ارائه می‌کند. تصمیمات اتخاذ شده در مورد ساختار یک سازمان عموماً به عنوان تصمیمات «طراحی سازمانی»[۳] شناخته می‌شوند. سازماندهی همچنین شامل طراحی مشاغل فردی در سازمان است. تصمیماتی که در مورد ماهیت مشاغل در سازمان گرفته می‌شود، عموماً تصمیمات «طراحی شغل»[٤] نامیده می‌شوند (اصول مدیریت[٥] ۲۰۱۵، ۲۰).

---

[۱] Organizing
[۲] Ghosh
[۳] Organizational Design
[٤] Job Design
[٥] Principles of Management

۲-۲-۲-۱- اهمیت و ضرورت سازماندهی برای سازمان

سازماندهی، همانند سایر وظایف اصلی مدیریت (برنامه‌ریزی، به‌کارگماردن، رهبری و کنترل) برای دستیابی به اهداف سازمان ضروری است. سازماندهی منجر به تخصیص بهینه‌ی منابع شده و از اتلاف آن‌ها جلوگیری می‌کند. برای مثال، طراحی سازمانی، به عنوان یک اقدام در چارچوب سازماندهی، فرآیندی رسمی و هدایت‌شده برای یکپارچه‌سازی افراد، اطلاعات و فناوری یک سازمان است (اصول مدیریت[۱] ۲۰۱۵، ۱٤۸).

یکپارچه‌سازی اطلاعات، به عنوان یکی از دست‌آوردهای سازماندهی مناسب، مزیت‌های فراوانی برای سازمان ایجاد می‌کند. به عنوان مثال، یکپارچه‌سازی و سازماندهی مناسب اطلاعات، در مدیریت مناسب شبکه‌های اجتماعی، که به عنوان «سازمان نامرئی» شناخته می‌شوند، مؤثر است (اصول مدیریت ۲۰۱۵، ۱۵۰).

درنتیجه، روشی که در این تحقیق ارائه خواهد شد، با سازماندهی اطلاعات، زمینه‌ی لازم برای یکپارچه‌سازی آن‌ها را فراهم کرده و مزیت‌های بی‌شماری را برای سازمان به ارمغان می‌آورد.

۲-۲-۳- سازماندهی در علوم داده

علم داده، علم تجزیه و تحلیل سیستماتیک داده‌ها در یک کار چارچوب علمی است (لاروس[۲] و لاروس ۲۰۱۹، ۱). سازماندهی که یکی از اقدامات پرکاربرد در علوم داده است، یا به صورت طبقه‌بندی و یا در قالب خوشه‌بندی انجام می‌شود (لاروس و لاروس ۲۰۱۹، ۵). در ادامه، ابتدا به بررسی اجمالی مفهوم رگرسیون پرداخته خواهد شد و سپس مفاهیم طبقه‌بندی و خوشه‌بندی را معرفی می‌شوند.

۲-۲-۳-۱- رگرسیون

رگرسیون یک تکنیک آماری است که یک متغیر وابسته را به یک یا چند متغیر مستقل مرتبط می‌کند. یک مدل رگرسیون به تجزیه و تحلیل همبستگی بین متغیرها می‌پردازد (جرون[۳] ۲۰۱۹، ٦۵). به بیان بروس[٤]، بروس و گدک[۵] (۲۰۲۰، ۱٤۱) شاید متداول‌ترین هدف در آمار، پاسخ به این سوال باشد که «آیا متغیر(های) مستقل به وابسته مرتبط هستند؟»؛ اگر چنین است، «چه رابطه‌ای میان آن‌ها وجود دارد که به کمک آن بتوان متغیر وابسته

---

[۱] Principles of Management
[۲] Larose
[۳] Géron
[٤] Bruce
[۵] Gedeck

را پیش‌بینی کرد؟». هیچ‌جا پیوند بین آمار و علم داده قوی‌تر از حوزه‌ی پیش‌بینی، به‌ویژه، پیش‌بینی متغیر نتیجه (هدف) بر اساس مقادیر سایر متغیرهای «پیش‌بینی‌کننده»[۱]، نیست.

یکی از ارتباطات مهم بین علم داده و آمار در حوزه‌ی تشخیص ناهنجاری است. در این حوزه، تشخیص‌های مبتنی بر رگرسیون[۲]، که در اصل برای تجزیه و تحلیل داده‌ها و بهبود مدل رگرسیون در نظر گرفته شده است، می‌تواند برای شناسایی رکوردهای غیر معمول استفاده شود (بروس، بروس و گدک ۲۰۲۰، ۱٤۱).

تخمین ارزش گذاری قیمت خانه بر اساس ویژگی‌های خانه، مانند مساحت، تعداد اتاق خواب، موقعیت مکانی و... یک مثال معروف از رگرسیون است. مسئله رگرسیون با یک الگوریتم یادگیری رگرسیون حل می‌شود که مجموعه‌ای از نمونه‌های برچسب گذاری شده را به عنوان ورودی گرفته و مدلی تولید می‌کند که می‌تواند برای نمونه‌های بدون برچسب، خروجی هدفی را پیش‌بینی کند (بورکوف[۳] ۲۰۱۹، ۱۲). درنتیجه رگرسیون یک فرایند «نظارت شده» است. در بخش «یادگیری ماشین» انواع یادگیری‌ها از جمله یادگیری نظارت شده معرفی شده‌اند.

۲-۲-۳-۲- طبقه‌بندی[٤]

طبقه‌بندی، فرایندی است که در آن، ورودی‌ها در کلاس‌های مختلف، براساس برچسب‌های متناظرشان، قرار داده می‌شوند (راسل و نورویگ ۲۰۲۱، ۸۷٤). مدل رگرسیون، فرض می‌کند که متغیر پاسخ، کمی است. اما در بسیاری از موقعیت ها، متغیر پاسخ کیفی است. در طبقه‌بندی رویکردهایی برای پیش‌بینی پاسخ‌های کیفی، اتخاذ می‌شوند (جیمز[٥] و دیگران ۲۰۲۱، ۱۲۹). به بیان راسل و نورویگ (۲۰۲۱، ٦۷۰)، در فرایند طبقه‌بندی، خروجی به شکل یک مجموعه‌ی محدود از مقادیر (مثلاً خوب، متوسط، بد) است. در حالیکه در رگرسیون خروجی به شکل یک عدد (مثلاً یک عدد صحیح) است. به عبارت دیگر خروجی طبقه‌بندی، کیفی و خروجی رگرسیون، کمی است.

تکنیک‌هایی که از آنها برای پیش‌بینی پاسخ کیفی در فرایند طبقه‌بندی استفاده می‌شود، **طبقه‌بندی‌کننده**[٦] نامیده می‌شوند اغلب طبقه‌بندی‌کننده‌ها، احتمال تعلق مشاهده به هر یک از دسته‌بندی‌های یک متغیر کیفی را

---

[۱] Predictor

[۲] تشخیص‌های رگرسیون (Regression Diagnostics)، مراحل مختلفی هستند که برای ارزیابی میزان تناسب مدل با داده‌ها طی می‌شوند. اکثر این مراحل بر اساس تجزیه و تحلیل مقادیر باقیمانده هستند (بروس، بروس و گدک ۲۰۲۰، ۱۷٦).

[۳] Burkov

[٤] Classification

[٥] James

[٦] Classifier

جداگانه پیش‌بینی می‌کنند که مبنایی برای طبقه‌بندی است. از این نظر آنها نیز مانند روش‌های رگرسیون رفتار می‌کنند (جیمز و دیگران ۲۰۲۱، ۱۲۹). در قسمت «یادگیری ماشین»، طبقه‌بندی‌کننده‌های مدنظر در تحقیق، به طور اجمالی معرفی شده‌اند.

طبقه‌بندی، که اصلی‌ترین رویکرد مورد استفاده در ارائه‌ی روش مدنظر است، در زمینه‌های بسیاری کاربرد دارد. به‌عنوان مثال، فرایند طبقه‌بندی، در گروه‌بندی بیماران با توجه به علائم بیماریشان، دسته‌بندی توالی‌های دی‌ان‌ای‌ها برای پیش‌بینی و پیشگیری از وقوع بیماری‌ها و یا کلاس‌بندی مشتریان با توجه به ویژگی‌های مربوطه، از جمله کاربردهای طبقه‌بندی است.

۲-۲-۳-۳- خوشه‌بندی[۱]

خوشه‌بندی، فرایند دسته‌بندی ورودی‌ها براساس شباهت‌های آن‌ها است. تفاوت خوشه‌بندی و طبقه‌بندی در آن است که فرایند خوشه‌بندی، بدون کمک برچسب‌های کلاس (که از پیش تعریف شده‌اند) انجام می‌گیرد و اصطلاحاً، فرایند بدون نظارت است (راسل و نورویگ ۲۰۲۱، ۸۷٤). فرایند یادگیری بدون نظارت، در قسمت «یادگیری ماشین»، معرفی شده است.

یک خوشه مجموعه‌ای از رکوردها است که شبیه یکدیگر و بی‌شباهت به رکوردهای دیگر خوشه‌ها هستند. همانطور که اشاره شد، در خوشه‌بندی برخلاف طبقه‌بندی، هیچ متغیر هدفی برای گروه‌بندی وجود ندارد. در فرایند خوشه‌بندی سعی نمی‌شود تا مقدار یک متغیر هدف، طبقه‌بندی، تخمین یا پیش‌بینی شود. بلکه، الگوریتم‌های خوشه‌بندی به دنبال تقسیم کل مجموعه داده‌ها به زیرگروه‌ها یا خوشه‌های نسبتاً همگن هستند. نقطه‌ی ایده‌آل جایی است که شباهت رکوردهای درون خوشه به حداکثر رسیده و شباهت به رکوردهای خارج از این خوشه، به حداقل می‌رسد (لاروس و لاروس ۲۰۱۹، ۱٤۱).

به عنوان مثال، در مقاله‌ای که توسط بیابیانی[۲] و دیگران (۲۰۲۰، ۱۵-۱)، با عنوان «طراحی یک متریک فاصله‌ی خبره[۳] برای خوشه‌بندی آب و هوا: موردی از (میزان) بارندگی در آنتیل‌های کوچک» توسعه یافته است، رکوردها در پنج خوشه گروه‌بندی شده‌اند. این خوشه‌ها برچسب یا عنوان از پیش تعریف شده‌ای ندارند. در این تحقیق، از عواملی مانند میزان بارندگی در طول یک زمان خاص، برای گروه‌بندی رکوردها استفاده شده است.

---

[۱] Clustering

[۲] Biabiany

[۳] Expert Distance Metric: متریک فاصله، یک معیار برای اندازه‌گیری میزان شباهت (یا اصطلاحا «فاصله») داده‌ها است (بروس، بروس و گدک ۲۰۲۰، ۲٤۱).

استفاده از مفاهیم ذکر شده بستگی به نوع مسئله و هدف مدنظر دارد (راسل و نورویگ ۲۰۲۱، ۸۷٤). الگوریتم‌های استفاده شده‌ی برای ارائه‌ی روش مدنظر در این تحقیق، مبتنی بر عمل طبقه‌بندی خواهند بود. در همین فصل، ذیل قسمت یادگیری ماشین، رویکردهای رایج طبقه‌بندی در علوم داده معرفی خواهند شد. در این تحقیق، همه‌ی این رویکردها مورد استفاده قرار گرفته و عملکرد آن‌ها تجزیه و تحلیل شده است.

۲-۲-٤- هوش مصنوعی[۱]

برای آشنایی با مفهوم هوش مصنوعی ابتدا باید مفاهیم الگوریتم و عامل درک شوند.

«الگوریتم»[۲]، یک فرایند محاسباتی از پیش تعریف شده است که براساس مقدار یا مجموعه‌ی مقادیر ورودی، مقدار یا مجموعه‌ای از مقادیر را به عنوان خروجی، در مدت زمانی محدود، تولید می‌کند. به عبارت دیگر، یک الگوریتم دنباله‌ای از مراحل محاسباتی است که ورودی را به خروجی تبدیل می‌کند (کورمن[۳] و دیگران ۲۰۲۲، ۵).

«عامل»[٤] هر چیزی است که از طریق حسگرها، محیط خود را درک کرده و روی آن محیط از طریق عملگرها واکنش نشان دهد (راسل و نورویگ ۲۰۲۱، ٥٤).

در کتاب «هوش مصنوعی: یک رویکرد نوین» که یکی از کتاب‌های مرجع این زمینه است، هوش مصنوعی به عنوان مفهومی برای مطالعه‌ی عوامل هوشمند[٥] معرفی شده است. یک عامل هوشمند ایده‌آل، بهترین اقدام ممکن را در یک موقعیت خاص انجام می‌دهد (راسل و نورویگ ۲۰۲۱، ۵۲).

در هوش مصنوعی مقصود، بهره‌گیری از الگوریتم‌های خاص، در راستای ایجاد نوعی هوش در ماشین‌ها، برای تقلید از اعمال انسان است (دی، بهاردواج و وی ۲۰۱۸، ۸). مقصود از ماشین در اینجا عمدتاً سخت‌افزارها می‌باشند. شرکت‌های بزرگ فناوری معمولاً در حوزه‌های پردازش زبان طبیعی، انجام اتوماتیک کارها، یادگیری ماشین و... از هوش مصنوعی بهره می‌گیرند.

---

[۱] Artificial Intelligence
[۲] Algorithm
[۳] Cormen
[٤] Agent
[٥] Intelligent Agents

۲-۲-٤-۱- پردازش زبان طبیعی (انال‌پی)[۱]

به بیان راسل و نورویگ (۲۰۲۱، ۸۷٤)، انال‌پی یکی از زیر شاخه‌های هوش مصنوعی بوده که به معنای برنامه دادن به کامپیوتر برای تجزیه و تحلیل زبان طبیعی است. اهداف اصلی انال‌پی شامل برقراری ارتباط با انسان، یادگیری و پیشبرد درک علمی زبان‌ها است.

لازم به ذکر است که یکی از کاربردهای مدل‌های مبتنی بر یادگیری ماشین، که در ادامه توضیح داده خواهد شد، پردازش زبان طبیعی می‌باشد. بدیهی است که برای ساخت مدل مدنظر، اساس کار ما، پردازش زبان طبیعی خواهد بود.

۲-۲-٤-۱-۱- نمایه‌سازی[۲]

نمایه‌سازی در لغت، عمل تهیه‌کردن نمایه (=فهرستی از عناوین و آدرس مربوط به آن) برای یک کتاب یا مجموعه است (فرهنگ لغت الکترونیکی کمبریج، واژه‌ی نمایه‌سازی).

در علوم داده، نمایه‌سازی به معنای مشخص‌سازی کلیدواژه‌هایی برای بهبود عمل فراخوانی مقادیر خاص می‌باشد (گروس[۳] ۲۰۱۹، ۳٤۳). این عمل به انواع مختلف (شامل توکن کردن، دگرنمایی واژه، بسته‌ی کلمات و...) در زمینه‌ی پردازش زبان طبیعی، انجام می‌گیرد.

۲-۲-۵- یادگیری ماشین[٤]

در نیمه دوم قرن بیستم، یادگیری ماشینی به عنوان زیرشاخه‌ای از هوش مصنوعی تکامل یافت. یادگیری ماشین الگوریتم‌های خودآموزی را دربر می‌گیرد که ما را قادر به استخراج دانش، از داده‌ها می‌سازند. داده‌ها منابعی هستند که به وفور انسان را احاطه کرده‌اند (راشکا[۵]، لیو[٦] و میرجلیلی ۲۰۲۲، ۱). درنتیجه یادگیری ماشین با فراهم آوردن امکان استخراج کارآمدتر دانش از داده‌های بسیار دسترس‌پذیر، می‌تواند به بهبود تصمیم‌گیری‌ها کمک کند. از آنجا که تصمیم‌گیری جایگاه والایی در علم مدیریت دارد، اهمیت بالای مفهوم یادگیری ماشین در این حوزه، غیر قابل انکار است.

---

[۱] Natural Language Processing (NLP)
[۲] Indexing
[۳] Grus
[٤] Machine Learning
[۵] Raschka
[٦] Liu

در لغت، یادگیری ماشین فرآیندی است که رایانه‌ها با یادگیری از داده‌های جدید، روش انجام وظایف خود را تغییر می‌دهند، بدون اینکه به ارائه دستورالعمل‌های انسان در قالب برنامه، احتیاج داشته باشند (فرهنگ لغت الکترونیکی کمبریج، واژه‌ی یادگیری ماشین).

اگر یک عامل پس از انجام مشاهدات در مورد جهان، عملکرد خود را بهبود بخشد، در حال یادگیری است. یادگیری می‌تواند از موارد بی اهمیت، مانند یادداشت کردن لیست خرید، تا عمیق، مانند زمانی که آلبرت انیشتین نظریه جدیدی از جهان را استنباط کرد، متغیر باشد. وقتی عامل یک کامپیوتر باشد، آن را یادگیری ماشینی می‌نامیم. یک کامپیوتر برخی از داده‌ها را مشاهده می‌کند، مدلی را بر اساس داده‌های یادگیری ماشین می‌سازد، و از مدل به عنوان فرضیه‌ای درباره جهان و نرم‌افزاری که می‌تواند مشکلات را حل کند، استفاده می‌کند (راسل و نورویگ ۲۰۲۱، ٦٦٩).

به طور کلی، یادگیری ماشین سه نوع دارد:

۲-۲-۵-۱- یادگیری نظارت شده[۱]

در این یادگیری، عامل، جفت ورودی-خروجی را مشاهده کرده و تابعی را می‌آموزد که از ورودی به خروجی نگاشت می‌کند. برای مثال، ورودی‌ها می‌توانند تصاویر دوربین باشند که هر کدام با یک خروجی مثل «اتوبوس» یا «عابر پیاده» و... همراه هستند. خروجی‌هایی مانند این برچسب نامیده می‌شوند. عامل، تابعی را می‌آموزد که وقتی تصویر جدیدی به آن داده می‌شود، برچسب مناسب را پیش‌بینی کند. پس در یادگیری نظارت شده، همواره «خروجی صحیح» (خروجی مدنظر) مشخص می‌شود و مبتنی بر آن عمل یادگیری انجام می‌گیرد. به عبارت دیگر، محیط معلم است و توسط آن، خروجی مناسب هر ورودی، آشکارا معرفی می‌شود (راسل و نورویگ ۲۰۲۱، ٦٧١).

به عبارت دیگر، در یادگیری نظارت شده، فرآیند آموزش یک مدل بر روی داده‌هایی که نتیجه‌ی آن مشخص است، برای کاربرد بعدی در داده‌هایی که نتیجه مشخص نیست انجام می‌گیرد (بروس، بروس و گدک ۲۰۲۰، ١٤١). رگرسیون و طبقه‌بندی فرایندهایی هستند که در آن‌ها الگوریتم‌ها یادگیری نظارت‌شده به‌کار برده می‌شود؛ چرا که در این دو فرایند، ابتدا داده‌هایی با نتایج یا «برچسب‌های» مشخص وجود دارد که عمل یادگیری روی

[۱] Supervised Learning

آن‌ها انجام می‌شود. مجموعه‌ای که به کمک آن عمل یادگیری انجام می‌شود، «مجموعه داده‌ی آموزش»[۱] و مجموعه‌ای که روی آن عمل آزمون انجام می‌شود، «مجموعه داده‌ی آزمایش»[۲] نام دارد (هوین[۳] ۲۰۲۲، ۸۱).

الگوریتم‌های متعددی از یادگیری ماشین، به صورت نظارت شده انجام می‌گیرند. رایج‌ترین این الگوریتم‌ها به شرح زیر می‌باشد:

۲-۲-۵-۱-۱- الگوریتم کِی-نزدیک‌ترین همسایه (کِی‌ان‌ان)[٤]

کی‌ان‌ان یکی از روش‌های ساده‌تر (نسبت به سایر روش‌ها) که در طبقه‌بندی و پیش‌بینی (در فرایند رگرسیون) کاربرد دارد. است. در این روش هیچ مدلی برای تناسب وجود ندارد. هرچند این بدان معنا نیست که استفاده از کی‌ان‌ان بدون انجام اقدامات جانبی امکان‌پذیر است. نتایج پیش‌بینی به نحوه‌ی مقیاس‌بندی ویژگی‌ها، نحوه‌ی اندازه‌گیری شباهت و میزان بزرگی «کِی»[٥] بستگی دارد. همچنین تمام پیش‌بینی‌کننده‌های مورد استفاده در این روش باید به صورت عددی باشند.

مدل کی‌ان‌ان علاوه بر اینکه یک مدل طبقه‌بندی‌کننده به حساب می‌آید، در حل مسائل رگرسیون نیز قابل استفاده است. الگوریتم کی‌ان‌ان استفاده شده برای رگرسیون، رگرسیون کی‌ان‌ان[٦] نیز نامیده می‌شود (بروس، بروس و گدک ۲۰۲۰، ۲۳۸).

۲-۲-۵-۱-۲- طبقه‌بندی‌کننده‌ی بِیز ساده[۷]

روش‌های طبقه‌بندی بیز ساده، مبتنی بر قضیه بیز است که توسط کشیش توماس بیز[۸] توسعه داده شده است. در قضیه‌ی بیز، «توزیع احتمال پیشین»[۹]، که در آن اطلاعات پیش‌بینی‌کننده در نظر گرفته نشده است، با اطلاعات جدید بدست آمده از داده‌های مشاهده شده، ترکیب می‌شوند. این ترکیب، منجر به دستیابی به «توزیع احتمال

---

[۱] Training Data Set

[۲] Test Data Set

[۳] Huyen

[٤] K-Nearest Neighbors (KNN)

[٥] کِی (K) تعداد همسایگان در نظر گرفته شده در فرایند محاسبه‌ی نزدیکترین همسایه است.

[٦] KNN Regression

[۷] Naive Bayes Classifier

[۸] Thomas Bayes

[۹] Prior Probability Distribution

پسین»[1] شده که در آن اطلاعات پیش‌بینی‌کننده در نظر گرفته شده است. با توجه به قضیه‌ی بیز، رابطه‌ی ۲–۱ بیانگر «احتمال شرطی»[2] برای خروجی y^* است (لاروس و لاروس ۲۰۱۹، ۱۱۳).

$$p(Y = y^*|X^*) = \frac{p(X^*|Y = y^*)p(Y = y^*)}{p(X^*)}$$

(۱–۲)

با توجه به ماهیت روش بیز ساده، این روش در طبقه‌بندی کاربرد دارد و استفاده از آن برای محاسبه‌ی رگرسیون مناسب نیست.

۲-۲-۵-۱-۳- رگرسیون خطی[3]

رگرسیون خطی ساده[4]، مدلی از رابطه بین بزرگی یک متغیر (مستقل) و یک متغیر (وابسته) را ارائه می‌دهد. برای مثال، نشان می‌دهد که با افزایش X، Y نیز افزایش یافته و یا کاهش می‌یابد. همبستگی[5] روش دیگری برای اندازه‌گیری نحوه ارتباط دو متغیر است. تفاوت در این است که در حالی که همبستگی *قدرت* ارتباط بین دو متغیر را اندازه‌گیری می‌کند، رگرسیون *ماهیت* رابطه را کمی می‌کند (بروس، بروس و گدک ۲۰۲۰، ۱٤۱). رابطه‌ی ۲–۲ معادله‌ی رگرسیون خطی ساده را نشان می‌دهد (بروس، بروس و گدک ۲۰۲۰، ۱٤۳).

$$Y = b_0 + b_1X$$

(۲–۲)

در **رگرسیون خطی چندگانه**[6]، بیش از یک متغیر مستقل (پیش‌بینی کننده) وجود دارد؛ درنتیجه معادله تعمیم می‌یابد تا آنها را در نظر بگیرد. رابطه‌ی ۲–۳ نمایشگر این معادله می‌باشد.

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

(۳–۲)

رابطه بین پاسخ (متغیر وابسته) و متغیر(ها) پیش‌بینی‌کننده (مستقل) لزوماً خطی نیست. مثلاً تقاضا برای یک محصول، تابع خطی میزان هزینه‌ی اختصاص یافته برای بازاریابی آن نیست و در نقطه‌ای تقاضا اشباع می‌شود.

---

[1] Posterior Probability Distribution

[2] احتمال مشاهده یک رویداد (مثلا $X = i$) با توجه به رویداد دیگری (مثلا $Y = i$) که به صورت $P(X_i|Y_i)$ نوشته می‌شود؛ احتمال شرطی (Conditional probability) نام دارد.

[3] Linear Regression

[4] Simple Linear Regression

[5] Correlation

[6] Multiple Linear Regression

راه‌های زیادی وجود دارد که می‌توان رگرسیون را برای ثبت این اثرات غیرخطی تعمیم داد (بروس، بروس و گدک ۲۰۲۰، ۱۸۷).

در **رگرسیون چند جمله‌ای**[۱]، عبارات چند جمله‌ای در یک معادله رگرسیونی گنجاندن می‌شوند. استفاده از رگرسیون چند جمله‌ای، تقریباً همزمان با توسعه‌ی مفهوم رگرسیون در مقاله‌ای توسط جرگون[۲] در سال ۱۸۱۵ معرفی شد. برای مثال، رگرسیون درجه دوم بین پاسخ Y و پیش‌بینی‌کننده‌ی X به شکل رابطه‌ی ۲–٤ است (بروس، بروس و گدک ۲۰۲۰، ۱۸۸).

$$Y = b_0 + b_1X_1 + b_2X^2 + \ldots + b_pX^p \quad (٤–٢)$$

همانطور که از نام رگرسیون خطی برمی‌آید، از این روش برای حل مسائل رگرسیون استفاده می‌شود. خروجی رگرسیون خطی، یک عدد پیوسته است و درنتیجه استفاده از آن برای حل مسائل طبقه‌بندی، غیر منطقی است.

۲-۲-۵-۱-٤- رگرسیون لجستیک

زمانی که پاسخ یک متغیر دوسویی (وجود یا عدم وجود، بله یا خیر و..) باشد، در این حالت می‌توان از رگرسیون لجستیک بهره برد. در رگرسیون لجستیک، پاسخ در یکی از دو دسته‌ی مدنظر قرار می‌گیرد. رگرسیون لجستیک، بجای اینکه مستقیما مدلی از پاسخ ارائه دهد، مدلی از احتمال متعلق بودن پاسخ به یک دسته‌ی خاص را ارائه می‌دهد (جیمز و دیگران ۲۰۲۱، ۱۳۳). در مدل رگرسیون لجستیک، از تابع لجستیک، که در رابطه‌ی ۲–۵ آورده شده، استفاده می‌شود (جیمز و دیگران ۲۰۲۱، ۱۳٤).

$$p(X) = \frac{e^{\beta_0+\beta_1X}}{1 + e^{\beta_0+\beta_1X}} \quad (٥–٢)$$

هرچند رگرسیون لجستیک یک الگوریتم رگرسیون است، اما عمدتاً در فرایند طبقه‌بندی مورد استفاده قرار می‌گیرد. به عبارت دیگر، معمولاً از رگرسیون لجستیک برای تخمین زدن احتمال تعلق یک نمونه به یک کلاس خاص استفاده می‌شود. به عنوان مثال، برای محاسبه‌ی احتمال تعلق داشتن یک ایمیل به کلاس هرزنامه، می‌توان از رگرسیون لجستیک بهره‌برد. اگر احتمال تخمین زده شده بیشتر از ٪۵۰ باشد، مدل پیش‌بینی می‌کند که نمونه متعلق به آن کلاس است (به نام کلاس مثبت، با برچسب «۱»). در غیر این صورت پیش‌بینی می‌کند که تعلق

[۱] Polynomial regression
[۲] Gergonne

ندارد (یعنی متعلق به کلاس منفی است؛ با برچسب «۰»). در حقیقت رگرسیون لجستیک، به عنوان یک طبقه‌بندی‌کننده‌ی دودویی (باینری) شناخته می‌شود (جرون[۱] ۲۰۱۹، ۲۰۵).

۲-۲-۵-۱-۵- ماشین بردار پشتیبانی (اس‌وی‌ام)[۲]

ماشین بردار پشتیبان، رویکردی برای طبقه‌بندی است که در دهه‌ی ۱۹۹۰ در جامعه علوم کامپیوتر توسعه یافت و از آن زمان محبوبیت آن افزایش یافته است. اس‌وی‌ام‌ها در زمینه‌های مختلف عملکرد خوبی دارند و اغلب به عنوان یکی از بهترین طبقه‌بندی‌کننده‌های «خارج از جعبه» در نظر گرفته می‌شوند (جیمز و دیگران ۲۰۲۱، ۳٦۷).

در حقیقت، اس‌وی‌ام یک مدل طبقه‌بندی متمایزکننده است که با یادگیری مرزهای تصمیم‌گیری خطی یا غیرخطی، کلاس‌های متفاوت را از هم جدا می‌کند. علاوه بر فراهم آوردن قابلیت حداکثر رسانی تفکیک‌پذیری دو کلاس، اس‌وی‌ام‌ها ظرفیت بالایی برای منظم‌سازی ارائه می‌دهند. به عنوان مثال، این مدل امکان کنترل کردن پیچیدگی را فراهم می‌آورد تا عملکرد تعمیم مدل، بهینه‌تر شود. درنتیجه‌ی این توانایی منحصر به فرد در منظم‌سازی فرایند یادگیری، اس‌وی‌ام‌ها می‌توانند به‌دور از خطر «بیش‌برازش»[۳]، مدل‌هایی بسیار گویا تولید کنند. بنابراین، اس‌وی‌ام‌ها توجه زیادی را در جامعه یادگیری ماشین به خود جلب کرده‌اند. از این مدل معمولاً در زمینه‌ها متعدد، از تشخیص دست‌خط گرفته تا طبقه‌بندی متن، استفاده می‌شود (تن[٤] و دیگران ۲۰۱۹، ٤۷۸).

همانطور که اشاره شده، ماشین‌های بردار پشتیبان، یک طبقه‌بندی‌کننده به حساب می‌آیند. با این وجود، بسطی از این مدل برای رگرسیون (با خروجی کمی) وجود دارد که «رگرسیون بردار پشتیبان»[۵] نام دارد.

۲-۲-۵-۱-٦- درخت تصمیم[٦]

درخت تصمیم مدلی پرکاربرد در زمینه‌ی یادگیری ماشین است. همانطور که از نام «درخت تصمیم» پیداست، در این مدل، مجموعه‌ای از سؤالات در هر مرحله مطرح شده و مبتنی بر آن‌ها تصمیماتی اتخاذ می‌شود. به این ترتیب، داده‌های مورد بررسی در اثر نتیجه‌ی تصمیمات اتخاذ شده، تجزیه و تحلیل می‌شوند (راشکا، لیو و میرجلیلی ۲۰۲۲، ۸٦).

---

[۱] Géron

[۲] Support Vector Machine (SVM)

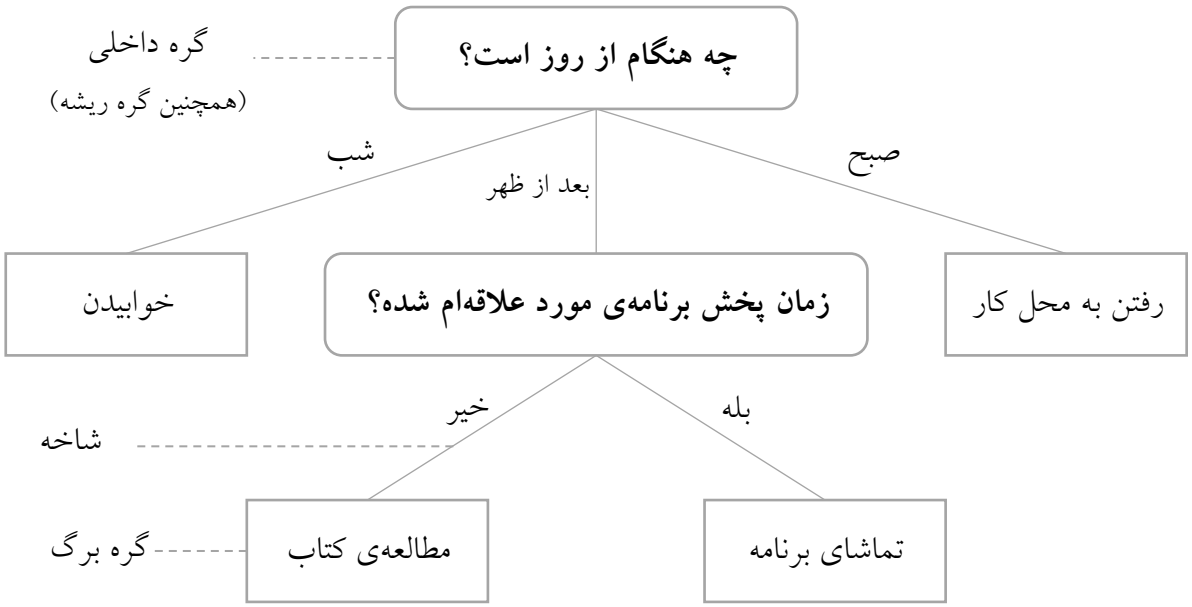
[۳] بیش‌برازش (Overfitting): تولید مدلی که بیش از حد بر روی مجموعه داده‌های آموزش متمرکز شده است؛ درنتیجه فقط روی داده‌های آموزش، عملکرد مناسبی دارد اما به طور ضعیفی به مجموعه داده‌ی جدید تعمیم می‌یابد (گروس ۲۰۱۹، ۱٤۹).

[٤] Tan

[۵] Support Vector Regression (SVR)

[٦] Decision Tree

مدل درخت تصمیم، بر اساس ویژگی‌های موجود در مجموعه داده‌ی آموزشی، یک سری سؤالات یاد می‌گیرد تا بتواند به کمک آن‌ها، برچسب‌های کلاس نمونه‌ها را استنتاج کند. برای استفاده از الگوریتم تصمیم، از ریشه درخت شروع کرده و داده‌ها بر روی ویژگی‌ای تقسیم می‌شوند که بیشترین «کسب اطلاعات»[۱] را به همراه داشته باشد. سپس می‌توان این روش تقسیم را در هر «گره»[۲] فرزند تکرار کرد تا زمانی که «برگ‌ها»[۳] خالص شوند؛ یعنی نمونه‌های آموزشی در هر گره، همه متعلق به یک کلاس شوند. در عمل، تکرار بیش از حد این فرایند، می‌تواند منجر به یک درخت بسیار عمیق با گره‌های بسیار شده و به راحتی به «بیش‌برازش» منتج شود. بنابراین، معمولاً باید با تعیین محدودیت برای حداکثر عمق درخت، اصطلاحاً آن‌را «هرس»[٤] کنیم (راشکا، لیو و میرجلیلی ۲۰۲۲، ۸۷). نمونه‌ای از مدل درخت تصمیم در نمودار ۲–۱ آورده شده است.



(نمودار ۲–۱) نمونه‌ای از درخت تصمیم برای تصمیم‌گیری درباره‌ی فعالیت روز.

اگر به تفسیرپذیری اهمیت دهیم، طبقه‌بندی‌کننده‌های درخت تصمیم مدل‌های جذابی هستند (راشکا، لیو و میرجلیلی ۲۰۲۲، ۸٦). دیگر مزیت مدل درخت تصمیم، انعطاف‌پذیری آن است. درنتیجه الگوریتم درخت تصمیم با ویژگی‌های دلخواه کار می‌کند و اگر با داده‌های غیرخطی سروکار داریم، نیازی به تغییر ویژگی‌ها نیست؛ زیرا درخت‌های تصمیم به جای تحلیل ترکیبی ویژگی‌ها (و وزن‌دار کردن آن‌ها)، در هر لحظه فقط یک ویژگی خاص را تحلیل می‌کنند.

---

[۱] Information Gain
[۲] Node
[۳] Leaves
[٤] Prune

درنتیجه، طبقه‌کننده‌ی درخت تصمیم[۱] و رگرسیون درخت تصمیم[۲] هر دو رویکردهایی مناسب به حساب می‌آیند (راشکا، لیو و میرجلیلی ۲۰۲۲، ۳۰۰). به روش‌های آماری برای ساخت پیش‌بینی‌کننده درخت (یا درخت تصمیم) با هدف حل مسائل رگرسیون و طبقه‌بندی، «درخت‌های رگرسیون و طبقه‌بندی»[۳] یا «کارت» گفته می‌شود (گنور[٤] و پوگی[٥] ۲۰۲۰، ۹).

۲-۲-۵-۱-۷- جنگل تصادفی[٦]

در سال ۱۹۹٤، لئو بریمن[۷]، پروفسور دانشگاه برکلی، یک سال پس از بازنشستگی، گزارش فنی کوچکی به نام «پیش‌بینی‌کنندگان بسته‌بندی»[۸] منتشر کرد که یکی از تأثیرگذارترین ایده‌ها در یادگیری ماشین مدرن بود. در این گزارش، روشی برای تولید نسخه‌های چندگانه یک پیش‌بینی‌کننده و استفاده از آن‌ها برای بدست آوردن یک پیش‌بینی‌کننده‌ی تجمیع‌شده، معرفی شده است. در سال ۲۰۰۱، بریمن نشان داد که این رویکرد برای ساخت مدل‌ها، به‌خصوص زمانیکه روی الگوریتم‌های ساخت درخت تصمیم اعمال می‌شود، بسیار قدرتمند است. او این روش را جنگل تصادفی نامید. بریمن که در گزارش ابتدایی‌اش، فقط از زیر مجموعه‌های تصادفی ردیف‌ها استفاده می‌کرد، این بار از زیر مجموعه‌های تصادفی ستون‌ها نیز، هنگام انتخاب هر تقسیم در هر درخت تصمیم، بهره‌برد (هوارد و گوگر ۲۰۲۰، ۲۹۸).

جنگل‌های تصادفی نمونه‌ای از «روش‌های تجمیعی»[۹] هستند، که در آن مدل‌هایی از درخت تصمیم بر اساس یک سری زیر مجموعه‌ی تصادفی از داده‌های مورد بررسی، آموزش داده می‌شوند. درنتیجه، چندین مدل آموزش دیده ساخته می‌شود که هر کدام از آن‌ها، یک خروجی خاص پیش‌بینی می‌کنند. سپس میانگین این پیش‌بینی‌ها به عنوان خروجی نهایی معرفی می‌گردد (لاروس و لاروس ۲۰۱۹، ۹۱).

امروزه شاید پرکاربردترین و عملاً مهم‌ترین روش یادگیری ماشین (به‌خصوص در حل مسائل رگرسیون و طبقه‌بندی)، جنگل‌های تصادفی باشند (هوارد و گوگر ۲۰۲۰، ۲۹۸).

---

[۱] Decision Tree Classifier

[۲] Decision Tree Regression

[۳] Classification And Regression Trees (CART)

[٤] Genuer

[٥] Poggi

[٦] Random Forest

[۷] Leo Breiman

[۸] Bagging Predictors

[۹] روش‌های تجمیعی (Ensemble Methods) دسته‌ای از تکنیک‌های مدل‌سازی هستند که خروجی چندین مدل را برای رسیدن به یک پاسخ در نظر می‌گیرند (لاروس و لاروس ۲۰۱۹، ۹۱).

۲-۲-۵-۲- یادگیری بدون نظارت[۱]

در این یادگیری که «یادگیری خودسازمانده» نیز نامیده می‌شود، عامل، الگوهای موجود در ورودی‌ها را بدون هیچ بازخورد صریحی می‌آموزد. رایج‌ترین کاربرد یادگیری بدون نظارت، عمل خوشه‌بندی است. در این مورد، عامل، خوشه‌های بالقوه مفید نمونه‌های ورودی را شناسایی می‌کند. برای مثال، هنگامی که میلیون‌ها تصویر گرفته شده از اینترنت نشان داده می‌شود، یک سیستم بینایی رایانه‌ای می‌تواند خوشه‌ای از تصاویر مشابه را شناسایی کند که یک انگلیسی‌زبان آن‌ها را «گربه» می‌نامد (راسل و نورویگ ۲۰۲۱، ٦٧١).

۲-۲-۵-۳- یادگیری تقویتی[۲]

در این یادگیری که «یادگیری پاداش و تاوان» نیز نامیده می‌شود، عامل از مجموعه‌ای از عوامل تقویتی شامل پاداش‌ها و تنبیه‌ها یاد می‌گیرد. به عنوان مثال، در پایان یک بازی شطرنج به عامل گفته می‌شود که برنده شده (پاداش) یا شکست خورده (تنبیه) است. این به عامل بستگی دارد که تصمیم بگیرد، کدام یک از اقدامات قبلی مسئول اصلی این نتیجه بوده است و اقدامات خود را در راستای دستیابی به پاداش‌های بیشتر در آینده، تغییر دهد (راسل و نورویگ ۲۰۲۱، ٦٧١).

در برخی رویکردهای ضمنی مورد استفاده در مدل‌های یادگیری نظارت شده، یادگیری تقویتی مشاهده می‌شود.

۲-۲-٦- یادگیری عمیق[۳]

نوعی یادگیری ماشین که از الگوریتم‌ها بر اساس نحوه عملکرد مغز انسان استفاده می‌کند (فرهنگ لغت الکترونیکی کمبریج، واژه‌ی یادگیری عمیق).

در یادگیری عمیق، از فعالیت‌های لایه‌های نورون در نئوکورتکس (تشکیل دهنده‌ی ٪۸۰ قسمت چروکیده مغز که عمل فکر کردن در آن رخ می‌دهد)، تقلیدی صورت می‌گیرد. برای دستیابی به این منظور، عمل یادگیری ماشین مبتنی بر ساختارهای سلسله مراتبی و سطوح نمایان و انتزاعی انجام می‌شود. این اقدامات در جهت درک الگوهای داده‌هایی که از منابعی مانند تصاویر، صداها و متون استخراج می‌شوند، است (دی، بهاردواج و وی ۲۰۱۸، ۸).

---

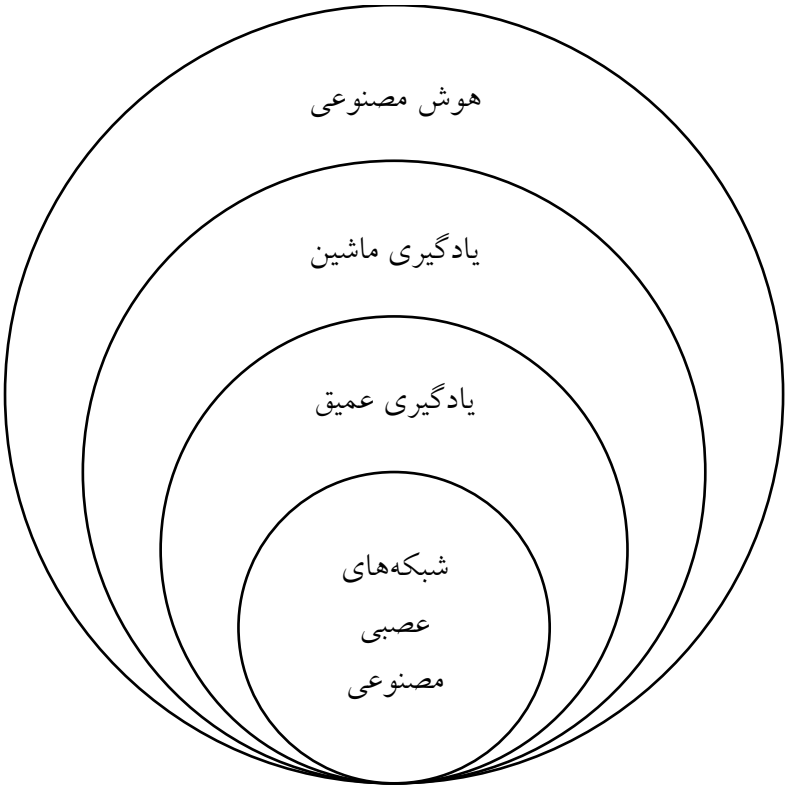
[۱] Unsupervised Learning

[۲] Reinforcement Learning

[۳] Deep Learning

۲-۲-۶-۱- شبکه‌های عصبی مصنوعی[۱]

شبکه‌هایی که با روش‌های یادگیری عمیق آموزش داده می‌شوند، اغلب شبکه‌های عصبی مصنوعی (یا به طور ساده‌تر شبکه‌های عصبی) نامیده می‌شوند، حتی اگر شباهت آن‌ها به سلول‌ها و ساختارهای عصبی واقعی سطحی باشد (راسل و نورویگ ۲۰۲۱، ۸۰۱). در شکل ۲–۲ رابطه‌ی سلسله مراتبی میان شبکه‌های عصبی مصنوعی و هوش مصنوعی قابل مشاهده است.



(شکل ۲–۲) رابطه‌ی سلسله مراتبی میان هوش مصنوعی و شبکه‌های عصبی مصنوعی.

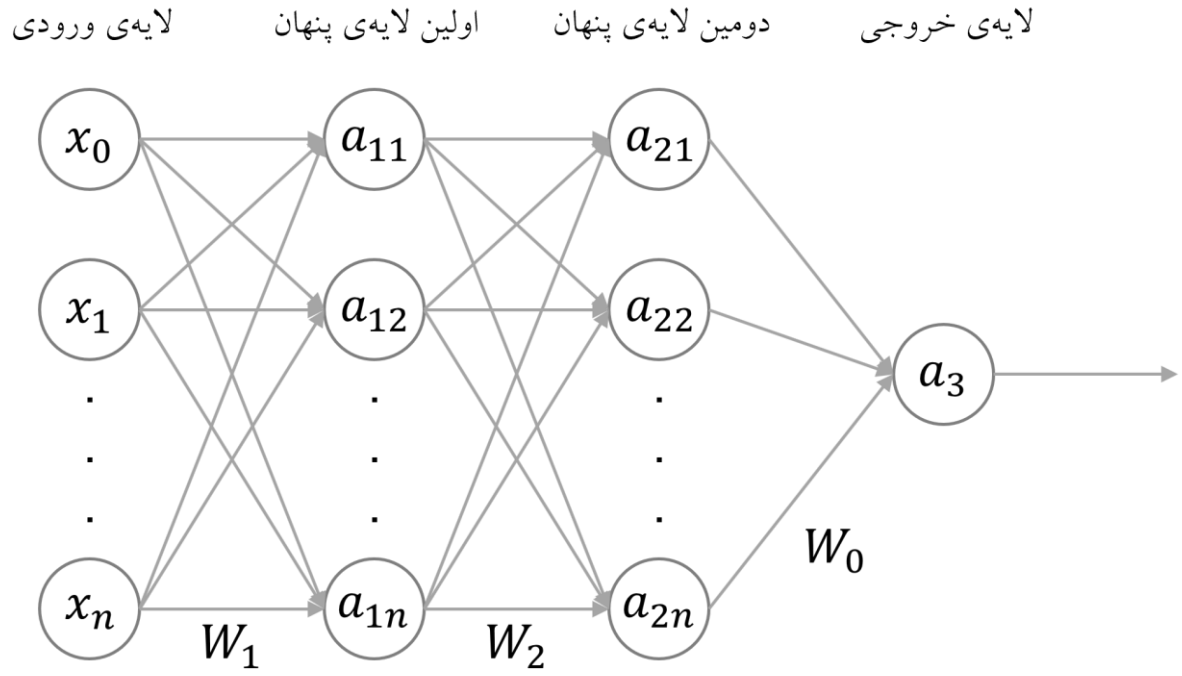
استفاده از سه نوع شبکه‌ی عصبی مصنوعی، متداول‌تر است. این شبکه‌ها عبارتند از:

۲-۲-۶-۱-۱- شبکه‌های عصبی پیشخور[۲]

این نوع شبکه‌های عصبی، اولین و ساده‌ترین شبکه‌های عصبی هستند و رویکرد پیش‌فرض در شبکه‌های مصنوع عصبی، به حساب می‌آید. «پرسپترون چندلایه»[۳] یک نوع ساده از این شبکه‌ها است که یک «لایه‌ی

[۱] Artificial Neural Networks (ANN)
[۲] Feedforward Neural Networks (FNN)
[۳] Multilayer Perceptron

ورودی»[۱]، یک «لایه خروجی»[۲] و چند (بیش از یک) «لایه‌ی پنهان»[۳] دارد. هر لایه دارای چندین نورون بوده و لایه‌های مجاور به طور کامل به هم متصل هستند. سیگنال‌های لایه‌های قبلی از طریق وزنه‌های متصل (سیناپس‌ها) به نورون لایه بعدی به جلو رانده می‌شوند. برای هر نورون مصنوعی، مجموع وزنی همه ورودی‌های با ضرب سیگنال در وزن‌ها و اضافه کردن یک «بایِس»[٤] (آفسِت)، محاسبه می‌کند. سپس بر اساس مجموع وزنی بدست آمده، تابعی به نام «تابع فعال‌سازی»[٥]، سیگنال‌های خروجی برای سطح بعدی را مشخص می‌کند. به عنوان مثال، یک شبکه عصبی پیشخور، در نمودار ۲–۲ نشان داده شده است (دی، بهاردواج و وی ۲۰۱۸، ٤٧).



(نمودار ۲–۲) نمونه‌ای از یک شبکه عصبی پیشخور ساده.

بر اساس نوشته‌ی دی، بهاردواج و وی (۲۰۱۸، ٦٦) در شبکه‌های عصبی، به طور کلی دو نوع جریان وجود دارد:

---

[۱] Input Layer

[۲] Output Layer

[۳] Hidden Layer

[٤] بایس (Bias) ثابتی است که به حاصل ضرب ویژگی‌ها و وزن‌ها اضافه می‌شود تا نتیجه را متعادل سازد و به آن آفسِت نیز گفته می‌شود (استیونز، آنتیگا و ویهمن ۲۰۲۰، ۲۷).

[٥] Activation Function

۱. اولین جریان «انتسار به جلو»[۱] نام دارد که در آن اساساً محاسبه‌ی ضرب داده‌های ورودی در وزن ضرب شده، با بایس جمع می‌شود و پس از عبور از تابع فعال‌سازی، به لایه‌ی بعدی منتقل می‌گردد.

۲. دومین جریان «انتشار به عقب»[۲] نام داشته و در آن با مقایسه‌ی مقادیر پیش‌بینی شده و مقادیر صحیح، شبکه از خطاهای انجام شده، یاد می‌گیرد و سپس وزن و پارامترهای قبلی را، براساس یک «تابع هزینه»[۳]، به‌روز می‌کند. تابع هزینه، بهبود را بر اساس «امتیاز زیان»[٤]، انجام می‌دهد.

این دو نوع جریان فقط مختص به شبکه‌های پیشخور نبوده و به طور کلی در سایر شبکه‌های عصبی نیز مشاهده می‌شوند. انتشار به عقب جریانی است که شبکه‌های عصبی را به یک مدل جذاب تبدیل کرده؛ چرا که در این جریان، عمل یادگیری، مشابه به آنچه در مغز انجام می‌شود، رخ می‌دهد.

۲-۲-٦-۱-۲- شبکه‌های عصبی پیچشی[٥]

نوعی شبکه عصبی مصنوعی است که حداقل در لایه‌های اولیه دارای اتصالات محلی بوده و حاوی الگوهایی از وزن‌ها است که در سراسر واحدهای هر لایه تکرار می‌شوند. الگویی از وزن‌ها که در چندین ناحیه‌ی محلی تکرار می‌شود، یک هسته[٦] نامیده می‌شود. فرآیند اعمال هسته بر روی پیکسل‌های تصویر (یا روی واحدهای سازمان‌یافته فضایی در لایه‌ی بعدی) کانولوشن (پیچیدگی) نامیده می‌شود (راسل و نورویگ ۲۰۲۱، ۸۱۱).

با توجه به ماهیت مسئله‌ی مدنظر، در این تحقیق استفاده‌ای از شبکه‌ها عصبی پیچشی نشده است. درنتیجه به معرفی اجمالی انجام شده اکتفا می‌کنیم.

۲-۲-٦-۱-۳- شبکه‌های عصبی بازگشتی[۷]

همانطور که اشاره شد، در یک شبکه‌ی عصبی پیچشی یا شبکه‌ی پیشخور معمولی، اطلاعات از طریق یک سری عملیات ریاضی که بر روی گره‌ها انجام می‌شود، بدون حلقه بازخورد یا هرگونه در نظر گرفتن ترتیب سیگنال، انجام می‌شود. بنابراین، آنها قادر به در نظر گرفتن تولی ورودی‌ها نیستند. با این حال، در عمل،

---

[۱] Forward Propagation

[۲] Back Propagation

[۳] تابع هزینه (Cost Function) پیش‌بینی‌های شبکه و هدف واقعی (آنچه می‌خواهیم شبکه بدست بیاورد) را می‌گیرد و امتیاز فاصله را محاسبه می‌کند تا نشان می‌دهد که شبکه در یک مثال خاص چقدر خوب عمل کرده است (شوله ۲۰۲۱، ۹). این تابع همچنین «تابع زیان» (Loss Function) نیز خوانده می‌شود.

[٤] امتیاز زیان (Loss score) مقداری است که میزان پیش‌بینی اشتباه مدل را مشخص می‌کند و هرچه مقدار آن پایین‌تر باشد، به معنای عملکرد بهتر مدل است. اگر همه‌ی پیش‌بینی‌های مدل صحیح باشد؛ آنگاه مقدار زیان آن برابر صفر خواهد بود (شوله ۲۰۲۱، ۱۰).

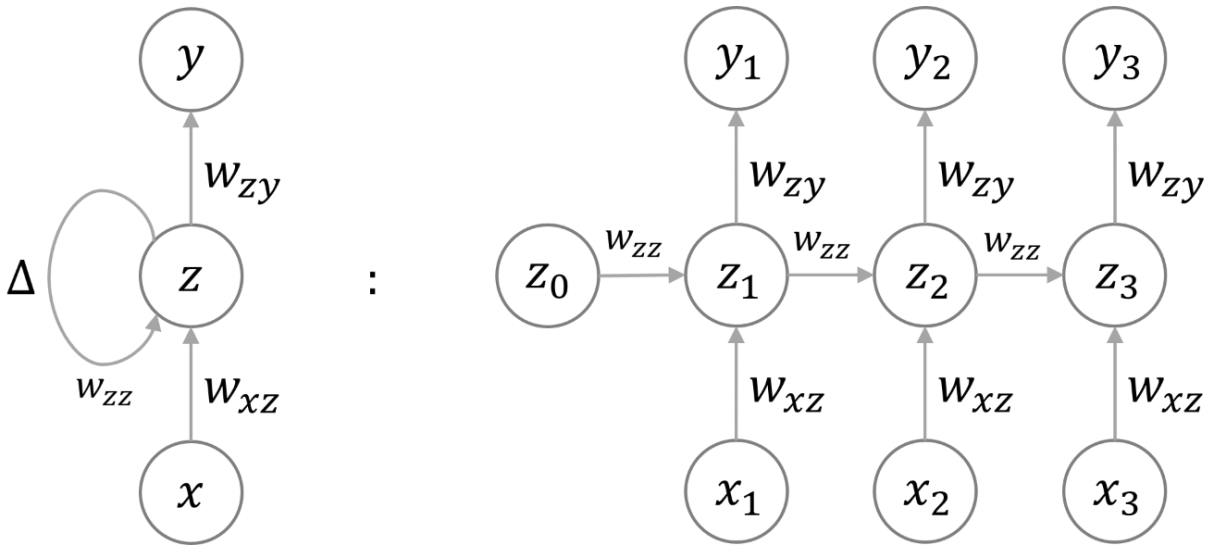
[٥] Convolutional Neural Networks (CNN)

[٦] Kernel

[۷] Recurrent Neural Networks (RNN)

داده‌های متوالی زیادی مانند جملات، ژنوم، اعداد مربوط به بازار سهام و... وجود دارند که در آن‌ها مقدار بعدی در ردیف، اغلب تا حد زیادی به زمینه‌ی گذشته (طولانی یا کوتاه مدت) مرتبط است. شبکه‌های عصبی بازگشتی، شکل جدیدی از شبکه‌ی عصبی مصنوعی هستند که به طور خاص برای این نوع داده‌ها طراحی شده‌اند.

بنابراین، شبکه‌های عصبی بازگشتی از شبکه‌های پیش‌خور متمایز هستند؛ زیرا امکان بهره‌گیری از چرخه‌ها را در نمودار محاسباتی فراهم می‌کنند. از آنجا که واحدها ممکن است مقداری را که از خروجی خودشان در مرحله قبلی محاسبه شده، دوباره به عنوان ورودی بگیرند؛ هر چرخه دارای میزانی تاخیر است. تأخیر لحاظ می‌شود تا از قرارگیری مدار چرخه در حالت ناسازگار، جلوگیری کند. درنتیجه شبکه‌های عصبی بازگشتی حالت داخلی یا حافظه[1] دارند تا تأثیر ورودی‌های دریافت شده در مراحل زمانی قبلی بر پاسخ شبکه‌های عصبی بازگشتی به ورودی فعلی، امکان‌پذیر شود (راسل و نورویگ ۲۰۲۱، ۸۲۳–۸۲٤). نمودار شماتیک یک شبکه‌ی بازگشتی پایه، در نمودار ۲–۳ نمایش داده شده است.



(آ) (ب)

(نمودار ۲–۳) (آ) نمودار شماتیک یک شبکه‌ی بازگشتی پایه که در آن لایه پنهان z دارای اتصالات مکرر است. نماد Δ نشان دهنده تاخیر است. (ب) همان شبکه طی سه مرحله زمانی باز شد تا یک شبکه پیشخور ایجاد کند. توجه داشته باشید که وزن‌ها در تمام مراحل زمانی به اشتراک گذاشته می‌شوند (اقتباس از راسل و نورویگ ۲۰۲۱، ۸۲٤).

[1] Memory

۲-۲-۷- زبان‌های برنامه‌نویسی پرکاربرد در یادگیری ماشین

زبان‌ها برنامه‌نویسی مجموعه‌ای از لغات و نشانه‌ها هستند که از آن‌ها برای ساخت برنامه‌های کامپیوتری استفاده می‌شود (داونی[۱] و میفیلد[۲] ۲۰۲۰، ۲). در زیر، به دو زبان برنامه‌نویسی پر استفاده در زمینه‌ی یادگیری ماشین، پرداخته شده است:

۲-۲-۷-۱- زبان برنامه‌نویسی پایتون[۳]

پایتون یک زبان برنامه‌نویسی سطح بالا، شی‌گرا و پویا است. سطح بالا بودن در کنار ساختار پویای آن، پایتون را برای توسعه سریع برنامه‌ها بسیار جذاب می‌کند (بنیاد نرم‌افزار پایتون). بنابراین به دلیل سادگی و دسترس‌پذیری پایتون و کتابخانه‌های بی‌شمار آن، این زبان برنامه‌نویسی طرفداران زیادی را، به‌خصوص در زمینه‌ی یادگیری ماشین، پیدا کرده است.

۲-۲-۷-۲- زبان برنامه‌نویسی آر[٤]

آر یک زبان و محیطی برای محاسبات آماری و گرافیکی است. این زبان توسط جان چمبرز[٥] و همکارانش در آزمایشگاه‌های بل توسعه داده شد. آر، طیف گسترده‌ای از تکنیک‌های آماری (مدل‌سازی خطی و غیرخطی، آزمون‌های آماری کلاسیک، طبقه‌بندی و...) را در کنار تکنیک‌های گرافیکی ارائه می‌کند. این زبان بسیار توسعه‌پذیر است (بنیاد زبان آر).

۲-۲-۸- کتابخانه‌های نرم‌افزار پرکاربرد در یادگیری ماشین

کتابخانه‌ی نرم‌افزار، مجموعه‌ی گسترده‌ای از بسته‌ها، کلاس‌ها و توابع است که برای استفاده در برنامه‌های دیگر در دسترس هستند (داونی و میفیلد ۲۰۲۰، ٤٠). برخی از کتابخانه‌های پرطرفدار در زمینه‌ی یادگیری ماشین، که در تحقیق حاضر مورد استفاده قرار گرفتند، به شرح زیر می‌باشند:

---

[۱] Downey
[۲] Mayfield
[۳] Python Programming Language
[٤] R Programming Language
[٥] John Chambers

۲-۲-۸-۱- کتابخانه‌ی سایکیت-لرن[۱]

کتابخانه‌ی سایکیت-لرن، یک کتابخانه‌ی متن‌باز پایتون برای یادگیری ماشینی است که بر روی کتابخانه‌ی سای‌پای[۲] ساخته شده است (کورناپو[۳] ۲۰۲۲، صفحه‌ی رسمی سایکیت-لرن در گیت‌هاب). این کتابخانه، بسیاری از الگوریتم‌های یادگیری ماشین را به طور مؤثر پیاده‌سازی می‌کند و استفاده از آن بسیار راحت می‌باشد (جرون ۲۰۱۹، ۷).

در این تحقیق، از این کتابخانه برای پیاده‌سازی مدل‌های سنتی یادگیری ماشین بهره گرفته شد و مدل‌های یادگیری عمیق (مبتنی بر شبکه‌های عصبی) بر پایه‌ی کتابخانه‌هایی که در ادامه معرفی خواهند شد، پیاده‌سازی شده‌اند.

۲-۸-۲-۲- کتابخانه‌ی تنسورفلو[٤]

تنسورفلو، یک کتابخانه پیچیده‌تر برای محاسبات عددی توزیع شده است. این کتابخانه، با توزیع محاسبات در بین صدها سرور بالقوه با «واحد پردازش گرافیکی (جی‌پی‌یو)»[٥] چندگانه، آموزش و اجرای شبکه‌های عصبی بسیار بزرگ را، به طور کارآمد، ممکن می‌سازد. تنسورفلو توسط گوگل توسعه داده شده و در سال ۲۰۱۵ به صورت متن‌باز، در اختیار عموم قرار گرفته است. بسیاری از برنامه‌های یادگیری ماشین بزرگ‌مقیاس گوگل، از این کتابخانه پشتیبانی می‌کنند (جرون ۲۰۱۹، ۷-۸).

گوگل همچنین محصولی به نام «کولب»[٦] ارائه کرده است که محیطی برای نوشتن و اجرای کدهای پایتون، از طریق مرورگر است. این پلتفرم، امکان استفاده کردن از کتابخانه‌های یادگیری ماشین، به‌ویژه تنسورفلو را فراهم کرده و برای یادگیری ماشین، تجزیه و تحلیل داده‌ها و به اشتراک گذاری پروژه‌ها بسیار مناسب است. این سرویس، دسترسی رایگان به منابع محاسباتی از جمله جی‌پی‌یوها را فراهم می‌کند (وبسایت گوگل – بخش تحقیقات).

---

[۱] Scikit-Learn Library
[۲] SciPy
[۳] Cournapeau
[٤] TensorFlow Library
[٥] Graphics Processing Unit (GPU)
[٦] Google Colab

۲-۲-۸-۳- کتابخانه‌ی کِرس[۱]

کرس، یک کتابخانه‌ی یادگیری عمیق سطح بالا است که آموزش و اجرای شبکه‌های عصبی را تسهیل می‌کند. این کتابخانه، که همچنین «واسط برنامه‌نویسی کاربردی»[۲] نیز محسوب می‌شود، می‌تواند در بستر تنسورفلو، «تیانو»[۳] یا «مجموعه ابزار مایکروسافت کاگنتیو»[٤] اجرا شود. تنسورفلو با پیاده‌سازی خاص خود از این واسط باعث شده که کرس از برخی ویژگی‌های پیشرفته‌ی تنسورفلو، مثل امکان بارگذاری بهینه‌ی داده‌ها، برخوردار شود (جرون ۲۰۱۹، ۸).

۲-۳- بررسی مقالات اخیر مرتبط با موضوع

در این بخش، مقالات مرتبط و به‌روز انتشاریافته از مجلات معتبر، مورد بررسی قرار خواهند گرفت. این مقالات، در حوزه‌ی سازماندهی مبتنی بر شبکه‌های عصبی انتشار یافته و از رویکردهایی مرتبط با رویکردهای مدنظر در این تحقیق، بهره برده‌اند. این مقالات در دو بخش مقالات فارسی و مقالات لاتین مورد بررسی قرار گرفته‌اند و در انتها رویکردهای تحقیق آن‌ها، در قالب یک جدول با یکدیگر مقایسه شده‌اند.

۲-۳-۱- مقالات فارسی

با توجه به جستجوهای انجام شده، علی‌رغم اهمیت زیاد موضوع، مقالات معتبر و جدید زیادی در این زمینه به زبان فارسی منتشر نشده‌اند. لذا محقق به بررسی سه مقاله‌ی فارسی در این قسمت پرداخته است.

یک روش هوشمند برای طبقه بندی ترک در سازه‌های بتنی بر اساس شبکه‌های عصبی عمیق (بیگدلی، جباری، شجاعی ۱٤۰۰، ۱٦-۱):

در این مقاله، محقق با توجه به چالش برانگیز بودن شناسایی و بررسی انواع ترک‌ها در سازه‌های بتنی، روشی مبتنی بر شبکه‌های عصبی پیچشی برای طبقه‌بندی آن‌ها ارائه داده است. برای اینکار ابتدا ویدئویی از سازه‌های بتنی تهیه شده و سپس به تصویری سراسرنما تبدیل شده است. در مرحله بعد تصویر سراسرنما به تصاویر با ابعاد کوچک و با رزولوشن مطلوب تجزیه شده و پیش‌پردازش‌هایی برای آماده‌سازی آن انجام گرفته است. در انتها با استفاده از لایه‌های شبکه‌ی عصبی پیچشی، ویژگی‌های مربوطه استخراج و عمل طبقه‌بندی به سه دسته‌ی بدون ترک، ترک ساده و چند شاخگی در ترک انجام شده است.

---

[۱] Keras Library

[۲] Application Programming Interface (API)

[۳] Theano

[٤] Microsoft Cognitive Toolkit

در این مقاله از ۱۲۰۰۰ تصویر استفاده شده که ٪۷۵ آن مجموعه داده‌ی آموزش ٪۲۵ باقی مجموعه داده تست بوده‌اند. یافته‌های مقاله حکایت از دقت ٪۹۹.۳ و فراخوانی ٪۹۹.۵ روش ارائه شده دارد.

در انتها محقق نتیجه گرفته که روش ارائه شده در حالیکه نسبت به سایر رویکردهای مورد مطالعه در تحقیق از پیچیدگی کمتری برخوردار است، دقت و کارایی بالاتری دارد. محدودیت روش ارائه شده، نیاز آن به سخت‌افزار قوی (با توجه به زمان بالای اجرا) برشمرده شده است.

شناسایی موارد ابتلا به کووید-۱۹ با استفاده از شبکه‌ی عصبی فازی نوع ۲ عمیق براساس تصاویر اشعه‌ی ایکس قفسه سینه (صباحی و دیگران ۱٤۰۰، ۱۵-۱):

در این مقاله با توجه به ضرورت تشخیص به‌هنگام بیماری کووید-۱۹، روشی مبتنی بر شبکه‌های عصبی پیچشی و توابع فعال‌سازی فازی نوع دو ارائه شده است. در این روش ابتدا پیش‌پردازش داده‌ها شامل عملیات نرمال‌سازی، سگمنت‌بندی و افزایش داده‌ها روی تصاویر انجام شده و سپس عملیات انتخاب و استخراج ویژگی به منظور طبقه‌بندی انجام شده است. در این تحقیق، عمل طبقه‌بندی در دو زمینه انجام شده است. در زمینه‌ی اول افراد در دو دسته‌ی سالم و مبتلا به کویید-۱۹ و در زمینه‌ی دوم افراد در سه دسته‌ی سالم، مبتلا به سینه‌پهلو و مبتلا به کووید-۱۹ طبقه‌بندی شده‌اند.

بررسی نتایج نشان داد که صحت نهایی برای زمینه‌ی اول ٪۹۹ و برای زمینه‌ی دوم ٪۹۵ بوده است.

محقق نتیجه گرفته که با توجه به عملکرد مطلوب روش پیشنهادی، می‌توان از آن برای تشخیص افراد مبتلا به کووید-۱۹ استفاده کرد.

تشخیص بیماران اسکیزوفرنی با استفاده از شبکه‌های عصبی پیچشی از تصاویر ارتباطات مؤثر مغزی سیگنال‌های چندکاناله الکتروانسفالوگرام (باقرزاده، مقصودی و شالباف ۱٤۰۰، ۷٦۱-۷٥٤):

از آنجا که تشخیص بیماری اسکیزوفرنی با استفاده از تست‌های شناختی توسط روان‌پزشک به شدت به تجربه و دانش وی وابسته است؛ هدف این تحقیق طراحی یک چارچوب کاملاً خودکار برای تشخیص اسکیزوفرنی از روی سیگنال الکتروانسفالوگرام با استفاده از ارتباطات مؤثر مغزی و روش‌های یادگیری عمیق است.

در این تحقیق، ابتدا سیگنال‌های الکتروانسفالوگرام ۱۹ کاناله از ۱٤ بیمار مبتلا به اسکیزوفرنی و ۱٤ فرد سالم ثبت و پیش‌پردازش شده است. سپس، معیار ارتباطات مؤثر با استفاده از روش آنتروپی انتقالی، از سیگنال‌های الکتروانسفالوگرام تخمین زده شده و ماتریس ارتباطات نامتقارن ۱۹×۱۹ ساخته شده و با یک نقشه رنگی به‌عنوان یک تصویر نشان داده می‌شود. در نهایت این تصاویر به عنوان ورودی شبکه‌های عصبی پیچشی برای تشخیص بیماران اسکیزوفرنی استفاده می‌شوند.

یافته‌های این تحقیق نشان داد که بالاترین میانگین صحت و امتیاز اف برای طبقه‌بندی دو کلاس اسکیزوفرنی و سالم با استفاده از مدل شبکه‌ی عصبی «اینسپشن»[1]، با مقادیر به ترتیب ٪۹۶.۵۲ و ٪۹۵.۸۹ در ارزیابی مستقل از فرد و ٪۹۸.۵۱ و ٪۹۸.۵۱ در ارزیابی متقابل با ۱۰ دسته می‌باشد.

محقق نتیجه گرفت که با توجه به نتایج، مدل ارائه شده می‌تواند کمک شایانی به روان‌پزشکان در تشخیص دقیق افراد اسکیزوفرنیا داشته باشد.

۲-۳-۲- مقالات لاتین

با توجه به اهمیت موضوع، مقالات لاتین زیادی در زمینه‌ی شبکه‌های عصبی انتشار یافته که برخی از آن‌ها با رویکرد طبقه‌بندی از این مفهوم بهره‌مند شده‌اند. در ادامه به بررسی برخی منابع جدید و معتبر در این زمینه پرداخته شده است.

طبقه‌بندی آریتمی بین بیماران متکی بر شبکه عصبی پیچشی باقیمانده‌ی عمیق بهبودیافته (لی[2]، کیان[3] و لی ۲۰۲۲، ۱۰-۱):

در این تحقیق با توجه به افزایش مرگ و میر ناشی از بیماری‌های قلبی-عروقی و درنتیجه‌ی آن ضرورت تشخیص زودهنگام آریتمی‌ها، روشی ارائه شده که با بهره‌گیری از یک شبکه عصبی پیچشی باقیمانده‌ی بهبودیافته، طبقه‌بندی خودکار آریتمی‌ها را انجام می‌دهد.

در مرحله اول، از روش تقسیم‌بندی همپوشانی برای تقسیم سیگنال‌های «نوار قلب»[4] استفاده شده تا برای غلبه بر عدم تعادل بین کلاس‌ها، آن‌ها را به بخش‌های پنج ثانیه‌ای تقسیم کند. در مرحله‌ی دوم بخش‌های تقسیم شده‌ی سیگنال‌های نوار قلب دوباره برچسب‌گذاری می‌شوند. در مرحله‌ی سوم از «تبدیل موجک گسسته»[5] برای حذف نویز این بخش‌ها استفاده شده و شبکه عصبی پیچشی باقیمانده‌ی بهبودیافته برای طبقه‌بندی آریتمی استفاده می‌شود.

نتایج نشان دادند که مدل پیشنهادی سه عامل حساسیت ٪۹٤.٥٤، پیش‌بینی مثبت ٪۹۳.۳۳ و ویژگی ٪۸۰.۸ در بخش عادی است. دقت کلی روش هم ٪۸۸.۹۹ محاسبه شده است.

---

[1] Inception : مدل از پیش آموزش دیده گوگل که بیش از ۱۰۰۰ کلاس و بیش از ۱.٤ میلیون تصویر آموزش دیده است. این مدل، یک مدل تشخیص تصویر برای استخراج ویژگی با کمک شبکه‌های عصبی پیچشی است (جوشی و دیگران ۲۰۲۰، ۳).

[2] Li

[3] Qian

[4] Electrocardiography (ECG)

[5] discrete wavelet transform (DWT)

محقق نتیجه گرفته که مدل شبکه‌ی عصبی ارائه شده که مبتنی بر روش تقسیم‌بندی همپوشانی آموزش داده شده است، نتایجی قابل مقایسه با یک روش کلاسیک و سه روش پیشرفته دارد. علاوه بر این، در این تحقیق نشان داده شد که عملکرد شبکه‌ی عصبی ارائه شده در صورتی که از فقدان کانونی به عنوان تابع هزینه استفاده کند، بهبود می‌یابد.

تکنیک یادگیری دو فازی در شبکه‌ی عصبی مدولار برای طبقه‌بندی الگوی الفبای دست‌نویس هندی (سینگ[۱] و سینگ ۲۰۲۱، ۱۵-۱):

در این مقاله از شبکه‌ی عصبی مدولار به علت انعطاف‌پذیری آن، که از عدم وابستگی این رویکرد به تنها یک ساختار اساسی نشات می‌گیرد، در طبقه‌بندی الگوی الفبای دست‌نویس هندی استفاده شده است. در این مقاله ۲٤ شبکه عصبی فرعی مجزا برای محاسبات فاز اول در نظر گرفته شده و در فاز دوم خروجی‌های فاز اول به عنوان ورودی شبکه عصبی استفاده شده‌اند. در نهایت، خروجی فاز دوم نتایج طبقه‌بندی مدنظر را ارائه می‌دهد.

دو روش در این تحقیق استفاده شده است. اولی تلفیق قدرت اتصال به‌روز شده وزن‌ها با وزن جدید اختصاص داده شده است که میانگین ٪۹۸ داشته است. روش دوم فقط بر اساس قدرت اتصال به‌روز شده وزن‌ها است که میانگین دقت ٪۹۰ دارد.

در نهایت با توجه به یافته‌ها محقق نتیجه گرفته که، در حالیکه هر دو روش از دقت بالایی برخوردارند، روش اول دقیق‌تر است اما روش دوم در ازای کمی دقت پایین‌تر، کاراتر می‌باشد.

طبقه‌بندی اختلال طیف اوتیسم برای کودکان با استفاده از شبکه عصبی عمیق (موهانتی[۲]، پریداب[۳] و پاترا[٤] ۲۰۲۱، ٦-۱):

در این مقاله، محقق تلاش در جهت ارائه‌ی روشی برای شناسایی کودکان دچار اختلال طیف اوتیسم کرده است. شناسایی افراد دچار این اختلال در سنین پایین‌تر، مزیت‌هایی را در درمانشان به‌دنبال دارد.

برای دستیابی به هدف، محقق اپلیکیشنی توسعه داده که با بهره‌گیری از شبکه‌های عصبی بازگشتی به بررسی داده‌ها پرداخته و یک پیش‌بینی برای طبقه‌بندی کودکان مبتلا به اوتیسم ارائه می‌دهد.

---

[۱] Singh
[۲] Mohanty
[۳] Paridab
[٤] Patraa

یافته‌های این تحقیق بیانگر حساسیت ۱، تشخیص ۰.۹، خطای میانگین مربعات ۰.۲٤، خطای جذر میانگین مربعات ۰.٦، ضریب تعیین ۰.۷۵ و دقت ۰.۹٤ می‌باشد.

محقق از یافته‌ها نتیجه‌گرفته که روش پیشنهادی از نظر بالینی قابل قبول می‌باشد.

طبقه‌بندی تصاویر فراطیفی در شرایط ناکافی بودن نمونه و یادگیری ویژگی با استفاده از شبکه‌های عصبی عمیق (مقاله‌ی مروری) (وامبوگو[۱] و دیگران ۲۰۲۱، ۱۱-۱):

در این تحقیق، با توجه به اهمیت موضوع ۲٤ مدل شبکه‌های عصبی ارائه شده از سال ۲۰۱۷ تا ۲۰۲۱ بررسی شده‌اند.

نوع حسگر، تعداد کلاس‌ها، بازه‌ی طول موج، تعداد باندها و وضوح فضایی هر مدل مورد تجزیه و تحلیل قرار گرفته و در نهایت ضریب کاپا و میزان دقت کلی آن‌ها مقایسه شده است. در نهایت مشخص شد که مدل شبکه‌ی «هیبریداس‌ان»[۲] با دقت کلی ٪۹۹.۹۸ و ضریب کاپای ٪۹۹.۹۸ بهترین عملکرد کلی را داشته است. اما مدل‌های دیگری نیز بودند که عملکردی نزدیک به این میزان داشتند.

در نهایت محقق نتیجه گرفت که مدل‌های ارائه شده در این زمینه، بسیار اثربخش و کارا هستند که این مطلب کار را برای ارائه‌ی روش‌های جدید سخت می‌کند.

شبکه‌های عصبی موجک ابرگراف برای طبقه بندی اشیاء سه بعدی (نانگ[۳] و دیگران ۲۰۲۱، ۱٤-۱):

در این مقاله تلاش شده تا با ارائه‌ی مدلی جدید موسوم به شبکه‌های عصبی موجک ابرگراف، پاسخی برای عدم انعطاف‌پذیری شبکه‌های عصبی ابرگراف در مدل‌سازی و استخراج روابط مرتبه بالا میان داده‌ها، فراهم شود.

در این مقاله، روش پیشنهادی در ٤ مرحله توضیح داده شده است. در مرحله اول، نحوه‌ی ساخت ابرگراف توضیح داده شده است. در این مرحله برای ساخت ابرگراف، محقق الگوریتم «دورترین نقطه‌ی نمونه‌برداری»[٤] را با مدل «کوئری توپ»[٥] تلفیق کرده است. در مرحله‌ی دوم چگونگی دگرگونی‌سازی موجک ابرگراف طیفی را در راستای آماده کردن آن برای مراحل بعد توضیح داده است. در مرحله‌ی سوم، محقق درباره‌ی پیچش دادن

---

[۱] Wambugu
[۲] HybridSN
[۳] Nong
[٤] Farthest point sampling
[٥] ball query

موجک طیفی روی ابرگراف توضیح داده است. در مرحله‌ی آخر هم مطالب مربوط به آموزش شبکه‌ی عصبی و تنظیم‌کننده‌های ابرگراف توضیح داده شده‌اند.

آزمایشات انجام شده حکایت از دقت ٪۹۷.۸ در مجموعه داده «مدل‌نت۴۰»[۱] و دقت ٪۹۱.۳ در مجموعه داده‌ی دانشگاه ملی تایوان داشته است که کمی دقیق‌تر از مدل‌های جایگزین است.

محقق نتیجه گرفت که علی‌رغم دقت بالای مدل ارائه شده، ضعف آن در سرعت بسیار پایین‌ترش نسبت به برخی مدل‌های جایگزین مثل مدل «شبکه‌های عصبی ابرگراف»[۲] است. این درحالی است که دقت مدل ارائه شده نیز بسیار نزدیک به دقت برخی مدل‌های جایگزین سریع‌تر می‌باشد.

طبقه‌بندی دستگاه‌های سیستم کنترل صنعتی با استفاده از ویژگی‌های ترافیک شبکه و امبدینگ‌های شبکه عصبی (چاکرابورتی[۳]، کلی[۴] و گالاگر[۵] ۲۰۲۱، ۱۰-۱):

در این مقاله با توجه به اهمیت طبقه‌بندی دستگاه‌های «سیستم کنترل صنعتی (آی‌سی‌اس)»[۶] سایبری-فیزیکی نوین، در ارزیابی میزان امنیتشان مدلی ارائه شده که اولاً دستگاه‌های آی‌سی‌اس و غیر آی‌سی‌اس را جدا و دوماً انواع خاص دستگاه‌های آی‌سی‌اس را شناسایی کند.

در این مقاله دو مجموعه داده در نظر گرفته شده که میان آن‌ها جریان شبکه‌ای برقرار است. جریان شبکه‌ای بین دو مجموعه داده با نوع دستگاه، آدرس آی‌پی، شماره پورت و... ارتباط دارد. دو روش برای هدف اول ارائه شده است. یکی مبتنی بر آی‌پی‌تووک[۷] و دیگری روشی جدید بر مبنای ویژگی‌های ترافیک شبکه می‌باشد. برای هدف دوم از «نسخه‌ی سوم پروتکل شبکه‌ی توزیع شده»[۸] بهره گرفته است.

یافته‌ها نشان از دقت کلی ٪۷۵ برای روش مبتنی بر ویژگی‌های شبکه و دقت ٪۲۵ برای روش مبتنی بر آی‌پی‌تووک، در مورد هدف اول یعنی تشخیص دستگاه‌های آی‌سی‌اس از غیر آی‌سی‌اس دارد. برای هدف دوم

---

[۱] مدل‌نت۴۰ (ModelNet40) یک مجموعه داده از سری مجموعه داده‌های مدل نت پرینستون برای آزمایش مدل‌های یادگیری عمیق است که برای عموم در دسترس می‌باشد (https://modelnet.cs.princeton.edu).

[۲] Hypergraph neural networks

[۳] Chakraborty

[۴] Kelley

[۵] Gallagher

[۶] Industrial Control System (ICS)

[۷] آی‌پی‌تووک (IP2Vec) یک تکنیک برای شناسایی شباهت آی‌پی‌ها است(رینگ و دیگران ۲۰۱۷، ۱).

[۸] Distributed Network Protocol, version 3

یعنی شناسایی نوع خاص دستگاه‌های آی‌سی‌اس، با بهره‌گیری از هسته‌ی خطی، میزان دقت در هر دو مجموعه داده ٪۱۰۰ گزارش شده است.

محقق نتیجه گرفته که در مقایسه با روش‌هایی مثل شبکه‌های پیچشی نموداری[۱] مدل ارائه شده از لحاظ محاسباتی پیچیدگی کمتری داد. محدودیت روش یادشده زمانی است که پیام‌ها به صورت غیر رمز شده در دسترس نباشند (پیام‌های رد و بدلی بین شبکه‌ها رمز شده باشند).

طبقه‌بندی داده محور انواع زمین لغزش در مقیاس ملی با استفاده از شبکه‌های عصبی مصنوعی (آماتو[۲]، پالمبی[۳] و ریموندی[٤] ۲۰۲۱، ۱۰-۱):

از آنجا که طبقه‌بندی نوع زمین لغزه اهمیت زیادی در مدیریت ریسک دارد، در این مقاله این مبحث مورد بررسی قرار گرفته است. در این مقاله روش جدید داده محوری ارائه شده که با استفاده از داده‌های مکانی، که برای عموم در دسترس‌اند، به طبقه‌بندی نوع زمین لغزش‌های ایتالیا در مقیاس ملی مبتنی بر شبکه‌ی عصبی مصنوعی ساده‌ای می‌پردازد.

پنج دسته‌ی زمین لغزش در این تحقیق در نظر گرفته شده است. یافته‌های حاصله نشان دهنده‌ی دقت ٪۷٦ در آزمایش روی ۲۷۵ هزار زمین لغزش بوده می‌باشد.

محقق نتیجه گرفته که عملکرد روش ارائه شده در مقیاس ملی بسیار خوب است. محقق بیان کرده که از آنجا که شبکه‌های عصبی مصنوعی بدون اتکا به اطلاعات جزئی وابسته به مورد عمل می‌کنند، لذا روش ارائه شده قابلیت اجرا شدن روی هر جسم چند ضلعی دیگر را نیز دارد.

تشخیص و طبقه‌بندی خودکار کووید-۱۹ مبتنی بر اشعه ایکس و سی تی اسکن با استفاده از شبکه‌های عصبی پیچشی (تاکور[۵] و کومار[٦] ۲۰۲۱، ۷-۱):

در حالی که ویروس کووید-۱۹ میزان مرگ و میر پایینی دارد، اما مشکل این است که بسیار عفونی بوده و درنتیجه تعداد زیادی از افراد را مبتلا کرده است. بنابراین شناسایی زودهنگام کووید-۱۹ در بیماران بسیار حیاتی است. هدف این تحقیق استفاده از تصاویر اشعه ایکس و تصاویر توموگرافی کامپیوتری برای معرفی

---

[۱] Graph convolutional network
[۲] Amato
[۳] Palombi
[٤] Raimondi
[۵] Thakur
[٦] Kumar

یک استراتژی یادگیری عمیق مبتنی بر شبکه عصبی پیچشی است که کار شناسایی و طبقه‌بندی خودکار بیماری کووید-۱۹ را انجام دهد.

در این مقاله محقق با بهره‌گیری از ۳,۸۷۷ تصویر که ۱,۹۱۷ عدد از آن‌ها مربوط به افراد مبتلا به کووید-۱۹ و بقیه مربوط به افراد مبتلا به ذات الریه است، شبکه‌ی عصبی را آموزش داده است. برای اینکار ابتدا تصاویر پیش‌پردازش شدند تا برای استفاده شدن به عنوان داده‌های ورودی در شبکه‌ی عصبی پیچشی آماده باشند. در نهایت با بهره‌گیری از فرمول تعداد صحیح به کل، میزان دقت مدل ارائه شده محاسبه شده است.

یافته‌ها حکایت از دقت ٪۹۹.٦٤ مدل با امتیاز اف-یک معادل ٪۹۹.۵۹ دارد و منحنی مشخصه‌ی عملکرد سیستم ٪۱۰۰ می‌باشد.

محقق نتیجه گرفته که با توجه به اینکه ویروس کویید-۱۹ بسیار مسری است، مدل ارائه شده که دقت آن بسیار بالا است، می‌تواند در موارد اورژانسی کمک زیادی کند. هرچند علی‌رغم دقت و سرعت بالای مدل ارائه شده، این مدل محدودیت‌هایی نیز دارد. این مدل وابسته به زاویه‌ی تصویر برداری ایکس ری است و مثلاً در مواردی که تصویر به طور جانبی گرفته شده، این مدل عملکرد مناسبی نخواهد داشت. همچنین این مدل تصاویری که دارای چندین نشانه بیماری همراه با کووید-۱۹ هستند را نمی‌توان به‌درستی طبقه بندی کند. در ضمن این مدل باید روی مجموعه داده‌های بیشتری آزمایش شود تا بتوان به طور دقیق درباره‌ی عملکرد آن اظهار نظر کرد.

طبقه‌بندی باکیفیت تصاویر ورزشی با استفاده از نسخه‌ی سوم مدل اینسپشن و شبکه‌های عصبی (جوشی[1] و دیگران ۲۰۲۰، ۷-۱):

این مقاله یک مدل برای طبقه‌بندی تصاویر ورزشی بر اساس محیط اطراف ارائه کرده است. در این مقاله، از نسخه‌ی سوم مدل اینسپشن برای استخراج ویژگی‌ها و از شبکه‌های عصبی برای طبقه‌بندی دسته‌های مختلف ورزشی استفاده شده است. شش دسته‌ی مدنظر در این مقاله راگبی، تنیس، کریکت، بسکتبال، والیبال و بدمینتون بوده است.

در این مقاله، محقق چند ویدئوی مربوط به ورزش‌های ذکر شده را از وبسایت یوتیوب دریافت کرده و از ٪۷۰ از آن‌ها به عنوان ورودی‌هایی برای آموزش دادن شبکه‌ی عصبی استفاده کرده است. ٪۳۰ باقیمانده نیز به عنوان مجموعه داده‌ی آزمایش مورد استفاده قرار گرفته است. محقق ابتدا ویژگی‌ها مربوط به هر تصویر را با

[1] Joshi

بهره‌گیری از نسخه‌ی سوم مدل اینسپشن استخراج و سپس با بهره‌گیری از شبکه‌های عصبی پیچشی به طبقه‌بندی داده‌ها پرداخته است.

دقت مدل ارائه شده ٪۹۶.۶۴ اندازه‌گیری شده که نسبت به الگوریتم‌های جنگل تصادفی، کی-نزدیک‌ترین همسایه و ماشین بردار پشتیبانی، کمی دقت بالاتری داشته است.

محقق نتیجه گرفت که مدل ارائه شده از دقت و کارایی قابل قبولی برخوردار است و می‌توان از این مدل در زمینه‌ی تجزیه و تحلیل فعالیت و جستجو در ویدئوهای مربوط به این حوزه استفاده کرد.

لایه ورودی باکیفیت در شبکه‌های عصبی برای طبقه‌بندی فراطیفی داده‌ها با باندهای گمشده (فسنکت[۱] و دیگران ۲۰۲۰، ۵-۱):

به گفته‌ی محقق در مواردی که بخشی از طیف داده‌های تصویر در فرایند طبقه‌بندی فراطیفی مبتنی بر شبکه‌های عصبی مفقود است، به‌عنوان مثال زمانی که بخش‌هایی از تصویر بیش از حد نوردهی شده یا تحت تأثیر پیکسل‌های بی‌جا قرار گرفته‌اند، عملکرد روش‌های فعلی در طبقه‌بندی داده‌ها رضایت‌بخش نیست. تمرکز این مقاله روی ارائه‌ی روشی است که در این موارد عمل طبقه‌بندی را به‌خوبی انجام دهد. برای انجام این کار یک لایه ورودی با رویکرد تبدیل هار[۲] برای شبکه‌های عصبی مصنوعی که برای طبقه‌بندی داده‌های ابرطیفی استفاده می‌شود، ارائه می‌کند. این لایه ورودی برای عملکرد مؤثر با داده‌های ناقص طراحی شده است و مستقل از باندهای خاص استفاده شده توسط دوربین است.

در مرحله‌ی آموزش دادن برای اینکه شبکه‌ی عصبی راه‌های متعددی را برای شناسایی داده‌های مشابه بیاموزد، به طور تصادفی برخی از ضرایب در نظر گرفته نمی‌شود تا سایر ضرایب آن‌ها را جبران کنند. لذا در موارد مفقود بودن برخی داده‌ها، شبکه‌های عصبی عملکرد مناسبی خواهند داشت. در نهایت آموزش روی ٪۸۵ مجموعه داده‌ها، اعتبارسنجی با بهره‌گیری از ٪۱۰ مجموعه داده‌ها و آزمایش روی ٪۵ مجموعه داده‌ها انجام شده است.

یافته‌های مقاله نشان دادند که در شرایط کامل بودن داده‌ها دقت طبقه‌بندی ٪۹۸ و در صورت مفقود بودن ٪۳۰ از داده‌ها، دقت طبقه‌بندی ٪۷۰ است که نسبت به سایر روش‌ها دقیق‌تر است.

محقق نتیجه گرفت، لایه ورودی ارائه شده که مبتنی بر توابع هار است، نیازهای محاسباتی کمی دارد؛ بنابراین بسیار کاربردی است. لایه‌ی ورودی ارائه شده در مقایسه با روش‌های جایگزین (که در تحقیق بررسی

---

[۱] Fasnacht

[۲] Haar transform

شده‌اند) در مواردی که داده‌ها کامل است، دقتی مشابه دارد و در موارد مفقود بودن داده‌ها عملکردش بهتر است. همچنین اجرای روش ارائه شده ساده‌تر از روش‌های جایگزین است.

۲-۳-۳- مقایسه‌ی رویکرد مقالات ارزیابی شده

رویکردهای اصلی استفاده شده در روش پیشنهادی مقالات بررسی شده، در جدول ۲—۱ آورده شده‌اند.

(جدول ۲—۱) مقایسه‌ی رویکرد مقالات ارزیابی شده.

عنوان مقاله	رویکردهای اصلی استفاده شده در روش پیشنهادی
یک روش هوشمند برای طبقه بندی ترک در سازه‌های بتنی بر اساس شبکه‌های عصبی عمیق	شبکه‌های عصبی پیچشی
شناسایی موارد ابتلا به کووید-۱۹ با استفاده از شبکه‌ی عصبی فازی نوع ۲ عمیق براساس تصاویر X-Ray قفسه سینه	شبکه‌های عصبی پیچشی و توابع فازی نوع دو
تشخیص بیماران اسکیزوفرنی با استفاده از شبکه‌های عصبی پیچشی از تصاویر ارتباطات مؤثر مغزی سیگنال‌های چندکاناله الکتروانسفالوگرام	شبکه‌های عصبی پیچشی
طبقه‌بندی آریتمی بین بیماران متکی بر شبکه عصبی پیچشی باقیمانده‌ی عمیق بهبودیافته	شبکه عصبی پیچشی باقیمانده
تکنیک یادگیری دو فازی در شبکه‌ی عصبی مدولار برای طبقه‌بندی الگوی الفبای دست‌نویس هندی	شبکه‌های عصبی مدولار
طبقه‌بندی اختلال طیف اوتیسم برای کودکان با استفاده از شبکه عصبی عمیق	شبکه‌های عصبی بازگشتی
طبقه‌بندی تصاویر فراطیفی در شرایط ناکافی بودن نمونه و یادگیری ویژگی با استفاده از شبکه‌های عصبی عمیق	- (مقاله‌ی مروری)
شبکه‌های عصبی موجک ابرگراف برای طبقه بندی اشیاء سه بعدی	شبکه‌های عصبی ابرگراف، الگوریتم دورترین نقطه‌ی نمونه‌برداری و مدل کوئری توپ

طبقه‌بندی دستگاه‌های سیستم کنترل صنعتی با استفاده از ویژگی‌های ترافیک شبکه و جاسازی شبکه‌های عصبی	امبدینگ‌های شبکه‌ی عصبی، IP2Vec و نسخه‌ی سوم پروتکل شبکه‌ی توزیع شده
طبقه بندی داده محور انواع زمین لغزش در مقیاس ملی با استفاده از شبکه‌های عصبی مصنوعی	شبکه‌های عصبی پیشخور
تشخیص و طبقه‌بندی خودکار کووید-۱۹ مبتنی بر اشعه ایکس و سی تی اسکن با استفاده از شبکه‌های عصبی پیچشی (CNN)	شبکه‌های عصبی پیچشی
طبقه‌بندی باکیفیت تصاویر ورزشی با استفاده از نسخه‌ی سوم مدل Inception و شبکه‌های عصبی	نسخه‌ی سوم مدل Inception و شبکه‌های عصبی پیچشی
لایه ورودی باکیفیت در شبکه‌های عصبی برای طبقه‌بندی فراطیفی داده‌ها با باندهای گمشده	رویکرد تبدیل هار برای لایه‌های ورودی شبکه‌ی عصبی مصنوعی

۲-۳-٤- نتیجه‌گیری

تحقیق حاضر، رویکردی متفاوت و جدید نسبت به مقالات بررسی شده دارد. در این تحقیق، قصد بر ارائه‌ی مدلی جامع و کاربردی در زمینه‌ی طبقه‌بندی محتوا در سیستم مدیریت دانش است. این رویکرد، براساس نوین‌ترین فناوری‌های حوزه‌ی یادگیری عمیق می‌باشد که در سایر تحقیقات، کمتر به آن‌ها پرداخته شده است.

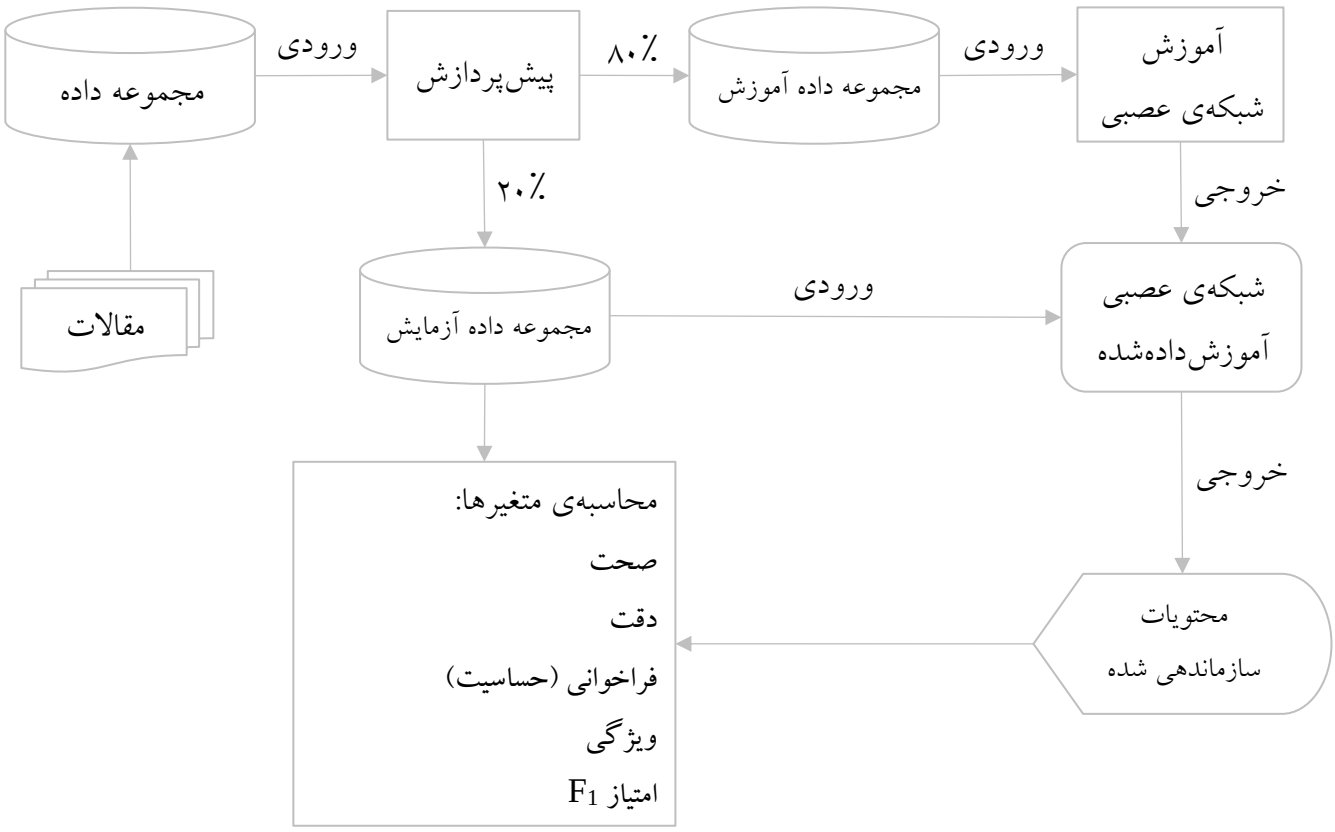
هم‌چنین تمام مراحل اجرای این تحقیق، از مرحله‌ی آماده‌سازی داده گرفته تا مراحل پیاده‌سازی مدل‌ها، به صورت متن‌باز در اختیار همگان قرار خواهد گرفت. لذا، از بابت جامع بودن و دسترس‌پذیر بودن شیوه‌های اجرا و پیاده‌سازی، این تحقیق تمایز ویژه‌ای نسبت به پژوهش‌های پیشین دارد.

علاوه بر موارد ذکر شده، یک مورد از تفاوت‌های مهم دیگر این تحقیق با پژوهش‌های پیشین، پیاده‌سازی شدن مدل‌های جایگزین مطرح و بررسی عملکرد آن‌ها است. لذا در این تحقیق، مدل‌های مطرح حوزه‌ی یادگیری ماشین ارزیابی و امکان مقاسیه‌ی آن‌ها فراهم می‌شود.

ضمناً، بررسی مدل‌های نوین حوزه‌ی یادگیری ماشین در سیستم‌های مدیریت دانش و با دید مدیریتی، مقوله‌ای است که کمتر به آن پرداخته شده است.

به این ترتیب، مطالب بیان شده در این تحقیق می‌تواند، هم در موارد کاربردی و هم در موارد پژوهشی، ارزش زیادی داشته باشد. این تحقیق حاوی اطلاعاتی ارزشمندی است که در پژوهش‌های پیشین، به این شکل به آن‌ها پرداخته نشده است.

شمای کلی مراحل اجرای تحقیق، در قالب یک مدل مفهومی در نمودار ۲–٤ قبل رؤیت است:



(نمودار ۲–٤) مدل مفهومی.

فصل سوم: روش اجرای تحقیق

- مقدمه
- نوع و روش تحقیق
- جامعه‌ی آماری و حجم نمونه

۳-۱- مقدمه

همانطور که اشاره شد، پژوهش حاضر، پژوهشی کاربردی است که هدف اصلی آن دستیابی به روشی کاربردی و باکیفیت مبتنی بر فناوری‌های روز، برای سازماندهی محتوا می‌باشد. در این فصل، روش دستیابی به این مدل و ابعاد اجرای تحقیق، تبیین خواهد شد.

پیاده‌سازی مدل‌های مدنظر، بر اساس هفت فاز اصلی داده‌کاوی در علم داده طی شده است. این هفت فاز در این فصل، مرحله به مرحله توضیح داده شده‌اند و شامل درک و مشخص‌سازی مسئله، آماده‌سازی داده، بررسی داده‌های آماده شده، راه‌اندازی، مدل‌سازی، ارزیابی و پیاده‌سازی و بازخورد می‌شوند.

در این فصل، بیان خواهد شد که چگونه داده‌های خام گردآوری و آماده می‌شوند. ابتدا چندین اقدام انجام می‌شود تا رکوردها، مناسب استفاده شدن به‌عنوان ورودی در مدل‌ها گردند؛ سپس مدل‌های مدنظر پیاده‌سازی می‌شوند.

مدل‌های پیاده‌سازی شده در این تحقیق، شامل کی‌ان‌ان، بیز ساده، رگرسیون لجستیک، اس‌وی‌ام، اس‌وی‌ام هسته‌ای، درخت تصمیم، جنگل تصادفی، شبکه‌ی عصبی پیشخور و شبکه‌ی عصبی بازگشتی هستند. در این فصل فرایند پیاده‌سازی این مدل‌ها تبیین شده است.

در نهایت، معیارهای اندازه‌گیری عملکرد مدل، معرفی خواهند شد. این معیارها شامل صحت، دقت، حساسیت یا بازیابی، ویژگی و امتیازهای اف می‌شوند. در فصل چهارم از این معیارها استفاده می‌شود تا نتایج حاصله مشخص گردند.

لازم به ذکر است، اطلاعات مربوط به چگونگی دسترسی به کد منبع، داده‌ها ورودی و خروجی و سایر فایل‌های مرتبط با فرایند انجام پروژه، در پیوست ۱ آورده شده است.

۳-۲- نوع و روش تحقیق

اجرای روش مدنظر بر اساس هفت فاز متدلوژی علم داده و مبتنی بر شرح لاروس و لاروس (۲۰۱۹، ۲)، انجام می‌شود. در ادامه، روش تحقیق به تفکیک هر یک از این فازها بیان شده است:

۳-۲-۱- فاز درک و مشخص‌سازی مسئله

در این فاز، قصد بر این است که ضمن شفاف‌سازی مسئله، اهداف اجرایی را نیز مشخص کنیم. هدف اجرایی اصلی در مسیر دستیابی به روش مدنظر، سازماندهی کارا و اثربخش محتویات ورودی به سیستم مدیریت دانش سازمانی می‌باشد. برای اینکار ابتدا باید مسئله را به‌گونه‌ای شفاف، مشخص سازیم.

در فرایند مدیریت دانش سازمانی، سیستم مدیریت دانش، همواره ورودی‌هایی را دریافت می‌کند. این ورودی‌ها به شکل کلماتی در قالب جملات، به سیستم وارد می‌شوند. جملات مرتبط نیز در پاراگراف‌هایی گنجانده می‌شوند که هر پاراگراف، موضوع خاصی را دنبال می‌کند. بر این اساس، مسئله‌ای مدنظر، طبقه‌بندی مناسب این پاراگراف‌ها در کلاس‌های مربوطه است. برای این منظور ابتدا باید تعدادی از پاراگراف‌ها توسط انسان طبقه‌بندی شوند تا مدل مدنظر، الگوهای نهان را شناسایی کرده و بر اساس آن، عمل یادگیری را انجام دهد. به عبارت دیگر، هدف اصلی این است که با اتکا بر رکوردهای مجموعه داده‌ی آموزش، عمل یادگیری مدل مدنظر به‌گونه‌ای انجام شود که بتواند با بالاترین دقت ممکن، در فرایندی بهینه و مقیاس‌پذیر، کلاس صحیح مرتبط به هر رکورد موجود در مجموعه داده‌ی آزمایش را پیش‌بینی کند.

۳-۲-۲- فاز آماده‌سازی داده‌ها

این فاز در دو بخش انجام می‌گیرد. ابتدا داده‌های خام جمع‌آوری شده و سپس این داده‌ها «تمیز»[۱] می‌شوند. در ادامه اقدامات انجام شده در این دو بخش بیان خواهند شد:

۳-۲-۲-۱- جمع‌آوری داده‌ها

برای دستیابی به نتایجی که قابلیت تعمیم و اتکای بالایی را داشته باشند؛ تصمیم گرفته شده تا از داده‌هایی استفاده شود که هم‌اکنون در حال استفاده شدن در دنیای واقعی هستند. به این منظور، داده‌ها را از مقالات وبسایت ویکی‌پدیا استخراج و جمع‌آوری کردیم. دلایل متعددی برای انجام اینکار وجود داشت؛ از جمله:

۱. تا دوم آذرماه ۱۴۰۱، تعداد صفحات ویکی‌پدیا انگلیسی زبان ۵۶,۹۸۰,۳۲۸ اندازه‌گیری شد که از این تعداد ۶,۵۷۹,۰۸۰ (٪۱۱.۵۵) مقاله هستند (ویکی‌پدیا[۲]، ۲۰۲۲). درنتیجه، این وبسایت، تعداد قابل توجهی داده‌ی بالقوه در اختیار قرار می‌دهد.

۲. ویکی‌پدیا خود یک سیستم مدیریت دانش است (دیویس ۲۰۲۰، ۴۳۹). در ماه اکتبر ۲۰۲۲، تعداد بازدید از صفحات این وبسایت بیش از ۲۴ میلیارد دفعه گزارش شده است. در همین بازه‌ی زمانی، صفحات ویکی‌پدیا

[۱] Data cleansing

[۲] https://www.wikipedia.org/

بیش از ۴۸ میلیون بار ویرایش شده‌اند (آمارهای ویکی‌پدیا[۱]، ۲۰۲۲). لذا، داده‌های این وبسایت نمونه‌ای از داده‌های مورد استفاده در یک سیستم مدیریت دانش محبوب، در دنیای واقعی هستند و نمونه‌ی آماری نابی را برای این تحقیق فراهم می‌آورند.

۳. یکی از خواسته‌های مدنظر در این تحقیق، مقایسه‌ی دقت مدل در سازماندهی ورودی‌های فارسی و انگلیسی زبان است. از آنجا که تعداد زیادی از مقالات ویکی‌پدیا در هر دو زبان فارسی و انگلیسی در دسترس‌اند؛ درنتیجه می‌توان با بهره‌گیری از داده‌های این مقالات، مقایسه‌ای مناسب میان دقت مدل در سازماندهی محتویات فارسی و انگلیسی انجام داد.

از آنجا که جمع‌آوری اطلاعات به صورت تصادفی و عاری از تأثیر قضاوت‌های نگارنده، می‌تواند نتایج قابل اتکاتری را به همراه داشته باشد؛ درنتیجه یک برنامه‌ی کامپیوتری «گزینش‌گر» پیاده‌سازی شد که داده‌ها را به صورت خودکار برگزیند. این برنامه به زبان سی‌شارپ[۲]، مبتنی بر نسخه‌ی ٦ چارچوب نرم‌افزاری دات‌نت[۳] و به صورت متن‌باز، توسعه داده شده است. اطلاعات مربوط به چگونگی دسترسی به این برنامه، در «پیوست ۱» آورده شده است. این برنامه، قابلیت سفارشی‌سازی بالایی داشته و برای سایر داده‌کاوان نیز، زمینه‌ی مناسبی در جهت فراهم کردن نمونه‌ی آماری مناسب، ایجاد می‌کند.

همانطور که اشاره شد، فرایند انتخاب داده‌ها در این تحقیق، تصادفی انجام می‌شود. با این وجود، جهت دستیابی به داده‌هایی باکیفیت‌تر، شاخص‌هایی در برنامه‌ی گزینش‌گر تعریف شدند تا با فیلتر کردن داده‌های بالقوه، خروجی‌های مناسب‌تری را در اختیار قرار دهد. از جمله‌ی این شاخص‌ها، حداقل تعداد کلمات در پاراگراف بود که روی عدد ۲۵ تنظیم شد. این شاخص لحاظ شد تا از مرتبط بودن پاراگرف مدنظر به کلاس مربوطه، اطمینان حاصل شود.

در ضمن، از آنجا که یکی از اهداف مهم در این تحقیق، ارائه دادن مدلی کاربردی است؛ لذا استفاده از داده‌هایی که منعکس‌کننده‌ی محتویات در دنیای واقعی باشند، ضرورت دارد. بنابراین شاخصی در برنامه‌ی گزینش‌گر تعریف شد که بجای قرار دادن تعداد یکسانی از رکوردها در همه‌ی کلاس‌ها، با توجه به تعداد مقالات بالقوه‌ی مربوط به هر کلاس، تعداد رکوردهای آن را تعیین کند. این کار منجر می‌شود که کلاس‌های تحقیق ما، همچون کلاس‌های موجود در دنیای واقعی، ناهمگن شوند. این اقدام شاید چالش‌هایی جدید برای

[۱] https://stats.wikimedia.org/

[۲] C# 10

[۳] .Net 6

مدل‌های مدنظر به وجود آورد؛ اما در نهایت نتایج بدست آمده برای تعمیم به دنیای واقعی، قابل اتکاتر خواهند بود.

برنامه‌ی گزینش‌گر، علاوه بر فیلتر کردن داده‌های بالقوه، برخی از لغات و علایم زاید موجود در پاراگرف‌ها را نیز حذف می‌کند که در بخش تمیزسازی داده‌ها به آن اشاره خواهد شد.

۳-۲-۲-۲- تمیزسازی داده‌ها

در این فاز، اقداماتی روی داده‌های خام انجام می‌شود تا آن‌ها آماده‌ی استفاده شدن در مدل شوند. به بیان لاروس و لاروس (۲۰۱۹، ٤)، «تمیز کردن» و آماده سازی داده‌ها احتمالاً پر زحمت‌ترین مرحله‌ی کل فرآیند علم داده است. در این فاز، مراحل زیر طی می‌شود:

۳-۲-۲-۲-۱- حذف علایم، حروف و لغات زاید

پاراگراف‌های مقالات ویکی‌پدیا، داده‌های خام را تشکیل می‌دهند. از آنجا که این داده‌ها در قالب کدهای اچ‌تی‌ام‌ال[۱] هستند؛ درنتیجه حاوی برخی تگ‌های زاید می‌باشند. این تگ‌ها، کیفیت فرایند یادگیری مدل‌ها را کاهش می‌دهند. درنتیجه، برنامه‌ی گزینش‌گر به گونه‌ای پیاده‌سازی شد تا تگ‌های زاید اچ‌تی‌ام‌ال را حذف نموده و پاراگراف‌ها را جداگانه و به شکل مطلوب استخراج نماید.

برای انجام اقدامات آتی، از زبان برنامه‌نویس پایتون و کتابخانه‌های این زبان استفاده خواهد شد. اجرای برنامه‌های نوشته شده، در پلتفرم گوگل کولب انجام می‌گیرد.

پس از استخراج پاراگراف‌ها در فرمت مطلوب، حال نوبت به حذف واژه هایی می‌رسد که در مسیر انجام عمل سازماندهی، نه تنها مفید نخواهند بود بلکه ممکن است گمراه کننده باشند. به این نوع واژه‌ها اصطلاحاً «واژه‌های ایستاده‌شده»[۲] یا «واژه‌های پالایشی» گویند. این واژه‌ها متناسب با نوع مسئله و داده‌ها، در هر تحقیق متغیرند. این واژه‌های شامل کلمات ربط مثل «و»، «به»، «با» و... می‌شوند. هرچند این واژه‌ها نقشی کلیدی برای تفهیم منظور در ارتباطات میان انسانی دارند اما با توجه به ماهیت عمل الگویابی در مدل‌های یادگیری ماشین و همچنین یادگیری عمیق، این کلمات تأثیر به‌سزایی در دستیابی به هدف نداشته و حتی ممکن است کیفیت الگویابی مدل‌ها را نیز پایین آورند.

---

[۱] HTML

[۲] Stop Words

۳-۲-۲-۲-۲- استانداردسازی داده‌ها

در این مرحله از تمیزسازی داده‌ها، چندین اقدام مهم صورت می‌گیرد. ابتدا فاصله‌های اضافی میان کلمات حذف می‌شود و از نیم‌فاصله در جاهای مناسب استفاده می‌شود. این کار، که اصطلاحاً «عادی سازی»[۱] نام دارد، به مدل‌های یادگیری ماشین مدنظر کمک می‌کند تا عمل ارتباط‌یابی میان کلمات را، به‌خصوص در مورد کلمات فارسی، بهتر انجام دهند.

پس از نشانه‌گذاری رکوردها، در قدم بعدی، عمل «لماتی»[۲] کلمات انجام می‌گیرد. در این قدم، واژه‌ها به بُن‌شان بازگردانده می‌شوند. برای مثال واژه‌ی «درخت‌ها» به «درخت» و واژه‌ی «می‌روم» به «رفت–رو» تبدیل می‌شود. تفاوت لماتی کردن با «ریشه‌یابی»[۳] آن است که در فرایند لماتی کردن، معنا و موقعیت کلمه در کنار سایر کلمات در جمله، در نظر گرفته می‌شود. به این ترتیب با لماتی کردن واژه‌ها، زمینه‌ای برای گروه‌بندی مناسب‌تر و درنتیجه‌ی آن ارتباط‌یابی بهتر میان کلمات، فراهم می‌شود.

با انجام اقدامات فوق، اکنون رکوردهای موجود به شکلی استاندارد درآمده و برای بررسی مناسب شده‌اند. با وجود اقدامات انجام شده، هنوز داده‌ها به طور کامل برای استفاده به عنوان ورودی در مدل‌های مدنظر مناسب نیستند و برخی اقدامات دیگر باید انجام گیرد.

۳-۲-۳- فاز بررسی داده‌های آماده شده

لازم به ذکر است، در این تحقیق تلاش شده که مدل ارائه شده تعمیم‌پذیر بوده و در سایر زمینه‌ها نیز عملکرد مناسبی داشته باشد. داده‌های فعلی یک مجموعه داده‌ی نمونه‌اند که مدل مدنظر روی آن‌ها آزمایش می‌شود تا کیفیت آن در دستیابی به اهداف مشخص گردد. پس مدل مدنظر، نباید بر اساس عواملی جزئی که صرفا منحصر به داده‌های فعلی است، ارائه شود.

در این تحقیق ۱۷ کلاس اصلی برای رکوردها در نظر گرفته شده است. رکوردها در دو مجموعه داده مجزا، یکی به زبان فارسی و دیگری به زبان انگلیسی قرار دارند. کلاس‌های متناظر هر یک از این رکوردها، در یکی از ستون‌های مجموعه داده مشخص شده‌اند.

در هر مجموعه داده سه ستون شامل آی‌دی، محتوا و برچسب وجود دارد. هر ردیف شامل اطلاعات مربوط به یک رکورد است. در جدول ۳–۱ نمونه‌ای از یک رکورد آورده شده است:

---

[۱] Normalizing
[۲] Lemmatizing
[۳] Stemming

(جدول ۳–۱) نمونه‌ای از یک رکورد، پس از فاز آماده‌سازی داده‌ها.

آی‌دی	محتوا	برچسب
۳	علم شیمی موضوعات چگونگی برهم‌کنش اتم مولکول طریق پیوند شیمیایی تشکیل ترکیبات شیمیایی پرداخت#پرداز پیوند شیمیایی داشت#دار ترکیبات مختلف دارا #هست پیوند کووالانسی پیوند یونی پیوند هیدروژن واندروالسی پیوند فلز	۰

با توجه به مطالب ذکر شده و نوع مسئله، می‌بایست از مدل‌های یادگیری نظارت شده برای دستیابی به هدف مدنظر بهره‌برد. همانطور که اشاره شده، مدل‌های مختلفی در زمینه‌ی یادگیری ماشین به شیوه‌ی یادگیری نظارت شده وجود دارند. جهت ارائه‌ی مدلی باکیفیت‌تر در راستای هدف، مدل‌های مبتنی بر یادگیری عمیق، مدل‌هایی مناسب‌تر به شمار می‌آیند. با بهره‌گیری از مدل شبکه‌های عصبی مصنوعی که در کتابخانه‌ی نرم‌افزار تنسورفلو ارائه شده‌اند، در جهت دقیق و بهینه‌تر سازی عمل سازماندهی محتوا، کوشش شده است.

در مسیر اجرای مدل مدنظر، مدل‌ها و رویکردهای سنتی نیز بررسی خواهند شد تا شاخصی مناسب برای مقایسه‌ی مدل مدنظر با مدل‌های سنتی بدست آید.

۳–۲–٤– فاز راه‌اندازی

در این فاز اقدامات انتهایی لازم برای انجام فرایند مدل‌سازی انجام می‌گیرد. اقدامات مراحل قبلی منجر به آماده سازی داده‌ها به شکلی مطلوب و عاری از محتویات زاید شده‌اند. اکنون می‌بایست برای بهینه‌سازی استفاده از منابع سخت‌افزاری و کارآمدتر کردن الگوریتم‌های مدل‌های مدنظر، رکوردها را در فرمت مناسب قرار دهیم.

۳–۲–٤–۱– ساخت «بسته‌ی کلمات»[۱]

بسته‌ی کلمات در حقیقت یک «آرایه»[۲] از اعداد صحیح است که به ازای هر پاراگراف ساخته می‌شود و محتویات آن را برای محاسبات کامپیوتری، بهینه‌تر می‌سازد. برای ساخت بسته‌ی کلمات لازم است، تعداد واژه‌های مدنظر برای بررسی را محدود سازیم. اینکار بر اساس دفعات تکرار کلمات انجام می‌شود. انتخاب میزان محدودسازی برای هر تحقیق متغیر است اما معمولاً، برای دستیابی به دقتی قابل قبول، محدود کردن کلمات به ده هزار عدد، مناسب است. پس از مشخص کردن میزان محدودسازی، برای هر پاراگراف یک آرایه با اندازه‌ای برابر این میزان ساخته می‌شود که هر عضو آن نمایانگر یکی از کلمات پرتکرار است. مقدار هر عضو آرایه به صورت پیش‌فرض برابر عدد صفر است. در ادامه کلمات هر پاراگراف بررسی می‌شوند تا

[۱] Bag of Words
[۲] Array

درصورتیکه جزو کلمات پرتکرار مدنظر باشند، مقدار جایگاه مربوط به آن کلمه، در آرایه‌ی مربوط به آن پاراگراف، به اندازه‌ی یک واحد افزایش داده شود. برای مثال، اگر جملات زیر رکوردهای ورودی باشند:

- علی از تماشای تلویزیون لذت می‌برد. احمد نیز از تماشای تلویزیون لذت می‌برد.
- احمد از فوتبال بازی کردن، لذت می‌برد.

در شکل ۳–۱، بسته‌ی کلمات برای رکورد اول نشان داده شده است.

علی	از	تماشا	تلویزیون	لذت	می‌برد	احمد	نیز	فوتبال	بازی	کردن
۱	۲	۲	۲	۲	۲	۱	۱	۰	۰	۰

(شکل ۳–۱) نمونه‌ای از بسته‌ی کلمات.

همانطور که مشاهده می‌شود، در بسته‌ی کلمات، دنباله‌ی کلمات به هم می‌خورد. با توجه به ماهیت مدل‌های یادگیری ماشین مدنظر، دنباله‌ی کلمات، اهمیت چندانی ندارد. در این مدل‌ها تلاش بر این است که الگوی موجود میان کلمات، فارق از مکان آن‌ها در یک جمله، شناسایی شود. درنتیجه، استفاده از بسته‌ی کلمات رویکردی مناسب برای آموزش مدل‌های مدنظر است.

۳–۲–٤–۲– تبدیل برچسب‌ها به آرایه‌های عددی

در مدل‌های یادگیری عمیق، به‌خصوص مدل‌های شبکه‌ی عصبی مصنوعی، بهتر است برچسب‌ها به شکل آرایه‌ای از اعداد درآیند. با این کار، مدل‌ها، بهتر متوجه ماهیت برچسب‌های طبقه‌بندی‌کننده می‌شوند. برای این منظور از روشی موسوم به «رمزگذاری وان–هات»[۱] استفاده شده است. در این روش برای هر برچسب، یک آرایه‌ی متناظر در نظر گرفته می‌شود که اندازه‌ی آن، برابر تعداد کل برچسب‌ها موجود است. تنها مقدار یکی از اعضای آرایه‌ی متناظر هر برچسب برابر یک است و مقدار مابقی اعضا، صفر می‌باشد. مثلاً برای برچسب اول که عنوان آن «شیمی» است، آرایه‌ی متناظر [۱, ۰, ۰, ۰, ۰, ۰, ۰, ۰, ۰, ۰, ۰, ۰, ۰, ۰, ۰, ۰, ۰, ۰, ۰, ۰] در نظر گرفته می‌شود.

۳–۲–٤–۳– تقسیم مجموعه داده به مجموعه داده‌ی آموزش و مجموعه داده‌ی آزمایش

در این مرحله که آخرین قدم در راستای فراهم کردن شرایط مناسب برای مدل‌سازی می‌باشد، مجموعه داده را به دو دسته‌ی آزمایش و آموزش تقسیم می‌کنیم. این فرایند می‌بایست کاملاً تصادفی انجام شود تا نتایج

[۱] One-Hot Encoding

بدست آمده اعتبار داشته باشند. به این منظور، از یکی از توابع پر استفاده‌ی کتابخانه‌ی نرم‌افزار سایکیت–لرن استفاده شده تا تقسیم تصادفی مجموعه داده‌ها، به شکلی مناسب انجام گیرد.

در مسائل مربوط به یادگیری ماشین مرسوم است که ۲۰ تا ۳۰ درصد داده‌ها را به عنوان مجموعه داده آزمایش در نظر گرفته و از باقی داده‌ها برای آموزش مدل استفاده شود. طبق روال مرسوم، اندازه‌ی مجموعه‌ی آزمایش ۲۰ درصد اندازه‌ی کل مجموعه داده در نظر گرفته شد. لازم به ذکر است که این فرایند برای هر دو مجموعه داده‌ی فارسی و انگلیسی و به طور جداگانه انجام می‌گیرد.

با انجام اقدامات فوق، اکنون شرایط برای انجام مدل‌سازی فراهم شده است. در پیوست ۲، نمونه‌ای از خروجی حاصل از این مرحله آورده شده است.

۳–۲–۵– فاز مدل‌سازی

در این فاز، رکوردهای مجموعه داده‌ی آموزش در اختیار مدل‌های مدنظر قرار گرفته و آن‌ها به‌گونه‌ای تنظیم می‌شوند تا به بهترین نحو ممکن، عمل یادگیری را انجام دهند. در این فاز، ابتدا مدل‌های سنتی آموزش داده می‌شوند و سپس مدل‌های شبکه‌ی عصبی مصنوعی برای عمل یادگیری تنظیم می‌شوند.

۳–۲–۵–۱– مدل کی‌ان‌ان

همانطور که اشاره شده، کی‌ان‌ان بر مبنای الگوریتمی بسیار ساده است. بروس، بروس و گدک (۲۰۲۰، ۲۳۸) برای طبقه‌بندی یا پیش‌بینی یک رکورد با این روش، مراحل کلی زیر را بیان کرده‌اند:

۱. پیدا کردن رکوردهایی که ویژگی‌های مشابهی دارند (مثلاً پیش‌بینی‌کننده‌های مشابه).

۲. برای طبقه‌بندی، باید کلاس اکثریت در بین آن رکوردهای مشابه را یافت و آن کلاس را به رکورد جدید اختصاص داد.

۳. در فرایند رگرسیون، مقدار پیش‌بینی برابر با میانگین رکوردهای مشابه خواهد بود.

برای پیاده‌سازی این مدل، از تابع طبقه‌بندی‌کننده‌ی نزدیکترین همسایه‌ی ارائه شده در کتابخانه‌ی نرم‌افزار سایکیت–لرن استفاده شده است. تعداد همسایه‌های مورد بررسی برای هر همسایه را عدد پنج در نظر گرفتیم. برای محاسبه فاصله نیز از «متریک مینکوفسکی»[۱] استفاده کرده و متغیر قدرت برای متریک مینکوفسکی را عدد دو در نظر گرفتیم.

---

[۱] «فاصله مینکوفسکی» یا «متریک مینکوفسکی» (Minkowski Metric) یک متریک در فضای برداری نُرم‌دار است که می‌توان آن را به عنوان تعمیم فاصله اقلیدسی و فاصله‌ی منهتن در نظر گرفت. نام این متریک، به افتخار ریاضیدان آلمانی هرمان مینکوفسکی گذاشته شده است (راشکا[۱]، لیو[۱] و میرجلیلی ۲۰۲۲، ۱۰۱).

۳-۲-۵-۲- مدل بیز ساده

برای انجام فرایند طبقه‌بندی به این روش کافی است بعد از محاسبه‌ی توزیع احتمال پسین، رکورد مدنظر را به کلاس دارای بیشترین احتمال، اختصاص داد. به طور کلی می‌توان مراحل زیر را برا طبقه‌بندی‌کننده‌ی بیز ساده متصور بود (بروس، بروس و گدک ۲۰۲۰، ۱۹۸):

۱. ابتدا برای پاسخ دودویی $Y = y^*$، احتمالات شرطی فردی برای هر پیش‌بینی‌کننده محاسبه می‌شود؛ یعنی $p(X^*|Y = y^*)$. این احتمال با محاسبه‌ی نسبت مقادیر X^* در میان رکوردهای $Y = y^*$ در مجموعه‌ی آموزشی، تخمین زده می‌شود.

۲. احتمالات بدست آمده از مرحله‌ی یک در یکدیگر ضرب شده و سپس در نسبت رکوردهای متعلق به $Y = y^*$ ضرب می‌شوند.

۳. مراحل یک و دو برای همه کلاس‌ها تکرار می‌شوند.

٤. با گرفتن مقدار محاسبه شده در مرحله دو برای کلاس y^* و تقسیم آن بر مجموع احتمالات مربوط به سایر کلاس‌ها، یک احتمال برای نتیجه y^* برآورد می‌شود.

۵. در نهایت، رکورد مدنظر به کلاسی با بیشترین احتمال حاصله از این مجموعه مقادیر پیش‌بینی‌کننده، اختصاص داده می‌شود.

برای پیاده‌سازی این مدل از تابع «بیز ساده‌ی گاوسی»[۱] کتابخانه‌ی نرم‌افزار سایکیت-لرن، بهره بردیم.

۳-۲-۵-۳- مدل رگرسیون لجستیک

همانطور که جرون (۲۰۱۹، ۲۰٦) اشاره کرده، در مدل رگرسیون لجستیک، برای هر کلاس عمل پیش‌بینی به طور جداگانه انجام می‌شود. اگر عدد پیش‌بینی شده برای یک برچسب مساوی یا بیش از ۵۰ درصد باشد؛ آنگاه آن برچسب به عنوان برچسب متناظر انتخاب می‌شود. این مطلب در رابطه‌ی ۳-۱ نمایش داده شده است.

$$y = \begin{cases} 0 \; if \; p < 0.5 \\ 1 \; if \; p \geq 0.5 \end{cases} \quad (۳-۱)$$

از آنجا که در این تحقیق برای هر رکورد فقط یک کلاس مورد قبول در نظر گرفته شده است؛ درنتیجه کلاسی که نسبت به سایر کلاس‌ها بالاترین مقدار پیش‌بینی شده را به خود اختصاص می‌دهد، به‌عنوان کلاس متناظر، انتخاب می‌شود.

---

[۱] Gaussian Naive Bayes

برای پیاده‌سازی مدل رگرسیون لجستیک نیز از تابع مربوطه‌ی ارائه شده کتابخانه‌ی نرم‌افزار سایکیت–لرن استفاده می‌کنیم. آموزش مدل با رویکرد چند جمله‌ای انجام شده و از الگوریتم «ال–بی‌اف‌جی‌اس»[1] به عنوان حل‌کننده استفاده شده است.

۳–۲–۵–٤– مدل اس‌وی‌ام

همانطور که اشاره شد، اس‌وی‌ام با یادگیری مرزهای تصمیم‌گیری خطی یا غیرخطی، کلاس‌های متفاوت را از هم جدا می‌کند. در اس‌وی‌ام از «ابرصفحه‌های جداکننده»[2] برای تعیین مرز میان کلاس‌ها و درنتیجه طبقه‌بندی آن‌ها استفاده می‌شود. مسلماً ابرصفحه‌های زیادی وجود دارند که ممکن است داده‌ها را طبقه‌بندی کنند. از میان این ابرصفحه‌های بالقوه، انتخاب ابرصفحه‌ای که بیشترین جدایی یا حاشیه را بین دو کلاس ایجاد کند، انتخابی معقول است. بنابراین ابرصفحه باید طوری انتخاب شود که فاصله آن تا نزدیکترین نقطه داده در هر طرف به حداکثر برسد. اگر چنین ابرصفحه‌ای وجود داشته باشد، به عنوان «ابرصفحه حداکثر حاشیه»[3] شناخته شده و طبقه‌بندی‌کننده‌ی مبتنی بر آن، به عنوان «طبقه‌بندی‌کننده‌ی حداکثر حاشیه»[4] شناخته می‌شود (جیمز و دیگران ۲۰۲۱، ۳۷۱–۳٦۹).

برای پیاده‌سازی این مدل، از تابع اس‌وی‌ام موجود در کتابخانه‌ی سایکیت–لرن استفاده شده است و از نوع «خطی» هسته در الگوریتم استفاده شده است.

۳–۲–۵–۵– مدل اس‌وی‌ام هسته‌ای

رویکرد هسته‌ای این امکان را فراهم می‌کند که فضای ویژگی خود را بزرگ کنیم تا یک مرز غیر خطی بین کلاس‌ها ایجاد شود. ماشین‌های هسته‌ای، دسته‌ای از الگوریتم‌ها برای تحلیل الگو هستند که در زمینه‌های خوشه‌بندی، طبقه‌بندی، رتبه‌بندی و... کاربرد دارند (جیمز و دیگران ۲۰۲۱، ۳۸۰).

---

[1] ال–بی‌اف‌جی‌اس (Limited Memory-BFGS) یک الگوریتم بهینه‌سازی در خانواده روش‌های شبه نیوتنی است که تقریب الگوریتم برودن–فلچر–گلدفارب–شانو (BFGS) را با استفاده از مقدار محدودی از حافظه‌ی رایانه محاسبه می‌کند. این یک الگوریتم محبوب برای تخمین پارامتر، در مدل‌های یادگیری ماشین است (جرون ۲۰۱۹، ۲۱۵).

[2] ابرصفحه‌های جداکننده (Separating Hyperplane) در یک فضای p–بعدی، یک زیرفضای همبسته مسطح به بعد $p - 1$ است (جیمز و دیگران ۲۰۲۱، ۳٦۸).

[3] Maximum-Margin Hyperplane

[4] Maximum-Margin Classifier

در این مدل همه‌ی مراحل مشابه مدل اس‌وی‌ام طی می‌شود اما به‌جای رویکرد خطی، از هسته‌ی «تابع پایه شعاعی»[1] یا به اختصار هسته‌ی «آربی‌اف» استفاده می‌شود. این کار باعث می‌شود که مدل ارائه شده برای طبقه‌بندی چندگانه (که ماهیت کار در این تحقیق است) مناسب‌تر شود.

۳-۲-۵-٦- مدل درخت تصمیم

همانطور که اشاره شد، برای دستیابی به بهینه‌سازی مدل درخت تصمیم، باید گره‌ها در ویژگی‌هایی که بیشترین اطلاعات را به‌دنبال دارند، تقسیم شوند. راشکا، لیو و میرجلیلی (۲۰۲۲، ۸۸)، تابع هدف کلی درخت تصمیم برای حداکثر سازی کسب اطلاعات را به شکل رابطه‌ی ۳–۲ تعریف کرده‌اند.

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j)$$

(۲–۳)

f : ویژگی مدنظر برای انجام تقسیم　　D_p : مجموعه داده‌های گره والد

D_j : مجموعه داده‌های فرزند j ام　　I : معیار ناخالصی[2]

N_p : تعداد کل نمونه‌های آموزشی در گره والد　　N_j : تعداد کل نمونه‌های آموزشی در گره والد

همانطور که مشخص است، بدست آوردن میزان دستیابی اطلاعات برای هر ویژگی، به سادگی محاسبه‌ی تفاوت میان ناخالصی گره والد و مجموع ناخالصی‌های گره فرزند است. هر چه ناخالصی‌های گره‌های فرزند کمتر باشد، میزان دستیابی اطلاعات بیشتر می‌شود. با این حال، برای سادگی و کاهش فضای جستجوی ترکیبی، اکثر کتابخانه‌ها (از جمله سایکیت–لرن) درخت‌های تصمیم دودویی را پیاده‌سازی می‌کنند (راشکا، لیو و میرجلیلی ۲۰۲۲، ۸۸).

برای ساخت این مدل نیز از تابع طبقه‌بندی‌کننده‌ی کتابخانه‌ی سایکیت–لرن استفاده شده و برای محاسبه‌ی ناخالصی (شاخص جداسازی)، از «ناخالصی جینی»[3] استفاده شده است.

---

[1] تابع پایه شعاعی (Radial Basis Function یا RBF) یک تابع هسته‌ی محبوب است که در الگوریتم‌های یادگیری هسته‌ای مختلف، به‌خصوص در طبقه‌بندی اس‌وی‌ام، استفاده می‌شود (جرون ۲۰۱۹، ۲۲۵).

[2] معیار ناخالصی (Impurity) بیانگر میزان همگن بودن برچسب‌ها در گره است. یک گره «خالص» است اگر همه‌ی نمونه‌های آموزشی به یک کلاس تعلق داشته باشند (جرون[2] ۲۰۱۹، ۲٤۲).

[3] ناخالصی جینی (Gini Impurity) یکی از شاخص‌های پر استفاده در ساخت درخت تصمیم است که برای تعیین چگونگی تقسیم گره‌ها، مبتنی بر ویژگی‌های یک مجموعه داده، مورد استفاده قرار می‌گیرد. علاوه بر این شاخص، «شاخص آنتروپی» (entropy) و «شاخص خطای طبقه‌بندی» (classification error) نیز پر استفاده‌اند.

۳-۲-۵-۷- مدل جنگل تصادفی

بر اساس نوشته‌ی هوارد و گوگر (۲۰۲۰، ۲۹۸) مراحل کلی رویکرد جنگل تصادفی به شرح زیر است:

۱. به طور تصادفی زیرمجموعه‌ای از مجموعه داده‌های انتخاب می‌شوند.

۲. یک مدل درخت تصمیم، با استفاده از این زیر مجموعه آموزش داده می‌شود.

۳. آن مدل ذخیره و سپس چند بار مرحله یک و دو تکرار می‌شوند تا درخت‌های تصمیم بیشتری به دست‌آید.

٤. خروجی مدل‌های آموزش دیده‌ی حاصله، هر کدام یک پیش‌بینی به‌خصوص به حساب می‌آیند. سپس با محاسبه‌ی میانگین آن پیش‌بینی‌ها، خروجی نهایی مدل جنگ تصادفی بدست می‌آید.

مراحل بالا بر اساس روشی موسوم به بسته‌بندی[۱] استنتاج شده‌اند. در این روش، هرچند هر یک از مدل‌های آموزش‌دیده بر روی زیرمجموعه‌ای از داده‌ها، نسبت به مدلی که بر روی مجموعه داده‌ی کامل آموزش داده شود، خطای بیشتری دارند؛ اما این خطاها با یکدیگر همبستگی نخواهند داشت. مدل‌های مختلف خطاهای متفاوتی خواهند داشت اما میانگین آن خطاها صفر است. بنابراین هر چه تعداد بیشتری مدل داشته باشیم، میانگین پیش‌بینی‌های آن‌ها به پاسخ صحیح نزدیک‌تر می‌شود. این یک دستاورد خارق‌العاده است که می‌تواند تقریباً دقت هر نوع الگوریتم یادگیری ماشینی را افزایش دهد (هوارد و گوگر ۲۰۲۰، ۲۹۸).

در این مدل هم، که آخرین مدل سنتی مورد بررسی است، از تابع طبقه‌بندی‌کننده‌ی جنگل تصادفی ارائه شده در کتابخانه‌ی سایکیت-لرن، استفاده شده است. این بار از شاخص آنتروپی برای اندازه‌گیری کیفیت تقسیم استفاده شده و تعداد درختان هر جنگل(تعداد تخمین‌کننده‌ها)، ۱۰۰ عدد در نظر گرفته شده است.

۳-۲-۵-۸- مدل شبکه‌ی عصبی پیشخور

برای پیاده‌سازی این مدل، از واسط برنامه‌نویسی کِرس، که در کتابخانه‌ی نرم‌افزار تنسورفلو ارائه شده است، استفاده شده است. با بهره‌گیری از این واسط، مدلی «ترتیبی»[۲] ساخته شد. بعد از بررسی‌های چندباره و با توجه به نوع داده‌ها و هدف، این نتیجه حاصل شد که بهتر است تنها یک لایه‌ی پنهان در مدل تعریف شود. در این لایه، از تابع فعال‌سازی «یکسوساز»[۳]، با اندازه‌ی خروجی ۱۰۰ بُعد، استفاده کردیم.

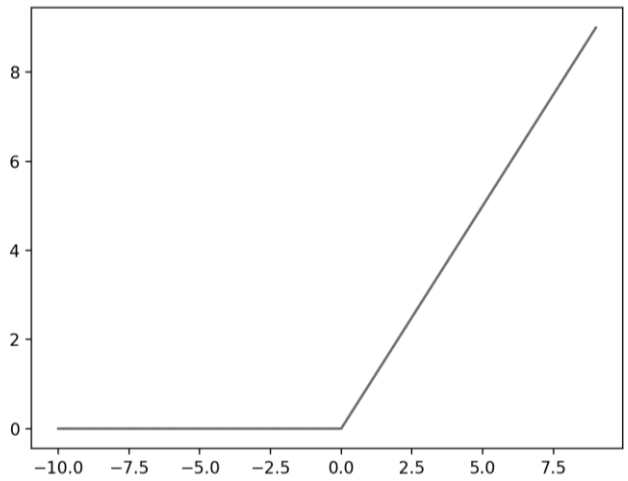
---

[۱] Bagging

[۲] Sequential

[۳] Rectifier (Rectified Linear Unit - ReLU)

تابع فعال‌سازی یکسوساز، تابعی پر استفاده است که در مقایسه با توابع جایگزین (مثل سیگموئید[۱] یا تانژانت هذلولی[۲]) محاسبه آن بسیار ساده و کارآمدتر است. این تابع، همگرایی را تا شش برابر بهبود می‌بخشد که این امر احتمالاً به دلیل ماهیت غیرخطی و فرم اشباع‌نپذیر آن است. همین سادگی و کارآمدی باعث شده که امروزه، تقریباً تمام مدل‌های یادگیری عمیق از این تابع بهره‌گیرند (دی، بهاردواج و وی ۲۰۱۸، ٦٤). نمودار این تابع در نمودار ۳–۱ قابل مشاهده است.



(نمودار ۳–۱) نمودار تابع یکسوساز (اقتباس از ویدمن ۲۰۱۹، ۱۰۸).

فرمول ریاضی تابع یکسوساز، در رابطه‌ی ۳–۳ دیده می‌شود.

$$\sigma(x) = \begin{cases} x \geq 0 \rightarrow \max(0, x) & , x \geq 0 \\ 0 & , x < 0 \end{cases}$$

(۳–۳)

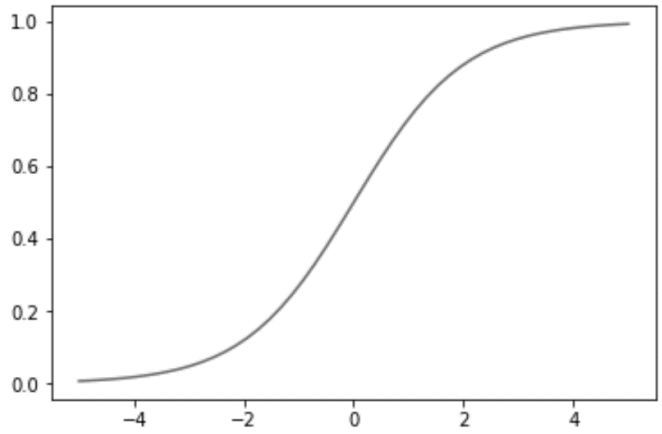
علی‌رغم مزیت‌های ذکر شده، که تابع یکسوساز را گزینه‌ای ایده‌آل برای استفاده در لایه‌ی پنهان می‌سازند، محدودیت این تابع این است که خروجی مستقیم آن در فضای احتمال نیست. درنتیجه نمی‌توان از آن در لایه‌ی خروجی بهره‌برد. در ضمن استفاده از تابع یکسوساز می‌تواند «نورون‌های مرده»[۳] را باعث شود؛ یعنی

[۱] Sigmoid
[۲] Hyperbolic Tangent (Tanh)
[۳] Dead Neurons

وزن‌ها را به گونه‌ای به‌روز کند که یک نورون هرگز در هیچ نقطه‌ی داده‌ی دیگری، فعال نباشد (دی، بهاردواج و وی ۲۰۱۸، ٦٤). در این حالت، نورون دارای «گرادیان»[1] صفر است.

با توجه به مطالب ذکر شده، در لایه‌ی خروجی از تابع فعال‌ساز پرطرفدار «بیشینه فعال»[2] استفاده کردیم. بدیهی است که اندازه‌ی ابعاد خروجی این تابع می‌بایست برابر تعداد کل کلاس‌ها (۱۷ عدد برای تحقیق ما) باشد.

تابع بیشینه فعال بسطی به تابع سیگموئید است. این تابع توسعه یافته شده تا محدودیت تابع سیگموئید در حل مسائلی با کلاس‌های چندگانه (بیش از دو کلاس) را برطرف سازد. تابع سیگموئید مخصوص طبقه‌بندی دودویی در مسائل دارای دو کلاس می‌باشد و نمودار آن در نمودار ۳–۲ قابل مشاهده است (ویدمن ۲۰۱۹، ۵۹).



(نمودار ۳–۲) نمودار تابع سیگموئید (اقتباس از ویدمن ۲۰۱۹، ۵۹).

تابع بیشینه فعال، مقدار حداکثر را نسبت به سایر مقادیر، به‌شدت تقویت می‌کند و شبکه عصبی را مجبور می‌کند تا در فرایند طبقه‌بندی، در راستای پیش‌بینی مقداری که احتمالاً صحیح است، «کمتر خنثی» عمل کند. به طور کلی این تابع، ضمن نرمال کردن مقادیر، تابع حداکثر را اعمال می‌کند که به همین دلیل به آن تابع «بیشینه فعال» اطلاق می‌شود (ویدمن ۲۰۱۹، ۱۰٤–۱۰۳). رابطه‌ی ۳–٤، نمایشگر معادله‌ی این تابع است.

[1] در یادگیری ماشین، گرادیان (Gradient) مشتقی از تابعی است که بیش از یک متغیر ورودی دارد. گرادیان که از نظر ریاضی به عنوان شیب تابع شناخته می‌شود، تغییرات تمام وزن‌ها را با توجه به تغییرات خطا، اندازه‌گیری می‌کند (رابرتز، یادا و هانین ۲۰۲۲، ۳۵۸).

[2] Softmax

$$p(y = j|z_j) = \emptyset(z_j) = \frac{e^{z_j}}{\sum_j^K z_j}$$

(۳–٤) رابطه‌ی کلی تابع بیشینه فعال (دی، بهاردواج و وی ۲۰۱۸، ٦٥).

پس از تعریف لایه‌ها در مدل، اکنون می‌بایست مدل را اصطلاحاً «کامپایل»[۱] یا راه‌اندازی کنیم. برای اینکار لازم است از یک تابع هزینه (تابع زیان) برای مدل استفاده کنیم. همانطور که قبل‌تر توضیح داده شد، در شبکه‌های عصبی از تابع هزینه برای یافتن خطا یا انحراف در فرآیند یادگیری استفاده می‌شود. با توجه به اینکه کلاس‌های رکوردها چندگانه می‌باشند؛ درنتیجه از تابع «آنتروپی متقاطع طبقه‌بندی»[۲] موجود در واسط برنامه‌نویسی کرس استفاده کردیم.

همچنین برای بهینه‌سازی مدل از تابع پرطرفدار «اَدم»[۳]، موجود در واسط کرس، استفاده کردیم. به بیان کینگما[٤] و با[٥] (۲۰۱۷، ۱)، این روش از نظر محاسباتی کارآمد بوده، نیاز به حافظه‌ی کمی داشته و نسبت به تغییرات مقیاس قطر گرادیان‌ها ثابت می‌باشد. درنتیجه روشی بسیار مناسب به حساب می‌آید.

در نهایت، با توجه به نوع مسئله، از متریک دقت طبقه‌بندی برای محاسبه‌ی عملکرد مدل، حین فرایند یادگیری، استفاده کردیم. این تابع نیز در واسط برنامه‌نویسی کرس ارائه شده است.

حال شرایط برای فرایند یادگیری مدل فراهم شده است. برای دستیابی به نتایج بهتر، داده‌های آموزش در چندین «دسته»[٦] قرار داده شد تا در هر دسته، جداگانه فرایند آموزش انجام شده و نتایج بدست آمده منجر به بهبود مدل شود. با توجه به شاخص‌ها و عوامل موجود در این تحقیق، اندازه‌ی دسته ۱۰۰ عدد و اندازه‌ی «مجموعه داده‌ی اعتبارسنجی»[۷] ۵ درصد کل مجموعه داده‌ی آموزش، در نظر گرفته شد. مجموعه داده‌ی اعتبارسنجی، مجموعه داده‌ای است که از مجموعه داده‌ی آزمایش جدا شده و در فرایند یادگیری حضور پیدا نمی‌کند. از این مجموعه داده پس از هر «چرخه»[۸]، برای محاسبه‌ی عملکرد مدل در آن چرخه مورد استفاده قرار گرفته و اطلاعات مفیدی را در اختیار داده‌کاو قرار می‌دهد. هر چرخه شامل یک جریان به جلو و یک

---

[۱] Compile

[۲] Categorical Crossentropy

[۳] تابع ادم (Adam) یک تابع بهینه ساز است که مبتنی بر الگوریتمی با همین اسم پیاده سازی شده است. بهینه‌سازی ادم یک روش گرادیان کاهشی تصادفی (Stochastic Gradient Descent) است که مبتنی بر تخمین تطبیقی گشتاور مرتبه اول و مرتبه دوم است (وبسایت رسمی کرس، ۲۰۲۲).

[٤] Kingma

[٥] Ba

[٦] Batch

[۷] Validation Dataset

[۸] Epoch

جریان به عقب است که روی تمامی دسته‌های مجموعه داده‌ی آزمایش، انجام می‌گیرد. تعداد هشت چرخه برای آموزش مدل، مناسب دیده شد.

۳-۲-۵-۹- مدل شبکه‌ی عصبی بازگشتی

با توجه به نحوه‌ی عملکرد این مدل، ابتدا از تابع «توکن‌ساز» واسط کرس، برای قرار دادن محتویات در فرمت مناسب ورودی، استفاده کردیم. این تابع در حقیقت عمل ساخت بسته‌ی کلمات را انجام می‌دهد. با کمک این تابع، تعداد کلمات را به پنج هزار عدد و تعداد توالی کلمات در هر پاراگراف را به ۲۵۰ عدد، محدود ساختیم.

پس از قرار دادن داده‌ها در قالب مناسب، یک مدل ترتیبی ساخته و لایه‌هایی به آن اضافه کردیم. لایه‌ی «دگرنمایی واژه»[۱] را به عنوان لایه‌ی ابتدایی معرفی کردیم. در این لایه بسته‌های کلمات ایجاد شده به شکل آرایه‌های اندازه ثابت در می‌آیند تا ورودی‌های مناسب‌تری برای مدل باشند. تعداد ابعاد آرایه‌ی خروجی از این لایه، ۱۰۰ عدد در نظر گرفته شد.

برای لایه‌ی دوم، از لایه‌ی «حافظه‌ی طولانی کوتاه-مدت»[۲] یا ال‌اس‌تی‌ام، که مهم‌ترین لایه در مدل شبکه‌ی عصبی بازگشتی به حساب می‌آید، استفاده کردیم. برای ارائه‌ی مدل، از لایه‌ی ال‌اس‌تی‌ام واسط نرم‌افزار کرس استفاده و اندازه‌ی بعد خروجی ۱۰۰ واحد در نظر گرفته شد.

به بیان راسل و نورویگ (۲۰۲۱ ، ۸۲٦)، ال‌اس‌تی‌ام یک مدل شبکه‌ی عصبی بازگشتی است که با هدف حفظ اطلاعات در چندین مرحله طراحی شده است. این مدل شامل سه «واحد دروازه‌ای»[۳] است. این دروازه‌ها بردارهایی هستند که جریان اطلاعات را از طریق ضرب درایه‌ای بردار اطلاعات مربوطه، کنترل می‌کنند:

- «دروازه‌ی فراموشی»[٤] در خصوص هر عنصر «سلول حافظه»[٥]، به خاطر سپرده شدن (کپی شدن در مرحله‌ی بعدی) یا فراموش کردن (بازنشانی به صفر) آن را تعیین می‌کند.
- «دروازه‌ی ورودی»[٦] در خصوص هر عنصر سلول حافظه، تعیین می‌کند که آن عنصر، در مرحله‌ی زمانی فعلی، مبتنی بر اطلاعات جدید از بردار ورودی، به‌روز شود یا نشود.

---

[۱] Word Embedding

[۲] Long Short-Term Memory (LSTM)

[۳] Gating Unit

[٤] Forget Gate

[٥] سلول حافظه (Memory Cell) واحد سازنده‌ی حافظه‌ی بلند مدت یک ال‌اس‌تی‌ام، که اساساً از یک مرحله‌ی زمانی به مرحله‌ی دیگر، کپی می‌شود (راسل و نورویگ ۲۰۲۱ ، ۸۲٦).

[٦] Input Gate

• «دروازه‌ی خروجی»[1] در خصوص هر عنصر سلول حافظه، منتقل شدن آن‌را به حافظه کوتاه مدت تعیین می‌کند. این دروازه نقشی مشابه حالت پنهان (ورودی‌های رمزگذاری شده که وابستگی‌ها زمانی در آن لحاظ و کنترل می‌شود) در شبکه‌های عصبی بازگشتی اصلی، ایفا می‌کند.

راسل و نورویگ (۲۰۲۱ ، ۸۲٦) همچنین اظهار داشتند که مقادیر در واحدهای دروازه، همیشه در محدوده ۰ و ۱ هستند. این مقادیر از اعمال تابع سیگموئید به ورودی فعلی و حالت پنهان قبلی بدست می‌آیند. معادلات به‌روزرسانی ال‌اس‌تی‌ام، در روابط ۳–۵، ۳–٦، ۳–۷، ۳–۸ و ۳–۹ آورده شده است.

$$\mathrm{f_t} = \sigma_g(W_{x,f}x_t + W_{z,f}z_{t-1})$$

(۳–۵)

$$\mathrm{i_t} = \sigma_g(W_{x,j}x_t + W_{z,i}z_{t-1})$$

(۳–٦)

$$o_\mathrm{t} = \sigma_g(W_{x,o}x_t + W_{z,o}z_{t-1})$$

(۳–۷)

$$\mathrm{c_t} = c_{t-1} \odot f_t + i_t \odot \tanh\left(W_{x,c}x_t + W_{z,c}z_{t-1}\right)$$

(۳–۸)

$$\mathrm{z_t} = \tanh(x_t) \odot \mathrm{o_t}$$

(۳–۹)

f : دروازه‌ی فراموشی	i : دروازه‌ی ورودی	o : دروازه‌ی خروجی
c : سلول حافظه	z : حافظه‌ی کوتاه مدت	$\odot$: ضرب درایه‌ای
σ_g : تابع سیگموئید	x : ورودی	z_{t-1} : حالت پنهان قبلی
W : وزن		

پس از اضافه کردن مدل ال‌اس‌تی‌ام، حال زمان اضافه کردن لایه‌ی خروجی، به عنوان آخرین لایه‌ی مدل است. مشابه مدل شبکه عصبی پیشخور، از لایه‌ی متراکم با تابع فعال‌سازی بیشینه فعال استفاده و اندازه‌ی ابعاد خروجی آن برابر با تعداد کل کلاس‌ها، ۱۷ عدد لحاظ شد.

[1] Output Gate

برای کامپایل مدل هم مشابه قبل، برای بهینه‌سازی مدل از تابع ادم و از آنتروپی متقاطع طبقه‌بندی، به عنوان تابع هزینه، بهره بردیم. اینبار برای محاسبه‌ی عملکرد، از متریک دقت ارائه شده در واسط کرس استفاده کردیم.

برای آموزش مدل ساخته شده، داده‌های را در دسته‌های ۵۰۰ عددی قرار داده و تعداد چرخه‌ها را ۱۰ عدد در نظر گرفتیم. همچنین، اندازه‌ی مجموعه داده‌ی اعتبارسنجی را ۵ درصد کل مجموعه داده‌ی آزمایش در نظر گرفتیم.

۳-۲-٦- فاز ارزیابی

پس از ساخت مدل‌ها و انجام دادن فرایند آموزش آن‌ها، حال باید عملکرد آن‌ها محاسبه تا میزان دستیابی به اهداف مدنظر مشخص شود. برای دستیابی به این خواسته، باید از روابط اندازه‌گیری عملکرد مدل در زمینه‌ی یادگیری ماشین، بهره‌برداری شود. در ادامه این روابط معرفی خواهند شد.

همانطور که لاروس و لاروس (۲۰۱۹، ۹۸) ذکر کرده‌اند، «صحت»، اندازه‌گیری کلی نسبت طبقه‌بندی‌های صحیح توسط مدل را نشان می‌دهد و رابطه‌ی ۳—۱۰ نمایانگر معادله‌ی مربوط به آن است.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

(۱۰—۳)

جائیکه:

TP: مثبت صحیح[۱] TN: منفی صحیح[۲] FP: مثبت کاذب[۳] FN: منفی کاذب[٤]

با توجه به رابطه‌ی فوق، «نرخ خطا» با رابطه‌ی ۳—۱۱ محاسبه می‌شود (لاروس و لاروس ۲۰۱۹، ۹۹).

$$Error\ Rate = 1 - Accuracy$$

(۱۱—۳)

---

[۱] مثبت صحیح (True Positive) مقادیری هستند که توسط مدل، به‌درستی، وجودشان در یک وضعیت یا مشخصه پیش‌بینی شده است.

[۲] منفی صحیح (True Negative) مقادیری هستند که توسط مدل، به‌درستی، عدم وجودشان در یک وضعیت یا مشخصه پیش‌بینی شده است.

[۳] مثبت کاذب (False Positive) مقادیری هستند که توسط مدل، به‌اشتباه، وجودشان در یک وضعیت یا مشخصه پیش‌بینی شده است (در حالیکه به آن مشخصه تعلق ندارند).

[٤] منفی کاذب (False Negative) مقادیری هستند که توسط مدل، به‌اشتباه، عدم وجودشان در یک وضعیت یا مشخصه پیش‌بینی شده است (در حالیکه به آن مشخصه تعلق دارند).

با وجود اینکه صحت و نرخ خطا، اطلاعات مهمی را در اختیار می‌گذارند؛ این معیارها بین انواع مختلف خطاها یا انواع مختلف تصمیمات صحیح، تمایز قائل نمی‌شوند. درنتیجه برای بدست آوردن اطلاعات بیشتر، باید از معیار حساسیت (بازیابی) و ویژگی، مبتنی بر روابط ۳–۱۲ و ۳–۱۳ استفاده شود (لاروس و لاروس ۲۰۱۹، ۹۹).

$$Sensitivity = Recall = \frac{TP}{TP + FN}$$

(۳–۱۲)

$$Specificity = \frac{TN}{FP + TN}$$

(۳–۱۳)

همانطور که مشاهده می‌شود، حساسیت توانایی مدل را در شناسایی رکوردهای مثبت و ویژگی توانایی مدل را در شناسایی رکوردهای منفی، اندازه‌گیری می‌کند.

علاوه بر معیارهای یادشده، برای آنکه متوجه شویم که چه تعداد از رکوردهای طبقه‌بندی شده توسط مدل، به‌درستی به کلاس متناظرشان مرتبط شده‌اند، از معیار «دقت» استفاده می‌کنیم. این معیار نشان می‌دهد که چه نسبتی از رکوردهای طبقه‌بندی شده مرتبط هستند و رابطه‌ی ۳–۱٤، بیانگر معادله‌ی آن است (لاروس و لاروس ۲۰۱۹، ۱۰۰):

$$Precision = \frac{TP}{TP + FP}$$

(۳–۱٤)

امتیازات اف ترکیب معیار دقت و بازیابی در یک معیار است که رابطه‌ی ۳–۱۵، نحوه‌ی محاسبه‌ی آن‌را نشان می‌دهد:

$$F_\beta - score = (1 + \beta^2) \times \frac{Precision \times Recall}{\beta^2 \times Precision + Recall}$$

(۳–۱۵)

برای $\beta > 0$ ، حالات مختلف و معانی آن‌ها در جدول ۳–۲، آورده شده است.

(جدول ۳–۲) معنای مقادیر مختلف β در امتیاز اف.

مقدار	$\beta = ۱$	$\beta > ۱$	$\beta < ۱$
معنی	میانگین هارمونیک دقت و یادآوری	وزن یادآوری بیشتر است	وزن دقت بیشتر است

در نهایت، محاسبه‌ی کارایی روش ارائه شده با بررسی میزان زمان صرف شده و معرفی منابع سخت‌افزاری استفاده شده برای سازماندهی محتویات، انجام خواهد شد.

۳–۲–۷– فاز پیاده‌سازی و بازخورد

در این فاز، مدل‌ها ذخیره شده و برای مصرف نهایی انتشار داده می‌شود. پس از انتشار نیز باید، با تجزیه و تحلیل نتایج بدست آمده، از مدل بازخورد گرفت.

همانطور که اشاره شد، تمامی مدل‌های ساخته شده و مراحل ساخت آن‌ها، به صورت متن‌باز در وبسایت گیت‌هاب[۱] انتشار یافته‌اند و برای عموم دسترس‌پذیرند. اطلاعات مربوطه در پیوست ۱ آورده شده است.

۳–۳– جامعه‌ی آماری و حجم نمونه

جامعه‌ی آماری این تحقیق، محتویات موجود در سیستم مدیریت دانش سازمان‌ها (شامل شرکت‌ها، نهادهای اجتماعی، انجمن‌ها و...) است. روش نمونه‌گیری به صورت تصادفیِ طبقه‌بندی شده است. به این ترتیب که فضای نمونه‌ای تحقیق را مقالات تصادفی استخراج شده از وبسایت ویکی‌پدیا، به‌عنوان یک سیستم مدیریت محتوای عظیم، تشکیل می‌دهند.

مطابق فرمول کوکران[۲]، حجم نمونه برای جامعه آماری نامحدود، با میزان خطای ۰.۰۲ و انحراف معیار ۰.۵ باید حداقل ۲,٤۰۱ در نظر گرفته شود. فرمول کوکران در رابطه‌ی ۳–۱٦ قابل مشاهده شده است (کوکران ۱۹۷۷، ۷۵).

$$n = \frac{\frac{z^2 pq}{d^2}}{1 + \frac{1}{N}\left[\frac{z^2 pq}{d^2} - 1\right]}$$

(۳–۱٦)

[۱] گیت‌هاب (https://github.com) یک سرویس میزبانی اینترنتی برای توسعه نرم‌افزار و کنترل نسخه است که عمدتاً برای میزبانی پروژه‌های متن‌باز، مورد استفاده قرار می‌گیرد. این سرویس تحت مالکیت شرکت مایکروسافت می‌باشد.

[۲] William G. Cochran

با توجه به مطالب فوق، برای مجموعه داده‌ی فارسی، تعداد ۱٤,۲۲٤ رکورد و برای مجموعه داده‌ی انگلیسی، تعداد ۲۳,۸۵۰ (به‌دلیل دسترس‌پذیری بیشتر) رکورد، در نظر گرفته شد.

فصل چهارم: تجزیه و تحلیل داده‌ها و یافته‌ها

- مقدمه
- تجزیه و تحلیل رکوردها (ورودی‌ها)
- بررسی عملکرد مدل‌ها

٤-١- مقدمه

در این فصل از تحقیق، ابتدا رکوردها و داده‌های ورودی معرفی می‌شوند. سپس، نتایج عملکرد مدل‌های پیاده‌سازی شده، محاسبه و مشخص می‌شوند. این نتایج براساس معیارهایی که در فصل قبل توضیح داده شد، محاسبه می‌گردند. بنابراین، هدف اصلی در این فصل، مشخص‌سازی و بررسی کلی داده‌ها و یافته‌ها است. در فصل بعد، با دیدی عمیق‌تر، در مورد یافته‌ها بحث خواهد شد و نتایج حاصله اعلام می‌شود.

همانطور که قبل‌تر هم اشاره شد، در این تحقیق دو دسته مجموعه داده در نظر گرفته شده است. یکی از آن‌ها حاوی پاراگراف‌های فارسی و دیگری حاوی پاراگراف‌های انگلیسی است. در این فصل، ابتدا توزیع فراوانی رکوردهای این مجموعه داده‌ها در هر کلاس، بررسی می‌شود. سپس عملکرد مدل‌های پیاده‌سازی شده، در هر کدام از این مجموعه داده‌ها، به طور مجزا گزارش می‌شود. بنابراین، عملکرد مدل‌ها هم به صورت کلی و هم به تفکیک هر کلاس، گزارش خواهد شد.

عملکرد مدل شبکه‌ی عصبی پیشخور، که مدل اصلی مدنظر است، به طور مفصل بررسی شده است؛ اما عملکرد مدل‌های جایگزین به صورت اجمالی‌تر مورد بررسی قرار می‌گیرند. نحوه‌ی دسترسی به اطلاعات مفصل‌تر، در خصوص مدل‌های جایگزین، در پیوست ۱ آورده شده است.

برای تجزیه و تحلیل داده‌ها از زبان برنامه‌نویسی پایتون و کتابخانه‌های داده کاوی آن در کنار نرم‌افزار اکسل استفاده شده است.

٤-٢- تجزیه و تحلیل رکوردها (ورودی‌ها)

همانطور که ذکر شد، ورودی‌ها در این تحقیق در قالب دو مجموعه داده، یکی فارسی و دیگری انگلیسی می‌باشند. در مرحله‌ی نخست، رکوردهای موجود در این مجموعه داده‌ها بررسی خواهند شد.

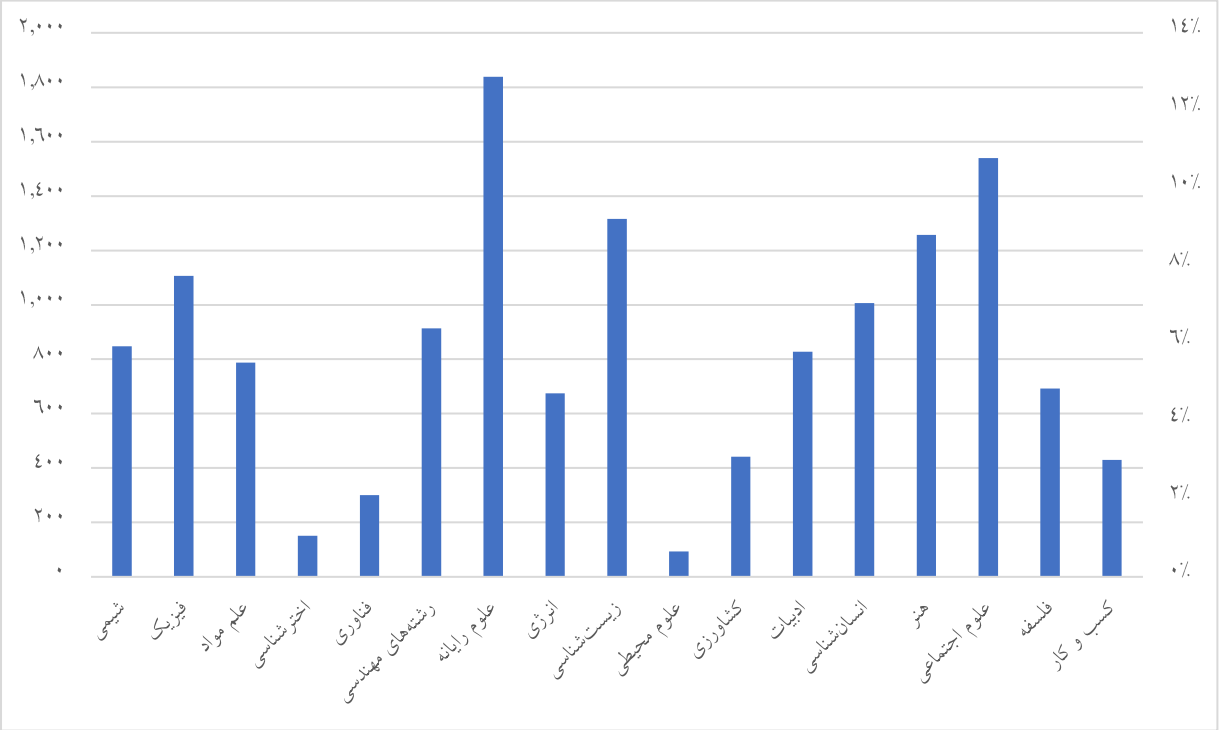
٤-٢-١- مجموعه داده‌ی فارسی

این مجموعه داده دارای ١٤,٢٢٤ رکورد است. جدول تعداد رکوردهای این مجموعه داده، به تفکیک کلاس، در جدول ٤—١ آورده شده است.

(جدول ٤–١) توزیع فراوانی رکوردهای مجموعه داده‌ی فارسی در هر کلاس.

برچسب	عنوان	فراوانی	درصد
۰	شیمی	۸٤۸	۵.۹٦٪
۱	فیزیک	۱,۱۰٦	۷.۷۸٪
۲	علم مواد	۷۸۸	۵.۵٤٪
۳	اخترشناسی	۱۵۱	۱.۰٦٪
٤	فناوری	۳۰۰	۲.۱۱٪
۵	رشته‌های مهندسی	۹۱۳	٦.٤۲٪
٦	علوم رایانه	۱,۸۳۹	۱۲.۹۳٪
۷	انرژی	٦۷۵	٤.۷۵٪
۸	زیست‌شناسی	۱,۳۱٦	۹.۲۵٪
۹	علوم محیطی	۹۳	۰.٦۵٪
۱۰	کشاورزی	٤٤۲	۳.۱۱٪
۱۱	ادبیات	۸۲۸	۵.۸۲٪
۱۲	انسان‌شناسی	۱,۰۰٦	۷.۰۷٪
۱۳	هنر	۱,۲۵۷	۸.۸٤٪
۱٤	علوم اجتماعی	۱,۵٤۰	۱۰.۸۳٪
۱۵	فلسفه	٦۹۲	٤.۸۷٪
۱٦	کسب و کار	٤۳۰	۳.۰۲٪
مجموع		۱٤,۲۲٤	۱۰۰.۰۰٪

نمودار میله‌ای تعداد رکوردها در هر کلاس، در نمودار ٤–۱ قابل مشاهده است.



(نمودار ٤–١) نمودار توزیع فراوانی رکوردهای مجموعه داده‌ی فارسی در هر کلاس.

همانطور که مشاهده می‌شود، کلاس علوم رایانه با ۱,۸۳۹ رکورد، حدود ٪۱۳ کل رکوردها را به خود اختصاص داده و از این بابت، کلاس دارای بیشترین تعداد رکورد به شمار می‌آید. از طرف دیگر، کلاس علوم محیطی با ۹۲ رکورد، تنها ٪۰.٦٥ رکوردها را به خود اختصاص داده و حاوی کمترین تعداد رکورد نسبت به سایر کلاس‌ها می‌باشد.

٤–۲–۲– مجموعه داده‌ی انگلیسی

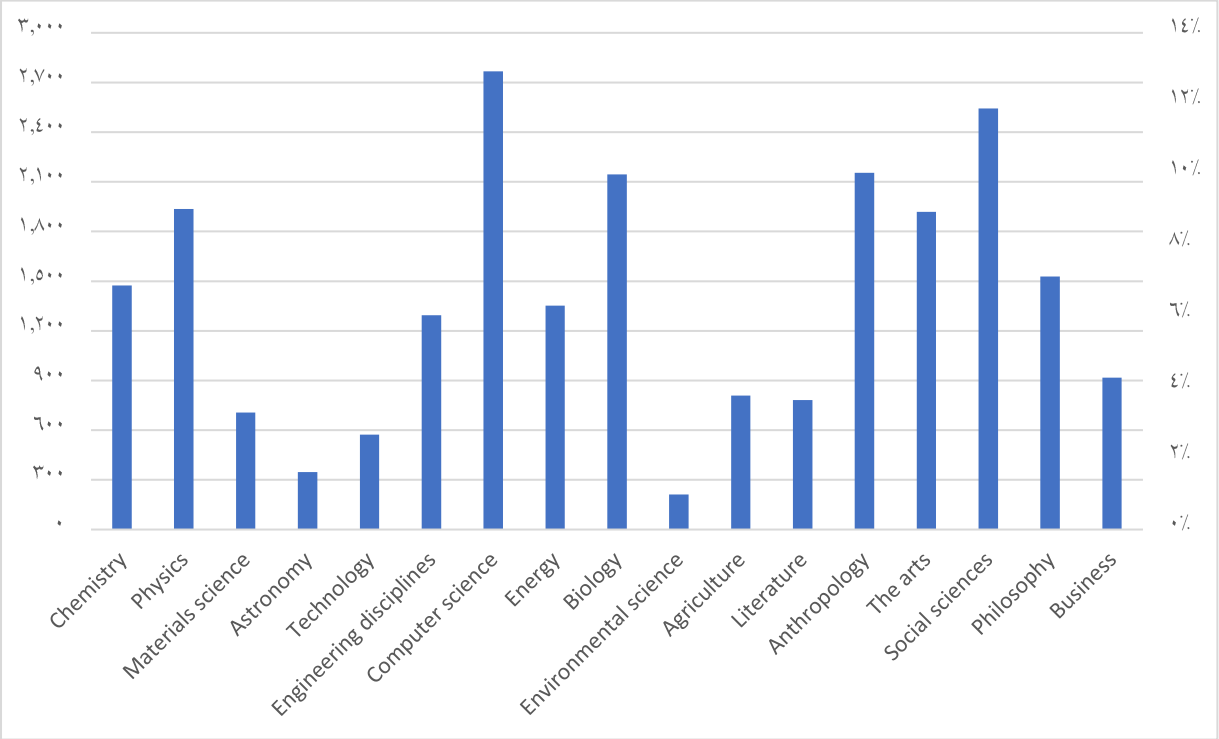
این مجموعه داده دارای ۲۳,٤٦٥ رکورد است. جدول تعداد رکوردهای این مجموعه داده، به تفکیک کلاس، در جدول ٤–۲ آورده شده است.

(جدول ٤–۲) توزیع فراوانی رکوردهای مجموعه داده‌ی انگلیسی در هر کلاس.

برچسب	عنوان	فراوانی	درصد
۰	Chemistry	۱,٤۷٤	٦.۲۸٪
۱	Physics	۱,۹۳٦	۸.۲٥٪
۲	Materials science	۷۰٦	۳.۰۱٪
۳	Astronomy	۳٤۸	۱.٤۸٪

٤	Technology	٥٧٤	٢.٤٥٪
٥	Engineering disciplines	١,٢٩٥	٥.٥٢٪
٦	Computer science	٢,٧٦٨	١١.٨٠٪
٧	Energy	١,٣٥٢	٥.٧٦٪
٨	Biology	٢,١٤٥	٩.١٤٪
٩	Environmental science	٢١٢	٠.٩٠٪
١٠	Agriculture	٨١٠	٣.٤٥٪
١١	Literature	٧٨٢	٣.٣٣٪
١٢	Anthropology	٢,١٥٤	٩.١٨٪
١٣	The arts	١,٩١٩	٨.١٨٪
١٤	Social sciences	٢,٥٤٤	١٠.٨٤٪
١٥	Philosophy	١,٥٢٨	٦.٥١٪
١٦	Business	٩١٨	٣.٩١٪
مجموع		٢٣,٤٦٥	١٠٠.٠٠٪

نمودار میله‌ای تعداد رکوردها در هر کلاس، در نمودار ٤—٢ آمده است.



(نمودار ٤–۲) نمودار میله‌ای توزیع فراوانی رکوردهای مجموعه داده‌ی انگلیسی در هر کلاس.

همانطور که مشاهده می‌شود، کلاس علوم رایانه با ۲,۷٦۸ رکورد، حدود ٪۱۲ کل رکوردها را به خود اختصاص داده و از این بابت، کلاس دارای بیشترین تعداد رکورد به شمار می‌آید. از طرف دیگر، کلاس علوم محیطی با ۲۱۲ رکورد، حدود ٪۱ رکوردها را به خود اختصاص داده و حاوی کمترین تعداد رکورد نسبت به سایر کلاس‌ها می‌باشد.

همانطور که در بخش جمع‌آوری داده‌ها در فصل سوم اشاره شد، رکوردها در هر کلاس به صورت ناهمگن قرار داده شدند تا منعکس‌کننده‌ی جامعه‌ی آماری واقعی باشند. تعداد رکوردهای هر کلاس می‌تواند به‌شدت متفاوت باشد. این تفاوت با زیادتر شدن تعداد کلاس‌ها، بیشتر هم می‌شود. بنابراین، یکی از شاخص‌های تأثیرگذار در تعیین میزان کارایی یک مدل، عملکرد آن در طبقه‌بندی مناسب کلاس‌هایی با تعداد رکوردهای متفاوت و گاهاً ناکافی[1] می‌باشد. ناهمگن بودن رکوردهای هر کلاس، این قابلیت را فراهم می‌کند تا عملکرد مدل در شرایط مختلف سنجیده شود.

[1] معمولاً توصیه می‌شود که تعداد رکوردهای هر کلاس، برای یادگیری ماشین، کمتر از سیصد الی پانصد عدد نباشد (شوله ۲۰۲۱، ۲۱٤). اما در دنیای واقعی، پیش‌آمد موارد «ناکافی» بودن تعداد داده‌ها، اجتناب ناپذیر است.

٤-٣- بررسی عملکرد مدل‌ها

در این بخش، معیارهای مربوطه در مورد هر مدل به تفکیک بررسی خواهند شد. نتایج هر مدل برای مجموعه داده‌ی فارسی و انگلیسی به طور جداگانه بیان خواهند شد. لازم به ذکر است که برای رکوردهای فارسی، تعداد رکوردها در مجموعه داده‌ی آموزش، ۱۱,۳۷۹ عدد و در مجموعه داده‌ی آزمایش، ۲,۸٤٥ عدد در نظر گرفته شده است. برای رکوردهای انگلیسی، تعداد ۱۸,۷۷۲ رکورد در مجموعه داده‌ی آموزش و ٤,٦۹۳ رکورد در مجموعه داده‌ی آزمایش، در نظر گرفته شده است.

در ادامه خواهیم دید که هر مدل تا چه اندازه در طبقه‌بندی رکوردها موفق است. نتایج مدل اصلی (مبتنی بر شبکه‌های عصبی پیشخور) به طور مفصل بررسی شده و نتایج مدل‌های جایگزین، به صورت خلاصه آورده شده است. اطلاعات جامع‌تر درباره‌ی عملکرد مدل‌های جایگزین، در صفحه‌ی گیت‌هاب این تحقیق آورده شده که چگونگی دسترسی به آن در فایل پیوست ۱، توضیح داده شده است.

همانطور که بیان شد، تمامی مدل‌ها در پلتفرم گوگل کولب پیاده‌سازی شده‌اند که مشخصات سخت‌افزاری قابل استفاده در آن، به این شرح است: پردازنده‌ی مجازی دو هسته‌ای با فرکانس ۲.۳ گیگاهرتز از خانواده‌ی «هاسوِل»[۱]، ۱۲ گیگابایت حافظه‌ی دسترسی تصادفی (رَم) و ۱۰۰ گیگابایت فضای ذخیره‌ی ابری.

نتایج حاصل از این بررسی‌ها نشان خواهد داد که مدل ارائه شده، به چه اندازه در دستیابی به اهداف مدنظر موفق بوده است. نخست به بررسی نتایج عملکرد مدل‌های جایگزین و سپس مدل اصلی پرداخته خواهد شد:

٤-٣-١- عملکرد مدل کی‌ان‌ان

همانطور که اشاره شد، کی‌ان‌ان مدلی بسیار ساده است. عملکرد این مدل، نخست در مجموعه داده‌ی فارسی و سپس در مجموعه داده‌ی انگلیسی، بررسی خواهد شد:

٤-٣-١-١- مجموعه داده‌ی فارسی

فرایند آموزش این مدل، حدود ۰.۰۳ ثانیه زمان برده است. پیش‌بینی ۲۸٤٥ رکورد مجموعه داده‌ی آزمایش توسط این مدل، ۲۲ ثانیه به طول انجامیده است؛ یعنی حدود ۷ میلی ثانیه برای هر رکورد (پاراگراف).

در جدول ٤–۳، عملکرد کلی مدل کی‌ان‌ان در طبقه‌بندی رکوردهای مجموعه داده‌ی فارسی، قابل مشاهده می‌باشد.

---

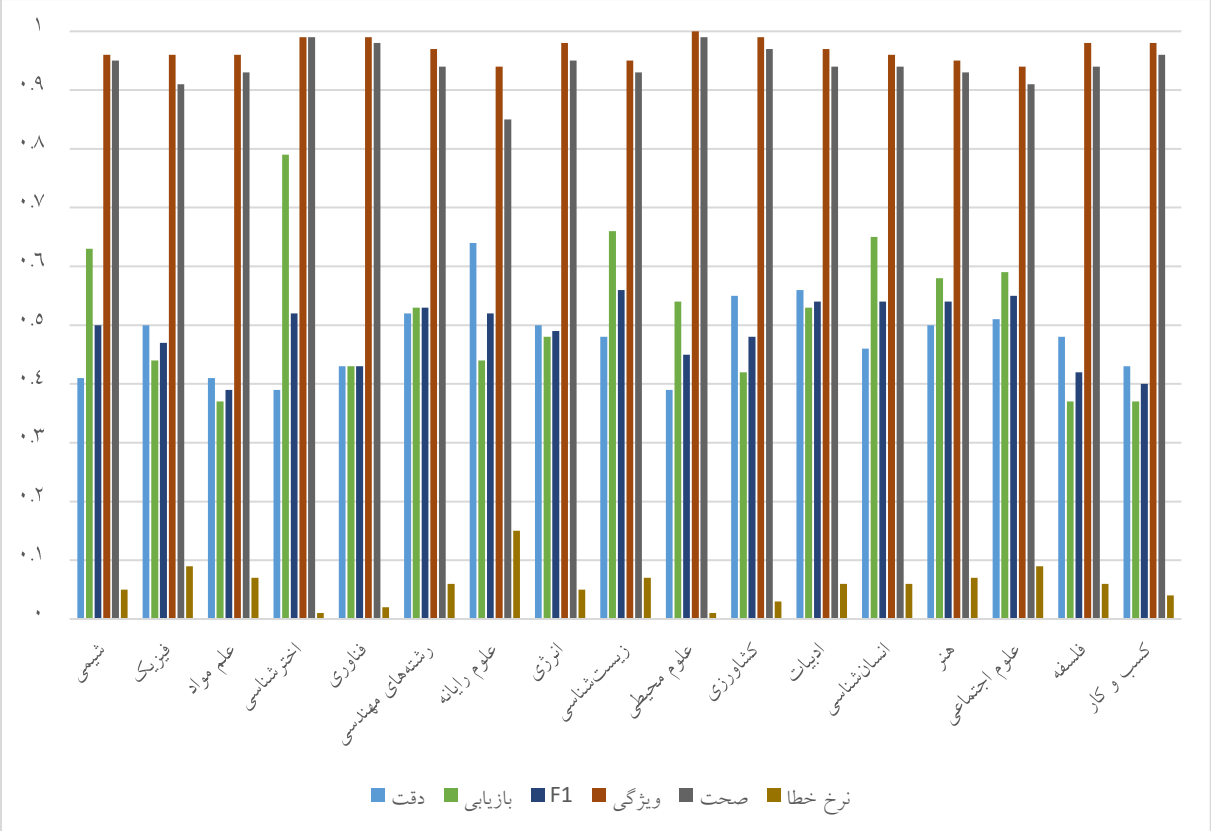
[۱] Haswell

(جدول ٤–٣) عملکرد کلی مدل کی‌انان در مجموعه داده‌ی فارسی.

+ ص	– ص	+ ک	– ک	دقت	بازیابی	F_1	ویژگی	صحت	خطا
١٤٣٥	٤٤١١٠	١٤١٠	١٤١٠	٠.٥٠	٠.٥٠	٠.٥٠	٠.٩٧	٠.٩٤	٠.٠٦

راهنمای جدول + ص: مثبت صحیح – ص: منفی صحیح + ک: مثبت کاذب – ک: منفی کاذب

بنابراین، دقت طبقه‌بندی مدل کی‌انان ٥٠٪ و صحت آن ٩٤٪، ارزیابی شده است. نمودار میله‌ای عملکرد مدل به تفکیک هر کلاس، در نمودار ٤–٣ آورده شده است.



(نمودار ٤–٣) نمودار عملکرد مجموعه داده‌ی فارسی مدل کی‌انان، به تفکیک کلاس.

همانطور که مشاهده می‌شود، این مدل در طبقه‌بندی رکوردهای کلاس زیست‌شناسی، با امتیاز اف–یک ٠.٥٦ و صحت ٠.٩٣، عملکرد مناسب‌تری را نسبت به سایر کلاس‌ها داشته است. اما در خصوص کلاس علم مواد، با امتیاز اف–یک ٠.٣٩، عملکرد ضعیف‌تری نسبت به سایر کلاس‌ها داشته است.

٤-٣-١-٢- مجموعه داده‌ی انگلیسی

فرایند آموزش این مدل در مجموعه داده‌ی انگلیسی نیز تقریباً به اندازه‌ی مجموعه داده‌ی فارسی طول کشید؛ یعنی حدود ۰.۰۳ ثانیه. این مدل پیش‌بینی کلاس ٤٦٩٣ رکورد مجموعه داده‌ی آزمایش را در ۵۵ ثانیه انجام داد؛ یعنی حدود ۱۲ میلی ثانیه برای هر رکورد.

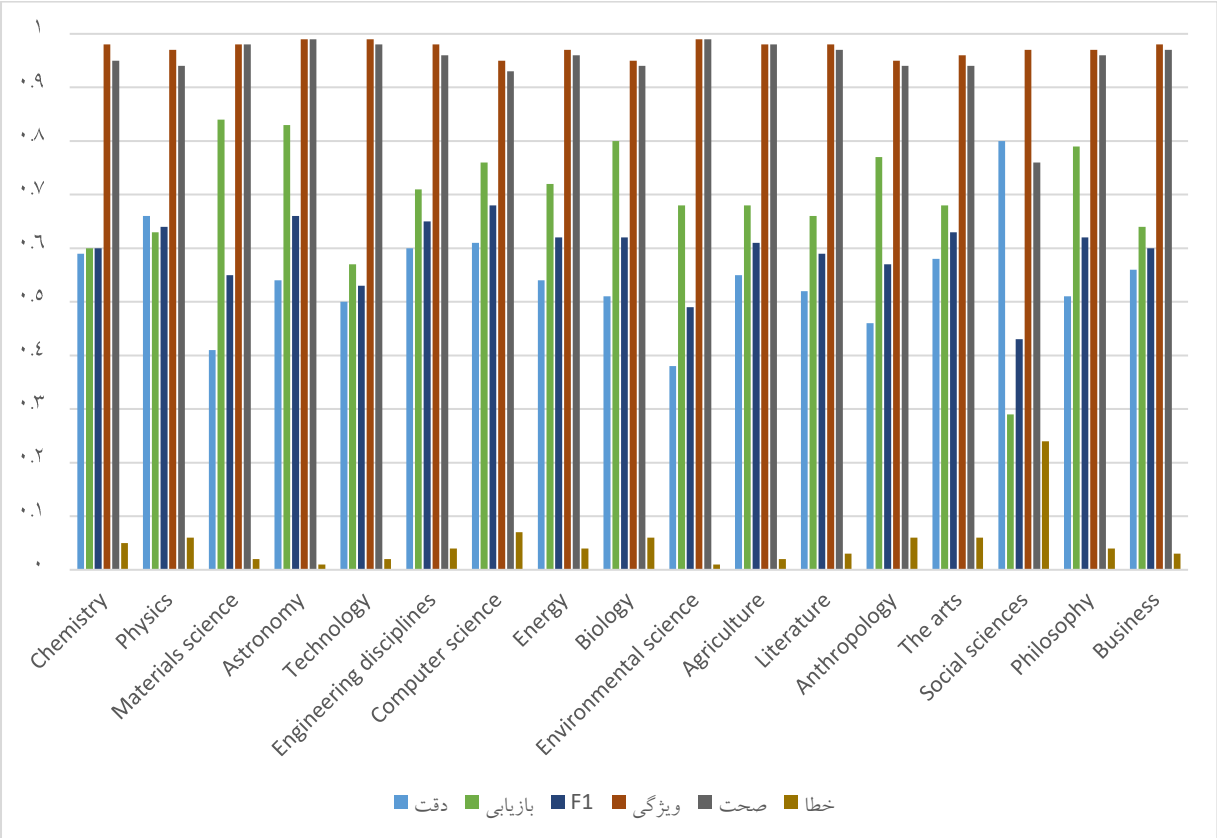
جدول ٤–٤ نشان‌دهنده‌ی عملکرد کلی مدل کی‌ان‌ان در طبقه‌بندی رکوردهای مجموعه داده‌ی انگلیسی است.

(جدول ٤–٤) عملکرد کلی مدل کی‌ان‌ان در مجموعه داده‌ی انگلیسی.

+ ص	– ص	+ ک	– ک	دقت	بازیابی	F_1	ویژگی	صحت	خطا
۲۷۱٦	۷۳۱۱۱	۱۹۷۷	۱۹۷۷	۰.۵۸	۰.۵۸	۰.۵۸	۰.۹۷	۰.۹۵	۰.۰۵

همانطور که دیده می‌شود، دقت کی‌ان‌ان در طبقه‌بندی رکوردهای انگلیسی ٪۵۸ و صحت آن ٪۹۵ محاسبه شده است.

نمودار میله‌ای عملکرد مدل به تفکیک هر کلاس، در نمودار ٤–٤، قابل مشاهده است.



(نمودار ٤–٤) نمودار عملکرد مجموعه داده‌ی انگلیسی مدل کی‌ان‌ان، به تفکیک کلاس.

درنتیجه، این مدل در طبقه‌بندی رکوردهای کلاس علوم رایانه، با امتیاز اف-یک ۰.٦٨ و صحت ۰.۹۳، بهترین عملکرد را در مقایسه با سایر کلاس‌ها داشته است. اما در خصوص کلاس علوم محیطی، با کسب امتیاز اف-یک ۰.٤۳، عملکرد ضعیف‌تری را نسبت به سایر کلاس‌ها داشته است.

٤-۳-۲- عملکرد مدل بیز ساده

بیز ساده در عین سادگی، مدلی باکیفیت به حساب می‌آید. نخست، عملکرد این مدل در مجموعه داده‌ی فارسی و سپس در مجموعه داده‌ی انگلیسی، بررسی خواهد شد:

٤-۳-۲-۱- مجموعه داده‌ی فارسی

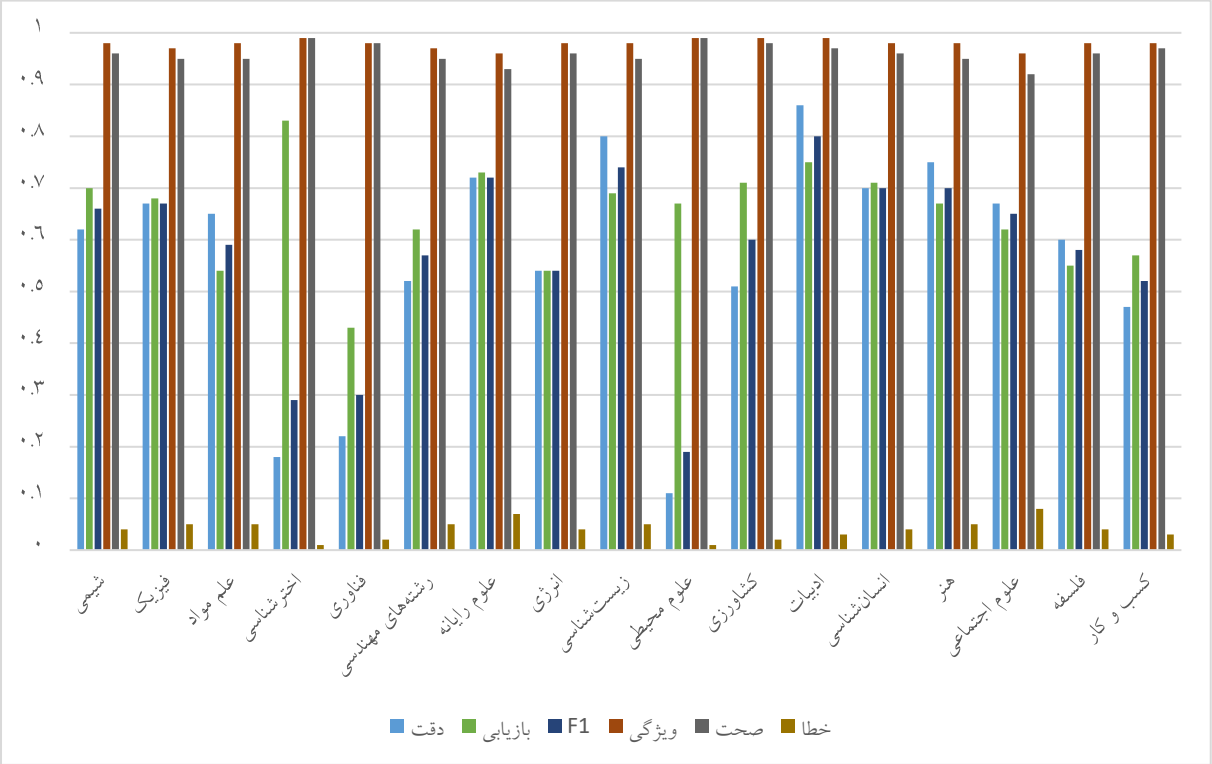
فرایند آموزش این مدل حدود ۲ ثانیه زمان برده است. پیش‌بینی ۲۸٤٥ رکورد مجموعه داده‌ی آزمایش، ۳ ثانیه به طول انجامیده است؛ یعنی تنها حدود ۱ میلی ثانیه برای هر رکورد.

در جدول ٤—٥، عملکرد کلی مدل بیز ساده در طبقه‌بندی رکوردهای مجموعه داده‌ی فارسی، آمده است.

(جدول ٤—٥) عملکرد کلی مدل بیز ساده در مجموعه داده‌ی فارسی.

+ ص	- ص	+ ک	- ک	دقت	بازیابی	F_1	ویژگی	صحت	خطا
۳۰۸۹	۷۳٤۸٤	۱٦۰٤	۱٦۰٤	۰.٦٦	۰.٦٦	۰.٦٦	۰.۹۸	۰.۹٦	۰.۰٤

بنابراین، دقت طبقه‌بندی این مدل برای رکوردهای فارسی ٦٦٪ و صحت آن ۹٦٪ بوده است. نمودار میله‌ای عملکرد این مدل، به تفکیک هر کلاس، در نمودار ٤—٥ آورده شده است:



(نمودار ٤–٥) نمودار عملکرد مجموعه داده‌ی فارسی مدل بیز ساده، به تفکیک کلاس.

همانطور که مشاهده می‌شود، این مدل در طبقه‌بندی رکوردهای کلاس ادبیات، با امتیاز اف–یک ۰.۸ و صحت ۰.۹۷، بهترین عملکرد را نشان داده است. اما ضعیف‌ترین عملکرد آن، در خصوص کلاس علوم محیطی، با امتیاز اف–یک ۰.۱۹، بوده است. این مدل همچنین، عملکرد ضعیفی در خصوص کلاس اخترشناسی داشته است.

٤–۳–۲–۲– مجموعه داده‌ی انگلیسی

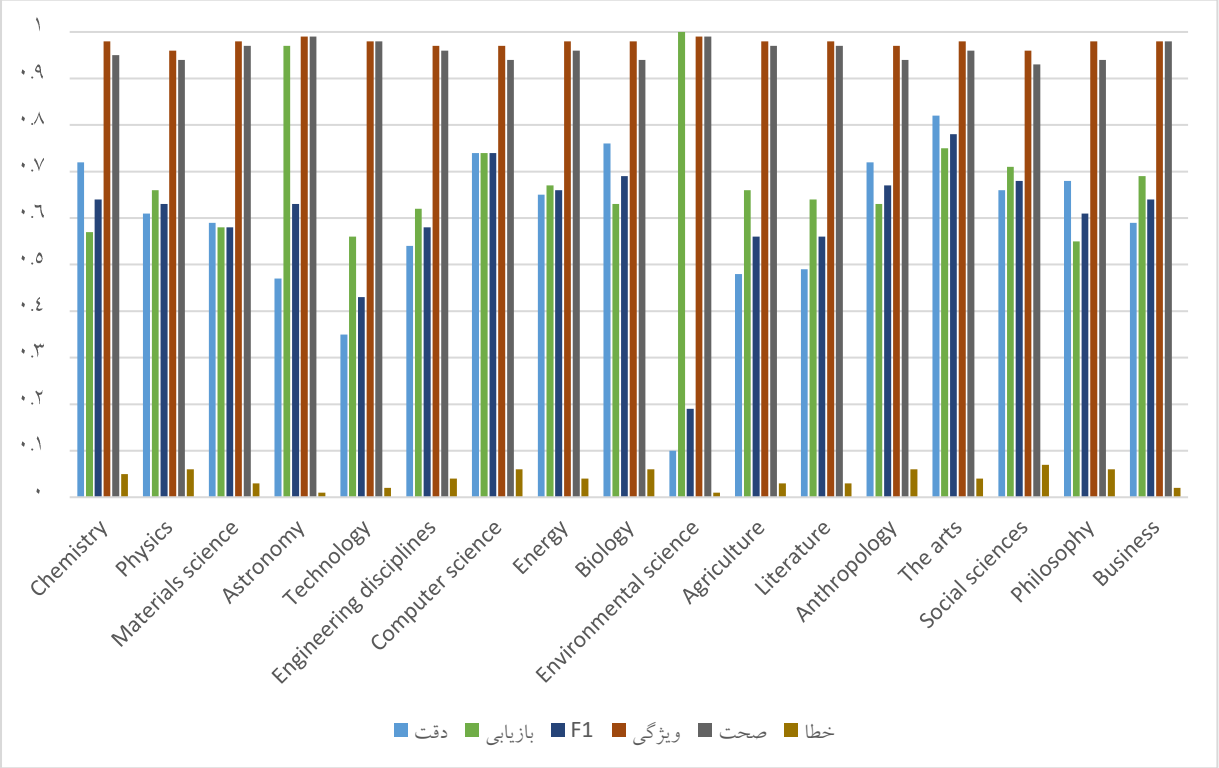
فرایند آموزش این مدل ۳ ثانیه زمان برده است. پیش‌بینی کلاس برای ٤٦۹۳ رکورد مجموعه داده‌ی آزمایش، ٤ ثانیه به طول انجامیده است؛ یعنی تنها حدود ۰.۲ میلی ثانیه برای هر رکورد.

جدول ٤–٦، نشان‌دهنده‌ی عملکرد کلی مدل بیز ساده در طبقه‌بندی رکوردهای مجموعه داده‌ی انگلیسی است:

(جدول ٤–٦) عملکرد کلی مدل بیز ساده در مجموعه داده‌ی انگلیسی.

+ ص	– ص	+ ک	– ک	دقت	بازیابی	F_1	ویژگی	صحت	خطا
۳۰۸۹	۷۳٤۸٤	۱٦۰٤	۱٦۰٤	۰.٦٦	۰.٦٦	۰.٦٦	۰.۹۸	۰.۹٦	۰.۰٤

بنابراین، دقت مدل بیز ساده در طبقه‌بندی رکوردهای انگلیسی ٦٦٪ و صحت آن ٩٦٪ بوده است. نمودار میله‌ای عملکرد مدل به تفکیک هر کلاس، در نمودار ٤–٦ آورده شده است.



(نمودار ٤–٦) نمودار عملکرد مجموعه داده‌ی انگلیسی مدل بیز ساده، به تفکیک کلاس.

همانطور که مشاهده می‌شود، این مدل در طبقه‌بندی رکوردهای کلاس هنر، با امتیاز اف–یک ۰.۷۸ و صحت ۰.۹٦، عملکرد مناسب‌تری را نسبت به سایر کلاس‌ها داشته است. اما در خصوص کلاس علوم محیطی، با امتیاز اف–یک ۰.۱۹، عملکرد ضعیف‌تری نسبت به سایر کلاس‌ها داشته است.

٤–۳–۳– عملکرد مدل رگرسیون لجستیک

این مدل طبقه‌بندی‌کننده‌ای بسیار قدرتمند است و تا حدودی به شبکه‌های عصبی نیز شبیه است. نخست عملکرد این مدل در مجموعه داده‌ی فارسی و سپس در مجموعه داده‌ی انگلیسی، بررسی خواهد شد:

٤–۳–۳–۱– مجموعه داده‌ی فارسی

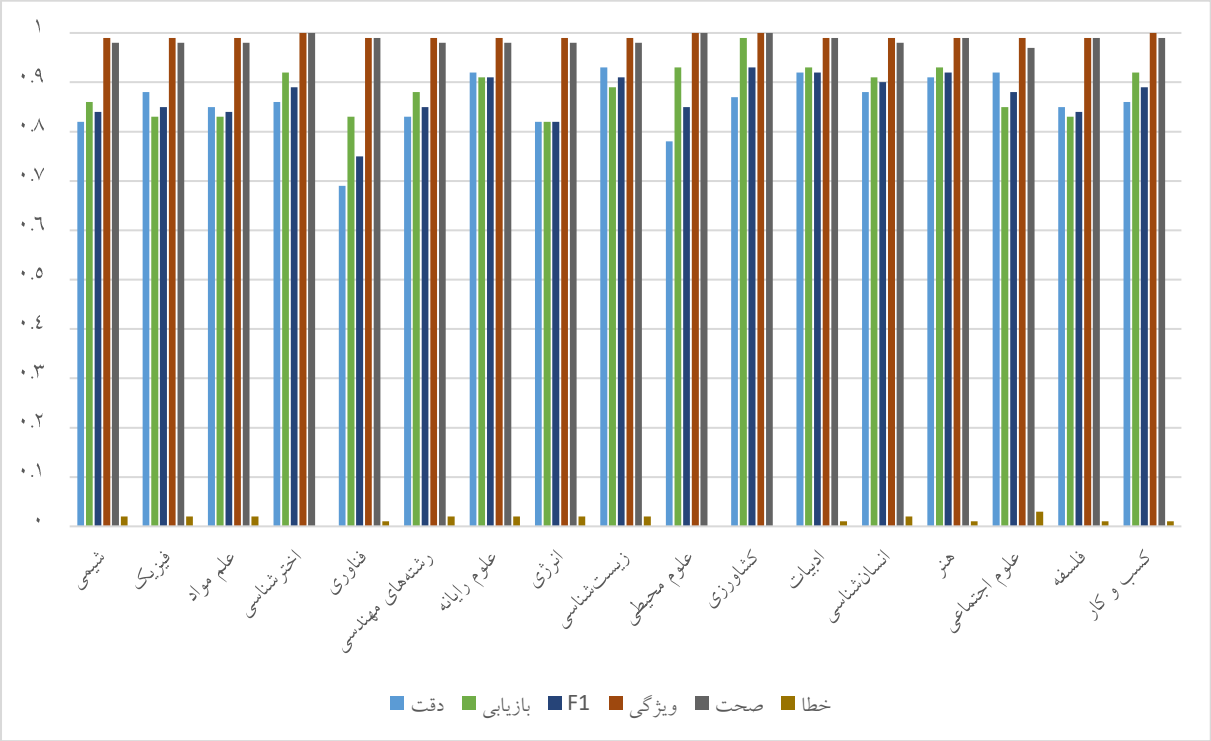
فرایند آموزش این مدل ۱۱۵ ثانیه زمان برد و به آن اجازه داده شد که تا جایی که سخت‌افزار گنجایش دارد، فرایند یادگیری را پیش‌ببرد. پیش‌بینی ۲۸٤٥ رکورد مجموعه داده‌ی آزمایش، حدود ۰.٥٥ ثانیه به طول انجامید؛ یعنی تنها حدود ۰.۲ میلی ثانیه برای هر رکورد.

در جدول ٤–٧، عملکرد کلی مدل رگرسیون لجستیک در طبقه‌بندی رکوردهای مجموعه داده‌ی فارسی، قابل مشاهده می‌باشد.

(جدول ٤–٧) عملکرد کلی مدل رگرسیون لجستیک در مجموعه داده‌ی فارسی.

+ ص	– ص	+ ک	– ک	دقت	بازیابی	F_1	ویژگی	صحت	خطا
٢٥٠٦	٤٥١٨١	٣٣٩	٣٣٩	٠.٨٨	٠.٨٨	٠.٨٨	٠.٩٩	٠.٩٩	٠.٠١

همانطور که از نتایج جدول پیداست، دقت طبقه‌بندی این مدل ٨٨٪ و صحت آن ٩٩٪ ارزیابی شده است. نمودار میله‌ای عملکرد مدل به تفکیک هر کلاس، در نمودار ٤–٧ آورده شده است.



(نمودار ٤–٧) نمودار عملکرد مجموعه داده‌ی فارسی مدل رگرسیون لجستیک، به تفکیک کلاس.

همانطور که مشاهده می‌شود، این مدل در تمامی کلاس‌ها، به‌جز کلاس فناوری و علوم محیطی، دقتی بالای ٨٠٪ داشته است. در خصوص کلاس علوم محیطی، این مدل توانسته طبقه‌بندی رکوردها را با دقت ٧٨٪ و امتیاز اف–یک ٠.٨٥ انجام دهد. بهترین عملکرد این مدل مربوط به کلاس کشاورزی بوده که رکوردهای آن‌را با امتیاز اف–یک ٠.٩٣ و صحت حدود ١٠٠٪، طبقه‌بندی کرده است.

٤-٣-٣-٢- مجموعه داده‌ی انگلیسی

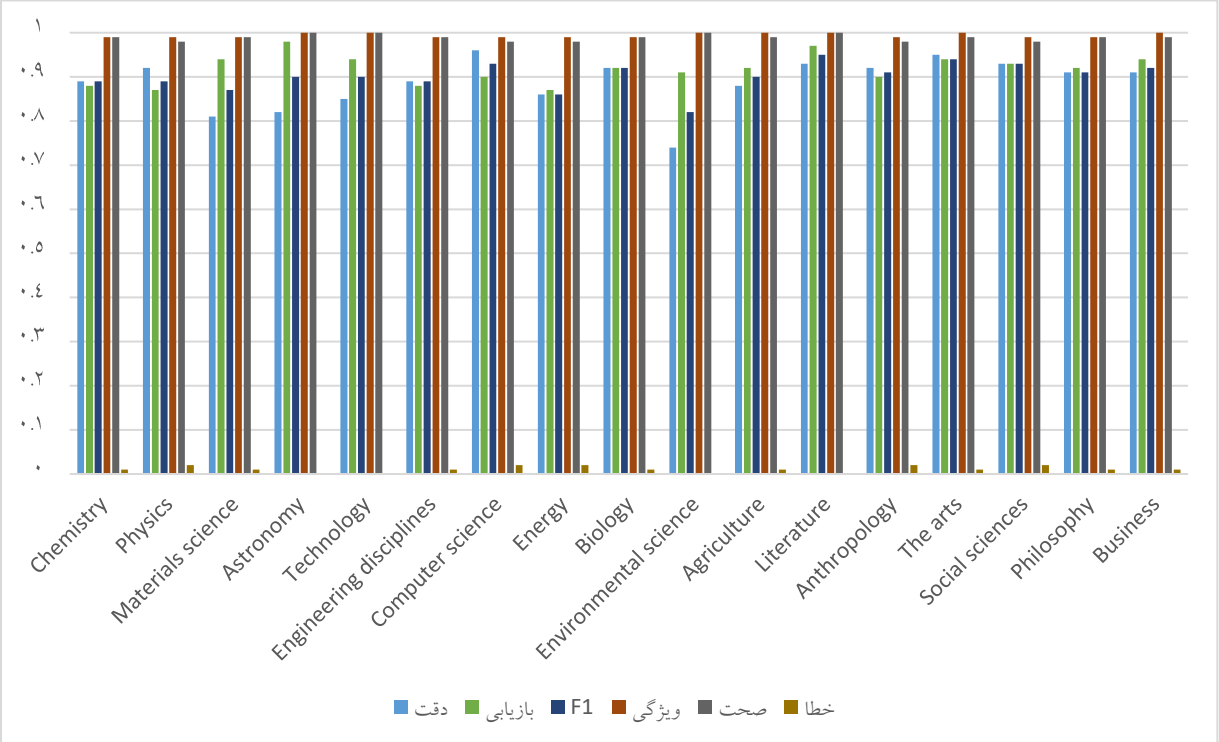
برای مجموعه داده‌ی انگلیسی، فرایند آموزش رگرسیون لجستیک، ۱۵۰ ثانیه زمان برد. پیش‌بینی ٤٦٩٣ رکورد مجموعه داده‌ی آزمایش، حدود ۰.٤ ثانیه به طول انجامید؛ یعنی تنها حدود ۰.۱ میلی ثانیه برای هر رکورد.

جدول ٤–۸، نشان‌دهنده‌ی عملکرد کلی مدل رگرسیون لجستیک در طبقه‌بندی رکوردهای مجموعه داده‌ی انگلیسی است.

(جدول ٤–۸) عملکرد کلی مدل رگرسیون لجستیک در مجموعه داده‌ی انگلیسی.

+ ص	– ص	+ ک	– ک	دقت	بازیابی	F_1	ویژگی	صحت	خطا
٤٢٧٤	٧٤٦٦٩	٤١٩	٤١٩	۰.۹۱	۰.۹۱	۰.۹۱	۰.۹۹	۰.۹۹	۰.۰۱

از جدول فوق پیداست که دقت طبقه‌بندی ٪۹۱ و صحت ٪۹۹، توسط این مدل حاصل شده است. نمودار میله‌ای عملکرد مدل به تفکیک هر کلاس، در نمودار ٤–۸ قابل مشاهده است.



(نمودار ٤–۸) نمودار عملکرد مجموعه داده‌ی انگلیسی مدل رگرسیون لجستیک، به تفکیک کلاس.

همانطور که ملاحظه می‌شود، بهترین عملکرد این مدل به کلاس ادبیات اختصاص دارد. این مدل توانسته که در این کلاس امتیاز اف–یک ۰.۹۵ و صحت تقریباً یک (کامل) را بدست آورد. ضعیف‌ترین عملکرد این مدل مربوط به کلاس علوم محیطی است که امتیاز اف–یک طبقه‌بندی رکوردهای آن ۰.۸۲ ارزیابی شده است. لازم به ذکر است که امتیاز اف–یک این مدل در طبقه‌بندی تمام کلاس‌ها، بالای ۰.۸ بوده است.

٤-٣-٤- عملکرد مدل اس‌وی‌ام

در این تحقیق، دو نوع مدل اس‌وی‌ام پیاده‌سازی شده‌اند که عملکرد اولی در این بخش بررسی می‌شود. نخست، عملکرد این مدل در مجموعه داده‌ی فارسی و سپس در مجموعه داده‌ی انگلیسی، بررسی خواهد شد:

٤-٣-٤-١- مجموعه داده‌ی فارسی

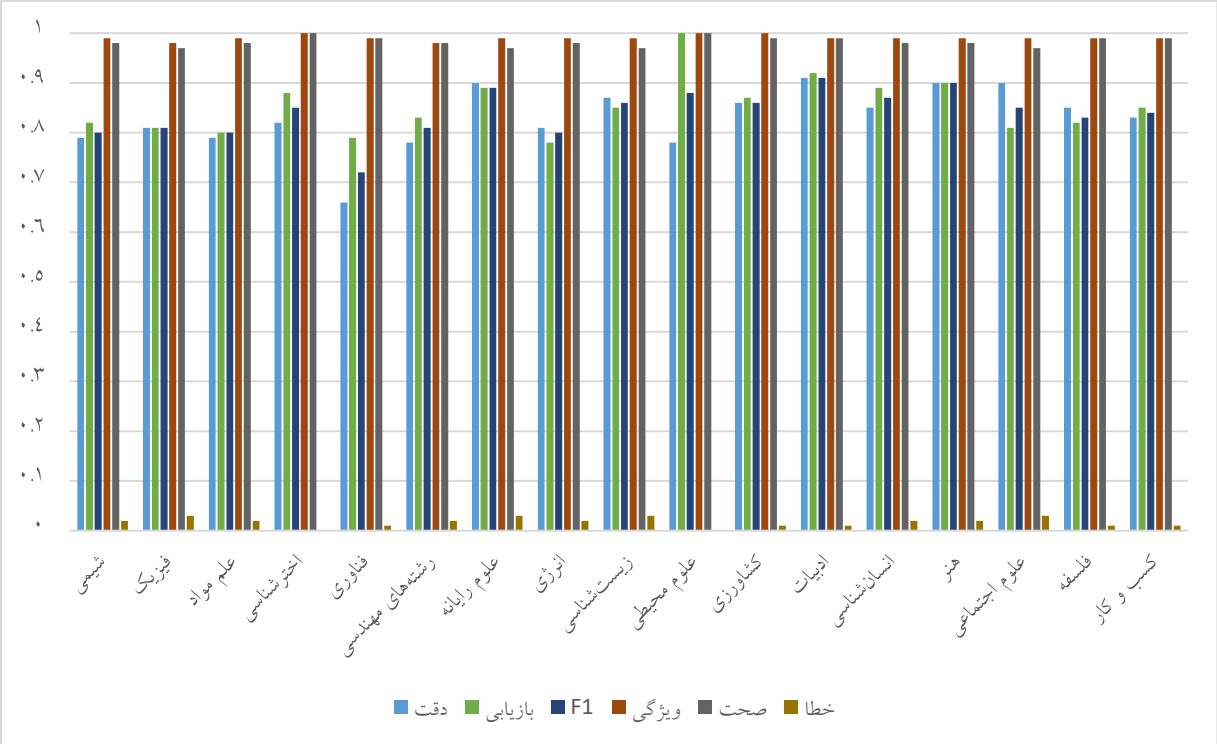
فرایند آموزش این مدل ٨٨٦ ثانیه (حدود ١٥ دقیقه) زمان برد. پیش‌بینی کلاس ٢٨٤٥ رکورد مجموعه داده‌ی آزمایش، ١٦٣ ثانیه به طول انجامید؛ یعنی حدود ٥٧ میلی ثانیه برای هر رکورد.

در جدول ٤-٩، عملکرد کلی مدل اس‌وی‌ام در طبقه‌بندی رکوردهای مجموعه داده‌ی فارسی، نشان داده شده است.

(جدول ٤-٩) عملکرد کلی مدل اس‌وی‌ام در مجموعه داده‌ی فارسی.

+ ص	- ص	+ ک	- ک	دقت	بازیابی	F_1	ویژگی	صحت	خطا
٢٤١٦	٤٥٠٩١	٤٢٩	٤٢٩	٠.٨٥	٠.٨٥	٠.٨٥	٠.٩٩	٠.٩٨	٠.٠٢

دقت طبقه‌بندی این مدل حدود ٨٥٪ و صحت طبقه‌بندی ٩٨٪ برآورد شده. نمودار میله‌ای عملکرد این مدل به تفکیک هر کلاس، در نمودار ٤-٩ آورده شده است.



(نمودار ٤-٩) نمودار عملکرد مجموعه داده‌ی فارسی مدل اس‌وی‌ام، به تفکیک کلاس.

همانطور که مشاهده می‌شود، این مدل رکوردهای کلاس ادبیات را با امتیاز اف-یک ۰.۹ و صحت ۰.۹۹، طبقه‌بندی کرده است. ضعیف‌ترین عملکرد آن، در خصوص کلاس فناوری است که امتیاز اف-یک ۰.۷۲ را کسب کرده است.

٤-۳-٤-۲- مجموعه داده‌ی انگلیسی

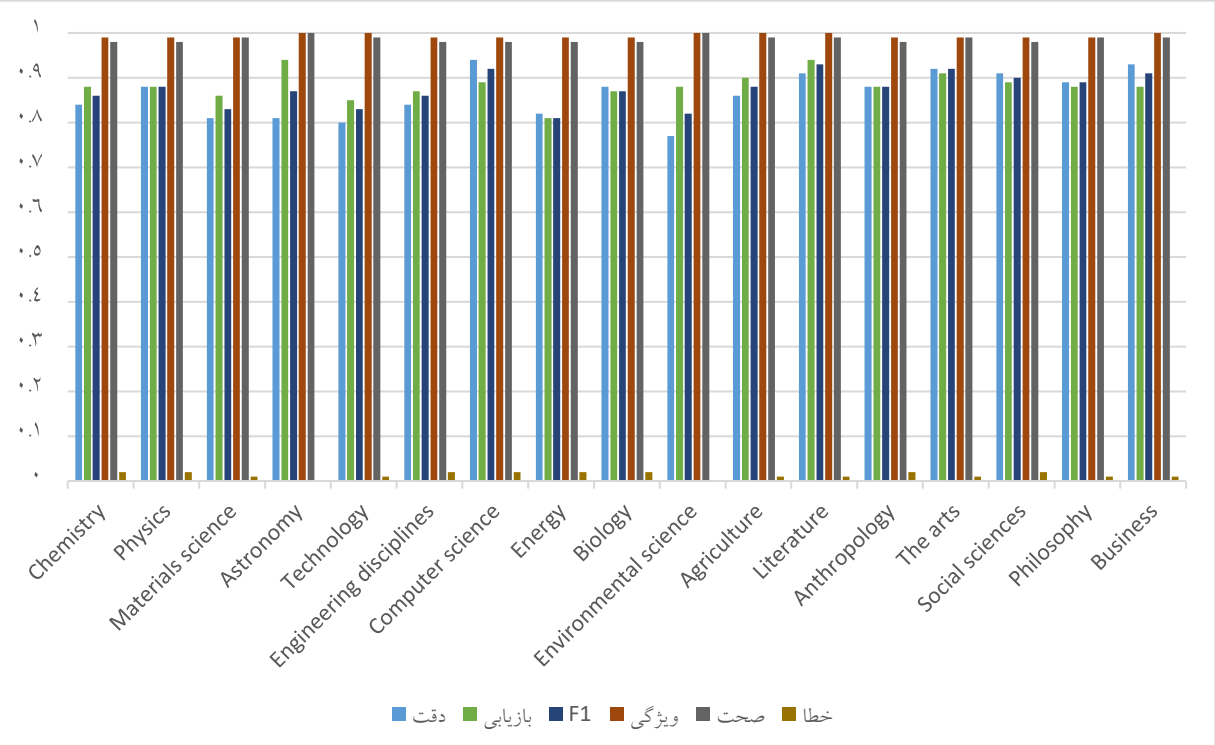
فرایند آموزش این مدل ۱۹٤٤ ثانیه (حدود ۳۲ دقیقه) زمان برد. پیش‌بینی کلاس ٤٦۹۳ رکورد مجموعه داده‌ی آزمایش توسط این مدل، ۵۵۳ ثانیه به طول انجامید؛ یعنی حدود ۱۱۸ میلی ثانیه برای هر رکورد.

جدول ٤-۱۰ نشان‌دهنده‌ی عملکرد کلی مدل اس‌وی‌ام در طبقه‌بندی رکوردهای مجموعه داده‌ی انگلیسی است:

(جدول ٤-۱۰) عملکرد کلی مدل اس‌وی‌ام در مجموعه داده‌ی انگلیسی.

+ ص	- ص	+ ک	- ک	دقت	بازیابی	F_1	ویژگی	صحت	خطا
٤۱٤۰	۷٤۵۳۵	۵۵۳	۵۵۳	۰.۸۸	۰.۸۸	۰.۸۸	۰.۹۹	۰.۹۹	۰.۰۱

بنابراین، مدل اس‌وی‌ام توانست به دقت ٪۸۸ و صحت ٪۹۹ در طبقه‌بندی رکوردهای مجموعه داده‌ی آزمایش انگلیسی، دست یابد. نمودار میله‌ای عملکرد مدل به تفکیک هر کلاس، در نمودار ٤-۱۰ آورده شده است.



(نمودار ٤-۱۰) نمودار عملکرد مجموعه داده‌ی انگلیسی مدل اس‌وی‌ام، به تفکیک کلاس.

همانطور که مشاهده می‌شود، این مدل در طبقه‌بندی رکوردهای کلاس ادبیات، با امتیاز اف-یک ۰.۹۲ و صحت ۰.۹۹، بهترین عملکرد را نسبت به سایر کلاس‌ها داشته است. بدترین عملکرد این مدل مربوط به کلاس انرژی است که امتیاز اف-یک ۰.۸۱ و صحت ۰.۹۸ را در طبقه‌بندی رکوردهای آن، کسب کرده است.

٤-۳-۵- عملکرد مدل اس‌وی‌ام هسته‌ای

در این بخش، عملکرد دومین مدل اس‌وی‌ام، مورد تحلیل قرار خواهد گرفت. نخست، عملکرد این مدل در مجموعه داده‌ی فارسی و سپس در مجموعه داده‌ی انگلیسی، بررسی خواهد شد:

٤-۳-۵-۱- مجموعه داده‌ی فارسی

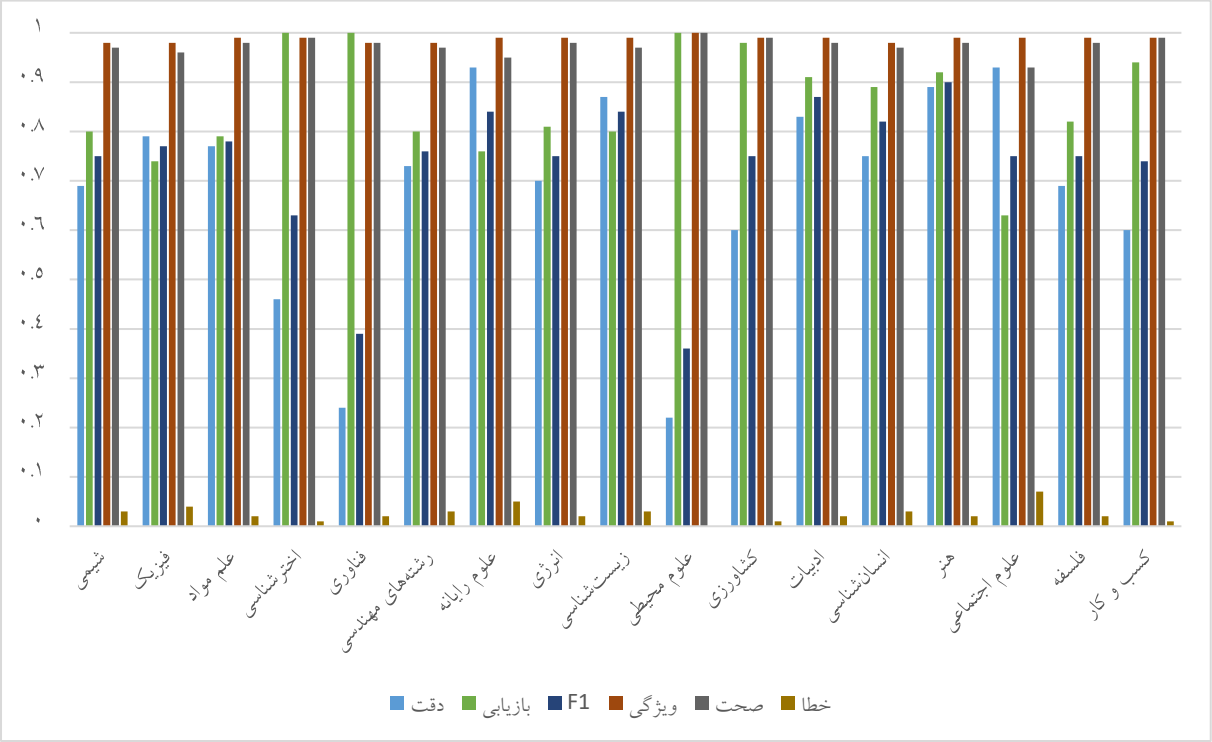
فرایند آموزش این مدل ۱۳۵۲ ثانیه (معادل ۲۲.۵ دقیقه) زمان برده است. پیش‌بینی ۲۸٤۵ رکورد مجموعه داده‌ی آزمایش، ۳٦۰ ثانیه به طول انجامید؛ یعنی حدود ۱۲٦ میلی ثانیه برای هر رکورد.

در جدول ٤-۱۱، عملکرد کلی مدل اس‌وی‌ام هسته‌ای در طبقه‌بندی رکوردهای مجموعه داده‌ی فارسی، آمده است.

(جدول ٤-۱۱) عملکرد کلی مدل اس‌وی‌ام هسته‌ای در مجموعه داده‌ی فارسی.

+ ص	- ص	+ ک	- ک	دقت	بازیابی	F_1	ویژگی	صحت	خطا
۲۲۵۲	٤٤۹۲۷	۵۹۳	۵۹۳	۰.۷۹	۰.۷۹	۰.۷۹	۰.۹۹	۰.۹۸	۰.۰۲

این مدل دقتی معادل ٪۷۹ و صحت ٪۹۸ در طبقه‌بندی رکوردهای فارسی داشته است. نمودار میله‌ای عملکرد مدل به تفکیک هر کلاس، در نمودار ٤-۱۱ آورده شده است.



(نمودار ٤–١١) نمودار عملکرد مجموعه داده‌ی فارسی مدل اس‌وی‌ام هسته‌ای، به تفکیک کلاس.

بهترین عملکرد این مدل در طبقه‌بندی رکوردهای کلاس هنر با امتیاز اف–یک ۰.۹ و صحت ۰.۹۸، بوده است. دقت طبقه‌بندی این مدل برای رکوردهای کلاس علوم محیطی، ۰.۲۲ و امتیاز اف–یک آن ۰.۳٦ بوده است که ضعیف‌ترین عملکرد این مدل در میان کلاس‌ها به‌حساب می‌آید.

٤–۳–٥–۲– مجموعه داده‌ی انگلیسی

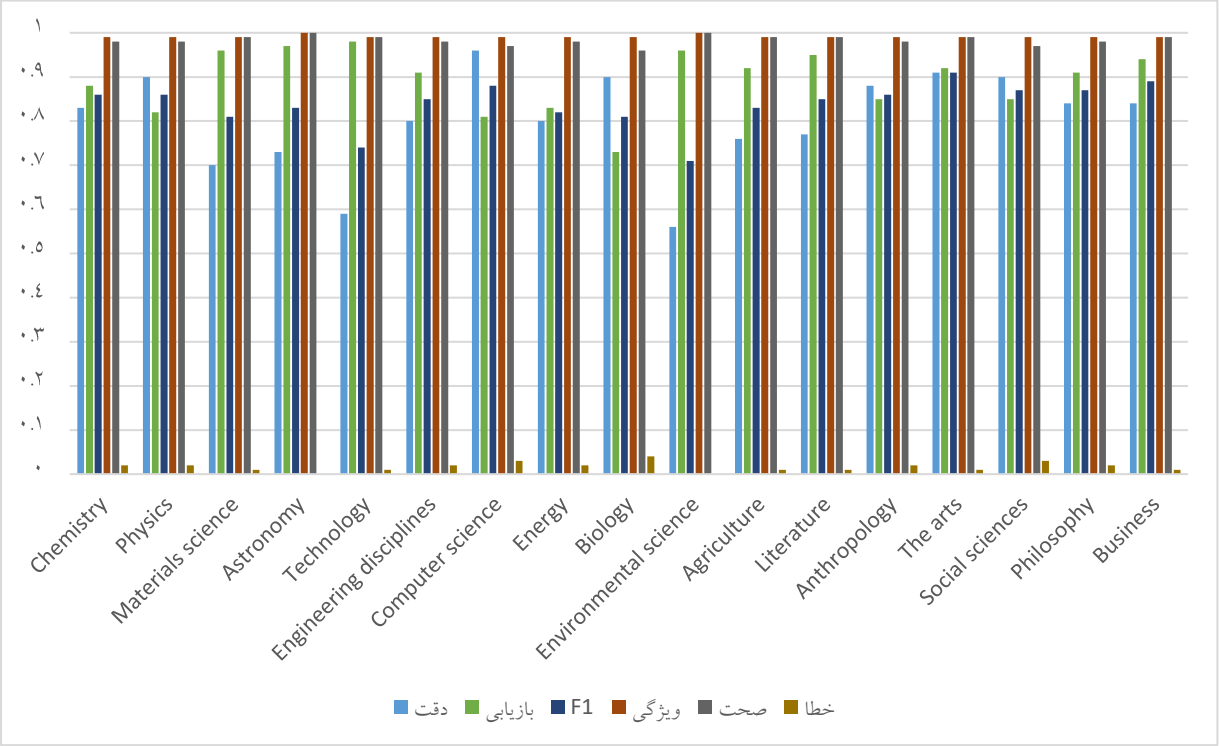
فرایند آموزش این مدل ٤٦٤۹ ثانیه (حدود ۱ ساعت و ۱۷ دقیقه) زمان برده است. پیش‌بینی کلاس ٤٦۹۳ رکورد مجموعه داده‌ی آزمایش، ۱۰۲۹ ثانیه به طول انجامید؛ یعنی حدود ۲۱۹ میلی ثانیه برای هر رکورد.

جدول ٤–۱۲، نشان‌دهنده‌ی عملکرد کلی مدل اس‌وی‌ام هسته‌ای در طبقه‌بندی رکوردهای مجموعه داده‌ی انگلیسی است.

(جدول ٤–۱۲) عملکرد کلی مدل اس‌وی‌ام هسته‌ای در مجموعه داده‌ی انگلیسی.

+ ص	– ص	+ ک	– ک	دقت	بازیابی	F_1	ویژگی	صحت	خطا
٤۰۱۷	۷٤٤۱۲	٦۷٦	٦۷٦	۰.۸٦	۰.۸٦	۰.۸٦	۰.۹۹	۰.۹۸	۰.۰۲

بنابراین، دقت طبقه‌بندی رکوردهای انگلیسی توسط اس‌وی‌ام هسته‌ای ٪۸٦ و صحت عملکرد آن ٪۹۸ محاسبه شده است. نمودار میله‌ای عملکرد مدل به تفکیک هر کلاس، در نمودار ٤–۱۲ آماده است.



(نمودار ٤–١٢) نمودار عملکرد مجموعه داده‌ی انگلیسی مدل اس‌وی‌ام هسته‌ای، به تفکیک کلاس.

همانطور که مشاهده می‌شود، این مدل در طبقه‌بندی رکوردهای کلاس هنر، با امتیاز اف–یک ۰.۹۱ و صحت ۰.۹۹، عملکرد مناسب‌تری را نسبت به سایر کلاس‌ها داشته است. ضعیف‌ترین عملکرد این مدل مربوط به کلاس علوم محیطی بوده که امتیاز اف–یک ۰.۷۱ را در طبقه‌بندی رکوردهای آن، بدست آورده است.

٤–۳–٦– عملکرد مدل درخت تصمیم

مدل درخت تصمیم مدلی نسبتاً ساده و سبک است که در این بخش، عملکرد آن مورد بررسی قرار خواهد گرفت. نخست، عملکرد این مدل در مجموعه داده‌ی فارسی و سپس در مجموعه داده‌ی انگلیسی، بررسی خواهد شد:

٤–۳–٦–۱– مجموعه داده‌ی فارسی

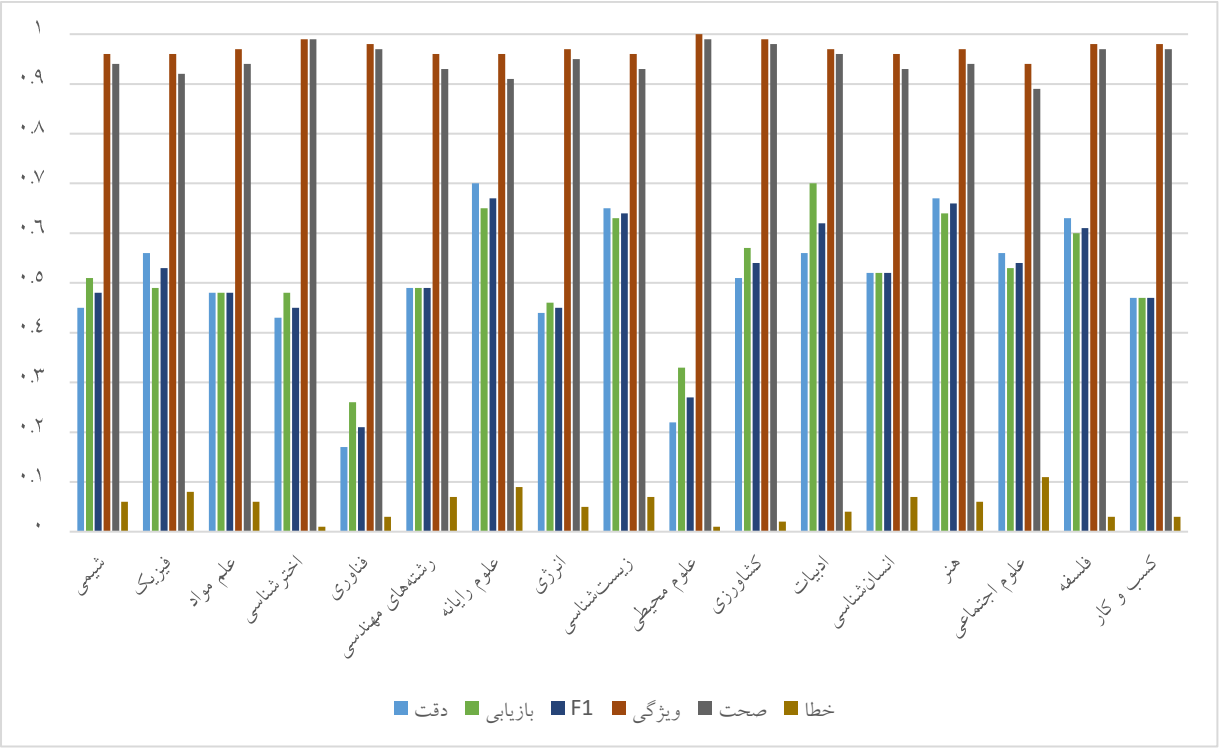
فرایند آموزش این مدل ۷۵ ثانیه زمان برد. فرایند پیش‌بینی کلاس ۲۸٤٥ رکورد مجموعه داده‌ی آزمایش، حدود ۰.۳ ثانیه به طول انجامید؛ یعنی تنها حدود ۰.۱ میلی ثانیه برای هر رکورد.

در جدول ٤–۱۳، عملکرد کلی مدل درخت تصمیم در طبقه‌بندی رکوردهای مجموعه داده‌ی فارسی، آمده است.

(جدول ٤–١٣) عملکرد کلی مدل درخت تصمیم در مجموعه داده‌ی فارسی.

+ ص	– ص	+ ک	– ک	دقت	بازیابی	F_1	ویژگی	صحت	خطا
۱۵۹۰	٤٤۲٦٥	۱۲٥٥	۱۲٥٥	۰.٥٦	۰.٥٦	۰.٥٦	۰.۹۷	۰.۹٥	۰.۰٥

همانطور که مشاهده می‌شود، دقت این مدل در طبقه‌بندی رکوردهای آزمایش فارسی ٥٦٪ و صحت آن ۹٥٪ بوده است. نمودار میله‌ای عملکرد مدل به تفکیک هر کلاس، در نمودار ٤–۱۳ نشان شده است:



(نمودار ٤–۱۳) نمودار عملکرد مجموعه داده‌ی فارسی مدل درخت تصمیم، به تفکیک کلاس.

با توجه به نمودار، این مدل در طبقه‌بندی رکوردهای کلاس علوم رایانه، با امتیاز اف–یک ۰.٦۷ و صحت ۰.۹۱، عملکرد بهتری نسبت به سایر کلاس‌ها داشته است. اما ضعیف‌ترین عملکرد آن، در خصوص کلاس فناوری، با امتیاز اف–یک ۰.۲۱ و دقت ۰.۱۷ می‌باشد.

٤–۳–٦–۲– مجموعه داده‌ی انگلیسی

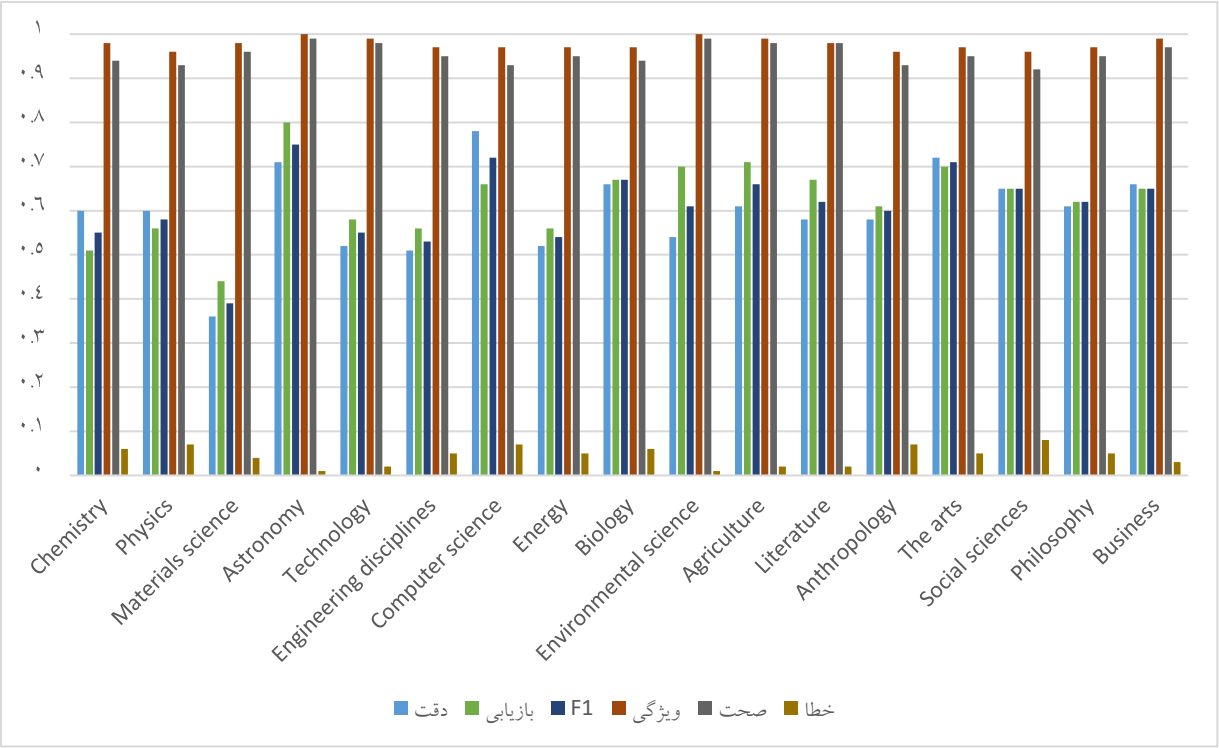
فرایند آموزش این مدل ۲۹۷ ثانیه (حدود ٥ دقیقه) زمان برده است. فرایند پیش‌بینی کلاس ٤٦۹۳ رکورد مجموعه داده‌ی آزمایش، حدود ۱ ثانیه به طول انجامید؛ یعنی تنها حدود ۰.۲ میلی ثانیه برای هر رکورد.

جدول ٤–۱٤ نشان‌دهنده‌ی عملکرد کلی مدل درخت تصمیم در طبقه‌بندی رکوردهای مجموعه داده‌ی انگلیسی است.

(جدول ٤–١٤) عملکرد کلی مدل درخت تصمیم در مجموعه داده‌ی انگلیسی.

+ ص	– ص	+ ک	– ک	دقت	بازیابی	F_1	ویژگی	صحت	خطا
۲۹۳۲	۷۳۳۲۷	۱۷۶۱	۱۷۶۱	۰.۶۲	۰.۶۲	۰.۶۲	۰.۹۸	۰.۹۶	۰.۰٤

بنابراین، این مدل با دقت ٦٢٪ و صحت ٩٦٪، رکوردهای انگلیسی را طبقه‌بندی کرده است. نمودار میله‌ای عملکرد مدل به تفکیک هر کلاس، در نمودار ٤–١٤ قابل مشاهده است.



(نمودار ٤–١٤) نمودار عملکرد مجموعه داده‌ی انگلیسی مدل درخت تصمیم، به تفکیک کلاس.

همانطور که مشاهده می‌شود، بهترین عملکرد این مدل در طبقه‌بندی رکوردهای کلاس اخترشناسی، با امتیاز اف–یک ۰.۷۵ و صحت ۰.۹۹ بوده است. ضعیف‌ترین عملکرد این مدل در خصوص کلاس علم مواد بوده که رکوردهای آن‌را با امتیاز اف–یک ۰.۳۹، طبقه‌بندی کرده است.

٤–۳–۷– عملکرد مدل جنگل تصادفی

عملکرد مدل جنگل تصادفی که مبتنی بر مدل درخت تصمیم است، در این بخش مورد بررسی قرار خواهد گرفت. نخست، عملکرد این مدل در مجموعه داده‌ی فارسی و سپس در مجموعه داده‌ی انگلیسی، بررسی خواهد شد:

٤-٣-٧-١- مجموعه داده‌ی فارسی

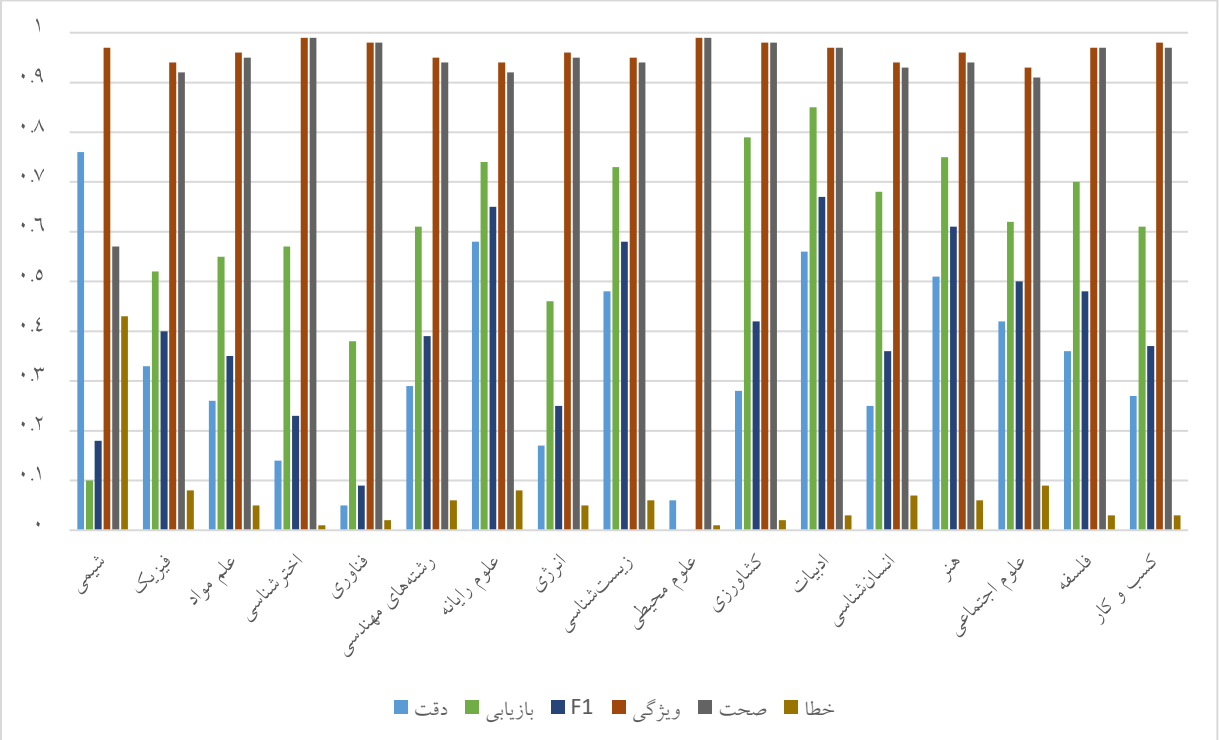
فرایند آموزش این مدل، تنها ۵.۵ ثانیه زمان برده است. پیش‌بینی کلاس ۲۸٤۵ رکورد مجموعه داده‌ی آزمایش، حدود ۰.۱ ثانیه به طول انجامید؛ یعنی تنها حدود ۰.۰۳ میلی ثانیه برای هر رکورد.

در جدول ٤-۱۵، عملکرد کلی مدل جنگل تصادفی در طبقه‌بندی رکوردهای مجموعه داده‌ی فارسی، نشان داده شده است.

(جدول ٤-۱۵) عملکرد کلی مدل جنگل تصادفی در مجموعه داده‌ی فارسی.

+ ص	- ص	+ ک	- ک	دقت	بازیابی	F_1	ویژگی	صحت	خطا
۱۱۷۱	٤۳۸٤٦	۱٦۷٤	۱٦۷٤	۰.٤۱	۰.٤۱	۰.٤۱	۰.۹٦	۰.۹۳	۰.۰۷

همانطور که ملاحظه می‌شود، دقت طبقه‌بندی رکوردهای فارسی توسط این مدل ٤۱٪ و صحت آن ۹۳٪ محاسبه شده است. نمودار میله‌ای عملکرد این مدل به تفکیک هر کلاس، در نمودار ٤-۱۵ آورده شده است.



(نمودار ٤-۱۵) نمودار عملکرد مجموعه داده‌ی فارسی مدل جنگل تصادفی ، به تفکیک کلاس.

همانطور که مشاهده می‌شود، بهترین عملکرد این مدل در خصوص کلاس ادبیات بوده که توانسته به امتیاز اف-یک ۰.٦۷ و صحت ۰.۹۷ دست یابد. بدترین عملکرد این مدل، با امتیاز اف-یک صفر و دقت ٦٪، مربوط به کلاس علوم محیطی است.

٤-٣-٧-٢- مجموعه داده‌ی انگلیسی

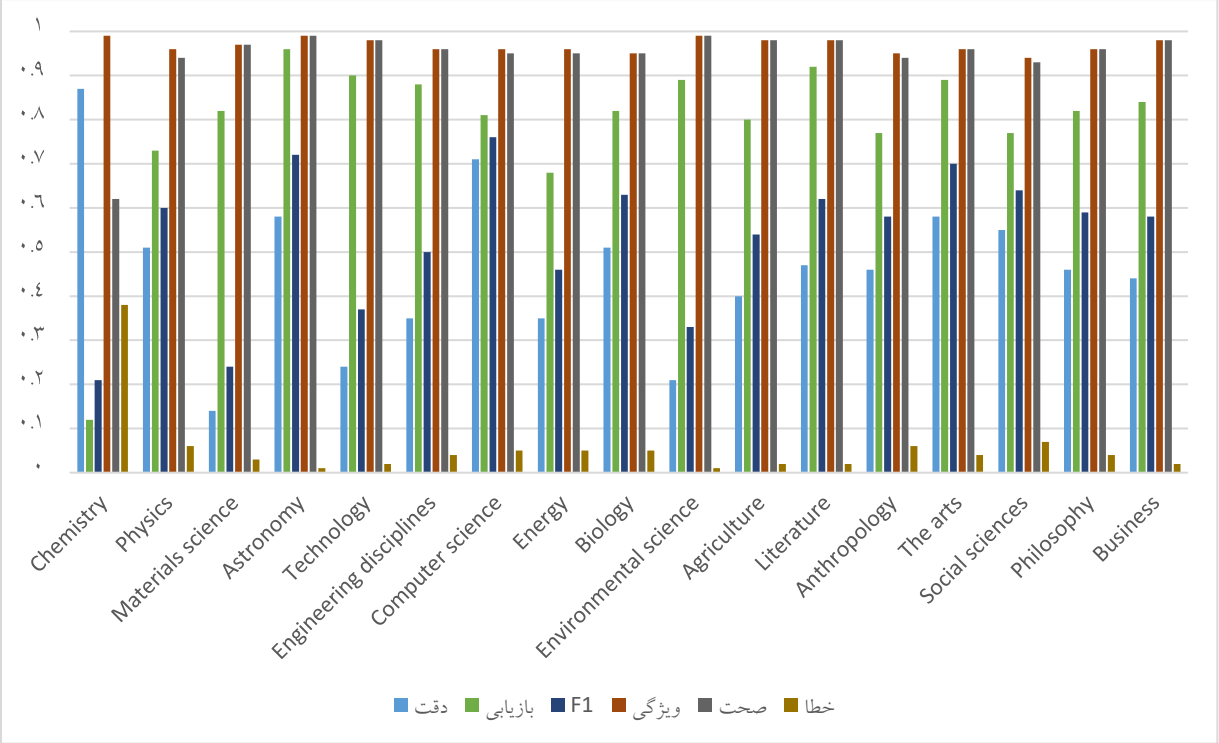
فرایند آموزش این مدل، حدود ١٤ ثانیه زمان برده است. پیش‌بینی کلاس ٤٦٩٣ رکورد مجموعه داده‌ی آزمایش، ٠.٢ ثانیه به طول انجامید؛ یعنی تنها حدود ٠.٠٤ میلی ثانیه برای هر رکورد.

جدول ٤-١٦ نشان‌دهنده‌ی عملکرد کلی مدل جنگل تصادفی در طبقه‌بندی رکوردهای مجموعه داده‌ی انگلیسی است:

(جدول ٤-١٦) عملکرد کلی مدل جنگل تصادفی در مجموعه داده‌ی انگلیسی.

+ ص	- ص	+ ک	- ک	دقت	بازیابی	F_1	ویژگی	صحت	خطا
٢٣٩٨	٧٢٧٩٣	٢٢٩٥	٢٢٩٥	٠.٥١	٠.٥١	٠.٥١	٠.٩٧	٠.٩٤	٠.٠٦

بنابراین، این مدل در طبقه‌بندی رکوردهای انگلیسی، با دقت ٥١٪ و صحت ٩٤٪ عمل کرده است. نمودار میله‌ای عملکرد این مدل به تفکیک هر کلاس، در نمودار ٤-١٦ قابل مشاهده است.



(نمودار ٤-١٦) نمودار عملکرد مجموعه داده‌ی انگلیسی مدل جنگل تصادفی، به تفکیک کلاس.

همانطور که مشاهده می‌شود، بدترین عملکرد این مدل به کلاس شیمی اختصاص دارد که امتیاز اف-یک ٠.٢١ و صحت ٠.٦٢ را در طبقه‌بندی رکوردهای آن کسب کرده است. بهترین عملکرد این مدل هم به کلاس

علوم کامپیوتر اختصاص داشته که توانسته با امتیاز اف-یک ۰.۷٦ و صحت ۰.۹۵، رکوردهای این کلاس را طبقه‌بندی کند.

٤-۳-۸- عملکرد مدل شبکه‌ی عصبی بازگشتی

عملکرد مدل شبکه‌ی عصبی بازگشتی، که از جدیدترین فناوری‌های حوزه‌ی یادگیری ماشین برای طبقه‌بندی محتوا به‌شمار می‌رود، در این بخش مورد تحلیل قرار خواهد گرفت. نخست، عملکرد این مدل در مجموعه داده‌ی فارسی و سپس در مجموعه داده‌ی انگلیسی، بررسی خواهد شد:

٤-۳-۸-۱- مجموعه داده‌ی فارسی

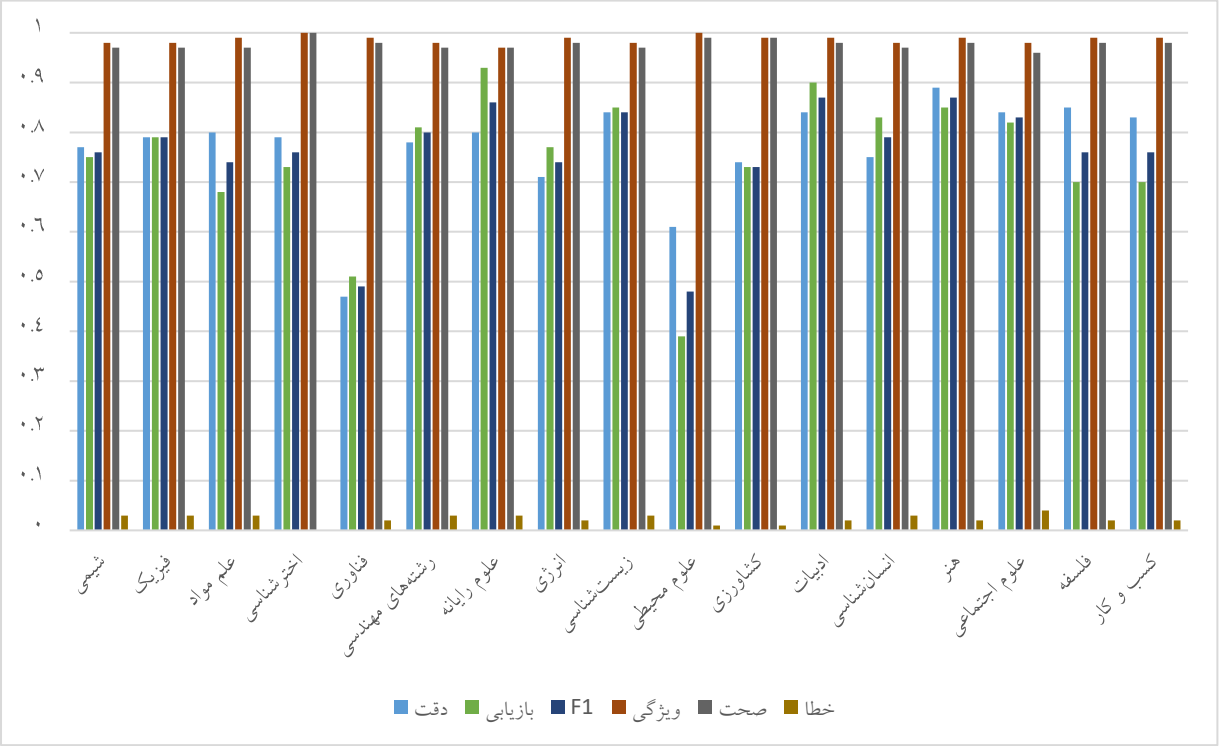
فرایند آموزش این مدل ۱۰۲۸ ثانیه (حدود ۱۷ دقیقه) زمان برده است. پیش‌بینی کلاس ۲۸٤۵ رکورد مجموعه داده‌ی آزمایش، ۱۱ ثانیه به طول انجامید؛ یعنی حدود ٤ میلی ثانیه برای هر رکورد.

در جدول ٤—۱۷، عملکرد کلی مدل شبکه‌ی عصبی بازگشتی در طبقه‌بندی رکوردهای مجموعه داده‌ی فارسی، نشان داده شده است:

(جدول ٤—۱۷) عملکرد کلی مدل شبکه‌ی عصبی بازگشتی در مجموعه داده‌ی فارسی.

+ ص	- ص	+ ک	- ک	دقت	بازیابی	F_1	ویژگی	صحت	خطا
۲۲۷٦	٤٤۹۵۱	۵٦۹	۵٦۹	۰.۸۰	۰.۸۰	۰.۸۰	۰.۹۹	۰.۹۸	۰.۰۲

همانطور که مشاهده می‌شود، دقت مدل در طبقه‌بندی رکوردهای فارسی، ٪۸۰ و صحت آن ٪۹۸ محاسبه شده است. نمودار میله‌ای عملکرد مدل به تفکیک هر کلاس، در نمودار ٤—۱۷ نمایش داده شده است.



(نمودار ٤–١٧) نمودار عملکرد مجموعه داده‌ی فارسی مدل شبکه‌ی عصبی بازگشتی، به تفکیک کلاس.

همانطور که مشاهده می‌شود، این مدل در طبقه‌بندی رکوردهای کلاس هنر، امتیاز اف–یک ۰.۸۷ و صحت ۰.۹۸ را بدست آورده که نسبت به سایر کلاس‌ها، عملکرد بهتری است. ضعیف‌ترین عملکرد این مدل در خصوص کلاس علوم محیطی بوده که امتیاز اف–یک ۰.٤٨ را بدست آورده است.

٤–۳–۸–۲– مجموعه داده‌ی انگلیسی

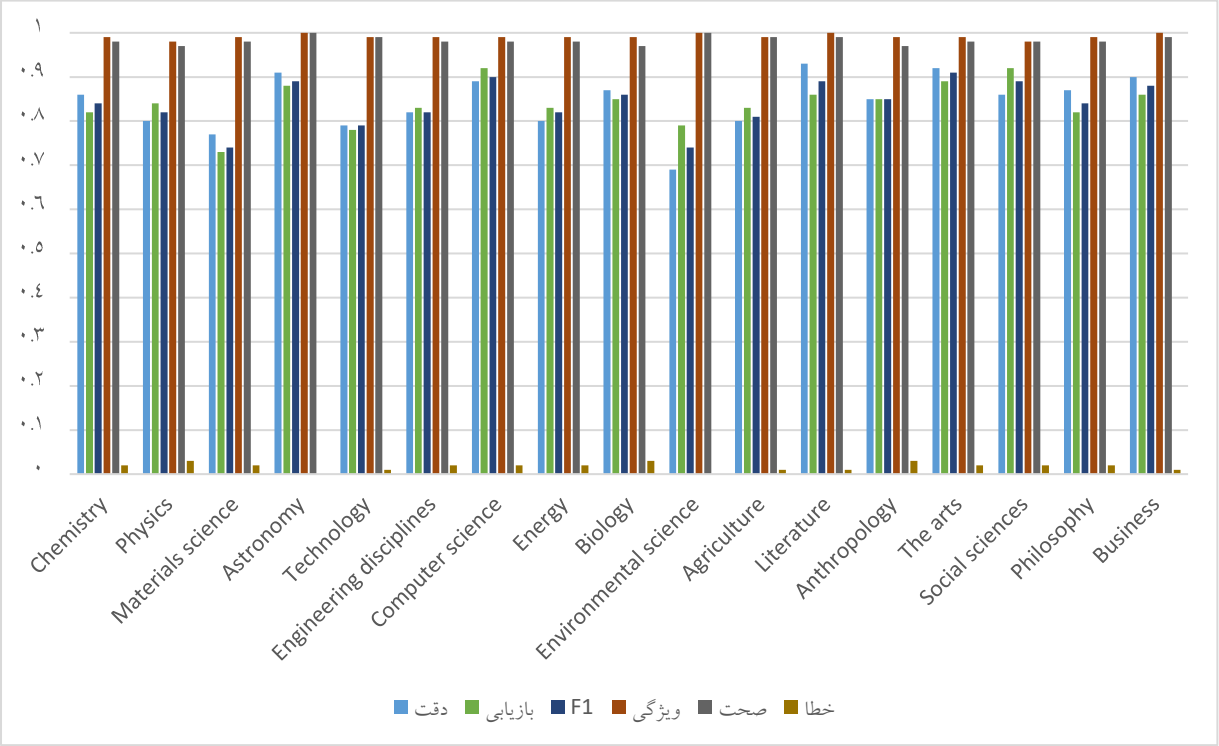
فرایند آموزش این مدل ۹۸٦ ثانیه (حدود ١٦.٥ دقیقه) زمان برده است. پیش‌بینی کلاس ٤٦۹۳ رکورد مجموعه داده‌ی آزمایش، حدود ۱۰ ثانیه به طول انجامید؛ یعنی حدود ۲ میلی ثانیه برای هر رکورد.

جدول ٤–١٨ نشان‌دهنده‌ی عملکرد کلی مدل شبکه‌ی عصبی بازگشتی در طبقه‌بندی رکوردهای مجموعه داده‌ی انگلیسی است.

(جدول ٤–١٨) عملکرد کلی مدل شبکه‌ی عصبی بازگشتی در مجموعه داده‌ی انگلیسی.

+ ص	– ص	+ ک	– ک	دقت	بازیابی	F_1	ویژگی	صحت	خطا
٤۰۱۳	۷٤٤۰۸	٦۸۰	٦۸۰	۰.۸٦	۰.۸٦	۰.۸٦	۰.۹۹	۰.۹۸	۰.۲

همانطور که مشاهده می‌شود، طبقه‌بندی رکوردهای انگلیسی توسط این مدل، با دقت ۰.۸٦ و صحت ۰.۹۸ انجام شده است. نمودار میله‌ای عملکرد مدل به تفکیک هر کلاس، در نمودار ٤–١٨ قابل مشاهده است.



(نمودار ٤–١٨) نمودار عملکرد مجموعه داده‌ی انگلیسی مدل شبکه‌ی عصبی بازگشتی، به تفکیک کلاس.

همانطور که مشاهده می‌شود، این مدل در طبقه‌بندی رکوردهای کلاس هنر، با امتیاز اف–یک ۰.۹۱ و صحت ۰.۹۸، عملکرد بهتری را نسبت به سایر کلاس‌ها داشته است. اما در خصوص کلاس علوم محیطی، با امتیاز اف–یک ۰.۷٤، عملکرد ضعیف‌تری نسبت به سایر کلاس‌ها نشان داده است.

٤–۳–۹– عملکرد مدل شبکه‌ی عصبی پیشخور (مدل اصلی)

سرانجام، در این مرحله عملکرد مدل اصلی مدنظر در طبقه‌بندی رکوردهای دو مجموعه داده‌ی تحقیق، بررسی خواهند شد. از آنجا که این مدل، مدل اصلی ارائه شده در این تحقیق است؛ نتایج مفصل‌تر از سایر مدل‌ها بیان خواهند شد. مشابه قبل، ابتدا عملکرد مدل را در طبقه‌بندی رکوردهای مجموعه داده‌ی فارسی و سپس مجموعه داده‌ی انگلیسی بررسی خواهد شد:

٤–۳–۹–۱– مجموعه داده‌ی فارسی

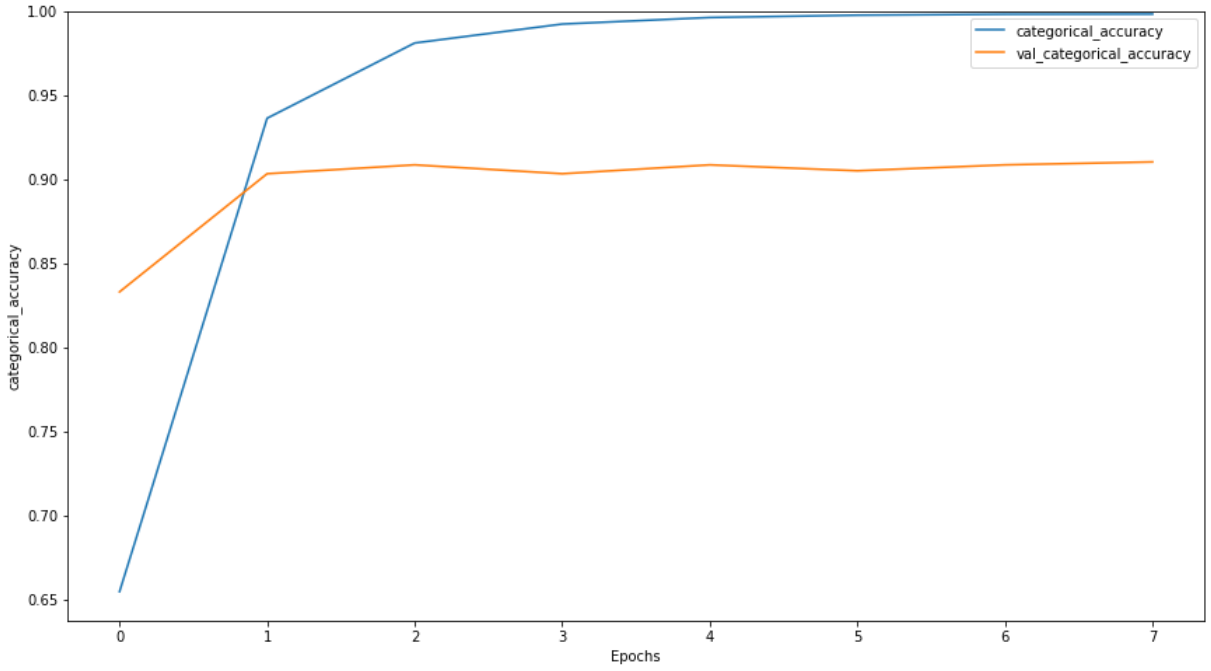
فرایند آموزش این مدل حدود ۱۸ ثانیه زمان برد که در شکل ٤–۱، مدت زمان یادگیری در هر چرخه قابل مشاهده است.

```
Epoch 1/8
109/109 [==============================] - 3s 21ms/step - loss: 1.5385 - categorical_accuracy: 0.6547 - val_loss: 0.8313 - val_categorical_accuracy: 0.8330
Epoch 2/8
109/109 [==============================] - 2s 20ms/step - loss: 0.4430 - categorical_accuracy: 0.9364 - val_loss: 0.5106 - val_categorical_accuracy: 0.9033
Epoch 3/8
109/109 [==============================] - 2s 20ms/step - loss: 0.1890 - categorical_accuracy: 0.9811 - val_loss: 0.4075 - val_categorical_accuracy: 0.9086
Epoch 4/8
109/109 [==============================] - 2s 19ms/step - loss: 0.0964 - categorical_accuracy: 0.9924 - val_loss: 0.3688 - val_categorical_accuracy: 0.9033
Epoch 5/8
109/109 [==============================] - 2s 20ms/step - loss: 0.0565 - categorical_accuracy: 0.9963 - val_loss: 0.3575 - val_categorical_accuracy: 0.9086
Epoch 6/8
109/109 [==============================] - 2s 19ms/step - loss: 0.0356 - categorical_accuracy: 0.9977 - val_loss: 0.3502 - val_categorical_accuracy: 0.9051
Epoch 7/8
109/109 [==============================] - 2s 20ms/step - loss: 0.0261 - categorical_accuracy: 0.9982 - val_loss: 0.3478 - val_categorical_accuracy: 0.9086
Epoch 8/8
109/109 [==============================] - 2s 20ms/step - loss: 0.0194 - categorical_accuracy: 0.9983 - val_loss: 0.3438 - val_categorical_accuracy: 0.9104

 total time: 17.817238330841064 seconds
```

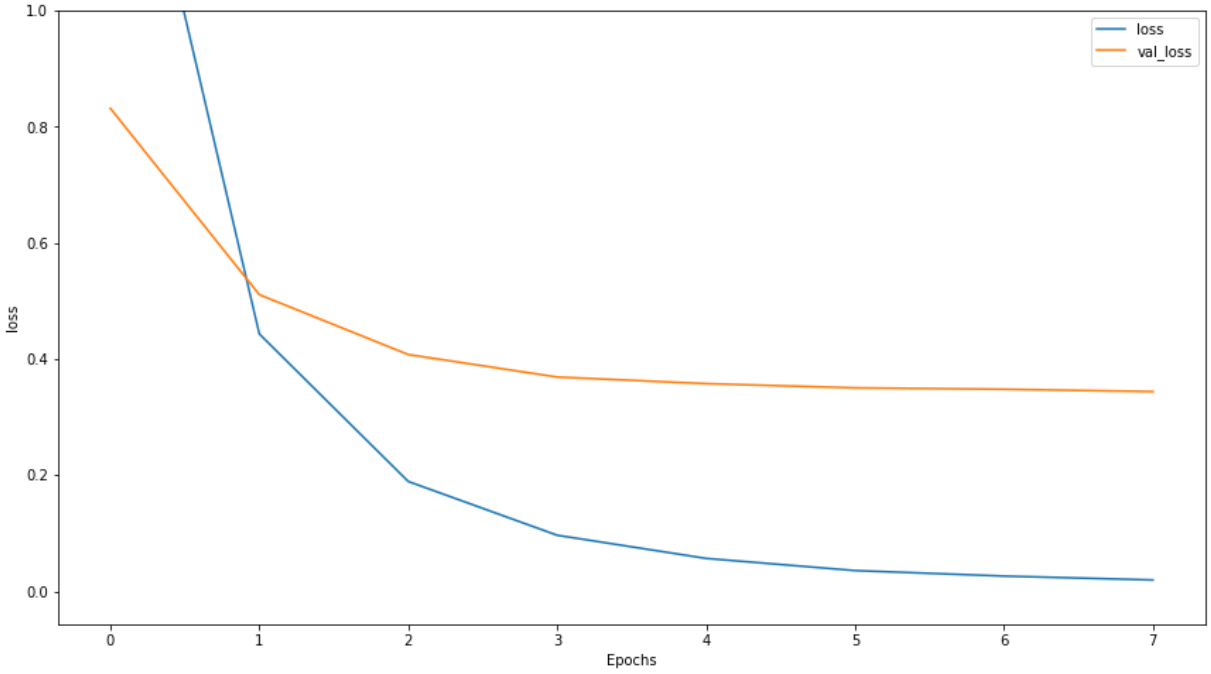
(شکل ٤–١) اطلاعات فرایند آموزش ای‌انان روی مجموعه داده‌ی فارسی در هر چرخه.

همانطور که دیده می‌شود، مدل در چرخه‌ی اول، به دقت حدود ٦٥٪ در طبقه‌بندی مجموعه داده‌ی آزمایش دست یافته و با ادامه دادن فرایند یادگیری، در چرخه‌ی هشتم، به دقت حدود ٩٩.٩٪ می‌رسد. دقت طبقه‌بندی مجموعه داده‌ی اعتبارسنجی در چرخه‌ی اول، حدود ٨٣٪ بوده که تا چرخه‌ی هشتم به حدود ٩١٪ می‌رسد. نمودار ٤–١٩ بهبود دقت طبقه‌بندی را، از چرخه‌ی ابتدایی تا انتهایی، نشان می‌دهد.



(نمودار ٤–١٩) نمودار فرایند یادگیری مدل شبکه‌ی عصبی پیشخور روی مجموعه داده‌ی فارسی.

همچنین بهبود مقدار زیان، حین فرایند یادگیری، در نمودار ٤–٢٠ آمده است:



(نمودار ٤–٢٠) نمودار مقدار زیان مدل شبکه‌ی عصبی پیشخور در مجموعه داده‌ی فارسی.

همانطور که از نمودار بالا پیداست، مقدار زیان ابتدا حدود ١.٥٤ بوده اما به حدود ٠.٠٢ تقلیل پیدا کرده است.

پیش‌بینی کلاس ٢٨٤٥ رکورد مجموعه داده‌ی آزمایش توسط این مدل، تنها ٠.٦٢٦٤ ثانیه زمان برد؛ یعنی حدود ٠.٢ میلی‌ثانیه برای هر رکورد (پاراگراف).

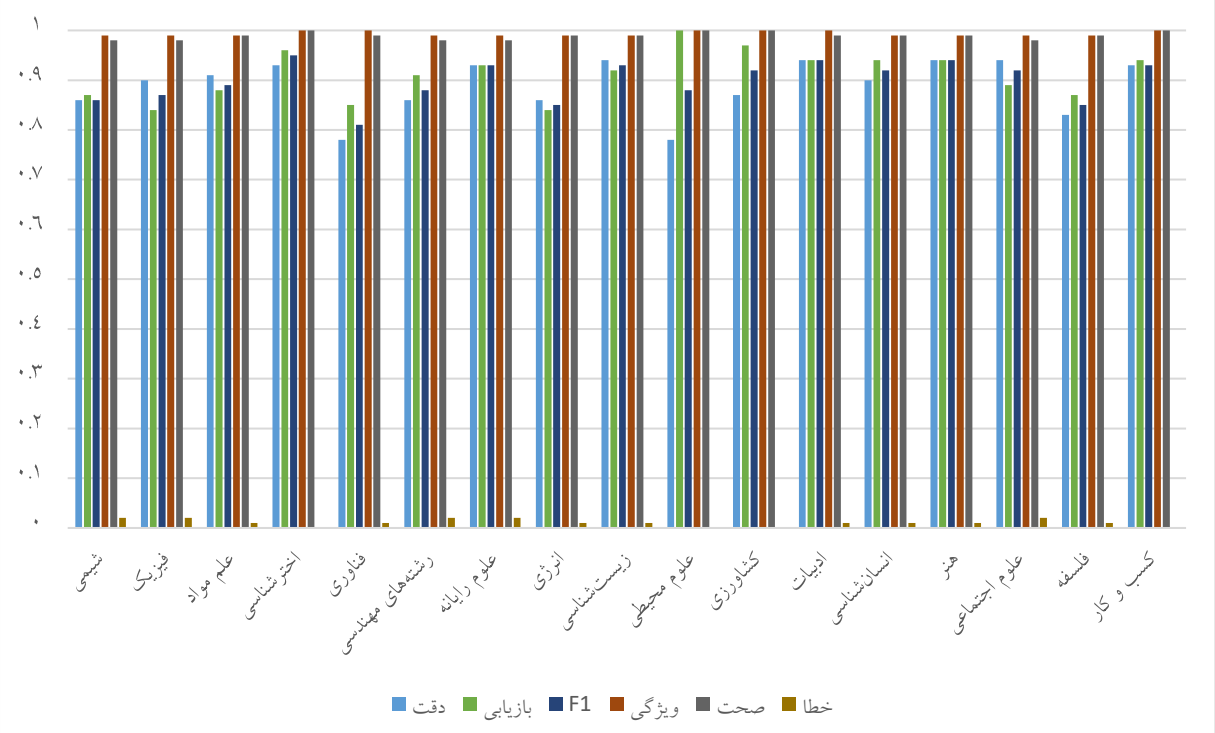
در جدول ٤–١٩، اطلاعات مربوط به عملکرد طبقه‌بندی این مدل، به تفکیک هر کلاس، آورده شده است.

(جدول ٤–١٩) عملکرد مدل شبکه‌ی عصبی پیشخور در مجموعه داده‌ی فارسی، به تفکیک کلاس.

#	عنوان	+ ص	– ص	+ غ	– غ	دقت	بازیابی	F_1	ویژگی	صحت
٠	شیمی	١٥٢	٢٦٤٥	٢٥	٢٣	٠.٨٦	٠.٨٧	٠.٨٦	٠.٩٩	٠.٩٨
١	فیزیک	٢٠١	٢٥٨٢	٢٣	٣٩	٠.٩٠	٠.٨٤	٠.٨٧	٠.٩٩	٠.٩٨
٢	علم مواد	١٤٣	٢٦٦٨	١٥	١٩	٠.٩١	٠.٨٨	٠.٨٩	٠.٩٩	٠.٩٩
٣	اخترشناسی	٢٦	٢٨١٦	٢	١	٠.٩٣	٠.٩٦	٠.٩٥	١.٠٠	١.٠٠
٤	فناوری	٤٥	٢٧٧٩	١٣	٨	٠.٧٨	٠.٨٥	٠.٨١	١.٠٠	٠.٩٩
٥	رشته‌های مهندسی	١٦٣	٢٦٣٨	٢٧	١٧	٠.٨٦	٠.٩١	٠.٨٨	٠.٩٩	٠.٩٨
٦	علوم رایانه	٣٣٦	٢٤٦١	٢٤	٢٤	٠.٩٣	٠.٩٣	٠.٩٣	٠.٩٩	٠.٩٨

۷	انرژی	۱۰۸	۲۶۹۹	۱۸	۲۰	۰.۸۶	۰.۸٤	۰.۸۵	۰.۹۹	۰.۹۹
۸	زیست‌شناسی	۲٤۷	۲۵٦۲	۱۵	۲۱	۰.۹٤	۰.۹۲	۰.۹۳	۰.۹۹	۰.۹۹
۹	علوم محیطی	۱٤	۲۸۲۷	٤	۰	۰.۷۸	۱.۰۰	۰.۸۸	۱.۰۰	۱.۰۰
۱۰	کشاورزی	٦۸	۲۷٦۵	۱۰	۲	۰.۸۷	۰.۹۷	۰.۹۲	۱.۰۰	۱.۰۰
۱۱	ادبیات	۱۷۰	۲٦۵۵	۱۰	۱۰	۰.۹٤	۰.۹٤	۰.۹٤	۱.۰۰	۰.۹۹
۱۲	انسان‌شناسی	۱۹۵	۲٦۱۷	۲۱	۱۲	۰.۹۰	۰.۹٤	۰.۹۲	۰.۹۹	۰.۹۹
۱۳	هنر	۲۳۱	۲۵۸٦	۱٤	۱٤	۰.۹٤	۰.۹٤	۰.۹٤	۰.۹۹	۰.۹۹
۱٤	علوم اجتماعی	۳۰۰	۲٤۹۱	۱۸	۳٦	۰.۹٤	۰.۸۹	۰.۹۲	۰.۹۹	۰.۹۸
۱۵	فلسفه	۱۰۳	۲۷۰٦	۲۱	۱۵	۰.۸۳	۰.۸۷	۰.۸۵	۰.۹۹	۰.۹۹
۱٦	کسب و کار	۷۷	۲۷۵۷	٦	۵	۰.۹۳	۰.۹٤	۰.۹۳	۱.۰۰	۱.۰۰
کلی (مدل)		**۲۵۷۹**	**٤۵۲۵٤**	**۲٦٦**	**۲٦٦**	**۰.۹۱**	**۰.۹۱**	**۰.۹۱**	**۰.۹۹**	**۰.۹۹**

بنابراین، مدل ارائه شده توانست، با دقت ۹۱٪ و صحت ۹۹٪، رکوردهای مجموعه داده‌ی آزمایش فارسی را طبقه‌بندی کند. نمودار میله‌ای عملکرد مدل به تفکیک هر کلاس، در نمودار ٤–۲۱ قابل مشاهده است.



(نمودار ٤–۲۱) نمودار عملکرد مجموعه داده‌ی فارسی شبکه‌ی عصبی پیشخور، به تفکیک کلاس.

نتایج نشان می‌دهند که امتیاز اف-یک مدل شبکه‌ی عصبی پیشخور، در همه‌ی کلاس‌ها بالای ۰.۸ بوده است. پایین‌ترین امتیاز اف-یک ۰.۸۱ بوده که مربوط به کلاس فناوری است. بالاترین امتیاز اف-یک اما، مربوط به کلاس اخترشناسی بوده است. طبقه‌بندی رکوردهای این کلاس، با امتیاز اف-یک ۰.۹۵ و صحت حدوداً یک (کامل) انجام شده.

«ماتریس درهم‌ریختگی»[1] این مدل، در جدول ٤–۲۰ قابل مشاهده است.

(جدول ٤–۲۰) ماتریس درهم‌ریختگی شبکه‌ی عصبی پیشخور در مجموعه داده‌ی فارسی.

	۰	۱	۲	۳	٤	٥	٦	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	۱٤	۱٥	۱٦
۰	۱٥۲	٦	٥	۰	۰	۱	۱	٥	٤	۰	۰	۰	۰	۰	۱	۲	۰
۱	٤	۲۰۱	۷	۰	۰	٤	۱	۱	۲	۰	۰	۰	۰	۲	۱	۱	۰
۲	٤	۷	۱٤۳	۰	۰	۱	۱	۱	۱	۰	۰	۰	۰	۰	۰	۰	۰
۳	۰	۰	۰	۲٦	۰	۰	۰	۰	۰	۰	۰	۱	۱	۰	۰	۰	۰
٤	۰	۰	۰	۰	٤٥	۱	٦	۱	۱	۰	۰	۰	۱	۱	۰	۰	۲
٥	۲	۱۰	۳	۰	۰	۱٦۳	٦	۲	۱	۰	۰	۰	۰	۰	۱	۲	۰
٦	۰	٦	۰	۰	۳	۲	۳۳٦	۱	۱	۰	۰	۰	۱	۰	۷	۲	۱
۷	۷	۳	۱	۰	۱	۳	۰	۱۰۸	۲	۰	۰	۰	۰	۱	۰	۰	۰
۸	۲	۲	۰	۰	۰	۲	۱	۳	۲٤۷	۰	۱	۰	۰	۰	۳	۱	۰
۹	۰	۰	۰	۰	۰	۰	۰	۰	۳	۱٤	۰	۰	۰	۰	۱	۰	۰
۱۰	۱	۰	۲	۰	۰	۰	۰	٥	۱	۰	٦۸	۰	۰	۱	۰	۰	۰
۱۱	۰	۰	۰	۰	۰	۰	۱	۰	۰	۰	۰	۱۷۰	۱	٤	۳	۱	۰
۱۲	۰	۰	۰	۰	۱	۰	۱	۰	۰	۰	۰	۳	۱۹٥	٤	۹	۳	۰
۱۳	۱	۱	۰	۱	۱	۱	۰	۰	۰	۰	۱	۳	۱	۲۳۱	۲	۰	۲
۱٤	۰	۰	۰	۰	۱	۰	۳	۱	٥	۰	۰	۰	٤	۱	۳۰۰	۳	۰
۱٥	۲	۳	۰	۰	۰	۱	۱	۰	۰	۰	۰	۳	۳	۰	۸	۱۰۳	۰
۱٦	۰	۱	۱	۰	۱	۱	۲	۰	۰	۰	۰	۰	۰	۰	۰	۰	۷۷

[1] در حوزه‌ی یادگیری ماشین، ماتریس سردرگمی (Confusion Matrix)، یک جدول خاص است که عملکرد یک مدل (معمولاً یک مدل مبتنی بر یادگیری تحت نظارت) را در طبقه‌بندی رکوردها نشان می‌دهد. هر ردیف از ماتریس، به مقادیر واقعی کلاس اشاره داشته؛ در حالی که هر ستون به مقادیر پیش‌بینی‌شده دلالت دارد. هرچند حالت عکس نیز، گاهاً استفاده می‌شود (لاروس و لاروس ۲۰۱۹، ۹۸).

همانطور که در ماتریس درهم‌ریختگی دیده می‌شود، بیشترین تعداد خطا بین دو کلاس رشته‌های مهندسی و فیزیک رخ داده که با توجه به شباهت این دو کلاس، منطقی به نظر می‌رسد. تعداد کل رکوردهایی کلاس رشته‌های مهندسی که به‌اشتباه به کلاس فیزیک نسبت داده شده‌اند، ۱۰ عدد می‌باشد که حدود ٪۶ کل داده‌های آن کلاس را تشکیل می‌دهند.

٤-۳-۹-۲- مجموعه داده‌ی انگلیسی

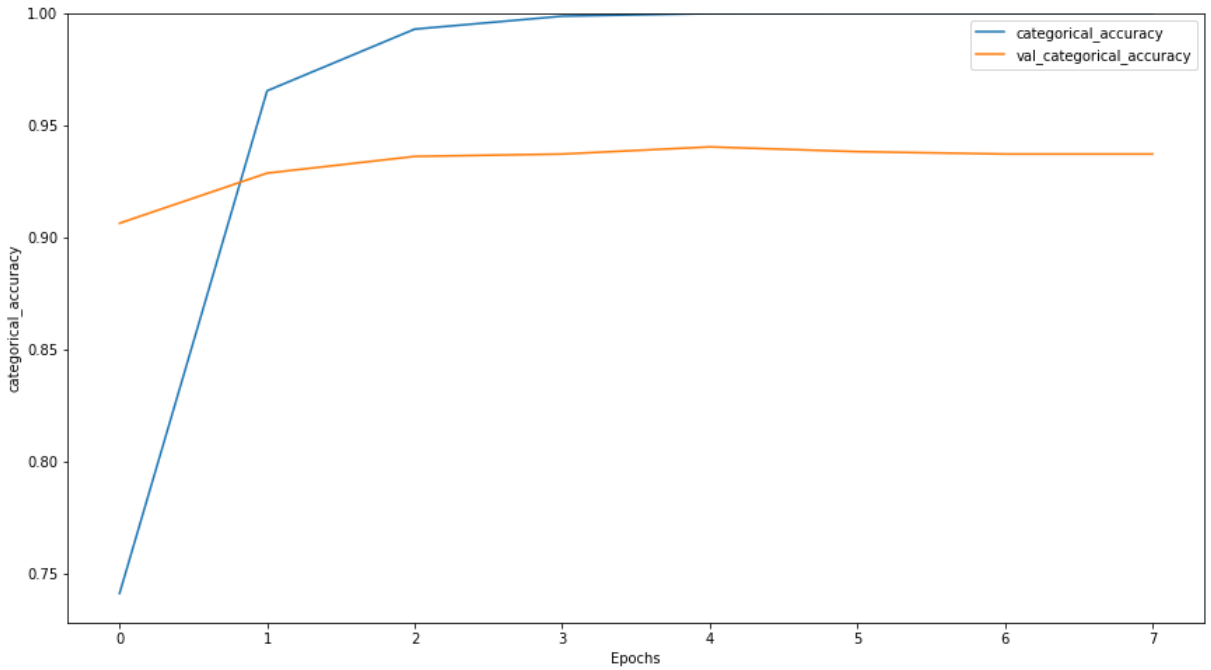
فرایند آموزش این مدل روی مجموعه داده‌های آزمایش انگلیسی، حدود ۳۳.٤ ثانیه زمان برده است که در شکل ٤-۲، مدت زمان یادگیری در هر چرخه قابل مشاهده است.

```
Epoch 1/8
179/179 [==============================] - 7s 35ms/step - loss: 1.2130 - categorical_accuracy: 0.7410 - val_loss: 0.4831 - val_categorical_accuracy: 0.9063
Epoch 2/8
179/179 [==============================] - 4s 20ms/step - loss: 0.2330 - categorical_accuracy: 0.9654 - val_loss: 0.3228 - val_categorical_accuracy: 0.9286
Epoch 3/8
179/179 [==============================] - 4s 22ms/step - loss: 0.0816 - categorical_accuracy: 0.9929 - val_loss: 0.2878 - val_categorical_accuracy: 0.9361
Epoch 4/8
179/179 [==============================] - 4s 21ms/step - loss: 0.0363 - categorical_accuracy: 0.9987 - val_loss: 0.2722 - val_categorical_accuracy: 0.9372
Epoch 5/8
179/179 [==============================] - 4s 21ms/step - loss: 0.0193 - categorical_accuracy: 0.9999 - val_loss: 0.2666 - val_categorical_accuracy: 0.9404
Epoch 6/8
179/179 [==============================] - 4s 22ms/step - loss: 0.0118 - categorical_accuracy: 1.0000 - val_loss: 0.2638 - val_categorical_accuracy: 0.9382
Epoch 7/8
179/179 [==============================] - 4s 21ms/step - loss: 0.0080 - categorical_accuracy: 1.0000 - val_loss: 0.2645 - val_categorical_accuracy: 0.9372
Epoch 8/8
179/179 [==============================] - 4s 21ms/step - loss: 0.0058 - categorical_accuracy: 1.0000 - val_loss: 0.2653 - val_categorical_accuracy: 0.9372

 total time: 33.39027786254883 seconds
```

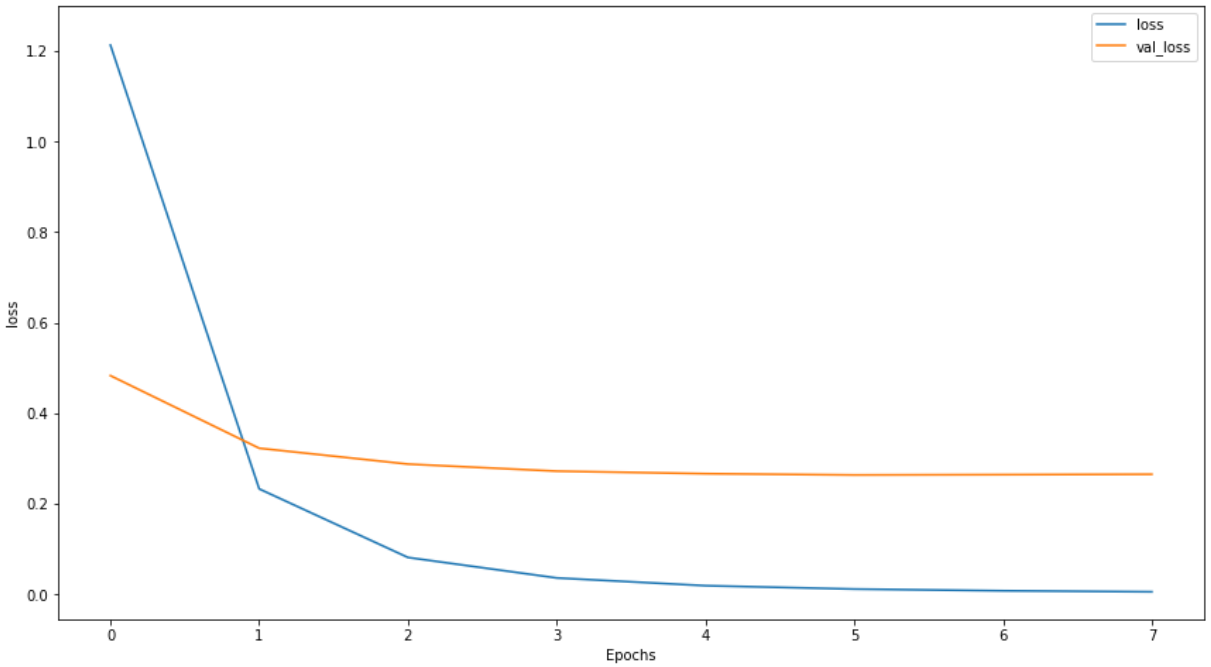
(شکل ٤-۲) اطلاعات فرایند آموزش ای‌ان‌ان روی مجموعه داده‌ی انگلیسی در هر چرخه.

همانطور که دیده می‌شود، مدل در چرخه‌ی اول، به دقت حدود ٪۷٤ در طبقه‌بندی مجموعه داده‌ی آزمایش دست یافته و با ادامه دادن فرایند یادگیری، در چرخه‌ی هشتم، به دقت ٪۱۰۰ می‌رسد. دقت طبقه‌بندی مجموعه داده‌ی اعتبارسنجی، در چرخه‌ی اول حدود ٪۹۱ بوده که تا چرخه‌ی هشتم به دقت حدود ٪۹٤ می‌رسد. نمودار ٤-۲۲ بهبود دقت طبقه‌بندی را، از چرخه‌ی ابتدایی تا نهایی، نشان می‌دهد.



(نمودار ٤–٢٢) نمودار فرایند یادگیری مدل شبکه‌ی عصبی پیشخور روی مجموعه داده‌ی انگلیسی.

همچنین بهبود مقدار زیان، حین فرایند یادگیری، در نمودار ٤–٢٣ آمده است.



(نمودار ٤–٢٣) نمودار مقدار زیان مدل شبکه‌ی عصبی پیشخور روی مجموعه داده‌ی انگلیسی.

همانطور که از نمودار بالا پیداست، مقدار زیان ابتدا حدود ۱.۲ بوده اما به حدود ۰.۰۰٦ تقلیل پیدا کرده است.

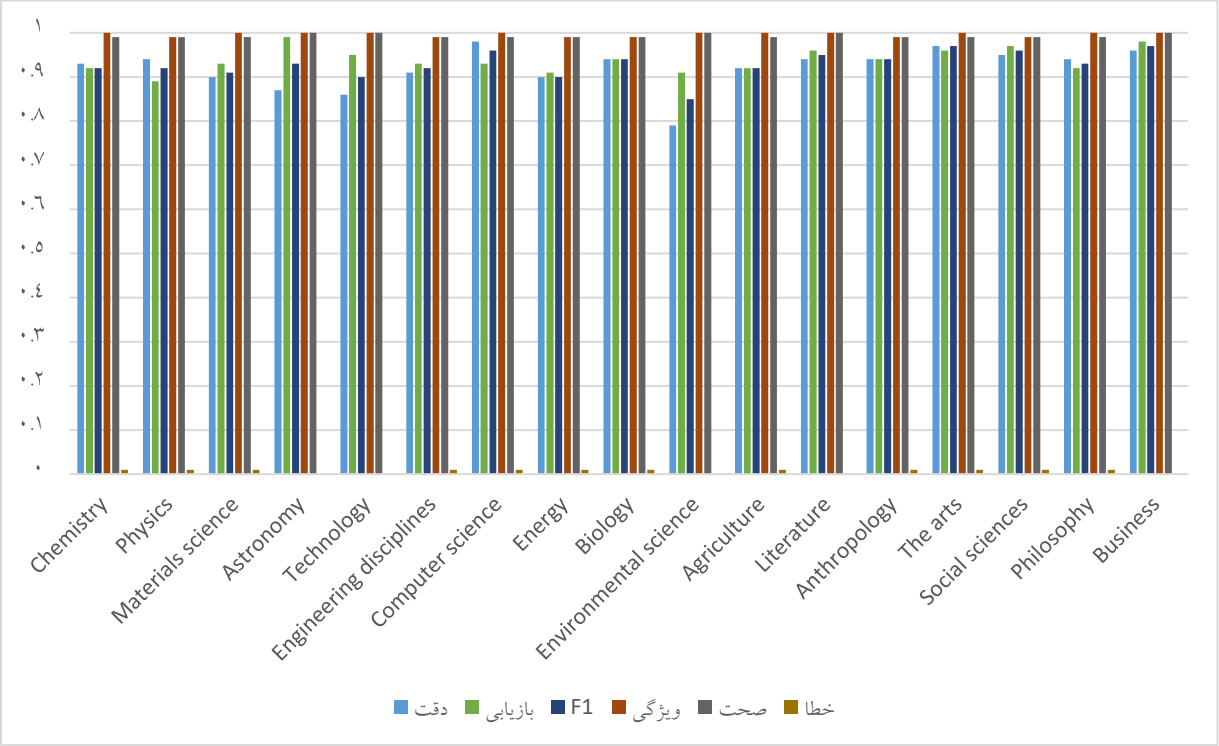
پیش‌بینی کلاس ٤٦٩٣ رکورد مجموعه داده‌ی آزمایش توسط این مدل، تنها ۱.۳٤ ثانیه زمان برد؛ یعنی حدود ۰.۳ میلی‌ثانیه برای هر رکورد (پاراگراف).

در جدول ٤–۲۱، اطلاعات مربوط به عملکرد طبقه‌بندی این مدل، به تفکیک هر کلاس، آورده شده است.

(جدول ٤–۲۱) عملکرد مدل شبکه‌ی عصبی پیشخور در مجموعه داده‌ی انگلیسی، به تفکیک کلاس.

#	عنوان	+ ص	− ص	+ غ	− غ	دقت	بازیابی	F_1	ویژگی	صحت
۰	Chemistry	۲۵۲	٤٤۰۰	۱۹	۲۲	۰.۹۳	۰.۹۲	۰.۹۲	۱	۰.۹۹
۱	Physics	۳٦۹	٤۲٥٦	۲٤	٤٤	۰.۹٤	۰.۸۹	۰.۹۲	۰.۹۹	۰.۹۹
۲	Materials science	۱٤۹	٤٥۱٦	۱۷	۱۱	۰.۹	۰.۹۳	۰.۹۱	۱	۰.۹۹
۳	Astronomy	٦۹	٤٦۱۳	۱۰	۱	۰.۸۷	۰.۹۹	۰.۹۳	۱	۱
٤	Technology	۹٥	٤٥۷۸	۱٥	٥	۰.۸٦	۰.۹٥	۰.۹	۱	۱
٥	Engineering disciplines	۲۳۳	٤٤۱۸	۲٤	۱۸	۰.۹۱	۰.۹۳	۰.۹۲	۰.۹۹	۰.۹۹
٦	Computer science	٥۳٥	٤۱۰۸	۱۲	۳۸	۰.۹۸	۰.۹۳	۰.۹٦	۱	۰.۹۹
۷	Energy	۲٤۳	٤۳۹۷	۲۸	۲٥	۰.۹	۰.۹۱	۰.۹	۰.۹۹	۰.۹۹
۸	Biology	۳۹۲	٤۲٥۱	۲٦	۲٤	۰.۹٤	۰.۹٤	۰.۹٤	۰.۹۹	۰.۹۹
۹	Environmental science	۳۱	٤٦٥۱	۸	۳	۰.۷۹	۰.۹۱	۰.۸٥	۱	۱
۱۰	Agriculture	۱٤۸	٤٥۱۹	۱۳	۱۳	۰.۹۲	۰.۹۲	۰.۹۲	۱	۰.۹۹
۱۱	Literature	۱٥٤	٤٥۲۲	۱۰	۷	۰.۹٤	۰.۹٦	۰.۹٥	۱	۱
۱۲	Anthropology	۳۸٥	٤۲٥۷	۲٦	۲٥	۰.۹٤	۰.۹٤	۰.۹٤	۰.۹۹	۰.۹۹
۱۳	The arts	۳۹٤	٤۲۷۱	۱۲	۱٦	۰.۹۷	۰.۹٦	۰.۹۷	۱	۰.۹۹
۱٤	Social sciences	٥۰۱	٤۱٤۸	۲٦	۱۸	۰.۹٥	۰.۹۷	۰.۹٦	۰.۹۹	۰.۹۹
۱٥	Philosophy	۲۸۰	٤۳۷۱	۱۹	۲۳	۰.۹٤	۰.۹۲	۰.۹۳	۱	۰.۹۹
۱٦	Business	۱٦۷	٤٥۱٦	۷	۳	۰.۹٦	۰.۹۸	۰.۹۷	۱	۱
	کلی (مدل)	٤۳۹۷	۷٤۷۹۲	۲۹٦	۲۹٦	۰.۹٤	۰.۹٤	۰.۹٤	۱	۰.۹۹

بنابراین، مدل ارائه شده توانست، با دقت ٪۹٤ و صحت ٪۹۹، رکوردهای مجموعه داده‌ی آزمایش انگلیسی را طبقه‌بندی کند. نمودار میله‌ای عملکرد مدل به تفکیک هر کلاس، در نمودار ٤–۲٤ قابل مشاهده است.



(نمودار ٤–٢٤) نمودار عملکرد مجموعه داده‌ی انگلیسی شبکه‌ی عصبی پیشخور، به تفکیک کلاس.

نتایج نشان می‌دهند که پایین‌ترین امتیاز اف–یک این مدل، به طبقه‌بندی رکورد کلاس علوم محیطی اختصاص دارد. رکوردهای این کلاس، با امتیاز اف–یک ۸۵ .۰، طبقه‌بندی شده‌اند. به‌جز این کلاس، امتیاز اف–یک طبقه‌بندی این مدل در سایر کلاس‌ها بالای ۰.۹ بوده است. بهترین عملکرد مدل، مربوط به طبقه‌بندی کلاس کسب و کار بوده که توانسته امتیاز اف–یک ۰.۹۷ و صحت حدوداً یک (کامل) را بدست آورد.

«ماتریس درهم‌ریختگی» این مدل در جدول ٤–٢٢ دیده می‌شود.

(جدول ٤–٢٢) ماتریس درهم‌ریختگی شبکه‌ی عصبی پیشخور در رکوردهای انگلیسی.

	۰	۱	۲	۳	٤	٥	٦	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	۱٤	۱٥	۱٦
۰	۲٥۲	۷	۱	۰	۰	۲	۰	٥	۲	۰	۱	۰	۰	۱	۰	۰	۰
۱	٥	۳٦۹	٤	۰	۰	۲	٤	۲	۲	۰	۰	۰	۰	۱	۰	٤	۰
۲	۳	۸	۱٤۹	۰	۰	۱	۱	۲	۰	۰	۰	۰	۰	۱	۰	۱	۰
۳	۰	٤	۰	٦۹	۰	۰	۲	۰	۰	۰	۰	۲	۱	۰	۰	۱	۰
٤	۰	۱	۱	۰	۹٥	۲	۳	٥	۰	۰	۰	۱	۰	۱	۱	۰	۰
٥	۰	٥	۲	۰	۱	۲۳۳	٥	٥	۱	۰	۱	۰	۲	۲	۰	۰	۰
٦	۰	۰	۰	۰	۰	٦	٥۳٥	۱	۰	۰	۰	۰	۱	۱	۱	۲	۰
۷	٤	۷	۲	۰	۳	۲	۲	۲٤۳	۱	۰	۳	۰	۲	۰	۱	۰	۱
۸	۲	۳	۱	۰	۰	۲	٥	۲	۳۹۲	۲	٦	۰	۳	۰	۰	۰	۰
۹	۰	۱	۰	۰	۰	۰	۰	۰	٤	۳۱	۱	۰	۰	۰	۲	۰	۰
۱۰	۲	۰	۰	۰	۰	۰	۲	۲	٤	۰	۱٤۸	۰	۱	۱	۱	۰	۰
۱۱	۰	۰	۰	۰	۰	۰	۲	۰	۰	۰	۰	۱٥٤	۳	۳	۱	۱	۰
۱۲	۰	۳	۰	۰	۰	۰	۲	۰	٥	۱	۱	۲	۳۸٥	۱	۷	۳	۱
۱۳	٤	۲	۰	۰	۰	۱	۱	۰	۰	۰	۰	۰	۳	۳۹٤	۰	۰	۱
۱٤	۱	۰	۰	۰	۰	۰	۳	۰	٤	۰	۰	۰	٥	۲	٥۰۱	۱۱	۰
۱٥	۱	۳	۰	۰	۰	۰	۲	۱	۱	۰	۰	۲	٤	۱	٤	۲۸۰	۰
۱٦	۰	۰	۰	۱	۱	۰	٤	۰	۰	۰	۰	۰	۰	۱	۰	۰	۱٦۷

همانطور که در ماتریس درهم‌ریختگی دیده می‌شود، بیشترین تعداد خطا بین دو کلاس علوم اجتماعی و فلسفه رخ داده است. تعداد کل رکوردهایی کلاس علوم اجتماعی که به‌اشتباه به کلاس فلسفه نسبت داده شده‌اند، ۱۱ عدد می‌باشد که حدود ٪۲ کل داده‌های آن کلاس را تشکیل می‌دهند.

فصل پنجم: نتیجه‌گیری و پیشنهادات

- مقدمه
- بحث در نتایج
- نتیجه‌گیری
- محدودیت‌های تحقیق
- پیشنهادات تحقیق

۵-۱- مقدمه

در این فصل، در مورد نتایج حاصله از فصل قبل و عملکرد مدل‌های پیاده‌سازی شده، بحث می‌شود. مجموعاً نُه مدل در این تحقیق پیاده‌سازی و عملکرد آن‌ها را در طبقه‌بندی دو دسته مجموعه داده‌ی فارسی و انگلیسی بررسی شده است. اکنون قصد بر این است که در خصوص اقدامات انجام شده و نتایج بدست آمده، نتیجه‌گیری انجام شود.

بنابراین، ابتدا یافته‌ها را مورد کنکاش قرار داده و درباره‌ی آن‌ها بحث می‌کنیم؛ سپس نتایج گرفته شده را اعلام و در نهایت محدودیت‌های تحقیق و پیشنهادات محقق بیان خواهند شد.

در قسمت بحث در نتایج، فرضیات و اهداف مشخص شده‌ی تحقیق، با توجه به نتایج بدست آمده، بررسی خواهند شد. در این قسمت تأیید یا رد شدن فرضیات و موفقیت یا عدم موفقیت دستیابی به اهداف تعیین شده، مشخص خواهد شد.

آخرین قسمت این فصل به پیشنهادات اختصاص یافته که در دو دسته‌ی پیشنهادات کاربردی و پیشنهادات پژوهشی، بیان شده‌اند.

۵-۲- بحث در نتایج

در این بخش، ابتدا در خصوص معیارهای اندازه‌گیری بحث می‌شود. سپس نتایج عملکرد مدل‌ها مورد بررسی نهایی قرار گرفته و آن‌ها با یکدیگر مقایسه خواهند شد. در انتها هم، رد یا تأیید شدن فرضیه‌های تحقیق بررسی می‌گردند.

۵-۲-۱- بحث در معیارهای اندازه‌گیری

با توجه به نتایج حاصله، برای هر مدل بررسی شده، مقادیر معیارهای دقت، بازیابی و امتیاز اف-یک کلی آن مدل، به عددی یکسان اشاره دارند. علت این امر، انتساب دادن تنها یک کلاس به رکوردهای مورد بررسی است. این اقدام با در نظر گرفتن اینکه هر رکورد تنها می‌تواند منتسب به یک کلاس باشد، انجام شده است. هرچند در بررسی این معیارها برای هر کلاس مجزا، مقادیر آن‌ها متفاوت می‌باشد.

همچنین، بالا بودن مقدار صحت در مدل‌های بررسی شده، به دلیل اختصاص داشتن تعداد زیادی از پیش‌بینی‌ها به گروه منفی صحیح است. برای هر رکورد، ۱۷ پیش‌بینی انجام می‌شود که ۱۶ عدد از آن‌ها عدم اختصاص داشتن (منفی) و تنها یک پیش‌بینی به اختصاص داشتن (مثبت)، اشاره دارد. درنتیجه، برای مدل‌های پیاده‌سازی شده در این تحقیق، با عنایت به نوع مسئله، معیار دقت، قابل اتکاترین معیار اندازه‌گیری عملکرد مدل‌ها است.

۵-۲-۲- بحث در نتایج عملکرد مدل‌ها

در این تحقیق نُه مدل رایج در زمینه‌ی یادگیری ماشین و یادگیری عمیق پیاده‌سازی شده و نتایج عملکرد آن‌ها بررسی شد. در ادامه، در مورد نتایج عملکرد هر مدل بحث خواهد شد:

آ- کی-نزدیک‌ترین همسایه

اولین مدل پیاده‌سازی شده، مدل کی-نزدیک‌ترین همسایه یا کی‌ان‌ان بود. این مدل، بسیار ساده و سبک است. فرایند آموزش این مدل روی مجموعه داده‌ی آزمایش در هر دو مجموعه داده‌ی فارسی و انگلیسی، کمتر از ۰.۰۵ ثانیه طول کشید. میانگین سرعت این مدل در پیش‌بینی کلاس هر رکود فارسی ۷ میلی‌ثانیه و برای هر رکورد انگلیسی ۱۲ میلی‌ثانیه بوده است. این نشان از سادگی و سبکی مدل کی‌ان‌ان دارد.

علی‌رغم مناسب بودن سرعت آموزش و پیش‌بینی مدل کی‌ان‌ان، دقت طبقه‌بندی آن در بسیاری از موارد قابل قبول نیست. دقت طبقه‌بندی این مدل تنها ٪۵۰ در مجموعه داده‌ی آزمایش فارسی و ٪۵۸ برای مجموعه داده‌ی انگلیسی ارزیابی شد. درنتیجه استفاده از این مدل در موقعیت‌های پیچیده که دقت مدل اهمیت بالایی دارد، مناسب نمی‌باشد.

ب- مدل بیز ساده

مدل بیز ساده نیز مدلی ساده و سبک به‌حساب می‌آید. فرایند یادگیری این مدل در مجموعه داده‌ی فارسی ۱.۷ ثانیه و در مجموعه داده‌ی انگلیسی حدود ۳ ثانیه، طول کشید. میانگین پیش‌بینی هر رکورد در مجموعه داده‌ی فارسی و انگلیسی حدود یک میلی ثانیه به طول انجامید. بنابراین، سرعت پیش‌بینی این مدل از مدل کی‌ان‌ان نیز پایین‌تر می‌باشد و در مجموع مدل بیز ساده بسیار سبک و سریع، عمل کرده است.

با این وجود، عملکرد این مدل چندان مناسب نبوده است. دقت پیش‌بینی کلی این مدل در هر دو مجموعه داده‌ی فارسی و انگلیسی، ٪۶۶ ارزیابی شده است. بین نتایج مربوط به مجموعه داده‌ی فارسی و انگلیسی هیچ مدل مورد بررسی دیگری در این تحقیق، تا این اندازه شباهت وجود نداشته است. در مجموع، مدل بیز ساده، مدلی سریع اما با دقتی نه‌چندان خوب به‌حساب می‌آید.

ج- مدل رگرسیون لجستیک

مدل رگرسیون لجستیک مدلی بسیار قدرتمند می‌باشد. این مدل را می‌توان یک شبکه‌ی عصبی تک لایه دانست. فرایند یادگیری این مدل در مجموعه داده‌ی فارسی حدود ۲ دقیقه و در مجموعه داده‌ی انگلیسی حدود ۲.۵ دقیقه زمان برد. میانگین زمان صرف شده‌ی این مدل برای پیش‌بینی کلاس رکوردهای فارسی، تنها حدود ۰.۲ ثانیه و برای رکوردهای انگلیسی، حدود ۰.۱ ثانیه محاسبه شده است. درنتیجه سرعت این مدل در طبقه‌بندی رکوردها، بسیار بالا بوده است.

از طرف دیگر، دقت طبقه‌بندی این مدل حدود ٪۸۸ برای مجموعه داده‌ی فارسی و حدود ٪۹۱ برای مجموعه داده‌ی انگلیسی، ارزیابی شده است. صحت عملکرد در هر دو مجموعه داده، حدود ٪۹۹ بوده است. بنابراین، مدل رگرسیون، سریع و دقیق عمل طبقه‌بندی را انجام داده است.

د- مدل ماشین بردار پشتیبان

مدل ماشین بردار پشتیبان یا اس‌وی‌ام، مدلی قدرتمند است که در این تحقیق به دو شکل پیاده‌سازی شد: مدل اول، خطی (که اس‌وی‌ام خوانده می‌شود)؛ مدل دوم، هسته‌ای (که اس‌وی‌ام هسته‌ای خوانده می‌شود).

فرایند یادگیری اولین مدل اس‌وی‌ام پیاده‌سازی شده، حدود ۱۵ دقیقه روی مجموعه داده‌ی فارسی و ۳۲ دقیقه روی مجموعه داده‌ی انگلیسی طول کشید که با توجه به بیشتر بودن رکوردهای مجموعه داده‌ی انگلیسی، این امر طبیعی است. میانگین زمان طبقه‌بندی هر رکورد توسط این مدل، در مجموعه داده‌ی آزمایش فارسی ۵۷ میلی‌ثانیه و در مجموعه داده‌ی انگلیسی حدود ۱۱۸ میلی‌ثانیه اندازه‌گیری شد.

دقت طبقه‌بندی این مدل در مجموعه داده‌ی فارسی ٪۸۵ و در مجموعه داده‌ی انگلیسی ٪۸۸ محاسبه شد. بنابراین، دقت و سرعت مدل اس‌وی‌ام، هر چند به خوبی مدل رگرسیون لجستیک نبوده، اما بسیار مناسب و قابل قبول می‌باشد. این مدل، حتی در خصوص کلاس‌هایی با داده‌های کم (مثل اخترشناسی و علوم محیطی)، عملکرد مناسبی داشته است.

ه- مدل اس‌وی‌ام هسته‌ای

مدل اس‌وی‌ام هسته‌ای، به‌عنوان دومین مدل اس‌وی‌ام بررسی شده در این تحقیق، مدلی قدرتمند و رایج به‌حساب می‌آید. فرایند آموزش این مدل حدود ۲۲.۵ دقیقه روی مجموعه داده‌ی فارسی و حدود ۱ ساعت و ۱۷ دقیقه روی مجموعه داده‌ی انگلیسی به طول انجامید. بنابراین، مدت زمان لازم برای آموزش مدل هسته‌ای، بیشتر از مدل خطی است که نشان از پیچیدگی بیشتر آن دارد. میانگین زمان صرف شده توسط مدل اس‌وی‌ام هسته‌ای برای طبقه‌بندی هر رکورد فارسی ۱۲۶ میلی‌ثانیه و برای هر رکورد انگلیسی ۲۱۹ میلی‌ثانیه بوده است.

دقت این مدل در طبقه‌بندی رکوردهای فارسی ٪۷۹ و در طبقه‌بندی رکوردهای انگلیسی ٪۸٦ ارزیابی شده است. برخلاف انتظار، دقت مدل اس‌وی‌ام هسته‌ای پایین‌تر از دقت مدل اس‌وی‌ام خطی است. با این وجود، عملکرد کلی مدل اس‌وی‌ام هسته‌ای نیز مناسب است.

و- مدل درخت تصمیم

فرایند یادگیری مدل درخت تصمیم، ۷۵ ثانیه برای مجموعه داده‌ی فارسی و ۲۹۷ ثانیه (حدود ۵ دقیقه) برای مجموعه داده‌ی انگلیسی انجام شده است. میانگین زمان صرف شده توسط این مدل برای طبقه‌بندی هر رکورد فارسی ۰.۱ میلی‌ثانیه و برای هر رکورد انگلیسی ۰.۲ میلی‌ثانیه بوده است. درنتیجه این مدل در طبقه‌بندی رکوردها، بسیار سریع است.

با این وجود، دقت این مدل در طبقه‌بندی رکوردها چندان مناسب نیست. دقت طبقه‌بندی این مدل در مجموعه داده‌ی فارسی ٪٥٦ و در مجموعه داده‌ی انگلیسی ٪٦۲ محاسبه شده است. بنابراین، علی‌رغم سرعت بالای این مدل در پیش‌بینی کلاس رکوردها، اما دقت طبقه‌بندی در بسیاری از موقعیت‌ها، قابل قبول نیست.

ز- مدل جنگل تصادفی

آخرین مدل سنتی بررسی شده، مدل جنگل تصادفی است. این مدل در مدت زمان ۵.۵ ثانیه روی مجموعه داده‌ی آموزش فارسی و در مدت ١٤ ثانیه روی مجموعه داده‌ی آموزش انگلیسی، عمل یادگیری را انجام داده است. میانگین زمان صرف شده توسط این مدل در طبقه‌بندی هر رکورد فارسی ۰.۰۳ میلی‌ثانیه و برای هر رکورد انگلیسی ۰.۰٤ میلی‌ثانیه محاسبه شده است.

دقت طبقه‌بندی این مدل برای مجموعه داده‌ی فارسی ٪٤۱ و برای مجموعه داده‌ی انگلیسی ٪۵۱ ارزیابی شده است. به این ترتیب، مدل جنگل تصادفی سریع‌ترین مدل پیاده‌سازی شده در طبقه‌بندی رکوردها است؛ اما دقت آن در مقایسه با سایر مدل‌ها، پایین‌تر می‌باشد. این مدل، در مجموعه داده‌ی فارسی، بیش از نصف رکوردها را به کلاس اشتباه نسبت داده است؛ لذا دقت طبقه‌بندی مدل جنگل تصادفی، اصلاً مناسب نیست.

ح- مدل شبکه‌ی عصبی بازگشتی

همانطور که اشاره شد، شبکه‌ی عصبی بازگشتی، یک مدل مبتنی بر شبکه‌های عصبی است که بر خلاف شبکه‌ی عصبی پیش‌خور، دنباله‌ی محتویات موجود در رکورد را در نظر می‌گیرد. درنتیجه این مدل نسبت به شبکه‌ی عصبی پیشخور، کمی پیچیده‌تر می‌باشد.

فرایند آموزش این مدل، برای مجموعه داده‌ی فارسی، حدود ۱۷ دقیقه و برای مجموعه داده‌ی انگلیسی، به رغم بیشتر بودن تعداد رکوردهای آزمایش آن، حدود ١٦.٥ دقیقه به طول انجامید. میانگین زمان صرف شده

توسط این مدل برای طبقه‌بندی هر رکورد فارسی ٤ میلی‌ثانیه و برای هر رکورد انگلیسی ۲ میلی‌ثانیه محاسبه شده است.

دقت این مدل در طبقه‌بندی رکوردهای مجموعه داده‌ی آزمایش فارسی ٪۸۰ و در طبقه‌بندی مجموعه داده‌ی آزمایش انگلیسی ٪۸٦ محاسبه شده است. به این ترتیب، مدل شبکه‌ی عصبی بازگشتی، با سرعت و دقتی قابل قبول، طبقه‌بندی رکوردها را انجام داده است. این مدل یکی از محبوب‌ترین مدل‌های طبقه‌بندی متن محسوب می‌شود و انتظار می‌رفت که دقت آن، از مقدار محاسبه شده نیز بالاتر باشد. هر چند، این مدل در شرایطی که تعداد رکوردها و زمان و منابع اختصاص یافته، بیشتر باشد؛ عملکرد بسیار بهتری خواهد داشت.

ط- مدل شبکه‌ی عصبی پیشخور

مدل شبکه‌ی عصبی پیشخور، به‌عنوان مدل اصلی ارائه شده در این تحقیق، یکی از نوین‌ترین و پرطرفدارترین مدل‌های ارائه شده در حوزه‌ی یادگیری عمیق می‌باشد. پیاده‌سازی این مدل، متکی بر یکی از شناخته شده‌ترین رابط‌های برنامه‌نویسی، یعنی رابط کِرِس که در کتابخانه‌ی نرم‌افزار تنسورفلو ارائه شده است، انجام شده است. تمام اقدامات انجام شده در مسیر پیاده‌سازی این مدل، مبتنی بر جدیدترین و بهترین فناوری‌های فعلی بوده است.

فرایند آموزش این مدل حدود ۱۸ ثانیه برای مجموعه داده‌ی فارسی و حدود ۳۳ ثانیه برای مجموعه داده‌ی انگلیسی زمان برد. میانگین زمان صرف شده توسط این مدل برای طبقه‌بندی هر رکورد فارسی ۰.۲ میلی‌ثانیه و برای هر رکورد انگلیسی ۰.۳ میلی‌ثانیه بوده است.

دقت این مدل در طبقه‌بندی رکوردهای مجموعه داده‌ی آزمایش فارسی ٪۹۱ و در طبقه‌بندی مجموعه داده‌ی آزمایش انگلیسی ٪۹٤ محاسبه شده است. بنابراین، مدل شبکه‌ی عصبی پیشخور، نه تنها رکوردها را با دقتی بالا طبقه‌بندی می‌کند؛ بلکه این کار را با سرعتی زیاد انجام می‌دهد. این مدل نسبت به سایر مدل‌های بررسی شده، با دقتی بالاتر عمل طبقه‌بندی را انجام داده است. صحت طبقه‌بندی رکوردها توسط این مدل، در هر دو مجموعه داده‌ی فارسی و انگلیسی، ٪۹۹ اندازه‌گیری شده است. این مدل در موقعیت‌های کمبود داده نیز عملکرد کاملاً مناسبی داشته است. مثلاً کلاس اخترشناسی را با وجود کمبود تعداد رکوردها، دقیق طبقه‌بندی کرده است (با امتیاز اف-یک ۰.۹۵ در مجموعه داده‌ی فارسی و ۰.۹۳ در مجموعه داده‌ی انگلیسی).

۵-۲-۳- مقایسه‌ی مدل‌ها

در جدول ۵—۱، مدل‌های پیاده شده، در کنار اطلاعات مربوط به معیارهای اندازه‌گیری عملکردشان و به ترتیب دقت طبقه‌بندی مجموعه داده‌ی فارسی، به شکل نزولی، آورده شده‌اند (دقیق‌ترین مدل، بالاتر از همه).

(جدول ۵–۱) مقایسه‌ی عملکرد مدل‌های پیاده‌سازی شده.

مدل	مجموعه داده‌ی فارسی		مجموعه داده‌ی انگلیسی	
	دقت ▼	صحت	دقت	صحت
ای‌ان‌ان (مدل اصلی)	۰.۹۱	۰.۹۹	۰.۹٤	۰.۹۹
رگرسیون لجستیک	۰.۸۸	۰.۹۹	۰.۹۱	۰.۹۹
اس‌وی‌ام	۰.۸۵	۰.۹۸	۰.۸۸	۰.۹۹
آران‌ان	۰.۸	۰.۹۸	۰.۸٦	۰.۹۸
اس‌وی‌ام هسته‌ای	۰.۷۹	۰.۹۸	۰.۸٦	۰.۹۸
بیز ساده	۰.٦٦	۰.۹٦	۰.٦٦	۰.۹٦
درخت تصمیم	۰.۵٦	۰.۹۵	۰.٦۲	۰.۹٦
کی‌ان‌ان	۰.۵	۰.۹٤	۰.۵۸	۰.۹۵
جنگل تصادفی	۰.٤۱	۰.۹۳	۰.۵۱	۰.۹٤

در جدول ۵–۲، مدت زمان صرف شده برای فرایند آموزش و طبقه‌بندی، در کنار میانگین مدت زمان صرف شده برای پیش‌بینی هر رکورد، به تفکیک هر مدل، قابل مشاهده است. مدل‌ها در این جدول به ترتیب میانگین مدت زمان پیش‌بینی هر رکورد فارسی و به شکل صعودی قید شده‌اند (سریع‌ترین مدل، بالاتر از همه).

(جدول ۵–۲) مقایسه‌ی مدت زمان صرف شده‌ی مدل‌ها در فرایندهای انجام شده (تمامی مدت زمان‌ها به ثانیه).

مدل	مجموعه داده‌ی فارسی			مجموعه داده‌ی انگلیسی		
	آموزش	طبقه‌بندی	میانگین ▲	آموزش	طبقه‌بندی	میانگین
جنگل تصادفی	٦	۰.۰۹	۰.۰۰۰۰۳	۱٤	۰.۲	۰.۰۰۰۰٤
درخت تصمیم	۷۵	۰.۳۱	۰.۰۰۰۱	۲۹۷	۱	۰.۰۰۰۲
رگرسیون لجستیک	۱۱۵	۰.۵۵	۰.۰۰۰۲	۱۵۰	۰.۳۸	۰.۰۰۰۱
ای‌ان‌ان	۱۸	۰.٦۳	۰.۰۰۰۲	۳۳	۱.۳٤	۰.۰۰۰۳
بیز ساده	۱.۷	۳.۳	۰.۰۰۱	۳	٤	۰.۰۰۰۲
آران‌ان	۱۰۲۸	۱۱	۰.۰۰٤	۹۸٦	۱۰.۲٦	۰.۰۰۲
کی‌ان‌ان	۰.۰۳	۲۲	۰.۰۰۷	۰.۰۳	۵۵	۰.۰۱۲
اس‌وی‌ام	۸۸٦	۱٦۳	۰.۰۵۷	۱۹٤٤	۵۵۳	۰.۱۱۸
اس‌وی‌ام هسته‌ای	۱۳۵۲	۳٦۰	۰.۱۲٦	٤٦٤۹	۱۰۲۹	۰.۲۱۹

۵-۲-٤- بحث در فرضیه‌های تحقیق

فرضیه‌های این تحقیق شامل یک فرضیه‌ی اصلی و سه فرضیه‌ی فرعی می‌باشند. فرضیه‌ی اصلی حاکی از این مطلب بود که «با بهره‌گیری از تکنیک‌های یادگیری عمیق ماشین، می‌توان روشی کاربردی و باکیفیت در راستای سازماندهی محتوا در فرایند مدیریت دانش سازمانی ارائه داد».

با توجه به یافته‌ها، مدل شبکه‌ی عصبی پیشخور که مدل اصلی ارائه شده‌ی در این تحقیق است، با دقتی بسیار عالی، حتی در موارد کمبود داده، رکوردهای فارسی و انگلیسی را طبقه‌بندی کرد. این مدل موفق شد که هر دو دسته مجموعه داده‌ی آزمایش را دقیق‌تر از سایر مدل‌های رایج حوزه‌ی یادگیری ماشین و یادگیری عمیق، طبقه‌بندی کند. دقت طبقه‌بندی این مدل بیش از ٪۹۰ برای هر دو مجموعه داده ارزیابی شد. این مقدار، با توجه به استانداردهای طبقه‌بندی در زمینه‌ی علوم داده، کاملاً مناسب است.

ضمن دقت بالای مدل ارائه شده، سرعت آن در طبقه‌بندی رکوردها نیز بسیار مناسب است. میانگین سرعت طبقه‌بندی برای هر رکورد (پاراگراف) آزمایش، در هر دو مجموعه داده‌ی مورد بررسی، کمتر از یک میلی‌ثانیه محاسبه شد.

با توجه به مطالب قید شده و با در نظر گرفتن یافته‌ها، می‌توان این ادعا را کرد که فرضیه‌ی اصلی تحقیق تأیید شده است؛ یعنی مدل شبکه‌ی عصبی پیشخور ارائه شده، روشی کاربردی و باکیفیت در راستای سازماندهی محتوا در فرایند مدیریت دانش سازمانی می‌باشد.

در جدول ۵—۳ رد یا تأیید شدن فرضیه‌های تحقیق، بررسی شده است.

(جدول ۵—۳) نتایج فرضیه‌های تحقیق.

#	فرضیه	نتیجه
۱	با بهره‌گیری از تکنیک‌های یادگیری عمیق ماشین، می‌توان روشی کاربردی و باکیفیت در راستای سازماندهی محتوا در فرایند مدیریت دانش سازمانی ارائه داد.	تأیید
۲	میزان اثربخشی (دقت) مدل ارائه شده در طبقه‌بندی رکوردهای آزمایش بیش از ۰.۹ خواهد بود.	تأیید
۳	زمان مورد نیاز برای بررسی و سازماندهی یک پاراگراف آزمایش توسط مدل ارائه شده، با داشتن ۱۰ هزار رکورد در پایگاه داده، کمتر از ۰.۱ ثانیه خواهد بود (کارایی مدل قابل قبول خواهد بود).	تأیید
٤	روش حاصل از به‌کارگیری فناوری‌های نوین، نسبت به رویکردهای سنتی مدنظر، سازماندهی محتوا را دقیق‌تر انجام می‌دهد.	تأیید

۵-۳- نتیجه‌گیری

پس از مقایسه‌ی مدل‌های سنتی‌تر یادگیری ماشین و مدل مبتنی بر شبکه‌ی عصبی پیشخور، می‌توان نتیجه گرفت که مطابق انتظار، مدل مبتنی بر شبکه‌ی عصبی، دقت طبقه‌بندی بهتری را ارائه می‌کند. این درحالی است که سرعت این مدل در پیش‌بینی کلاس‌ها نیز بسیار بالا و قابل قبول است. درنتیجه بهبود دقت ایجاد شده منجر به استفاده‌ی بیش از حد از منابع سخت‌افزاری نمی‌شود.

از طرفی، برخی مدل‌های جایگزین، به‌خصوص مدل رگرسیون لجستیک، که می‌توان آن‌را یک شبکه‌ی عصبی تک‌لایه دانست، عملکرد بسیار مناسبی دارند. رگرسیون لجستیک نه تنها با دقت بسیار قابل قبولی رکوردهای آزمایش را طبقه‌بندی کرد؛ بلکه این کار را در مدت زمانی کوتاه به انجام رساند. به همین دلیل است که این مدل، هنوز یکی از پرطرفدارترین مدل‌های حوزه‌ی یادگیری ماشین و علوم داده به‌حساب می‌آید.

با این وجود، مدل مبتنی بر شبکه‌ی عصبی توانست حتی دقتی فراتر از دقت طبقه‌بندی مدل قدرتمند رگرسیون لجستیک ارائه دهد. این نشان از توانایی منحصر به فرد مدل‌های مبتنی بر شبکه‌ی عصبی در زمینه‌ی طبقه‌بندی دارد. به همین دلیل بسیاری از افراد و سازمان‌های کوچک و بزرگ، هم‌اکنون به سمت بهره‌گیری از شبکه‌های عصبی سوق پیدا کرده‌اند.

بنابراین، مدل ارائه شده در این تحقیق، که در هفت فاز معرفی شد و مرحله‌ی مهم آماده‌سازی داده‌ها را نیز دربر گرفته است، کیفیت بسیار بالایی در زمینه‌ی سازماندهی محتویات ورودی به سیستم مدیریت دانش، ایجاد می‌کند. این مدل می‌تواند با ایجاد بهبود در کارایی و اثربخشی مدل‌های سنتی‌تر، مزیت‌های فراوانی را برای سازمان‌ها به ارمغان آورد.

۵-٤- محدودیت‌های تحقیق

شاید بزرگترین محدودیت در این تحقیق، نبود مجموعه داده‌ی فارسی استانداردی بود که بتوان از تمایز داشتن رکوردهای هر کلاس آن، اطمینان حاصل کرد. این محدودیت باعث شد که نرم‌افزاری جامع توسعه داده شود تا محتویات مقالات ویکی‌پدیا را استخراج و آن‌ها را از فرمت اچ‌تی‌ام‌ال به فرمت مناسب تبدیل کند. هرچند این نرم‌افزار مزیت‌های فراوانی را به همراه دارد؛ اما از تمایز داشتن پاراگراف‌ها و مرتبط بودن آن‌ها، تنها به یک کلاس تعریف شده، اطمینان حاصل نمی‌کند.

بنابراین ممکن است، برخی از رکوردها به چند کلاس مربوط باشند. درنتیجه این احتمال وجود دارد که در مواردی، مدل‌ها رکوردی را به کلاسی مرتبط نسبت داده‌اند اما این اقدامِ درست، به‌عنوان اقدامی نادرست تلقی شده است. لذا این امکان وجود دارد که دقت طبقه‌بندی رکوردها، از مقدار گزارش شده، کمی بالاتر باشد.

محدودیت‌های سخت‌افزاری، محدودیت بارز دیگری بود که در مسیر انجام این تحقیق وجود داشت. با تشکر از شرکت گوگل و بواسطه‌ی پلتفرم گوگل کولب، این امکان فراهم شد تا به منابع سخت‌افزاری نسبتاً مناسب، دسترسی پیدا شده و امکان به اشتراک‌گذاری پروژه وجود داشته باشد. اما این منابع سخت‌افزاری آنقدر قدرتمند نیستند که بتوان مبتنی بر آن‌ها شرایط پیچیده‌تر که تعداد داده‌ها در آن بسیار زیاد است را شبیه‌سازی کرد.

لذا نتایج حاصل از این تحقیق، در نشان دادن عملکرد مدل‌ها در مواجهه با کلان‌داده‌ها، قابلیت اتکای بالایی ندارد. به‌خصوص در مورد مدل‌های مبتنی بر شبکه‌ی عصبی (از جمله شبکه‌ی عصبی بازگشتی) می‌توان این اظهار را داشت که احتمالاً میزان تفاوت این مدل‌ها با مدل‌های جایگزین، در مواجهه با کلان‌داده‌ها، بسیار بیشتر می‌باشد. به عبارت دیگر، در این موقعیت‌ها، با توجه به ماهیت یادگیری مدل‌های شبکه‌ی عصبی، احتمالاً عملکرد مناسب‌تر آن‌ها نسبت به مدل‌های جایگزین، نمود بیشتری خواهد داشت.

۵-۵- پیشنهادات تحقیق

در مسیر این تحقیق و با بررسی یافته‌ها، تجربیاتی اندوخته و نتایجی حاصل شد که می‌تواند در زمینه‌های مختلف، بستری برای بهبود ایجاد کند. این تجربیات و نتایج اندوخته، در قالب پیشنهادات کاربردی و پیشنهادات پژوهشی، در اختیار مخاطب قرار داده شده است.

۵-۵-۱- پیشنهادات کاربردی

تغییر و تحول مداوم و پرسرعت معادلات، ساختارها و شرایط کلی دنیای امروز، غیر قابل انکار است. همه‌گیری کووید ۱۹ نشان داد که حوادث غیر مترقبه همواره در کمین هستند و نداشتن آمادگی مناسب در مواجهه با این حوادث می‌تواند، تبعات و پیامدهای منفی زیادی را به همراه داشته باشد. انعطاف‌پذیری در مقابل حوادث غیر مترقبه، برای سازمان‌ها که به‌شدت از محیط خارجی تأثیر پذیرند، اهمیت دوچندان دارد.

علوم داده و مفاهیم نوین این عرصه، شامل یادگیری ماشین و یادگیری عمیق، زمینه‌های بالفعل زیادی برای مدیریت مناسب‌تر سازمان‌ها فراهم کرده‌اند. یکی از مهم‌ترین زمینه‌های بالفعل، که می‌توان با بهره‌گیری از یادگیری ماشین بهبود قابل توجهی در آن ایجاد کرد، زمینه‌ی مدیریت دانش می‌باشد.

در این تحقیق نشان داده شد که چگونه مدل‌های مختلف حوزه‌ی یادگیری ماشین و به‌خصوص یادگیری عمیق می‌توانند در سازماندهی محتوای ورودی به سیستم مدیریت دانش، مفید واقع شوند. غفلت از به‌کارگیری مناسب این فناوری‌های نوین و توسعه‌ی سیستم مدیریت دانشی مناسب، بر پایه‌ی آن‌ها، منجر به محروم شدن

سازمان‌ها از یک مزیت رقابتی منحصر به فرد، در راستای دستیابی به هدف سازمانی شده و حتی بقای آن‌ها را نیز به خطر اندازد.

بنابراین، توصیه‌ی محقق به سازمان‌ها، به‌خصوص سازمان‌های واقع در ایران، این است که در راستای بهره‌گیری از مزیت‌های وافر فناوری‌های یادگیری ماشین، کوتاهی نکنند. با توجه به شرایط فعلی کشورمان، در نظر گرفتن سناریوهای محتمل و آینده‌نگری در فرایند اخذ تصمیم، بسیار مهم است. تصمیم‌گیری آینده‌نگرانه بر مبنای اطلاعات کلیدی سازمان انجام می‌گیرد. درنتیجه، یک سیستم مدیریت دانش مناسب می‌تواند شرایطی مطلوب برای کسب اطلاعات مدنظر و تصمیم‌گیری مناسب فراهم آورد. برای دستیابی به چنین سیستمی، بهره‌بردن مناسب از فناوری‌های حوزه‌ی یادگیری ماشین، یک الزام است.

مدل‌های مبتنی بر شبکه‌ی عصبی، عملکرد بهتری از مدل‌های جایگزین دارند. پس به‌روز کردن سیستم مدیریت دانش و سایر سیستم‌های اطلاعاتی به نحوی که از این رویکرد قدرتمند بهره‌مند شوند، یک اقدام ضروری است. در این تحقیق تلاش شده است که مسیر دستیابی به مدلی بهینه بر پایه‌ی شبکه‌های عصبی مصنوعی شرح داده شود. اطلاعات مربوط به این مدل در اختیار همگان قرار گرفته تا بتوانند از این اطلاعات، در راستای توسعه‌ی مدل‌های مبتنی بر شبکه‌ی عصبی، بهره‌مند شوند.

۵-۵-۲- پیشنهادات پژوهشی

همانطور که در بخش محدودیت‌های تحقیق اشاره شد، فقدان مجموعه داده‌های فارسی استاندارد برای داده‌کاوی، به‌خصوص برای آزمایشات مربوط به طبقه‌بندی، یکی از معضلات موجود در مسیر داده‌کاوان است. برای حل این مشکل، یک برنامه‌ی کامپیوتری توسعه داده شد که عمل استخراج و مناسب‌سازی رکوردها را انجام می‌دهد. محدودیت این نرم‌افزار در کسب اطمینان از مختص بودن رکوردها به یک کلاس واحد است. یکی از راه‌حل‌های ممکن برای کسب اطمینان از مرتبط بودن رکوردها صرفاً به یک کلاس خاص، حذف رکوردهایی است که توسط چند مدل، به کلاس‌های نادرست نسبت داده می‌شوند. آزمایش عملکرد این راهکار یا ارائه‌ی راهکاری برای بهبود فرایند جمع‌آوری مجموعه داده‌های مناسب فارسی، می‌تواند زمینه‌ی مطالعاتی مناسبی برای پژوهشگران باشد.

محدودیت دیگر این تحقیق، قابل اتکا نبودن نتایج آن در بازتاب عملکرد مدل‌ها در طبقه‌بندی کلان‌داده‌ها است. درنتیجه، زمینه‌ی دیگری که می‌تواند گزینه‌ی مناسبی برای تحقیقات بیشتر باشد، بررسی عملکرد مدل‌ها، به‌خصوص مدل‌های مبتنی بر شبکه‌ی عصبی بازگشتی، در طبقه‌بندی کلان‌داده‌ها است.

این تحقیق در زمینه‌ی طبقه‌بندی و یادگیری نظارت شده به انجام رسید. به نظر می‌رسد، با توجه به خلع تحقیقاتی به‌خصوص در پژوهش‌های فارسی زبان، نیاز به تحقیقات بیشتر در عرصه‌ی یادگیری بدون نظارت و خوشه‌بندی محتویات وجود دارد.

فهرست منابع

فهرست منابع فارسی

۱. باقرزاده، سارا، آرش مقصودی، احمد شالباف. ۱۴۰۰. تشخیص بیماران اسکیزوفرنی با استفاده از شبکه‌های عصبی پیچشی از تصاویر ارتباطات مؤثر مغزی سیگنال‌های چندکاناله الکتروانسفالوگرام. مجله دانشکده پزشکی دانشگاه علوم پزشکی تهران، دوره ۷۹، شماره ۷: ۷۶۳-۷۵۴.

۲. بیگدلی، نوشین، حامد جباری و مهدی شجاعی. ۱۴۰۰. یک روش هوشمند برای طبقه‌بندی ترک در سازه‌های بتنی. نشریه مهندسی عمران امیرکبیر، دوره ۵۳، شماره ۸: ۲۱-۳. doi:10.22060/ceej.2020.17738.6660.

۳. صباحی، کامل، سبحان شیخی‌وند، زهره موسوی و مهدی رجبیون. ۱۴۰۰. شناسایی موارد ابتلا به کووید-۱۹ با استفاده از شبکه‌ی عصبی فازی نوع ۲ عمیق براساس تصاویر X-Ray قفسه سینه. هوش محاسباتی در مهندسی برق. doi:10.22108/isee.2021.128829.1475.

فهرست منابع انگلیسی

٤. Aggarwal, Aggarwal C. 2018. Neural Networks and Deep: A Textbook. Berlin: Springer International Publishing.

٥. Amato, Gabriele, Lorenzo Palombi, and Valentina Raimondi. 2021. Data–driven classification of landslide types at a national scale by using Artificial Neural Networks. International Journal of Applied Earth Observation and Geoinformation 104: 102549. doi: https://doi.org/10.1016/j.jag.2021.102549.

٦. Beynon-Davies, Paul. 2020. Business Information Systems. 3rd. London: Palgrave Macmillan.

٧. Biabiany, Emmanuel, Didier C. Bernard, Vincent Page, and Hélène Paugam-Moisy. 2020. Design of an expert distance metric for climate clustering: The case of rainfall in the Lesser Antilles. Computers & Geosciences (Elsevier B.V.) 145: 104612. doi: https://doi.org/10.1016/j.cageo.2020.104612.

٨. Bruce, Peter, Andrew Bruce, and Peter Gedeck. 2020. Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python. 2nd. Sebastopol: O'Reilly Media, Inc.

٩. Burkov, Andriy. 2019. The Hundred Page Machine Learning Book. Andriy Burkov.

١٠. Chakraborty, Indrasis, Brian M. Kelley, and Brian Gallagher. 2021. Industrial control system device classification using network traffic features and neural network embeddings. Array 12: 100081. doi: https://doi.org/10.1016/j.array.2021.100081.

١١. Chollet, François. 2021. Deep Learning with Python. 2nd. Shelter Island: Manning Publications.

١٢. Cochran, William G. 1977. Sampling Techniques. 3rd. Hoboken: John Wiley & Sons.

١٣. Cormen, Thomas H., Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. 2022. Introduction to Algorithms. 4. London: The MIT Press.

١٤. Di, Wei, Anurag Bhardwaj, and Jianing Wei. 2018. Deep Learning Essentials. Birmingham: Packt Publishing Ltd.

١٥. Downey, Allen B., and Chris Mayfield. 2020. Think Java: How to Think Like a Computer Scientist. 2nd. Sebastopol: O'Reilly Media.

١٦. Ekambaram, Anandasivakumar, Anette Ø. Sørensen, Heidi Bull-Berg, and Nils O.E. Olsson. 2018. The role of big data and knowledge management in improving projects and project-based organizations. Procedia Computer Science (Elsevier) 138: 851-858. doi: https://doi.org/10.1016/j.procs.2018.10.111.

١٧. Genuer, Robin, and Jean-Michel Poggi. 2020. Random Forests with R. Cham: Springer Nature.

١٨. Géron, Aurélien. 2019. Hands-On Machine Learning. 2nd. Sebastopol: O'Reilly Media.

١٩. Ghosh, P.P. 2021. Principles and Practices of Management. New Dehli: Laxmi Publications.

٢٠. Grus, Joel. 2019. Data Science from Scratch: First Principles with Python. 2nd. Sebastopol: O'Reilly.

٢١. Howard, Jeremy, and Sylvain Gugger. 2020. Deep Learning for Coders with Fastai & PyTorch. Sebastopol: O'Reilly Media.

٢٢. Husain, Shabahat, and Jean-Louis Ermine. 2021. Knowledge Management Systems: Concepts, Technologies and Practices. Bingley: Emerald Publishing.

٢٣. Huyen, Chip. 2022. Designing Machine Learning Systems. Sebastopol: O'Reilly Media.

٢٤. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. An Introduction to Statistical Learning with Applications in R. 2nd. New York: Springer Science+Business Media.

٢٥. Joshi, Ketan, Tripathi Vikas, Chitransh Bose, and Chaitanya Bhardwaj. 2020. Robust Sports Image Classification Using InceptionV3 and Neural Networks. Procedia Computer Science 167: 2374-2381. doi: https://doi.org/10.1016/j.procs.2020.03.290.

٢٦. Kingma, Diederik P., and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. arXiv 1-15. doi:10.48550/ARXIV.1412.6980.

٢٧. Koç, Tuğba, Kadir Kurt, and Adem Akbıyık. 2019. A Brief Summary of Knowledge Management Domain: 10-Year History of the Journal of Knowledge Management. Procedia Computer Science 158: 891–898. doi: https://doi.org/10.1016/j.procs.2019.09.128.

٢٨. Larose, Chantal D., and Daniel T. Larose. 2019. Data Science Using Python and R. Hoboken: Wiley.

٢٩. Li, Yuanlu, Renfei Qian, and Kun Li. 2022. Inter-patient arrhythmia classification with improved deep residual. Computer Methods and Programs in Biomedicine 214: 106582. doi: https://doi.org/10.1016/j.cmpb.2021.106582.

٣٠. Mohanty, Ashima Sindhu, Priyadarsan Paridab, and Krishna Chandra Patraa. 2021. ASD classification for children using deep neural network. Global Transitions Proceedings 2 (2): 461-466. doi: 10.1016/j.gltp.2021.08.042.

٣١. Nong, Liping, Junyi Wang, Jiming Lin, Hongbing Qiu, Lin Zheng, and Wenhui Zhang. 2021. Hypergraph wavelet neural networks for 3D object classification. Neurocomputing 463: 580-595. doi: https://doi.org/10.1016/j.neucom.2021.08.006.

٣٢. Raschka, Sebastian, Yuxi Hayden Liu, and Vahid Mirjalili. 2022. Machine Learning with PyTorch and Scikit-Learn. Birmingham: Packt Publishing.

٣٣. Ring, M, A Dallmann, D Landes, and A Hotho. 2017. IP2Vec: Learning Similarities Between IP Addresses. 2017 IEEE international conference on data mining workshops (ICDMW). IEEE. 657-666. doi:10.1109/ICDMW.2017.93.

٣٤. Roberts, Daniel A., Sho Yaida, and Boris Hanin. 2022. The Principles of Deep Learning Theory. Cambridge: Cambridge University Press.

٣٥. Russell, Stuart J, and Peter Norvig. 2021. Artificial Intelligence: A Modern Approach. 4th. London: Pearson Education.

٣٦. Sarkar, Dipanjan. 2019. Text Analytics with Python. New York City: Apress.

٣٧. Singh, Manu Pratap, and Gunjan Singh. 2021. Two phase learning technique in modular neural network for pattern classification of handwritten Hindi alphabets. Machine Learning with Applications 6: 100174. doi: https://doi.org/10.1016/j.mlwa.2021.100174.

٣٨. Stevens, Eli, Luca Antiga, and Thomas Viehmann. 2020. Deep Learning with PyTorch. Shelter Island: Manning Publications.

٣٩. Tan, Pang-Ning, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. 2019. Introduction to Data Mining. Harlow: Pearson Education.

٤٠. Thakur, Samritika, and Aman Kumar. 2021. X-ray and CT-scan-based automated detection and classification of covid-19 using convolutional neural networks (CNN). Biomedical Signal Processing and Control 102920. doi: https://doi.org/10.1016/j.bspc.2021.102920.

٤١. University of Minnesota. 2015. Principles of Management. Minneapolis: University of Minnesota.

٤٢. Wambugu, Naftaly, Yiping Chen, Zhenlong Xiao, Kun Tan, Mingqiang Wei, Xiaoxue Liu, and Jonathan Li. 2021. Hyperspectral image classification on insufficient-sample and feature learning using deep neural networks: A review. International Journal of Applied Earth Observation and Geoinformation 102603. doi: https://doi.org/10.1016/j.jag.2021.102603.

٤٣. Weidman, Seth. 2019. Deep Learning from Scratch: Building with Python from First Principles. Sebastopol: O'Reilly Media.

فهرست منابع الکترونیک

٤٤. Cambridge University Press. n.d. Deep Learning. Cambridge dictionary. Accessed 1 15, 2022. (https://dictionary.cambridge.org/us/dictionary/english/deep-learning).

٤٥. Cambridge University Press. n.d. Indexing. Cambridge dictionary. Accessed 10 18, 2022. (https://dictionary.cambridge.org/us/dictionary/english/indexing).

٤٦. Cambridge University Press. n.d. Knowledge. Cambridge dictionary. Accessed 10 18, 2022. (https://dictionary.cambridge.org/us/dictionary/english/knowledge).

٤٧. Cambridge University Press. n.d. Machine Learning. Cambridge dictionary. Accessed 10 20, 2022. (https://dictionary.cambridge.org/us/dictionary/english/machine-learning).

٤٨. Cambridge University Press. n.d. Organizing. Cambridge dictionary. Accessed 20 10, 2022. (https://dictionary.cambridge.org/us/dictionary/english/organizing).

٤٩. Costello, Katie. 2019. Gartner Survey Shows 37 Percent of Organizations Have Implemented AI in Some Form. January 21. (https://www.gartner.com/en/newsroom/press-releases/2019-01-21-gartner-survey-shows-37-percent-of-organizations-have#:~:text=%E2%80%9CFour%20years%20ago%2C%20AI%20implementation,research%20vice%20president%20at%20Gartner).

٥٠. Cournapeau, David. 2022. scikit-learn: machine learning in Python. October 26. Accessed November 10, 2022. (https://github.com/scikit-learn/scikit-learn).

٥١. Google. n.d. Google Colab. Accessed December 7, 2022. (https://research.google.com/colaboratory/faq.html).

٥٢. Keras Contributors. 2022. Adam (Optimizer). June 22. Accessed December 4, 2022. (https://keras.io/api/optimizers/adam/).

٥٣. Python Software Foundation. n.d. What is Python? Executive Summary. Accessed December 7, 2023. (https://www.python.org/doc/essays/blurb/).

٥٤. The R Foundation. n.d. What is R? Accessed December 7, 2022. (https://www.r-project.org/about.html).

٥٥. Wikipedia Editors. 2022. Size of Wikipedia. November 22. Accessed November 25, 2022. (https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia).

٥٦. Wikipedia. 2022. Wikipedia Statistics. November 25. Accessed November 25, 2022. (https://stats.wikimedia.org/#/all-projects).

پیوست ۱: اطلاعات دسترسی به فایل‌های پروژه

صفحه‌ی گیت‌هاب پژوهش:

https://github.com/dev2223/Organizing-KMS-Content-via-ANN-Thesis

صفحه‌ی گوگل کولب (کد منبع مدل‌ها):

https://colab.research.google.com/drive/1tJuMjt3ePEr37NnfBFGG4L0w6aGFTykC?usp=sharing#scrollTo=7Aie1wQaQInd

صفحه‌ی نتایج عملکرد مدل‌ها:

https://docs.google.com/spreadsheets/d/1YXxynkC3qvrxzHYp8-_XNUw_2Q2A9BNWxVNzZ_OgeRc/edit?usp=sharing

پیوست ۲: نمونه‌ای از پاراگراف آماده شده برای ورود به مدل‌ها

آرایه‌ی زیر، نمونه‌ای از پاراگراف آماده شده برای ورود به مدل طبقه‌بندی کننده است. توجه داشته باشید که اندازه‌ی این آرایه، برای جلوگیری از جاگیری زیاد، محدود شده است.

[۱۹۵, ۱۵۲, ۲۲, ۱۲, ۱۱۷, ۵۲, ۵۰, ۱۶۴, ۷۸, ۷۶, ۳۷, ۷۵, ۶۶, ۲۲۱, ۷۱, ۲۱۱, ۱۴۹, ۷۷, ۶۹, ۸۵, ۱۸۱, ۱۰۶, ۱۲۳, ۲۴۷, ۳۸, ۱۹۲, ۶۱, ۱۴۳, ۲۴۹, ۱۶۳, ۲۶, ۲۳۲, ۲۵, ۷۱, ۱۰۷, ۱۷۳, ۲۱۵, ۵۷, ۵۳, ۸۴, ۲۱۲, ۱۴۳, ۱۳۹, ۱۳۹, ۱۶۰, ۱۸۸, ۱۳۰, ۳۲, ۲۰۹, ۱۲۸, ۲۲۵, ۲۳۴, ۶۱, ۱۸۰, ۲۲, ۱۵۳, ۸۳, ۷, ۱۴۹, ۹۱, ۱۳۱, ۱۶۴, ۱۰۵, ۱۶۴, ۱۰۲, ۳, ۲۰۵, ۳۶, ۴۸, ۲۲۹, ۲۴۶, ۱۵۰, ۹۲, ۲۱۷, ۱۶۸, ۴۲, ۱۷۷, ۱۶۶, ۱۳۵, ۷۰, ۲۵, ۹۷, ۱۲۸, ۲۴۶, ۴۲, ۲۱۳, ۲۰۷, ۲۳۷, ۱۹۴, ۱۹۷, ۶۳, ۲۴۸, ۱۹۲, ۲۲۵, ۱۸۳, ۲۲۲, ۱۲, ۱۵۲, ۱۸۴, ۱۵۴, ۲۹, ۲۴۲, ۱۷۸, ۳۴, ۳۲, ۱۰۱, ۷۶, ۵۱, ۱۵۲, ۱۹۲, ۳۰, ۳۱, ۴۵, ۲۹, ۵۲, ۱۴۰, ۱۰۱, ۲۵۲, ۱۳۵, ۱۰۰, ۱۱, ۳۶, ۱۴۲, ۲۵, ۲۱۹, ۲۵۲, ۹۰, ۲۳, ۱۴۱, ۱۳۳, ۱۰, ۲۱۹, ۱۲۶, ۱۷, ۱۳۲, ۱۴۹, ۷۳, ۴۰, ۱۱۴, ۱۷۵, ۹۰, ۲۳۶, ۳۵, ۲۵۵, ۱۶۰, ۲۲, ۱۴۴, ۴۷, ۱۹۱, ۱۹, ۴۷, ۳۲, ۲۴, ۸۸, ۴۴, ۱۸۸, ۱۷۲, ۲۳۹, ۵۲, ۱۹۵, ۲۲۸, ۶, ۴۹, ۱۶۸, ۴۶, ۲۶, ۷۱, ۱۳۵, ۵۶, ۶۶, ۱۱۶, ۲۳۱, ۱۴۸, ۸, ۱۴۴, ۶۶, ۲۹, ۱۴, ۹۰, ۲۴۹, ۸۶, ۱۹۱, ۹۳, ۱۲۵, ۱۲۱, ۲۳۴, ۲۱۸, ۱۵۲, ۲۴۱, ۱۰, ۷۰, ۳۵, ۸, ۱۹۷, ۲۳۹, ۲۲۴, ۱۵۰, ۲۲۰, ۲۵, ۱۶, ۱۲۵, ۱۲۸, ۲۳۳, ۵۱, ۵۲, ۲۴۸, ۴۶, ۲۴۱, ۲۳۷, ۲۳۶, ۲۱۹, ۲۲۶, ۱۷۹, ۲۱۲, ۲۳۷, ۶۰, ۹۲, ۱۱۲, ۱۴۷, ۲۴۶, ۱۵۹, ۳۰, ۲۲۹, ۲۴۱, ۵۰, ۲۰۰, ۱۲۵, ۸۰, ۱۳۷, ۱۷, ۵۳, ۱۱۵, ۱۹۹, ۱۶۲, ۳۴, ۱۵۸, ۳۵, ۱۷۳, ۱۰۱, ۱۶۸, ۱۲۲, ۸۳, ۱۹, ۱۱۲, ۲۰۲, ۲۵۴, ۱۳۵, ۵۴, ۲۶, ۱۱۵, ۱۱, ۱۹۸, ۱۶۳, ۱۱۷, ۱۳۴, ۲۳, ۲۸, ۴, ۱۳۵, ۲۱۹, ۶۲, ۱۳۲, ۴۵, ۲۵۱, ۲۱۱, ۱۳۲, ۱, ۲۰۰, ۱۳۶, ۹, ۲۲۴, ۲۰۲, ۲۴۲, ۱۳۹, ۴۹, ۱۹۵, ۱۴۹, ۵۵, ۴, ۲۴۳, ۲۶, ۱۷۳, ۱۷۱, ۱۳۰, ۱۷۵, ۸۰, ۱۲۳, ۱۳۳, ۱۴۴, ۲۴۹, ۶۶, ۱۰, ۲۲۰, ۲۳۲, ۲۵۰, ۹۸, ۸۲, ۱۵۹, ۲۱۱, ۲۱۹, ۱۷۵, ۹۱, ۱۱۰, ۱۹۷, ۶۸, ۱۲۰, ۱۲۰, ۱۲۸, ۷۷, ۲۴۱, ۱۷۱, ۲۲۶, ۲۰۱, ۱۱۱, ۱۴۹, ۲۵۳, ۲۰۷, ۳۵, ۳۷, ۲۴۳, ۲۰۵, ۶۹, ۱۷۸, ۱, ۹۶, ۱۵۸, ۱۰۵, ۲۸, ۲۱۱, ۱۶۲, ۱۱۹, ۶۲, ۲۳۶, ۲۱۳, ۴۷, ۴۴, ۴۷, ۱۹۰, ۱۵۶, ۲۱۶, ۱۹۰, ۱۳۱, ۱۸, ۲۱۰, ۱۷۲, ۲۳۹, ۷۹, ۹۸, ۱۷۷, ۲۱۴, ۱۴۳, ۲۴۵, ۲۱۳, ۲۳۱, ۲۲۴, ۱۴۰, ۱۴۵, ۱۹۱, ۲۳۳, ۱۸۲, ۱۰۶, ۶۹, ۲۴۸, ۱۵۲, ۱۲۸, ۱۰۴, ۲۳, ۳۲, ۱۵۰, ۱۸۲, ۱۴۰, ۳۶, ۲۲۵, ۱۲۸, ۲۴۹, ۱۲۶, ۱۵۵, ۳۱, ۱۲۸, ۷۶, ۱۴, ۷, ۱۰, ۲۳۸, ۲۲۴, ۵۴, ۲۴۱, ۵, ۱۸۵, ۹۳, ۱۵۸, ۴۴, ۱۹, ۱۷۷, ۲۰۶, ۱۴۳, ۴۰, ۱۹۰, ۷۰, ۲۳۵, ۱۰۵, ۲۴۱, ۴۲, ۱۵۷, ۱۳۰, ۱۸۵, ۸۷, ۴۴, ۲۱۵, ۳۵, ۱۹۰, ۸۸, ۵۶, ۱۵, ۱۸۰, ۷۹]

Abstract

Nowadays, the pace of world evolution is very high. Organizations constantly face new opportunities and threats. The ability to tolerate changes and make the most of the resulting capacities is essential for small and large organizations. The role of information and organizational knowledge becomes a key factor here. The contents of a knowledge management system provide the decision-making requirements at different management levels. Therefore, the proper and optimal classification of input data to knowledge management systems lays the foundation for making better decisions, & enables organizations to face diverse environmental conditions. Data science, especially the unique topic of deep learning, has created a big revolution in the organizing process of data and information. Taking advantage of deep learning concepts is not merely a competitive advantage but vital to the organization's survival. In this research, relying on the latest approaches of deep learning and artificial neural networks, a method has been presented that allows the classification of input contents to knowledge management systems with remarkable efficiency and effectiveness. We have compared the performance of this method with other prominent machine-learning counterparts and analyzed its pros and cons.

The presented method has performed better than alternative methods by achieving a classification accuracy of 94% for the English and 91% for the Persian dataset. Also, the average classification speed for each English and Persian paragraph was 0.3 and 0.2 milliseconds, respectively, which shows the extraordinary efficiency of this method.

Keywords: classification, knowledge management system, artificial intelligence, machine learning, deep learning, artificial neural networks, TensorFlow software library

ISLAMIC AZAD UNIVERSITY
SCIENCE AND RESEARCH BRANCH
Faculty of Management & Economics - Department of MBA

Thesis for receiving «M.A» degree on Business Administration
Field: Information Systems & Information Technology

Subject:

Presenting a method for content classification in the organizational knowledge management process with the artificial neural networks approach

Thesis Advisor:
Mohammadreza Movahedi Sefat Ph.D.

By:
Abbas Esmaili

Winter 2022