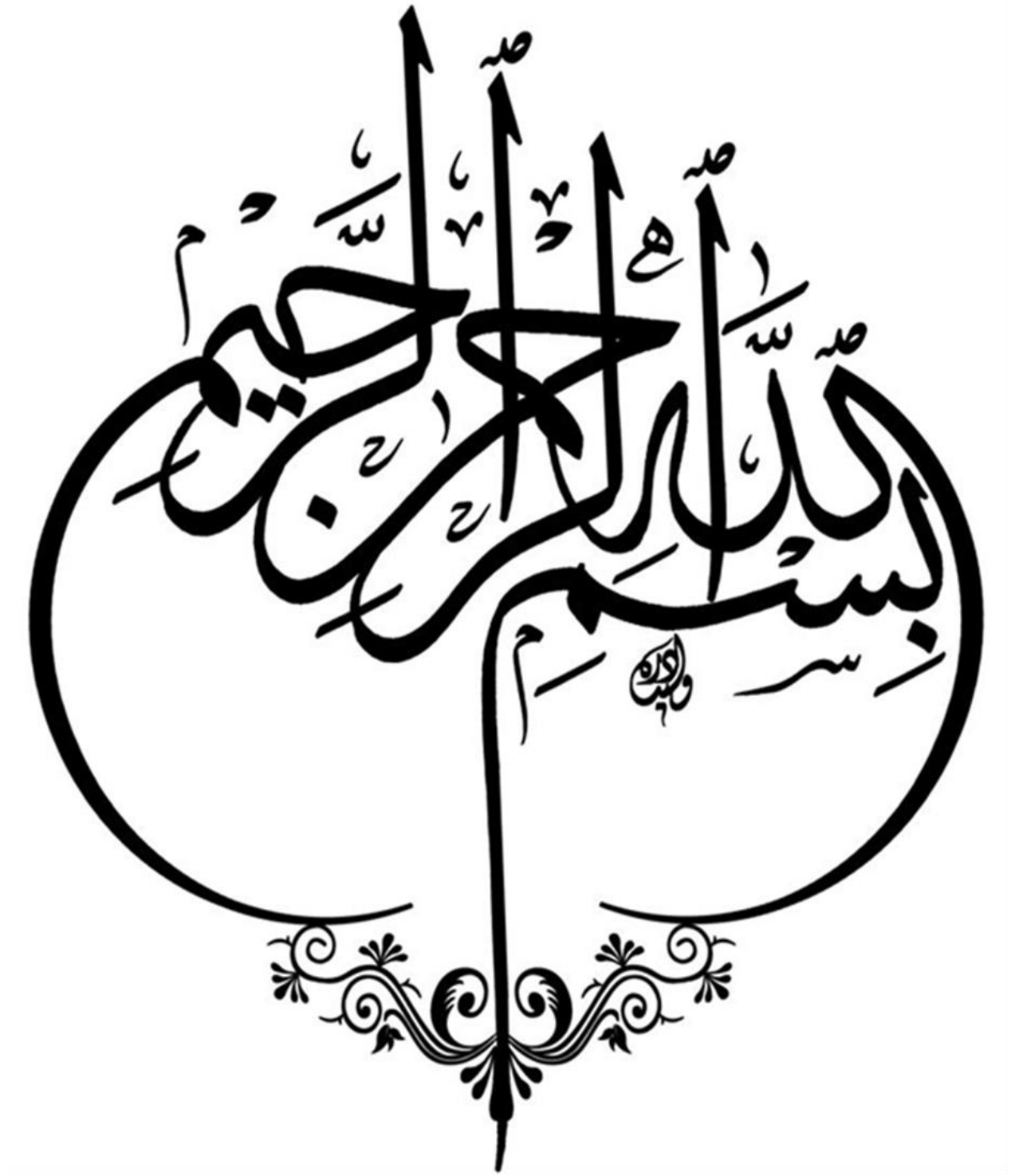

بسم الله الرحمن الرحيم

چکیده

زنجیره‌های تامین امروزی که پیوسته توسط جهانی‌سازی، برون‌سپاری، پیکربندی‌های محصول، نوسانات تقاضا، هزینه‌ها و SKUهای متنوع به چالش کشیده می‌شوند، فوق‌العاده پیچیده هستند. همه این فشارها شرکت‌ها را وادار می‌کند راهی برای ارضای نیازهای مشتریان بیابند که به‌موقع، کارآمد و سودآور باشد. گسترش سطح دسترسی به اینترنت و افزایش میل به خرید آنلاین، تعداد مشتریان را افزایش داده است. در شرایطی که تنوع کالا و تعداد مشتریان زیاد است، حل مسائلی مانند تحویل به‌موقع کالا یا خدمات، انتخاب و تعیین سفارش‌ها در انبارهای غیرمتمرکز، تخصیص انبار به مشتریان و... دشوار است. برای حل این چالش‌ها، استفاده از مدل‌سازی ریاضی با روش‌های حل فراابتکاری پیشنهاد شده است. اما به دلیل تعداد زیاد حالت‌های تخصیص، حل مدل‌های ریاضی بسیار پیچیده و زمان‌بر است. با بهبود قدرت محاسباتی و فضای ذخیره‌سازی، روش‌های مبتنی بر داده برای حل این چالش‌ها موردمطالعه و بررسی قرار گرفته است. ازاین‌رو در این پژوهش نیز از روش داده‌کاوی برای تجزیه‌وتحلیل داده‌های مربوط به زنجیره‌تامین یک فروشگاه زنجیره‌ای با فروش حضوری و آنلاین استفاده شده است و نتایج به‌دست‌آمده نشان می‌دهد که الگوریتم‌های یادگیری ماشین رگرسیون و دسته‌بندی به ترتیب عملکرد فوق‌العاده‌ای در پیش‌بینی میزان فروش هر سفارش و وجود ریسک ارسال با تأخیر سفارش دارند. این موضوع نشان می‌دهد که استفاده از الگوریتم‌های یادگیری ماشین می‌تواند عملکرد بسیار مؤثری در حوزه زنجیره تأمین فروشگاه داشته باشند.

واژگان کلیدی: عارضه‌یابی، زنجیره تأمین، فروشگاه زنجیره‌ای، داده‌کاوی، یادگیری ماشین

فهرست مطالب

فهرست اشکال

فهرست جداول

فهرست نمودارها

۱- مقدمه

تشخیص صحیح و به‌موقع مسائل ریشه‌ای سازمان، اولین گام به‌منظور ایجاد تحول، حفظ بقا و برتری سازمان‌ها است. در بسیاری از مباحث مدیریتی، سازمان به بدن انسان تشبیه می‌شود، چرا که سازمان نیز یک موجود پویا و زنده است. همان‌طوری که پیش‌نیاز انجام هر درمان و بهبود در بدن، انجام آزمایش‌های کاملی از وضعیت بدن انسان است و هرچه دقت این آزمایش‌ها بیشتر و دقیق‌تر باشد بهبودها و فرایند درمان مؤثرتر خواهد بود، در مورد ایجاد اصلاحات و حرکت به‌سوی رشد و پیشرفت سازمان نیز انجام فرایند عارضه‌یابی همین حکم را دارد و با انجام این فرایند سعی در یافتن فرصت‌های بهبودی و تنگناهایی خواهد شد که ممکن است به‌عنوان مانع حرکت سازمان، شناسایی گردد و در جهت بهبود و رفع این موانع برنامه‌ریزی و اقدام مناسب صورت گیرد.

در واقع زنجیره‌تامین یکی از حوزه‌های بسیار وسیع بوده که زیرمجموعه‌های مختلفی را در بردارد. در زنجیره تامین بحث کنترل زمان و هزینه از اهمیت کلیدی برخوردار است. تحویل محموله‌ها و کالا بر اساس زمان‌بندی مشخص شده، جهت حصول اطمینان از اینکه در نهایت محصولات در زمان مناسب به دست مشتری می‌رسند، موردتوجه است. همچنین مدیریت و کنترل هزینه‌های کل شبکه تامین، امری ضروری است، چرا که در صورت عدم کنترل مناسب و اتلاف هزینه‌ها در طول زنجیره، این هزینه‌ها بر روی قیمت تمام‌شده محصول یا خدمت ارائه شده اثر گذاشته و منجر به افزایش آن می‌گردد. درنهایت این افزایش قیمت تمام‌شده ممکن است به مشتری نهایی منتقل شده که مطلوب نیست و می‌تواند منجر به ایجاد نارضایتی و کاهش میزان تقاضا از جانب مشتریان و در نهایت کاهش فروش شرکت گردد. در حالت دیگر در صورت عدم افزایش قیمت محصول، باتوجه‌به اتلاف‌های صورت‌گرفته و افزایش قیمت تمام‌شده، حاشیه سود شرکت کاهش پیدا کرده و این موضوع منجر به نارضایتی سهام‌داران می‌شود. ازاین‌رو دراین‌بین داده‌کاوی می‌تواند به‌عنوان یک ابزار کمکی برای تحلیل بهتر هزینه‌ها و زمان‌ها در زنجیره‌تامین و ارتقای کارایی آن مؤثر واقع شود. برخی از کاربردهای داده‌کاوی در فضای عارضه‌یابی زنجیره‌تامین به‌صورت زیر می‌باشند:

- استفاده از تکنیک‌های داده‌کاوی برای شناسایی الگوهای مربوط تأخیر در تحویل محصولات یا ارائه خدمات
- استفاده از الگوریتم‌های داده‌کاوی به‌منظور کشف علل مربوط به افزایش هزینه‌های واقعی نسبت به هزینه‌های تخمینی
- و...

رویکرد داده‌محور، از مزیت دسترسی به منابع داده جدید برای کمک به شرکت‌ها در عارضه‌یابی زنجیره‌های تأمین استفاده می‌کند. زنجیره‌تامین مدرن به طور فزاینده‌ای برای کسب‌وکارها، پیچیده و جهانی می‌شود. این بدان معناست که شناسایی آسیب‌پذیری‌های بالقوه در زنجیره‌تامین، رفع تنگناها و واکنش سریع به اختلالات زنجیره تأمین، بسیار حیاتی است. زنجیره تأمین برای کسب‌وکارها، تنها یک زنجیره ساده

و خطی از فعالیت‌ها نیست. بلکه یک شبکه پویا از فرایندها، فناوری و افراد یکپارچه است. بسیاری از مدیران، توجه به زنجیره‌تامین را تا زمانی که مشکل بزرگی پیش نیاید در نظر نمی‌گیرند، درحالی‌که سازمان‌ها باید استراتژی‌های منسجمی را توسعه داده و تصمیمات مبتنی بر داده را برای بهبود عملکرد زنجیره‌تامین اتخاذ کنند. نقطه شروع برای بررسی زنجیره‌تامین، تأیید این مسئله است که چگونه عملکرد زنجیره‌تامین فعلی منجر به نتایج مالی و عملکردی بزرگ می‌شود. اتخاذ رویکردی سیستماتیک و ساختاریافته که معیارهای کلیدی عملکرد عملیاتی را با بازگشت سرمایه همسو می‌کند، اساسی است و می‌تواند نشان دهد که چگونه عملیات روزانه کسب‌وکار، منجر به نتایج مالی می‌شود. سپس با بینش مبتنی بر واقعیت، یک برنامه تحول متناسب که سودمندترین اقدامات را مورد هدف قرار می‌دهد، می‌تواند اتخاذ شود. بدون شک زنجیره تأمین داده‌محور می‌تواند برای بسیاری از شرکت‌ها فرصت رشد فراهم کند؛ چرا که اکثر شرکت‌ها، زنجیره‌های تأمین خود را تا زمانی که مشکلی پیش نیاید بررسی نمی‌کنند.

از رو در این تحقیق، یک راه‌حل مبتنی بر داده که از داده‌کاوی استفاده می‌کند؛ پیشنهاد شده است که بر روی داده‌های «یک فروشگاه زنجیره‌ای با فروش حضوری و آنلاین» پیاده‌سازی می‌شود؛ تا بررسی کند که آیا می‌توان مسائل و چالش‌های موجود در زنجیره‌تامین را با داده‌کاوی حل نمود؟ یا خیر. همچنین تلاش می‌شود در ابتدا به بررسی و تجزیه‌وتحلیل اکتشافی داده‌ها پرداخته شود و در ادامه با استفاده از الگوریتم‌های یادگیری ماشین عوامل مؤثر بر فروش و سایر ویژگی‌های تأثیرگذار بر عملکرد زنجیره‌تامین فروشگاه مدل‌سازی شود تا بتوان میزان فروش و همچنین عوامل مؤثر را پیش‌بینی و بررسی نمود.

۲- پیشینه تحقیق

۲-۱- مرور ادبیات

این مقاله به بررسی زنجیره تأمین خدمات آنلاین مبتنی بر داده از طریق دو دیدگاه دامنه و تأمین تقاضا می‌پردازد؛ همچنین عوامل مختلفی را که بر تقاضای خدمات آنلاین تأثیر می‌گذارند و راهبردهایی را که برای برآورده این تقاضا می‌توان در زنجیره تأمین به کار گرفت، مورد بررسی قرار داده و ویژگی‌های خدمات آنلاین و تأثیر آن‌ها بر زنجیره تأمین مانند سفارشی‌سازی خدمات و کنترل کیفیت خدمات را، تحلیل می‌کند؛ همچنین نقش داده‌ها در مدیریت زنجیره تأمین شامل جمع‌آوری، تجزیه‌وتحلیل و استفاده از داده‌ها، بررسی می‌شود. مقاله نشان می‌دهد که داده‌ها می‌توانند برای امور مختلفی چون بهینه‌سازی زنجیره تأمین و بهبود کیفیت خدمات به کار گرفته شوند. نویسندگان چالش‌ها و فرصت‌های زنجیره تأمین خدمات آنلاین مبتنی بر داده را نیز مورد بررسی قرار داده‌اند و همچنین، یک چارچوب برای مدیریت زنجیره تأمین خدمات آنلاین ارائه می‌دهند که شامل شناسایی عوامل تقاضا، توسعه راهبردهای تأمین تقاضا در زنجیره تأمین و بهره‌گیری از داده‌ها برای بهینه‌سازی زنجیره تأمین هستند [۱] .

این مقاله یک رویکرد مبتنی بر داده برای همگام‌سازی تطبیقی تقاضا و عرضه در زنجیره‌های عرضه خرده‌فروشی omni-channel را پیشنهاد می‌کند. نویسندگان چالش‌هایی را که خرده‌فروش‌های omni-channel در برآوردن تقاضای مشتری با آن مواجه هستند موردبحث قرار می‌دهند و استدلال می‌کنند که رویکردهای داده‌محور می‌توانند برای غلبه بر این چالش‌ها و بهبود عملکرد زنجیره تأمین مورداستفاده قرار گیرند. این مقاله چارچوبی برای همگام‌سازی تطبیقی ارائه می‌کند که شامل جمع‌آوری، تجزیه‌وتحلیل و تصمیم‌گیری در زمان واقعی است. نویسندگان استفاده از تکنیک‌های یادگیری ماشین را برای پیش‌بینی تقاضا و بهینه‌سازی تخصیص موجودی پیشنهاد می‌کنند. این مقاله شامل یک مطالعه موردی از یک خرده‌فروش برزیلی است که رویکرد پیشنهادی را اجرا کرد که منجر به بهبود دید زنجیره تأمین، کاهش انبارها و افزایش فروش شد. نویسندگان همچنین محدودیت‌ها و جهت‌گیری‌های تحقیقاتی آینده این رویکرد را موردبحث قرار می‌دهند [۲] .

این مقاله اندازه‌گیری، کاهش و پیشگیری از ضایعات مواد غذایی در زنجیره تأمین در خرید آنلاین را موردبحث قرار می‌دهد. نویسندگان چالش‌های ضایعات مواد غذایی را در زنجیره‌های تأمین، از جمله اثرات اقتصادی، زیست‌محیطی و اجتماعی را برجسته می‌کنند و استدلال می‌کنند که خرید آنلاین می‌تواند به‌عنوان ابزاری برای کاهش ضایعات مواد غذایی با ارائه پیش‌بینی دقیق‌تر تقاضا، کاهش بیش از حد سفارش، و امکان مدیریت بهتر موجودی استفاده شود. این مقاله چارچوبی برای اندازه‌گیری ضایعات مواد غذایی در زنجیره‌های تأمین ارائه می‌کند که شامل کمی‌کردن مقدار ضایعات، شناسایی علل آن و ارزیابی اثربخشی استراتژی‌های کاهش است. نویسندگان استفاده از تجزیه‌وتحلیل داده‌ها و یادگیری ماشین را برای بهبود پیش‌بینی تقاضا و بهینه‌سازی مدیریت موجودی پیشنهاد می‌کنند. همچنین این پژوهش، شامل یک مطالعه

موردی از یک خواربارفروشی آنلاین است که رویکرد پیشنهادی را پیاده‌سازی کرده و منجر به کاهش ضایعات مواد غذایی و بهبود کارایی زنجیره تأمین شده است [۳] .

این مقاله در مورداستفاده از تکنیک‌های یادگیری ماشین برای پیش‌بینی فروش در فروشگاه‌های خرده‌فروشی صحبت می‌کند و اشاره می‌کند که پیش‌بینی دقیق فروش برای مدیریت موجودی، برنامه‌ریزی عملیاتی و بهینه‌سازی درآمد ضروری است. آنها استفاده از الگوریتم‌های یادگیری ماشین مانند Random Forest، Support Vector Machine و Artificial Neural Network را برای پیش‌بینی فروش بر اساس داده‌های تاریخی و متغیرهای دیگر مانند تخفیف‌ها، تعطیلات و هواشناسی پیشنهاد می‌دهند. نویسندگان مزایا و محدودیت هر تکنیک را بررسی کرده و عملکرد آن‌ها را با استفاده از معیارهایی مانند میانگین خطای درصدی مطلق مقایسه می‌کنند. آنها همچنین اهمیت انتخاب ویژگی و پیش‌پردازش داده را در بهبود دقت پیش‌بینی موردبحث قرار می‌دهند. مقاله شامل یک مطالعه موردی از یک فروشگاه در هند است که رویکرد پیشنهادی پیاده‌سازی شده است و منجر به بهبود دقت پیش‌بینی فروش و مدیریت بهتر موجودی شده است [٤] .

این مقاله به بررسی پیش‌بینی فروش خرده‌فروشان در بازارهای مد می‌پردازد و بیان می‌کند که پیش‌بینی دقیق فروش برای مدیریت مؤثر موجودی، برنامه‌ریزی تولید و رضایت مشتری حیاتی است. آنها چالش‌ها و فرصت‌های پیش‌بینی فروش خرده‌فروشان صنعت مد، از جمله چرخه‌های عمر کوتاه محصول، تغییرات سریع در اولویت‌های مصرف‌کننده، و نیاز به تجزیه‌وتحلیل داده‌های بلادرنگ را موردبحث قرار می‌دهند. این مقاله مروری بر تکنیک‌های مختلف پیش‌بینی مورداستفاده در خرده‌فروشی مد، مانند روش‌های آماری، هوش مصنوعی و یادگیری ماشینی ارائه می‌کند و مزایا و محدودیت‌های هر روش را موردبحث قرار داده و عملکرد آنها را با استفاده از معیارهایی مانند دقت پیش‌بینی و میانگین درصد مطلق خطا مقایسه می‌کند. آنها همچنین در مورد اهمیت کیفیت داده‌ها و پیش‌پردازش داده‌ها در بهبود دقت پیش‌بینی بحث می‌کنند. نویسندگان به این نتیجه رسیدند که هیچ راه‌حلی برای پیش‌بینی فروش خرده‌فروشی مد وجود ندارد و ممکن است ترکیبی از تکنیک‌های مختلف برای دستیابی به پیش‌بینی‌های دقیق و به‌موقع موردنیاز باشد . [٥]

این مقاله یک مطالعه موردی از مدیریت اختلالات در زنجیره تامین خرده‌فروشی را ارائه می‌کند. نویسندگان اشاره می‌کنند که اختلالات زنجیره تامین می‌تواند منجر به ضررهای مالی قابل‌توجه و آسیب به شهرت برند شود. آنها چالش‌های موجود در مدیریت اختلالات، از جمله نیاز به اطلاعات به‌موقع و دقیق، ارتباطات مؤثر و برنامه‌ریزی اضطراری را موردبحث قرار می‌دهند. این مقاله یک مطالعه موردی از یک زنجیره تامین خرده‌فروشی را توصیف می‌کند که در آن یک تأمین‌کننده اصلی آتش‌سوزی در تأسیسات تولیدی خود را تجربه کرد که باعث اختلال در عرضه یک دسته محصول حیاتی شد. نویسندگان گام‌های برداشته‌شده توسط شرکت خرده‌فروشی برای مدیریت اختلال را موردبحث قرار می‌دهند. این مقاله همچنین استفاده از مدل‌های شبیه‌سازی زنجیره تامین را برای ارزیابی سناریوهای مختلف و ارزیابی تأثیر اختلالات احتمالی

موردبحث قرار می‌دهد. نویسندگان بر اهمیت مدیریت ریسک فعال و انعطاف‌پذیری زنجیره تأمین در به‌حداقل‌رساندن تأثیر اختلالات تأکید می‌کنند [٦] .

این مقاله کاربرد یادگیری ماشین و مدل‌های ترکیبی را در پیش‌بینی فروش خرده‌فروشی بررسی می‌کند. نویسندگان اشاره می‌کنند که پیش‌بینی دقیق فروش برای خرده‌فروشان برای تصمیم‌گیری آگاهانه در مورد مدیریت موجودی و برنامه‌ریزی تولید بسیار مهم است. این مقاله مروری بر مدل‌های مختلف یادگیری ماشین، از جمله شبکه‌های عصبی، درخت‌های تصمیم‌گیری و رگرسیون برداری پشتیبان و مزایا و محدودیت‌های آن‌ها در پیش‌بینی فروش خرده‌فروشی ارائه می‌کند. نویسندگان همچنین یک مدل ترکیبی پیشنهاد می‌کنند که نقاط قوت مدل‌های مختلف را برای بهبود دقت پیش‌بینی ترکیب می‌کند و در مورد اهمیت کیفیت داده‌ها و پیش‌پردازش در آموزش مدل یادگیری ماشین بحث می‌کنند و بر نیاز به نظارت مداوم و به‌روزرسانی مدل برای درنظرگرفتن تغییرات در رفتار مصرف‌کننده و روند بازار تأکید می‌کنند. این مقاله نتیجه‌گیری می‌کند که یادگیری ماشین و مدل‌های ترکیبی پتانسیل بالایی در پیش‌بینی فروش خرده‌فروشی دارند و تحقیقات بیشتری برای کشف قابلیت‌های کامل آنها در این زمینه موردنیاز است . [٧]

در این مقاله یک مدل بهینه‌سازی برای تخصیص منابع ارائه داده می‌شود و سپس یک الگوریتم مورچه[1] بهبودیافته برای حل آن ایجاد خواهد شد. الگوریتم مورچه دارای مزایای همگرایی سریع به راه‌حل بهینه است و الگوریتم مورچه بهبودیافته نیز دارای مزایای آشکاری در تعادل انعطاف‌پذیری بهینه‌سازی‌های چندهدفه، تنظیم سرعت همگرایی و تنظیم پارامترهای عملیات می‌باشد.

در ادامه، اثربخشی و امکان‌سنجی روش و الگوریتم بهینه‌سازی با یک شبیه‌سازی عددی نشان داده می‌شود. مدل بهینه‌سازی نه‌تنها هزینه خدمات و زمان تحویل موردنیاز را در تابع هدف منعکس می‌کند، بلکه بهینه‌سازی اثر اندازه خدمات و روابط مزایا و خطرات یکپارچه‌سازی را نیز در نظر می‌گیرد و نشان می‌دهد موفقیت خدمات خرید آنلاین مستلزم دو رابطه ضروری است؛ ١- الگوهای مختلف خدمات سفارشی باید با طرح‌های مختلف تخصیص منابع زنجیره تامین مطابقت داشته باشد و ٢- الگوهای مختلف ترکیب خدمات سفارشی شده نیز باید با طرح‌های تخصیص منابع زنجیره‌تامین مختلف مطابقت داشته باشند [٨] .

شرکت‌های خرید آنلاین B٢C باید یکپارچه‌سازی منابع زنجیره‌تامین[2] خود را برای ارائه ظرفیت‌های خدمات، بهبود تجربیات مشتریان خود و معرفی خدمات سفارشی رضایت‌بخش بهینه کنند. این مطالعه با تجزیه‌وتحلیل ویژگی‌ها و حالت‌های خدمات SCRI درB٢C، تعادل پویا بین ظرفیت‌های خدمات عرضه و تقاضا را در یک شرکت خرید آنلاین موردبحث قرار می‌دهد. بحث در مورد این ظرفیت‌ها نه‌تنها باید اهداف

[1] Ant algorithm
[2] SCRI: supply chain resources integration

بهینه‌سازی سنتی را در نظر بگیرد، بلکه باید مناسب‌بودن منابع را از زوایای خدمات عمومی، خدمات اضطراری و ظرفیت‌های بالقوه استراتژیک برای دستیابی به یک SCRI مؤثر را ارزیابی کند؛ بنابراین، این مطالعه ثبات جهت‌گیری هدف ظرفیت منبع را به‌عنوان یک هدف مهم بهینه‌سازی در نظر گرفته و با شناسایی عوامل مشخصه ظرفیت و معرفی آن‌ها به SCRI، این ثبات را ارزیابی می‌کند. بدین ترتیب یک مدل بهینه‌سازی و یک الگوریتم مورچه بهبودیافته برای حل فرایند SCRI پیشنهاد شدند که امکان‌سنجی و اعتبار این مدل و الگوریتم با نشان‌دادن کاربرد آن‌ها و همچنین نشان‌دادن ارزش تحقیق آن‌ها تأیید شدند [۹].

این مقاله، به پیش‌بینی فروش با استفاده از مدل‌های رگرسیون لجستیک و K- نزدیک‌ترین همسایه[۳] بر اساس یادگیری ماشین می‌پردازد. اطلاعات از منابع داده‌های بین‌المللی جمع‌آوری شده است. پیش‌بینی فروش در توسعه کسب‌وکار، برآورد نیاز مالی، صحت تصمیمات مدیریتی، همکاری و هماهنگی، کنترل و غیره کاربرد دارد و این نشان از اهمیت پیش‌بینی فروش در کسب‌وکار دارد. پس از ایجاد و تست مدل‌ها بر روی داده‌ها، از ماتریس سردرگمی و روش اقلیدسی نتایج مورد ارزیابی قرار گرفتند. نتایج نشان می‌دهند که دو مدل رگرسیون لجستیک و K- نزدیک‌ترین همسایه می‌توانند عملکرد خوبی در پیش‌بینی فروش به‌ویژه در بازارهای محصولات FMCG که تقاضای بالایی دارند و به‌سرعت فروخته می‌شوند، داشته باشند [۱۰].

برای شرکت‌های خرده‌فروشی مدرن که زنجیره عظیمی از کسب‌وکارها را اداره می‌کنند، پیش‌بینی دقیق فروش، کلید توسعه شرکت‌ها و حتی موفقیت یا شکست آن‌ها است. پیش‌بینی فروش به شرکت‌ها این امکان را می‌دهد تا به طور مؤثر منابعی از جمله جریان نقدی، تولید و برنامه کسب‌وکار را تخصیص دهند. در این مقاله، یک مدل پیش‌بینی فروش کارآمد و دقیق با استفاده از یادگیری ماشین پیشنهاد می‌شود. در ابتدا، مهندسی ویژگی برای استخراج ویژگی‌ها از داده‌های فروش تاریخی انجام می‌شود. علاوه بر این، از XGBoost که یک مدل درختی تقویت‌کننده گرادیان توزیع‌شده بهینه‌سازی‌شده می‌باشد، برای پیش‌بینی میزان فروش آینده استفاده شده است. نتایج از طریق معیار خطای جذر میانگین مربعات[۴] که نوعی از معیار میانگین خطای مقیاس مطلق است، ارزیابی شده است. نتایج آزمایش بر روی مجموعه‌داده کالاهای خرده‌فروشی عمومی والمارت[۵] نشان می‌دهد که مدل پیشنهادی برای پیش‌بینی فروش با زمان محاسباتی و منابع حافظه کمتر بسیار خوب عمل می‌کند [۱۱].

پیش‌بینی فروش چالش‌برانگیزترین کار برای مدیریت موجودی، بازاریابی، خدمات مشتری و برنامه‌ریزی مالی کسب‌وکار صنعت خرده‌فروشی است. در این مقاله، یک تجزیه‌وتحلیل پیش‌بینی‌کننده از خرده‌فروشی مجموعه‌داده Citadel POS با استفاده از تکنیک‌های مختلف یادگیری ماشین انجام شده است. در این

---

[۳] K-Nearest Neighbour
[۴] RMSE: Root Mean Squared Error
[۵] Walmart

مطالعه، مدل‌های رگرسیون مختلف (رگرسیون خطی، رگرسیون جنگل تصادفی، رگرسیون افزایش گرادیان) و مدل‌های سری زمانی (ARIMA و LSTM)، برای پیش‌بینی فروش پیاده‌سازی شده است و تجزیه‌وتحلیل و ارزیابی پیش‌بینی دقیق ارائه گردیده است. مجموعه‌داده مورداستفاده در این تحقیق از Citadel POS از سال ۲۰۱۳ تا ۲۰۱۸ به‌دست‌آمده است؛ نتایج نشان می‌دهد که XGboost از سری‌های زمانی و سایر مدل‌های رگرسیون بهتر عمل کرده و بهترین عملکرد را با میانگین خطای مطلق[۶] ۰.۵۱۶ و خطای ریشه میانگین مربع[۷] ۰.۶۳ به دست آورده است [۱۲] .

رویدادهای اخیر، همانند همه‌گیری مداوم، عدم‌قطعیت زنجیره تامین را افزایش داده است. این عوامل ذاتی (مثل کمبود کانتینر) و عوامل خارجی (مثل افزایش تقاضا) عدم‌قطعیت‌ها در مدیریت زنجیره تامین را تشدید کرده و اهمیت BDA[۸] را در[۹] SCMبیشتر کرده است. براین‌اساس، این مطالعه مروری سیستماتیک از مطالعات موجود در BDA را انجام می‌دهد. در حال حاضر، همراه با توسعه اخیر در زیرساخت‌های یادگیری ماشین و محاسبات، BDA در زنجیره‌تامین اهمیت زیادی پیدا کرده است. این مقاله چارچوبی از یک بررسی نظام‌مند ادبیات را از دیدگاه‌های بین‌رشته‌ای ارائه می‌کند. از دیدگاه سازمانی، این مطالعه به بررسی مبانی نظری و مدل‌های تحقیقاتی می‌پردازد که پایداری و عملکرد به‌دست‌آمده از طریق استفاده از تجزیه‌وتحلیل داده‌های بزرگ را توضیح می‌دهند؛ سپس، از دیدگاه فنی، این پژوهش انواع تجزیه‌وتحلیل داده‌های بزرگ، تکنیک‌ها، الگوریتم‌ها و ویژگی‌های توسعه‌یافته برای توابع زنجیره تامین افزایش‌یافته را تحلیل می‌کند و در نهایت، شکاف تحقیق را شناسایی کرده و جهت‌گیری‌های تحقیقاتی آتی را پیشنهاد می‌کند [۱۳] .

این تحقیق تلاش می‌کند تا عوامل مؤثر بر عملکرد زنجیره تأمین را برای تولیدکنندگان کوچک تحت سناریوی خرید الکترونیکی شناسایی کند. داده‌های اولیه مانند تقاضای مشتری از طریق توابع کنترل شده با ورودی تولید شده است و مفروضات بر اساس تجربه محقق تنظیم شده است. ادبیات موجود نشان می‌دهد که زمان کوتاه‌تر منجر به عملکرد بهتر زنجیره تامین می‌شود. بااین‌حال، از طریق شبیه‌سازی پیشنهاد می‌شود که زمان طولانی‌تر، سطح خدمات را با افزایش موجودی ایمنی و سطح موجودی عمومی بهبود می‌بخشد. نتایج نشان‌دهنده وجود اثر شلاقی در زنجیره‌تامین خرید الکترونیکی به دلیل سفارش‌های عقب‌افتاده و تصمیمات مدیریتی است. مقایسه بین دو استراتژی کنترل موجودی مختلف با الگوهای تکمیل نشان می‌دهد که روش سطح سهام هدف در سناریوی خرید الکترونیکی نسبت به روش کمیت سفارش اقتصادی بهتر عمل می‌کند [۱۴] .

در این مطالعه، یک مدل پیش‌بینی مبتنی بر خوشه‌بندی با ترکیب روش‌های خوشه‌بندی و یادگیری ماشین برای پیش‌بینی فروش خرده‌فروشی رایانه‌ای پیشنهاد شده است. ابتدا از تکنیک خوشه‌بندی برای

---

[۶] MAE: Mean Absolute Error
[۷] RMSE: Root Mean Square Error
[۸] Big Data Analytics
[۹] Supply Chain Management

خوشه‌بندی داده‌ها با ویژگی‌ها یا الگوهای مشابه در یک گروه استفاده شده است. همچنین، از تکنیک‌های یادگیری ماشین برای آموزش مدل پیش‌بینی هر گروه استفاده شده است. در مجموع شش مدل پیش‌بینی مبتنی بر خوشه‌بندی پیشنهاد شده است. داده‌های فروش واقعی برای رایانه‌های شخصی، رایانه‌های نوت‌بوک و نمایشگرهای کریستال مایع به‌عنوان‌مثال‌های تجربی استفاده شده‌اند. لازم به ذکر است که دقت پیش‌بینی را می‌توان با استفاده از مدل پیش‌بینی مبتنی بر خوشه‌بندی پیشنهادی افزایش داد. نتایج تجربی نشان می‌دهد که مدل ترکیبی از GHSOM و ELM عملکرد پیش‌بینی برتری را برای هر سه محصول در مقایسه با سایر روش‌ها داشته‌اند. این می‌تواند به طور مؤثر به‌عنوان یک مدل پیش‌بینی فروش مبتنی بر خوشه‌بندی برای خرده‌فروشی کامپیوتر استفاده شود [۱۵] .

ازآنجایی‌که صنعت مد به‌راحتی قابل‌پیش‌بینی نیست، پیش‌بینی دقیق فروش نیز ساده نیست. در این مطالعه، از روش‌های رگرسیون در یادگیری ماشین و تکنیک‌های تحلیل سری‌های زمانی برای پیش‌بینی میزان فروش بر اساس چندین ویژگی استفاده شده است. مدل‌ها بر روی داده‌های فروش والمارت در پلتفرم Microsoft Azure Machine Learning Studio اعمال شده‌اند. تکنیک‌های رگرسیون شامل رگرسیون خطی، رگرسیون بیزی، رگرسیون شبکه عصبی، رگرسیون جنگل تصادفی و رگرسیون درخت تصمیم تقویت شده بکار گرفته شده‌اند. علاوه بر این تکنیک‌های رگرسیون، روش‌های تحلیل سری‌های زمانی شامل ARIMA فصلی، ARIMA غیر فصلی، ETS فصلی، ETS غیر فصلی، روش متوسط و روش دریفت اجرا شده‌اند. نتایج می‌دهد که رگرسیون درخت تصمیم تقویت شده با ضریب تعیین ۰.۹۷ بهترین عملکرد را در این داده‌های فروش ارائه می‌دهد و عملکرد بهتری را در مقایسه با سایر روش‌ها برای پیش‌بینی فروش دارد [۱٦] .

کلید موفقیت در تجارت، امروزه کنترل زنجیره تامین خرده‌فروشی است. پیش‌بینی تقاضای مشتری برای مدیریت زنجیره‌تامین بسیار ضروری است. این مقاله، یک روش جدید با استفاده از یادگیری ماشین ارائه می‌کند که به پیش‌بینی دقیق تقاضا کمک می‌کند. در این مطالعه، داده‌های گذشته یک فروشگاه را جمع‌آوری شده است. از الگوریتم‌های K-Nearest Neighbor، Support Machine Vector، Gaussian Nave Bayes، Random Forest، Decision Tree Classifier و رگرسیون‌ها در این مطالعه مقایسه شده‌اند. این مقاله، از ترکیب موقعیت فروشگاه، ماه و مناسبت آن ماه و سایر داده‌های مرتبط استفاده کرده است. مدل ارائه شده تقاضای آزمایشی برای یک محصول خاص را پیش‌بینی می‌کند. پس از ایجاد یک مجموعه‌داده و اعمال الگوریتم‌ها، نتایج و دقت الگوریتم‌های مختلف مقایسه شده‌اند. نتایج نشان می‌دهد که Gaussian Nave Bayes بهترین دقت را در بین الگوریتم‌ها دارد [۱۷] .

این مقاله، یک چارچوب مفهومی برای درک نقشی که هوش مصنوعی می‌تواند در زنجیره ارزش خرده‌فروشی بازی کند، معرفی می‌کند. هدف اصلی این مقاله، درک بهتر یک زنجیره ارزش خرده‌فروشی مبتنی بر هوش مصنوعی است. به‌عنوان نقطه شروع، یک مرور مختصر از زنجیره ارزش سنتی خرده‌فروشی و فعالیت‌ها، ذی‌نفعان و فناوری درگیر در هر مرحله ارائه داده می‌شود. سپس ضعف‌های حاضر صنعت خرده‌فروشی توضیح داده می‌شود و به دنبال آن تمرکز ویژه‌ای بر نقشی که هوش مصنوعی در رفع اختلال

در این صنعت ایفا کرده است، مورد بررسی قرار می‌گیرد. سپس، فناوری‌های مختلف هوش مصنوعی مبتنی بر گارتنر را برای هر مرحله در زنجیره ارزش ترسیم می‌شود و نشان داده می‌شود که برخی از سرمایه‌گذاری‌های فناوری هوش مصنوعی می‌توانند اهداف متعددی را در زنجیره ارزش انجام دهند **[۱۸] .**

در این مطالعه، یک تابع تقاضای زنجیره تامین خرید آنلاین بر اساس درجه همجوشی آنلاین به آفلاین با درنظرگرفتن آگاهی کم‌کربن خریداران آنلاین، باتوجه‌به تصمیم‌گیری کم‌کربن زنجیره تامین، در ترکیب با پس‌زمینه عملیات آنلاین به آفلاین پیشنهاد شده است. براین‌اساس، یک مدل تصمیم‌گیری کم‌کربن مبتنی بر فروشگاه‌های آنلاین و تأمین‌کنندگان آن‌ها ایجاد شده و کاربرد مدل مورد تجزیه‌وتحلیل قرار می‌گیرد. نتایج نشان می‌دهد که ارتقای عملکرد کم‌کربن زمانی کارآمد است که ثبات عملیات آنلاین و آفلاین فروشگاه‌های آنلاین و تأمین‌کنندگان آن‌ها بالا باشد. علاوه بر این، تصمیم‌گیری با درنظرگرفتن هزینه انتشار اطلاعات کربن برای بهبود آگاهی خریداران آنلاین نسبت به کربن پایین بسیار مفید است. اگر سطح کربن پایین کالاهای خرید آنلاین کمتر از استاندارد کم‌کربن باشد، براین‌اساس، عملکرد کم‌کربن زنجیره تامین خرید آنلاین ترویج خواهد شد [۱۹] .

این مقاله به بررسی اهمیت لجستیک در تجارت الکترونیک و نقش آن در مدیریت زنجیره تأمین پرداخته است. همچنین بر روی نیاز به عملیات لجستیکی کارآمد و مؤثر برای اطمینان از رضایت مشتریان، بهبود فروش و حفظ مزیت رقابتی تأکید می‌کند. علاوه‌برآن چالش‌هایی که صاحبان کسب‌وکار در لجستیک تجارت الکترونیک و مدیریت زنجیره تأمین با آن روبرو هستند. نویسندگان این مقاله علاوه بر آنکه چندین راهبرد برای بهینه‌سازی لجستیک تجارت الکترونیک از جمله خودکارسازی، همکاری و سفارشی‌سازی ارائه می‌دهد؛ بر اهمیت تجزیه‌وتحلیل داده‌ها، ادغام فناوری و شیوه‌های پایدار در بهبود لجستیک و مدیریت زنجیره تأمین تأکید می‌کنند و بر تأثیر لجستیک تجارت الکترونیک بر ذی‌نفعان مختلف، بهبود مداوم، نوآوری و سازگاری با تقاضاهای تغییرپذیر لجستیک تجارت الکترونیک و مدیریت زنجیره تأمین تأکید می‌کند. در آخر؛ مقاله یک دیدگاه جامع از لجستیک تجارت الکترونیک در مدیریت زنجیره تأمین، چالش‌های آن و راهکارهای پتانسیل آن را ارائه می‌دهد [۲۰] .

این مقاله به بررسی رابطهٔ بین کیفیت زنجیره تأمین الکترونیکی و رضایت مشتری در خرید آنلاین می‌پردازد. نویسندگان یک مطالعه موردی بر روی یک وب‌سایت خرید آنلاین را انجام دادند و داده‌ها را با استفاده از پرسش‌نامه نظرسنجی جمع‌آوری کردند. آنها داده‌ها را با استفاده از روش‌های آماری تجزیه‌وتحلیل کردند و دریافتند که کیفیت زنجیره تأمین الکترونیکی تأثیر مثبت قابل‌توجهی بر رضایت مشتری دارد. این مطالعه همچنین چندین عامل کلیدی که بر کیفیت زنجیره تأمین الکترونیکی و رضایت مشتری در خرید آنلاین تأثیر می‌گذارد، مانند طراحی وب‌سایت، دردسترس‌بودن محصول، زمان تحویل، و خدمات مشتری را شناسایی می‌کند. نویسندگان توصیه‌هایی از جمله بهبود عملکرد وب‌سایت، بهینه‌سازی تدارکات و حمل‌ونقل و ارائه خدمات شخصی‌سازی‌شده را برای خرده‌فروشان آنلاین ارائه می‌کنند تا کیفیت زنجیره تأمین

الکترونیکی و رضایت مشتری را بهبود ببخشند و بر اهمیت کیفیت زنجیره تأمین الکترونیکی در موفقیت خرده‌فروشان آنلاین و نیاز به بهبود مستمر برای برآورده‌کردن انتظارات مشتری تأکید می‌کند [۲۱] .

این مقاله یک تحلیل پیش‌بینی فروش خرده‌فروشی با استفاده از تکنیک‌های یادگیری ماشینی ارائه می‌دهد. این مطالعه از داده‌های فروش تاریخی فروشگاهی در پاکستان برای توسعه مدل پیش‌بینی استفاده کرده است. نویسندگان پنج الگوریتم یادگیری ماشینی مختلف را به کار برده و عملکرد آن‌ها را مقایسه کردند. الگوریتم‌های استفاده شده عبارت‌اند از: Multiple Linear Regression، Decision Tree، Random Forest، Gradient Boosting، Artificial Neural Network

از اندازه‌گیری‌های آماری مانند خطای مطلق میانگین درصد[۱۰]و خطای میانگین مربعات ریشه[۱۱] برای ارزیابی عملکرد مدل‌ها استفاده کردند. نتایج نشان داد که الگوریتم افزایش گرادیان با کمترین مقدار MAPE و RMSE بهترین عملکرد را داشت. نویسندگان معتقدند که مدل پیشنهادی می‌تواند توسط خرده‌فروشان برای پیش‌بینی فروش محصولات و تصمیم‌گیری‌های اطلاعاتی در رابطه مدیریت موجودی و استراتژی فروش استفاده شود [۲۲] .

این مقاله یک چارچوب جدید برای پیش‌بینی فروش بر اساس مدل Convolutional Long Short-Term Memory ارائه می‌دهد. نویسندگان از مجموعه‌داده فروش روزانه یک فروشگاه زنجیره‌ای در چین برای آموزش و ارزیابی مدل استفاده کرده‌اند. آنها عملکرد این مدل را با مدل‌های سنتی پیش‌بینی سری زمانی مانند مدل ARIMA و ETS مقایسه کرده‌اند. نتایج نشان داد که مدل ConvLSTM در اصطلاح دقت و استحکام عملکرد مدل‌های سنتی را برتری داشته است. نویسندگان معتقدند که چارچوب پیشنهادی توسط خرده‌فروشان برای تصمیم‌گیری درباره مدیریت موجودی و استراتژی‌های فروش قابل‌استفاده است. این مطالعه اهمیت درنظرگرفتن عوامل خارجی مانند تعطیلات و رویدادهای تبلیغاتی در پیش‌بینی فروش را نیز بیان می‌کند. نویسندگان بیشتر می‌گویند که چارچوب می‌تواند به صنایع دیگر خارج از خرده‌فروشی مانند مالی و حمل‌ونقل توسعه یابد. به‌طورکلی، این مطالعه در مورد کاربرد تکنیک‌های یادگیری ماشین برای پیش‌بینی فروش در صنعت خرده‌فروشی ارزشمند است [۲۳] .

این مقاله، مدل‌های پژوهشی‌ای ارائه داده است که می‌توانند با برخی تغییرات، برای اندازه‌گیری کارایی زنجیره تامین خرده‌فروشی مورداستفاده قرار گیرند. مدل‌ها بر اساس رویکرد تحلیل پوششی داده‌ها DEA بودند. چهار گروه اصلی شناسایی شدند: مدل‌های استاندارد DEA، مدل‌های تجزیه کارایی، مدل‌های شبکه و مدل‌های مبتنی بر نظریه بازی. در بخش دوم مقاله، رویکردهای مختلف بر روی یک نمونه واقعی یک شرکت تجاری فعال در صربستان آزمایش شد. هفت زنجیره تامین مشاهده شد که هرکدام از یک مرکز

---

[۱۰] MAPE
[۱۱] RMSE

توزیع DC و فروشگاه خرده‌فروشی RC تشکیل شده است. متغیرهای مورداستفاده عبارت بودند از تعداد مکان‌های پالت، هزینه‌های لجستیک، تعداد تحویل، دقت تحویل و گردش مالی، نتایج مزایا و معایب رویکردهای مختلف را در مثال واقعی نشان داد. سهم اصلی این مقاله در رویکردهای منحصربه‌فرد برای اندازه‌گیری کارایی زنجیره تامین خرده‌فروشی نهفته است [۲٤] .

۲-۲- جدول مرور ادبیات

جدول (۱-۲) جدول مرور ادبیات

ردیف	عنوان مقاله	سال انتشار	نام مجله یا کنفرانس	نویسندگان	روش مورداستفاده	مطالعه موردی	یادگیری ماشین	الگوریتم											روش‌های ارزیابی			
								دسته‌بندی							خوشه‌بندی	سری‌های زمانی		سایر	Confusion Matrix	RMSSE	MAE	RMSE	سایر
								linear regression	K-Nearest Neighbour	XGBoost	Random Forest	Decision Tree	SVM	سایر	K-Means	سایر	ARIMA	سایر					
۱	Data-driven online service supply chain: a demand-side and supplyside perspective	۲۰۲۰	Journal of Enterprise Information Management	LeiLi et al.	بررسی زنجیره تامین از طریق دو دیدگاه دامنه تقاضا و تامین تقاضا																		ارائه چارچوب برای مدیریت خدمات آنلاین
۲	A data-driven approach to adaptive synchronization of demand and supply in omni-channel retail supply chains	۲۰۲۰	International Journal of Information Management	Marina Meireles Pereira et al.	همگام‌سازی تطبیقی داده و تحلیل داده بلادرنگ	*	*																ارائه چارچوب برای همگام‌سازی تطبیقی

۳	Measurement, mitigation and prevention of food waste in supply chains: An online shopping perspective	۲۰۱۹	Industrial Marketing Management	Vasco Sanchez Rodrigues ,et al.	یادگیری ماشین و تجزیه‌وتحلیل داده‌ها و کمی‌کردن داده‌های کیفی	*	*																							الگوریتم طراحی شده به‌منظور شناسایی اتلاف مواد غذایی									ارائه روش‌ها برای بهبود پیش‌بینی تقاضا و بهینه‌سازی
۴	Sales-forecasting of Retail Stores using Machine Learning Techniques	۲۰۱۸	International Conference on Computational Systems and Information Technology for Sustainable Solutions	Akshay Krishna et al.	پیش‌بینی فروش از طریق الگوریتم‌های یادگیری ماشین	*	*								*												*							مقایسه عملکرد روش‌های پیش‌بینی با استفاده از معیارها
۵	A survey on retail sales forecasting and prediction in fashion markets	۲۰۱۵	Systems Science & Control Engineering	Samaneh Beheshti-Kashi et al.	پیش‌بینی بر اساس روش‌های مختلف		*																					روش‌های آماری، هوش مصنوعی									مقایسه عملکرد راه‌حل‌های موجود

۶	Managing disruptions in supply chains: A case study of a retail supply chain	۲۰۱۹	Intnational Journal of Production Economics	Adegoke Oke , Mohan Gopalakrishnan.	استفاده از مدل‌های شبیه‌سازی برای ارزیابی سناریوهای مختلف	*																		ارزیابی سناریوهای مختلف و ارزیابی تأثیر اختلالات احتمالی
۷	Application of Machine Learning Model and Hybrid Model in Retail Sales Forecast	۲۰۲۱	Intnational Journal of Production Economics	Haichen Jiang, Jiatong Ruan, Jianmin Sun	پیش‌بینی بر اساس روش‌های مختلف		*	*				*		جمله شبکه‌های عصبی، درخت‌های تصمیم‌گیری و رگرسیون برداری										ارائه مدل ترکیبی
۸	Optimization Analysis of Supply Chain Resource Allocation in Customized Online Shopping Service Mode	۲۰۱۵	Mathematical Problems in Engineering	Yao, Jianming. Gu, Mengjie	مدل بهینه‌سازی مبتنی بر الگوریتم مورچه بهبودیافته														Ant algorithm					شبیه‌سازی عددی
۹	Supply chain resources integration optimisation in B۲C online shopping	۲۰۱۷	International Journal of Production Research	Yao, JM	یکپارچه‌سازی منابع زنجیره تامین بر اساس یک مدل بهینه‌سازی و یک الگوریتم مورچه بهبودیافته														Ant algorithm					ارائه کاربرد

۱۰	Sales Forecasting using Linear Regression and K-Nearest Neighbour	۲۰۱۹	International Journal of Innovative Research in Science , Engineering and Technology	Neha Sehgal , Deepika Garg	پیش‌بینی فروش به کمک مدل‌های یادگیری ماشین	*	*	*		*																		*					روش اقلیدسی
۱۱	Machine learning model for sales forecasting by using XGBoost	۲۰۲۱	۲۰۲۱IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)	Shilong, Zhang	پیش‌بینی فروش به کمک مدل‌های یادگیری ماشین	*	*			*																		*					
۱۲	A Predictive Analysis of Retail Sales Forecasting using Machine Learning Techniques	۲۰۲۱	Computer Science and Information Technology	Sajawal, Muhammad et al.	پیش‌بینی فروش به کمک مدل‌های یادگیری ماشین	*	*	*		*	*		*																	*	*		
۱۳	Big data analytics in supply chain management: a systematic literature review and research directions	۲۰۲۲	Big Data and Cognitive Computing	Lee, In. Mangalaraj, George	مرور ادبیات سیستماتیک از مطالعات موجود در BDA																												
۱۴	Impact of replenishment strategies on supply chain performance under e-shopping scenario	۲۰۱۶	Computers &Industrial Engineering	Li, Jiafu et al.	شبیه‌سازی استراتژی	*																											مقایسه آماری

۱۵	Sales forecasting by combining clustering and machine-learning techniques for computer retailing	۲۰۱۷	Neural Computing and Applications	Chen, I-Fei. Lu, Chi-Jie	پیش‌بینی فروش با ترکیب تکنیک‌های خوشه‌بندی و یادگیری ماشین	*	*										SOM - GHSOM								MAPE - RMSPE
۱۶	Benchmarking of regression algorithms and time series analysis techniques for sales forecasting	۲۰۱۹	Balkan Journal of Electrical and Computer Engineering	Catal, Cagatay et al.	پیش‌بینی فروش با استفاده از روش‌های رگرسیون در یادگیری ماشین و تکنیک‌های تحلیل سری‌های زمانی	*	*	*									*	ELM - SVR	رگرسیون شبکه عصبی - رگرسیون جنگل تصادفی - رگرسیون درخت تصمیم - رگرسیون بیزی							Coefficient of Determination
۱۷	Comparison study: Product demand forecasting with machine learning for shop	۲۰۱۹	۸ ۲۰۱۹th International Conference System Modeling and Advancement in Research Trends (SMART)	Arif, Md Ariful Islam et al.	پیش‌بینی تقاضای محصول به کمک یادگیری ماشین	*	*	*	*									Gaussian Nave Bayes		*	*	*				MAPE - MPE - Precision - Accuracy – Score F۱

۱۸	Artificial intelligence in retail: The AI-enabled value chain	۲۰۲۱	Australasian Marketing Journal	Oosthuizen, Kim et al.	بررسی نقش هوش مصنوعی در زنجیره ارزش خرده‌فروشی	*																		
۱۹	Low-carbon decision-making model of online shopping supply chain considering the O۲O model	۲۰۲۱	Journal of Retailing and Consumer Services	Wu, Yisheng et al.	ارائه مدل مبتنی بر محیط‌زیست برای خرید آنلاین																			
۲۰	E-commerce Logistics in Supply Chain Management: Practice Perspective	۲۰۱۶	ScienceDirect	Ying, Yua et al.	بهبود مدل‌ها و تکنیک‌های موجود														بهینه‌سازی لجستیک تجارت الکترونیک					ارائه راهبرد
۲۱	Customer satisfaction assessment of E-Supply chain quality in online shopping: a case study	۲۰۲۰	Preprints	Maryam, Abdirad. Krishna, Krishnan.	مدل‌سازی مفهومی با نظرسنجی	*								حداقل مربعات جزئی ساختاری PLS بوت‌استرپ bootstrap										ارائه مدل مفهومی

۲۲	A Predictive Analysis of Retail Sales Forecasting using Machine Learning Techniques	۲۰۲۱	LGU Research Journal of Computer Science & IT	Muhammad Sajawal et al.	تجزیه‌وتحلیل پیش‌بینی‌کننده با استفاده از تکنیک‌های یادگیری ماشین	*	*				*				مدل‌های رگرسیون مختلف (رگرسیون خطی، رگرسیون پیش‌فرض تصادفی،						*	*				*
۲۳	An Aggregate Store Sales Forecasting Framework based on ConvLSTM	۲۰۲۱	Association for Computing Machinery	Yanbing Liu et al.	ارائه چارچوب برای پیش‌بینی کوتاه‌مدت و بلندمدت	*	*						*									*				ارائه چارچوب پیشنهادی
۲۴	Modeling Retail Supply Chain Efficiency: Exploration and Comparative Analysis of Different Approaches	۲۰۲۳	MDPI	Milan Andreji´c	ارائه مدل‌های پژوهشی برای اندازه‌گیری کارایی	*									مدل‌های DEA مدل‌های تجزیه کارایی											ارائه رویکرد مختلف
مطالعه حاضر					داده‌کاوی و یادگیری ماشین	*	*	*		*	*	*	*	*	سایر الگوریتم‌ها											سایر معیارهای ارزیابی

۲-۳- شکاف تحقیق

باتوجه‌به بخش مرور ادبیات، نکته‌ای که حائز اهمیت است، حجم داده‌های مورد بررسی در مقالات است که غالباً از داده‌های محدودی استفاده شده است؛ بنابراین، استفاده از داده‌های حجیم‌تر برای افزایش دقت نتایج، به‌عنوان شکاف تحقیقاتی در این بررسی شناخته می‌شود. همچنین استفاده از الگوریتم‌های یادگیری ماشین به‌صورت ترکیبی تنها در یک مورد از مقالات، بررسی شده است و مطالعه مقالات نشان می‌دهد که همچنان نیاز به توسعه الگوریتم‌های یادگیری ماشین در این زمینه وجود دارد. به‌علاوه، تجزیه‌وتحلیل اکتشافی، در هیچ‌کدام از مقالات وجود ندارد و استخراج دانش به‌ندرت در مقاله‌های یادگیری ماشین صورت‌گرفته است.

در جدول مرور ادبیات مشاهده می‌شود که ۱۶ مورد از ۲۴ مقاله بررسی‌شده دارای مطالعه موردی هستند و باتوجه‌به اهمیت پیاده‌سازی روش‌ها در مطالعه موردی، در این پژوهش نیز از مطالعه موردی استفاده خواهد شد. براین‌اساس در این مطالعه، از مجموعه‌داده‌های یک فروشگاه زنجیره‌ای با فروش حضوری و اینترنتی استفاده شده است و تلاش می‌شود در ابتدا به بررسی و تجزیه‌وتحلیل اکتشافی داده‌ها پرداخته شود و در ادامه با استفاده از الگوریتم‌های یادگیری ماشین عوامل مؤثر بر فروش و سایر ویژگی‌های تأثیرگذار بر عملکرد زنجیره تامین فروشگاه مدل‌سازی شود تا بتوان میزان فروش و همچنین عوامل مؤثر را پیش‌بینی و بررسی نمود.

۳- مطالعه موردی

در این مطالعه، برای تجزیه‌وتحلیل زنجیره‌تامین یک فروشگاه با فروش حضوری و اینترنتی و پیاده‌سازی الگوریتم‌های یادگیری ماشین، ما از مجموعه‌داده اطلاعات زنجیره‌تامین یک فروشگاه با خرید اینترنتی و حضوری که توسط شرکت دیتاکو[۱۲] بین سال‌های ۲۰۱۵ تا ۲۰۱۸، منتشر شده است، استفاده کرده‌ایم. این مجموعه‌داده در ابتدا شامل ۵۳ ستون (ویژگی[۱۳]) و ۱۸۰۵۱۹ سطر (نمونه[۱۴]) بوده است که پس از انتخاب ویژگی‌ها[۱۵]، پاک‌سازی داده[۱۶] ، کاهش ابعاد[۱۷] و استخراج ویژگی‌های مجموعه‌داده، تعداد ستون‌ها به ۶۰ ستون و سطرها به ۱۸۰۵۱۶ سطر رسید.

در ادامه، در جدول (۴-۱) ویژگی‌های مجموعه‌داده پاک‌سازی شده، به‌صورت مختصر توضیح داده شده است.

جدول (۳-۱) تعریف ویژگی‌های مجموعه‌داده پاک‌سازی شده

نام ویژگی	تعریف
Type	نوع معامله انجام شده
Days for shipping (real)	روزهای ارسال واقعی محصول خریداری شده
Days for shipment (scheduled)	روزهای تحویل برنامه‌ریزی شده محصول خریداری شده
Benefit per order	درآمد به‌ازای هر سفارش ثبت شده
Sales per customer	مجموع فروش انجام شده به‌ازای هر مشتری
Delivery Status	وضعیت تحویل سفارش‌ها: ارسال پیش از زمان، تأخیر در تحویل، لغو ارسال، ارسال به‌موقع
Late_delivery_risk	متغیر طبقه‌بندی که نشان می‌دهد اگر ارسال دیر باشد (۱) و اگر دیر نشده است (۰).
Category Id	کد دسته محصول
Category Name	نام دسته‌بندی محصول
Customer City	شهری که مشتری خرید را انجام داده است.
Customer Country	کشوری که مشتری خرید را انجام داده است.
Customer Id	شناسه مشتری
Customer Segment	بخش‌بندی مشتریان: مصرف‌کننده، شرکت‌های بزرگ، شرکت‌های کوچک
Customer State	ایالتی که فروشگاهی که خرید در آن ثبت شده است، به آن تعلق دارد.
Customer Zipcode	کد پستی مشتری

[۱۲] Data-Co
[۱۳] Feature
[۱۴] Instance (Record)
[۱۵] Feature Selection
[۱۶] Data Cleaning
[۱۷] Dimension Reduction

Department Id	کد دپارتمان فروشگاه
Department Name	نام دپارتمان فروشگاه
Latitude	عرض جغرافیایی مربوط به محل فروشگاه
Longitude	طول جغرافیایی مربوط به محل فروشگاه
Market	بازار محل تحویل سفارش: آفریقا، اروپا، آمریکای لاتین، آسیا اقیانوسیه، ایالات متحده آمریکا
Order City	شهر مقصد سفارش
Order Country	کشور مقصد سفارش
Order Customer Id	کد سفارش مشتری
order date (DateOrders)	تاریخ و ساعتی که در آن سفارش انجام شده است.
Order Id	کد سفارش
Order Item Cardprod Id	کد محصول تولید شده از طریق خواننده فرکانس رادیویی[۱۸]
Order Item Discount	مقدار تخفیف کالای سفارشی
Order Item Discount Rate	درصد تخفیف کالای سفارشی
Order Item Id	کد کالای سفارشی
Order Item Product Price	قیمت محصولات بدون تخفیف
Order Item Profit Ratio	نرخ سود کالای سفارشی
Order Item Quantity	تعداد محصولات در هر سفارش
Sales	میزان فروش
Order Item Total	مبلغ کل به‌ازای هر سفارش
Order Profit Per Order	سود سفارش در هر سفارش
Order Region	منطقه‌ای از جهان که در آن سفارش تحویل داده می‌شود: آسیای جنوب شرقی، آسیای جنوبی، اقیانوسیه، آسیای شرقی، غرب آسیا، غرب ایالات متحده و غیره
Order State	ایالتی منطقه‌ای که سفارش در آن تحویل داده می‌شود.
Order Status	وضعیت سفارش: کامل، در انتظار، بسته، در انتظار پرداخت، لغو شده، در حال پردازش و غیره
Product Card Id	کد محصول
Product Category Id	کد دسته محصول
Product Name	نام محصول
Product Price	قیمت محصول
shipping date (DateOrders)	تاریخ و زمان دقیق حمل‌ونقل
Shipping Mode	حالت‌های حمل‌ونقل شامل کلاس استاندارد، کلاس اول، کلاس دوم، همان روز
DateOrders	تاریخی که در آن سفارش انجام شده است.

[۱۸] RFID: Radio Frequency Identification

TimeOrders	زمانی که در آن سفارش انجام شده است.
DayOrders	روزی از ماه که در آن سفارش انجام شده است.
MonthOrders	ماهی که در آن سفارش انجام شده است.
YearOrders	سالی که در آن سفارش انجام شده است.
DayOfWeekOrders	روزی از هفته که در آن سفارش انجام شده است.
HourOrders	ساعتی از روز که در آن سفارش انجام شده است.
MinutesOrders	دقیقه‌ای از ساعت که در آن سفارش انجام شده است.
shipping Date	تاریخ حمل‌ونقل
shipping Time	زمان حمل‌ونقل
shipping Day	روز حمل‌ونقل
shipping Month	ماه حمل‌ونقل
shipping Year	سال حمل‌ونقل
DayOfWeek Shipping	روز هفته حمل‌ونقل
shipping Hour	ساعت حمل‌ونقل
shipping Minute	دقیقه‌ی حمل‌ونقل

۴- مبانی نظری

۴-۱- زنجیره‌تامین

زنجیره تامین یک فرایند یکپارچه است که شامل تبدیل مواد خام به محصول نهایی و در نهایت تحویل به مشتری می‌شود. فرایند کامل زنجیره تامین به چهار سطح تقسیم می‌شود: تأمین‌کنندگان، تولیدکنندگان، توزیع‌کنندگان و مشتریان (ماتور و همکاران ۲۰۱۷).

۴-۲- عارضه‌یابی

عارضه‌یابی مجموعه‌ای از فرایندها است که به مفهوم عامی اشاره می‌کند و آن چیزی نیست جز شناسایی و کشف معضلات و عارضه‌های سازمان در بخش‌های مختلف به‌منظور رفع و توسعه ظرفیت‌های سازمان.

سازمان‌ها درگذر زمان با تأثیرگرفتن از تصمیمات مختلف و تغییرات محیطی درگیر مشکلات متعددی می‌شوند که برخی از این مشکلات در توسعه و رشد سازمان اثرگذار هستند. اغلب مشکلات و عارضه‌های سازمانی به‌صورت پنهان در سازمان باقی‌مانده و غالب آن‌ها به‌صورت محدودیت عمل می‌کنند. در واقع این مشکلات محدودیت نیستند؛ ولی چون رفع نشده‌اند به‌صورت محدودیت عمل کرده و فرایندهای سازمان را با اختلال همراه می‌کنند. حرکت سازمان به سمت بهبود و توسعه زمانی به طور صحیح و اثربخش انجام می‌گردد که عارضه‌های سازمانی، دلایل آن‌ها و اولویت دلایل به‌درستی شناسایی و مشخص گردند. از طرفی شناسایی نقاط ضعف یک سازمان زمانی می‌تواند برای سازمان اثربخش باشد که باعث شناسایی نقاط بهبود بالقوه گردد. پس از بررسی و شناسایی عارضه‌ها و نقاط ضعف سازمان، بهبودهای بالقوه شناسایی می‌شوند پس از این مرحله و با اولویت‌بندی، پروژه‌های بهبود برای بهبود در عملکردها و قابلیت‌های سازمان تعریف می‌گردد.

۴-۳- مدیریت زنجیره تامین

مدیریت زنجیره‌تامین، مدیریت جریان کالاها و خدمات، بین مشاغل و مکان‌ها است و شامل جابه‌جایی و ذخیره‌سازی مواد خام، موجودی در جریان کار و کالاهای نهایی و همچنین انجام سفارش از مبدأ تا نقطه مصرف می‌باشد. در واقع مدیریت زنجیره‌تامین ادغام فرایندهای تجاری کلیدی از کاربر نهایی از طریق تأمین‌کنندگان اصلی است که محصولات، خدمات و اطلاعاتی را ارائه می‌دهد که برای مشتریان و سایر ذی‌نفعان ارزش‌افزوده ایجاد می‌کند (لامبرت و همکاران، ۱۹۹۸؛ کوپر و همکاران، ۱۹۹۷).

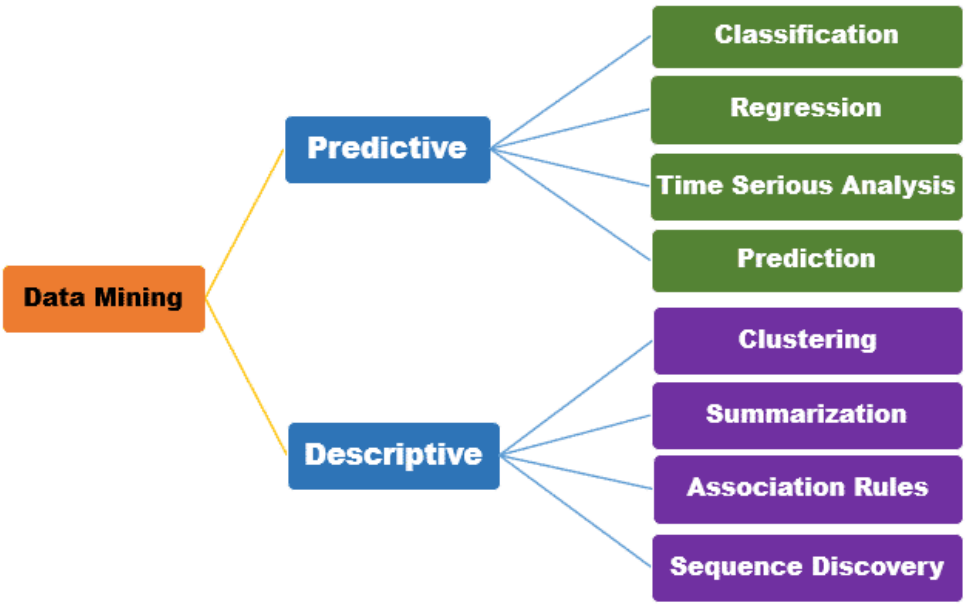
کریستوفر (۱۹۹۸) مدیریت زنجیره‌تامین را چنین تعریف کرد: «زنجیره‌تامین شبکه‌ای از سازمان‌ها است که از طریق پیوندهای بالادستی و پایین‌دستی درگیر فرایندهای مختلف و فعالیت‌هایی است که در قالب محصولات و خدمات در دست مشتری نهایی ارزش تولید می‌کنند.»

۴-۴- داده‌کاوی[۱۹]

به مجموعه‌ای از روش‌های قابل‌اعمال بر پایگاه‌داده‌های بزرگ و پیچیده به‌منظور کشف الگوهای پنهان و جالب‌توجه نهفته در میان داده‌ها، داده‌کاوی گفته می‌شود. علم میان‌رشته‌ای داده‌کاوی، پیرامون ابزارها، متدولوژی‌ها و تئوری‌هایی است که برای آشکارسازی الگوهای موجود در داده‌ها مورداستفاده قرار می‌گیرند و گامی اساسی در راستای کشف دانش محسوب می‌شود.

داده‌کاوی که با عنوان «کشف دانش از داده» [۲۰] نیز شناخته شده است، فرایند استخراج اطلاعات و دانش از داده‌های موجود در پایگاه‌داده یا انبار داده است.

۴-۴-۱- انواع داده‌کاوی

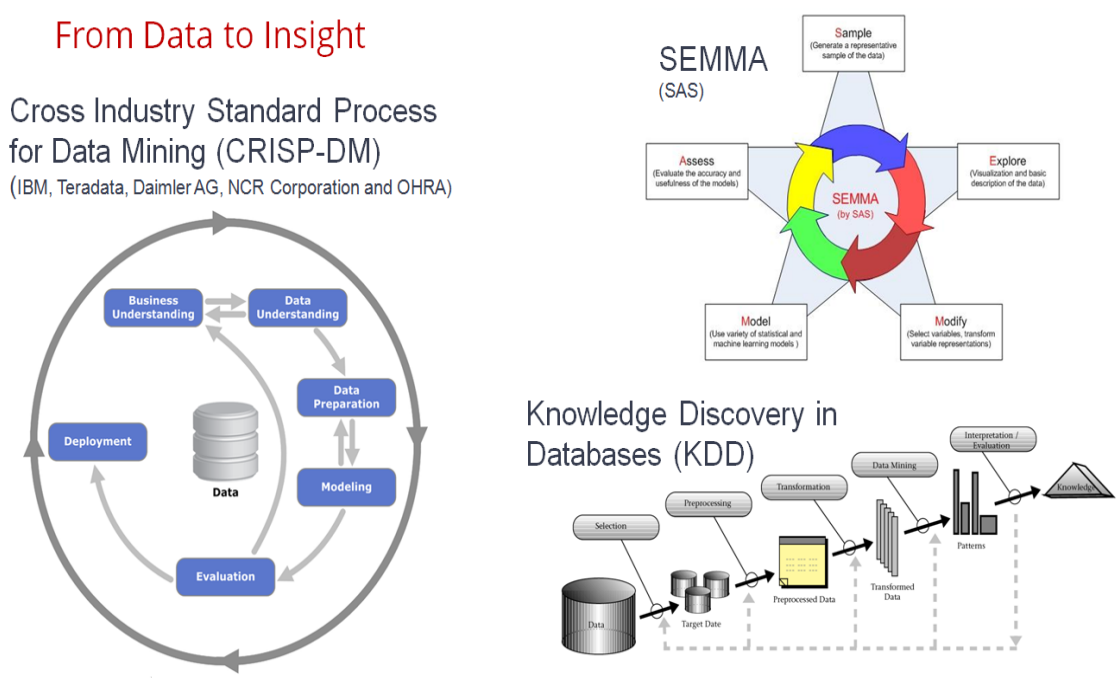


شکل (۴-۱) انواع داده‌کاوی

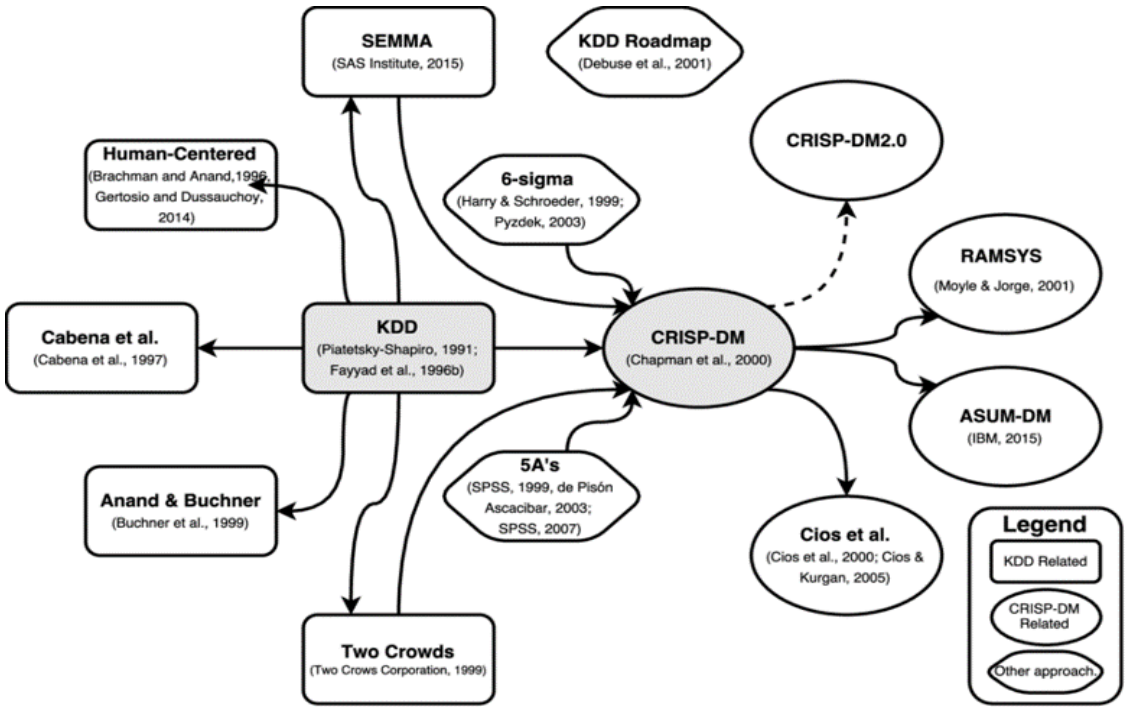
متدولوژی‌های متفاوتی برای داده‌کاوی از قبل سال ۱۹۹۶ ارائه شده است که در شکل‌های زیر مهم‌ترین آن‌ها آورده شده است:

---

[۱۹] Data Mining
[۲۰] Knowledge Discovery From Data | KDD



شکل (۴-۲) متدولوژی‌های داده‌کاوی



شکل (۴-۳) متدولوژی‌های داده‌کاوی

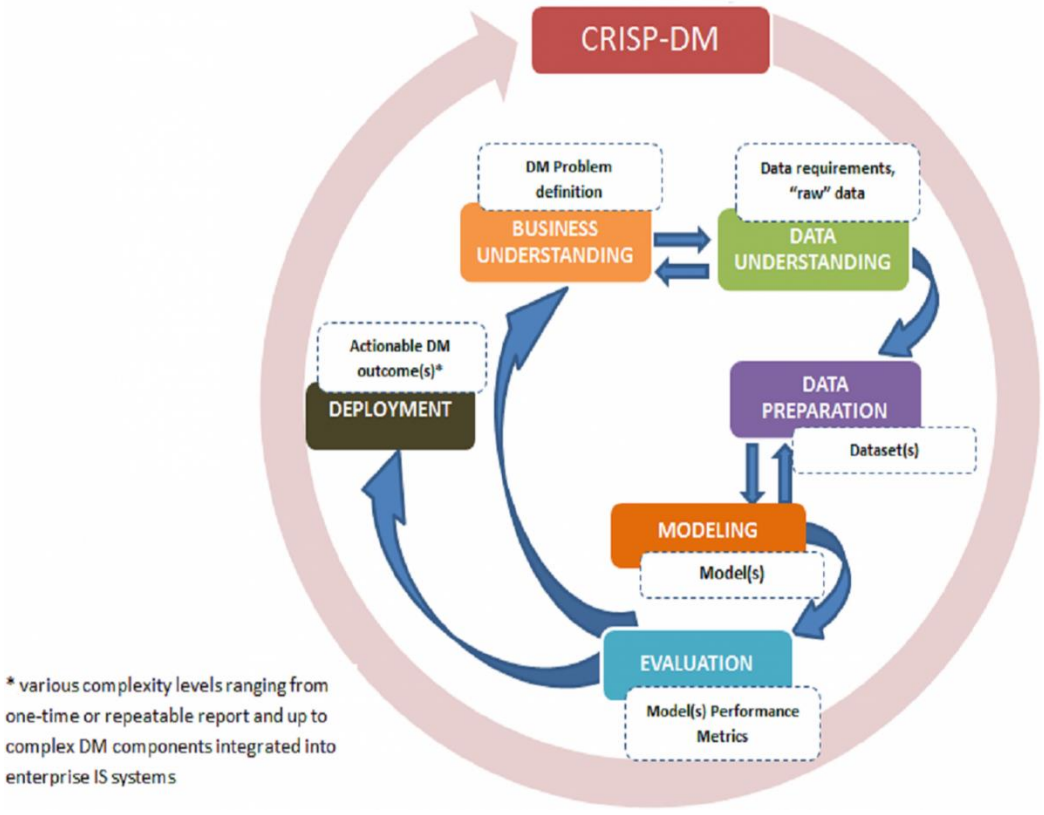
برخی از معروف‌ترین مدل‌های داده‌کاوی مدل‌های مبتنی بر CRISP-DM از شرکت IBM و مدل‌های مبتنی بر KDD است که در ادامه باتوجه‌به اهمیت مدل CRISP-DM توضیح مختصری در مورد آن داده خواهد شد.

۴-۵- مدل CRISP-DM

روش داده‌کاوی مبتنی بر CRISP-DM در ۶ گام زیر صورت می‌گیرد:

۱. درک کسب‌وکار
۲. بررسی و درک داده‌ها
۳. آماده‌سازی داده‌ها
٤. مدل‌سازی
٥. تست و ارزیابی مدل

توسعه مدل **نهایی و استقرار**



شکل (۴-۴) مدل داده‌کاوی CRISP-DM

۴-۵-۱- مرحله اول: درک کسب‌وکار

کاربران برای اتخاذ تصمیم‌های مناسب در هنگام ایجاد مدل‌های داده‌کاوی باید به درک صحیحی از داده‌ها برسند. در این مرحله مواردی همچون الزامات مربوط به کسب‌وکار، تعریف چارچوب مسئله، تعریف معیارهای مورداستفاده برای ارزیابی مدل و تعریف اهداف مشخص برای پروژه داده‌کاوی صورت می‌پذیرد.

۴-۵-۲- مرحله دوم: بررسی و درک داده‌ها

متخصص داده‌کاوی، داده‌های ثبت شده در کسب‌وکار کارفرما را از وی درخواست می‌کند و به بررسی داده‌ها می‌پردازد. متخصص داده‌کاوی باتوجه‌به حجم و کیفیت داده‌ها، مسئله طرح شده در مرحله قبل را تعدیل می‌کند تا نتیجه‌ی پروسه‌ی داده‌کاوی واقع‌بینانه‌تر به شبه طور به طور خلاصه می‌توان دیتاست‌ها را به موارد زیر تقسیم‌بندی نمود:

- **کمی:** همانند درجه حرارت و قد افراد
- **کیفی:** همانند دسته مدارک تحصیلی (دیپلم، لیسانس و...) یا گروه رنگ‌ها (زرد، قرمز و...)
- **ترتیبی:** چنین داده‌هایی دارای یک ترتیبت طبیعی هستند؛ همانند مدارج تحصیلی (دبستان، راهنمایی، دبیرستان، کارشناسی، کارشناسی ارشد، دکتری)
- **اسمی:** اسامی دسته‌ها همانند وضعیت تأهل، جنسیت و رنگ‌ها
- **عددی:** داده‌های عددی خود به دودسته فاصله‌ای و نسبتی تقسیم می‌شوند. داده‌های فاصله‌ای بر اساس مقیاس واحدهایی با اندازه برابر اندازه‌گیری می‌شوند. مقادیر ویژگی‌های عددی دارای ترتیب هستند و می‌توانند مثبت، صفر و یا منفی باشند. یک داده نسبتی، خصیصه عددی دارای یک صفر مطلق است. اگر اندازه‌ها نسبتی باشند، می‌توان از نسبت مقادیر با یکدیگر سخن گفت. به‌علاوه، مقادیر قابل مرتب‌سازی شدن هستند و می‌توان تفاضل بین آن‌ها، میانگین و مد را محاسبه نمود.

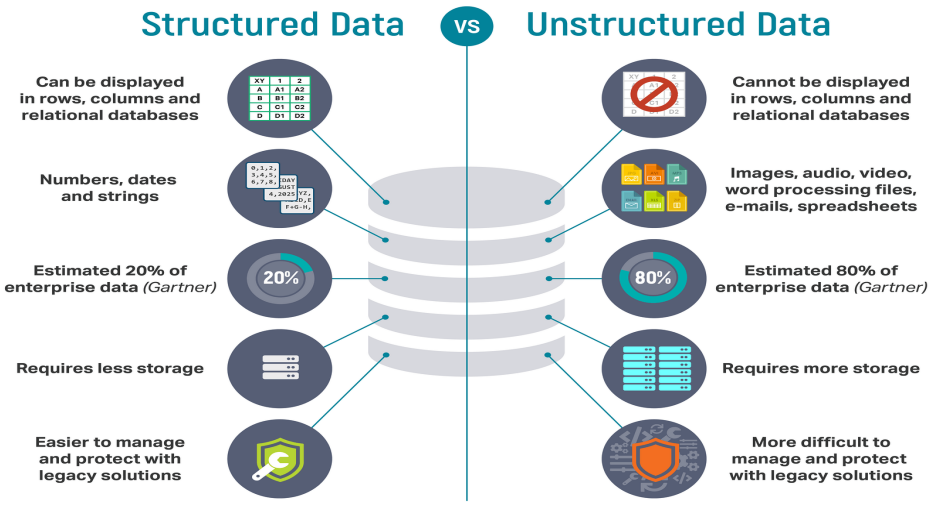
در بسیاری از مباحث داده‌کاوی، یادگیری ماشین و کلان‌داده‌ها، داده‌ها را می‌توان به دودسته تقسیم‌بندی کرد:

۱-داده‌های ساختاریافته[۲۱]

۲-داده‌های غیرساختاریافته[۲۲]

---

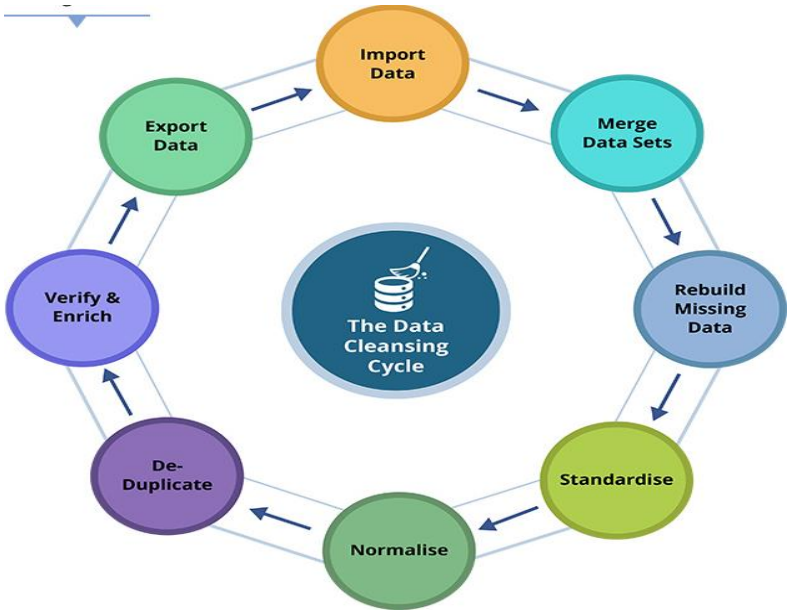
[۲۱] Structured Data
[۲۲] Un Structured Data



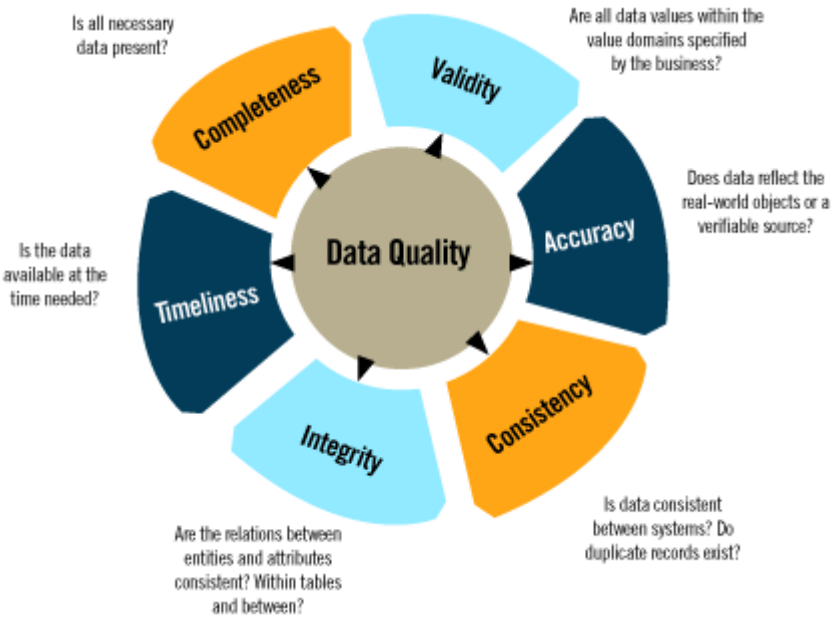
شکل (۴-۵) داده ساختاریافته و غیرساختاریافته

۴-۵-۳- مرحله سوم: آماده‌سازی یا پیش‌پردازش داده‌ها

این امکان وجود دارد که داده‌ها در سراسر سازمان توزیع شده و در قالب‌های مختلف ذخیره گردند و یا اینکه ممکن است شامل تناقضات و ناسازگاری‌هایی از جمله ورودی‌های نادرست یا ازدست‌رفته باشند.



شکل (۴-۶) پاک‌سازی داده



شکل (۴-۷) اعتبارسنجی داده

۴-۵-۴- مرحله چهارم: مدل‌سازی

قدم چهارم مدل‌سازی داده‌های آماده‌سازی شده است. باتوجه‌به متدهای متفاوت، مدل‌های متفاوتی ساخته می‌شود و بهترین مدل‌ها از نظر متخصص داده‌کاوی انتخاب می‌شود.

۴-۵-۵- مرحله پنجم: تست و ارزیابی مدل

پیش از پیاده‌سازی مدل در محیط عملیاتی باید نحوه عملکرد آن مورد بررسی قرار گیرد. به‌علاوه در هنگام تهیه مدل معمولاً باید چندین مدل با پیکربندی‌های متفاوت ارائه شوند تا پس از تست نمودن آن‌ها بتوان به مدلی دست‌یافت که بهترین نتیجه را در ارتباط با مشکلات و داده‌ها فراهم آورد.

مدل‌های ساخته شده تست و ارزیابی می‌شوند و بهترین مدل از نظر مسئلهٔ طرح شده در مرحله یک، انتخاب می‌شود.

۴-۵-۶- مرحله ششم: توسعه مدل نهایی و استقرار

پس از استقرار Mining Model در یک محیط عملیاتی می‌توان عملکردهای بسیاری را باتوجه‌به نیازها اجرا نمود. در زیر به برخی از این عملکردها اشاره می‌شود. استفاده از مدل‌ها برای فرایندهای پیش‌بینی که ممکن است در مراحل بعدی برای اتخاذ تصمیمات در کسب‌وکار نیز به کار گرفته شود.

- انجام Queryهای محتوا به‌منظور بازیابی اطلاعات آماری، قواعد یا فرمول‌های مربوط به مدل‌ها
- جای‌گذاری مستقیم عملکرد داده‌کاوی در برنامه‌های کاربردی

- ارائه گزارشی که امکان Query نمودن مستقیم در مدل داده‌کاوی موجود را برای کاربران فراهم می‌کند.

۴-۶- یادگیری ماشین[۲۳]

یادگیری ماشین شاخه‌ای از هوش مصنوعی و علوم کامپیوتر است که بر استفاده از داده‌ها و الگوریتم‌ها برای تقلید از روشی که انسان‌ها یاد می‌گیرند، تمرکز دارد و به‌تدریج دقت آن را بهبود می‌بخشد. یادگیری ماشین جزء مهمی از حوزه روبه‌رشد علم داده است که از طریق استفاده از روش‌های آماری، الگوریتم‌ها، برای دسته‌بندی یا پیش‌بینی و کشف بینش‌های کلیدی در پروژه‌های داده‌کاوی آموزش داده می‌شوند. این بینش‌ها متعاقباً تصمیم‌گیری را در برنامه‌ها و کسب‌وکارها هدایت می‌کنند و به طور ایده‌آل بر معیارهای رشد کلیدی تأثیر می‌گذارند. الگوریتم‌های یادگیری ماشین از داده‌های ساختاریافته و برچسب‌گذاری شده برای پیش‌بینی استفاده می‌کنند. به این معنی که ویژگی‌های خاصی از داده‌های ورودی برای مدل تعریف شده و در جداول سازمان‌دهی می‌شوند. این لزوماً به این معنی نیست که از داده‌های بدون ساختار استفاده نمی‌کنند. این فقط به این معنی است که معمولاً برای سازماندهی داده‌ها در قالبی ساختاریافته، داده‌ها فرایند پیش‌پردازش را طی می‌کنند.

۴-۶-۱- انواع یادگیری ماشین

- یادگیری با نظارت[۲۴]
- یادگیری بدون نظارت[۲۵]
- یادگیری تقویتی[۲۶]

۴-۶-۱-۱- الگوریتم‌های یادگیری ماشین با نظارت

این نوع از یادگیری، یک نوع یادگیری ماشین است که در آن الگوریتم از داده‌های برچسب‌دار یاد می‌گیرد. داده برچسب‌گذاری شده، به معنای مجموعه‌داده‌ای است که متغیر هدف مربوطه آن از قبل مشخص است. یادگیری با نظارت دو نوع دارد.

- **دسته‌بندی:** در این نوع از الگوریتم‌ها، کلاس مجموعه‌داده بر اساس متغیر ورودی مستقل پیش‌بینی می‌شود. کلاس مقادیر مقوله‌ای گسسته است. مثلاً تصویر حیوان گربه یا سگ است.
- **رگرسیون:** در این نوع از الگوریتم‌ها متغیرهای خروجی پیوسته بر اساس متغیر ورودی مستقل پیش‌بینی می‌شود. برای مثال پیش‌بینی قیمت مسکن بر اساس پارامترهای مختلف مانند سن خانه، فاصله از جاده اصلی، موقعیت مکانی، مساحت و غیره.

---

[۲۳] Machine Learning
[۲۴] Supervised Learning
[۲۵] Unsupervised Learning
[۲۶] Reinforcement Learning

۴-۶-۱-۲- الگوریتم‌های یادگیری ماشین بدون نظارت

در یادگیری بدون نظارت، الگوریتم باید خود به‌تنهایی به دنبال ساختارهای جالب موجود در داده‌ها باشد. به بیان ریاضی، یادگیری بدون نظارت مربوط به زمانی است که در مجموعه‌داده فقط متغیرهای ورودی وجود داشته باشند و هیچ متغیر داده خروجی موجود نباشد. به این نوع یادگیری، بدون نظارت گفته می‌شود. زیرا برخلاف یادگیری با نظارت، هیچ پاسخ صحیح داده شده‌ای وجود ندارد و ماشین خود باید به دنبال پاسخ باشد.

به بیان دیگر، هنگامی که الگوریتم برای کارکردن از مجموعه‌داده‌ای بهره گیرد که فاقد داده‌های برچسب‌دار (متغیرهای خروجی) است، از مکانیزم دیگری برای یادگیری و تصمیم‌گیری استفاده می‌کند. به چنین نوع یادگیری، بدون نظارت گفته می‌شود. یادگیری بدون نظارت قابل‌تقسیم به مسائل خوشه‌بندی و انجمنی است.

- **قوانین انجمنی:** یک مسئله یادگیری هنگامی قوانین انجمنی محسوب می‌شود که هدف کشف‌کردن قواعدی باشد که بخش بزرگی از داده‌ها را توصیف می‌کنند. مثلاً شخصی که کالای الف را خریداری کند، تمایل به خرید کالای ب نیز دارد.
- **خوشه‌بندی:** یک مسئله هنگامی خوشه‌بندی محسوب می‌شود که قصد کشف گروه‌های ذاتی (داده‌هایی که ذاتاً در یک گروه خاص می‌گنجند) در داده‌ها وجود داشته باشد. مثلاً، بخش‌بندی مشتریان بر اساس رفتار خرید آن‌ها.

۴-۶-۱-۳- یادگیری تقویتی

یک برنامه رایانه‌ای که با محیط پویا در تعامل است باید به هدف خاصی دست یابد (مانند بازی‌کردن با یک رقیب یا راندن خودرو). این برنامه بازخوردهایی را با عنوان پاداش‌ها و تنبیه‌ها فراهم و فضای مسئله خود را بر همین اساس هدایت می‌کند. با استفاده از یادگیری تقویتی، ماشین می‌آموزد که تصمیمات مشخصی را در محیطی که دائم در معرض آزمون‌وخطا است، اتخاذ کند.

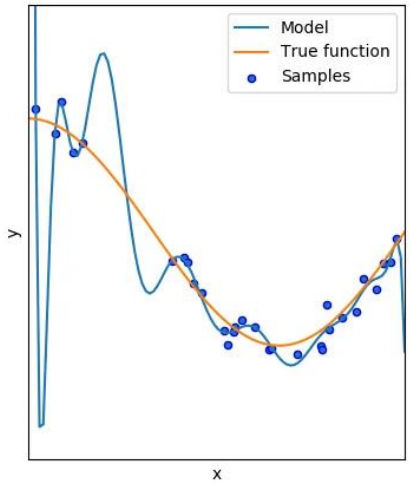
۴-۶-۲- بیش برازش[۲۷]، کم برازش[۲۸] و برازش مناسب

مدل بیش برازش، مدلی بسیار پیچیده برای داده‌ها است. به این معنی که در تحلیل رگرسیونی، مدلی با بیشترین پارامترها ایجاد می‌شود. در چنین حالتی، مدل با تغییرات جهشی سعی در پوشش داده‌های حاصل از نمونه و حتی مقدارهای نویز می‌کند. درحالی‌که چنین مدلی باید منعکس‌کننده رفتار جامعه باشد. در این گونه موارد، اگر مدل رگرسیون به‌دست‌آمده، برای پیش‌بینی نمونه دیگری به کار رود، مقدارهای پیش‌بینی شده اصلاً مناسب به نظر نخواهند رسید.

---

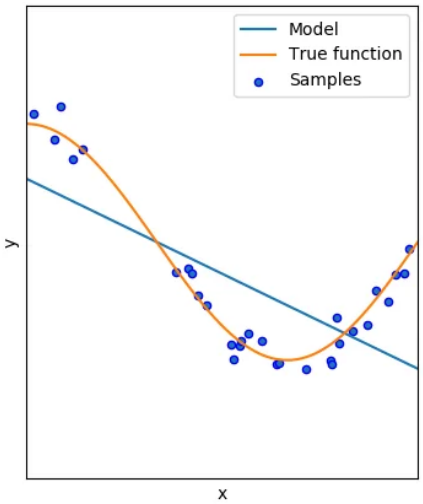
[۲۷] Overfitting
[۲۸] Underfitting

در تصویر زیر، نمودار حاصل از بیش برازش روی داده‌های حاصل از نمونه دیده می‌شود. خط آبی، نشان‌دهنده منحنی برازش شده روی داده‌ها است و خط نارنجی تابعی است که مدل واقعی جامعه آماری را نشان می‌دهد. نقاط آبی‌رنگ نیز نمونه‌های تصادفی از جامعه آماری را نشان می‌دهند. در مدل بیش برازش، نقطه‌های حاصل از نمونه بهترین برازش را دارند و خط آبی تقریباً از همه آن‌ها عبور کرده است.



نمودار (۴-۱) منحنی بیش برازش بر اساس چندجمله‌ای مرتبه ۱۵

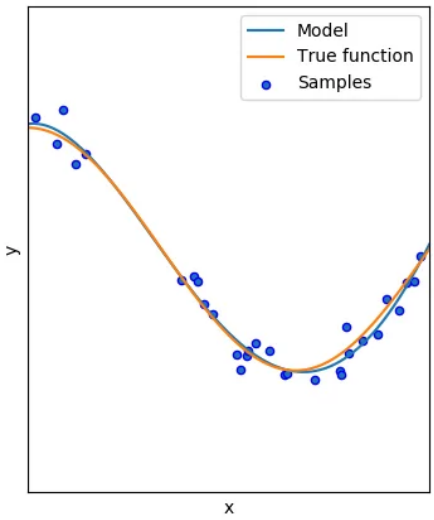
همچنین در زمانی که پارامترهای مدل رگرسیونی به‌صورت کم برازش برآورد می‌شوند، جانب احتیاط حفظ شده و مدل سعی می‌کند با کمترین پارامترها، عمل برازش را انجام دهد. در نتیجه خطای حاصل از این مدل حتی بر اساس نمونه‌های به‌کاررفته نیز بسیار زیاد است. در تصویر زیر، یک نمونه از مدل رگرسیونی کم برازش دیده می‌شود. درجه منحنی به‌کاررفته در این حالت ۱ است که معادله خط محسوب می‌شود.



نمودار (۴-۲) منحنی کم برازش بر اساس چندجمله‌ای مرتبه ۱

انتظار ما از یک تحلیل رگرسیون مناسب، ایجاد مدلی است که نه‌تنها بتواند برای داده‌های مربوط به نمونه، برازش مناسب را انجام دهد، بلکه برای داده‌هایی جدید نیز امکان برآورد مناسب وجود داشته باشد.

همان‌طور که در تصویر زیر دیده می‌شود، مدل مناسب دارای خطای کوچکی است و قابلیت پیش‌بینی برای داده‌های جدید را دارد.



نمودار (۴-۳) منحنی برازش مناسب بر اساس چندجمله‌ای مرتبه ۴

۴-۶-۳- موازنه واریانس و بایاس[۲۹]

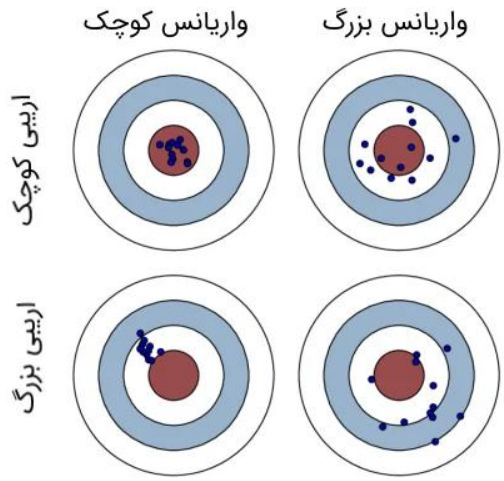
خطای بایاس: بایاس در واقع میزان اختلاف نقاط پیش‌بینی شده از متغیر هدف واقعی است. وجود فرضیه‌های مختلف روی مدل و الگوریتم یادگیری منجر به ایجاد خطای اریبی می‌شود. بزرگ‌بودن اریبی می‌تواند الگوریتم یا مدل آماری را از کشف روابط بین ویژگی‌ها و متغیر پاسخ باز دارد. اغلب بزرگ‌بودن خطای اریبی، منجر به کم برازش می‌شود.

خطای واریانس: واریانس میزان پراکندگی نقاط را نشان می‌دهد. هر چه واریانس بیشتر باشد، پراکندگی داده‌ها بیشتر است. حساسیت زیاد مدل با تغییرات کوچک روی داده‌های آموزشی، نشانگر وجود واریانس زیاد است. این امر نشانگر آن است که اگر مدل آموزش‌داده‌شده را روی داده‌های آزمایشی به کار گیریم، نتایج حاصل با داده‌های واقعی فاصله زیادی خواهند داشت. متأسفانه افزایش واریانس در این حالت منجر به مدل‌بندی مقادیر نویز[۳۰] شده و به‌جای پیش‌بینی صحیح، دچار پیچیدگی و مشکل بیش برازش می‌شود.

شکل زیر مصورسازی مفهوم موازنه واریانس و بایاس را نشان می‌دهد.

---

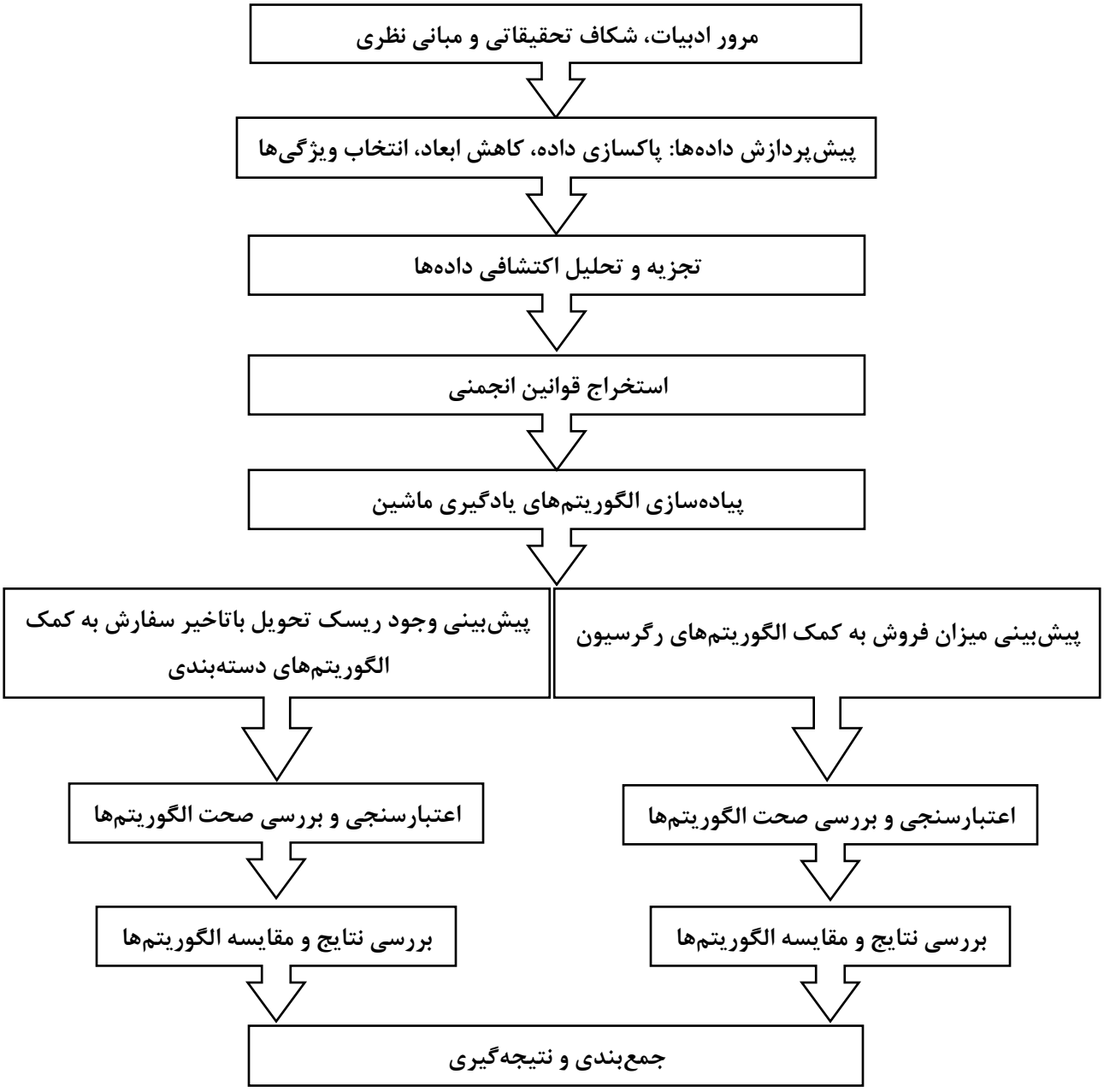
[۲۹] Bias-Variance Tradeoff
[۳۰] Noise



شکل (۴-۸) مصورسازی موازنه واریانس و بایاس برای برآوردگر

۵- روش تحقیق

روش تحقیق این مطالعه، در شکل زیر آمده است.



شکل (۵-۱) روش تحقیق مطالعه

۵-۱- پیش‌پردازش داده‌ها: پاک‌سازی داده، کاهش ابعاد، انتخاب ویژگی‌ها

مجموعه‌داده توسط شرکت دیتاکو منتشر شده است. سه سطر از داده‌ها که غلط وارد شده‌اند، حذف گردید. از ستون‌های تاریخ و زمان سفارش و تاریخ و زمان حمل‌ونقل، متغیرهای زمان به تفکیک استخراج

گردید و به ستون‌های مجموعه‌داده اضافه شدند و فرایند استخراج ویژگی انجام شد. ستون‌هایی که تعداد زیادی از داده‌هایشان وجود نداشت، حذف شدند و در نهایت سطرهایی که داده‌های خالی داشتند، حذف گردید.

۵-۲- تجزیه‌وتحلیل اکتشافی داده‌ها

در این بخش، تجزیه تحلیل فراوانی روی ستون‌های (ویژگی‌های) موجود در مجموعه‌داده صورت‌گرفته است و با استخراج دانش، پیشنهاداتی برای بهبود ارائه گردیده است. همچنین همبستگی و ارتباط بین ستون‌ها نیز در این بخش مورد بررسی قرار گرفته است.

۵-۳- استخراج قوانین انجمنی[۳۱]

استخراج قوانین انجمنی، تکنیکی است که برای کشف روابط پنهان‌بین متغیرها در مجموعه‌داده‌های بزرگ استفاده می‌شود. این یک روش محبوب در داده‌کاوی و یادگیری ماشین است و کاربردهای گسترده‌ای در زمینه‌های مختلف مانند تجزیه‌وتحلیل سبد بازار، تقسیم‌بندی مشتریان و کشف تقلب دارد. هدف از استخراج قوانین انجمنی، کشف قوانینی است که روابط بین متغیرهای مختلف در مجموعه‌داده را توصیف می‌کند.

برای مثال، مجموعه داده‌ای از معاملات در یک فروشگاه مواد غذایی را در نظر بگیرید. استخراج قوانین انجمنی می‌تواند برای شناسایی روابط بین اقلامی که اغلب با هم خریداری می‌شوند، استفاده شود. برای مثال، قانون «اگر مشتری نان بخرد، احتمالاً شیر هم می‌خرد.» یک قانون انجمنی است که می‌تواند از این مجموعه‌داده استخراج شود. ما می‌توانیم از چنین قوانینی برای اطلاع از تصمیم‌گیری‌ها در مورد چیدمان فروشگاه، قرارگیری محصول و بازاریابی استفاده کنیم.

الگوریتم‌های مختلفی برای استخراج قوانین انجمنی وجود دارند. در ادامه به پرکاربردترین آن‌ها اشاره می‌شود.

- **الگوریتم Apriori**: الگوریتم Apriori یکی از پرکاربردترین الگوریتم‌ها برای استخراج قوانین انجمنی است. این الگوریتم، ابتدا مجموعه موارد پرتکرار در مجموعه‌داده را شناسایی می‌کند (مجموعه‌هایی که در تعداد معینی از رکوردها ظاهر می‌شوند). سپس از این مجموعه موارد پرتکرار برای تولید قوانین انجمنی استفاده می‌کند. الگوریتم Apriori از یک رویکرد پایین‌به‌بالا استفاده می‌کند که از موارد جداگانه شروع می‌شود و به‌تدریج به مجموعه‌های موارد پیچیده‌تر می‌رسد.

[۳۱] Association Rule Mining

- **الگوریتم FP-Growth** [۳۲]: الگوریتم رشد الگوی پرتکرار، یکی دیگر از الگوریتم‌های محبوب برای استخراج قوانین انجمنی است که با ساختن یک ساختار درخت‌مانند به نام FP-tree کار می‌کند که مجموعه موارد پرتکرار در مجموعه‌داده را رمزگذاری می‌کند. سپس از FP-tree برای ایجاد قوانین انجمنی به روشی مشابه الگوریتم Apriori استفاده می‌شود. الگوریتم رشد الگوی پرتکرار به‌طورکلی سریع‌تر از الگوریتم Apriori است.
- **الگوریتم ECLAT** [۳۳]: الگوریتم خوشه‌بندی کلاس هم ارز و پیمایش شبکه از بالابه‌پایین، نوعی از الگوریتم Apriori است که از رویکرد بالابه‌پایین به‌جای رویکرد از پایین‌به‌بالا استفاده می‌کند. با تقسیم موارد به کلاس‌های معادل بر اساس پشتیبانی آن‌ها (تعداد رکوردهایی که در آن‌ها ظاهر می‌شوند) کار می‌کند. سپس قوانین انجمنی با ترکیب این کلاس‌های هم ارزی در یک ساختار شبکه مانند ایجاد می‌شود. این یک نسخه کارآمدتر و مقیاس‌پذیرتر از الگوریتم Apriori است.

در این مطالعه ما از الگوریتم Apriori برای استخراج قوانین انجمنی استفاده کرده‌ایم. در ادامه نحوه عملکرد این الگوریتم را بیان می‌کنیم.

۵-۳-۱- الگوریتم Apriori

این الگوریتم، با تنظیم حداقل آستانه پشتیبانی[۳۴] شروع می‌شود. این عدد حداقل تعداد دفعاتی است که یک مورد باید در پایگاه‌داده رخ دهد تا بتوان آن را به‌عنوان مجموعه موارد پرتکرار در نظر گرفت. سپس الگوریتم هر مجموعه مواردی را که حداقل آستانه پشتیبانی را برآورده نمی‌کنند، فیلتر می‌کند.

سپس الگوریتم لیستی از تمام ترکیبات ممکن از مجموعه موارد پرتکرار ایجاد می‌کند و تعداد دفعاتی که هر ترکیب در پایگاه‌داده ظاهر می‌شود را می‌شمارد. در ادامه الگوریتم فهرستی از قوانین مرتبط را بر اساس ترکیبات پرتکرار مجموعه موارد تولید می‌کند.

قدرت[۳۵] قانون انجمنی با استفاده از معیار اطمینان[۳۶] اندازه‌گیری می‌شود که احتمال وجود مورد ب با توجه به وجود مورد الف است. سپس الگوریتم قوانین انجمنی را که حداقل آستانه اطمینان را برآورده نمی‌کند، فیلتر می‌کند. از این قوانین به‌عنوان قوانین انجمنی قوی یاد می‌شود. در نهایت، الگوریتم لیستی از قوانین مرتبط قوی را به‌عنوان خروجی برمی‌گرداند. در ادامه به معیارهای ارزیابی و تحلیل قوانین انجمنی می‌پردازیم.

---

[۳۲] Frequent Pattern Growth
[۳۳] Equivalence Class Clustering and bottom-up Lattice Traversal
[۳۴] Minimum Support Threshold
[۳۵] Strength
[۳۶] Confidence

۵-۳-۲- معیارهای ارزیابی قوانین انجمن

در استخراج قوانین انجمنی، معمولاً از چندین معیار برای ارزیابی کیفیت و اهمیت قوانین کشف شده استفاده می‌شود. این معیارها را می‌توان برای ارزیابی کیفیت و اهمیت قوانین مرتبط و انتخاب مناسب‌ترین قوانین برای یک کاربرد خاص مورداستفاده قرارداد.

تفسیر نتایج معیارهای استخراج قواعد انجمنی مستلزم درک معنا و مفاهیم هر معیار و همچنین نحوه استفاده از آن‌ها برای ارزیابی کیفیت و اهمیت قوانین مرتبط کشف شده است. در اینجا چند دستورالعمل برای تفسیر نتایج معیارهای استخراج قانون انجمنی اصلی آورده شده است.

- **Support:** پشتیبانی معیاری است که نشان می‌دهد یک مورد یا مجموعه موارد به‌دفعات در مجموعه‌داده ظاهر می‌شود و به شکل زیر محاسبه می‌گردد. پشتیبانی زیاد نشان می‌دهد که یک مورد یا مجموعه موارد در مجموعه‌داده مشترک است، درحالی‌که پشتیبانی کم نشان‌دهنده نادر بودن آن است.

$$Support(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Total\ number\ of\ transactions}$$

- **Confidence:** معیاری برای سنجش قدرت ارتباط بین دو مورد است. به‌عنوان تعداد رکوردهای حاوی هر دو مورد تقسیم بر تعداد رکوردهای حاوی اولین مورد محاسبه می‌شود. اطمینان بالا نشان می‌دهد که وجود مورد اول یک پیش‌بینی‌کننده قوی برای حضور مورد دوم است.

$$Confidence(\{X\} \rightarrow \{Y\}) = \frac{Transactions\ containing\ both\ X\ and\ Y}{Transactions\ containing\ X}$$

- **Lift:** این معیار اندازه‌گیری قدرت ارتباط بین دو مورد با درنظرگرفتن فراوانی هر دو مورد در مجموعه‌داده است. به‌عنوان معیار اطمینان تقسیم بر معیار پشتیبانی مورد دوم محاسبه می‌شود. این معیار، برای مقایسه قدرت ارتباط بین دو مورد باقدرت مورد انتظار انجمن درصورتی‌که موارد مستقل باشند، استفاده می‌شود. مقدار بیشتر از ۱ نشان می‌دهد که ارتباط بین دو مورد قوی‌تر از حد انتظار بر اساس فراوانی اقلام است. این نشان می‌دهد که این ارتباط ممکن است معنی‌دار باشد و ارزش بررسی بیشتر را داشته باشد. مقدار کمتر از ۱ نشان می‌دهد که ارتباط ضعیف‌تر از حد انتظار است و احتمالاً کمتر قابل‌توجه باشد.

$$Lift(\{X\} \rightarrow \{Y\}) = \frac{(Transactions\ containing\ both\ X\ and\ Y)/(Transactions\ containing\ X)}{Fraction\ of\ transactions\ containing\ Y}$$

۵-۴- پیاده‌سازی الگوریتم‌های یادگیری ماشین

۵-۴-۱- آماده‌سازی مجموعه‌داده برای پیاده‌سازی الگوریتم‌های یادگیری ماشین

ازآنجایی‌که ستون‌هایی که واریانس کمی دارند در مدل‌های پیش‌بینی تأثیر چندانی ندارند، برای کاهش زمان پردازش مدل، از مجموعه‌داده حذف شدند. همچنین ستون‌های تاریخ و زمان سفارش و تاریخ و زمان

حمل‌ونقل به علت اینکه نوع داده‌های آن‌ها از نوع عددی[۳۷] نمی‌باشند نیز از مجموعه‌داده حذف گردیدند. زیرا الگوریتم‌های یادگیری ماشین تنها روی داده‌های عددی قابل‌اجرا هستند.

برای تبدیل سایر ستون‌های غیرعددی طبقه‌بندی شده[۳۸]، به ستون‌های عددی از دو روش کدگذاری طبقه‌ای[۳۹] استفاده شده است. کدگذاری طبقه‌ای، فرایندی است که در آن داده‌های طبقه‌بندی به داده‌های عددی تبدیل می‌شوند. تکنیک‌های کدگذاری طبقه‌بندی زیادی وجود دارد، در این مطالعه، ما از دو روش کدگذاری وان‌هات[۴۰] و کدگذاری هَش[۴۱] استفاده کرده‌ایم.

کدگذاری وان‌هات، تکنیکی است که تمام عناصر یک ستون طبقه‌بندی را به ستون‌های جدیدی تبدیل می‌کند که با ۰ یا ۱ نشان داده می‌شوند تا وجود مقدار دسته را نشان دهد. جدول زیر مثالی از فرایند این روش را نشان می‌دهد.

جدول (۵-۱) فرایند کدگذاری وان‌هات

Original categorical column

Origin
USA
Japan
Europe
USA
Europe

One-Hot encoded columns

Origin_USA	Origin_Japan	Origin_Europe
1	0	0
0	1	0
0	0	1
1	0	0
0	0	1

کدگذاری هش، مبتنی بر تابع هش انجام می‌شود و داده‌های طبقه‌بندی را به عددی تبدیل می‌کند. مزیت اصلی استفاده از کدگذاری هش این است که می‌توان تعداد ستون‌های عدد مطلوب را کنترل کرد. این کار باعث می‌شود تا هنگام پردازش مجموعه‌داده توسط مدل‌های یادگیری ماشین، میزان حافظه[۴۲] کمتری اشغال گردد و مدل در زمان کمتری اجرا شود.

در این مطالعه، ما فرایند کدگذاری وان‌هات را روی ستون‌های نوع معامله انجام شده، وضعیت تحویل سفارش‌ها، کشور محل سفارش، بازار محل تحویل سفارش و حالت حمل‌ونقل پیاده‌سازی کرده‌ایم و این

---

[۳۷] Numerical
[۳۸] Categorical
[۳۹] Categorical Encoding
[۴۰] One-Hot Encoding (OHE)
[۴۱] Hash Encoding
[۴۲] Memory

ستون‌های غیرعددی را به مقادیر عددی تبدیل کرده‌ایم. همچنین کدگذاری هش را روی ستون‌های شهر محل سفارش، بخش‌بندی مشتریان، ایالت محل سفارش، شهر مقصد سفارش، کشور مقصد سفارش، منطقه مقصد سفارش، ایالت مقصد سفارش و وضعیت سفارش پیاده‌سازی کرده‌ایم و تمام ستون‌های غیرعددی را به ستون‌های عددی تبدیل کرده‌ایم.

۵-۴-۲- تقسیم مجموعه‌داده به سه قسمت آموزشی[۴۳]، اعتبارسنجی[۴۴] و تست[۴۵]

پس از آماده‌سازی مجموعه‌داده برای پیاده‌سازی الگوریتم‌های یادگیری ماشین، مجموعه‌داده را به سه قسمت آموزشی، اعتبارسنجی و تست تقسیم کرده‌ایم. ۷۰ درصد مجموعه‌داده، به‌عنوان مجموعه‌داده آموزشی، ۱۰ درصد برای اعتبارسنجی اولیه هر مدل و ۲۰ درصد از مجموعه‌داده به‌عنوان مجموعه‌داده تست در نظر گرفته شده‌اند که این مجموعه‌داده برای اندازه‌گیری صحت و دقت مدل‌ها کنار گذاشته می‌شود و اصطلاحاً توسط مدل‌ها دیده نمی‌شود و مدل در نهایت روی آن تست می‌گردد.

۵-۴-۳- پیاده‌سازی الگوریتم‌های یادگیری ماشین برای پیش‌بینی میزان فروش هر سفارش

ازآنجایی‌که متغیر هدف ما متغیر پیوسته می‌باشد، برای پیش‌بینی میزان فروش هر سفارش، از الگوریتم‌های رگرسیون با نظارت زیر استفاده شده است.

۵-۴-۳-۱- رگرسیون خطی[۴۶]

رگرسیون خطی، نوعی از الگوریتم یادگیری ماشین با نظارت است که رابطه خطی بین یک متغیر وابسته و یک یا چند ویژگی مستقل را محاسبه می‌کند. هنگامی که تعداد ویژگی مستقل، ۱ باشد، به آن رگرسیون خطی تک‌متغیره و در مورد بیش از یک ویژگی، به‌عنوان رگرسیون خطی چندمتغیره شناخته می‌شود. هدف این الگوریتم، یافتن بهترین معادله خطی است که بتواند مقدار متغیر وابسته را بر اساس متغیرهای مستقل پیش‌بینی کند. این الگوریتم، معادله یک خط مستقیم را ارائه می‌دهد که نشان‌دهنده رابطه بین متغیرهای وابسته و مستقل است. در معادله ارائه شده، شیب‌خط نشان می‌دهد که متغیر وابسته برای یک واحد در متغیر(های) مستقل چقدر تغییر می‌کند.

رگرسیون خطی در بسیاری از زمینه‌های مختلف از جمله مالی، اقتصاد و روان‌شناسی برای درک و پیش‌بینی رفتار یک متغیر خاص استفاده می‌شود. به‌عنوان‌مثال، در امور مالی، رگرسیون خطی ممکن است برای درک رابطه بین قیمت سهام شرکت و درآمد آن یا برای پیش‌بینی ارزش آتی یک ارز بر اساس عملکرد گذشته آن استفاده شود.

---

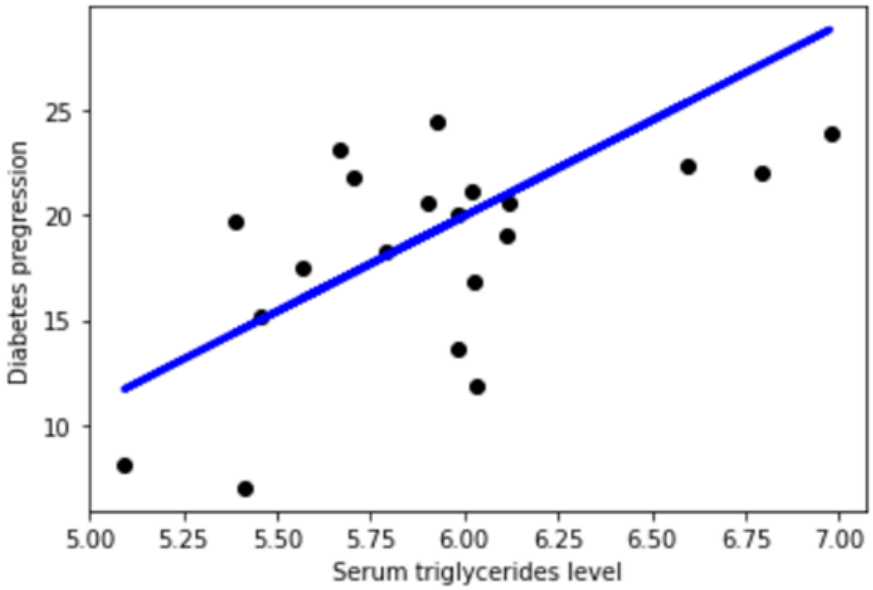
[۴۳] Train Dataset
[۴۴] Validation Dataset
[۴۵] Test Dataset
[۴۶] Linear Regression

مدل رگرسیون خطی، معادله خطی را پیدا می‌کند که به بهترین شکل همبستگی متغیرهای مستقل را با متغیر وابسته توصیف می‌کند. این کار با برازش یک خط به داده‌ها با استفاده از روش حداقل مربعات به دست می‌آید. این خط سعی می‌کند مجموع مجذورهای باقی‌مانده را به حداقل برساند. باقی‌مانده فاصله بین خط و مقدار واقعی متغیر توضیحی است. در زیر نمونه‌ای از معادله رگرسیون خطی حاصل آمده است:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + e$$

در فرمول بالا، y متغیر وابسته و x_1 و x_2 و غیره متغیرهای مستقل هستند. ضرایب b_1 و b_2 و غیره همبستگی متغیرهای مستقل را با متغیر وابسته نشان می‌دهند. علامت ضرایب (+/-) مشخص می‌کند که آیا متغیر همبستگی مثبت یا منفی دارد. b_0 عرض از مبدایی است که مقدار متغیر وابسته را با فرض ۰ بودن همه متغیرهای مستقل نشان می‌دهد.

در تصویر زیر، نمودار یک مدل رگرسیون خطی نشان‌داده‌شده است. این مدل رابطه بین متغیر وابسته، پیشروی دیابت، و متغیر مستقل، سطح تری گلیسیرید سرم[۴۷] را توصیف می‌کند. همان‌طور که مشخص است یک همبستگی مثبت نشان‌داده‌شده است. این مثال یک مدل رگرسیون خطی با دو متغیر را نشان می‌دهد. اگرچه نمی‌توان مدل‌هایی با بیش از سه متغیر را تجسم کرد، اما عملاً یک مدل می‌تواند هر تعداد متغیر داشته باشد.



شکل (۵-۲) مثالی از پیاده‌سازی روش رگرسیون خطی

[۴۷] Serum Triglycerides Level

۵-۴-۳-۲- رگرسیون ستیغی [۴۸]

در مباحث مربوط به رگرسیون چندگانه، تعیین تعداد متغیرهای مستقلی که باید در مدل به کار گرفته شوند، یک مشکل محسوب می‌شود. با افزایش تعداد متغیرها، بیش برازش[۴۹] رخ‌داده و با کاهش آن‌ها نیز ممکن است با مسئله کم برازش[۵۰] مواجه شویم. درصورتی‌که مدل رگرسیونی دچار بیش برازش شود، خطای آن برای برآورد مقدارهای جدید متغیر وابسته زیاد خواهد بود. درحالی‌که وجود متغیرهای کمتر از حد لازم در مدل (کم برازش) واریانس مدل را افزایش می‌دهد؛ بنابراین با افزایش تعداد متغیرها مشکل هم‌خطی و بیش برازش ظاهر شده و با کاهش آن‌ها، واریانس مدل افزایش خواهد یافت. یکی از روش‌های غلبه بر این مسائل در رگرسیون چندگانه، استفاده از مدل رگرسیون ستیغی است. ازآنجایی‌که در زمانی که متغیرهای مدل، زیاد و یا هم‌خطی چندگانه وجود داشته باشد، واریانس برآوردگرها متورم شده و به شکل قله (ستیغ) در می‌آید، از همین روی، این روش رگرسیونی که بر این مشکل غلبه می‌کند، رگرسیون ستیغی نام‌گذاری شده است.

رگرسیون ستیغی، روشی برای تخمین ضرایب مدل‌های رگرسیون چندگانه در سناریوهایی است که متغیرهای مستقل همبستگی بالایی دارند. این الگوریتم در بسیاری از زمینه‌ها از جمله اقتصادسنجی، شیمی و مهندسی استفاده شده است. این روش به‌ویژه برای کاهش مشکل چندخطی بودن در رگرسیون خطی مفید است که معمولاً در مدل‌هایی با تعداد پارامترهای زیاد رخ می‌دهد. به‌طورکلی، این روش کارایی بهبودیافته‌ای را در مسائل تخمین پارامتر در ازای مقدار قابل تحملی سوگیری فراهم می‌کند.

رگرسیون ستیغی به‌عنوان یک راه‌حل ممکن برای عدم دقت برآوردگرهای حداقل مربع توسعه داده شد، زمانی که مدل‌های رگرسیون خطی دارای متغیرهای مستقل چندخطی (بسیار همبسته) هستند، با ایجاد یک برآوردگر رگرسیون ستیغی، یک تخمین دقیق‌تری از پارامترهای برجستگی ارائه می‌دهد، زیرا واریانس آن و برآوردگر میانگین مربع اغلب کوچک‌تر از برآوردگرهای حداقل مربعی است که قبلاً ارائه شده بود.

این الگوریتم، با جریمه‌کردن بزرگیِ ضرایب ویژگی‌ها و به‌حداقل‌رساندن خطا بین مشاهدات پیش‌بینی شده و واقعی کار می‌کند که به آن منظم‌سازی[۵۱] می‌گویند. رگرسیون ستیغی، روش منظم‌سازی L_2 را انجام می‌دهد، یعنی جریمه‌ای معادل مجذور بزرگی ضرایب اضافه می‌کند. هدف این الگوریتم کمینه‌سازی هدف حداقل مربعات به‌اضافه حاصل‌ضرب ضریب ثابت a در مجموع مجذور ضرایب است که رابطه آن به‌صورت زیر است.

Minimization objective = least squares objective + α * (sum of square of coefficients)

---

[۴۸] Ridge Regression
[۴۹] Overfitting
[۵۰] Underfitting
[۵۱] Regularization

۵-۴-۳-۳- رگرسیون لاسو[۵۲]

رگرسیون لاسو در تحلیل‌های چندمتغیره خطی کاربرد ساده‌سازی مسائل به نظر پیچیده را دارد. در واقع با استفاده از رگرسیون لاسو سعی می‌کنیم، روش مناسب برای مدل‌سازی متغیر پاسخ بر اساس کمترین و البته مناسب‌ترین تعداد متغیرهای مستقل را ارائه دهیم. ازآنجایی‌که کلمه لاسو[۵۳] به معنی طناب و کمند انداختن است، این روش رگرسیونی نیز سعی دارد تقریباً با استفاده از کمنداندازی متغیرهای مناسب‌تر را از بقیه متغیرها جدا کرده و مدل ساده‌تری ارائه دهد.

در رگرسیون لاسو (عملگر گزینش و انقباض کمترین قدرمطلق[۵۴]) که توسط رابرت تیب‌شیرانی[۵۵] در سال ۱۹۹۶ معرفی شد، در واقع این روش یعنی کاهش بعد متغیرها به‌کاررفته و از کمینه‌سازی مجموع مربعات تغییریافته استفاده می‌شود. به‌این‌ترتیب با استفاده از یک تابع جریمه[۵۶] روی جمع قدرمطلق ضرایب مدل رگرسیونی، تعداد پارامترها کنترل می‌شود. مجموع مربعات خطای رگرسیونی لاسو به‌صورت زیر نوشته می‌شود.

$$\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j} |\beta_j|$$

دراین‌رابطه λ پارامتر تنظیم‌کننده است، به این معنی که اگر مقدارش برابر با صفر باشد، مدل به رگرسیون عادی تبدیل شده و همه متغیرها در آن حضور خواهند داشت و اگر مقدار آن افزایش یابد تعداد متغیرهای مستقل در مدل کاهش خواهند یافت؛ بنابراین با انتخاب ∞ برای λ عملاً هیچی متغیری در مدل وجود ندارد. تعیین مقدار برای این پارامتر معمولاً توسط روش اعتبارسنجی متقابل[۵۷] انجام می‌شود. همچنین ممکن است به‌جای استفاده از پارامتر λ، شرطی که باعث کنترل تعداد متغیرهای مستقل می‌شود را به‌صورت زیر نوشت که در آن t پارامتر تنظیم‌کننده مدل است. ولی به‌هرحال باید توجه داشت که تعداد پارامترها بر اساس مجموع قدرمطلق ضرایب کنترل می‌شوند.

$$\sum_{j} |\beta_j| < t$$

در زیر، تفاوت بین الگوریتم‌های رگرسیون ستیغی و لاسو قابل‌مشاهده است.

---

[۵۲] Lasso Regression
[۵۳] Lasso
[۵۴] Least Absolute Shrinkage and Selection Operator
[۵۵] Robert Tibshirani
[۵۶] Penalty
[۵۷] Cross Validation

$$RIDGE: \quad RSS + \lambda \sum_{i=1}^{n} \beta_i^2$$

$$LASSO: \quad RSS + \lambda \sum_{i=1}^{n} |\beta_i|$$

که در آن مجموع مربعات باقی‌مانده (انحرافات پیش‌بینی‌شده از روی مقادیر تجربی واقعی) از رابطه زیر به دست می‌آید.

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

۵-۴-۳-۴- رگرسیون درخت تصمیم[۵۸]

روش درخت تصمیم برای دسته‌بندی و پیش‌بینی مقادیر گسسته بکار می‌رود که در بخش‌های بعدی به‌تفصیل به آن پرداخته خواهد شد. به طور خلاصه، فرایند پیاده‌سازی الگوریتم درخت تصمیم از گره ریشه شروع می‌شود و توسط یک درخت منشعب دنبال می‌شود که در نهایت به یک گره برگ منتهی می‌شود که حاوی پیش‌بینی یا نتیجه نهایی الگوریتم است. ساخت درخت تصمیم معمولاً از بالابه‌پایین با انتخاب متغیری در هر مرحله که به بهترین نحو مجموعه موارد را تقسیم می‌کند، انجام می‌شود. هر زیر درخت از مدل درخت تصمیم را می‌توان به‌عنوان یک درخت دودویی نشان داد که در آن یک گره تصمیم بر اساس شرایط به دو گره تقسیم می‌شود. در درخت تصمیم متغیر هدف یا گره پایانی می‌تواند مقادیر پیوسته (معمولاً اعداد واقعی) باشد که به آن، درخت تصمیم رگرسیون گفته می‌شود.

۵-۴-۳-۵- رگرسیون جنگل تصادفی[۵۹]

جنگل تصادفی یک روش یادگیری ترکیبی برای دسته‌بندی و رگرسیون می‌باشد که بر اساس ساختاری متشکل از شمار بسیاری درخت تصمیم، بر روی زمان آموزش و خروجی کلاس‌ها (دسته‌بندی) یا برای پیش‌بینی‌های هر درخت به شکل مجزا، کار می‌کنند. جنگل‌های تصادفی برای درختان تصمیم که در مجموعه آموزشی دچار بیش برازش می‌شوند، مناسب هستند. عملکرد جنگل تصادفی معمولاً بهتر از درخت تصمیم است، اما این بهبود عملکرد تا حدی به نوع داده هم‌بستگی دارد. هنگامی که متغیر هدف پیوسته است از رگرسیون جنگل تصادفی استفاده می‌شود.

---

[۵۸] Decision Tree Regression
[۵۹] Random Forest Regression

۵-۴-۳-۶- رگرسیون بیزی[۶۰]

در رگرسیون خطی بیزی، میانگین یک پارامتر با مجموع وزنی سایر متغیرها مشخص می‌شود. هدف این نوع مدل‌سازی شرطی تعیین توزیع قبلی رگرسیون‌ها و همچنین سایر متغیرهایی است که تخصیص رگرسیون را توصیف می‌کنند و در نهایت امکان پیش‌بینی خارج از نمونه رگرسیون و مشروط به مشاهدات ضرایب رگرسیون را می‌دهد.

هنگامی که مجموعه‌داده دارای داده‌های بسیار کم یا ضعیف است، رگرسیون بیزی ممکن است بسیار مفید باشد. برخلاف روش‌های رگرسیون مرسوم که در آن خروجی تنها از یک عدد از هر ویژگی مشتق می‌شود، خروجی مدل رگرسیون بیزی از توزیع احتمال مشتق می‌شود.

۵-۴-۴- پیاده‌سازی الگوریتم‌های یادگیری ماشین برای پیش‌بینی وجود ریسک ارسال با تأخیر

برای پیش‌بینی وجود ریسک ارسال با تأخیر سفارش از الگوریتم‌های دسته‌بندی با نظارت زیر استفاده شده است. عدد ۱ به معنی وجود ریسک ارسال با تأخیر و عدد ۰ به معنی عدم وجود ریسک ارسال با تأخیر هستند.

۵-۴-۴-۱- درخت تصمیم‌گیری[۶۱]

درخت تصمیم یک مدل سلسله‌مراتبی پشتیبانی تصمیم است که از یک مدل درخت‌مانند از تصمیمات و پیامدهای احتمالی آن‌ها، از جمله نتایج رویدادهای شانسی، هزینه‌های منابع و مطلوبیت استفاده می‌کند. درخت‌های تصمیم معمولاً در تحقیقات عملیاتی، به‌ویژه در تجزیه‌وتحلیل تصمیم‌گیری برای کمک به شناسایی استراتژی که به‌احتمال زیاد به یک هدف می‌رسد، استفاده می‌شوند.

درخت تصمیم دارای اجزای زیر است:

- **گره اصلی**[۶۲]: ویژگی کلیدی در مجموعه‌داده
- **گره داخلی**[۶۳]: گره‌هایی که یک یال ورودی و دو یا چند یال خروجی دارند.
- **گره برگ**[۶۴]: گره پایانی بدون یال خروجی

درخت تصمیم از یک گره اصلی شروع می‌شود و با بررسی شرایط مختلف و اختصاص آن به سایر گره‌ها ادامه می‌یابد. درخت تصمیم زمانی کامل می‌شود که تمام شرایط به یک گره برگ منتهی شوند. گره برگ حاوی برچسب طبقه‌بندی می‌باشد.

---

[۶۰] Bayesian Regression
[۶۱] Decision Tree
[۶۲] Root Node
[۶۳] Internal Node
[۶۴] Leaf Node

برای تقسیم بهینهٔ ویژگی‌ها دو روش وجود دارد:

- **روش شاخص جینی**[۶۵]: ناخالصی جینی تعداد برچسب‌گذاری اشتباه هر عنصر مجموعه‌داده را هنگامی که به طور تصادفی برچسب‌گذاری می‌شود، اندازه‌گیری می‌کند. در شکل زیر فرمول شاخص جینی مشاهده می‌شود که در آن p_j احتمال کلاس j است. حداقل مقدار شاخص جینی ۰ است که زمانی اتفاق می‌افتد که گره خالص باشد، به این معنی که تمام عناصر موجود در گره از یک کلاس منحصربه‌فرد هستند؛ بنابراین، این گره دوباره تقسیم نخواهد شد. تقسیم بهینه توسط ویژگی‌هایی با شاخص جینی کمتر انتخاب می‌شود. علاوه بر این، زمانی که احتمال دو کلاس یکسان باشد، حداکثر مقدار (۰.۵) را دریافت می‌کند.

$$GiniIndex = 1 - \sum_j p_j^2$$

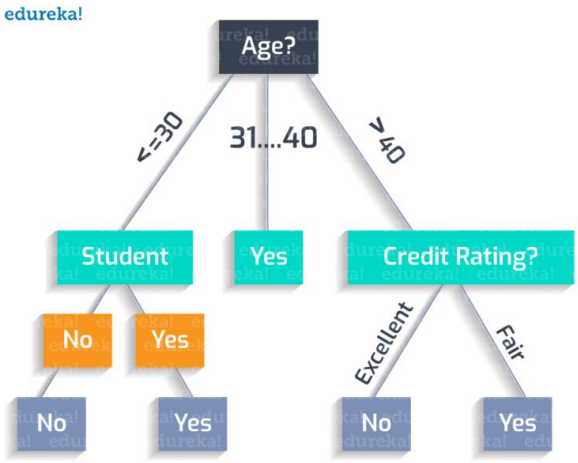
- **روش آنتروپی**[۶۶]: آنتروپی معیاری از اطلاعات است که نشان‌دهنده بی‌نظمی ویژگی‌ها با متغیر هدف است. مشابه شاخص جینی، تقسیم بهینه توسط ویژگی با آنتروپی کمتر انتخاب می‌شود. مقدار آنتروپی زمانی حداکثر مقدار خود (۱) را به دست می‌آورد که احتمال دو کلاس یکسان باشد و هنگامی که یک گره خالص باشد، مقدار آنتروپی حداقل مقدار خود یعنی ۰ است. فرمول محاسبه آنتروپی به شکل زیر است که در آن p_j احتمال کلاس j است.

$$Entropy = -\sum_j p_j \cdot log_2 \cdot p_j$$

شکل زیر نمونه‌ای از یک درخت تصمیم با تقسیم ویژگی‌ها با روش شاخص جینی را نشان می‌دهد که هدف آن پیش‌بینی خرید لپ‌تاپ توسط کاربر می‌باشد. همان‌طور که مشاهده می‌شود، ویژگی سن[۶۷] به‌عنوان گره اصلی انتخاب شده است و سایر ویژگی‌ها در گره‌های داخلی قرار دارند و با بررسی شرایط مختلف، گره‌های برگ مشخص شده‌اند.

---

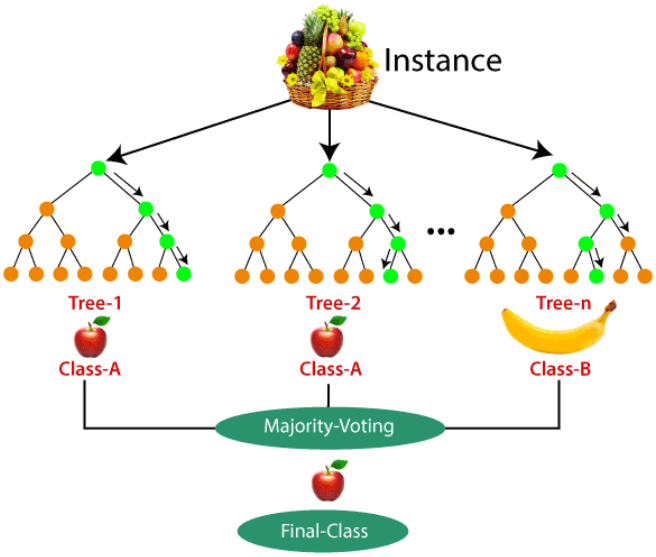
[۶۵] Gini Index
[۶۶] Entropy
[۶۷] Age



شکل (۵-۳) مثالی از پیاده‌سازی الگوریتم درخت تصمیم برای طبقه‌بندی داده‌ها

۵-۴-۴-۲- جنگل تصادفی[۶۸]

جنگل تصادفی یک روش یادگیری ترکیبی برای دسته‌بندی و رگرسیون می‌باشد که بر اساس ساختاری متشکل از شمار بسیاری درخت تصمیم، در زمان آموزش عمل می‌کند. عملکرد الگوریتم جنگل تصادفی معمولاً بهتر از الگوریتم درخت تصمیم است، اما این بهبود عملکرد تا حدی به نوع داده هم‌بستگی دارد. برای کاربرد دسته‌بندی، خروجی جنگل تصادفی، کلاسی است که توسط اکثر درختان انتخاب شده است. شکل زیر مثالی ساده از پیاده‌سازی الگوریتم جنگل تصادفی را بر روی نمونه‌ای از میوه‌ها نشان می‌دهد.



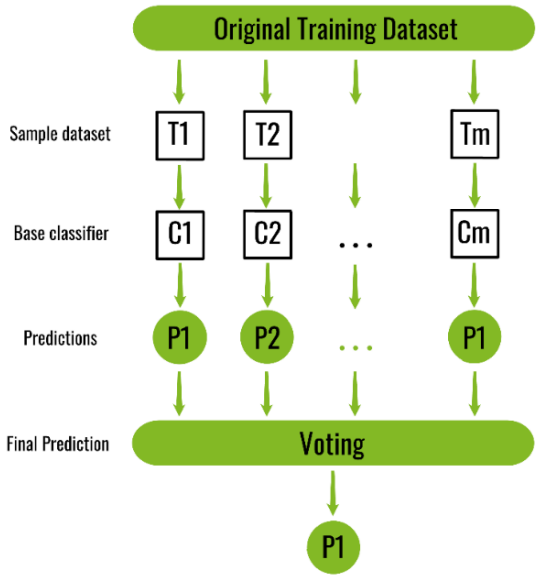
شکل (۵-۴) مثالی از پیاده‌سازی الگوریتم جنگل تصادفی برای دسته‌بندی داده‌ها

---

[۶۸] Random Forest

۵-۴-۴-۳- دسته‌بندی کیسه‌ای[۶۹]

الگوریتم دسته‌بندی کیسه‌ای یک فرا برآوردگر[۷۰] ترکیبی[۷۱] است که هر کدام از طبقه‌بندی‌کننده‌های پایه را بر روی زیرمجموعه‌های تصادفی مجموعه‌داده اصلی قرار می‌دهد و سپس پیش‌بینی‌های فردی آن‌ها (چه با رأی‌گیری[۷۲] یا با میانگین‌گیری[۷۳]) را جمع‌آوری می‌کند تا یک پیش‌بینی نهایی را تشکیل دهد. چنین فرا برآوردگر معمولاً می‌تواند به‌عنوان راهی برای کاهش واریانس تخمین‌گر جعبه سیاه[۷۴] (به‌عنوان‌مثال، درخت تصمیم)، با ورود تصادفی به مراحل ایجاد آن و سپس ساختن مجموعه‌ای از آن استفاده شود.



شکل (۵-۵) نحوه عملکرد الگوریتم دسته‌بندی کیسه‌ای

۵-۴-۴-۴- دسته‌بندی تقویتی گرادیان[۷۵]

طبقه‌بندی‌کننده تقویتی گرادیان یک الگوریتم یادگیری ماشین است که بسیاری از مدل‌های یادگیری ضعیف را با هم ترکیب می‌کند تا یک مدل پیش‌بینی قوی ایجاد کند. معمولاً هنگام انجام الگوریتم دسته‌بندی تقویتی گرادیان از درختان تصمیم استفاده می‌شود. مدل تقویتی گرادیان به دلیل اثربخشی در طبقه‌بندی مجموعه‌داده‌های پیچیده، محبوب شده‌اند.

---

[۶۹] Bagging Classifier
[۷۰] Meta-Estimator
[۷۱] Ensemble
[۷۲] Voting
[۷۳] Averaging
[۷۴] Black-Box Estimator
[۷۵] Gradient Boosting Classifier Algorithm

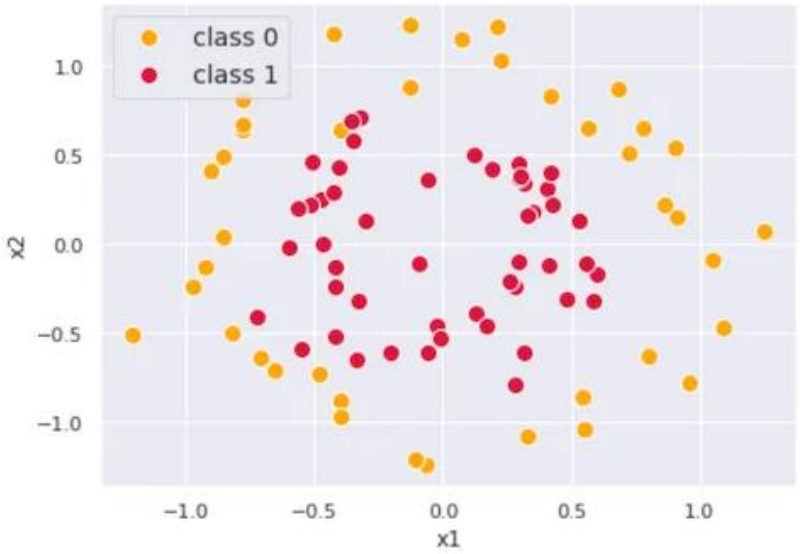
مدل تقویت گرادیان ترکیبی خطی از یک سری مدل‌های ضعیف است که به‌صورت تناوبی برای ایجاد یک مدل نهایی قوی ساخته شده است. این روش به خانواده الگوریتم‌های یادگیری گروهی تعلق دارد و عملکرد آن همواره از الگوریتم‌های اساسی یا ضعیف (مثلاً درخت تصمیم) یا روش‌های بر اساس ریسه‌گذاری (مانند جنگل تصادفی) بهتر است؛ اما این موضوع تا حدی از مشخصات داده‌های ورودی تأثیر می‌پذیرد.

روش این الگوریتم بدین ترتیب است که تابع هزینه[۷۶] را به کمینه‌ترین مقدار خود برساند. در علم آمار، معمولاً تابع هزینه برای اینکه مشخص شود تخمین پارامترمان تا چه حد موفق بوده، استفاده می‌شود. تابعی است که برای سنجش میزان موفقیت تخمین‌گر از تخمین پارامتر نسبت به مقادیر واقعی از آن استفاده می‌شود. در مسائل طبقه‌بندی، تابع هزینه در اصل به‌نوعی تعداد طبقه‌بندی‌های اشتباه توسط تخمین‌گر را نمایان می‌کند.

الگوریتم یادگیری تقویتی یک الگوریتم تقویتی قدرتمند است که چندین یادگیرنده ضعیف را با یادگیرندگان قوی ترکیب می‌کند که در آن هر مدل جدید برای به‌حداقل‌رساندن تابع هزینه مانند میانگین مربعات خطایا آنتروپی متقابل مدل قبلی، با استفاده از گرادیان نزول آموزش داده می‌شود. در هر تکرار، الگوریتم گرادیان تابع هزینه را باتوجه‌به پیش‌بینی‌های مجموعه فعلی محاسبه می‌کند و سپس یک مدل ضعیف جدید را برای به‌حداقل‌رساندن این گرادیان آموزش می‌دهد. سپس پیش‌بینی‌های مدل جدید به مجموعه اضافه می‌شود و این فرایند تا زمانی که یک معیار توقف برآورده شود، تکرار می‌شود. در این الگوریتم، وزن نمونه‌های آموزشی بهینه‌سازی نشده است، در عوض، هر پیش‌بینی‌کننده با استفاده از خطاهای باقی‌مانده قبلی به‌عنوان برچسب آموزش داده می‌شود.

در ادامه روند پیاده‌سازی مدل تقویتی گرادیان در قالب یک مثال آموزشی توضیح داده خواهد شد. شکل زیر نمایش داده‌های دسته‌بندی در شکل زیر نشان‌داده‌شده است.

[۷۶] Loss Function

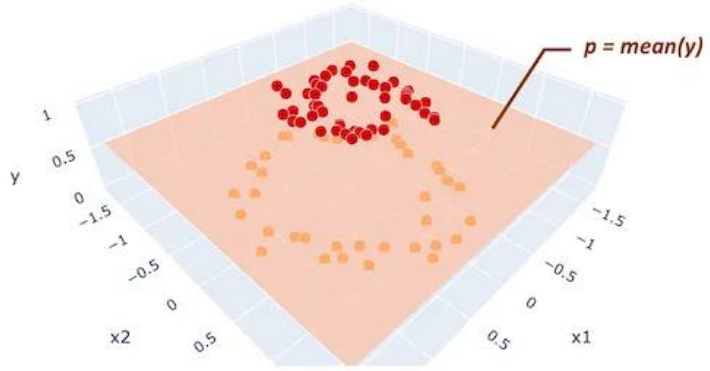


شکل (۵-۶) مثالی از مسئله دسته‌بندی دو کلاسه

هدف ساخت یک مدل تقویتی گرادیان است که داده‌ها را به دودسته دسته‌بندی کند. اولین گام، ایجاد یک پیش‌بینی یکنواخت بر روی احتمال کلاس ۱ (ما آن را p می‌نامیم) برای تمام نقاط داده است که در واقع همان میانگین کلاس می‌باشد.

$$p = P(y = 1) = \bar{y}$$

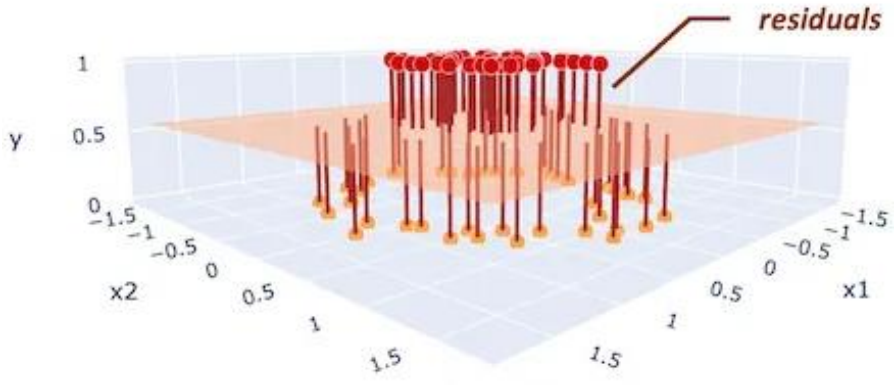
در اینجا یک نمایش سه‌بعدی از داده‌ها و پیش‌بینی اولیه آمده است. در این لحظه، پیش‌بینی فقط صفحه‌ای است که همیشه مقدار یکنواخت p = mean(y) را در محور y دارد.



شکل (۵-۷) نمایش صفحه پیش‌بینی به شکل سه‌بعدی

در این مثال، میانگین y، ۰.۵۶ است. ازآنجایی‌که بزرگ‌تر از ۰.۵ است، همه چیز با این پیش‌بینی اولیه در کلاس ۱ طبقه‌بندی می‌شود. ممکن است به نظر برسد که این پیش‌بینی ارزش یکسان، منطقی نیست، لازم به ذکر است که با اضافه‌کردن مدل‌های ضعیف بیشتر به آن، پیش‌بینی بهبود می‌یابد.

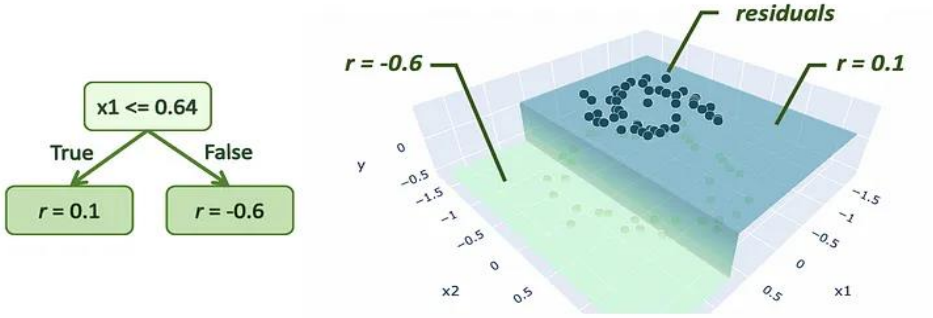
برای بهبود کیفیت پیش‌بینی، ممکن است روی باقی‌مانده‌ها (خطای پیش‌بینی) از پیش‌بینی اولیه تمرکز کنیم، زیرا این همان چیزی است که باید به حداقل برسد. باقی‌مانده‌ها به‌صورت $r_i = y_i - p$ تعریف می‌شوند (i نشان‌دهنده شاخص هر نقطه داده است). در شکل زیر باقیمانده‌ها به‌صورت خطوط قهوه‌ای نشان داده شده‌اند که خطوط عمود از هر نقطه داده به صفحه پیش‌بینی هستند.



شکل (۵-۸) شکل نمایش باقی‌مانده‌ها

برای به‌حداقل‌رساندن این باقی‌مانده‌ها، یک مدل درخت رگرسیون با x_1 و x_2 به‌عنوان ویژگی‌های آن و باقی‌مانده r به‌عنوان هدف آن باید ساخته شود. اگر بتوان درختی ساخت که الگوهایی را بین x و r پیدا کند، می‌توان با استفاده از آن الگوهای یافت شده، باقی‌مانده‌های حاصل از پیش‌بینی اولیه p را کاهش داد.

برای ساده‌کردن نمایش، درختان بسیار ساده‌ای که هر کدام فقط دارای یک تقسیم و دو گره برگ هستند، ساخته شده‌اند که به آن «استامپ[۷۷]» می‌گویند. لازم به ذکر است که درخت‌های تقویت‌کننده گرادیان معمولاً درختان کمی عمیق‌تر مانند درخت‌هایی با ۸ تا ۳۲ گره برگ دارند. در اینجا اولین درخت ایجاد شده باقی‌مانده‌ها را با دو مقدار مختلف $r = \{۰،۱,-۰،۶\}$ پیش‌بینی می‌کند.



شکل (۵-۹) درخت ساخته شده برای متغیرهای x و باقی‌مانده r

[۷۷] Stump

در ادامه گاما طبق فرمول زیر محاسبه می‌شود. مقادیر گاما را به پیش‌بینی اولیه خود اضافه می‌کنیم تا باقی‌مانده‌ها را کاهش دهیم.

$$\gamma_j = \frac{\sum_{x_i \in R_j} (y_i - p)}{\sum_{x_i \in R_j} p(1-p)}$$

γ is computed for each terminal node j

Aggregating for all the data points x_i that belongs to terminal node j

مقادیر گاما ۱ و گاما ۲ بدین ترتیب محاسبه می‌شوند.

$$\gamma_1 = \frac{\sum_{x_i \in R_1} (y_i - 0.56)}{\sum_{x_i \in R_1} 0.56 \cdot (1 - 0.56)} = 0.3$$

$$\gamma_2 = \frac{\sum_{x_i \in R_2} (y_i - 0.56)}{\sum_{x_i \in R_2} 0.56 \cdot (1 - 0.56)} = -2.2$$

برای اینکه گاما را به مقدار p اضافه شود، به شکل زیر عمل می‌کنیم. ابتدا مقدار log(odds) را از p به دست می‌آوریم (به آن F(x) گفته می‌شود). سپس گاما را به آن اضافه می‌کنیم.

$$log(odds) = log\left(\frac{p}{1-p}\right)$$

برای اینکه مدل بیش از حد از آموزش[۷۸] نبیند و خطای آن کاهش یابد می‌توان مقدار گاما را با یک‌وزنی (بین ۰ تا ۱) که به آن نرخ یادگیری[۷۹] ν گفته می‌شود، ضرب کرد و سپس به مقدار log(odds) یا همان F(x) اضافه نمود تا پیش‌بینی بروز شود.

$$F_1(x) = F_0(x) + \nu \cdot \gamma$$

Updated prediction — Initial prediction — Learning rate

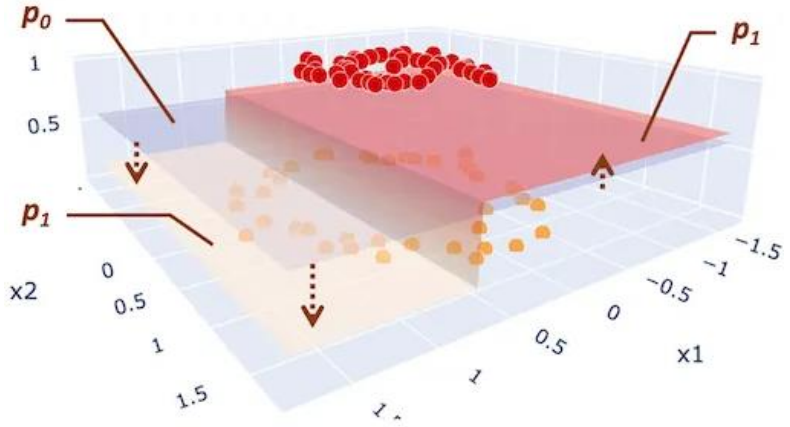
در این مثال، ما از نرخ یادگیری نسبتاً بزرگ $\nu = ۰,۹$ استفاده می‌کنیم تا فرایند بهینه‌سازی را آسان‌تر درک کنیم، اما معمولاً قرار است مقادیر بسیار کوچک‌تری مانند ۰.۱ در نظر گرفته شود. با جایگزینی مقادیر واقعی برای متغیرهای سمت راست معادله بالا، پیش‌بینی به‌روز $F_1(x)$ را به دست می‌آید.

---

[۷۸] Overfit
[۷۹] Learning Rate

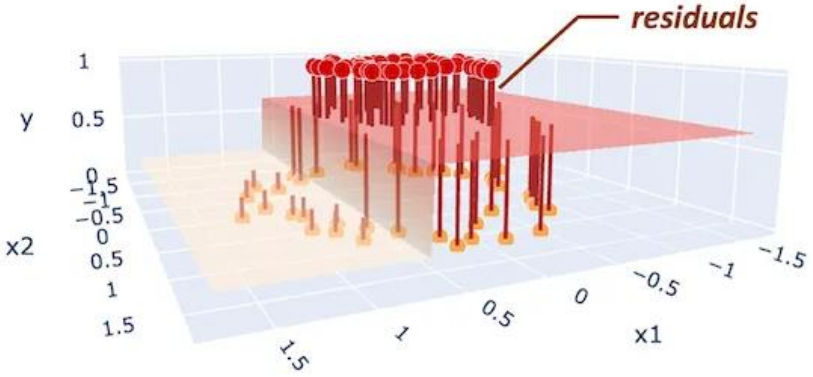
$$F_1(x) = \begin{cases} log\left(\frac{0.56}{1-0.56}\right) + 0.9 \cdot 0.3 = 0.5 & if\ x_1 \leq 0.64 \\ log\left(\frac{0.56}{1-0.56}\right) - 0.9 \cdot 2.2 = -1.7 & otherwise \end{cases}$$

اگر log(odds) را دوباره به p تبدیل کنیم. شکی پله‌مانند از داده‌ها مانند شکل زیر به دست می‌آید.



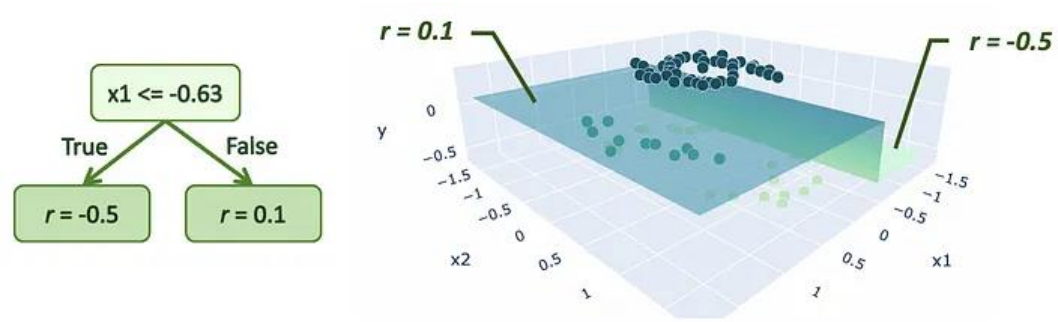
شکل (۵-۱۰) نمایش پیش‌بینی بروز شده

اکنون، باقی‌مانده‌های بروز شده r به شکل زیر است.



شکل (۵-۱۱) نمایش باقی‌مانده‌های بروز شده

مجدداً یک درخت رگرسیون با استفاده از همان x_1 و x_2 به‌عنوان ویژگی‌های ورودی برای باقی‌مانده‌های بروز شده ایجاد می‌کنیم.



شکل (۵-۱۲) درخت ساخته شده برای متغیرهای x و باقی‌مانده r بروز شده

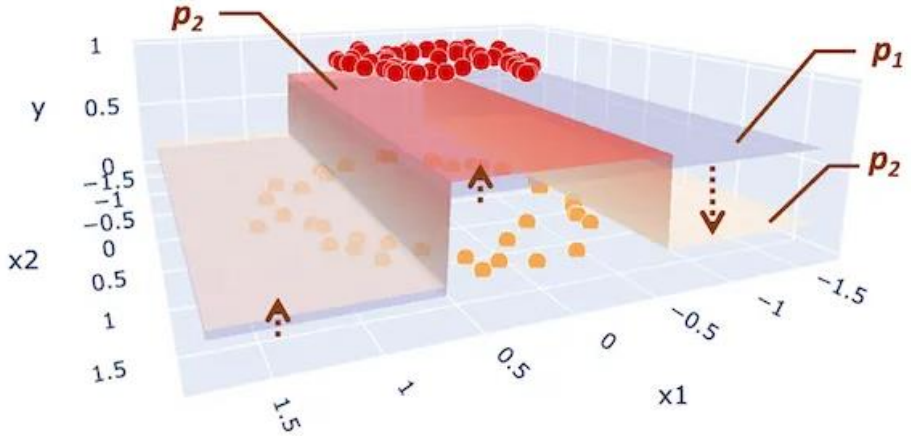
حال مجدداً گاما به همان روش قبل محاسبه کرده و $F_2(x)$ را به دست می‌آوریم.

$$F_2(x) = \begin{cases} F_1(x) - \nu \cdot 2.3 = 0.5 - 0.9 \cdot 2.3 = -1.6 & if\ x_1 \le -0.63 \\ F_1(x) + \nu \cdot 0.4 = 0.5 + 0.9 \cdot 0.4 = 0.9 & else\ if\ -0.63 < x_1 \le 0.64 \\ F_1(x) + \nu \cdot 0.4 = -1.7 + 0.9 \cdot 0.4 = -1.3 & otherwise \end{cases}$$

These are γ computed with this formula:

$$\gamma_j = \frac{\sum_{x_i \in R_j} (y_i - p)}{\sum_{x_i \in R_j} p(1-p)}$$

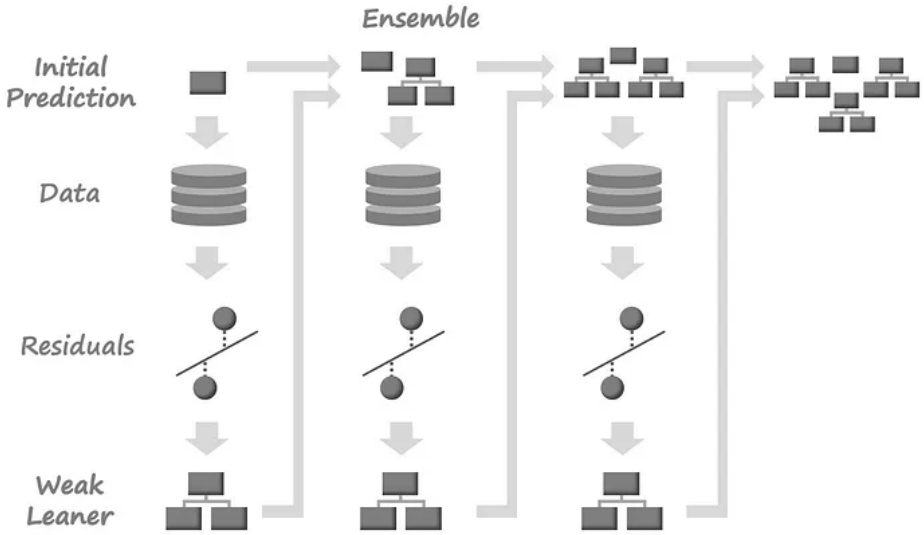
سپس $F_2(x)$ را به $p_2(x)$ تبدیل می‌کنیم و به شکل زیر می‌رسیم.



شکل (۵-۱۳) نمایش پیش‌بینی بروز شده

سپس، این مراحل را تکرار می‌کنیم تا زمانی که پیش‌بینی مدل متوقف شود. می‌توان دید که پیش‌بینی ترکیبی $p(x)$ به هدف ما نزدیک‌تر می‌شود؛ زیرا درخت‌های بیشتری را به مدل ترکیبی اضافه می‌کنیم. این روشی است که الگوریتم تقویتی گرادیان برای پیش‌بینی اهداف پیچیده با ترکیب چندین مدل ضعیف انجام می‌دهد.

تصویر زیر به طور خلاصه کل فرایند این الگوریتم را نشان می‌دهد.



شکل (۵-۱۴) فرایند الگوریتم تقویتی گرادیان

۵-۴-۴-۵- دسته‌بندی XGboost [۸۰]

XGBoost یک الگوریتم یادگیری ماشین مبتنی بر درخت تصمیم است که از یک چارچوب تقویت گرادیان استفاده می‌کند. شکل زیر فرایند بهینه‌سازی الگوریتم ماشین تقویت گرادیان[۸۱] را توسط XGboost نشان می‌دهد.

شکل (۵-۱۵) نحوه بهینه‌سازی در الگوریتم XGboost

[۸۰] XGboost Classifier
[۸۱] Gradient Boosting Machines (GBMs)

XGBoost مخفف واژه تقویت گرادیان شدید[۸۲] است و به دلیل توانایی آن در مدیریت مجموعه‌داده‌های بزرگ و توانایی آن برای دستیابی به عملکرد پیشرفته در بسیاری از وظایف یادگیری ماشین، به یکی از محبوب‌ترین و پرکاربردترین الگوریتم‌های یادگیری ماشین تبدیل شده است.

در این الگوریتم درخت‌های تصمیم به‌صورت متوالی ایجاد می‌شوند. وزن‌ها نقش مهمی در XGBoost دارند. وزن‌ها به همه متغیرهای مستقل اختصاص داده می‌شوند که سپس به درخت تصمیم که نتایج را پیش‌بینی می‌کند، وارد می‌شوند. وزن متغیرهای پیش‌بینی‌شده اشتباه توسط درخت، افزایش می‌یابد و این متغیرها سپس به درخت تصمیم دوم تغذیه می‌شوند. سپس این طبقه‌بندی‌کننده‌ها یا پیش‌بینی‌کننده‌های منفرد برای ارائه یک مدل قوی و دقیق‌تر جمع می‌شوند.

۵-۵- اعتبارسنجی و بررسی صحت الگوریتم‌های رگرسیون و دسته‌بندی

پس از پیاده‌سازی و آموزش مدل‌ها روی داده‌های آموزشی، اعتبارسنجی الگوریتم‌ها با داده‌های اعتبارسنجی انجام شد. در ادامه به معیارهای اعتبارسنجی الگوریتم‌ها می‌پردازیم.

۵-۵-۱- میانگین مربعات خطا[۸۳]

یکی از معیارهای ارزیابی الگوریتم‌های رگرسیون می‌باشد. فرمول آن به‌صورت زیر است.

$$\mathrm{MSE}(Y,\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n} {e_i}^2$$

در واقع در این معیار ارزیابی، خطاها به توان دو رسیده، با هم جمع می‌شوند و بر تعداد نقاط تقسیم می‌شوند. نکته‌ای که در مورد این معیار وجود دارد، وزن‌دهی بیشتر به خطاهای بزرگ‌تر است، به‌طوری‌که اگر دو داده با خطاهای ۱ و ۳ وجود داشته باشد، اثرگذاری داده دوم بیشتر خواهد بود.

۵-۵-۲- جذر میانگین مربعات خطا[۸۴]

معیار جذر میانگین مربعات خطا، یک معیار ارزیابی رگرسیون است. با استفاده از رابطه زیر از میانگین مربعات خطا محاسبه می‌شود.

$$RMSE(Y,\hat{Y}) = \sqrt{MSE(Y,\hat{Y})}$$

دلیل استفاده از جذر میانگین مربعات خطا، یکسان بودن بُعد و مقیاس آن با ویژگی هدف است. برای مثال، اگر یک مدل برای پیش‌بینی وزن افراد برحسب Kg ایجاد کرده باشیم، واحد میانگین مربعات خطا

---

[۸۲] Extreme Gradient Boosting
[۸۳] MSE: Mean Squared Error
[۸۴] RMSE: Root Mean Squared Error

Kg^2 خواهد بود، درحالی‌که که واحد جذر مربعات خطا Kg است. این معیار اغلب در گزارش نتایج استفاده می‌شود و به‌عنوان تابع هزینه استفاده نمی‌شود.

۵-۵-۳- میانگین قدرمطلق خطا[۸۵]

در این معیار ارزیابی رگرسیون، به‌جای به توان ۲ رساندن خطاها، از تابع قدرمطلق استفاده می‌شود و به‌صورت زیر محاسبه می‌شود.

$$\mathrm{MAE}(Y,\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right| = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

۵-۵-۴- ضریب امتیاز R^2

این معیار برخلاف ۳ معیار قبلی، با افزایش، دقت بالای مدل را نشان می‌دهد. برای محاسبه آن به شکل زیر عمل می‌کنیم.

$$R^2(Y,\widehat{Y}) = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} = 1 - \frac{MSE(Y,\widehat{Y})}{\sigma^2(Y)}$$

این معیار همواره عددی کوچک‌تر از ۱ است. اگر مدلی همواره میانگین ویژگی هدف را در خروجی تولید کند، مقدار این معیار برابر ۰ خواهد بود. ضریب تعیین معمولاً به‌صورت درصد بیان می‌شود. از ضریب تعیین تنها برای مقایسه مدل‌ها و گزارش نتایج استفاده می‌شود.

۵-۵-۵- مربع R تنظیم شده[۸۶]

این معیار ارزیابی رگرسیون، نوع بهبودیافته معیار ارزیابی مربع R به‌حساب می‌آید. مشکلی که روش قبل دارد این است که با افزایش ویژگی‌های مدل، مقدار R^2 نیز افزایش می‌یابد که البته باعث به‌وجودآمدن مدل خوبی نیز می‌شود. اما مربع R تنظیم شده برای حل این مشکل ارائه شده است. این روش فقط ویژگی‌هایی را در نظر می‌گیرد که برای مدل مهم هستند و براین‌اساس می‌تواند بهبود واقعی مدل را نشان دهد. همچنین، معیار R تنظیم شده همیشه از معیار R^2 کمتر است. معادله این معیار به شکل زیر است.

$$R^2adjusted = 1 - frac(1-R^2)(N-1)N-P-1$$

در رابطه فوق، R^2 نشان‌دهنده تعداد نمونه‌های مربع R، آرگومان p نشان‌دهنده تعداد پیش‌بینی‌ها و N اندازه کل نمونه‌ها هستند.

---

[۸۵] MAE: Mean Absolute Error
[۸۶] Adjusted R squared

۵-۵-۶- ماتریس اغتشاش[۸۷]

برای اینکه بتوانیم نتایج دسته‌بندی الگوریتم را با داده‌های واقعی مقایسه کنیم، از ماتریس اغتشاش استفاده می‌کنیم. جدول زیر، ماتریس اغتشاش را نشان می‌دهد.

جدول (۵-۲) ماتریس اغتشاش

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

سطرها مقادیر واقعی و ستون‌ها مقادیر پیش‌بینی شده را نشان می‌دهند. سلول‌های این ماتریس مفاهیم زیر را ارائه می‌دهند.

- **مثبت – صحیح**[۸۸]: نشان می‌دهد که مدل یک نتیجه مثبت را پیش‌بینی کرده است و مشاهده واقعی درست بوده است.
- **مثبت – کاذب**[۸۹]: نشان می‌دهد که مدل یک نتیجه مثبت را پیش‌بینی کرده است، اما مشاهده واقعی نادرست بوده است.
- **منفی – کاذب**[۹۰]: نشان می‌دهد که مدل یک نتیجه منفی را پیش‌بینی کرده است، درحالی‌که مشاهده واقعی درست بوده است.
- **منفی – صحیح**[۹۱]: نشان می‌دهد مدل یک نتیجه منفی را پیش‌بینی کرده است، و نتیجه واقعی نیز نادرست بوده است.

۵-۵-۷- دقت[۹۲]

دقت معمولاً برای قضاوت در مورد عملکرد مدل استفاده می‌شود، فرمول محاسبه دقت به شکل زیر است.

---

[۸۷] Confusion Matrices
[۸۸] True Positive
[۸۹] False Positive
[۹۰] False Negative
[۹۱] True Negative
[۹۲] Accuracy

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

در واقع دقت، میزان پیش‌بینی درست مدل را بر کل محاسبه می‌کند.

۵-۵-۸- صحت[۹۳]

این معیار، اندازه‌گیری مثبت‌های واقعی نسبت به تعداد کل مثبت‌های پیش‌بینی شده توسط مدل را محاسبه می‌کند. در واقع این معیار، میزان مثبت‌بودن پیش‌بینی‌های مثبت مدل را اندازه‌گیری می‌کند.

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

۵-۵-۹- پوشش[۹۴]

معیار پوشش قادر به سنجش مثبت پیش‌بینی شده مدل نسبت به تعداد پیامدهای مثبت واقعی است. با استفاده از این معیار، می‌توان ارزیابی کرد که مدل چقدر قادر به شناسایی نتایج واقعی است.

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

۵-۵-۱۰- امتیاز F۱ [۹۵]

این معیار، میانگین هارمونیک بین دقت و پوشش است. فرمول آن به شکل زیر است.

$$F1\ Score = \frac{۲ * (Precision * Recall)}{Precision + Recall}$$

این معیار هنگامی که داده‌ها به‌صورت نامتوازن پخش شده‌اند، دید بهتری از عملکرد مدل ارائه می‌دهد.

۵-۵-۱۱- اعتبارسنجی متقابل[۹۶]

اعتبارسنجی متقابل یک روش آماری است که برای تخمین عملکرد مدل‌های یادگیری ماشین استفاده می‌شود. این روش برای ارزیابی چگونگی تعمیم نتایج یک تحلیل آماری به یک مجموعه‌داده دیده نشده است. این روش، آموزش بیش از حد مدل را شناسایی می‌کند و با بررسی داده‌های دیده نشده نتیجه دقیق‌تری راجع به عملکرد مدل، ارائه می‌دهد.

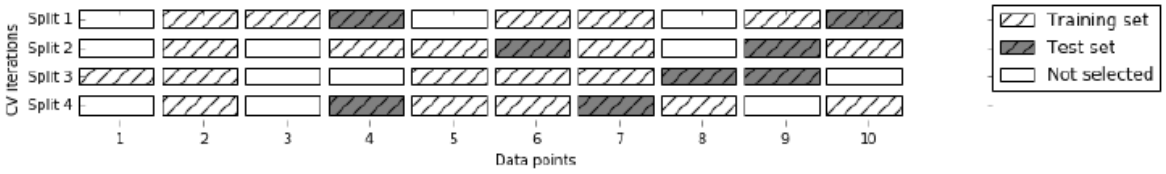
در این پژوهش ما از روش اعتبارسنجی متقابل مونت‌کارلو[۹۷] برای ارزیابی نهایی عملکرد الگوریتم‌ها روی داده‌های دیده نشده، استفاده کرده‌ایم. در ادامه به نحوه عملکرد این معیار اعتبارسنجی می‌پردازیم.

این معیار، یک استراتژی بسیار انعطاف‌پذیر برای اعتبارسنجی متقابل است. در این تکنیک، مجموعه‌داده‌ها به طور تصادفی به مجموعه‌های آموزشی و اعتبارسنجی تقسیم می‌شوند. درصدی از مجموعه داده‌ای را که قرار است به‌عنوان مجموعه آموزشی استفاده شود و درصدی که به‌عنوان مجموعه اعتبارسنجی

---

[۹۳] Precision
[۹۴] Recall
[۹۵] F۱ Score
[۹۶] Cross Validation
[۹۷] Monte Carlo Cross-Validation (Shuffle Split)

استفاده می‌شود، را مشخص می‌کنیم. اگر مجموع درصدها به ۱۰۰ نرسد، از مجموعه‌داده باقی‌مانده استفاده نمی‌شود. سپس این تقسیم‌بندی به تعداد دفعاتی که مشخص می‌کنیم، تکرار می‌شود و دقت هر تکرار محاسبه می‌گردد. می‌توان میانگین نهایی دقت دفعات تکرار را به‌عنوان میزان دقت نهایی مدل روی داده‌های دیده نشده در نظر گرفت.



شکل (۵-۱۶) نحوه عملکرد اعتبارسنجی متقابل مونت‌کارلو

۵-۵-۱۲- منحنی AUC – ROC [۹۸]

این معیار ارزیابی روی داده‌های دیده نشده برای مقایسه نهایی مدل‌ها، پیاده‌سازی شده است. برای درک بهتر این معیار، مفاهیم زیر مطرح می‌گردد.

- **نرخ مثبت صحیح**[۹۹]: یک معیار ارزیابی عملکرد می‌باشد. همان مفهوم معیار پوشش می‌باشد و مشخص می‌کند که به چه نسبتی پیش‌بینی صحیح صورت‌گرفته است. فرمول آن به شکل زیر است.

$$\mathrm{TPR} = \frac{TP}{P} = \frac{TP}{TP + FN}$$

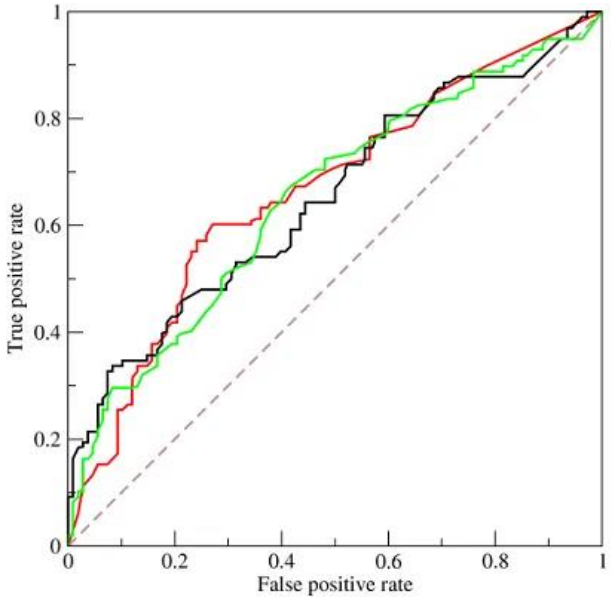
- **نرخ مثبت کاذب**[۱۰۰]: یک معیار ارزیابی عملکرد می‌باشد و نشانگر تعداد شناسایی‌های مثبت از میان مشاهدات منفی است. فرمول آن به شکل زیر است.

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

- **منحنی مشخصه عملکرد**[۱۰۱]: یک منحنی مشخصه عملکرد، یک نمودار برای نمایش توانایی ارزیابی یک سیستم دسته‌بندی دودویی محسوب می‌شود که آستانه تشخیص آن نیز متغیر است. این منحنی توسط ترسیم نرخ مثبت صحیح برحسب نرخ مثبت کاذب، ایجاد می‌شود. نمودار زیر منحنی مشخصه عملکرد را برای سه مدل دسته‌بندی مختلف نشان می‌دهد.

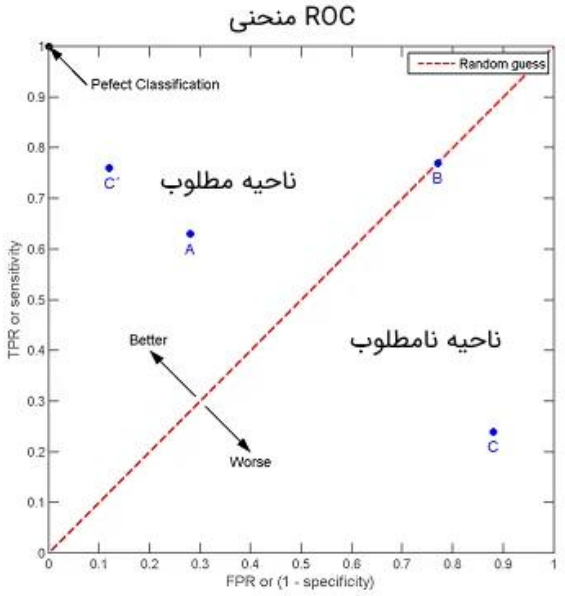
---

[۹۸] AUC (Area Under the Curve) - ROC (Receiver Operating Characteristics) curve
[۹۹] True Positive Rate (TPR)
[۱۰۰] False Positive Rate (FPR)
[۱۰۱] ROC: Receiver Operating Characteristics curve



نمودار (۵-۱) منحنی مشخصه عملکرد برای سه روش مختلف دسته‌بندی

باتوجه‌به نمودار زیر، بهترین عملکرد دسته‌بندی در این نمودار در نقطه‌ای با مختصات (۰,۱) رخ خواهد داد که در آن کمترین نرخ اشتباه و بیشترین نرخ بازیابی یا حساسیت را داریم. این نقطه بیانگر «بهترین دسته‌بندی[۱۰۲]» است.



نمودار (۵-۲): نواحی مطلوب و نامطلوب در منحنی ROC

[۱۰۲] Perfect Classification

همچنین در نمودار فوق، خط منقطعی که از میان نمودار عبور کرده و نقطه (۰,۰) را به (۱,۱) پیوند می‌دهند، حدس تصادفی است که به‌صورت ناحیه ۵۰٪-۵۰٪ نیز شناخته می‌شود. اگر نقطه‌ای روی این خط منقطع قرار گرفته باشد، تشخیص درستی نسبت به قرارگیری در هر گروه، برایش وجود ندارد. در حقیقت در نیمی از موارد می‌تواند در یک دسته و در نیمی از موارد نیز در دسته دیگری، طبقه‌بندی شود و نقشی در تعیین خطا نخواهد داشت. یکی از نمونه‌های معروف برای دسته‌بندی به‌صورت تصادفی، تصمیم تعلق نقطه به هر یک از دو گروه به‌وسیله پرتاب سکه است. هر چه تعداد نمونه‌ها در دسته‌بندی تصادفی بیشتر شود، این خط به قطر نواحی ROC نزدیک‌تر خواهد شد.

AUC: مساحت زیر منحنی مشخصه عملکرد را AUC می‌گویند که نشان‌دهنده درجه یا معیار تفکیک‌پذیری است. این معیار نشان می‌دهد که مدل چقدر می‌تواند بین کلاس‌ها تمایز قائل شود. یک مدل عالی دارای AUC = ۱ است

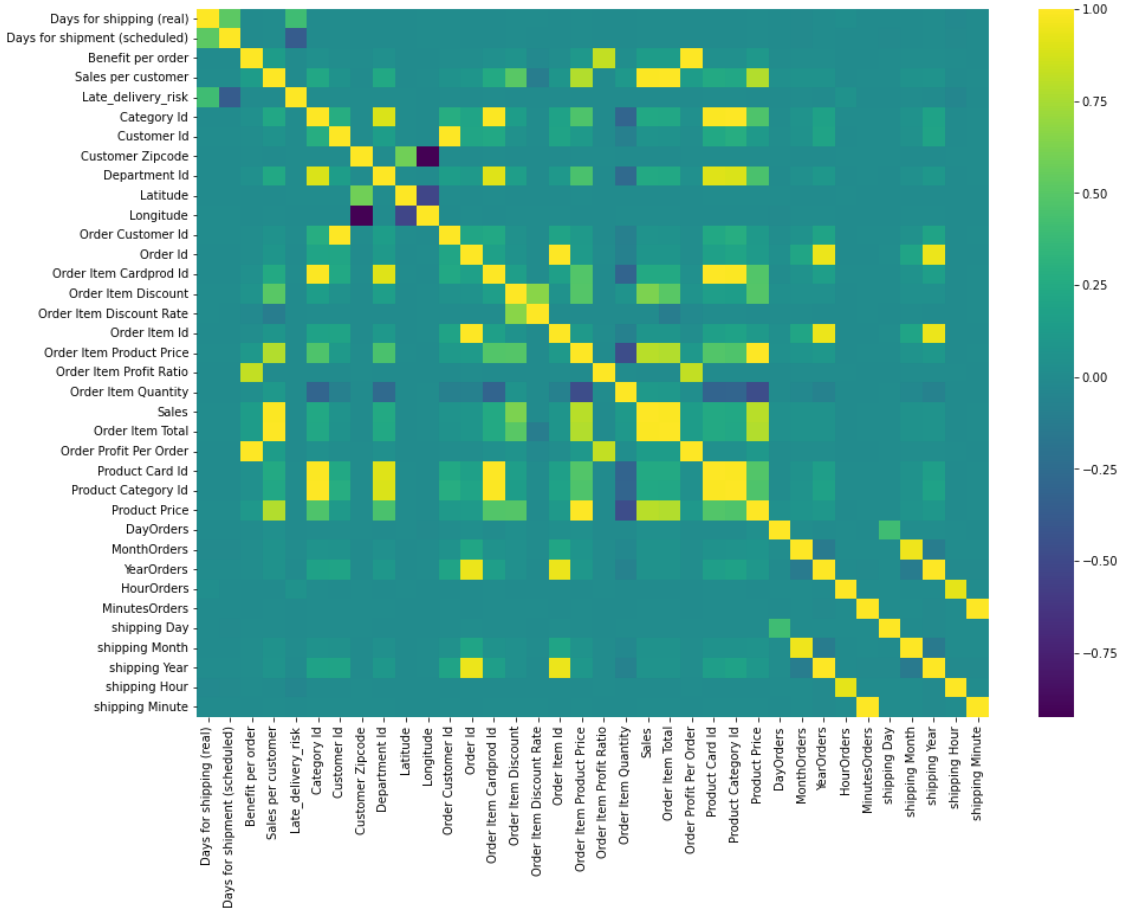
۵-۶- بررسی نتایج و مقایسه الگوریتم‌ها

در این بخش، مقایسه الگوریتم‌ها بر اساس معیارهای اعتبارسنجی، صورت گرفت و نتایج در قالب جدول ارائه گردیده است.

۶- یافته‌های تحقیق

۶-۱- تجزیه‌وتحلیل اکتشافی

ابتدا همبستگی میان ویژگی‌ها بررسی شده است. شکل زیر، همبستگی بین ویژگی‌ها را نشان می‌دهد. همبستگی یک معیار آماری است که میزان ارتباط خطی دو متغیر را بیان می‌کند (به این معنی که آن‌ها با هم با یک نرخ ثابت تغییر می‌کنند). این یک ابزار رایج برای توصیف روابط ساده بدون اظهارنظر در مورد علت و معلول است. همبستگی، عددی بین ۱- و ۱+ است. همبستگی مثبت، نشان‌دهنده میزان افزایش یا کاهش آن متغیرها به‌صورت موازی است و همبستگی منفی نشان‌دهنده میزان افزایش یک متغیر با کاهش متغیر دیگر است.



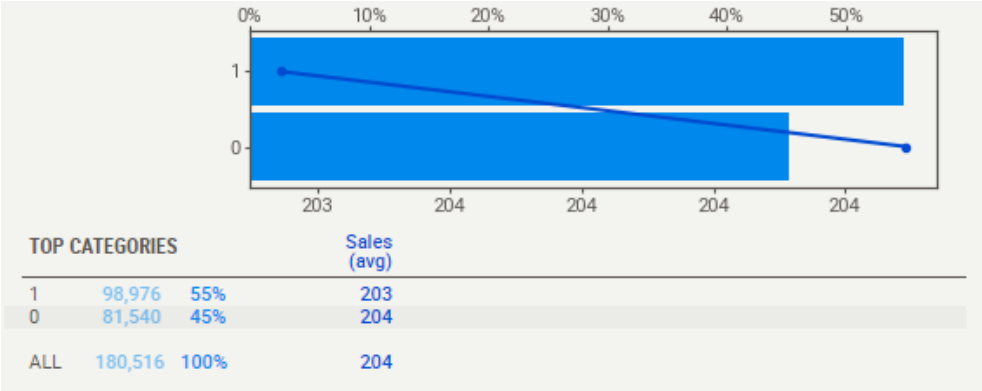
شکل (۶-۱) همبستگی ویژگی‌ها در مجموعه‌داده

همان‌طور که از شکل فوق مشاهده می‌گردد، ویژگی‌های مجموع فروش انجام شده به‌ازای هر مشتری، مقدار تخفیف کالای سفارشی، قیمت محصولات بدون تخفیف، مبلغ کل به‌ازای هر سفارش و قیمت محصول

با میزان فروش همبستگی بالایی (بین ۰.۷۵ تا ۱) دارند. همچنین ویژگی روزهای ارسال واقعی محصول خریداری شده با ویژگی‌های روزهای تحویل برنامه‌ریزی شده محصول خریداری شده و وجود ریسک تأخیر در ارسال همبستگی نسبتاً بالایی دارد (بین ۰.۵ تا ۰.۷۵).

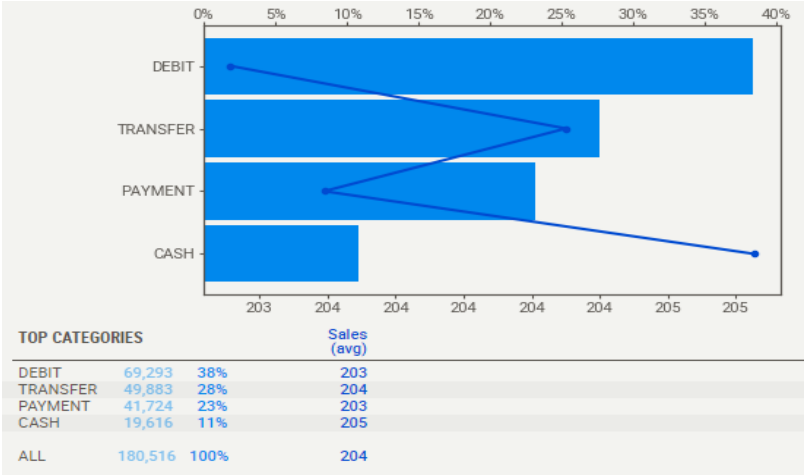
ویژگی درآمد به‌ازای هر سفارش ثبت شده، با ویژگی‌های نرخ سود کالای سفارشی و سود سفارش در هر سفارش همبستگی بالایی (بین ۰.۷۵ تا ۱) دارد. ویژگی مجموع فروش انجام شده به‌ازای هر مشتری، با ویژگی‌های قیمت محصولات بدون تخفیف، میزان فروش و مبلغ کل به‌ازای هر سفارش همبستگی بالایی (بین ۰.۷۵ تا ۱) دارد. همبستگی سایر ویژگی‌ها نیز در شکل فوق نشان‌داده‌شده است.

متغیر هدف دوم در این مطالعه، وجود ریسک ارسال با تأخیر می‌باشد. همان‌طور که در شکل زیر مشاهده می‌شود، در مجموعه‌داده مورد بررسی، ۵۵ درصد سفارش‌ها ریسک ارسال با تأخیر را دارند و ۴۵ درصد سفارش‌ها ریسک ارسال با تأخیر ندارند. این موضوع نشان می‌دهد که مجموعه‌داده مورد بررسی متعادل بوده و نیازی به متعادل‌سازی مجموعه‌داده نمی‌باشد.



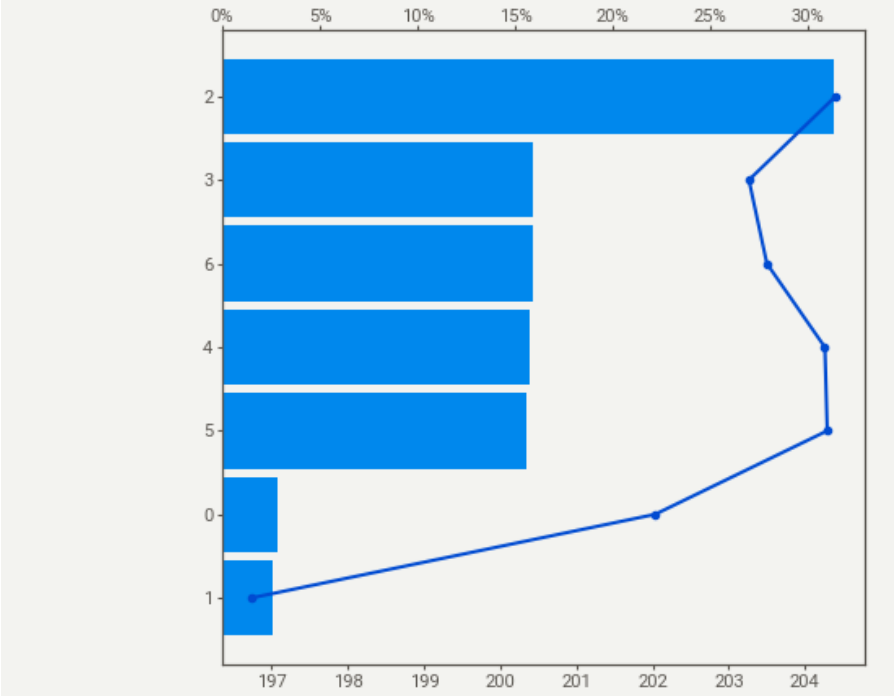
نمودار (۶-۱) متغیر هدف وجود ریسک ارسال با تأخیر

در ادامه به بررسی نوع معاملات انجام شده در مجموعه‌داده پرداخته شده است.



نمودار (۶-۲) نمودار انواع معاملات انجام شده

همان‌طور که مشاهده می‌شود، بیشترین میزان معاملات در این مجموعه‌داده، مربوط پرداخت وکالتی (۶۹۲۹۳ سفارش معادل ۳۸ درصد سفارش‌ها) می‌باشد. سفارش‌ها در حال پرداخت (۲۸ درصد معادل ۴۹۸۸۳ سفارش)، سفارش‌ها پرداخت نشده (۲۳ درصد معادل ۴۱۷۲۴ سفارش) و پرداخت آنی (۱۱ درصد معادل ۱۹۶۱۶ سفارش) رتبه‌های دوم تا چهارم را در میان انواع معاملات انجام شده در اختیار دارند. نمودار زیر روزهای ارسال واقعی محصول خریداری شده نشان می‌دهد.

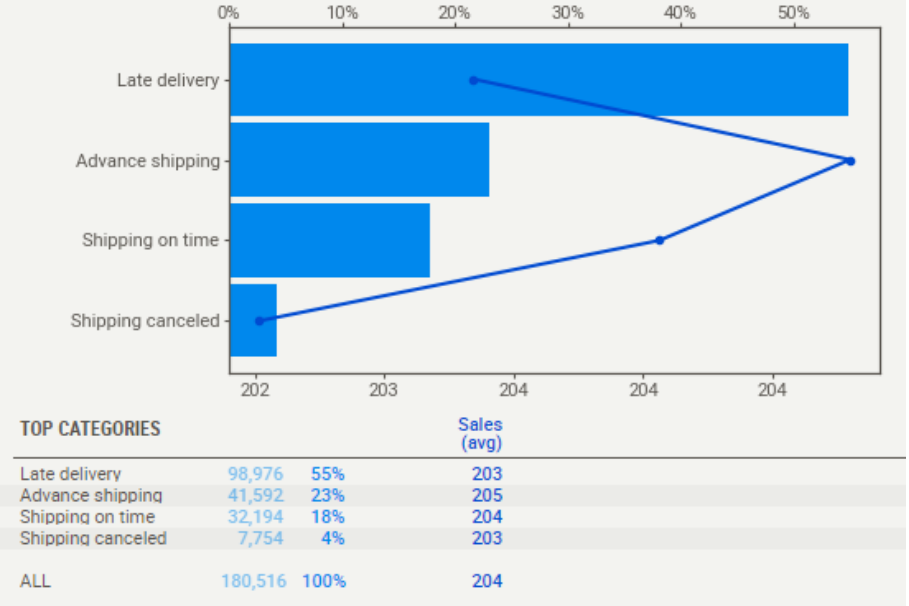


TOP CATEGORIES			Sales (avg)
2	56,617	31%	204
3	28,765	16%	203
6	28,723	16%	203
4	28,511	16%	204
5	28,163	16%	204
0	5,080	3%	202
1	4,657	3%	197
ALL	180,516	100%	204

نمودار (۶-۳) نمودار روزهای ارسال واقعی محصول خریداری شده

همان‌طور که از نمودار فوق قابل‌مشاهده است، ۳۱ درصد سفارش‌ها در ۲ روز به مشتری تحویل داده شده‌اند. سایر روزهای ارسال واقعی محصول خریداری شده در نمودار فوق نشان‌داده‌شده است.

در ادامه به توزیع ویژگی وضعیت تحویل سفارش‌ها در مجموعه‌داده مورد بررسی می‌پردازیم.

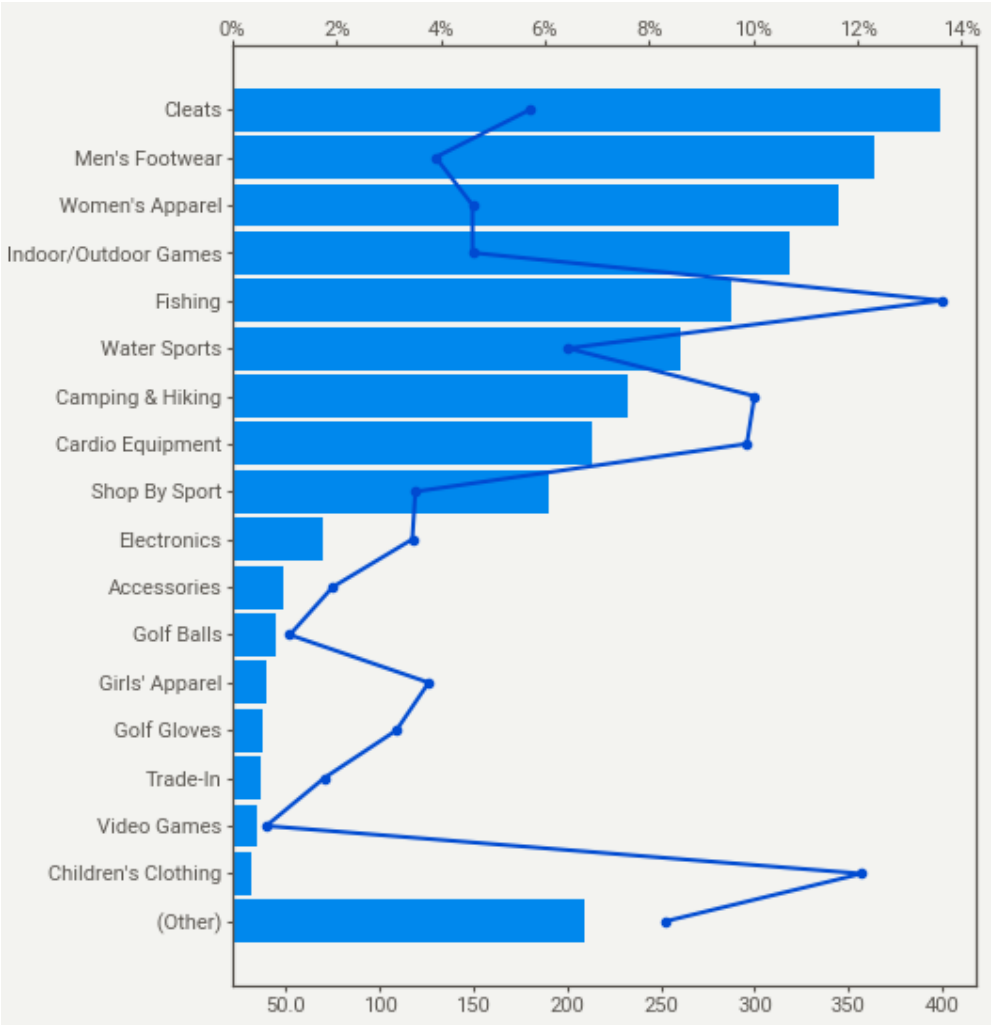


TOP CATEGORIES			Sales (avg)
Late delivery	98,976	55%	203
Advance shipping	41,592	23%	205
Shipping on time	32,194	18%	204
Shipping canceled	7,754	4%	203
ALL	180,516	100%	204

نمودار (۶-۴) نمودار وضعیت تحویل سفارش‌ها در مجموعه‌داده

همان‌طور که در نمودار فوق مشاهده می‌گردد، ۵۵ درصد سفارش‌ها تأخیر در زمان تحویل دارند و در زمان برنامه‌ریزی شده به مشتری تحویل داده نشده‌اند. ۲۳ درصد سفارش‌ها پیش از زمان مقرر و برنامه‌ریزی شده به مشتری تحویل داده شده‌اند. همچنین ۱۸ درصد سفارش‌ها در زمان مقرر به مشتری تحویل داده شده است و تنها ۴ درصد سفارش‌ها توسط مشتری لغو شده‌اند که باید علل ریشه‌ای لغو این سفارش‌ها مورد بررسی قرار گیرد تا برای اجرای پروژه‌های بهبود، اصلاحات لازم صورت بگیرد. در اینجا تمام سفارشاتی که لغو شده‌اند، سفارش‌ها در حال پرداخت بوده‌اند. ممکن است یکی از علل لغو سفارش، اختلال در درگاه پرداخت فروشگاه، بوده باشد.

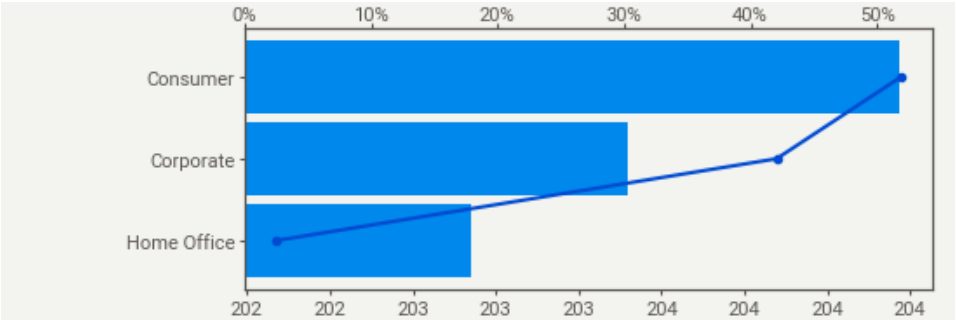
در ادامه به بررسی دسته‌بندی محصولات می‌پردازیم. نمودار فراوانی دسته‌بندی محصولات در ادامه آمده است.



نمودار (۶-۵) نمودار فراوانی دسته‌بندی محصولات سفارش داده شده

باتوجه‌به نمودار فوق، می‌توان متوجه شد که ۱۴ درصد محصولات سفارش داده شده کفش ورزشی می‌باشد. ۱۲ درصد محصولات سفارش داده شده کفش مردانه و ۱۲ درصد محصولات نیز پوشاک زنانه هستند. همچنین ۱۱ درصد محصولات سفارش داده شده مربوط به بازی‌های داخل و خارج از خانه هستند. ماهیگیری در رتبه بعدی دسته‌بندی محصولات سفارش داده است. سایر محصولات به ترتیب مربوط به ورزش‌های آبی، تجهیزات کمپینگ و پیاده‌روی، تجهیزات ورزشی و غیره هستند. باتوجه‌به تعداد سفارش‌ها انجام شده می‌توان متوجه شده که محصولات کفش ورزشی، کفش مردانه و پوشاک زنانه نزدیک به ۴۰ درصد سفارش‌ها فروشگاه را شامل می‌شوند که فروشگاه باتوجه‌به این موضوع باید برای فراهم‌کردن این محصولات برنامه‌ریزی دقیق‌تری بر اساس سلیقه مصرف‌کننده داشته باشد.

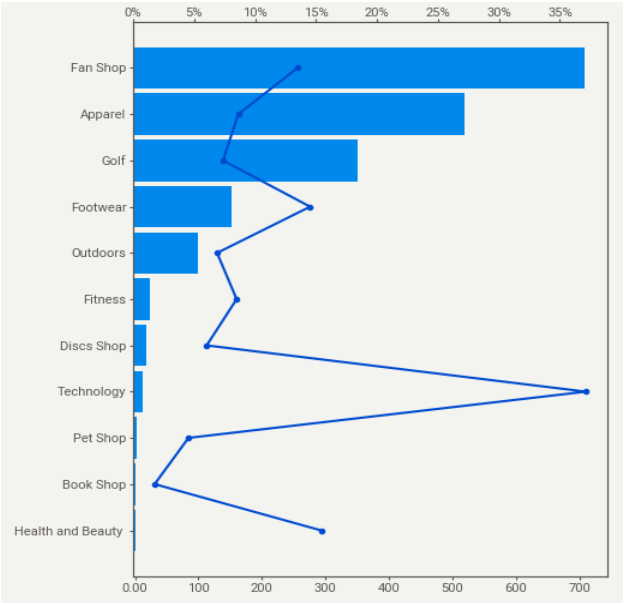
نمودار بخش‌بندی مشتریان فروشگاه در زیر آمده است.



TOP CATEGORIES			Sales (avg)
Consumer	93,502	52%	204
Corporate	54,788	30%	204
Home Office	32,226	18%	202
ALL	180,516	100%	204

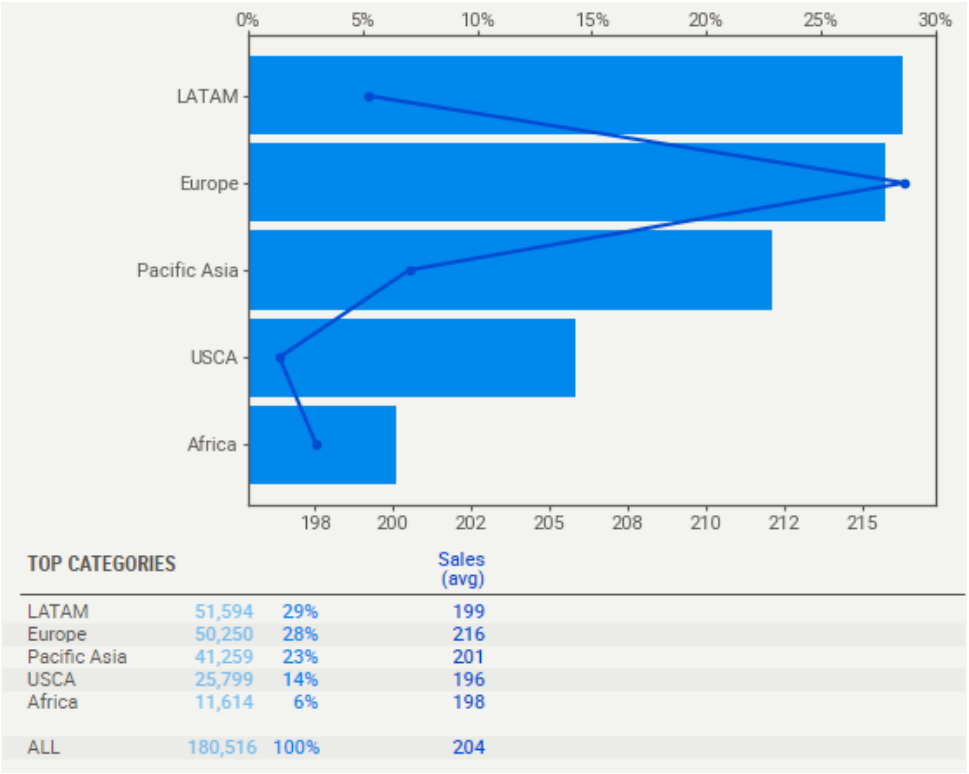
نمودار (۶-۶) نمودار بخش‌بندی مشتریان فروشگاه

باتوجه‌به نمودار فوق، ۵۲ درصد مشتریان فروشگاه مصرف‌کنندگان هستند. ۳۰ درصد مشتریان، شرکت‌های بزرگ و ۱۸ درصد مشتریان فروشگاه، شرکت‌های کوچک هستند. توجه به هر کدام از این مشتریان از اهمیت خاص خود برخوردار است. ازآنجایی‌که میانگین فروش این مشتریان تقریباً با یکدیگر برابر است، بنابراین فروشگاه باید برای جلب نظر تمام بخش‌های مشتریان تلاش کند و محصولات درخواستی آن‌ها را فراهم آورد. همچنین با خدمات پس از فروش مناسب در افزایش رضایت مشتریان تلاش نماید.



نمودار (۶-۷) نمودار فراوانی دپارتمان‌های فروشگاه

همان‌طور که از نمودار فوق مشاهده می‌شود، فروشگاه‌های ورزشی مخصوص هواداران، پوشاک و گلف حدود ۸۲ درصد دپارتمان‌های فروشگاه را تشکیل می‌دهند سفارش در آن‌ها ثبت می‌شود.

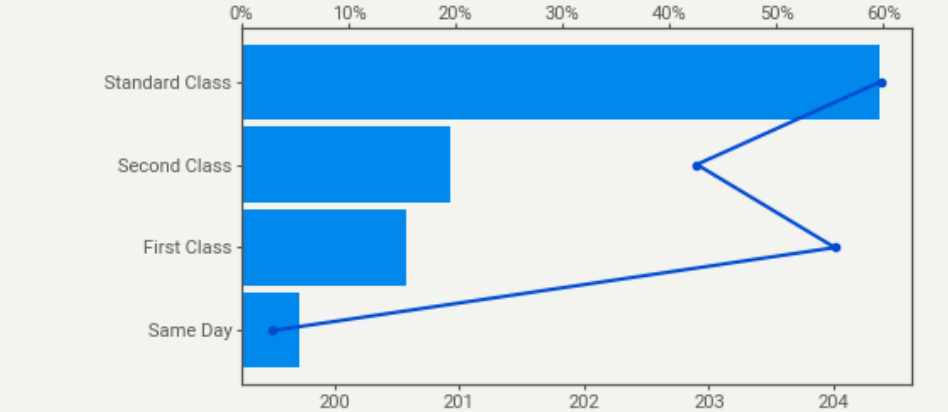


TOP CATEGORIES			Sales (avg)
LATAM	51,594	29%	199
Europe	50,250	28%	216
Pacific Asia	41,259	23%	201
USCA	25,799	14%	196
Africa	11,614	6%	198
ALL	180,516	100%	204

نمودار (۶-۸) نمودار فراوانی بازار محل تحویل سفارش

نمودار فوق، فراوانی بازارهای محل تحویل سفارش را نشان می‌دهد. همان‌طور که در این نمودار قابل‌مشاهده است، بیشترین سفارش‌ها در بازار آمریکای لاتین تحویل داده شده‌اند که ۲۹ درصد سفارش‌ها را به خود اختصاص داده است. ۲۸ درصد سفارش‌ها در بازار اروپا تحویل داده شده‌اند. همچنین ۲۳ درصد سفارش‌ها نیز در بازار آسیا و اقیانوسیه توسط مشتریان تحویل گرفته شده‌اند. ایالات متحده و آفریقا نیز ۲۰ درصد بازار محل تحویل سفارش‌ها را به خود اختصاص داده‌اند. گستردگی بازارها و درصد سفارش‌ها نشان‌دهنده عملکرد وسیع فروشگاه در بازارهای جهانی است. پراکندگی سفارش‌ها خود می‌تواند دلیلی بر ارسال با تأخیر سفارشات باشد که در این مجموعه‌داده در زمان تحویل تاخیر دارند.

در ادامه بررسی فراوانی حالت‌های حمل‌ونقل سفارش‌ها می‌پردازیم.

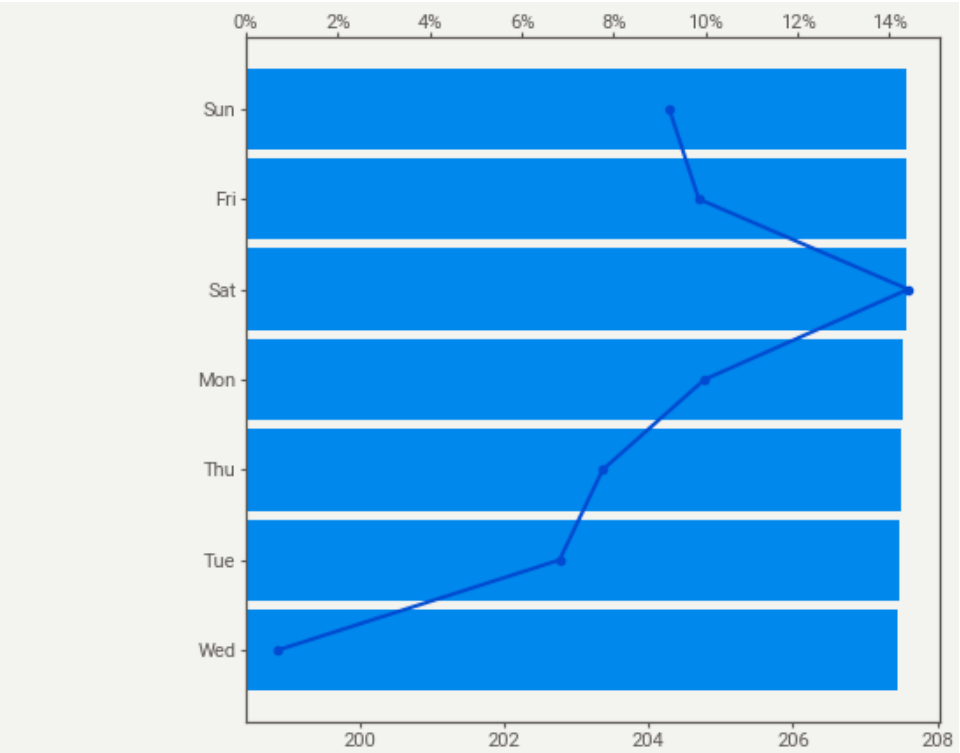


TOP CATEGORIES			Sales (avg)
Standard Class	107,750	60%	204
Second Class	35,216	20%	203
First Class	27,813	15%	204
Same Day	9,737	5%	199
ALL	180,516	100%	204

نمودار (۶-۹) نمودار فراوانی حالت‌های حمل‌ونقل سفارش‌ها

باتوجه‌به نمودار فوق، ۶۰ درصد از سفارش‌ها باکلاس استاندارد حمل‌ونقل آن‌ها صورت‌گرفته است. ۲۰ درصد سفارش‌ها با حمل‌ونقل درجه‌دو و ۱۵ درصد سفارش‌ها با حمل‌ونقل درجه یک تحویل داده شده‌اند. تنها ۵ درصد سفارش‌ها در همان روز سفارش تحویل مشتری داده شده‌اند.

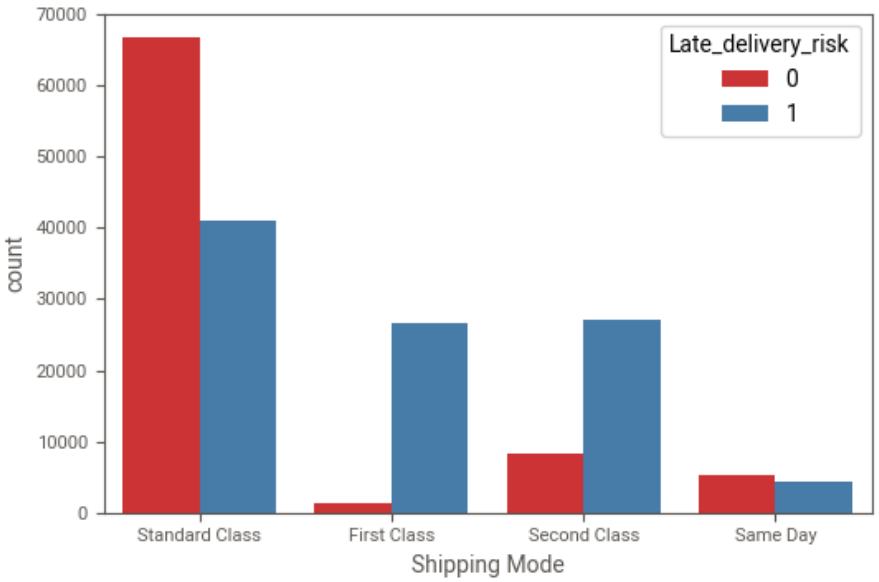
همچنین باتوجه‌به نمودار زیر، می‌توان فهمید که توزیع سفارش‌ها در روزهای هفته تقریباً به‌صورت یکنواخت بوده و روزانه بین ۲۵۹۰۰ الی ۲۶۰۰۰ سفارش در این فروشگاه ثبت می‌شود.



TOP CATEGORIES			Sales (avg)
Sun	25,945	14%	204
Fri	25,925	14%	205
Sat	25,901	14%	208
Mon	25,786	14%	205
Thu	25,750	14%	203
Tue	25,622	14%	203
Wed	25,587	14%	199
ALL	180,516	100%	204

نمودار (۶-۱۰) نمودار توزیع سفارش‌ها فروشگاه طی روزهای هفته

نمودار زیر حالت‌های حمل‌ونقل سفارش‌ها را به تفکیک وجود ریسک ارسال با تأخیر نمایش می‌دهد. همان‌طور که مشاهده می‌شود، سفارش‌ها با حمل‌ونقل درجه یک و درجه‌دو، ریسک ارسال با تاخیر در آن‌ها بیشتر است. اما در حالت حمل‌ونقل کلاس استاندارد، در بیشتر سفارش‌ها ریسک وجود ارسال با تأخیر را ندارند.



نمودار (۶-۱۱) نمودار حالت‌های حمل‌ونقل سفارش‌ها به تفکیک وجود ریسک ارسال با تأخیر

۶-۲- استخراج قوانین انجمنی

الگوریتم Apriori برای استخراج قوانین انجمنی در این مطالعه استفاده شده است. حداقل حد آستانه پشتیبانی، ۰.۳ در نظر گرفته شده است. جدول زیر تعدادی از قوانین را نشان می‌دهد که بیشترین تکرار را در مجموعه‌داده داشته‌اند. در واقع بیشترین میزان پشتیبانی را دارند.

جدول (۶-۱) ترتیب قوانین انجمنی استخراج شده بر اساس معیار پشتیبانی از زیاد به کم

	support	itemsets
5	0.615696	(Customer Country_USA)
9	0.596900	(Shipping Mode_Standard Class)
10	0.548295	(Delivery Status_Late delivery, Late_delivery_...
0	0.548295	(Late_delivery_risk)
2	0.548295	(Delivery Status_Late delivery)
6	0.517971	(Customer Segment_Consumer)
4	0.384304	(Customer Country_Puerto Rico)
1	0.383861	(Type_DEBIT)
3	0.369884	(Customer City_Caguas)
14	0.369884	(Customer City_Caguas, Customer Country_Puerto...
16	0.366433	(Shipping Mode_Standard Class, Customer Countr...
11	0.337837	(Late_delivery_risk, Customer Country_USA)
13	0.337837	(Delivery Status_Late delivery, Customer Count...
18	0.337837	(Delivery Status_Late delivery, Late_delivery_...
8	0.329550	(Order Status_COMPLETE)
12	0.329550	(Order Status_COMPLETE, Type_DEBIT)
15	0.320348	(Customer Segment_Consumer, Customer Country_USA)
17	0.309950	(Customer Segment_Consumer, Shipping Mode_Stan...
7	0.303508	(Customer Segment_Corporate)

همان‌طور که مشاهده می‌شود، کشور ایالات متحده به‌عنوان کشور محل سفارش مشتری و حالت حمل‌ونقل استاندارد، در میان موارد تکی بیشترین تعداد را داشته‌اند. جدول زیر، قوانین انجمنی استخراج شده از جدول فوق را نشان می‌دهد که معیارهای Confidence و Lift در آن محاسبه شده است. ترتیب این قوانین بر اساس بیشترین میزان معیار Lift می‌باشد.

جدول (۶-۲) قوانین انجمنی استخراج شده از مجموعه‌داده، ترتیب بر اساس معیار Lift از زیاد به کم

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
4	(Order Status_COMPLETE)	(Type_DEBIT)	0.329550	0.383861	0.329550	1.000000	2.605112	0.203049	inf	0.918993
5	(Type_DEBIT)	(Order Status_COMPLETE)	0.383861	0.329550	0.329550	0.858514	2.605112	0.203049	4.738628	1.000000
8	(Customer City_Caguas)	(Customer Country_Puerto Rico)	0.369884	0.384304	0.369884	1.000000	2.602107	0.227736	inf	0.977116
9	(Customer Country_Puerto Rico)	(Customer City_Caguas)	0.384304	0.369884	0.369884	0.962478	2.602107	0.227736	16.793327	1.000000
0	(Delivery Status_Late delivery)	(Late_delivery_risk)	0.548295	0.548295	0.548295	1.000000	1.823836	0.247668	inf	1.000000
20	(Late_delivery_risk)	(Delivery Status_Late delivery, Customer Count...	0.548295	0.337837	0.337837	0.616159	1.823836	0.152603	1.725099	1.000000
19	(Delivery Status_Late delivery)	(Late_delivery_risk, Customer Country_USA)	0.548295	0.337837	0.337837	0.616159	1.823836	0.152603	1.725099	1.000000
18	(Late_delivery_risk, Customer Country_USA)	(Delivery Status_Late delivery)	0.337837	0.548295	0.337837	1.000000	1.823836	0.152603	inf	0.682166
17	(Delivery Status_Late delivery, Customer Count...	(Late_delivery_risk)	0.337837	0.548295	0.337837	1.000000	1.823836	0.152603	inf	0.682166
1	(Late_delivery_risk)	(Delivery Status_Late delivery)	0.548295	0.548295	0.548295	1.000000	1.823836	0.247668	inf	1.000000
11	(Customer Country_USA)	(Customer Segment_Consumer)	0.615696	0.517971	0.320348	0.520303	1.004502	0.001436	1.004861	0.011662
10	(Customer Segment_Consumer)	(Customer Country_USA)	0.517971	0.615696	0.320348	0.618468	1.004502	0.001436	1.007265	0.009298
14	(Customer Segment_Consumer)	(Shipping Mode_Standard Class)	0.517971	0.596900	0.309950	0.598394	1.002502	0.000774	1.003719	0.005178
15	(Shipping Mode_Standard Class)	(Customer Segment_Consumer)	0.596900	0.517971	0.309950	0.519267	1.002502	0.000774	1.002696	0.006192
16	(Delivery Status_Late delivery, Late_delivery_...	(Customer Country_USA)	0.548295	0.615696	0.337837	0.616159	1.000753	0.000254	1.001207	0.001665
7	(Customer Country_USA)	(Delivery Status_Late delivery)	0.615696	0.548295	0.337837	0.548708	1.000753	0.000254	1.000914	0.001957
6	(Delivery Status_Late delivery)	(Customer Country_USA)	0.548295	0.615696	0.337837	0.616159	1.000753	0.000254	1.001207	0.001665
3	(Customer Country_USA)	(Late_delivery_risk)	0.615696	0.548295	0.337837	0.548708	1.000753	0.000254	1.000914	0.001957
2	(Late_delivery_risk)	(Customer Country_USA)	0.548295	0.615696	0.337837	0.616159	1.000753	0.000254	1.001207	0.001665

در سطر اول جدول مشاهده می‌گردد، وضعیت سفارش کامل با نوع معامله وکالتی به‌هم‌وابسته هستند و در کنار یکدیگر می‌آیند. سایر موارد و قوانین استخراج شده نیز در فایل پیوست کدها قابل‌مشاهده است.

۶-۳- ارزیابی و مقایسه عملکرد الگوریتم‌های یادگیری ماشین

در جدول زیر، مقایسه ارزیابی الگوریتم‌های اجرا شده روی داده‌های دیده نشده توسط مدل صورت‌گرفته است. ابتدا مقایسه الگوریتم‌های رگرسیون برای پیش‌بینی میزان فروش هر سفارش را بررسی می‌کنیم.

جدول (۶-۳) مقایسه الگوریتم‌های رگرسیون یادگیری ماشین برای پیش‌بینی فروش هر سفارش

	MLA used	Train Accuracy	Test Accuracy	MAE	MSE	RMSE	R^2 Score	Adjusted R^2 Score
0	LinearRegression	1.0000	1.0000	0.000593	0.000002	0.001563	1.000000	1.000000
1	Ridge	1.0000	1.0000	0.000593	0.000002	0.001563	1.000000	1.000000
2	Lasso	1.0000	1.0000	0.012694	0.000407	0.020177	1.000000	1.000000
3	DecisionTreeRegressor	1.0000	1.0000	0.012253	0.656473	0.810230	0.999963	0.999963
5	BayesianRidge	1.0000	1.0000	0.000593	0.000002	0.001563	1.000000	1.000000
4	RandomForestRegressor	0.8585	0.8636	38.873632	2444.683456	49.443740	0.863609	0.862842

همان‌طور که مشاهده می‌شود، از ۶ مدل پیاده‌سازی شده، تنها الگوریتم رگرسیون جنگل تصادفی عملکرد ضعیف‌تری داشته است. الگوریتم‌های رگرسیون خطی، رگرسیون ستیغی و رگرسیون بیزی با معیار مربع R تنظیم شده یکسان ۱، بهترین مدل‌ها برای پیش‌بینی فروش هر سفارش هستند.

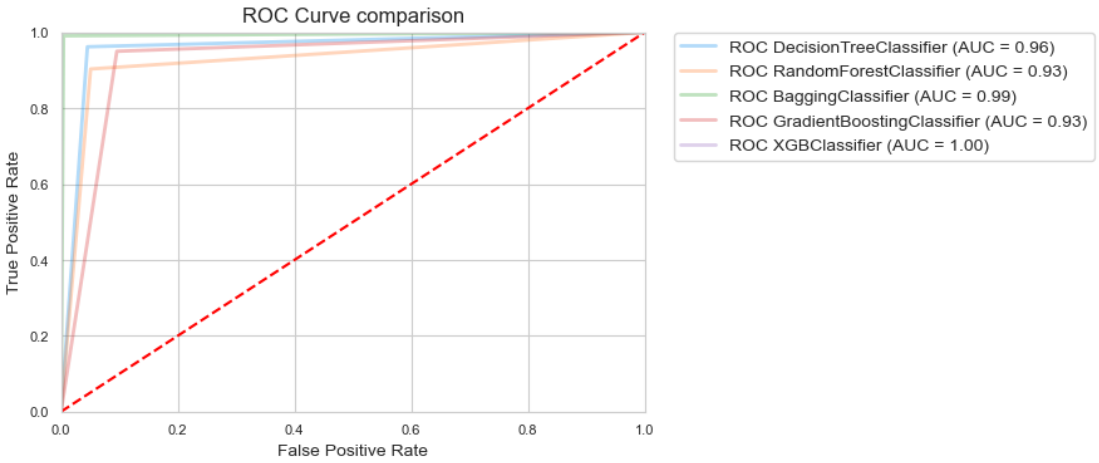
در ادامه، به مقایسه الگوریتم‌های دسته‌بندی برای پیش‌بینی وجود ریسک ارسال با تأخیر برای هر سفارش می‌پردازیم.

جدول (۶-۴) مقایسه الگوریتم‌های دسته‌بندی یادگیری ماشین برای پیش‌بینی وجود ریسک ارسال با تأخیر

	MLA used	Train Accuracy	Test Accuracy	Precission	Recall	AUC	F1-Score
4	XGBClassifier	1.0000	0.9998	0.999695	1.000000	0.999818	0.999847
2	BaggingClassifier	0.9998	0.9838	0.993476	0.976635	0.984490	0.984983
0	DecisionTreeClassifier	1.0000	0.9587	0.962169	0.961924	0.958391	0.962047
3	GradientBoostingClassifier	0.9317	0.9292	0.922267	0.950013	0.927221	0.935934
1	RandomForestClassifier	1.0000	0.9244	0.955167	0.903385	0.926387	0.928555

مشاهده می‌شود که الگوریتم XGboost با امتیاز F۱ نزدیک به ۱ به‌خوبی پیش‌بینی را انجام داده است و سایر الگوریتم‌ها نیز در محدوده ۰.۹۲ تا ۰.۹۹ قرار دارند که این موضوع نشان از کارایی بسیار خوب الگوریتم‌های یادگیری ماشین در این زمینه دارد.

همچنین نمودار زیر، منحنی ROC الگوریتم‌های دسته‌بندی بکار گرفته شده را نشان می‌دهد. مساحت زیر محنی مربوط به مدل XGboost بیشترین مقدار را در بین الگوریتم‌های پیاده‌سازی شده در اختیار دارد.



نمودار (۶-۱۲) منحنی ROC برای الگوریتم‌های دسته‌بندی بکار گرفته شده

۷- نتیجه‌گیری و پیشنهاد

در این مطالعه، برای تجزیه‌وتحلیل زنجیره تامین یک فروشگاه خرده‌فروشی با فروش حضوری و اینترنتی و پیاده‌سازی الگوریتم‌های یادگیری ماشین، ما از مجموعه‌داده اطلاعات زنجیره تامین یک فروشگاه اینترنتی که توسط شرکت دیتاکو[۱۰۳] بین سال‌های ۲۰۱۵ تا ۲۰۱۸، منتشر شده است، استفاده کرده‌ایم. تجزیه‌وتحلیل اکتشافی داده‌ها و نتایج این مطالعه نشان می‌دهد:

- ویژگی‌های مجموع فروش انجام شده به‌ازای هر مشتری، مقدار تخفیف کالای سفارشی، قیمت محصولات بدون تخفیف، مبلغ کل به‌ازای هر سفارش و قیمت محصول با میزان فروش همبستگی بالایی (بین ۰.۷۵ تا ۱) دارند. در واقع با افزایش هر کدام از این ویژگی‌ها می‌توان انتظار داشت که میزان فروش افزایش یابد. یکی از راه‌های افزایش فروش می‌تواند ایجاد کمپین‌های تبلیغاتی دوره‌ای و افزایش میزان تخفیفات کالاها باشد.
- در وضعیت تحویل سفارش‌ها باید به طور ویژه به آن پرداخته شود و با بررسی علل ریشه‌ای هر کدام از سفارش‌ها، پروژه‌های بهبود را برای حمل‌ونقل سفارش‌ها بادقت بیشتری دنبال نمود.
- باتوجه‌به تعداد سفارش‌ها انجام شده می‌توان متوجه شده که محصولات کفش ورزشی، کفش مردانه و پوشاک زنانه نزدیک به ۴۰ درصد سفارش‌ها فروشگاه را شامل می‌شوند که فروشگاه باتوجه‌به این موضوع باید برای فراهم‌کردن این محصولات برنامه‌ریزی دقیق‌تری بر اساس سلیقه مصرف‌کننده داشته باشد.
- الگوریتم‌های یادگیری ماشین رگرسیون و دسته‌بندی به ترتیب عملکرد فوق‌العاده‌ای در پیش‌بینی میزان فروش هر سفارش و وجود ریسک ارسال با تأخیر سفارش داشته‌اند. این موضوع نشان می‌دهد که استفاده از الگوریتم‌های یادگیری ماشین می‌تواند عملکرد بسیار مؤثری در حوزه زنجیره تامین فروشگاه اینترنتی داشته باشند.
- در مطالعات آینده، پیشنهاد می‌شود از داده‌های حجیم‌تر برای پیاده‌سازی الگوریتم‌های یادگیری ماشین استفاده شود. همچنین به دلیل بهبود پیوسته و مداوم الگوریتم‌ها، می‌توان از الگوریتم‌های موجود دیگر مانند Adaboost برای پیش‌بینی استفاده کرد و نتایج را با نتایج این پژوهش مقایسه نمود.
- حوزه زنجیره تامین، در حال حاضر به‌عنوان یک ترند جهانی به طور پیوسته در حال بهبود است و به‌کارگیری هوش مصنوعی در فرایندهای مختلف زنجیره تامین می‌تواند بسیار مؤثر واقع گردد.

---

[۱۰۳] Data-Co

استفاده از حوزه‌های یادگیری عمیق[۱۰۴] ، پردازش زبان طبیعی[۱۰۵] و بینایی ماشین[۱۰۶] در مطالعات آینده می‌تواند موردتوجه قرار گیرد.

[۱۰۴] Deep Learning
[۱۰۵] Natural Language Process
[۱۰۶] Computer Vision

۸- منابع

.۱ Li, L., et al., *Data-driven online service supply chain: a demand-side and supply-side perspective.* Journal of Enterprise Information Management, ۲۰۲۱. ۳٤(۱): p. ۳٦٥-۳۸۱.

.۲ Pereira, M.M. and E.M. Frazzon, *A data-driven approach to adaptive synchronization of demand and supply in omni-channel retail supply chains.* International Journal of Information Management, ۲۰۲۱. ٥۷: p. ۱۰۲۱٦٥.

.۳ Rodrigues, V.S., et al., *Measurement, mitigation and prevention of food waste in supply chains: An online shopping perspective.* Industrial Marketing Management, ۲۰۲۱. ۹۳: p. ٥٤٥-٥٦۲.

.٤ Krishna, A., et al. *Sales-forecasting of retail stores using machine learning techniques*. in *2018 3rd international conference on computational systems and information technology for sustainable solutions (CSITSS)*. ۲۰۱۸. IEEE.

.٥ Beheshti-Kashi, S., et al., *A survey on retail sales forecasting and prediction in fashion markets.* Systems Science & Control Engineering, ۲۰۱٥. ۳(۱): p. ۱٥٤-۱٦۱.

.٦ Oke, A. and M. Gopalakrishnan, *Managing disruptions in supply chains: A case study of a retail supply chain.* International journal of production economics, ۲۰۰۹. ۱۱۸(۱): p. ۱٦۸-۱۷٤.

.۷ Jiang, H., J. Ruan, and J. Sun. *Application of machine learning model and hybrid model in retail sales forecast*. in ۲۰۲۱ *IEEE 6th international conference on big data analytics (ICBDA)*. ۲۰۲۱. IEEE.

.۸ Yao, J. and M. Gu, *Optimization analysis of supply chain resource allocation in customized online shopping service mode.* Mathematical Problems in Engineering, ۲۰۱٥. ۲۰۱٥.

.۹ Yao, J., *Supply chain resources integration optimisation in B2C online shopping.* International Journal of Production Research, ۲۰۱۷. ٥٥(۱۷): p. ٥۰۷۹-٥۰۹٤.

.۱۰ Kück, M. and M. Freitag, *Forecasting of customer demands for production planning by local k-nearest neighbor models.* International Journal of Production Economics, ۲۰۲۱. ۲۳۱: p. ۱۰۷۸۳۷.

.۱۱ Shilong, Z. *Machine learning model for sales forecasting by using XGBoost*. in *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*. ۲۰۲۱. IEEE.

.۱۲ Sajawal, M., et al., *A Predictive Analysis of Retail Sales Forecasting using Machine Learning Techniques.* Lahore Garrison University Research Journal of Computer Science and Information Technology, ۲۰۲۲. ٦(۰٤): p. ۳۳-٤٥.

.۱۳ Lee, I. and G. Mangalaraj, *Big data analytics in supply chain management: a systematic literature review and research directions.* Big Data and Cognitive Computing, ۲۰۲۲. ٦(۱): p. ۱۷.

.۱٤ Li, J., A. Ghadge, and M.K. Tiwari, *Impact of replenishment strategies on supply chain performance under e-shopping scenario.* Computers & Industrial Engineering, ۲۰۱٦. ۱۰۲: p. ۷۸-۸۷.

.١٥ Chen, I.-F. and C.-J. Lu, *Sales forecasting by combining clustering and machine-learning techniques for computer retailing.* Neural Computing and Applications, ٢٠١٧. ٢٨: p. ٢٦٣٣-٢٦٤٧.

.١٦ Catal, C., et al., *Benchmarking of regression algorithms and time series analysis techniques for sales forecasting.* Balkan Journal of Electrical and Computer Engineering, ٢٠١٩. ٧(١): p. ٢٠-٢٦.

.١٧ Arif ,M.A.I., et al. *Comparison study: Product demand forecasting with machine learning for shop*. in *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*. ٢٠١٩. IEEE.

.١٨ Oosthuizen, K., et al., *Artificial intelligence in retail: The AI-enabled value chain.* Australasian Marketing Journal, ٢٠٢١. **٢٩**(٣): p. ٢٦٤-٢٧٣.

.١٩ Wu, Y., et al., *Low-carbon decision-making model of online shopping supply chain considering the O2O model.* Journal of Retailing and Consumer Services, ٢٠٢١ . :٥٩p. ١٠٢٣٨٨.

.٢٠ Yu, Y., et al., *E-commerce logistics in supply chain management: Practice perspective.* Procedia Cirp, ٢٠١٦. **٥٢**: p. ١٧٩-١٨٥.

.٢١ Abdirad, M. and K. Krishnan, *Examining the impact of E-supply chain on service quality and customer satisfaction: a case study.* International Journal of Quality and Service Sciences, ٢٠٢٢. ١٤(٢): p. ٢٧٤-٢٩٠.

.٢٢ Chen, F. and T. Ou, *Sales forecasting system based on Gray extreme learning machine with Taguchi method in retail industry.* Expert Systems with Applications, ٢٠٢١. ٣٨(٣): p. ١٣٣٦-١٣٤٥.

.٢٣ Liu, Y., et al. *An Aggregate Store Sales Forecasting Framework based on ConvLSTM*. in *2021 The 5th International Conference on Compute and Data Analysis*. ٢٠٢١.

.٢٤ Andrejić, M., *Modeling Retail Supply Chain Efficiency* :*Exploration and Comparative Analysis of Different Approaches.* Mathematics, ٢٠٢٣. ١١(٧): p. ١٥٧١.