*** دو روش برای ارزیابی مدل های یادگیری ماشین وجود دارد :**ارزیابی آفلاین ،ارزیابی آنلاین**
ارزیابی آفلاین مدل را بر اساس معیارهایی که از داده‌های تاریخی، ایستا و توزیع‌شده یاد گرفته و ارزیابی شده‌اند، اندازه‌گیری می‌کند. معیارهایی مانند دقت و *precision-recall* معمولاً در مرحله آموزش آفلاین استفاده می‌شوند. تکنیک‌های ارزیابی آفلاین شامل روش *hold-back* و اعتبارسنجی متقاطع *n-fold* هستند.

ارزیابی آنلاین به ارزیابی معیارها پس از استقرار مدل اشاره دارد. نکته کلیدی این است که این معیارها ممکن است با معیارهای مورد استفاده برای ارزیابی عملکرد مدل در حالت استقرار زنده متفاوت باشند. ارزیابی آنلاین، به ویژه در عصر دیجیتال، می‌تواند از تست‌های چندمتغیره برای درک بهترین مدل‌های عملکردی پشتیبانی کند.حلقه‌های بازخورد *(Feedback Loops)* برای اطمینان از عملکرد درست و مورد نظر سیستم ها کلیدی هستند و به درک بهتر مدل در زمینه استفاده از آن کمک می‌کنند. این بازخوردها می‌توانند از طریق یک عامل انسانی، یک عامل خودکار هوشمند مبتنی بر زمینه ویا کاربران مدل جمع‌آوری شوند.

*** چگونه می‌توانم مدل خود را ارزیابی کنم؟
تعیین معیارهای ارزیابی مناسب برای سنجش عملکرد مدل بر اساس هدفی که مدل برای آن طراحی شده است،میباشد. اغلب، داده‌های استفاده‌شده از انتخاب الگوریتم مهمتر هستند؛ هرچه ویژگی‌های استفاده‌شده بهتر باشند، عملکرد مدل نیز بهبود می‌یابد.برای معیار های ارزیابی از کتابخانه‌هایی مانند پکیج *metrics* در *R* و *scikit-learn* در پایتون می توان استفاده کرد.

- **<u>برای ارزیابی مدل‌های طبقه‌بندی *(Classification)*، معیارهای رایج عبارتند از:</u>**

** دقت *(Accuracy)*** سادەترین تکنیک برای شناسایی این است که آیا یک مدل پیش‌بینی‌های صحیح انجام می‌دهد یا خیر. این معیار به‌عنوان درصد پیش‌بینی‌های صحیح از کل پیش‌بینی‌های انجام‌شده محاسبه می‌شود. *Accuracy = number of correct predictions/number of total predictions*

** ماتریس درهم‌ریختگی *(Confusion Matrix)*** طبقه‌بندی‌های صحیح و نادرست مدل را تفکیک کرده و آن‌ها را به برچسب‌های مناسب نسبت می‌دهد. در نتیجه، دقت را می‌توان به شکل زیر بازنویسی کرد. *(TP + TN)/(TP + TN + FP + FN)*

** صحت *(Precision)*** بررسی می‌کند که از بین مواردی که مدل به‌عنوان مرتبط پیش‌بینی کرده است، چند مورد واقعاً مرتبط هستند.
** فراخوانی *(Recall)*** بررسی می‌کند که از بین مواردی که واقعاً مرتبط هستند، مدل چند مورد را به‌درستی پیش‌بینی کرده است.
Precision: (correctly predicted Positive)/(total predicted Positive) = TP/TP + FP
Recall: (correctly predicted Positive)/(total correct Positive observation) = TP/TP + FN
- ** حساسیت *(Sensitivity)*:** این معیار همان فراخوانی *(Recall)* است و نشان می‌دهد که از بین موارد واقعاً مثبت، مدل چند مورد را به درستی شناسایی کرده است.
- ** اختصاصیت *(Specificity)*:** این معیار بررسی می‌کند که مدل چقدر خوب می‌تواند موارد منفی را به درستی شناسایی کند.
Specificity: (correctly predicted Negative)/(total Negative observation) = TN/TN + FP

** معیار *F-measure/F1 score*:** این معیار فراتر از میانگین حسابی رفته و میانگین هارمونیک صحت و فراخوانی را محاسبه می‌کند که در آن *p* نشان‌دهنده دقت *(Precision)* و *r* نشان‌دهنده فراخوانی *(Recall)* است. این معیار برای متعادل‌سازی دقت و فراخوانی در مواردی که توزیع کلاس‌ها نامتوازن است، مفید است.

$$F = \frac{1}{\frac{1}{2}\left(\frac{1}{p}+\frac{1}{r}\right)} = \frac{2pr}{p+r}$$

** لگاریتم زیان *(Logarithmic Loss* (یا *Log-loss*)** برای مسائلی استفاده می‌شود که در آن‌ها یک احتمال پیوسته پیش‌بینی می‌شود، نه یک برچسب کلاس. *Log-loss* یک معیار احتمالی برای سنجش اطمینان از دقت ارائه می‌دهد و آنتروپی بین توزیع برچسب‌های واقعی و پیش‌بینی‌ها را در نظر می‌گیرد. برای یک مسئله طبقه‌بندی دودویی به صورت زیر محاسبه می‌شود :

$$\text{Log-loss} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p_i + (1-y_i)\log(1-p_i)$$

** AUC (مساحت زیر منحنی)** نرخ مثبت‌های صحیح *(True Positives)* را در مقابل نرخ مثبت‌های کاذب *(False Positives)* رسم می‌کند. این معیار امکان تجسم حساسیت *(Sensitivity)* و اختصاصیت *(Specificity)* طبقه‌بندی را فراهم می‌کند و نشان می‌دهد که چند طبقه‌بندی مثبت صحیح می‌توان با تحمل مثبت‌های کاذب به دست آورد. .این منحنی به عنوان** منحنی مشخصه عملکرد سیستم *(Receiver Operating Characteristic Curve)* یا *ROC* شناخته می‌شود

- **<u>برای ارزیابی مدل‌های رگرسیون *(Regression)*، معیارهای رایج عبارتند از:</u>**

** *(Root-Mean-Squared Error (RMSE .1**** این معیار جذر میانگین مربعات خطاها است و میزان اختلاف بین مقادیر پیش‌بینی‌شده و مقادیر واقعی را اندازه‌گیری می‌کند. *RMSE* به خطاهای بزرگ حساس‌تر است، زیرا خطاها قبل از میانگین‌گیری به توان ۲ می‌رسند.

$$RMSE = \sqrt{(1/n) * \Sigma(y_i - \hat{y}_i)^2}$$

** *(Mean-Squared Error (MSE .2**** این معیار میانگین مربعات خطاها است و مشابه *RMSE* عمل می‌کند، اما جذر گرفته نمی‌شود. *MSE* نیز به خطاهای بزرگ حساس است.

$$MSE = (1/n) * \Sigma(y_i - \hat{y}_i)^2$$

** *(Mean-Absolute Error (MAE .3**** این معیار میانگین قدر مطلق خطاها است و نسبت به *RMSE* و *MSE* به خطاهای بزرگ حساسیت کمتری دارد، زیرا خطاها به توان ۲ نمی‌رسند.

$$MAE = (1/n) * \Sigma|y_i - \hat{y}_i|$$

4. Median Absolute Percentage Error (MAPE) این معیار میانه درصد خطاها را محاسبه می‌کند و برای داده‌هایی که دارای مقادیر پرت (Outliers) هستند، مناسب است.

$$MAPE = median\left(\left|\left(y_t - \hat{y}_t\right) / \left(y_t\right)\right|\right),$$

5. R^2 (ضریب تعیین) این معیار نشان می‌دهد که مدل چقدر از واریانس داده‌ها را توضیح می‌دهد. مقدار R^2 بین ۰ و ۱ است. 1: مدل تمام واریانس داده‌ها را توضیح می‌دهد. 0: مدل هیچ واریانسی را توضیح نمی‌دهد.

$$R^2 = 1 - (SSR/SST) = 1 - \Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2$$

*** یک دانشمند داده باتجربه همه داده‌ها را با شک و تردید بررسی می‌کند.داده‌های نامتعادل (Skewed Datasets)، anomalies (ناهنجاری‌ها)، داده‌های نادر (Rare Data)، داده‌های ناسازگار (Inconsistent)، نمونه‌های کلاس نامتعادل (Imbalanced Class Examples)، و outliers (مقادیر پرت) همگی می‌توانند به طور قابل توجهی بر عملکرد مدل تأثیر بگذارند. وجود نمونه‌های بیشتر در یک کلاس نسبت به کلاس دیگر می‌تواند منجر به عملکرد ضعیف مدل شود.تأثیر outliers بزرگ را می‌توان با استفاده از درصدهای خطا (percentiles of error) کاهش داد.پاکسازی خوب داده‌ها، حذف outliers، و نرمال‌سازی متغیرها از جمله اقدامات مهم برای بهبود کیفیت داده‌ها هستند.

<u>تنظیم هایپر پارامترها (Hyperparameter Tuning)</u>

تنظیم هایپر پارامترها به معنای انتخاب مجموعه‌ای از هایپر پارامترهای بهینه برای یک مدل یادگیری ماشین است. مقادیر بهینه هایپر پارامترها دقت پیش‌بینی مدل را به حداکثر می‌رسانند. این فرآیند شامل آموزش مدل، ارزیابی دقت کلی و تنظیم مناسب هایپر پارامترها است. با آزمایش مقادیر مختلف هایپر پارامترها، بهترین مقادیر برای مسئله مشخص می‌شوند که دقت کلی مدل را بهبود می‌بخشند. نمونه‌هایی از هایپر پارامترها در شبکه‌های عصبی (Neural Networks) تعداد و اندازه لایه‌های پنهان، وزن‌ها، نرخ یادگیری (Learning Rate) و غیره و یا درخت‌های تصمیم (Decision Trees) عمق درخت و تعداد برگ‌ها میباشد.تنظیم هایپر پارامترها شبیه به آموزش یک مدل یادگیری ماشین است و به عنوان یک مسئله بهینه‌سازی در نظر گرفته می‌شود.رایج‌ترین روش‌های تنظیم هایپر پارامترها 1. جستجوی شبکه‌ای (Grid Search): بررسی تمام ترکیبات ممکن از هایپر پارامترها. 2. جستجوی تصادفی (Random Search): بررسی تصادفی ترکیبات مختلف از هایپر پارامترها میباشد.حالت اول بسیار دقیق تر عمل میکند اما حالت دوم سریع تر و با محاسبات کمتر این کار را انجام میدهد.

** آزمون چندمتغیره (Multivariate Testing)** یک روش بسیار مفید برای تعیین بهترین مدل برای مسئله خاص است. این روش که به عنوان* آزمون فرضیه آماری* نیز شناخته می‌شود، تفاوت بین فرضیه صفر (Null Hypothesis) و فرضیه جایگزین (Alternative Hypothesis) را بررسی می‌کند. -**فرضیه صفر:** مدل جدید تأثیری بر میانگین مقدار معیار عملکرد ندارد. -**فرضیه جایگزین:** مدل جدید میانگین مقدار معیار عملکرد را تغییر می‌دهد. این روش مدل‌های مشابه را مقایسه می‌کند تا بهترین عملکرد را شناسایی کند یا مدل جدید را با مدل قدیمی‌تر مقایسه می‌نماید. معیارهای عملکرد مربوطه مقایسه شده و تصمیم گرفته می‌شود که با کدام مدل ادامه دهند.

اگرچه این فرآیند ساده به نظر می‌رسد، اما چند نکته کلیدی برای توجه وجود دارد.

کدام معیار را برای ارزیابی استفاده کنم؟انتخاب معیار مناسب برای ارزیابی مدل به مورد استفاده بستگی دارد. تکرار آزمایش و انجام ارزیابی‌های مکرر می‌تواند احتمال نتایج گمراه‌کننده را کاهش دهد.

همبستگی به معنای علیت نیست.همبستگی بین دو متغیر به این معنا نیست که یکی باعث دیگری می‌شود.در مدل‌های با چندین ویژگی، ممکن است عوامل پنهانی وجود داشته باشند که باعث حرکت همزمان دو متغیر می‌شوند.

چه مقدار تغییر به عنوان تغییر واقعی در نظر گرفته می‌شود؟ تعیین مقدار تغییر مورد نیاز برای رد فرضیه صفر به مورد استفاده بستگی دارد. بهتر است در ابتدای پروژه یک مقدار مشخص تعیین کنید و به آن پایبند باشید.

چند مشاهده مورد نیاز است؟ تعداد مشاهدات مورد نیاز به قدرت آماری مورد نیاز پروژه بستگی دارد. بهتر است این مقدار در ابتدای پروژه تعیین شود.

مدت زمان اجرای آزمون چند متغیره چه قدر باید باشد؟ مدت زمان اجرای آزمون باید به اندازه‌ای باشد که تعداد مشاهدات کافی برای دستیابی به قدرت آماری تعیین‌شده را فراهم کند.

شناسایی تغییرات توزیع (Distribution Drift):پس از استقرار مدل، نظارت بر عملکرد مدل بسیار مهم است. تغییرات در داده‌ها و توسعه سیستم نیازمند تایید مدل در مقایسه با خط پایه است. اگر تغییر قابل توجهی در معیار اعتبار سنجی مشاهده شود، مدل باید با داده‌های جدید آموزش داده شود.

ثبت تمام تغییرات مدل و یادداشت‌های مربوط به آن‌ها نه تنها به عنوان یک گزارش تغییر برای ذینفعان مفید است، بلکه سابقه‌ای فیزیکی از نحوه تغییر سیستم در طول زمان ارائه می‌دهد.