در یادگیری نظارت‌شده، دو رویکرد اصلی وجود دارد: *سیستم‌های مبتنی بر *Regression* *سیستم‌های مبتنی بر *Classification*

- رگرسیون *(Regression)*: برای پیش بینی متغیرهای کمی و پیوسته استفاده می‌شود. این روش ضریب و شیب خط را محاسبه می‌کند .در رگرسیون سری زمانی، بعد زمان نیز اضافه می‌شود. به‌عنوان مثال، از قند خون قبلی بیمار می‌توان برای پیش‌بینی هموگلوبین *A1C* استفاده کرد یا از علائم بالینی برای پیش‌بینی بدتر شدن وضعیت بیمار استفاده کرد.یا ریسک عوارض قلبی عروقی مریض را بر اساس عوامل بالینی فورکست کرد.در یک سیستم مبتنی بر رگرسیون، عملکرد با ارزیابی تعداد خطاهای پیش‌بینی اندازه‌گیری می‌شود.عملکرد مدل با معیارهایی مانند **ضریب تعیین** *(R^2)*، **میانگین خطای مطلق** *(MAE)*، **خطای نسبی مطلق** *(RAE)*، و **ریشه میانگین مربعات خطا** *(RMSE)* ارزیابی می‌شود.

*در **رگرسیون خطی** *(linear regression)*، هدف تنظیم فاصله بین نقاط واقعی و پیش‌بینی‌شده برای حداقل کردن خطاست. خطا به صورت مجموع مربعات تفاضل مقادیر واقعی و پیش‌بینی‌شده محاسبه می‌شود. مدل می‌تواند خطی *(درجه یک)* یا غیرخطی *(درجه دو یا سه)* باشد که دقت بیشتری ایجاد می‌کند. شرط برابری واریانس‌ها در بسیاری از داده‌ها برقرار نیست و نویز در مدل‌های خطی می‌تواند بالا باشد.مدل خطی که به صورت $f(x)=ax+b$ تعریف می‌شود، ضرایب a و b به‌گونه‌ای تنظیم می‌شوند که تخمین به دقیق‌ترین و نزدیک‌ترین مقدار ممکن برسد. این کار به این صورت انجام می‌شود که مدل در ابتدا دو عدد تصادفی به ضرایب اختصاص می‌دهد و طی فرایند آموزش *(Training)*، مقادیر را به‌روزرسانی می‌کند تا به مقادیر مطلوب برسد. **مولتیپل رگرسیون یا مولتی وریبل** شامل چند متغیر مستقل است، درحالی‌که **مولتی‌وریت آنالیز** به معنای داشتن چندین متغیر وابسته به عنوان پیامد است.

<u>قانون اعداد بزرگ</u> در یادگیری ماشین نشان می‌دهد که نتیجه تعداد زیادی آزمایش مشابه به یک مقدار مشخص همگرا می‌شود. هدف، به حداقل رساندن خطای مدل و ارائه دقیق‌ترین پیش‌بینی‌هاست.این فرآیند تکرار می‌شود تا سیستم به بهترین مقادیر برای a و b برسد. به این ترتیب، ماشین از طریق تجربه یاد می‌گیرد و برای استفاده آماده می‌شود.در بسیاری از زمینه‌های یادگیری نظارت‌شده، هدف تنظیم دقیق تابع پیش‌بینی‌کننده همان فرضیه میباشد. یادگیری شامل استفاده از الگوریتم‌های ریاضی برای نمایش داده‌های ورودی x در یک حوزه خاص است. به عنوان مثال، x می‌تواند نشان‌دهنده روز باشد وتابع می‌تواند پیش‌بینی مدت زمان انتظار در یک بیمارستان باشد.

دو رویکرد اصلی برای انتخاب ویژگی‌ها وجود دارد.

1. **روش فیلتر** ***(Filter Method)***: ویژگی‌ها به‌طور مستقل و بدون وابستگی به داده‌ها ارزیابی و قبل از ساخت مدل فیلتر می‌شوند.این روش بر اساس الگوریتمی ثابت عمل می‌کند و وابسته به ویژگی‌های خاص داده‌ها نیست.
2. **روش روکش‌دار** ***(Wrapper)***: با استفاده از الگوریتم یادگیری ماشین، زیرمجموعه‌ای از داده‌ها بر اساس دقت مدل انتخاب می‌شود. این روش مشخص می‌کند که کدام متغیرها و آستانه‌ها *(Cutoff)* برای بهینه‌سازی مدل استفاده شوند، به‌طوری‌که دقت و تقاطع مدل با خط y بهینه باشد.

*رگرسیون لجستیک *(logistic regression)*: احتمال وقوع را بین صفر و یک قرار می‌دهد و داده‌ها را به دو دسته طبقه‌بندی می‌کند. اسم آن رگرشن می باشد ولی در واقع برای تقسیم بندی استفاده میشود دقیقا چون لگاریتم کار میکند بین صفر و یک می باشد و مثلاً بیان میکند که اگر احتمال کمتر از نیم باشد به سمت صفر میرود و اگر احتمال بیشتر از نیم باشد به سمت یک میرود و می‌تواند در دو گروه صفر و یک طبقه‌بندی کند.

- طبقه‌بندی *(Classification)*: برای دسته بندی کردن متغیرهای کیفی استفاده می‌شود. به‌عنوان مثال، در مورد هموگلوبین *A1C*، مقدار ۷.۸ درصد به‌عنوان دیابت در نظر گرفته می‌شود چون بالاتر از کات اف تعیین شده میباشد همچنین برای تشخیص وضعیت‌های مختلف مانند پیامدهای دو یا چند حالته مثل بهبودی و عدم بهبودی ، دیابتی بودن یا نبودن، مرگ و زندگی می‌توان از این مورد استفاده کرد یا اینکه بر اساس پروفایل سلامت تشخیص بدیم آیا رتینوپاتی دیابتی وجود دارد یا ندارد. مناسب‌ترین روش درمانی چیست مثلاً ادامه درمان است صبر کردن هست یا تیپر کردن دار..دسته‌بندی فرآیندی است که در آن یک یا چند آیتم *(x)* به یکی از دسته‌های از پیش تعریف‌شده و گسسته *(یا برچسب‌ها، y)* تخصیص داده می‌شود.الگوریتم‌های دسته‌بندی از ویژگی‌ها *(یا صفات)* برای تعیین چگونگی تخصیص یک آیتم به حداقل یک کلاس دودویی *(دو دسته)* یا دو یا چند کلاس استفاده می‌کنند.داده‌های آموزشی به صورت یک بردار ورودی ارائه می‌شوند. این بردار شامل برچسب‌هایی برای ویژگی‌های آیتم است. به عنوان مثال، یک آیتم ممکن است به این صورت تعریف شود: *{"HbA1c, 7.8%", "دیابتی"}* به شکل *{x_training, output}*

***الگوریتم‌هایی که فرآیند دسته‌بندی را انجام می‌دهند شامل درخت‌های تصمیم‌گیری*(Decision Tree)*،جنگل تصادفی *(Random Forest)* بیز ساده *(Naïve Bayesian)*، رگرسیون لجستیک، نزدیک‌ترین همسایگان *(kNN)*، ماشین بردار پشتیبان *(SVM)*، و غیره هستند.

درخت تصمیم *(Decision Tree)*:

درخت تصمیم یک نمودار درختی است که با تقسیم داده‌های بزرگ به بخش‌های کوچک‌تر و تعیین قوانین *if-then*، تصمیم‌گیری را انجام می‌دهد. .ساختار درخت شامل ریشه به‌عنوان گره اولیه، شاخه‌ها برای گسترش درخت، و برگ‌ها به‌عنوان گره‌های نهایی است که نتیجه تصمیم‌گیری را مشخص می‌کنند.یکی از مشکلات درخت تصمیم، افزایش گره‌ها و شاخه‌ها است که احتمال بیش‌برازش *(Overfitting)* و پیچیدگی محاسبات را افزایش می‌دهد. برای رفع این مشکل از فرایند هرس *(Pruning)* استفاده می‌شود.سه الگوریتم اصلی برای ساخت درخت تصمیم وجود دارد که تفاوت آن‌ها در معیار هزینه‌ای است که برای انتخاب گره‌ها یا ویژگی‌ها استفاده می‌شود.یکی از ویژگی آنها این است که اولویت‌ها و روابط در تصمیم‌گیری به وضوح مشخص می باشند و خروجی به شکل مجموعه‌ای از قوانین بیان می شود. در نهایت، این درخت‌ها فرآیند تصمیم‌گیری را به‌صورت قوانین شفاف و قابل‌اجرا نمایش می‌دهند.

جنگل تصادفی *(Random Forest)* :

جنگل تصادفی به‌جای یک درخت تصمیم، از مجموعه‌ای از درخت‌ها تشکیل شده که هر کدام متغیری خاص را نشان می‌دهند.این روش به‌عنوان یک تکنیک تجمعی *(Ensemble)* شناخته می‌شود، زیرا با ترکیب پیش‌بینی‌های آن‌ها دقت مدل را افزایش و واریانس را کاهش می‌دهد. در رگرسیون، میانگین پیش‌بینی‌ها و در طبقه‌بندی، دسته‌ای که اکثریت آرا را دارد نتیجه نهایی را تعیین می‌کنند. این روش نسبت به درخت تصمیم دقت بیشتری دارد، کمتر دچار بیش‌برازش می‌شود، با داده‌های ناقص بهتر عمل می‌کند و تخمین‌های بهتری ارائه می‌دهد.در واقع، هر درخت در جنگل تصادفی به‌طور مستقل عمل کرده و پیش‌بینی خاص خود را ارائه می‌دهد، اما نتیجه نهایی بر اساس ترکیب این پیش‌بینی‌ها مشخص می‌شود جنگل تصادفی رویکردی قوی‌تر و مناسب برای مسائل پیچیده طبقه‌بندی و رگرسیون با متغیرهای متنوع است.

مدل‌های یادگیری تجمعی *(Ensemble Learning)*

مدل‌های ترکیبی *(Ensemble Methods)* برای بهبود دقت پیش‌بینی از ترکیب چند مدل استفاده می‌کنند وبه‌طور همزمان از مزایای مدل‌های مختلف بهره‌برداری کنند و در نتیجه عملکرد بهتری نسبت به مدل‌های تک‌بعدی داشته باشند. شامل دو روش اصلی ***Boosting*** و ***Bagging*** هستند:

1. ***Boosting***: در این روش، مدل‌ها به‌صورت ترتیبی آموزش داده می‌شوند و هر مدل جدید سعی می‌کند خطاهای مدل قبلی را اصلاح کند.در هر مرحله، مدل جدید نقاطی که مدل قبلی قادر به پیش‌بینی صحیح آن‌ها نبوده را شناسایی کرده و سعی می‌کند این نقاط را درست پیش‌بینی کند. این فرآیند باعث افزایش تدریجی دقت و کاهش بیش‌برازش *(Overfitting)* می‌شود. ویژگی بارز *Boosting*، یادگیری مدل جدید از خطاهای مدل قبلی است. یکی از تکنیک‌های پیشرفته در این حوزه ***Gradient Boosting*** است که از درخت‌های تصمیم‌گیری و پیش‌بینی‌کننده‌های ضعیف برای کاهش خطا استفاده می‌کند. نسخه پیشرفته‌تر آن ***XGBoost***، با اضافه کردن منظم‌سازی *(Regularization)*، قادر است از **واریانس بالا** جلوگیری کرده و دقت پیش‌بینی را بهبود ببخشد.عملکرد بهتری در داده‌های پیچیده دارد.
2. ***Bagging***: هدف این روش کاهش واریانس و جلوگیری از بیش‌برازش است. مدل‌ها به‌طور مستقل و موازی آموزش داده می‌شوند و پیش‌بینی نهایی از میانگین یا ترکیب نتایج آن‌ها به‌دست می‌آید. این روش از نمونه‌گیری بوت‌استرپ برای آموزش مدل‌ها استفاده می‌کند، که به‌طور تصادفی داده‌ها را برای آموزش مدل‌ها انتخاب می‌کند . باعث کاهش رفتارهای غیرعادی در پیش‌بینی می‌شود.زیرا پیش‌بینی‌ها از مدل‌های مختلف با داده‌های متفاوت میانگین‌گیری می‌شوند. ***Random Forest*** یکی از معروف‌ترین الگوریتم‌های *Bagging* است که از ترکیب درخت‌های تصمیم بهره می‌گیرد.

ماشین بردار پشتیبان *(SVM)(Support Vector Machine)* :

SVM یک الگوریتم طبقه‌بندی خطی باینری غیر احتمالاتی است که برای مسائل طبقه‌بندی و رگرسیون به کار می‌رود. این الگوریتم با یافتن ابرصفحه‌ای که بهترین تفکیک را بین دو کلاس دارد عمل می‌کند. بردارهای پشتیبان، نقاط داده نزدیک به ابرصفحه هستند که اگر حذف شوند، موقعیت ابرصفحه تغییر می‌کند هرچه مقدار حاشیه یا فاصله نقاط داده از ابرصفحه بیشتر باشد، اطمینان بیشتری وجود دارد که داده‌ها به درستی طبقه‌بندی شده‌اند. هدف *SVM* بیشینه کردن حاشیه یا فاصله نقاط داده از ابرصفحه است تا اطمینان بیشتری از طبقه‌بندی صحیح داشته باشد.این الگوریتم از "هسته‌سازی" برای نگاشت داده‌ها به فضاهای با ابعاد بالاتر و تسهیل طبقه‌بندی استفاده می‌کند. داده‌ها با استفاده از ترفند هسته به صورت تکراری به ابعاد بالاتری نگاشت می‌شوند تا یک ابرصفحه برای طبقه‌بندی آن‌ها تشکیل شود این روش در مجموعه داده‌های کوچک و در مسائلی مانند تحلیل متن، تحلیل احساسات، تشخیص تصویر و شناسایی دست‌نوشته‌های عددی عملکرد خوبی دارد. اما با افزایش ابعاد داده‌ها، قابلیت درک مدل کاهش یافته و زمان آموزش افزایش می‌یابد. همچنین در مواجهه با داده‌های پر نویز عملکرد کمتری دارد.

نایو بیز *(Naïve Bayes)*:

نایو بیز الگوریتمی برای محاسبه احتمال وقوع یک رویداد بر اساس وقوع رویدادی دیگر است. این روش با فرض استقلال متغیرها، که در واقعیت کمتر دیده می‌شود، عمل می‌کند و برای مجموعه داده‌های با ابعاد بالا کاربرد دارد.این الگوریتم از فرمول زیر استفاده می‌کند. *(P(A|B) = (P(B|A) * P(B)) / P(A))*با این فرمول می‌توان احتمال تعلق یک نمونه به یک کلاس را تعیین کرد؛ مثلاً احتمال خطر بیماری با توجه به وضعیت سلامتی بیمار محاسبه می‌شود و کلاسی که بیشترین احتمال را دارد، انتخاب می‌گردد.نایو بیز ساده، اما بسیار کارآمد است مدل‌های بایس ساده به راحتی ساخته می‌شوند و برای مجموعه داده‌های بزرگ بسیار مفید هستند. علی‌رغم سادگی ، بایس ساده به خاطر عملکرد بهتر نسبت به روش‌های طبقه‌بندی پیچیده‌تر شناخته شده است.

نزدیک ترین همسایه *(K-NN(K-Nearest Neighbour)*:

نباید روش *kNN* را با خوشه‌بندی *k-means* اشتباه گرفت. روش *kNN* یک شیء ناشناخته را با برچسب اکثریت نزدیک‌ترین *k* همسایگان طبقه‌بندی می‌کند. در واقع یک تکنیک غیر پارامتری برای طبقه‌بندی و رگرسیون است که به جای یادگیری مدل، داده‌های آموزشی را ذخیره کرده و طبقه‌بندی نمونه جدید را بر اساس یادگیری با قیاس انجام می‌دهد. از آنجا که مدلی یاد گرفته نمی‌شود، *kNN* به عنوان یک یادگیرنده تنبل شناخته می‌شود.فاصله بین شیء و همسایه‌ها با استفاده از فاصله اقلیدسی محاسبه می‌شود.در واقع الگوریتم فاصله بین هر دو نقطه را محاسبه کنید.نزدیک‌ترین همسایگان را بر اساس این فواصل زوجی پیدا کنید.برای تعیین برچسب کلاس، بر اساس لیست نزدیک‌ترین همسایگان رای‌گیری اکثریت انجام دهید.پیش‌بینی در زمان درخواست انجام می‌شود. در مسائل رگرسیون، میانگین یا میانه *k*-مشابه‌ترین نمونه‌ها استفاده می‌شود. در مسائل طبقه‌بندی، کلاسی که بیشترین فراوانی را در میان *k*-مشابه‌ترین نمونه‌ها دارد، انتخاب می‌شود

از معایب آن می‌توان به موارد زیر اشاره کرد. *1)* نیاز به محاسبات سنگین برای هر پیش‌بینی چون برای طبقه‌بندی یک شیء، فاصله همه همسایگان در مجموعه داده آموزشی باید محاسبه شود *2)* عدم سازگاری با داده‌های با ابعاد بالا چون هر متغیر پیش‌بین به‌عنوان یک بعد از فضای ورودی در نظر گرفته می‌شود. افزایش ابعاد به‌صورت نمایی حجم فضای ورودی را افزایش می‌دهد .*3)* عدم توانایی در مواجهه با داده‌های ناقص اشاره کرد، زیرا فاصله بین بردارها در صورت وجود داده‌های ناقص نمی‌تواند محاسبه شود

<u>نتیجه‌گیری</u>: مدل‌های مختلف هر کدام نقاط قوت و ضعف خودشان را دارند و در شرایط خاصی کاربرد دارند و انتخاب بهترین مدل به نوع داده‌ها و اهداف تحقیق بستگی دارد.