**فصل ١١. ارزیابی مدل**

**١١.٠. مقدمه**

در این فصل به بررسی استراتژی‌های ارزیابی کیفیت مدل‌های ایجاد شده از طریق الگوریتم­های یادگیری می‌پردازیم. شاید عجیب به نظر برسد که در مورد ارزیابی مدل، قبل از بحث در مورد چگونگی ایجاد آنها صحبت کنیم اما روشی برای رفع این سردرگمی وجود دارد. مدل­ها تنها به اندازه‌ی کیفیت پیش‌بینی­هایشان مفید هستند و بنابراین اساساً هدف ما ایجاد مدل(که آسان است) نیست، بلکه ایجاد مدل­های با کیفیت بالا (که سخت است) است. بنابراین، قبل از اینکه هزاران الگوریتم یادگیری را بررسی کنیم، ابتدا نحوه ارزیابی مدل­های تولید شده توسط آن­ها را بررسی می­کنیم.

**١١.١. مدل­های اعتبار سنجی متقاطع (متقابل)**

**مسئله**

می­خواهید ارزیابی کنید که مدل شما در دنیای واقعی چقدر خوب کار خواهد کرد.

**راه حل**

یک خط انتقال داده ایجاد کنید که داده­ها را پیش‌پردازش می­کند، مدل را آموزش می­دهد و سپس آن را با استفاده از اعتبارسنجی متقابل ارزیابی می­کند:



****

**بحث**

در نگاه اول، ارزیابی مدل­های یادگیری تحت نظارت ممکن است ساده به نظر برسد به این صورت که یک مدل را آموزش دهید و سپس با استفاده از معیارهای عملکرد (دقت، مربع خطاها و غیره) میزان صحت عملکرد آن را بررسی کنید. با این حال، این رویکرد اساساً ناقص است. اگر مدلی را با استفاده از داده­های خود آموزش دهیم و سپس ارزیابی کنیم که چقدر روی آن داده­ها خوب عمل کرده است، به هدف مورد نظر خود نمی­رسیم. هدف ما ارزیابی این نیست که مدل تا چه حد بر روی داده­های آموزشی ما خوب عمل می­کند، بلکه این است که تا چه حد بر روی داده­هایی که قبلا ندیده است (به عنوان مثال، یک مشتری جدید، یک جرم جدید، یک تصویر جدید) خوب عمل می­کند. به همین دلیل، روش ارزیابی باید به ما کمک کند تا بفهمیم مدل­ها تا چه حد قادر به پیش‌بینی از داده­هایی هستند که تاکنون ندیده است.

یک استراتژی ممکن است نگه داشتن بخشی از داده­ها برای آزمایش باشد. این روش اعتبار سنجی(نگه داشتن) نامیده می‌شود. در اعتبارسنجی(validation)، مشاهدات ما (ویژگی‌ها و اهداف) به دو مجموعه تقسیم می‌شوند که به طور سنتی مجموعه آموزشی و مجموعه آزمون(تست) نامیده می‌شوند. مجموعه تست را انتخاب می‌کنیم و کنار می­گذاریم و وانمود می­کنیم که قبلاً آن را ندیده‌ایم. سپس مدل خود را با استفاده از مجموعه آموزشی خود، با استفاده از ویژگی­ها و بردار هدف آموزش می­دهیم تا به مدل یاد دهیم که چگونه بهترین پیش‌بینی را انجام دهد. در نهایت، با ارزیابی اینکه مدل آموزش دیده ما در مجموعه آموزشی چگونه در مجموعه آزمایشی عمل می­کند شبیه سازی می­کنیم که تا به حال، داده­های خارجی را ندیده‌ایم. با این حال، رویکرد اعتبارسنجی دو نقطه ضعف عمده دارد. اولین مورد اینکه، عملکرد مدل می­تواند بسیار به این وابسته باشد که مشاهدات کمی برای مجموعه‌ی تست انتخاب شده است. مورد دوم اینکه، مدل با استفاده از تمام داده­های موجود آموزش داده نمی­شود، و بر روی تمام داده­های موجود ارزیابی نمی­شود.

یک استراتژی بهتر، که بر این نقاط ضعف غلبه می‌کند، اعتبارسنجی متقاطع k-fold cross.validation (KFCV) نامیده می‌شود. در KFCV، داده­ها را به k بخش به نام "fold"(دسته) تقسیم می‌کنیم. سپس مدل با استفاده از k-1 دسته که در یک مجموعه آموزشی ترکیب می‌شوند، آموزش داده می‌شود و سپس آخرین دسته به عنوان یک مجموعه تستی استفاده می‌شود و مدل با این دسته تست می‌شود. ما این­کار را k بار، هر بار با استفاده از یک fold متفاوت به عنوان مجموعه تست تکرار می­کنیم. سپس عملکرد مدل برای هر یک از k تکرار میانگین گیری می­شود تا یک اندازه‌گیری کلی ایجاد شود.

در راه حل خود، اعتبارسنجی متقاطع k-fold را با استفاده از 10 fold انجام دادیم و نتایج ارزیابی را به cv\_results ارائه کردیم:





هنگام استفاده از KFCV سه نکته‌ی مهم وجود دارد که باید در نظر بگیریم. اول، KFCV فرض می‌کند که هر مشاهده، مستقل از دیگری ایجاد شده است(یعنی داده‌ها به طور یکسان توزیع شده‌اند [IID]). اگر داده‌ها IID هستند، ایده خوبی است که هنگام تخصیص به foldها، مشاهدات را جا به جا کنید. در sikit learn می‌توانیم shuffle=True­­ را برای انجام جابجایی تنظیم کنیم.

دوم، هنگامی که از KFCV برای ارزیابی یک دسته‌بندی‌کننده استفاده می‌کنیم، اغلب داشتن fold­های حاوی درصد یکسان مشاهدات از هر یک از کلاس‌های هدف مختلف(به نام k-fold دسته‌بندی شده) مفید است. به عنوان مثال، اگر بردار هدف ما حاوی جنسیت باشد و 80 درصد مشاهدات مذکر باشد، در این صورت هر دسته شامل 80 درصد مشاهدات مرد و 20 درصد زنان خواهد بود. در scikit-learn، می‌توانیم اعتبارسنجی متقاطع k-fold دسته‌بندی شده را با جایگزین کردن کلاس KFold با StratifiedKFold انجام دهیم.

در نهایت، هنگامی که از مجموعه‌ی اعتبارسنجی متقاطع استفاده می‌کنیم، پیش‌پردازش داده‌ها براساس مجموعه‌ی آموزشی و سپس اعمال این تبدیلات برای مجموعه‌ی آموزشی و تست مهم است. برای مثال، هنگامی که با هدف استاندارد سازی، standardizer، برازش می‌کنیم، تنها میانگین و واریانس مجموعه‌ی آموزشی را محاسبه می‌کنیم. سپس آن تبدیل(با استفاده از transform) را برای مجموعه­های آموزشی و تست اعمال می­کنیم:



دلیل این امر این است که ما وانمود می­کنیم که مجموعه‌ی آزمایشی، داده‌ی ناشناخته است. اگر ما هر دو پیش پردازش خود را با استفاده از مشاهدات هر دو مجموعه‌ی آموزشی و تست هماهنگ کنیم، برخی از اطلاعات مجموعه‌ی تست به مجموعه‌ی آموزشی ما تراوش می‌کند. این قانون برای هر مرحله‌ی پیش‌پردازش مانند انتخاب ویژگی اعمال می‌شود.

روند کاری پکیج pipeline در scikit-learn انجام این کار را با استفاده از تکنیک­های اعتبارسنجی متقابل آسان می­کند. ابتدا یک خط لوله ایجاد می­کنیم که داده­ها را از قبل پردازش می­کند(به عنوان مثال، استانداردساز) و سپس یک مدل(رگرسیون لوجستیکی، لاجیت) را آموزش می­دهد:



سپس KFCV را با استفاده از آن روند کاری اجرا می‌کنیم و scikit همه کارها را برای ما انجام می­دهد:



cross\_val\_score دارای سه پارامتر شایان ذکر است که ما در مورد آنها صحبت نکردیم. cv روش اعتبارسنجی متقاطع ما را تعیین می‌کند. K-fold تا کنون رایج‌ترین روش است، اما موارد دیگری نیز وجود دارد، مانند leave-one-out-cross-validation که در آن تعداد fold­های k برابر با تعداد مشاهدات است. پارامتر scoring معیار ما را برای موفقیت تعریف می‌کند که تعدادی از آن‌ها در دستور العمل‌های دیگر در این فصل مورد بحث قرار می‌گیرند. در نهایت، n\_jobs=-1 به scikit-learn استفاده از هر هسته‌ی موجود را بیان می­کند. به عنوان مثال، اگر کامپیوتر شما چهار هسته دارد(یک عدد رایج برای لپ تاپ­ها)، پس scikit-learn از هر چهار هسته به صورت هم زمان برای سرعت بخشیدن به عملیات استفاده خواهد کرد.

**برای مطالعه بیشتر**

* [Why every statistician should know about cross.validation](https://robjhyndman.com/hyndsight/crossvalidation/)
* [Cross.Validation Gone Wrong](https://betatim.github.io/posts/cross-validation-gone-wrong/)

**١١.٢. ایجاد یک مدل رگرسیون پایه**

**مسئله**

شما یک مدل رگرسیون پایه ساده می­خواهید تا با مدل خود مقایسه کنید.

**راه حل**

برای ایجاد یک مدل ساده که بتوان از آن به عنوان یک خط پایه استفاده کرد، از دستور DummyRegressor در scikit-learn استفاده کنید:





برای مقایسه، مدل خود را آموزش می‌دهیم و امتیاز عملکرد را ارزیابی می‌کنیم:





**بحث**

DummyRegressor به ما اجازه می‌دهد تا یک مدل بسیار ساده ایجاد کنیم که بتوانیم از آن به عنوان یک پایه(اساس) برای مقایسه با مدل واقعی خود استفاده کنیم. این امر اغلب می‌تواند برای شبیه‌سازی یک فرآیند پیش‌بینی "ساده" در یک محصول یا سیستم مفید باشد. به عنوان مثال، ممکن است یک محصول در اصل طوری کدگذاری شده باشد تا تصور کنید که همه کاربران جدید در ماه اول، صرف نظر از ویژگی­های آن­ها، ۱۰۰ دلار هزینه خواهند کرد. اگر این فرض را در یک مدل پایه رمزگذاری کنیم، می‌توانیم مزایای استفاده از رویکرد یادگیری ماشین را به طور مشخص بیان کنیم.

DummyRegressor از پارامتر strategy برای تعیین روش پیش‌بینی، شامل مقدار میانگین یا میانه در مجموعه‌ی آموزشی استفاده می‌کند. علاوه بر این، اگر strategy را ثابت(constant) قرار دهیم و از پارامتر ثابت(constant) استفاده کنیم، می­توانیم رگرسیون ساخته‌شده را برای پیش‌بینی مقداری ثابت برای هر مشاهده تنظیم کنیم:





یک نکته کوچک در مورد score. به طور پیش‌فرض، score مقدار امتیاز ضریب توضیح(R.squared،R2) را برمی­گرداند:

که در آن ، ارزش واقعی مشاهده‌ی هدف است، ، مقدار پیش‌بینی‌شده است و ، مقدار میانگین برای بردار هدف است. هر چه به ۱ نزدیک‌تر باشد، واریانس در بردار هدف که توسط ویژگی­ها توضیح داده می­شود، بیشتر است.

**١١.٣. ایجاد یک مدل دسته‌بندی پایه**

**مسئله**

شما یک دسته‌بندی پایه‌ی ساده می‌خواهید تا با مدل خود مقایسه کنید.

**راه حل**

از DummyClassifier در scikit-learn استفاده کنید.





با مقایسه‍ی دسته‌بندی پایه با دسته‌بندی آموزش‌دیده خود، می­توانیم بهبود را ببینیم:





**بحث**

معیار معمول عملکرد یک دسته‌بندی‌کننده این است که چقدر بهتر از حدس تصادفی است. Dummy Classifier در scikit-learn این مقایسه را آسان می­کند. پارامتر strategy به ما چند گزینه برای تولید مقادیر می‌دهد. دو استراتژی بسیار مفید وجود دارد. ابتدا، stratifiedپیش‌بینی‌هایی را انجام می‌دهد که متناسب با نسبت‌های کلاس بردار هدف مجموعه آموزشی است(به عنوان مثال، اگر 20٪ از مشاهدات در داده‌های آموزشی زنان باشند، DummyClassifier زنان را در 20٪ موارد پیش‌بینی می‌کند). دوم، uniform، پیش‌بینی­ها را به طور یکنواخت و تصادفی بین کلاس­های مختلف تولید می‌کند. به عنوان مثال، اگر ۲۰ % مشاهدات، زن و ۸۰ % مرد باشند، uniform پیش‌بینی­هایی را ایجاد می­کند که ۵۰ % زن و ۵۰ % مرد هستند.

**برای مطالعه بیشتر**

* [scikit-learn documentation: DummyClassifier](https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html)

**۱۱.۴. ارزیابی پیش‌بینی­های دسته‌بندی‌کننده دودویی**

**مسئله**

با توجه به یک مدل دسته‌بندی آموزش دیده، می­خواهید کیفیت آن را ارزیابی کنید.

**راه حل**

از scikit-learn’s cross\_val\_score برای انجام اعتبارسنجی متقاطع استفاده کنید در حالی که از پارامتر scoring برای تعریف یک معیار از تعدادی معیار عملکرد، از جمله صحت، دقت، فراخوانی و امتیاز استفاده کنید.

دقت یک معیار عملکرد رایج است. به بیان ساده، نسبت مشاهداتی است که به درستی پیش‌بینی‌شده اند:

که در آن:

TP تعداد مثبت­های حقیقی است. مشاهداتی که بخشی از طبقه مثبت هستند(بیماری دارند، محصول را خریده اند و غیره) و اینکه ما به درستی پیش‌بینی کرده‌ایم.

TN تعداد منفی­های حقیقی است. مشاهداتی که بخشی از طبقه منفی هستند (بیماری ندارند، محصول را نخریده اند و غیره) و اینکه ما به درستی پیش‌بینی کرده‌ایم.

FP تعداد مثبت­های کاذب است. همچنین خطای نوع اول(*I*) نیز نامیده می­شود. این مشاهدات پیش‌بینی‌شده از کلاس مثبت هستند که در واقع اشتباه پیش‌بینی‌شده‌اند و از کلاس منفی هستند.

FN تعداد منفی­های کاذب است. همچنین خطای نوع دوم(II ) نیز نامیده می­شود. این مشاهدات پیش‌بینی‌شده از کلاس منفی هستند که در واقع اشتباه پیش‌بینی‌شده‌اند و از کلاس مثبت هستند.

ما می­توانیم صحت را در اعتبارسنجی متقاطع سه fold(تعداد پیش‌فرض foldها) با تنظیم scoring='accuracy' اندازه‌گیری کنیم:





جذابیت صحت در این است که یک توضیح شهودی و ساده‌ دارد: نسبت مشاهداتی که به درستی پیش‌بینی‌شده اند. با این حال، در دنیای واقعی، اغلب داده­های ما دارای کلاس­های نامتعادل هستند(به عنوان مثال، 99.9٪ مشاهدات از کلاس 1 و تنها 0.1٪ کلاس 2 هستند). در حضور کلاس­های نامتعادل، صحت از پارادوکسی رنج می‌برد که در آن یک مدل، بسیار دقیق است اما فاقد قدرت پیش‌بینی است. به عنوان مثال، تصور کنید ما در حال تلاش برای پیش‌بینی وجود یک سرطان بسیار نادر هستیم که در 0.1٪ از جمعیت رخ می­دهد. پس از آموزش مدل متوجه می‌شویم که صحت آن ۹۵ درصد است. این در حالی است که ۹۹.۹ درصد افراد به این سرطان مبتلا نیستند: اگر ما به سادگی مدلی بسازیم که "پیش‌بینی کرده باشد" که هیچ‌کس چنین شکلی از سرطان ندارد، مدل ساده‌ی ما ۴.۹ درصد دقیق‌تر خواهد بود، اما به وضوح قادر به پیش‌بینی هیچ چیز نیست. به همین دلیل، ما اغلب برای استفاده از معیارهای دیگر مانند دقت، فراخوانی و امتیاز تحریک می‌شویم.

دقت، نسبت هر مشاهده‌ای است که پیش‌بینی می­شود مثبت باشد که در واقع مثبت است. ما می­توانیم آن را به عنوان یک نویز اندازه‌گیری در پیش‌بینی­های خود در نظر بگیریم . یعنی وقتی ما پیش‌بینی می­کنیم چیزی مثبت است، چقدر احتمال دارد که درست بگوییم. مدل­هایی که دقت بالایی دارند به این دلیل بدبین هستند که تنها زمانی یک مشاهده را پیش‌بینی می­کنند که در مورد آن بسیار مطمئن باشند. به طور رسمی، دقت عبارت است از:





فراخوانی(بازیابی) نسبت هر مشاهده‌ی مثبت است که واقعا ً مثبت است و توانایی مدل برای شناسایی مشاهده کلاس مثبت را، اندازه‌گیری می‌کند. مدل­هایی با فراخوانی(بازیابی) بالا خوش‌بین هستند که بار پایینی برای پیش‌بینی اینکه یک مشاهده در کلاس مثبت است دارند:





اگر این اولین باری است که با دقت و یادآوری(recall) مواجه می­شوید، قابل درک است اگر کمی طول بکشد تا آن­ها را به طور کامل درک کنید. این یکی از معایب دقت است؛ دقت و یادآوری کم‌تر شهودی هستند. تقریبا همیشه نوعی تعادل بین دقت و یادآوری(recall) می­خواهیم و این نقش با امتیاز F1 پر می­شود. امتیاز F1 یک میانگین هارمونیک است(نوعی میانگین که برای نسبت­ها استفاده می­شود):

این معیاری از درستی است که در پیش‌بینی مثبت به دست می­آید. یعنی از مشاهداتی که به عنوان مثبت برچسب‌گذاری شده اند، چه تعداد واقعا مثبت هستند:





**بحث**

به عنوان یک معیار ارزیابی، صحت، ویژگی­های ارزشمندی دارد، به خصوص شهود ساده‌ی آن. با این حال، معیارهای بهتر، اغلب شامل تعادل استفاده از دقت(percision) و یادآوری(recall) است . یعنی توازنی بین خوش بینی و بدبینی مدل ما. F1 نشان دهنده تعادل بین بازیابی و دقت است، جایی که سهم نسبی هر دو برابر است.

همچنین برای استفاده از cross\_val\_score، اگر مقادیر y حقیقی و مقادیر y پیش‌بینی‌شده را داشته باشیم، می‌توانیم معیارهایی مانند دقت و یادآوری را مستقیما محاسبه کنیم:





**برای مطالعه بیشتر**

* [Accuracy paradox](https://en.wikipedia.org/wiki/Accuracy_paradox)

**١١.٥. ارزیابی آستانه­های دسته‌بندی‌کننده‌ی دودویی**

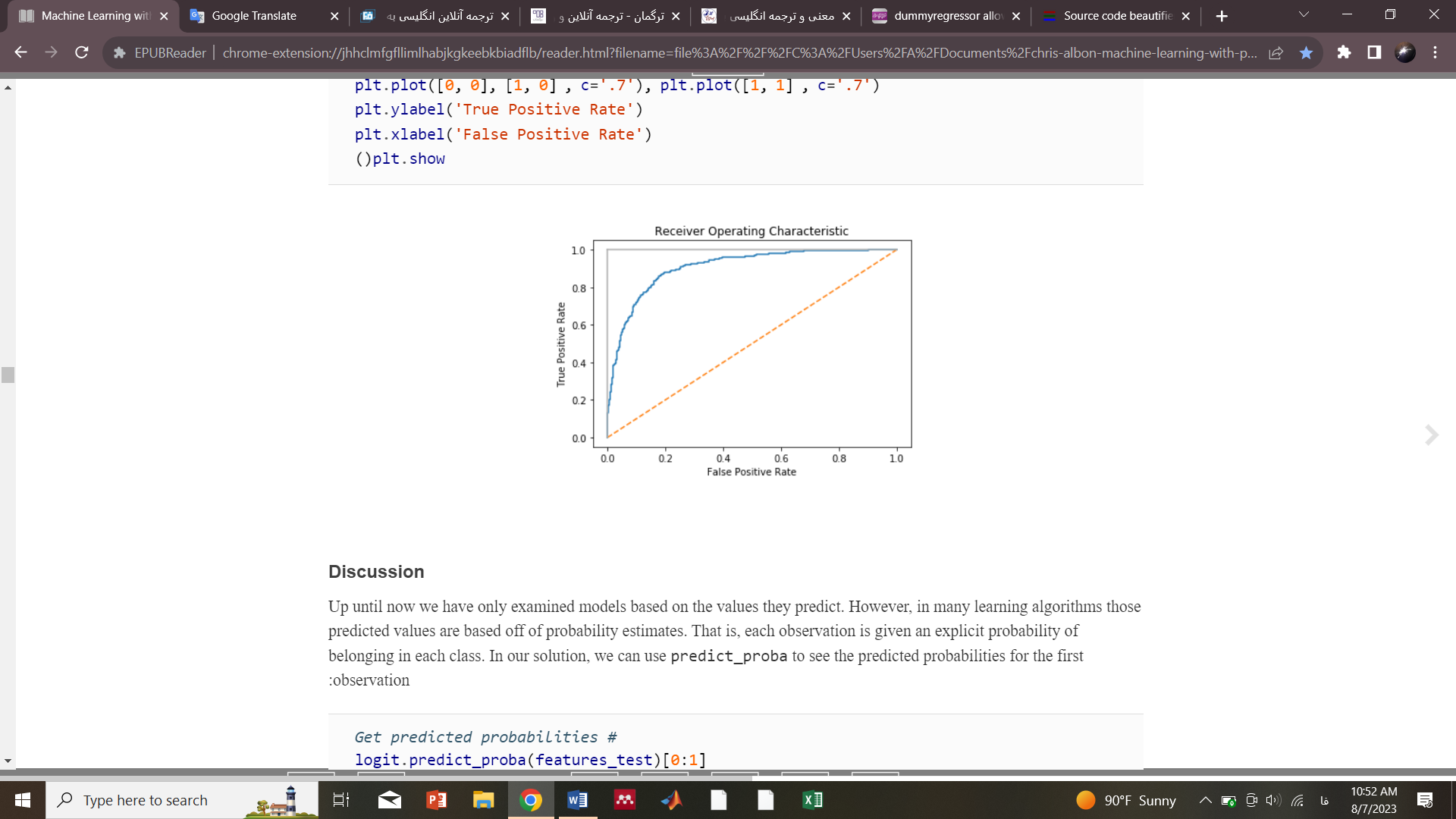
**مسئله**

می­خواهید یک دسته‌بندی‌کننده‌ی باینری و آستانه­های احتمالی مختلف را ارزیابی کنید.

**راه حل**

منحنی دریافت‌کننده مشخصه‌ی عملیاتی (ROC) یک روش رایج برای ارزیابی کیفیت یک دسته‌بندی‌کننده‌ی باینری(دودویی) است. ROC حضور مثبت­های حقیقی و مثبت­های کاذب را در هر آستانه احتمال مقایسه می‌کند (یعنی احتمال اینکه یک مشاهده در آن یک کلاس باشد). با رسم منحنی ROC، می‌توانیم عملکرد مدل را مشاهده کنیم. دسته‌بندی‌کننده‌ای که هر مشاهده را به درستی پیش‌بینی می­کند، مانند خط خاکستری روشن در نمودار زیر است که بلافاصله به سمت بالا می­رود. یک دسته‌بندی‌کننده که به صورت تصادفی پیش‌بینی می­کند، به صورت خط مورب ظاهر خواهد شد. هر چه مدل بهتر باشد، به خط ثابت نزدیک‌تر است. در scikit-learn، می‌توانیم از roc\_curve برای محاسبه مثبت‌های درست و نادرست در هر آستانه استفاده کنیم، سپس آنها را رسم کنیم:





**بحث**

تا کنون ما فقط مدل­ها را بر اساس مقادیری که پیش‌بینی می­کنند بررسی کرده‌ایم. با این حال، در بسیاری از الگوریتم­های یادگیری، آن مقادیر پیش‌بینی‌شده براساس برآورده‌ای از احتمالات هستند. یعنی به هر مشاهده یک احتمال صریح وابسته به هر کلاس داده می­شود. در این راه حل، ما می­توانیم ازpredict\_proba  استفاده کنیم تا احتمال­های پیش‌بینی‌شده برای اولین مشاهده را ببینیم:





ما می­توانیم کلاس­ها را با استفاده از classes\_ ببینیم:





در این مثال، اولین مشاهده دارای 87% احتمال قرار گرفتن در کلاس منفی (0) و 13% احتمال قرار گرفتن در کلاس مثبت (1) است. به طور پیش‌فرض، scikit-learn پیش‌بینی می­کند که یک مشاهده بخشی از کلاس مثبت است اگر احتمال آن بیشتر از ۰.۵ باشد (که آستانه نامیده می­شود).با این حال، به جای یک حد وسط، اغلب می‌خواهیم به طور صریح مدل خود را برای استفاده از یک آستانه‌ی متفاوت به دلایل متعدد به کار ببریم. به عنوان مثال، اگر مثبت کاذب برای کار ما بسیار پرهزینه باشد، ممکن است مدلی را ترجیح دهیم که آستانه احتمال بالایی دارد. ما نمی‌توانیم برخی از موارد مثبت را پیش‌بینی کنیم، اما وقتی یک مشاهده مثبت پیش‌بینی می‌شود، می‌توانیم بسیار مطمئن باشیم که پیش‌بینی درست است. این موازنه در نرخ مثبت حقیقی (TPR) و نرخ مثبت کاذب (FPR) نشان داده می­شود. نرخ مثبت حقیقی تعداد مشاهداتی است که به درستی پیش‌بینی‌شده اند و تقسیم بر تمام مشاهدات مثبت حقیقی می­شود:

نرخ مثبت کاذب، تعداد مثبت­های پیش‌بینی‌شده نادرست تقسیم بر تمام مشاهدات منفی واقعی است:

منحنی ROC نشان‌دهنده‌ی TPR و FPR مربوطه برای هر آستانه‌ی احتمال است. به عنوان مثال، در راه حل ما، یک آستانه تقریباً 0.50 دارای TPR 0.81 and an FPR of\0.15 است:





با این حال، اگر آستانه را به ۸۰ % افزایش دهیم (یعنی، میزان اطمینان مدل را قبل از اینکه یک مشاهده را مثبت پیش‌بینی کند افزایش دهیم)TPR به طور قابل توجهی کاهش می­یابد و به همین صورت FPR نیز کاهش می­یابد:





این به این دلیل است که نیاز بیشتر ما برای پیش‌بینی در کلاس مثبت باعث شده‌است که مدل تعدادی از مشاهدات مثبت (TPR پایین‌تر) را شناسایی نکند، بلکه نویز حاصل از مشاهدات منفی را به صورت مثبت (FPR پایین‌تر) کاهش دهد.

منحنی ROC علاوه بر اینکه می‌تواند مبادله‌ی بین TPR و FPR را تجسم کند، می‌تواند به عنوان یک معیار کلی برای یک مدل نیز استفاده شود. هر چه یک مدل بهتر باشد، منحنی بالاتر و در نتیجه مساحت زیر منحنی بیشتر است. به همین دلیل، محاسبه‌ی مساحت زیر منحنی ROC (AUCROC) برای مقایسه در مورد برابری کلی یک مدل در تمام آستانه‌های ممکن، معمول است. هر چه AUCROC به 1 نزدیکتر باشد، مدل بهتر است. در scikit-learn می‌توانیم AUCROC را با استفاده از roc\_auc\_score محاسبه کنیم:





**برای مطالعه بیشتر**

* [ROC Curves in Python and R](https://community.alteryx.com/t5/Data-Science/ROC-Curves-in-Python-and-R/ba-p/138430)
* [The Area Under an ROC Curve](https://darwin.unmc.edu/dxtests/roc3.htm)

**١١.٦. ارزیابی پیش‌بینی­های دسته‌بندی‌کننده چند طبقه**

**مسئله**

مدلی دارید که سه یا چند کلاس را پیش‌بینی می­کند و می­خواهید عملکرد آن را ارزیابی کنید.

**راه حل**

از اعتبارسنجی متقاطع با معیار ارزیابی‌ای که قادر به مدیریت بیش از دو کلاس است استفاده کنید:





**بحث**

وقتی کلاس‌های متعادلی داریم (مثلاً تعداد تقریباً مساوی از مشاهدات در هر کلاس از بردار هدف)، صحت(accuracy)، درست مانند تنظیمات کلاس باینری، یک انتخاب ساده و قابل تفسیر برای یک متریک ارزیابی است. صحت عبارت است از تعداد پیش‌بینی­های صحیح تقسیم بر تعداد مشاهدات که درست مانند تنظیمات باینری عمل می­کند. با این حال، زمانی که کلاس‌های نامتعادل داریم (یک سناریوی رایج)، باید به استفاده از معیارهای ارزیابی دیگر تمایل داشته باشیم.

بسیاری از معیارهای داخلی scikit-learn برای ارزیابی طبقه‌بندی‌کننده‌های باینری هستند. با این حال، بسیاری از این معیارها را می­توان برای استفاده در زمانی که بیش از دو کلاس داریم گسترش داد. دقت، یادآوری و امتیاز F1 معیارهای مفیدی هستند که قبلا به طور مفصل در دستورالعمل­های قبلی پوشش داده شده اند. در حالی که همه‌ی آنها در ابتدا برای طبقه‌بندی‌کننده‌های باینری طراحی شده بودند، می‌توانیم با در نظر گرفتن داده‌هایمان به‌عنوان مجموعه‌ای از کلاس‌های باینری، آن‌ها را در تنظیمات چند طبقه اعمال کنیم. انجام این کار، ما را قادر می­سازد تا معیارها را برای هر کلاس به گونه‌ای اعمال کنیم که گویی تنها کلاس موجود در داده­ها است، و سپس با میانگین‌گیری از آن­ها امتیازات ارزیابی را برای تمام کلاس­ها جمع آوری کنیم:





در این کد، \_macro به روش استفاده شده برای میانگین نمرات ارزیابی از کلاس­ها اشاره دارد:

macro میانگین امتیازات متریک را برای هر کلاس محاسبه می­کند و هر کلاس را به طور مساوی وزن می­کند.

weighted میانگین امتیازات متریک را برای هر کلاس محاسبه می­کند و هر کلاس را متناسب با اندازه‌ی آن در داده‌ها وزن می­کند.

micro میانگین امتیازات متریک را برای هر ترکیب کلاس مشاهده محاسبه می­کند.

**١١.٧. نمایش عملکرد یک دسته‌بندی‌کننده**

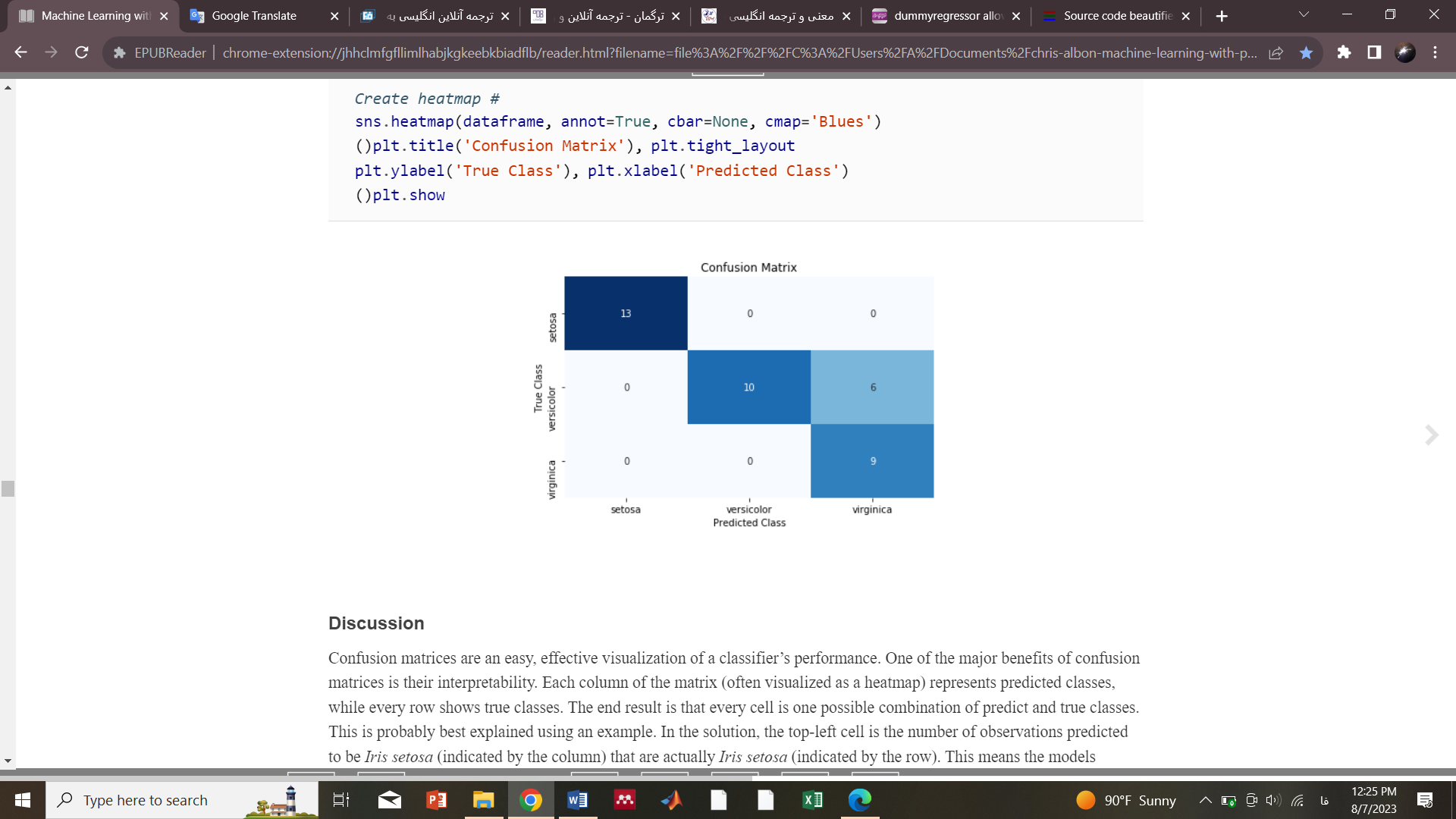
**مسئله**

با توجه به کلاس­های پیش‌بینی‌شده و کلاس­های واقعی داده­های آزمون، شما می­خواهید کیفیت مدل را به صورت بصری مقایسه کنید.

**راه حل**

از یک ماتریس درهم‌ریختگی(Confusion Matrix) استفاده کنید که کلاس­های پیش‌بینی‌شده و کلاس­های واقعی را مقایسه می­کند:





**بحث**

ماتریس­های درهم‌ریختگی نمایشی آسان و موثر از عملکرد یک دسته‌بندی‌کننده هستند. یکی از مزایای عمده‌ی این ماتریس­ها، تفسیرپذیری آنهاست. هر ستون از ماتریس (که اغلب به صورت یک نقشه حرارتی تجسم می­شود) کلاس­های پیش‌بینی‌شده را نشان می­دهد، در حالی که هر سطر کلاس­های واقعی را نشان می­دهد. نتیجه نهایی این است که هر سلول یک ترکیب ممکن از کلاس‌های پیش‌بینی و واقعی است. احتمالا ً بهتر است با استفاده از یک مثال توضیح داده شود. در این راه حل، سلول بالا سمت چپ، تعداد مشاهداتی است که پیش‌بینی می­شود زنبق هستند (با ستون نشان داده می­شود) که در واقعیت نیز زنبق (نشان داده شده توسط ردیف) هستند. این یعنی مدل­ها به درستی تمام گل­های زنبق را پیش‌بینی کردند. با این حال، این مدل در پیش‌بینی میخک به خوبی عمل نمی­کند. سلول پایین سمت راست نشان می‌دهد که مدل با موفقیت یازده مشاهدات Iris virginica را پیش‌بینی کرده، اما (با نگاه کردن به یک سلول بالا) یک گل را پیش‌بینی کرد که در واقع رنگارنگ زنبق بود.

سه نکته در مورد ماتریس­های درهم‌ریختگی قابل توجه است. اول، یک مدل کامل مقادیری در امتداد قطر و همچنین مقدار صفر در هر جای دیگر خواهد داشت. یک مدل بد، نشان می­دهد که تعداد مشاهدات به طور یکنواخت در اطراف سلول­ها پخش شده. دوم، یک ماتریس درهم‌ریختگی به ما اجازه می­دهد تا نه تنها ببینیم که مدل کجا اشتباه بوده است، بلکه ببینیم که چگونه اشتباه بوده است. یعنی می‌توانیم به الگوهای دسته‌بندی نادرست نگاه کنیم. به عنوان مثال، مدل ما به آسانی زنبق و میخک را تشخیص داد اما میخک و زنبق رنگین‌کمانی را به دشواری تشخیص می­داد. سوم، ماتریس­های درهم‌ریختگی با هر تعداد کلاس کار می­کنند (اگر چه اگر یک میلیون کلاس در بردار هدف خود داشته باشیم، تجسم ماتریس در هم ریخنگی ممکن است دشوار باشد).

**برای مطالعه بیشتر**

* [Confusion matrix, Wikipedia](https://en.wikipedia.org/wiki/Confusion_matrix" \t "_blank)
* [scikit-learn documentation: Confusion Matrix](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html)

**١١.٨.ارزیابی مدل­های رگرسیون**

**مسئله**

شما می­خواهید عملکرد یک مدل رگرسیون را ارزیابی کنید.

**راه حل**

استفاده از میانگین مربعات خطا (MSE):





یکی دیگر از معیارهای رایج رگرسیون، ضریب تعیین، R2 است:





**بحث**

MSE یکی از رایج­ترین معیارهای ارزیابی برای مدل­های رگرسیون است. MSE عبارت است از:

که در آن n تعداد مشاهدات، yi مقدار حقیقی هدفی است که می­خواهیم برای مشاهده i پیش‌بینی کنیم و مقدار پیش‌بینی‌شده توسط مدل برای yi است. MSE اندازه‌گیری مجموع مربع تمام فواصل بین مقادیر پیش‌بینی‌شده و واقعی است. هر چه مقدار MSE بیشتر باشد، خطای مجذور کل بیشتر و در نتیجه مدل بدتر می­شود. تعدادی از مزایای ریاضی برای مجموع مربعات خطا وجود دارند؛ از جمله اینکه تمام مقادیر خطا را مجبور می­کند که مثبت باشند؛ اما یک مفهوم اغلب محقق نشده این است که مربع کردن، چند خطای بزرگ را بیشتر از بسیاری از خطاهای کوچک جریمه می­کند(به چشم می‌آورد). حتی اگر قدر مطلق خطاها یکسان باشد. برای مثال، دو مدل A و B را تصور کنید که هر کدام دارای دو مشاهده هستند:

مدل A دارای خطاهای 0 و 10 است و بنابراین MSE آن 100= 02 + 102 است.

مدل B هر کدام دو خطای 5 دارد و بنابراین MSE آن 50= 52 + 52  است.

هر دو مدل خطای کل یکسان برابر با 10 دارند؛ با این حال، MSE مدل (MSE = 100) A را بدتر از مدل B (MSE=50) در نظر می­گیرد. در عمل این مفهوم به ندرت یک مساله است (و در واقع می‌تواند از لحاظ نظری سودمند باشد) و MSE به عنوان یک معیار ارزیابی کاملا خوب عمل می­کند.

یک نکته‌ی مهم: به طور پیش‌فرض در آرگومان‌های scikit-learn پارامتر امتیازدهی فرض می­شود که مقادیر بالاتر، بهتر از مقادیر پایین­تر هستند. با این حال، این مورد برای MSE صادق نیست، که در آن مقادیر بالاتر به معنای مدل بدتر است. به همین دلیل، scikit-learn با استفاده از آرگومان neg\_mean\_squared\_errorبه MSE منفی نگاه می­کند.

یک معیار ارزیابی رگرسیون جایگزین رایج R2 است، که مقدار واریانس در بردار هدف را اندازه‌گیری می‌کند که توسط مدل توضیح داده می­شود:

در این فرمول، yi مقدار هدف واقعی مشاهده i ام است، مقدار پیش‌بینی‌شده برای مشاهده iام است و مقدار میانگین بردار هدف است. هرچه به 1.0 نزدیکتر باشد، مدل بهتر است.

برای مطالعه بیشتر

* [Mean squared error, Wikipedia](https://en.wikipedia.org/wiki/Mean_squared_error)
* [Coefficient of determination](https://en.wikipedia.org/wiki/Coefficient_of_determination)

**۱۱.۹. ارزیابی مدل­های خوشه‌بندی**

**مسئله**

شما از یک الگوریتم یادگیری بدون نظارت برای خوشه‌بندی داده­های خود استفاده کرده‌اید. حالا می­خواهید بدانید که مدل شما چقدر خوب عمل کرده است.

**راه حل**

پاسخ کوتاه این است که احتمالا نمی­توانید، حداقل نه آنطور که مد نظر دارید.

به این ترتیب، یک گزینه‌ی ارزیابی خوشه‌بندی با استفاده از ضرایب سیلوئت است که کیفیت خوشه‌ها را اندازه‌گیری می‌کند:





**بحث**

ارزیابی مدل نظارت شده، پیش‌بینی­ها (به عنوان مثال، کلاس­ها یا مقادیر کمی) را با مقادیر واقعی متناظر در بردار هدف مقایسه می­کند. با این حال، رایج‌ترین انگیزه برای استفاده از روش‌های خوشه‌بندی این است که داده‌های شما بردار هدف ندارند. تعدادی از معیارهای ارزیابی خوشه‌بندی وجود دارند که نیاز به بردار هدف دارند، اما باز هم، استفاده از روش­های یادگیری بدون نظارت مانند خوشه‌بندی زمانی که شما یک بردار هدف در دسترس دارید، احتمالا به طور غیرضروری خود را ناتوان می­کند.

در حالی که اگر بردار هدف نداشته باشیم نمی­توانیم پیش‌بینی­ها را در مقابل مقادیر واقعی ارزیابی کنیم و می­توانیم ماهیت خود خوشه­ها را ارزیابی کنیم. به طور شهودی، ما می­توانیم خوشه­های "خوب" را تصور کنیم که فاصله­های بسیار کمی بین مشاهدات در یک خوشه (یعنی خوشه­های متراکم) و فاصله­های زیاد بین خوشه­های مختلف (یعنی خوشه­های به خوبی جدا شده) دارند. ضرایب سیلوئت(Silhouette)، یک مقدار واحد برای اندازه‌گیری هر دو صفت فراهم می‌کنند. ضریب سیلوئت مشاهده i ام برابر است با:

که در آن si ضریب سیلوئت برای مشاهده‌ی iام، میانگین فاصله بین i و تمام مشاهدات یک کلاس و bi میانگین فاصله بین i و تمام مشاهدات از نزدیک‌ترین خوشه یک کلاس متفاوت است. مقدار برگردانده شده توسط silhouette\_score میانگین ضریب سیلوئت برای همه مشاهدات است. ضرایب سیلوئت بین 1- و 1 است که 1 نشان دهنده خوشه­های متراکم و به خوبی جدا شده است.

**برای مطالعه بیشتر**

* [scikit-learn documentation: silhouette\_score](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html)

**١١.١٠.ایجاد معیار ارزیابی ویژه**

**مسئله**

شما می­خواهید یک مدل را با استفاده از معیاری که ایجاد کرده‌اید ارزیابی کنید.

**راه حل**

اندازه را به عنوان یک تابع ایجاد کنید و آن را با استفاده از make\_scorerدر scikit-learn به یک تابع امتیاز دهنده تبدیل کنید:





**بحث**

در حالی که scikit-learn تعدادی معیار داخلی برای ارزیابی عملکرد مدل را دارد، اغلب تعریف معیارهای خود نیز مفید است. scikit-learn این کار را با استفاده از make\_scorer آسان می­کند. ابتدا، تابعی را تعریف می‌کنیم که دو آرگومان می‌گیرد. بردار هدف اصلی و مقادیر پیش‌بینی‌شده‌ی ما، و مقداری امتیاز را به دست می‌آورد (در خروجی به ما میدهد). دوم، ما از make\_scorer برای ایجاد یک امتیازدهنده استفاده می‌کنیم، و مطمئن می‌شویم که نمره‌های بالاتر یا پایین‌تر مطلوب هستند (با استفاده از پارامتر greater\_is\_better). معیار سفارشی (ویژه) در راه حل (custom\_metric) یک مثال اسباب بازی است زیرا به سادگی یک معیار داخلی برای محاسبه امتیاز 2R را پوشش می­دهد. در یک موقعیت واقعی، ما تابع custom\_metric را با هر متریک سفارشی‌ای که می‌خواهیم جایگزین می‌کنیم. با این حال، مشاهده می­شود که متریک سفارشی‌ای که 2R را محاسبه می­کند، با مقایسه نتایج با روش ساخته شده‌ی scikit-learn’s r2\_score کار می­کند:



**برای مطالعه بیشتر**

* [scikit-learn documentation: make\_scorer](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.make_scorer.html)

**۱۱.۱۱. نمایش تاثیر اندازه مجموعه آموزشی**

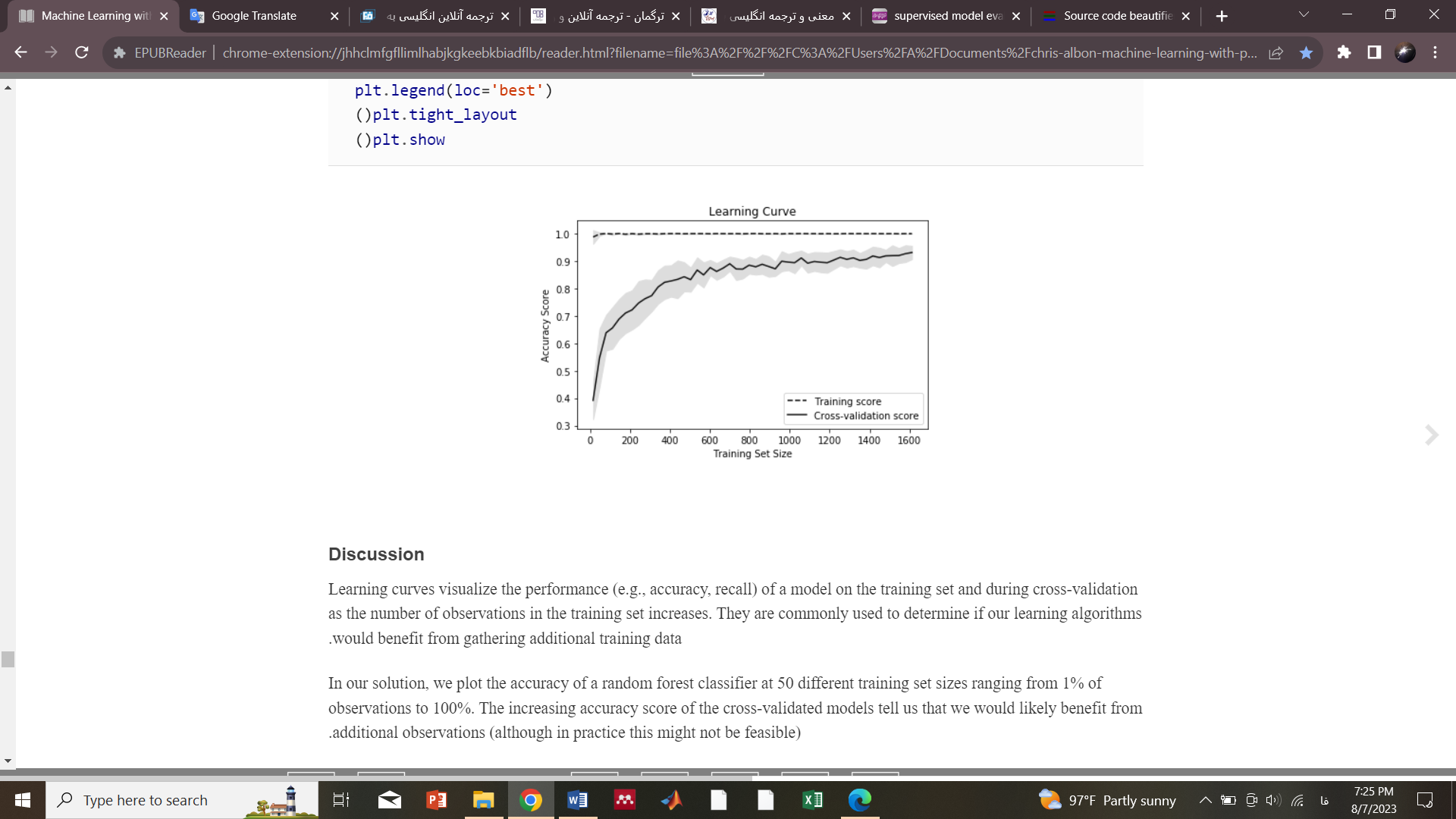
**مسئله**

شما می­خواهید تأثیر تعداد مشاهدات در مجموعه آموزشی خود را بر روی برخی متریک­ها (صحت، F1 و غیره) ارزیابی کنید.

**راه حل**

منحنی یادگیری را رسم کنید:





**بحث**

منحنی­های یادگیری، عملکرد (به عنوان مثال، صحت، یادآوری) یک مدل را در مجموعه آموزشی و در طول اعتبارسنجی متقابل با افزایش تعداد مشاهدات در مجموعه آموزشی، تصویرسازی می­کنند. آن‌ها معمولا ً برای تعیین اینکه آیا الگوریتم‌های یادگیری ما از جمع‌آوری داده‌های آموزشی اضافی بهره خواهند برد، استفاده می‌شوند.

در راه حل ما، ما دقت یک طبقه‌بندی‌کننده جنگل تصادفی(random forest) را در 50 اندازه‌ی مجموعه‌ی آموزشی مختلف از 1٪ مشاهدات تا 100٪ ترسیم می‌کنیم. افزایش امتیاز صحت مدل­های تایید شده‌ی متقاطع به ما می­گوید که ما احتمالا از مشاهدات اضافی بهره خواهیم برد (اگرچه این مورد در عمل ممکن است عملی نباشد).

برای مطالعه بیشتر

* [scikit-learn documentation: Learning Curve](https://scikit-learn.org/stable/modules/learning_curve.html" \l "learning-curve)

**١١.١٢.ایجاد یک گزارش متنی از معیارهای ارزیابی**

**مسئله**

شما یک توصیف سریع از عملکرد یک دسته‌بندی می­خواهید.

**راه حل**

از classification\_report در scikit-learn استفاده کنید:





**بحث**

classification\_report راهکاری سریع برای ما فراهم می‌کند تا برخی از معیارهای ارزیابی رایج از جمله دقت، یادآوری و امتیاز F1 را ببینیم (که قبلاً در این فصل توضیح داده شد). پشتیبانی(support) به تعداد مشاهدات در هر کلاس اشاره دارد.

**برای مطالعه بیشتر**

* [Precision and recall](https://en.wikipedia.org/wiki/Precision_and_recall)

**١١.١٣. تجسم اثر مقادیر فراپارامتر**

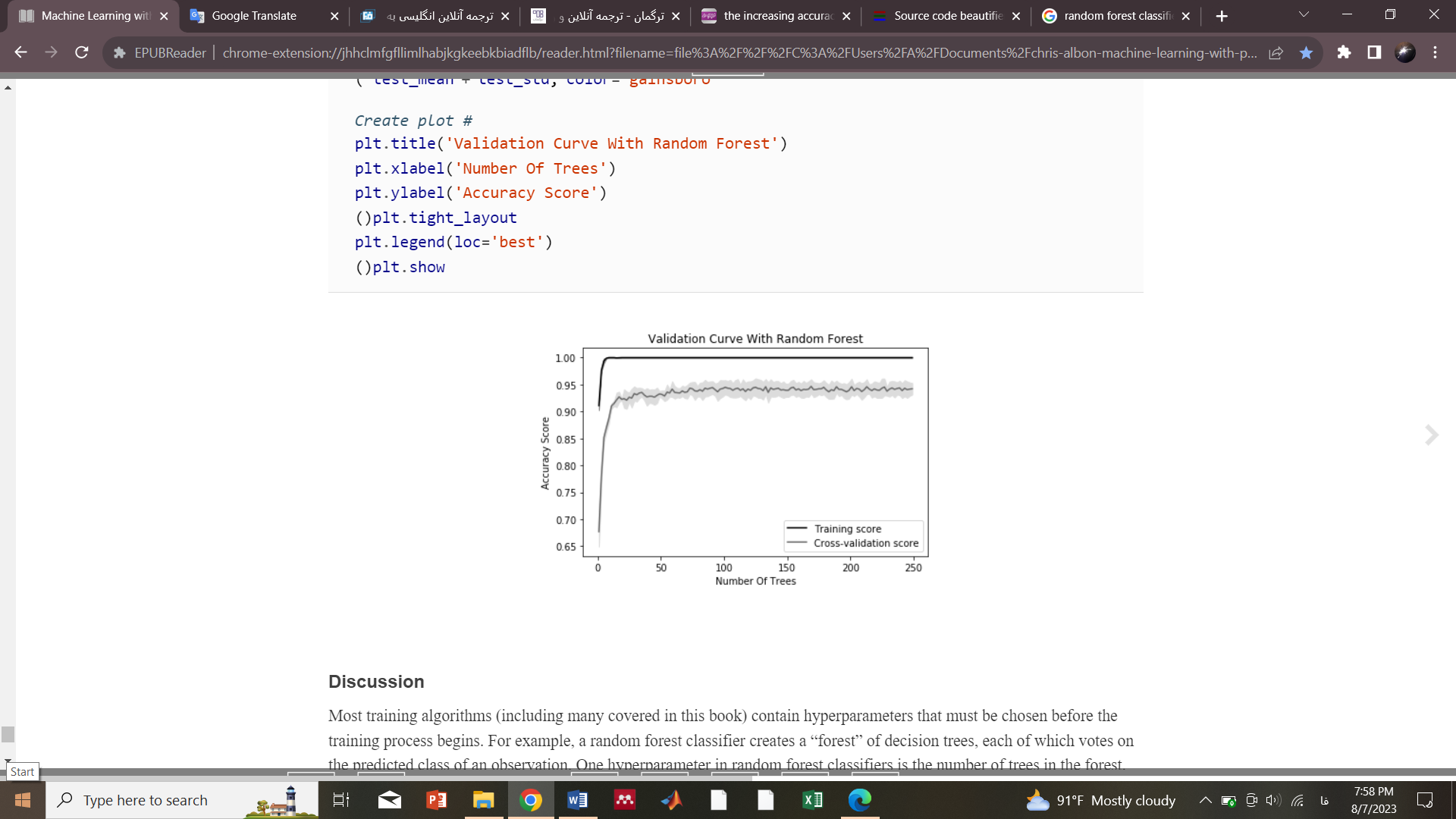
**مسئله**

شما قصد دارید متوجه چگونگی تغییر عملکرد یک مدل با تغییر مقدار برخی ­هایپرپارامترها شوید.

**راه حل**

منحنی اعتبار‌سنجی را رسم کنید:





**بحث**

بیشتر الگوریتم‌های آموزشی (از جمله بسیاری از الگوریتم‌های مورد بحث در این کتاب) حاوی هایپرپارامترهایی هستند که باید قبل از شروع فرآیند آموزش انتخاب شوند. برای مثال، یک دسته‌بندی‌کننده‌ی جنگل تصادفی(random forest) یک "جنگل" از درختان تصمیم ایجاد می­کند که هر کدام از آن­ها به کلاس پیش‌بینی‌شده‌ی یک مشاهده رای می­دهند. یک هایپرپارامتر در دسته‌بندی‌کننده­های جنگل تصادفی، تعداد درختان در جنگل است. اغلب مقادیر هایپرپارامتر در طول انتخاب مدل انتخاب می­شوند (به فصل 12 مراجعه کنید). با این حال، گاهی اوقات تصویرسازی نحوه‌ی تغییر عملکرد مدل با تغییر مقدار هایپرپارامتر مفید است. در راه حل ما، ما تغییرات در صحت را برای یک دسته‌بندی‌کننده‌ی جنگل تصادفی برای مجموعه‌ی آموزشی و در طول اعتبارسنجی متقابل با افزایش تعداد درختان رسم می­کنیم. هنگامی که تعداد کمی درخت داریم، هم امتیاز آموزش و هم امتیاز اعتبارسنجی متقاطع پایین است، که نشان می‌دهد مدل ضعیف است. با افزایش تعداد درختان به ۲۵۰ عدد، دقت هر دو سطح کاهش می­یابد که نشان می­دهد احتمالا ارزش چندانی در هزینه محاسباتی آموزش یک جنگل عظیم وجود ندارد.

در scikit-learn می‌توانیم منحنی اعتبارسنجی را با استفاده از validation\_curve محاسبه کنیم که شامل سه پارامتر مهم است:

aram\_name نام هایپرپارامتر برای تغییر است.

param\_range مقدار هایپرپارامتر مورد استفاده است.

scoring معیار ارزیابی مورد استفاده برای قضاوت در مورد مدل است.

**برای مطالعه بیشتر**

* [scikit-learn documentation: Validation Curve](https://scikit-learn.org/stable/modules/learning_curve.html" \l "validation-curve)