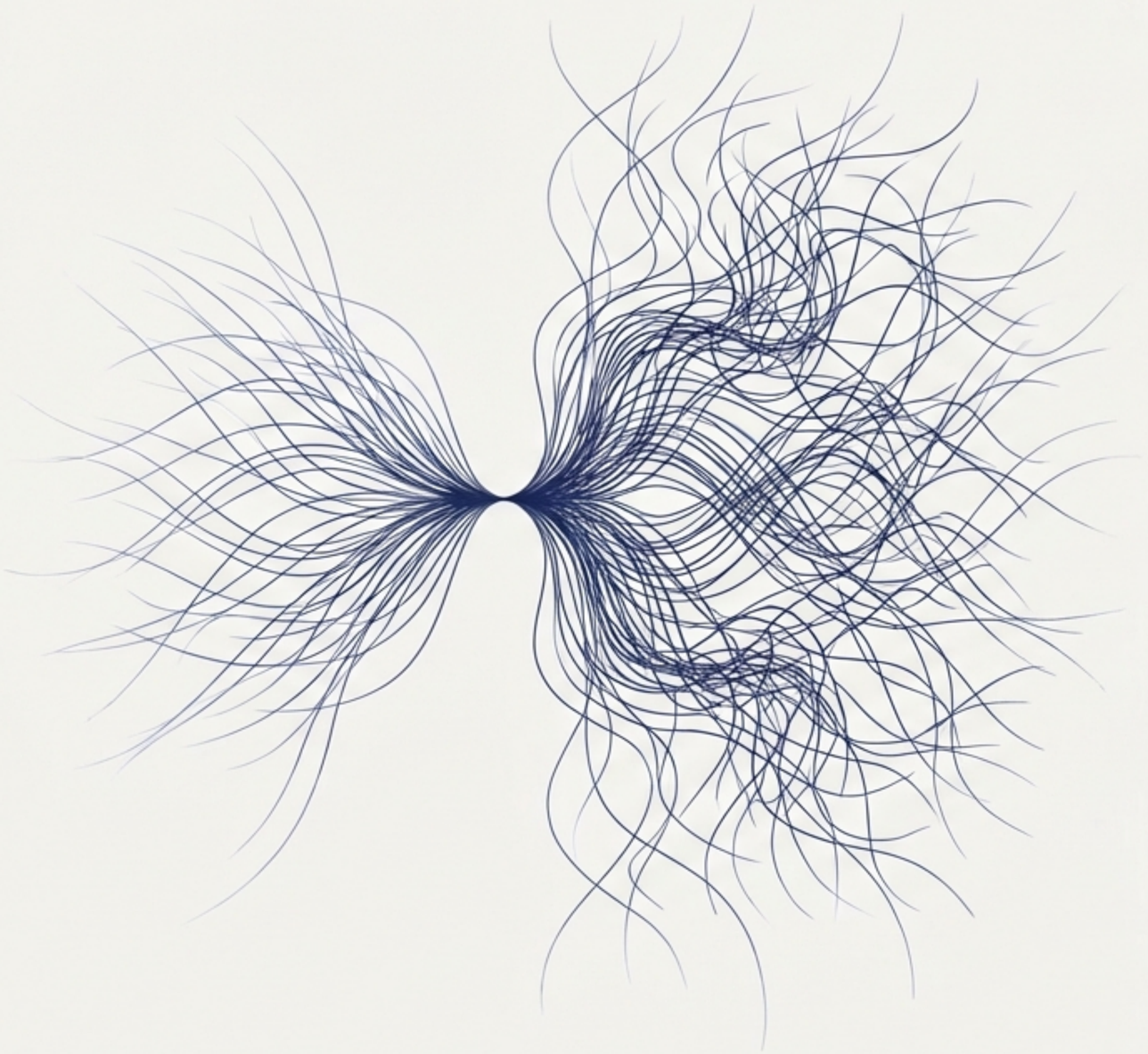


چالشی به نام «ناهار رایگان» وجود ندارد

- در دنیای یادگیری ماشین، هیچ مدلی برای همه‌ی مسائل بهترین نیست. (اشاره به قضیه No Free Lunch)
- داده‌های واقعی همیشه با چالش‌هایی همراه هستند: ابعاد بالا، نویز، و ساختارهای پیچیده.
- هدف ما: یافتن اینکه کدام الگوریتم‌های کلاسیک در سخت‌ترین شرایط، قابل اعتمادتر هستند.
- ما یک آزمون بزرگ طراحی کردیم تا قهرمانان واقعی را شناسایی کنیم.

NotebookLM

معرفی رقبا: هفت جنگجوی کلاسیک

هفت الگوریتم پرکاربرد و کلاسیک را برای این نبرد انتخاب کرده‌ایم. هر کدام با استراتژی منحصربه‌فرد خود وارد میدان می‌شوند.

درخت تصمیم

رگرسیون لجستیک

kNN: نزدیک‌ترین همسایه

SVM: ماشین بردار پشتیبان

شبکه عصبی (MLP)

مدل خطی

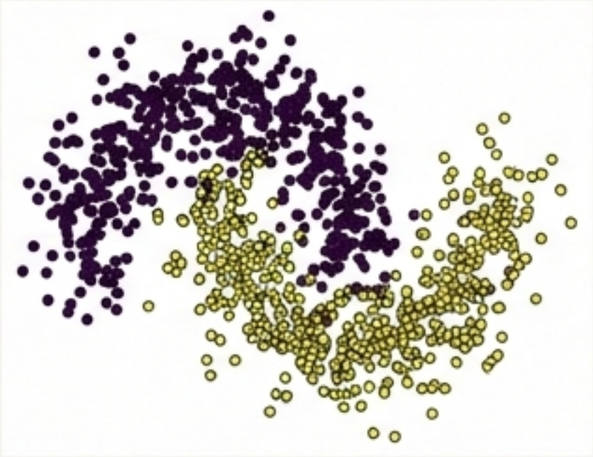
بیز ساده

NotebookLM

سه میدان نبرد، سه آزمون نهایی

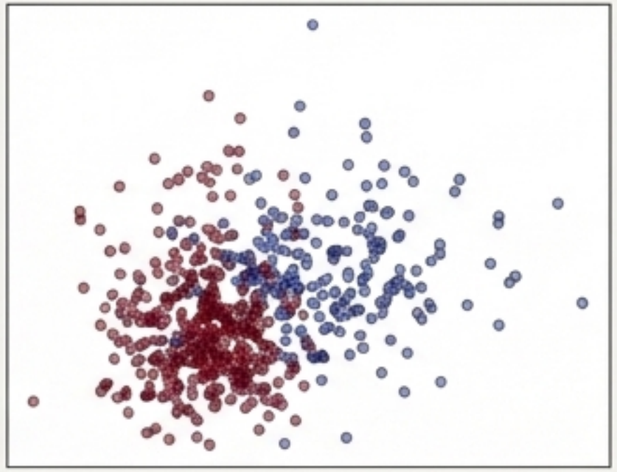
برای سنجش واقعی توانایی‌ها، سه مجموعه داده با چالش‌های ساختاری متفاوت طراحی کردیم:

میدان سوم: مقابله با پیچیدگی غیرخطی (Moons)



میدان سوم: دانه مستقلی
چالش: داده‌های غیرقابل تفکیک با خط آزمون واقعی برای درک الگوهای پیچیده. پیچیده.

میدان دوم: آزمون پایداری در برابر نویز (Breast Cancer + Noise)



میدان دوم: آزمون پایداری
چالش: افزودن نویز مصنوعی.
کدام مدل سیگنال را از نویز تشخیص می‌دهد؟

میدان اول: نفرین ابعاد (Digits)

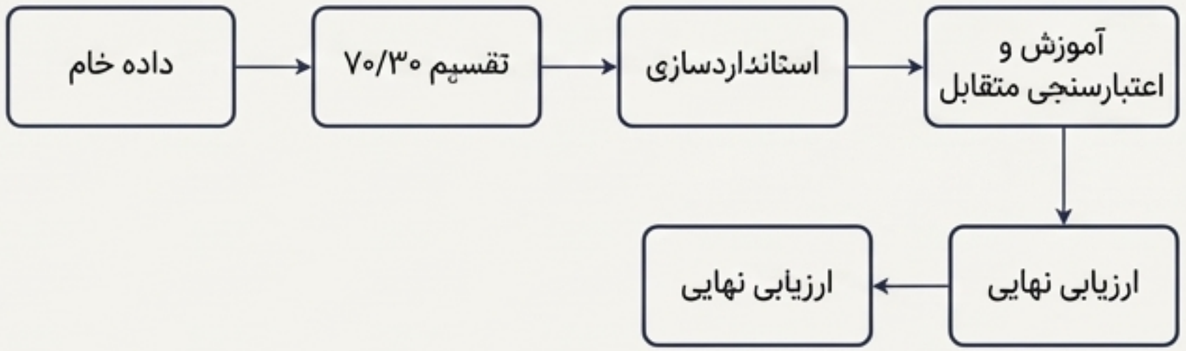
میدان اول: نفرین ابعاد
چالش: ۶۴ ویژگی (پیکسل). آیا مدل‌ها در فضای ابعاد بالا مسیر خود را گم می‌کنند؟

NotebookLM

قوانین یک نبرد عادلانه: متدولوژی آزمایش

برای تضمین یک مقایسه علمی و منصفانه، چارچوب دقیقی را دنبال کردیم:

- **آماده‌سازی دقیق:**
  - **تقسیم داده‌ها:** ۷۰٪ آموزش، ۳۰٪ تست (Stratified) برای جلوگیری از Data Leakage.
  - **استانداردسازی** (StandardScaler): حذف اثر مقیاس ویژگی‌ها، حیاتی برای SVM و kNN.
- **انتخاب هوشمندانه:**
  - تنظیم **فراپارامترها** با Grid Search برای یافتن یافتن بهترین نسخه از هر مدل.
  - **اعتبارسنجی متقابل** (Cross-Validation 5-Fold Fold) برای نتایج قابل اعتماد.




```python
def preprocess_data(X, y):
    # Split
    X_train, X_test, y_train, y_test = train_test_split(
        X, y, test_size=0.3, random_state=42, stratify=y
    )
    # Scale
    scaler = StandardScaler()
    X_train_scaled = scaler.fit_transform(X_train)
    X_test_scaled = scaler.transform(X_test)

    return X_train_scaled, X_test_scaled, y_train, y_test
```

NotebookLM

دور اول: نبرد در نفرین ابعاد

در فضای ۶۴ بعدی دیتاست Digits، فاصله معنای خود را از دست می‌دهد.

بازنده اصلی

درخت تصمیم: با **دقت ۸۲.۲٪** شکست خورد. این مدل در تقسیم فضا بر اساس تک‌پیکسل‌ها دچار مشکل شده و به بیش‌برازش می‌رسد.

برندگان

SVM (با کرنل RBF): با دقت **۹۸.۳٪**، برتری خود را درمدیریت فضاهای با ابعاد بالا از طریق طریق بیشینه‌سازی حاشیه نشان داد.

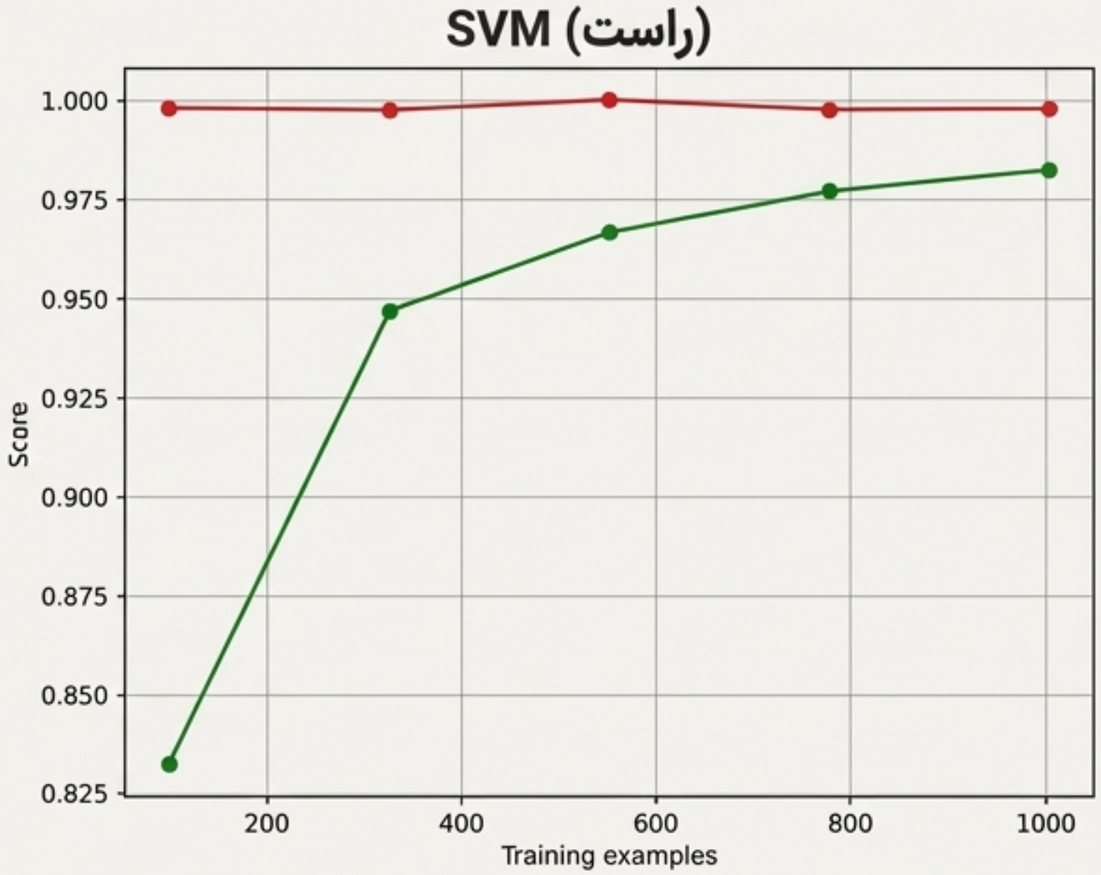
MLP: با دقت ۹۷.۹٪ عملکردی نزدیک و قدرتمند داشت.

نکته کلیدی: مدل‌هایی که ساختار کلی فضا را درک می‌کنند، بر مدل‌های حریص و محلی غلبه کردند.

NotebookLM

تحلیل استراتژی: چرا SVM پیروز شد؟

منحنی‌های یادگیری، داستان تعمیم‌پذیری مدل را روایت می‌کنند.



همگرایی سریع‌تر و فاصله‌ی کمتر بین دو منحنی، نشان‌دهنده تعادل بهتر بین بایاس و واریانس و قابلیت تعمیم بالاتر است.

فاصله‌ی قابل توجه بین منحنی آموزش (قرمز) و اعتبارسنجی (سبز) نشان‌دهنده «واریانس بالا» است. مدل داده‌های آموزشی را «حفظ» کرده است.

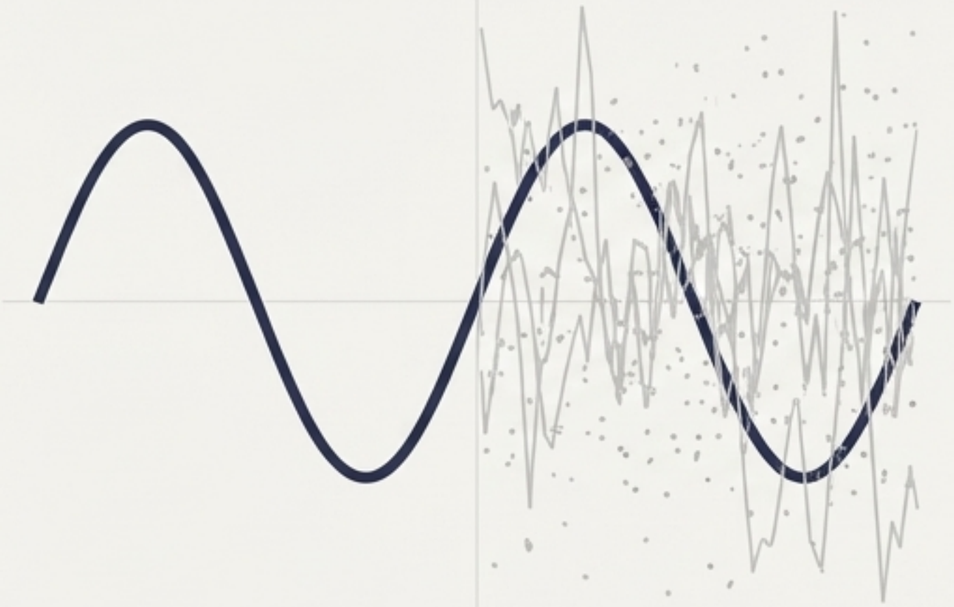
NotebookLM

دور دوم: آزمون پایداری در برابر نویز

با افزودن نویز گاوسی به دیتاست Breast Cancer، مرز بین کلاس‌ها مخدوش شد. این یک آزمون سخت برای سنجش تعمیم‌پذیری است.

قربانی بیش‌برازش

MLP*: دقت آموزش ۱۰۰٪ اما دقت تست ۹۱.۸٪.
این مدل به شدت نویز را به عنوان الگو یاد گرفته و دچار Overfitting شده است.



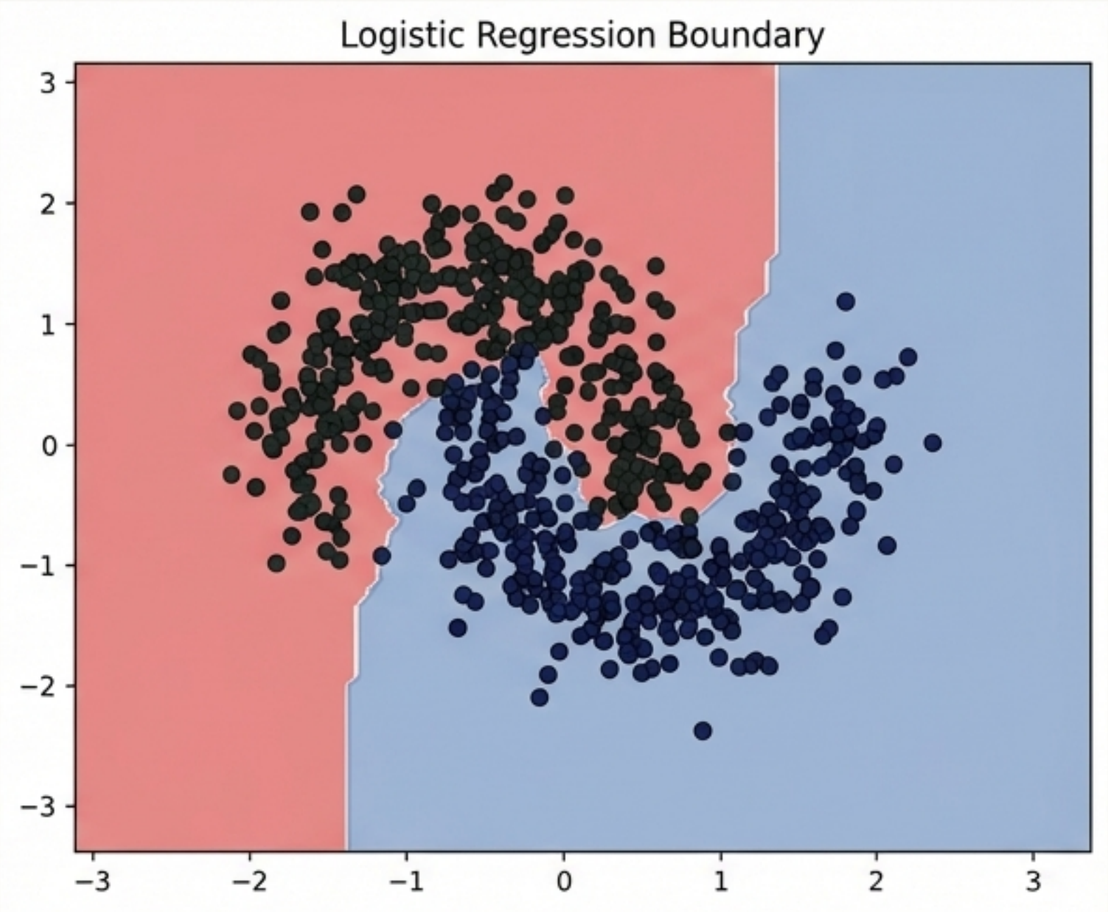
قهرمانان پایداری

SVM* و رگرسیون لجستیک:
با دقت بالای ۹۵٪، پایدارترین عملکرد را داشتند.
مکانیزم‌های Regularization به آن‌ها اجازه داد تا نویز را نادیده بگیرند و بر ساختار کلی تمرکز کنند.

NotebookLM

دور سوم: شکستن منحنی‌های غیرخطی

در دیتاست Moons، هیچ خط مستقیمی قادر به جداسازی دو کلاس نیست. اینجاست که ضعف مدل‌های خطی آشکار می‌شود.



شکست مدل‌های خطی

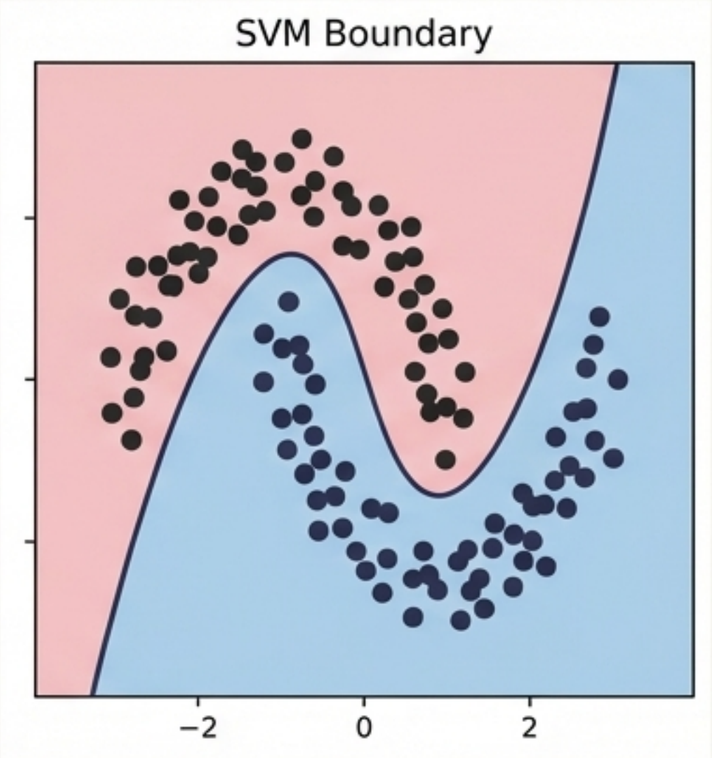
رگرسیون لجستیک و مدل خطی با دقت زیر ۹۰٪ شکست خوردند. آن‌ها ذاتاً برای این هندسه طراحی نشده‌اند و دچار «بایاس بالا» (High Bias) هستند.

پیروزی انعطاف‌پذیری

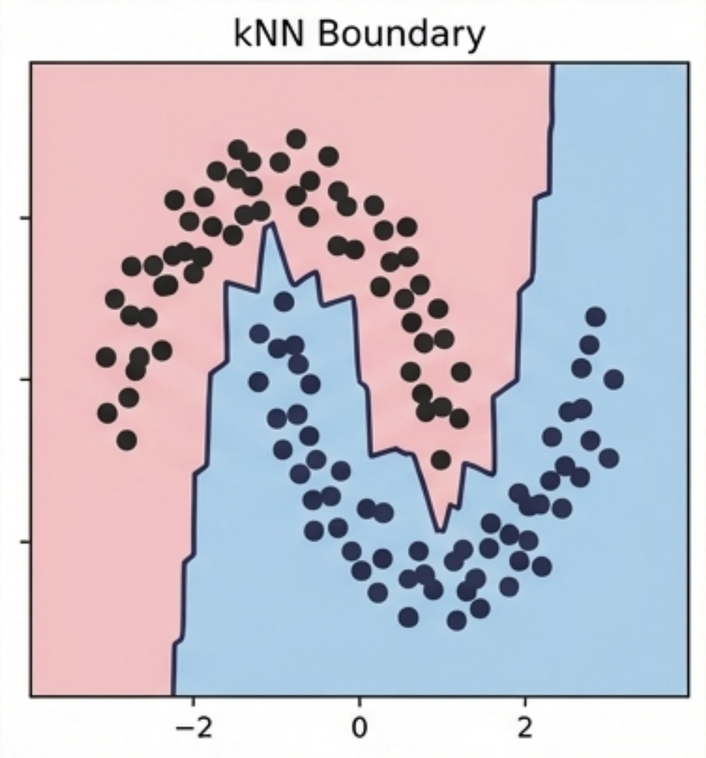
SVM (با کرنل RBF)، kNN و MLP با دقت بالای ۹۸٪ پیروز شدند. این مدل‌ها توانستند مرزهیم پیچیده و منحنی را به خوبی یاد بگیرند.

NotebookLM

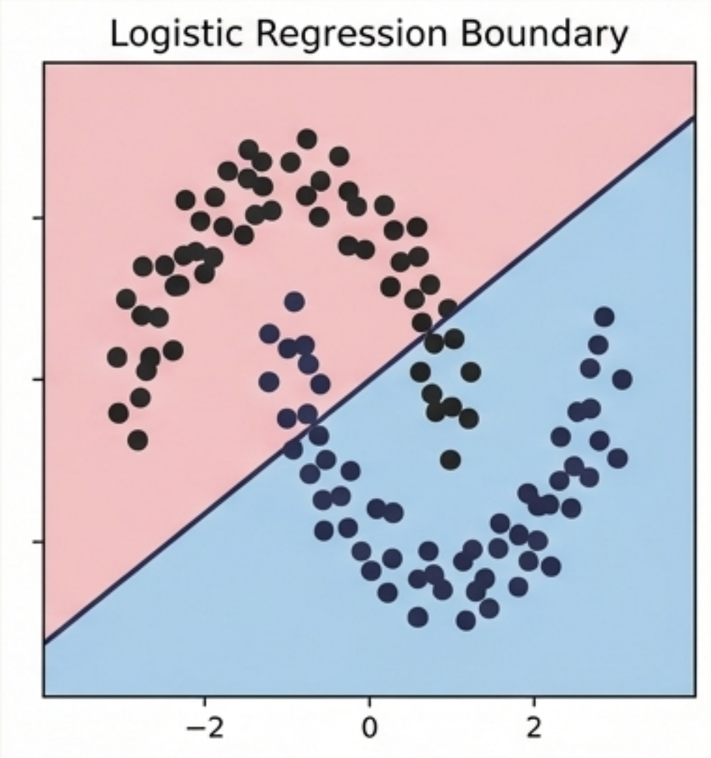
مرزهای تصمیم: داستان در تصاویر



SVM (راست): با استفاده از کرنل RBF، داده‌ها را به فضایی بالاتر نگاشت داده و یک مرز جداسازی نرم و بهینه ایجاد کرده است.



kNN (وسط): ایجاد یک مرز محلی و دقیق که از داده‌ها پیروی می‌کند.



رگرسیون لجستیک (چپ): تلاش ناموفق برای تفکیک با یک خط راست.

NotebookLM

جدول نهایی نتایج: قهرمان آزمون بزرگ کیست؟

پس از سه دور رقابت نفس‌گیر، نتایج نهایی عملکرد هر مدل را در هر سه میدان نبرد خلاصه می‌کند.

	نفرین ابعاد	پایداری در برابر نویز	پیچیدگی غیرخطی
SVM	98.3%	95.3%	98.7%
kNN	97.4%	93.0%	98.3%
رگرسیون لجستیک	98.1%	93.6%	89.7%
درخت تصمیم	82.2%	91.2%	96.0%
بیز ساده	78.7%	91.8%	89.3%
مدل خطی	95.4%	93.0%	87.0%
شبکه عصبی (MLP)	98.0%	91.8%	98.3%

تحلیل سریع:

- SVM در هر سه چالش در بین سه مدل برتر قرار دارد.
- مدل‌های خطی فقط در شرایط ساده و بدون نویز موفق هستند.
- درخت تصمیم و بیز ساده در چالش‌های پیچیده دچار مشکل شدند.

NotebookLM

قهرمان همه‌کاره: برتری ماشین بردار پشتیبان (SVM)

در مجموع سه آزمایش، **SVM** به عنوان قوی‌ترین، پایدارترین و همه‌کاره‌ترین مدل شناخته شد.

کرنل‌ها (Kernel Trick)

توانایی حل مسائل غیرخطی بدون افزایش پیچیدگی محاسباتی.

بیشینه‌سازی حاشیه (Margin Maximization)

ایجاد مرز تصمیمی که بیشترین فاصله را از نقاط داده دارد و منجر به تعمیم‌پذیری بالا در مقابل ابعاد بالا و نویز می‌شود.

تنظیم‌پذیری (Regularization)

کنترل پیچیدگی مدل (پارامتر C) برای جلوگیری از بیش‌برازش.

NotebookLM

دفترچه استراتژی: کدام الگوریتم برای کدام نبرد؟

این پژوهش نشان داد که انتخاب الگوریتم باید بر اساس ماهیت داده و نوع چالش باشد.

برای داده‌های پیچیده با الگوی ناشناخته و غیرخطی:

انتخاب اول: SVM و kNN گزینه‌های امنی هستند.

با احتیاط: MLP، نیازمند تنظیم دقیق برای جلوگیری از بیش‌برازش.

برای مسائل سریع و قابل تفسیر با ساختار ساده:

انتخاب اول: رگرسیون لجستیک.

برای داده‌های با ابعاد بالا (مانند تصاویر):

انتخاب اول: SVM با کرنل RBF.

***قانون نهایی:** بهترین مدل وجود ندارد، بهترین استراتژی برای هر مسئله وجود دارد.*

NotebookLM