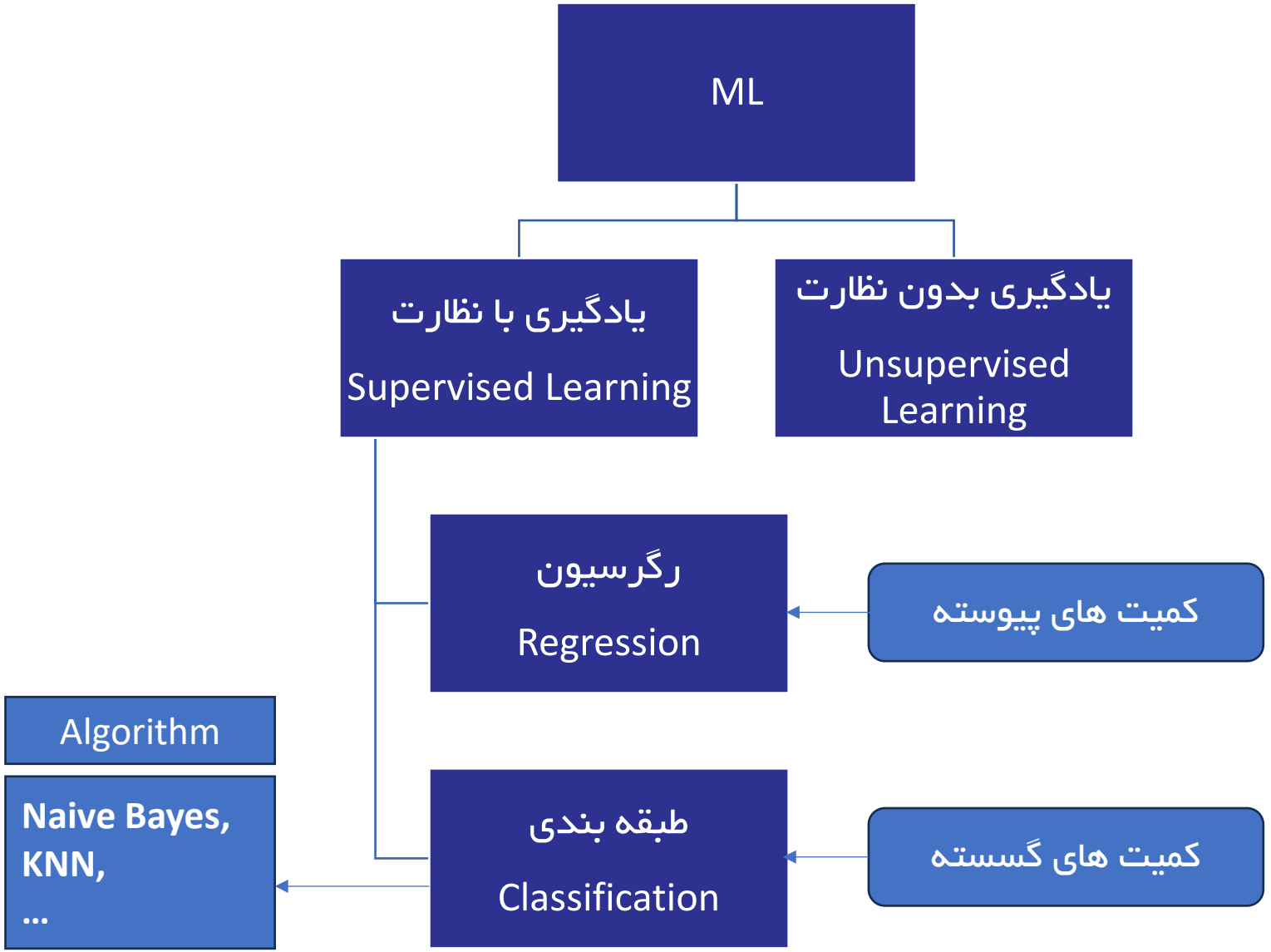
یادگیری ماشین
Machine Learning
مهر ۱۴۰۴

KNN Algorithm

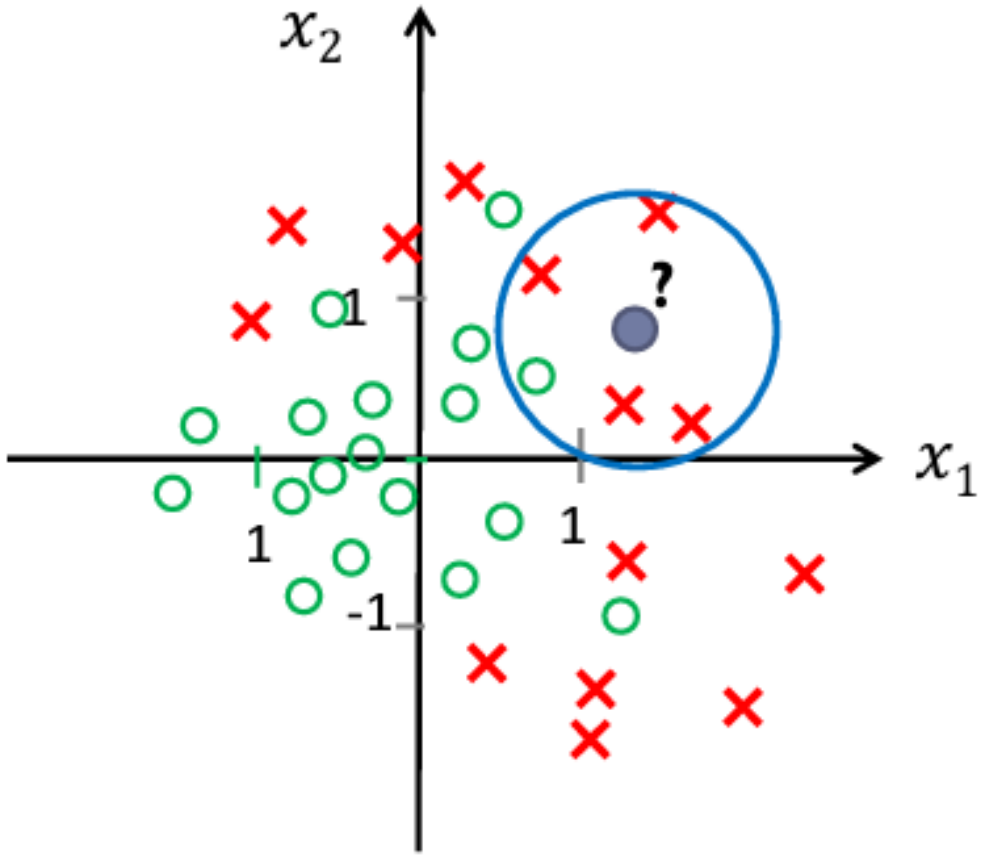
k-nearest neighbors

استاد : محسن عباسی

موسسه آموزش عالی خراسان


ML
یادگیری با نظارت
Supervised Learning
یادگیری بدون نظارت
Unsupervised Learning
رگرسیون
Regression
کمیت های پیوسته
طبقه بندی
Classification
کمیت های گسسته
Algorithm
Naive Bayes,
KNN,
...

KNN



ایده ی اصلی

برای پیش‌بینی برچسب یک داده ی جدید، به kتا از «نزدیک‌ترین» داده‌های آموزش نگاه می‌کنیم و بر اساس رأی اکثریت (در طبقه‌بندی) خروجی را تعیین می‌کنیم.

KNN Algorithm

1

محاسبه ی فاصله

بین نقطه ی جدید و همه ی نمونه‌های آموزش (معمولاً فاصله اقلیدسی)

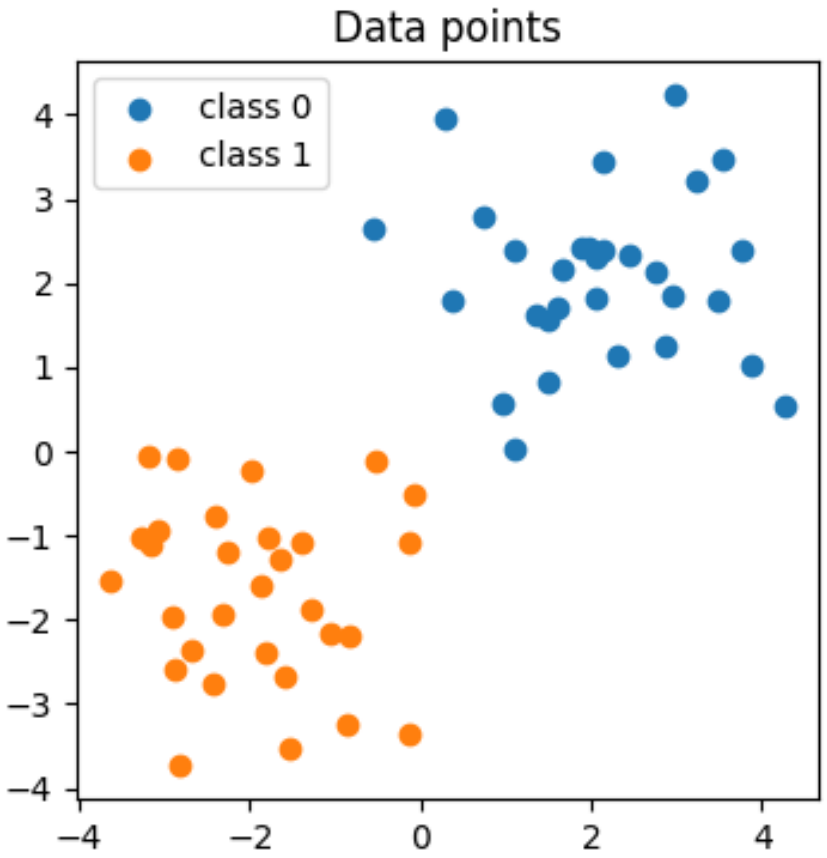
2

نمونه ی k انتخاب نزدیک‌تر

3

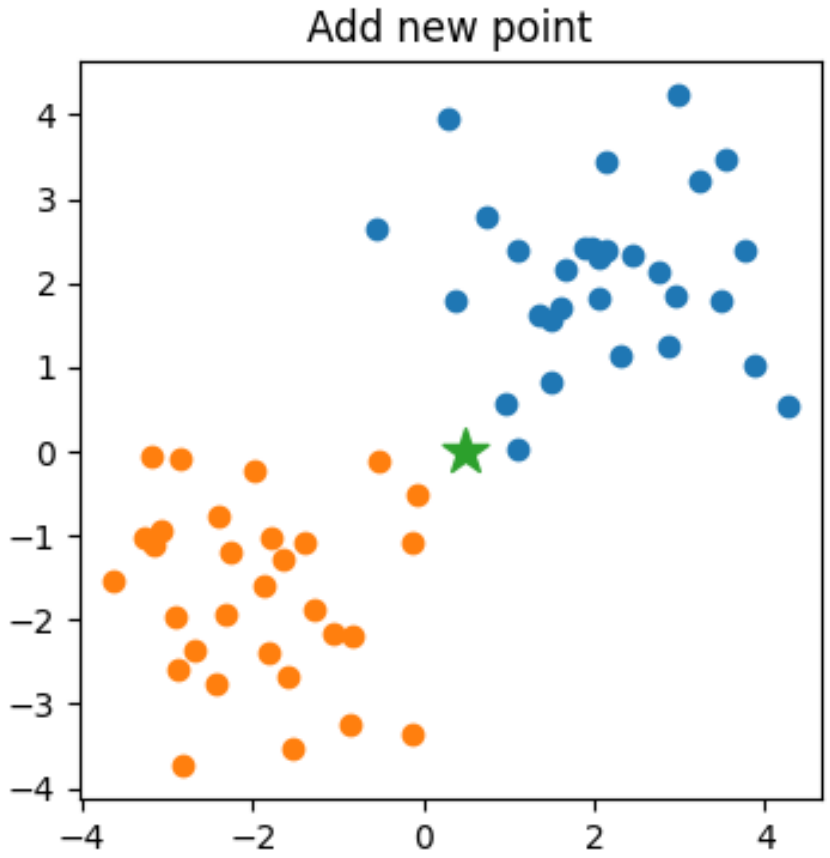
خروجی‌گیری

رأی‌گیری بین برچسب‌ها

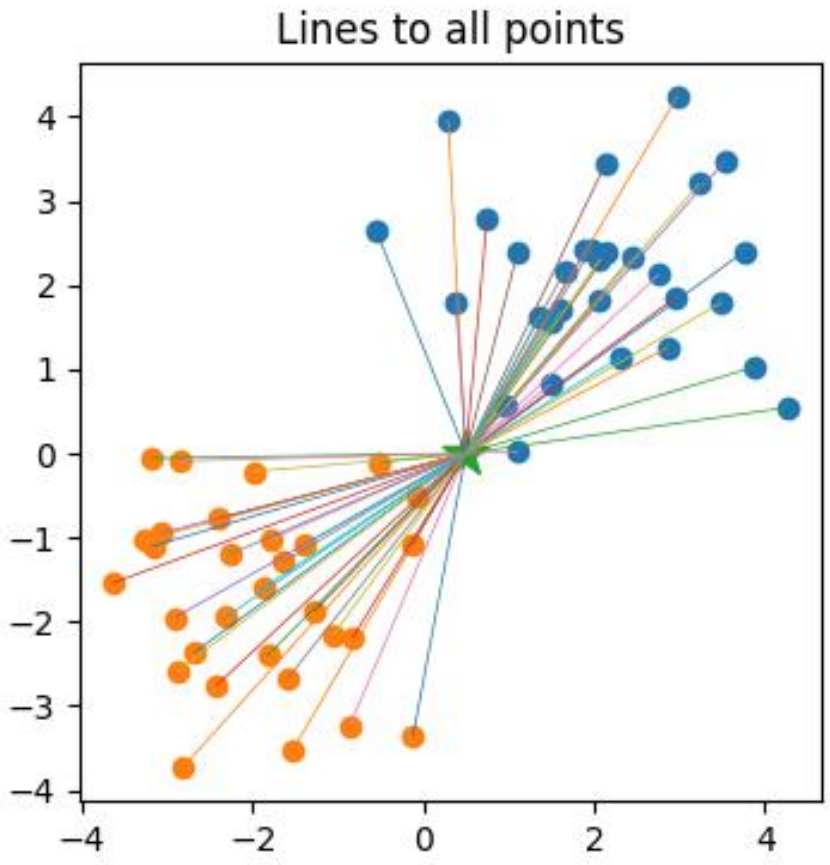


۱ – دیتای آموزش

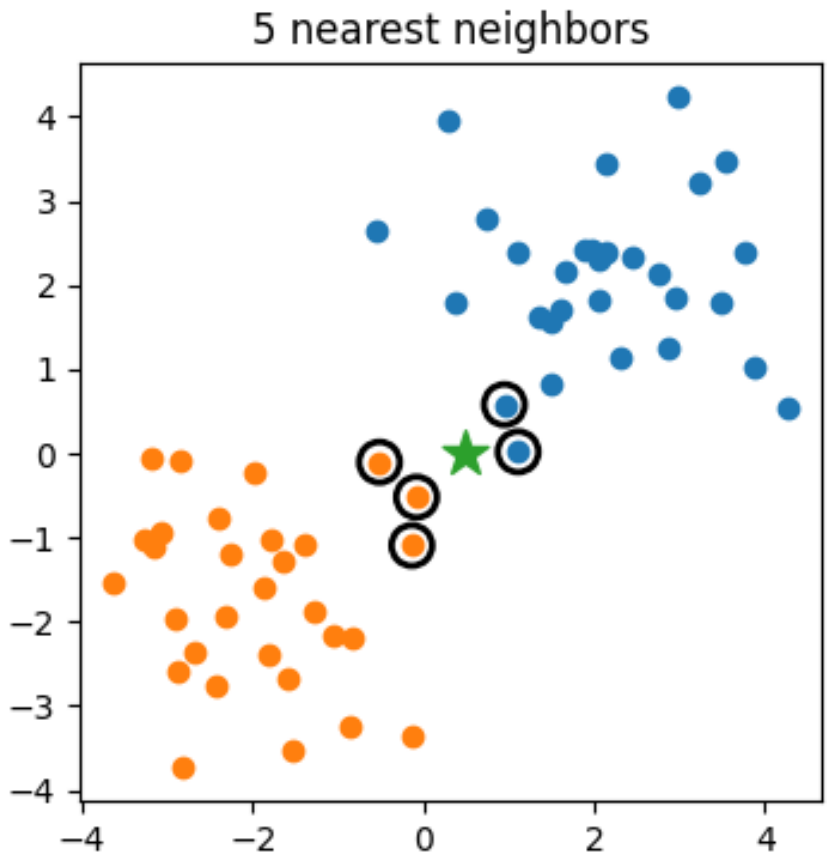
بر اساس برچسب



۲–ورود دیتای تست



۳–محاسبه فاصله تمام نقاط آموزش تا نقطه تست



۴–یافتن k=5 همسایه دیتای تست

معایب

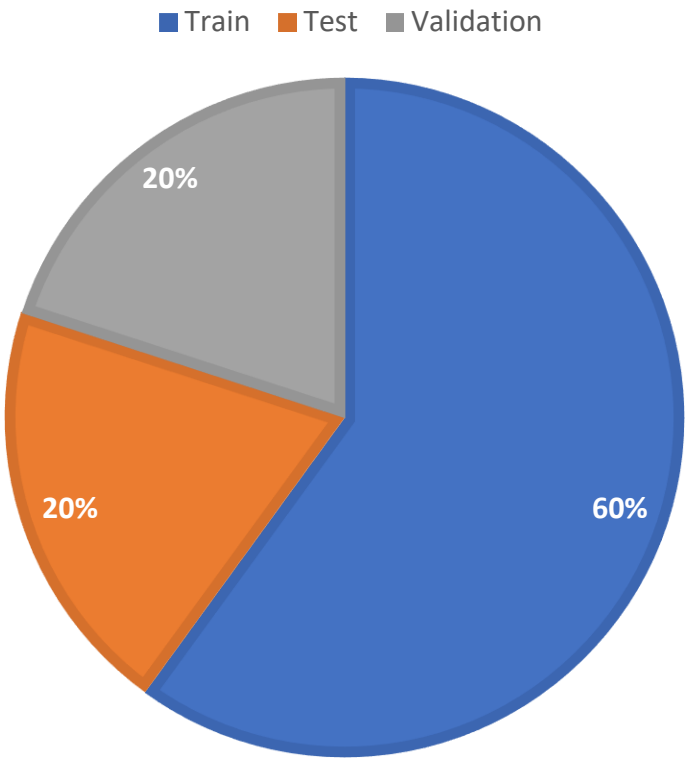
الگوریتم تنبل (در مرحله آموزش اتفاقی نمی افتد و همه محاسبات در حین پیش بینی محاسبه میشود.)

کند روی دیتاست‌های بزرگ (نیاز به محاسبه فاصله با همه)

حساس به مقیاس ویژگی‌ها (نیاز به نرمال‌سازی)

کارایی ضعیف در ابعاد زیاد curse of dimensionality

Validation



تقسیم داده به Train / Validation / Test
برای ارزیابی منصفانه ی مدل و جلوگیری از overfitting، داده به سه بخش جدا تقسیم می‌شود:

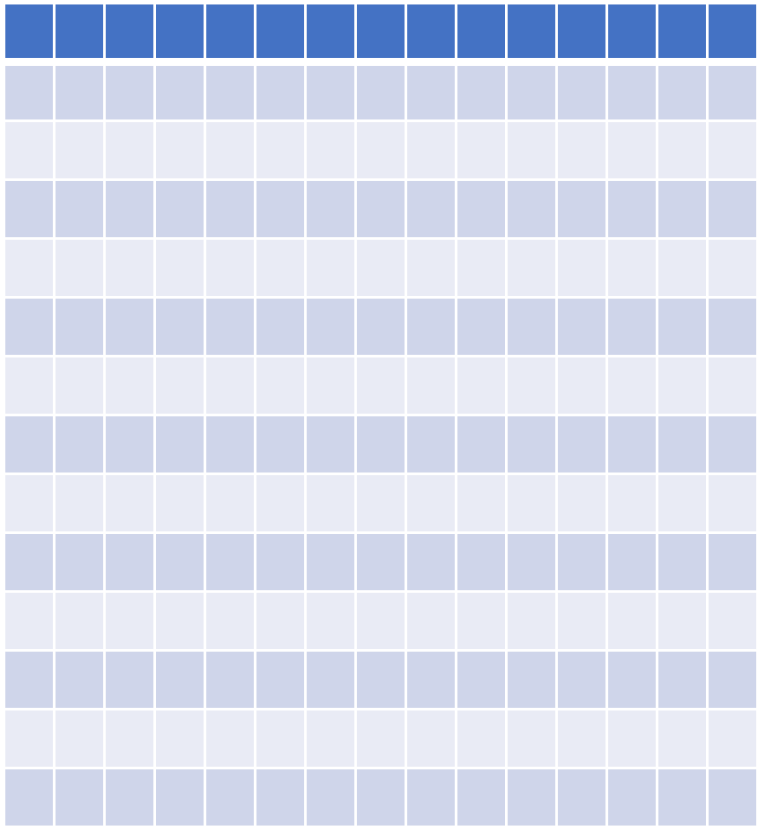
1. Train | آموزش
   مدل با استفاده از این بخش یاد می‌گیرد و پارامترهای داخلی‌اش تنظیم می‌شود.

2. Validation | اعتبارسنجی
   برای انتخاب و تنظیم هایپرپارامترها (مثلاً مقدار k در KNN یا نرخ یادگیری)، بدون اینکه مدل داده ی تست را ببیند.

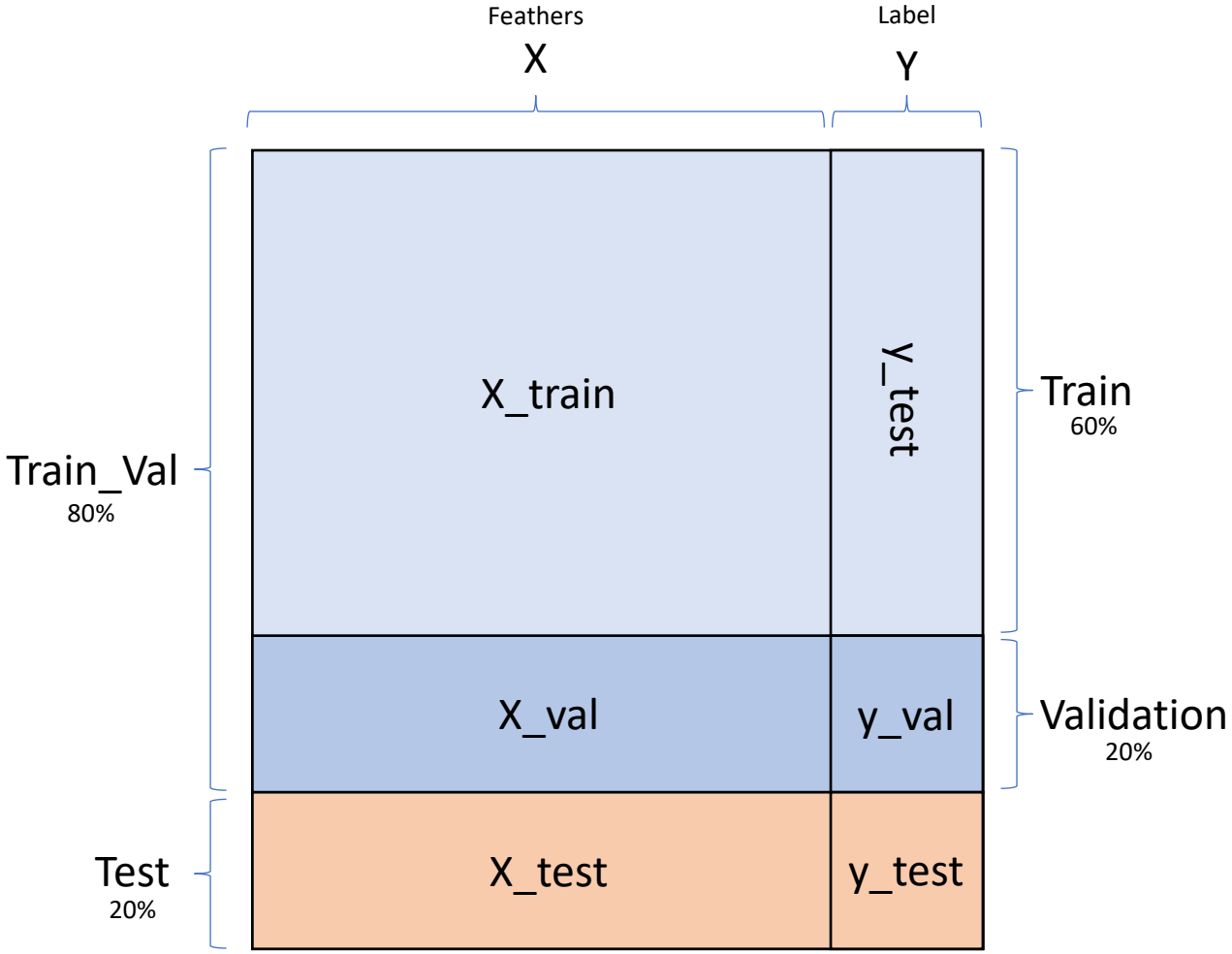
3. Test | آزمون نهایی
   فقط یک بار در انتهای کار برای سنجش نهایی عملکرد مدل روی داده ی دیده‌نشده.
   استفاده از تست در طول توسعه ممنوع است تا گزارش نهایی واقعی باشد.

Split

df : Data Frame



Split



انتخاب k

kکوچک → مدل به نویز حساس می‌شود، احتمال overfittingبالا

kبزرگ → تصمیم‌گیری نرم‌تر اما احتمال underfittingو کاهش دقت روی جزئیات

معمولاً انتخاب kبا روش‌های Cross-Validationانجام می‌شود

مقدار kاغلب عددی فرد (برای جلوگیری از تساوی در رأی‌گیری) انتخاب می‌شود

انتخاب kوابسته به تراکم و اندازه ی داده است؛ با افزایش داده ی آموزشی، معمولاً kهم کمی بزرگ‌تر انتخاب می‌شود

F1 Score

F1-Scoreیک معیار ترکیبی برای ارزیابی مدل‌های طبقه‌بندی است.

تعادل بین Precision و Recall را نشان می‌دهد:

$$F = \frac{1}{\frac{1}{2P} + \frac{1}{2R}} = \frac{2PR}{P+R}$$

Precision:درصد پیش‌بینی‌های درست مثبت نسبت به کل پیش‌بینی‌های مثبت

Recall:درصد پیش‌بینی‌های درست مثبت نسبت به کل نمونه‌های مثبت واقعی

F1-Score نزدیک به ۱ ← مدل دقیق و حساس است

F1-Score نزدیک به ۰ ← مدل عملکرد ضعیف دارد

پروژه

۱ – انتخاب دیتاست

هر نفر یک دیتاست مولتی‌کلاس انتخاب کند. (با حداقل ۳ کلاس)ویژگی‌ها و برچسب‌ها مشخص باشند.

۲– تقسیم‌بندی داده

داده به سه بخش تقسیم شود:
validation – train – test

۳– اجرای KNN

انتخاب k با استفاده از داده validation

۴– پیاده‌سازی مدل KNN

۵– ارزیابی مدل

محاسبه معیارهای Accuracy وF1-Score

۶– رسم Confusion Matrix

۷– گزارش نهایی

انتخاب k به چه صورت انجام شد
معیارهای عملکرد مدل
تحلیل ماتریس در هم‌ریختگی: چه کلاس‌هایی با هم اشتباه گرفته شدند؟

تهیه کننده:
وحید سیدی