\* انجام یادگیری ماشین مشابه روش‌شناسی‌های معمول الگوریتمی است و شامل مراحل مختلفی است که به ترتیب عبارتند از:

1.مشخص کردن مسئله 2. آماده‌سازی داده‌ها 3.انتخاب روش یادگیری 4.انتخاب مدل‌ها 5.اعمال روش یادگیری و آموزش مدل 6. ارزیابی روش و نتایج 7.بهینه‌سازی 8. گزارش نتایج

**\* آماده‌سازی داده‌ها (پیش‌پردازش داده‌ها) یا preprocessing :** پیش‌پردازش داده‌ها برای داشتن داده‌های منظم و معتبر لازم است. داده‌های مرتب و معتبر کلید دستیابی به نتایج قوی و قابل‌اعتماد هستند. به همین دلیل، زمان صرف‌شده برای پیش‌پردازش داده‌ها اهمیت زیادی دارد. برای رسیدن به نتایج دقیق و درست، داشتن داده‌های تمیز و مناسب ضروری است.این مرحله شامل قالب‌بندی صحیح داده‌ها، حذف داده‌های اشتباه، رفع داده‌های گمشده، نمونه‌گیری از مجموعه داده‌ها و موارد مشابه است. همچنین، این مرحله برای کاهش تأثیر خطاها مانند دسته‌بندی‌های نادرست، داده‌های گمشده، اطلاعات پیش‌زمینه نادرست و داده‌های تکراری ضروری است.

\*از جمله نکاتی که باید مورد توجه قرار گیرد : **- \*انتخاب ویژگی‌ها (Feature Selection):** این شامل فیلترینگ و پیدا کردن متغیرهای مهم و تأثیرگذار بر نتایج پروژه است که دانش در آن حوزه اهمیت زیادی دارد، به ویژه در یادگیری نظارت‌شده یا Supervised Learning.  **-\*تبدیل داده‌ها (Data Transforming):** شامل اطمینان از ناشناس بودن داده‌های بیماران، تبدیل متن به عدد برای بهبود تحلیل داده‌ها (مثلاً تبدیل "دیابت دارد" یا "ندارد" به 0 یا 1) و اطمینان از یکسان بودن واحدها در مقیاس بین‌المللی است. **- \*تجمیع (Aggregation):** برای ایجاد ویژگی‌های جدید، مثلاً استفاده از یک کلاس به جای افراد جداگانه. **- \*تجزیه (Decomposition):** برای تقسیم داده‌ها، مثلاً به جای استفاده از نژاد، کشور و زبان به‌عنوان متغیر، این ویژگی‌ها به‌طور جداگانه مطرح می‌شوند.  **- \*مقیاس‌بندی (Scaling):** تنظیم مقیاس متریک‌ها و نشانگرهای بالینی.  **- \*مهندسی ویژگی‌ها (Feature Engineering):** ساخت ویژگی‌های جدیدتر بر اساس ویژگی‌های موجود.  **- \*وزن‌دهی (Weightings):** تعیین اولویت یا تأثیر یک ویژگی خاص.  **- \*مقدار داده‌ها (Amount of Data):** بستگی به پیچیدگی مسئله و الگوریتم دارد، که معمولاً روش‌های غیرخطی و غیرپارامتریک نیاز به داده‌های بیشتر دارند.  **- \*روش یادگیری (Learning Method):** نمایشی از راه‌حلی است که مدل برای حل مسئله از آن استفاده می‌کند و این در الگوریتم‌های مختلف متفاوت است. **- \*انتخاب الگوریتم‌ها(model selection):**انتخاب الگوریتم‌ها بستگی به اندازه، کیفیت و ماهیت داده‌ها، ضرب‌الاجل انجام وظیفه، منابع مالی موجود، میزان مهارت اعضای تیم و هدف طرح دارد.

\*داده‌ها معمولا به سه بخش تقسیم میشوند: داده‌های آموزشی (Train)، داده‌های اعتبارسنجی (Validation) و داده‌های تست (Test). همچنین، استفاده از\*Cross Validation\* برای ارزیابی و بهینه‌سازی ویژگی‌ها ضروری است. علاوه بر این بسیار مهم است که **\*External Validation\*** نیزانجام شود تا مدل روی داده‌های متفاوت از داده‌های اصلی تست شود و تنها به استفاده از داده‌های مشابه نپردازد.

\*هدف یک الگوریتم یادگیری ماشین این است که به‌خوبی تعمیم یابد، به‌طوری که نه دچار کم‌برازش (Underfitting) شود و نه بیش‌برازش (Overfitting). کم‌برازش به این معناست که دقت مدل پایین است و روی داده‌های آموزشی و داده‌های تست به‌خوبی عمل نمی‌کند. بیش‌برازش زمانی اتفاق می‌افتد که مدل روی داده‌های آموزشی به‌خوبی عمل کند، اما در دنیای واقعی و روی داده‌های جدید عملکرد خوبی نداشته باشد، زیرا مدل به‌جای یادگیری از داده‌ها، تنها یاد گرفته است که جزئیات داده‌های آموزشی را حفظ کند و برای کاهش خطاها و جلوگیری از این موارد باید حتما به تعادل بین بایاس و واریانس مدل توجه کرد.

\*در ارزیابی مدل از **\*\*Confusion Matrix\*\*** استفاده می‌شود که معیارهای ارزیابی مانند F1 Score، Precision، Recall و Accuracy را مورد بررسی قرار می‌دهد. در واقع تعمیم یا **generalization** به عملکرد مدل با حداکثر دقت روی نمونه‌هایی اشاره دارد که در حین آموزش دیده نشده‌اند.که با با معیار accuracy سنجیده میشود.



\*تئوری "No Free Lunch" در یادگیری ماشین بیان می‌کند که هیچ الگوریتم واحدی وجود ندارد که برای تمام مسائل بهترین باشد. پیش‌بینی اینکه کدام رویکرد بهترین عملکرد را روی داده‌ها خواهد داشت تقریباً غیرممکن است. به همین دلیل، معمولاً چندین مدل بررسی می‌شوند و بهترین مدل برای رسیدن به نتیجه مطلوب انتخاب می‌شود.

\*فقط به این دلیل که یک مدل خوب عمل می‌کند، به این معنا نیست که بهترین و تنها مدل است. بعد از انجام این موارد از راه های مختلف میتوان برای برای بهینه‌سازی الگوریتم‌ها(optimization)، از روش‌های مختلفی استفاده کرد، از جمله : - تنظیم مدل با استفاده از تغییر هایپر پارامترها - ارزیابی مجموعه‌ای از روش‌ها و الگو ها - به‌دست آوردن داده‌های بیشتر و باکیفیت تر - نمونه‌برداری مجدد از داده‌ها - انتخاب ویژگی‌ها - مهندسی ویژگی‌ها

\*گزارش نتایج شامل موارد زیر می باشد : دلیل انجام مطالعه و طرح مسئله ، راه‌حل و مدل آموزش داده‌ شده ، یافته‌ها ، معیارهای ارزیابی و عملکرد ، نتایج نهایی.