فهرست

مقدمه

در دهه‌های اخیر، یادگیری ماشین به یکی از ارکان بنیادین تحول در علوم داده، هوش مصنوعی و سامانه‌های تصمیم‌یار تبدیل شده است. این حوزه، که در مرز میان علوم رایانه، ریاضیات، آمار و مهندسی شکل گرفته و توسعه یافته است، امروزه نقشی تعیین‌کننده در تحلیل داده‌های پیچیده، استخراج الگوهای پنهان و پیش‌بینی هوشمندانه پدیده‌ها ایفا می‌کند. گسترش روزافزون کاربردهای یادگیری ماشین در حوزه‌هایی همچون پزشکی، مهندسی، اقتصاد، پردازش تصویر، پردازش زبان طبیعی، علوم زیستی، سامانه‌های توصیه‌گر و فناوری‌های خودکار، نشان‌دهنده جایگاه راهبردی آن در علم و فناوری معاصر است. در چنین بستری، دستیابی به درکی دقیق، استاندارد و نظام‌مند از مفاهیم، تعاریف و واژگان این حوزه، ضرورتی انکارناپذیر برای آموزش، پژوهش و توسعه علمی به شمار می‌آید.

رشد سریع الگوریتم‌ها، چارچوب‌های محاسباتی و روش‌های تحلیلی در سال‌های اخیر، ادبیات یادگیری ماشین را به‌گونه‌ای چشمگیر گسترش داده است. این گسترش، اگرچه به غنای علمی و کاربردی این حوزه افزوده، اما هم‌زمان موجب شکل‌گیری مجموعه‌ای گسترده از اصطلاحات تخصصی، تعابیر فنی و مفاهیم چندلایه شده است که در بسیاری از موارد، میان منابع مختلف با تفاوت‌هایی در بیان، دامنه معنا یا نحوه کاربرد همراه‌اند. چنین وضعیتی ممکن است به ابهام مفهومی، برداشت‌های ناهمگون و دشواری در برقراری ارتباط علمی منجر شود؛ به‌ویژه برای دانشجویان، پژوهشگران جوان و متخصصانی که از حوزه‌های دیگر وارد عرصه یادگیری ماشین می‌شوند. از این‌رو، فقدان یک مرجع منسجم، دقیق و علمی برای اصطلاح‌شناسی این دانش، می‌تواند مانعی جدی در مسیر یادگیری مؤثر، پژوهش روشمند و انتقال دانش میان‌رشته‌ای باشد.

کتاب حاضر با هدف تبیین، تعریف، طبقه‌بندی و یکپارچه‌سازی اصطلاحات رایج و تخصصی یادگیری ماشین تدوین شده است. این اثر می‌کوشد با ارائه تعاریفی روشن، دقیق و مبتنی بر ادبیات علمی معتبر، زمینه‌ای مناسب برای فهم عمیق‌تر مفاهیم بنیادی و پیشرفته این حوزه فراهم آورد. در این راستا، اصطلاحاتی نظیر یادگیری نظارت‌شده، یادگیری بدون نظارت، یادگیری نیمه‌نظارتی، یادگیری تقویتی، یادگیری عمیق، بیش‌برازش، کم‌برازش، اعتبارسنجی متقابل، تابع هزینه، تابع زیان، بهینه‌سازی، تنظیم‌گری، تعمیم‌پذیری، تفسیرپذیری، توضیح‌پذیری، استخراج ویژگی، مهندسی ویژگی و انتخاب مدل، نه صرفاً به‌عنوان واژه‌هایی فنی، بلکه به‌مثابه اجزای یک زبان علمی پویا و ساختاریافته معرفی می‌شوند. درک صحیح این زبان برای هر دانش‌پژوه، مدرس، پژوهشگر و متخصصی که در این حوزه فعالیت می‌کند یا با آن در تعامل است، ضرورتی بنیادین دارد.

اهمیت اصطلاح‌شناسی در یادگیری ماشین تنها به تعریف واژگان محدود نمی‌شود، بلکه با ایجاد نوعی انسجام معرفتی و روش‌شناختی در مواجهه با نظریه‌ها، مدل‌ها و کاربردها نیز پیوند دارد. بسیاری از مفاهیم این حوزه، بسته به بستر استفاده، سطح تحلیل یا سنت علمی حاکم، ممکن است دارای معانی نزدیک اما متمایز باشند. برای مثال، برخی اصطلاحات در متون آماری، علوم رایانه یا کاربردهای مهندسی با بار مفهومی متفاوتی به کار می‌روند، در حالی‌که در ظاهر، معادل یا هم‌معنا به نظر می‌رسند. از این منظر، تعیین مرزهای مفهومی هر اصطلاح و روشن‌ساختن نسبت آن با مفاهیم همجوار، نقش مهمی در جلوگیری از سوءبرداشت، افزایش دقت علمی، بهبود کیفیت نگارش تخصصی و ارتقای توان تحلیل انتقادی ایفا می‌کند. این مسئله در محیط‌های آموزشی و پژوهشی، که دقت در فهم مفاهیم زیربنای تولید دانش معتبر است، از اهمیت ویژه‌ای برخوردار است.

این ضرورت در مطالعات میان‌رشته‌ای اهمیتی دوچندان می‌یابد. امروزه یادگیری ماشین به‌طور گسترده در تعامل با حوزه‌هایی نظیر تصویربرداری پزشکی، فیزیک پزشکی، بیوانفورماتیک، علوم اعصاب، اپیدمیولوژی، رباتیک و علوم شناختی قرار گرفته است. در این فضاها، انتقال صحیح مفاهیم میان حوزه‌های مختلف نیازمند زبانی مشترک، دقیق و استاندارد است. هرگونه ابهام یا ناهماهنگی در کاربرد اصطلاحات می‌تواند به تفسیر نادرست نتایج، ضعف در طراحی مطالعات، یا حتی اختلال در بازتولید و ارزیابی یافته‌های علمی منجر شود. ازاین‌رو، اصطلاح‌شناسی را باید نه یک فعالیت حاشیه‌ای، بلکه بخشی از زیرساخت نظری و ارتباطی علم یادگیری ماشین دانست.

در تدوین این کتاب تلاش شده است که اصطلاحات نه صرفاً به‌صورت واژه‌نامه‌ای و فهرست‌وار، بلکه در پیوند با بنیان‌های نظری، زمینه‌های کاربردی و جایگاه آن‌ها در ساختار دانش یادگیری ماشین معرفی شوند. رویکرد کتاب بر آن است که میان دقت علمی، شفافیت مفهومی، انسجام ساختاری و قابلیت استفاده آموزشی توازن برقرار کند. بر این اساس، هر اصطلاح تا حد امکان در بستر مفهومی مناسب خود تبیین می‌شود تا خواننده بتواند علاوه بر آشنایی با تعریف آن، نسبت آن را با دیگر مفاهیم، روش‌ها و کاربردهای مرتبط نیز درک کند. این شیوه به‌ویژه برای افرادی که در مراحل آغازین یادگیری قرار دارند، و نیز برای پژوهشگرانی که به دنبال استناد دقیق و به‌کارگیری سنجیده مفاهیم هستند، سودمند خواهد بود

در مجموع، این اثر با این فرض بنیادین شکل گرفته است که فهم عمیق هر دانش، مستلزم فهم دقیق زبان آن دانش است. یادگیری ماشین نیز به‌عنوان یکی از پویاترین و اثرگذارترین حوزه‌های علمی معاصر، بیش از هر زمان دیگری به منابعی نیاز دارد که بتوانند زبان تخصصی آن را به شکلی معتبر، منظم و روشن صورت‌بندی کنند. امید است این کتاب بتواند به‌عنوان منبعی قابل اعتماد برای دانشجویان، پژوهشگران، مدرسان و متخصصان، در فهم زبان تخصصی یادگیری ماشین، تسهیل تعامل علمی، ارتقای دقت مفهومی و استحکام‌بخشی به آموزش و پژوهش در این حوزه نقشی مؤثر و ماندگار ایفا کند.

فصل اول

مفاهیـم پایـــه
(Core Concepts)

دقت Accuracy

معیاری برای ارزیابی عملکرد مدل طبقه‌بندی است که نسبت تعداد پیش‌بینی‌های صحیح مدل به کل نمونه‌های مورد بررسی را بیان می‌کند. این معیار از تقسیم مجموع پیش‌بینی‌های درست (شامل نمونه‌های مثبت و منفی که به‌درستی شناسایی شده‌اند) بر کل تعداد نمونه‌ها محاسبه می‌شود. دقت برای داده‌های نامتوازن ممکن است گمراه‌کننده باشد، ازاین‌رو، در کاربردهای پزشکی توصیه می‌شود این شاخص همواره در کنار معیارهای تکمیلی مانند حساسیت و ویژگی تفسیر شود.

الگوریتم Algorithm

الگوریتم مجموعه‌ای از دستورات گام‌به‌گام جهت حل یک مسئله یا انجام یک وظیفه خاص است. در چارچوب یادگیری ماشین، الگوریتم‌ها نحوه آموزش مدل‌ها، به‌روزرسانی پارامترها و استخراج الگو از داده‌ها را تعیین می‌کنند (مانند الگوریتم درخت تصمیم، رگرسیون خطی، شبکه‌های عصبی). انتخاب و طراحی الگوریتم به عواملی همچون نوع داده‌ها، ماهیت مسئله (مانند طبقه‌بندی یا پیش‌بینی) و هدف نهایی مدل وابسته است. کارایی یک الگوریتم معمولاً بر اساس تعادلی میان پیچیدگی محاسباتی، دقت مدل و بهره‌وری در پردازش داده‌ها ارزیابی می‌شود.

(AI) Artificial Intelligence هوش مصنوعی

شاخه‌ای از علوم رایانه است که هدف آن طراحی و توسعه سامانه‌هایی است که بتوانند توانایی‌های شناختی انسان، از جمله یادگیری، استدلال، تصمیم‌گیری و حل مسئله را شبیه‌سازی کنند. این حوزه به‌صورت میان‌رشته‌ای و با اتکا بر مفاهیم ریاضیات، آمار، علوم شناختی و مهندسی کامپیوتر شکل گرفته است. یکی از زیرشاخه‌های کلیدی هوش مصنوعی، **یادگیری ماشین** است

که بر توسعه الگوریتم‌هایی تمرکز دارد که قادرند الگوها و روابط نهفته در داده‌ها را به‌صورت خودکار بیاموزند، بدون آنکه به‌طور صریح برنامه‌نویسی شوند. در سطح پیشرفته‌تر، **یادگیری عمیق** به‌عنوان شاخه‌ای از یادگیری ماشین، با بهره‌گیری از شبکه‌های عصبی مصنوعی چندلایه امکان مدل‌سازی روابط پیچیده و غیرخطی را فراهم می‌سازد. امروزه کاربردهای هوش مصنوعی طیف گسترده‌ای از حوزه‌ها، از جمله پردازش تصویر و گفتار، تحلیل داده‌های پزشکی، پشتیبانی تصمیم‌گیری بالینی و سامانه‌های خودران را در بر می‌گیرد و نقش فزاینده‌ای در توسعه فناوری‌های نوین سلامت ایفا می‌کند.

ویژگی Attribute

ویژگی یا صفت، یک مشخصه قابل اندازه‌گیری از یک شیء است که برای مدل‌سازی داده‌ها استفاده می‌شود. به ویژگی معمولا متغیر نیز گفته می‌شود و می‌تواند عددی یا طبقه‌ای باشد. در یادگیری ماشین، انتخاب مناسب ویژگی‌ها نقش تعیین‌کننده‌ای در عملکرد مدل دارد، به‌طوری‌که هر نمونه داده به‌صورت یک سطر شامل مجموعه‌ای از ویژگی‌ها و یک **برچسب** به‌عنوان متغیر هدف نمایش داده می‌شود.

Augmented Data
داده توسعه‌یافته

به مجموعه‌ای از روش‌های محاسباتی اطلاق می‌شود که با ایجاد تغییرات کنترل‌شده و مصنوعی بر روی داده‌های موجود، به افزایش حجم و تنوع داده‌های آموزشی برای مدل‌های یادگیری ماشین و یادگیری عمیق منجر می‌شوند. این فرایند معمولاً شامل اعمال تبدیل‌هایی مانند چرخش، برش، انتقال، تغییر روشنایی، وارونه‌سازی یا افزودن نویز است و با هدف بهبود دقت، پایداری و تعمیم‌پذیری مدل در مواجهه با تغییرات واقعی انجام می‌شود؛ به‌ویژه در حوزه‌هایی مانند بینایی ماشین و تصویربرداری پزشکی، افزایش داده نقش مؤثری در کاهش بیش‌برازش و ارتقای عملکرد مدل ایفا می‌کند.

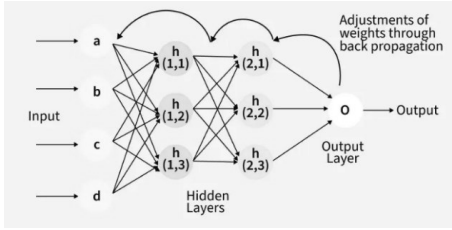
خودمختاری Autonomy

به قابلیت سیستم برای انجام وظایف بدون دخالت مستقیم انسان اشاره دارد. در سیستم‌های یادگیری ماشین و رباتیک، خودمختاری یعنی مدل بتواند تصمیم‌گیری و واکنش به محیط را به صورت مستقل انجام دهد. میزان خودمختاری را می‌توان با سطوح مختلف (کامل، جزئی) تعریف کرد؛ به‌طوری‌که سیستم‌های با خودمختاری بالا قادر به یادگیری مستمر، تصمیم‌گیری تطبیقی و اصلاح عملکرد خود بر اساس بازخورد محیط هستند، در حالی که در سطوح پایین‌تر، درجه استقلال مدل به عواملی مانند میزان وابستگی به داده‌های برچسب‌خورده، استفاده از الگوریتم‌های تنظیم خودکار و توانایی تعمیم به شرایط جدید وابسته است.

Backpropagation
انتشار به عقب

الگوریتمی بنیادی در یادگیری شبکه‌های عصبی مصنوعی است که برای به‌روزرسانی وزن‌ها به‌منظور کاهش خطای پیش‌بینی به‌کار

می‌رود. این روش بر پایه‌ی قانون زنجیره‌ای مشتق‌ها [1] عمل می‌کند و امکان انتقال خطا را از لایه خروجی به لایه‌های پیشین شبکه فراهم می‌سازد. در ابتدا، مدل طی انتشار روبه‌جلو[2] خروجی را تولید کرده و اختلاف میان خروجی پیش‌بینی‌شده و مقدار واقعی به‌صورت تابع هزینه[3] محاسبه می‌شود؛ سپس در مرحله‌ی انتشار به عقب، گرادیان‌های تابع هزینه نسبت به وزن‌ها محاسبه شده و برای تصحیح آن‌ها در جهت نزولی[4] استفاده می‌گردد. این فرایند تکراری موجب می‌شود شبکه به‌تدریج الگوهای داده را یاد بگیرد و انتشار به عقب را به هسته اصلی فرآیند یادگیری در شبکه‌های عصبی چندلایه تبدیل می‌کند.



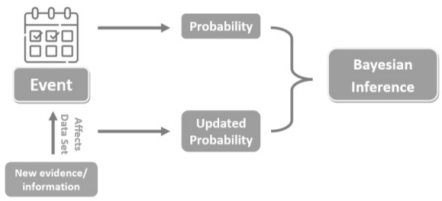
Batch بَـچ

به مجموعه‌ای از نمونه‌های داده گفته می‌شود که به صورت همزمان و طی یک مرحله به مدل داده می‌شوند. اندازه بچ[5] یکی از پارامترهای کلیدی در فرایند آموزش است که بر سرعت همگرایی، پایداری یادگیری و مصرف منابع محاسباتی تأثیر مستقیم دارد. در عمل، به‌جای استفاده از کل داده‌ها به‌طور هم‌زمان، مجموعه آموزشی به بخش‌های کوچک‌تر تقسیم می‌شود تا محاسبات کارآمدتر و پایدارتر انجام گیرد؛ ازاین‌رو، آموزش به‌صورت mini-batch رویکردی رایج است که تعادلی میان کارایی محاسباتی و دقت یادگیری مدل ایجاد می‌کند.

1. Chain rule for derivatives
2. Forward Propagation
3. Loss Function
4. Gradient Descent
5. Batch Size

Bayesian Inference استنتاج بیزی

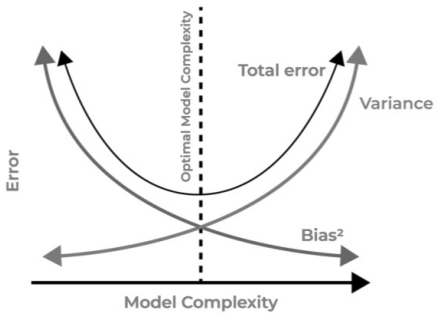
روشی آماری است که با استفاده از قضیه‌ی بیز[6] احتمال یک فرضیه را بر اساس داده‌های مشاهده‌شده به‌روزرسانی می‌کند. در این رویکرد، دانش پیشین یا توزیع پیشین[7] با شواهد جدید ترکیب می‌شود تا توزیع پسین[8] به‌دست آید. این روش به‌جای برآوردهای ثابت، توزیع‌های احتمالی را برای پارامترها ارائه می‌دهد و در نتیجه، عدم‌قطعیت و ابهام داده‌ها را به‌طور صریح مدل می‌کند. این رویکرد در یادگیری ماشین و تصویربرداری پزشکی برای تخمین پارامترها، مدل‌سازی نویز و پشتیبانی از تصمیم‌گیری تطبیقی به‌کار می‌رود و بر این اصل استوار است که یادگیری فرآیندی پویا و مبتنی بر احتمال است.



6. Bayes' Theorem
7. Prior Distribution
8. Posterior Distribution

بایاس Bias

به تفاوت میان پیش‌بینی مدل و واقعیت اشاره دارد که ناشی از خطاهای سیستماتیک در مدل است. در سطح مفهومی، بایاس میزان تمایل مدل به نادیده‌گرفتن پیچیدگی‌های داده برای دستیابی به تعمیم‌پذیری است. بایاس زیاد نشان‌دهنده‌ی مدل ساده و کم‌دقت[1] است، در حالی‌که بایاس کم معمولاً در مدل‌های پیچیده و دقیق‌تر دیده می‌شود،همچنین بایاس در داده‌ها می‌تواند به‌صورت سوگیری جمعیتی یا انتخابی ظاهر شود و منجر به تصمیم‌گیری ناعادلانه گردد. در کل، کنترل تعادل میان بایاس و واریانس یکی از چالش‌های اساسی در طراحی مدل‌های هوش مصنوعی است.



طبقه‌بندی Classification

یکی از فرایندهای اصلی در یادگیری ماشین است که طی آن، مدل داده‌های ورودی را به یکی از چند دسته یا برچسب ازپیش‌تعریف‌شده اختصاص می‌دهد. در این فرآیند، الگوریتم ابتدا بر اساس داده‌های برچسب‌دار آموزش می‌بیند تا الگوهای تفکیک‌کننده‌ی بین کلاس‌ها را بیاموزد و سپس در مرحله آزمون، قادر است داده‌های جدید و ناشناخته را در کلاس مناسب پیش‌بینی کند.به طور کلی طبقه‌بندی می‌تواند باینری[2] (دوتایی) مانند تشخیص سالم/ بیمار یا چندکلاسه[3] مانند شناسایی نوع بافت در تصاویر MRI باشد. در نهایت هدف اصلی این فرایند، دستیابی به مدلی با دقت و تعمیم‌پذیری بالا در تفکیک نمونه‌ها است.

خوشه‌بندی Clustering

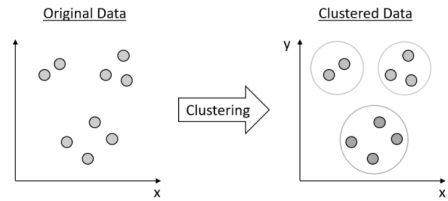
روشی در یادگیری ماشینِ بدون ناظر[4] است که هدف آن، گروه‌بندی داده‌ها بر اساس شباهت ذاتی یا الگوهای پنهان در آن‌ها است و برخلاف طبقه‌بندی، در خوشه‌بندی هیچ برچسب ازپیش‌تعریف‌شده‌ای وجود ندارد و الگوریتم باید ساختار درونی داده را کشف کند. هر خوشه شامل نمونه‌هایی است که بیشترین شباهت درون‌خوشه‌ای و کمترین شباهت بین‌خوشه‌ای را دارند. از الگوریتم‌های متداول می‌توان به K-Means، Hierarchical Clustering و DBSCAN اشاره کرد. این روش در تحلیل داده‌های پزشکی، ژنومیک، و تصویربرداری مغزی برای کشف الگوهای پنهان در بیماران یا نواحی مغزی کاربرد وسیعی دارد.

1. Underfitting

2. Binary

3. Multiclass

4. Unsupervised Learning



Computational Complexity
پیچیدگی محاسباتی

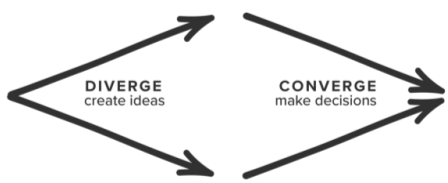
پیچیدگی محاسباتی شاخه‌ای از علوم رایانه است که کارایی یک الگوریتم را از نظر زمان اجرا و میزان حافظه مصرفی بررسی می‌کند این مفهوم معمولاً در قالب دو مؤلفه اصلی، پیچیدگی زمانی[1] و پیچیدگی فضایی[2] بیان می‌شود و هدف آن شناسایی الگوریتم‌هایی است که با حداقل مصرف منابع محاسباتی، به همان نتیجه مطلوب دست یابند. در حوزه یادگیری ماشین، پیچیدگی محاسباتی نقش تعیین‌کننده‌ای در **مقیاس‌پذیری مدل‌ها**، **سرعت آموزش و امکان‌پذیری به‌کارگیری الگوریتم‌ها بر روی مجموعه‌داده‌های بزرگ** ایفا می‌کند و یکی از معیارهای کلیدی در انتخاب و طراحی مدل‌های کارآمد به‌شمار می‌رود.

Confounding Variable
متغیر مخدوش‌کننده

متغیر مخدوش‌کننده عاملی است که به‌صورت هم‌زمان بر متغیر مستقل و متغیر وابسته اثر می‌گذارد و باعث تحریف رابطه‌ی واقعی بین آن‌ها می‌شود. در پژوهش‌های پزشکی یا علوم داده، وجود متغیرهای مخدوش‌کننده می‌تواند منجر به نتایج گمراه‌کننده یا علیت نادرست گردد، برای مثال، در مطالعه اثر فعالیت بدنی بر بیماری قلبی، سن ممکن است متغیر مخدوش‌کننده باشد و کنترل آن از طریق طراحی مطالعه، تصادفی‌سازی، تطبیق[3] یا روش‌های آماری مثل رگرسیون چندمتغیره انجام می‌شود.

Convergence همگرایی

در یادگیری ماشین و بهینه‌سازی به فرآیند نزدیک‌شدن تدریجی پارامترهای مدل به مقدار بهینه اشاره دارد. در این فرایند، الگوریتم‌های یادگیری مانند Gradient Descent با به‌روزرسانی تکرارشونده پارامترها، مقدار **تابع هزینه** را کاهش می‌دهند تا در نهایت به نقطه‌ای برسند که تغییرات پارامترها ناچیز شود. هم‌گرایی معمولاً نشان‌دهنده پایداری و اتمام موفق فرایند یادگیری است، هرچند تضمینی برای دستیابی به کمینه سراسری وجود ندارد. سرعت و کیفیت هم‌گرایی به عواملی نظیر **نرخ یادگیری**، ویژگی‌های داده، تابع هزینه و ساختار مدل بستگی دارد؛ به‌گونه‌ای که هم‌گرایی ناکامل می‌تواند به گیر افتادن در **کمینه‌های محلی**، نوسان در مقدار خطا یا واگرایی الگوریتم منجر شود.



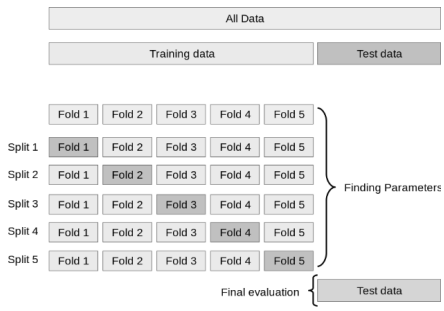
1. Time Complexity
2. Space Complexity
3. Matching

همبستگی Correlation

همبستگی رابطه‌ی آماری میان دو متغیر است که نحوه و شدت تغییر آن‌ها را نسبت به یکدیگر نشان می‌دهد. مقدار همبستگی معمولاً در بازه **۱- تا ۱+** قرار دارد؛ به‌طوری‌که ۱+ نشان‌دهنده همبستگی مثبت کامل و ۱- همبستگی منفی کامل و ۰ حاکی از عدم وجود رابطه آماری معنادار است. همبستگی بالا الزاماً به معنی رابطه‌ی علّی[1] نیست و فقط نوعی ارتباط خطی یا غیرخطی را توصیف می‌کند. برای سنجش همبستگی، شاخص‌های مختلفی از جمله **ضریب همبستگی پیرسون** برای روابط خطی، و **ضرایب رتبه‌ای اسپیرمن** و **کندال** برای روابط یکنواخت غیرخطی به‌کار می‌روند. این مفهوم در یادگیری ماشین و تحلیل داده، به‌منظور بررسی وابستگی ویژگی‌ها، انتخاب متغیرها و تحلیل ساختار داده‌ها اهمیت ویژه‌ای دارد.

اعتبارسنجی متقاطع Cross-validation

روشی آماری برای ارزیابی عملکرد مدل‌های یادگیری ماشین است که با تقسیم داده‌ها به چندین زیرمجموعه آموزشی و آزمونی انجام می‌شود. در این چارچوب، مدل به‌صورت تکرارشونده بر روی بخشی از داده‌ها آموزش داده شده و بر روی بخش دیگر مورد ارزیابی قرار می‌گیرد و در نهایت، **میانگین نتایج حاصل از تکرارها** به‌عنوان برآوردی از عملکرد مدل گزارش می‌شود. متداول‌ترین شکل این روش، **اعتبارسنجی k- بخشی** است که در آن داده‌ها به k بخش تقسیم شده و هر بخش یک‌بار به‌عنوان مجموعه آزمون مورد استفاده قرار می‌گیرد. اعتبارسنجی متقابل با کاهش خطر **بیش‌برازش**، تخمینی پایدار و واقع‌گرایانه از عملکرد مدل بر روی داده‌های جدید فراهم می‌کند و از ابزارهای کلیدی در انتخاب، تنظیم و مقایسه مدل‌ها به‌شمار می‌رود.



داده‌کاوی Data Mining

به فرایند استخراج اطلاعات، الگوها و روابط معنادار از مجموعه‌داده‌های بزرگ و پیچیده با استفاده از روش‌های آماری، الگوریتم‌های یادگیری ماشین و تکنیک‌های مدیریت پایگاه داده اطلاق می‌شود و هدف آن کشف روندهای پنهان، انجام پیش‌بینی و پشتیبانی از تصمیم‌گیری هوشمند بر مبنای داده‌هاست. داده‌کاوی مجموعه‌ای از روش‌ها از جمله **خوشه‌بندی**، **طبقه‌بندی**، رگرسیون[2] و کشف قوانین انجمنی[3] را دربر می‌گیرد و نقش مهمی در تحلیل داده‌های پزشکی، تصویربرداری، علوم زیستی و سامانه‌های پشتیبان تصمیم ایفا می‌کند.

1. Causation

2. Regression

3. Association Rules

Data Preprocessing
پیش‌پردازش داده

به گام ابتدایی و ضروری در فرایند تحلیل داده اطلاق می‌شود که هدف آن آماده‌سازی داده‌های خام برای استفاده در مدل‌های آماری و یادگیری ماشین است. این مرحله شامل مجموعه‌ای از عملیات شامل پاکسازی، نرمال‌سازی، ترکیب و تبدیل داده‌ها، کاهش نویز، حذف داده‌های ناقص و در صورت نیاز استخراج ویژگی است تا داده‌هایی سازگار، قابل‌اعتماد و باکیفیت فراهم شود. پیش‌پردازش مناسب نقش تعیین‌کننده‌ای در بهبود دقت، پایداری و تعمیم‌پذیری مدل‌های نهایی ایفا می‌کند و به‌عنوان پایه اصلی مراحل بعدی تحلیل داده شناخته می‌شود.

Data Split تقسیم‌بندی داده

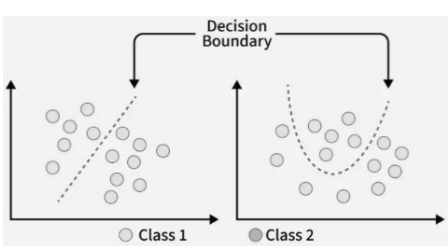
به فرآیند جداسازی مجموعه‌داده‌ی اصلی به زیرمجموعه‌های مستقل آموزش، اعتبارسنجی و آزمون اطلاق می‌شود که با هدف ارزیابی منصفانه و قابل اتکای عملکرد مدل انجام می‌گیرد. به‌طور متداول، داده‌ها با نسبت‌هایی مانند ٪۷۰ برای آموزش، ٪۱۵ برای اعتبارسنجی و ٪۱۵ برای آزمون تفکیک می‌شوند، هرچند این نسبت‌ها بسته به اندازه و ماهیت داده می‌توانند تغییر کنند. مجموعه‌ی آموزش برای یادگیری الگوها و تخمین پارامترهای مدل به کار می‌رود؛ مجموعه‌ی اعتبارسنجی جهت تنظیم اَبَرپارامترها، انتخاب مدل و کنترل پیچیدگی استفاده می‌شود؛ و مجموعه‌ی آزمون صرفاً برای ارزیابی نهایی عملکرد مدل روی داده‌های دیده‌نشده به کار می‌رود. این تفکیک ساختاریافته نقش اساسی در جلوگیری از بیش‌برازش داشته و امکان سنجش توان تعمیم‌پذیری مدل را فراهم می‌کند.

Dataset دیتاست

مجموعه‌ای طبقه‌بندی‌شده از داده‌ها است که در یک قالب ساختاریافته (جدول، تصویر، متن) برای آموزش و آزمون مدل‌های یادگیری ماشین استفاده می‌شود. دیتاست باید ویژگی‌ها و برچسب‌های مرتبط با وظیفه مورد نظر را داشته باشد.

Decision Boundary مرز تصمیم

در مدل‌های طبقه‌بندی به حد یا سطحی در فضای ویژگی‌ها اطلاق می‌شود که تعیین می‌کند هر نمونه‌ی جدید به کدام کلاس تعلق دارد. شکل این مرز می‌تواند خطی یا غیرخطی باشد و به نوع الگوریتم مورد استفاده، مانند ماشین بردار پشتیبان، رگرسیون لجستیک یا شبکه‌های عصبی بستگی دارد. مرز تصمیم ایده‌آل به‌گونه‌ای تعریف می‌شود که خطای تفکیک میان کلاس‌ها را به حداقل برساند و بیشترین توان تمایز را میان داده‌ها فراهم کند. در مسائل چندبعدی، این مرز به‌صورت یک ابرصفحه[1] نمایش داده می‌شود که فضای ویژگی‌ها را به نواحی متناظر با کلاس‌های مختلف تقسیم می‌کند.



1. Hyperplane