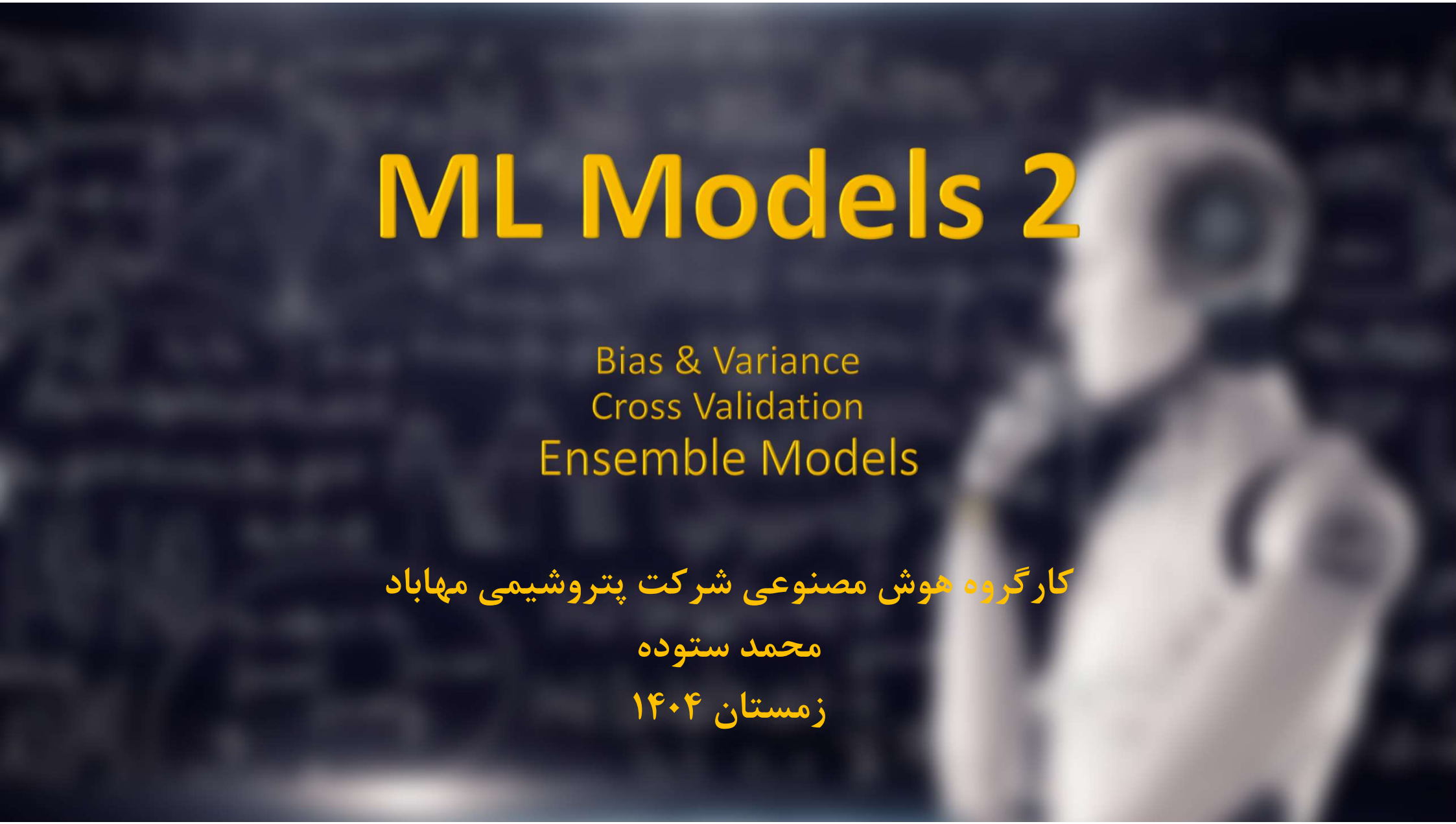
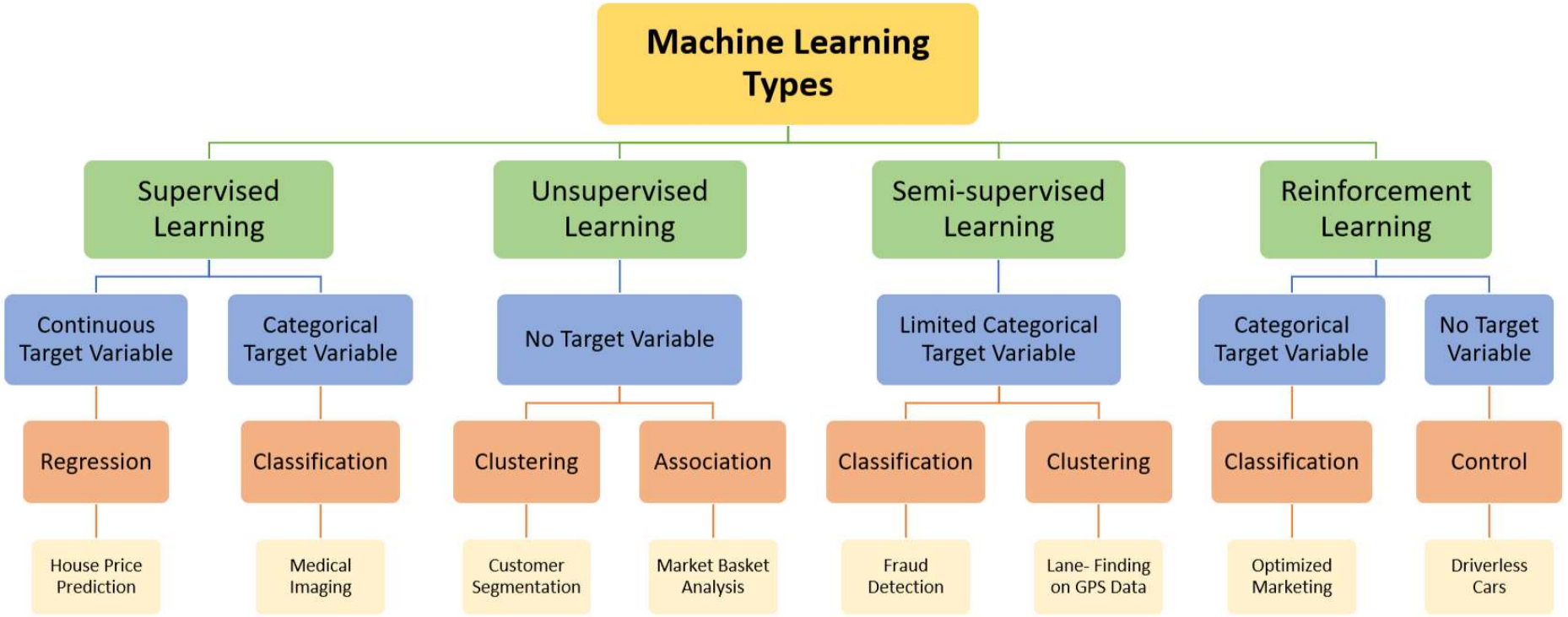

ML Models 2
Bias & Variance
Cross Validation
Ensemble Models
کارگروه هوش مصنوعی شرکت پتروشیمی مهاباد
محمد ستوده
زمستان ۱۴۰۴

ML

2. Machine Learning (ML)
2.1. Supervised Learning
2.2. Unsupervised Learning
2.3. Semi-Supervised & Self-Supervised Learning
2.4. Reinforcement Learning (RL)
2.5. Probabilistic & Bayesian Learning
2.6. Graph-based Learning
2.7. Representation Learning
2.8. Federated, Online & Data Stream Learning



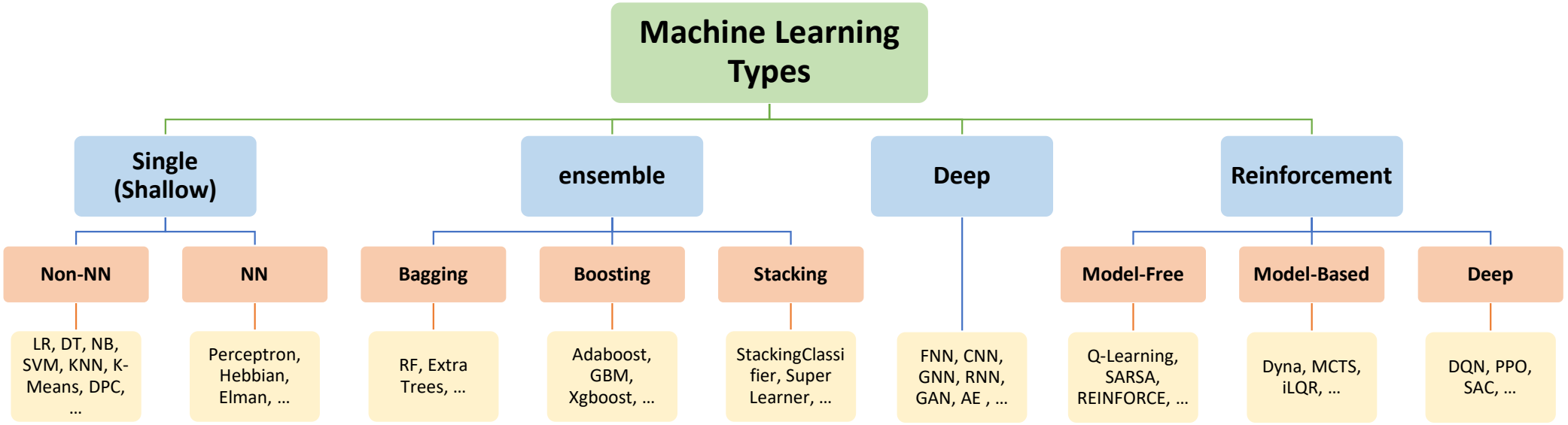
ML Models Taxonomy

«این تقسیم‌بندی برای یادگیریه، نه یک استاندارد آکادمیک سخت»

- از منظر جنس مدل (غیر عصبی- عصبی)
- از منظر استراتژی کاربست (تکی- جمعی)
- از منظر عمق مدل (سطحی- عمیق)

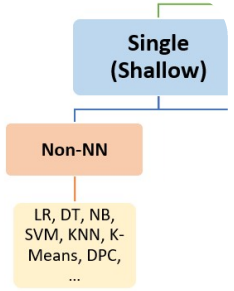
ML Models Taxonomy

«این تقسیم‌بندی برای یادگیریه، نه یک استاندارد آکادمیک سخت»



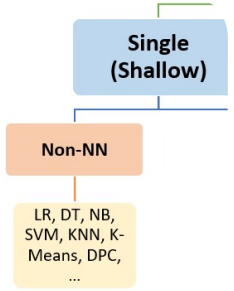
Self-supervised Learning, Graphical Models, and Probabilistic ML?! Maybe Later!

Single Shallow Non-NN



- Supervised
  - DT — درخت تصمیم:
  - NB — نایِیف بیز:
  - KNN — K نزدیکترین همسایگی:
  - SVM — ماشین بردار پشتیبان:
  - Liner Regression — رگرسیون خطی:
- Unsupervised
  - K-Means — K-Means:
  - DBSCAN — DBSCAN:

Single Shallow Non-NN



- Supervised
  - DT
  - NB
  - KNN
  - SVM
  - Liner Regression
- Unsupervised
  - K-Means
  - DBSCAN

درخت تصمیم: تقسیم بندی داده بر اساس سوال! بله/خیر

نایِیف بیز: ضرب احتمال‌های ساده‌ی هر ویژگی، برچسب با بیشترین احتمال

K نزدیکترین همسایگی: برچسب نزدیکترین نمونه ها به داده جدید

ماشین بردار پشتیبان: خط مرزی با بیشترین فاصله از نزدیک‌ترین نمونه‌های هر کلاس

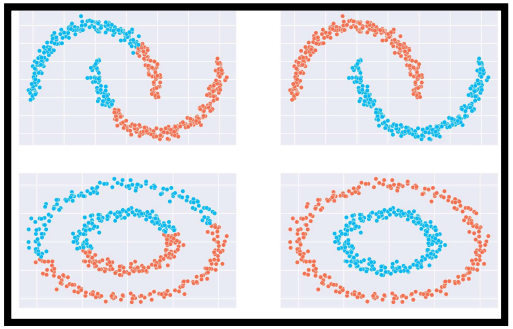
رگرسیون خطی: کدام خط / صفحه کمترین فاصله‌ی مجموع از همه‌ی نقاط را دارد؟

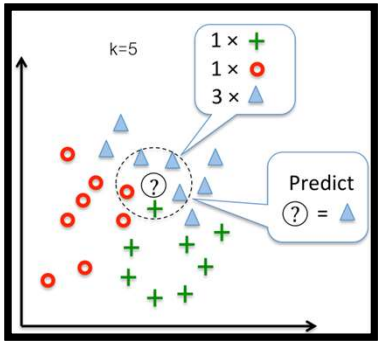
K-Means: هر نقطه به نزدیکترین مرکز تعلق دارد و مراکز خوشه میانگین مقادیر داده ها هستند.

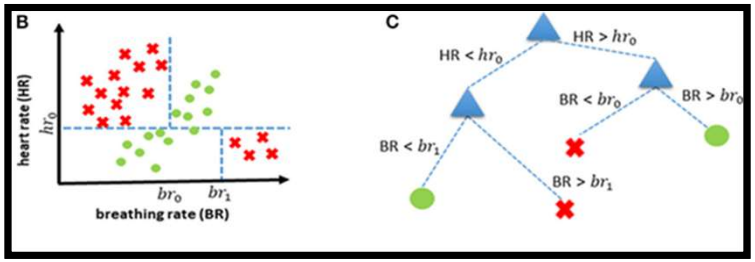
DBSCAN: تشخیص خوشه‌ها را بر اساس تراکم نقاط

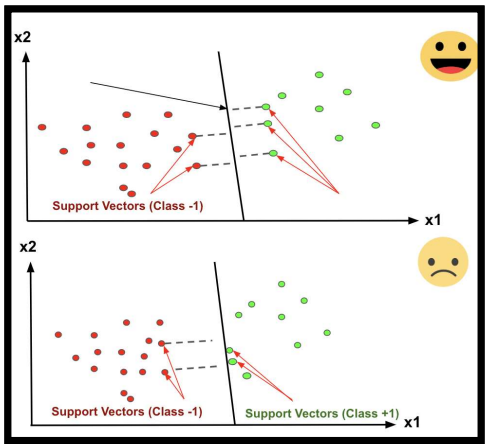
Single Shallow Non-NN

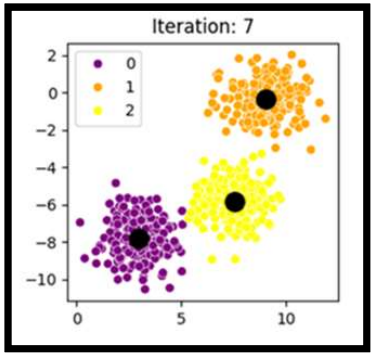
- Supervised
  - DT
  - NB
  - KNN
  - SVM
  - Liner Regression
- Unsupervised
  - K-Means
  - DBSCAN

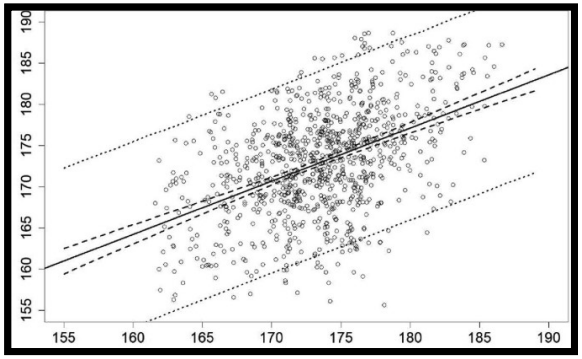








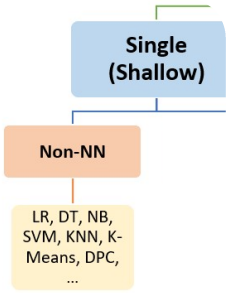




Outlook	Temp	Humidity	Windy	Play?
Rainy	Hot	High	False.	No
Rainy	Hot	High	True.	No
Overcast	Hot	High	False.	Yes
Sunny	Mild	High	False.	Yes
Sunny	Cool	Normal	False.	Yes
Sunny	Cool	Normal	True.	No
Overcast	Cool	Normal	True.	Yes
Rainy	Mild	High	False.	No
Rainy	Cool	Normal	False.	Yes
Sunny	Mild	Normal	False.	Yes
Rainy	Mild	Normal	True.	Yes
Overcast	Mild	High	True.	Yes
Overcast	Hot	Normal	False.	Yes
Sunny	Mild	High	True.	No

query:	Outlook	Temp	Humidity	Windy	Play?
	Rainy	Cool	High	TRUE	?
Play No	3/5	1/4	4/7	3/6	
Play Yes	2/5	3/4	3/7	3/6	

query:	Rainy	Cool	High	TRUE	Multiply		
Play No	0.60%	0.25%	0.57%	0.50%	0.0000000004286		
Play Yes	0.40%	0.75%	0.43%	0.50%	0.0000000006429	Yes	60.0%

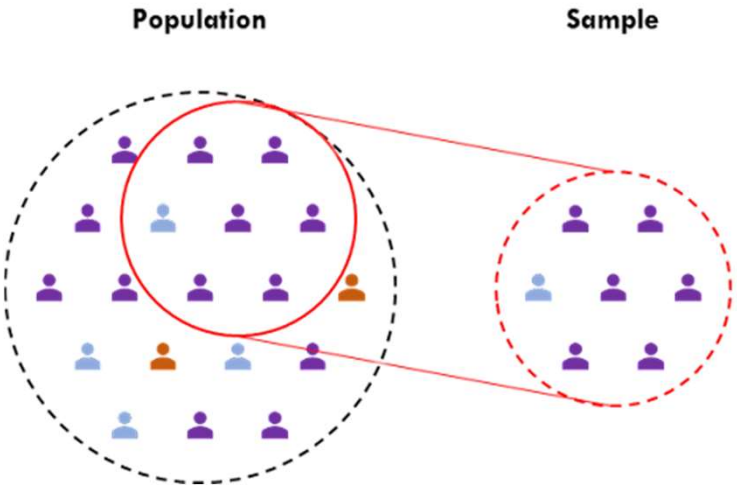


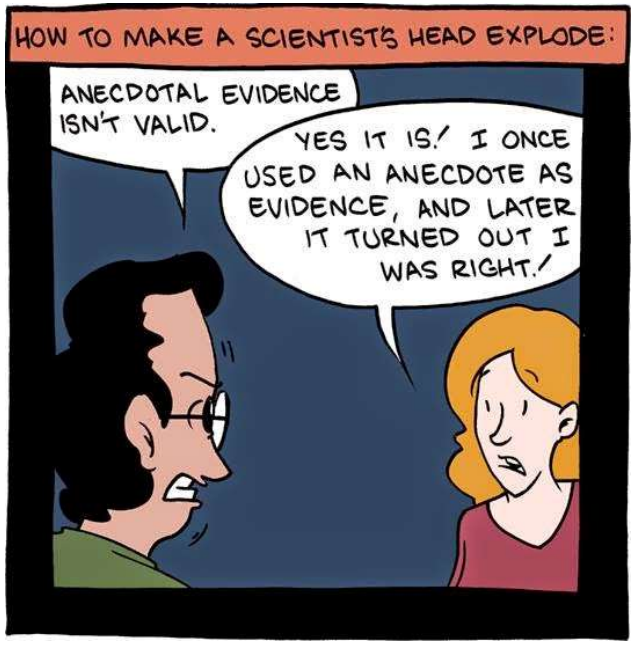
Bias & Variance

بایاس در داده

تعمیم جزء به کل!

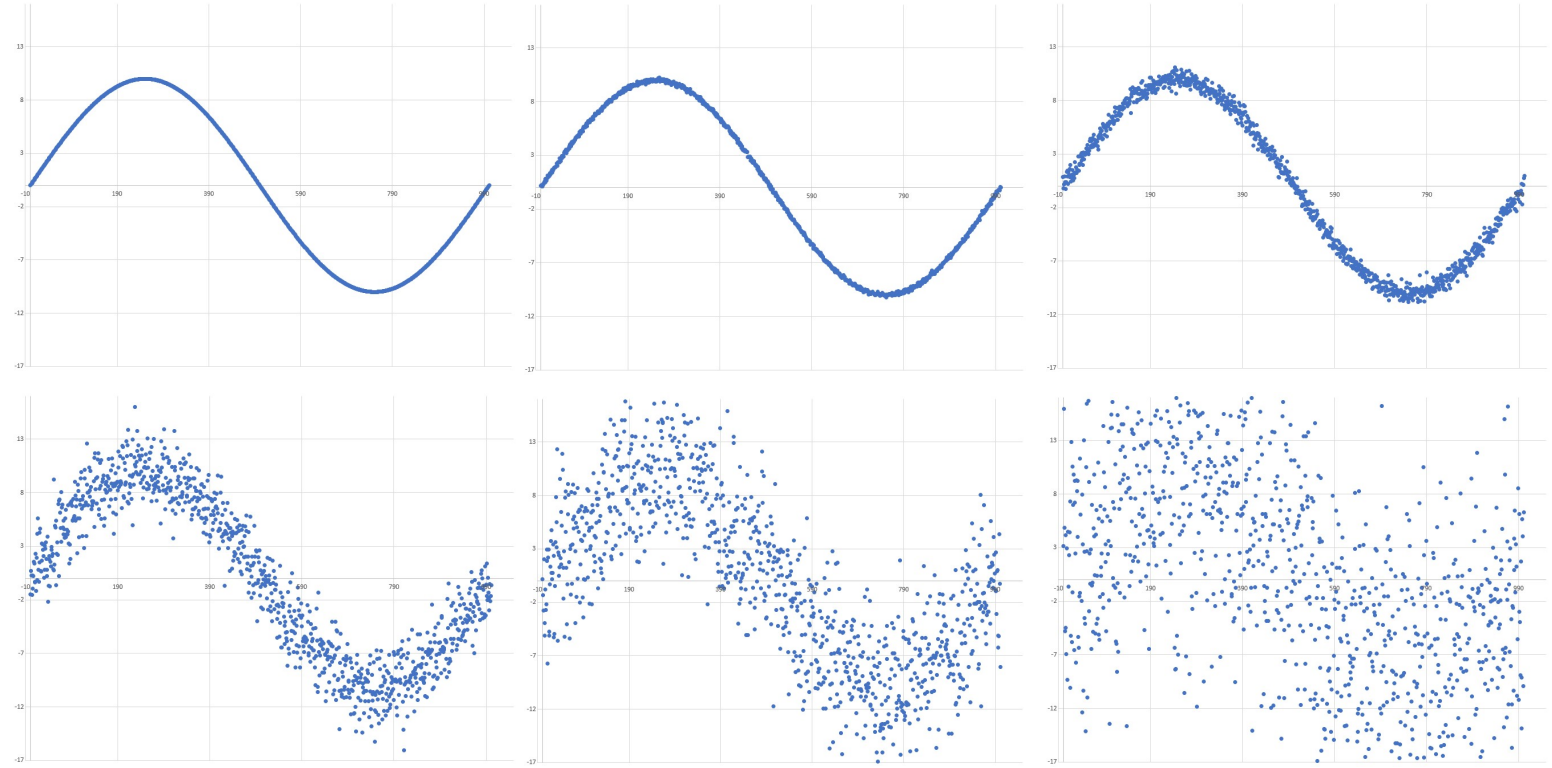
Hasty Generalization Fallecy

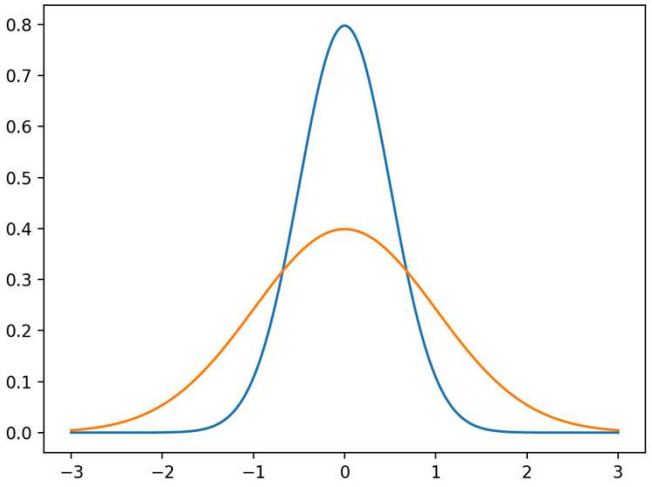




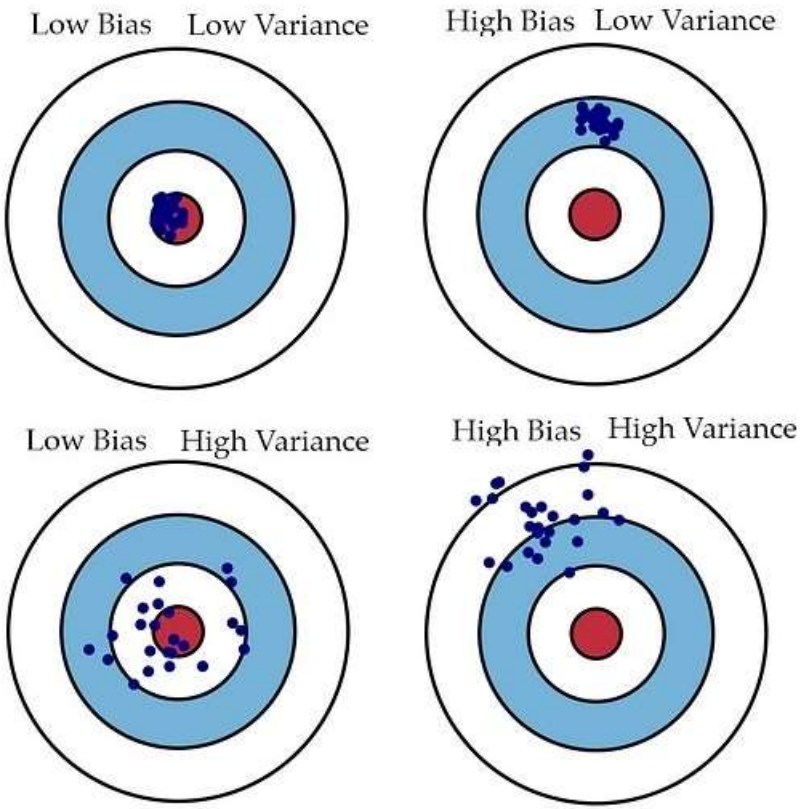
Bias & Variance

واریانس در داده



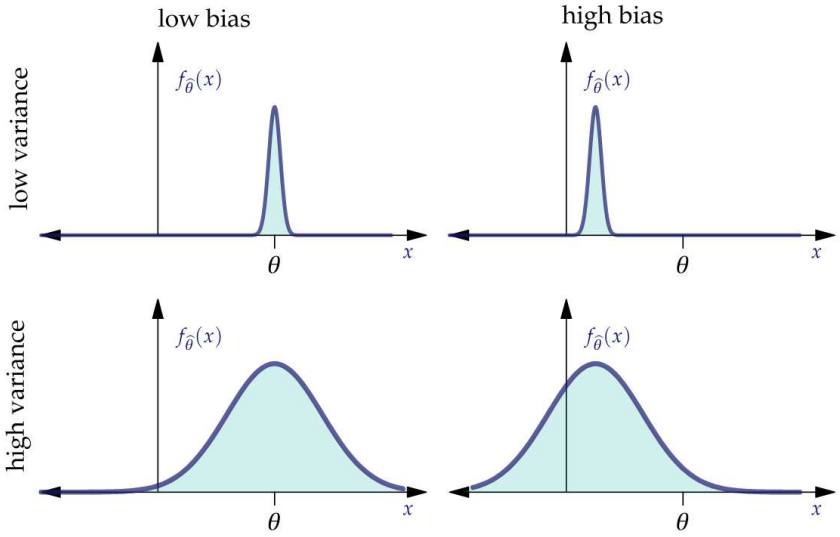


Bias & Variance



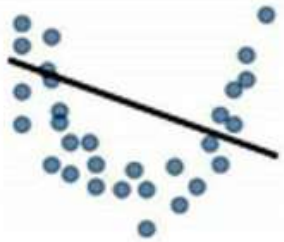
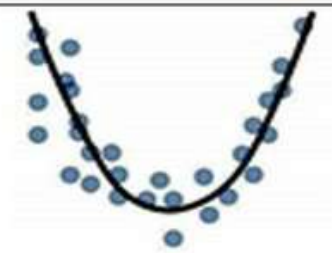
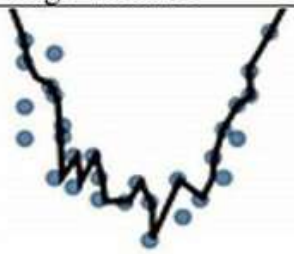
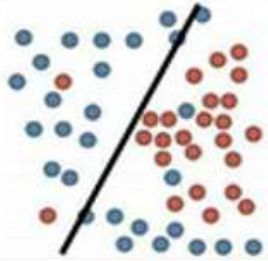
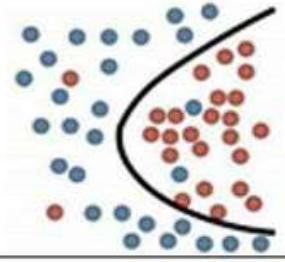
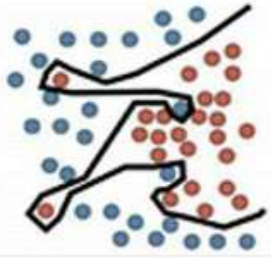
بایاس و واریانس در مدل

پراکندگی و دقت خروجی
مدل نسبت به واقعیت

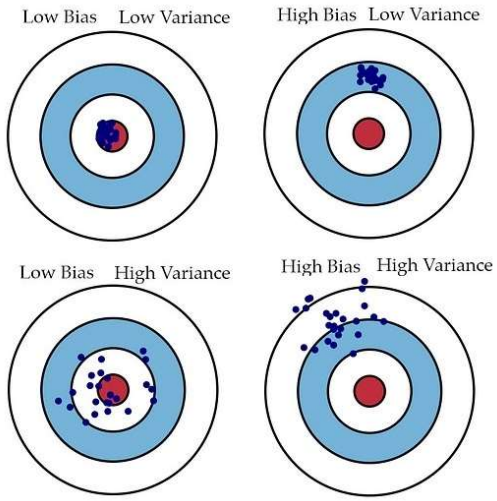


Bias & Variance

بایاس و واریانس در مدل

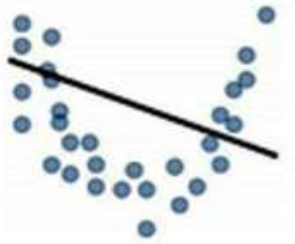
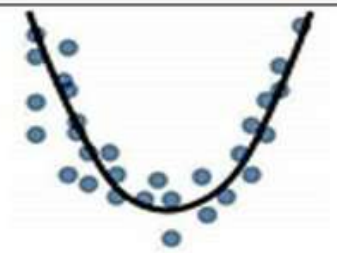
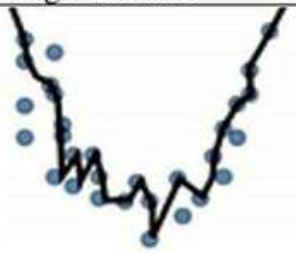
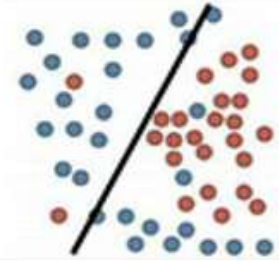
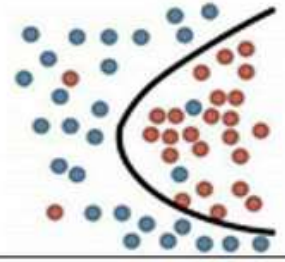
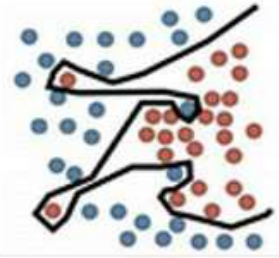
	Underfitting	Just right	Overfitting
Symptoms	• High training error • Training error clos to test error • High bias	• Training error slightly lower than test error	• Very low training error • Training error much lower than test error • High variance
Regression illustration			
Classification illustration			

پراکندگی و دقت خروجی مدل نسبت به واقعیت



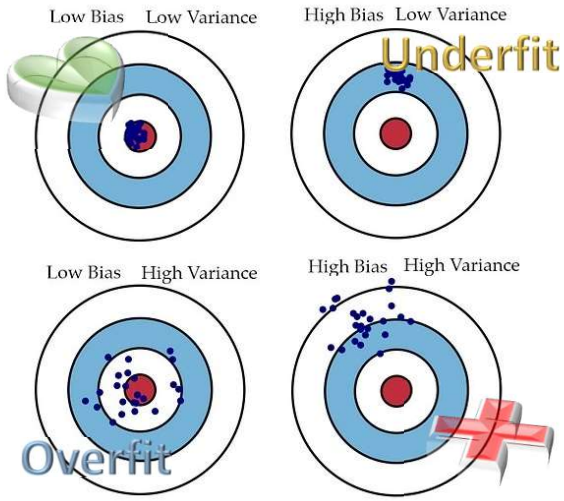
Bias & Variance

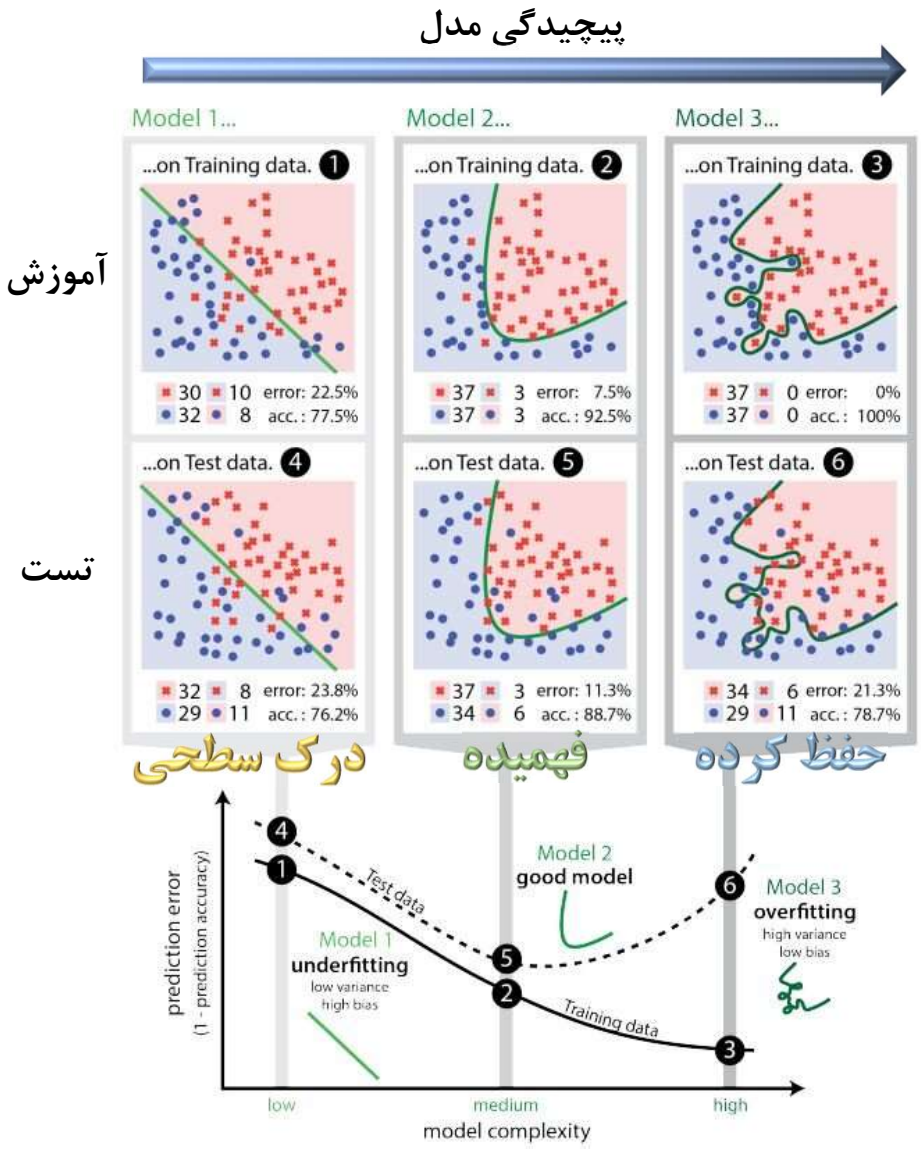
درک سطحی | فهمیده | حفظ کرده

	Underfitting	Just right	Overfitting
Symptoms	• High training error • Training error clos to test error • High bias	• Training error slightly lower than test error	• Very low training error • Training error much lower than test error • High variance
Regression illustration			
Classification illustration			

بایاس و واریانس در مدل

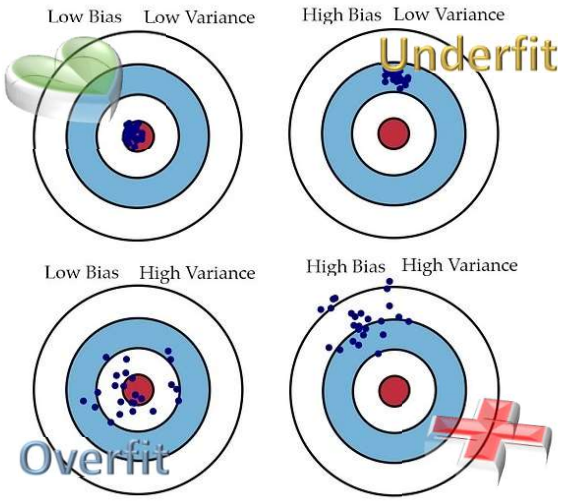
پراکندگی و دقت خروجی مدل نسبت به واقعیت





بایاس و واریانس در مدل

پراکندگی و دقت خروجی مدل نسبت به واقعیت

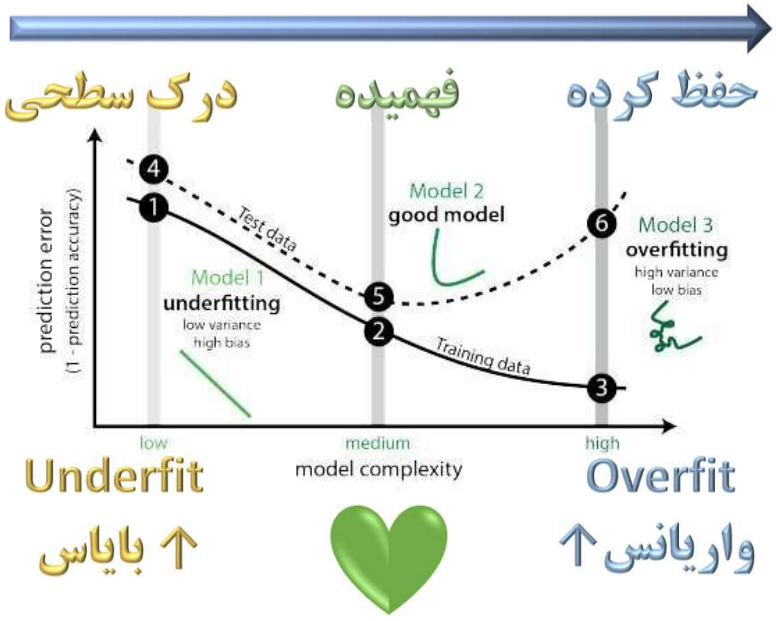


Bias & Variance

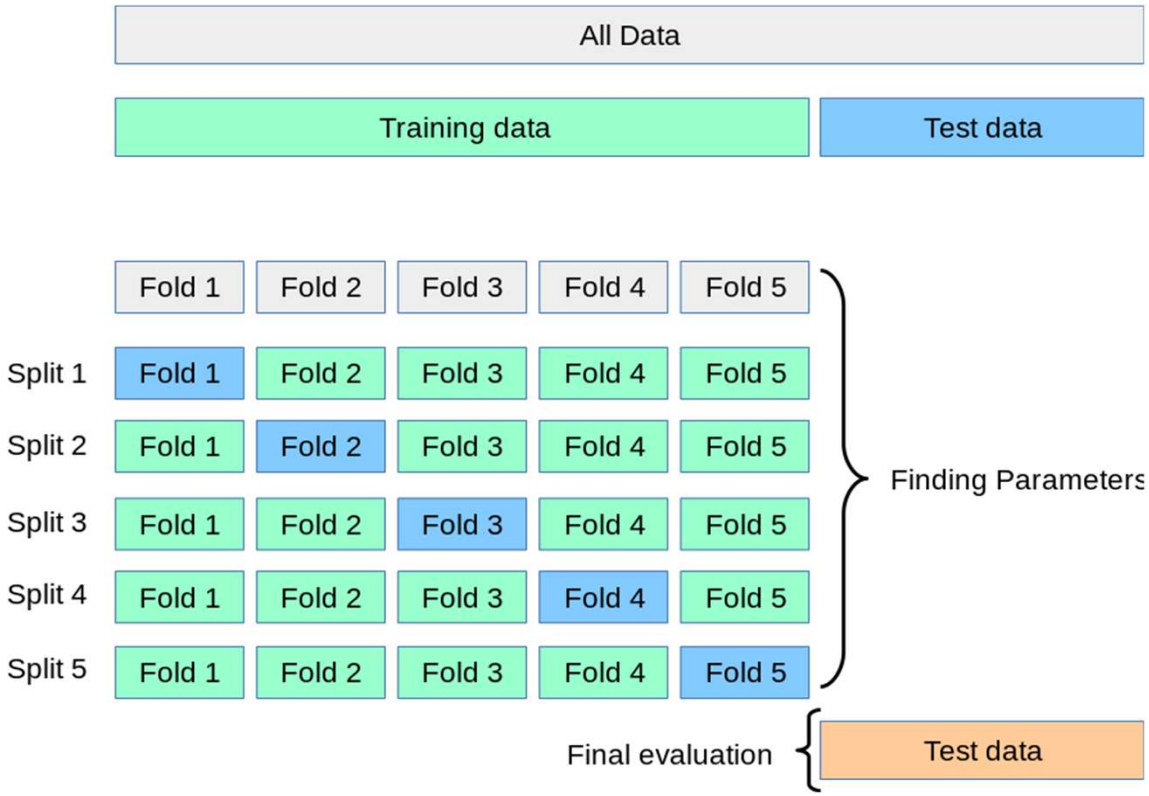
روش	بایاس	واریانس
Cross-Validation	کمک به مدیریت	کمک به مدیریت
Bagging	→	↓↓↓
Boosting	↓↓	↓ (اغلب)
Stacking	↓	↓
افزایش پیچیدگی مدل	↓	↑
کاهش پیچیدگی مدل	↑	↓
Regularization (L۱/L۲)	↑ (کمی)	↓
Feature Selection	↑ (ممکن است)	↓
Feature Engineering خوب	↓	↓
افزایش داده	→	↓
حذف نویز	→	↓
Early Stopping	↑ / →	↓
Dropout	↑ / →	↓

روش‌های مدیریت
بایاس و واریانس

پیچیدگی مدل



Cross Validation



CV مدل را قوی‌تر نمی‌کند؛ بلکه تخمینِ بهتر و انتخابِ **هایپرپارامترِ** مناسب را ممکن می‌سازد که منجر به مدلی با بایاس/ واریانس مطلوب‌تر می‌شود.

انواع:

k-fold CV

Stratified k-fold CV

Leave-One-Out CV (LOOCV)

Repeated k-fold CV

Hold-out (train/validation/test split)

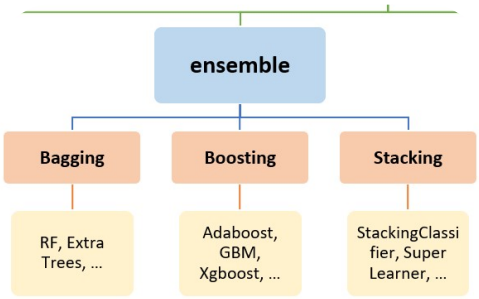
Time-series (forward chaining/ rolling-origin) CV

Nested CV

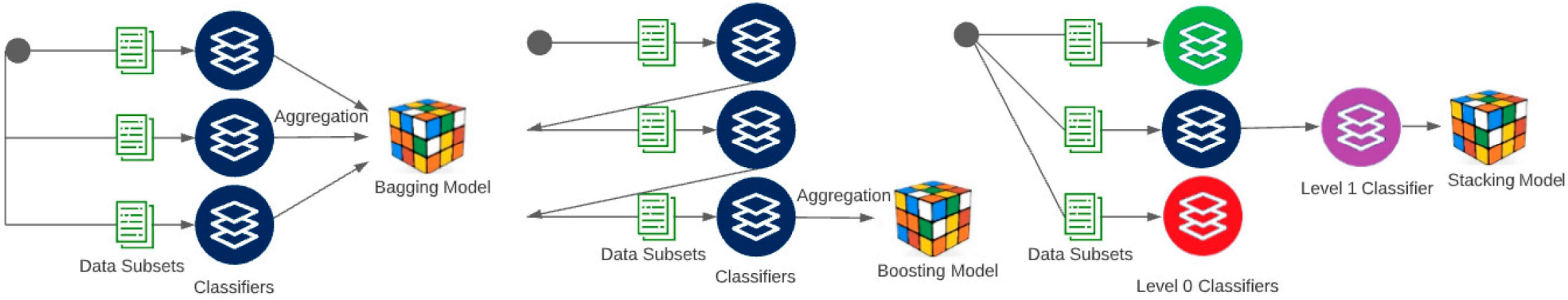
Monte Carlo (Random) CV / ShuffleSplit

Group k-fold CV

Ensemble Models

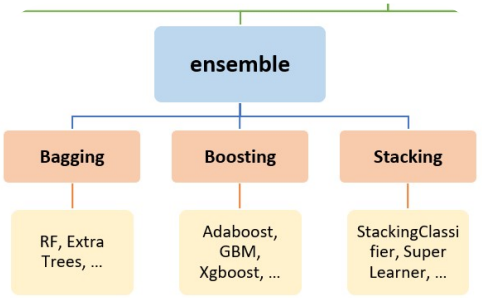


- Ensemble Strategies
  - Bagging
  - Boosting
  - Stacking



(**A**) Bagging (**B**) Boosting (**C**) Stacking

Ensemble Models



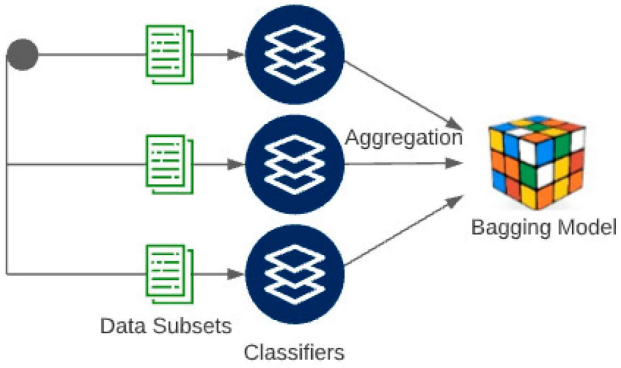
- Ensemble Strategies
  - Bagging (bootstrap aggregating)
  - Boosting
  - Stacking

بگینگ: مدل‌ها روی نمونه‌های بازگشت داده شده به مجموعه داده آموزش می‌بینند و خروجی‌ها تجمیع می‌شوند.

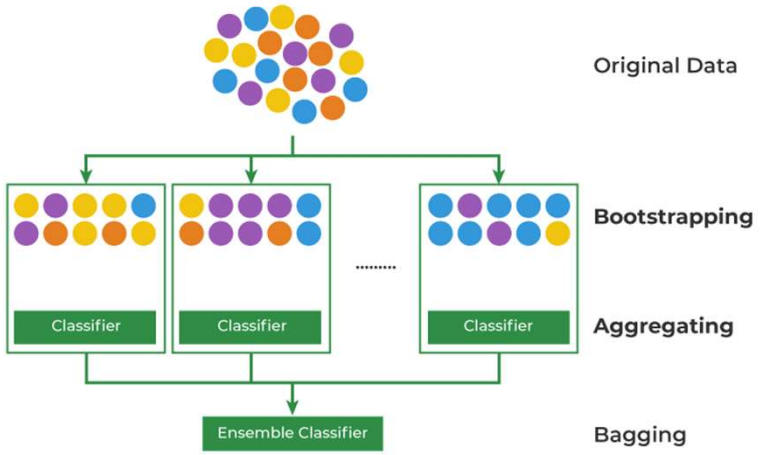
آموزش موازی چند مدل مشابه

داده ها با نمونه برداری تصادفی و جایگذاری مجدد

خروجی با رای گیری

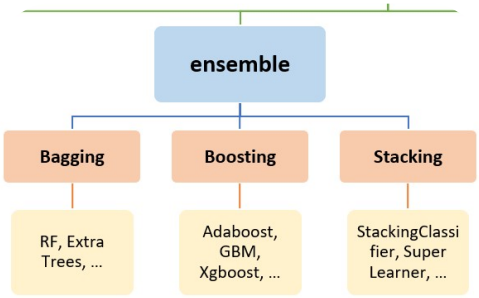


(A) Bagging



هدف: کاهش واریانس و افزایش پایداری مدل بدون افزایش بایاس قابل‌توجه.

Ensemble Models



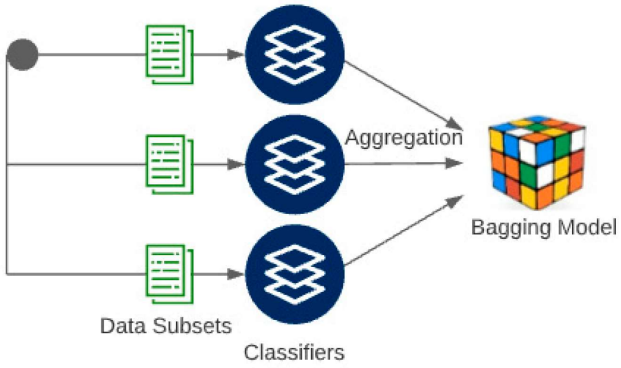
- Ensemble Strategies
  - Bagging (bootstrap aggregating)
  - Boosting
  - Stacking

مدل‌های معروف/معمول بگینگ

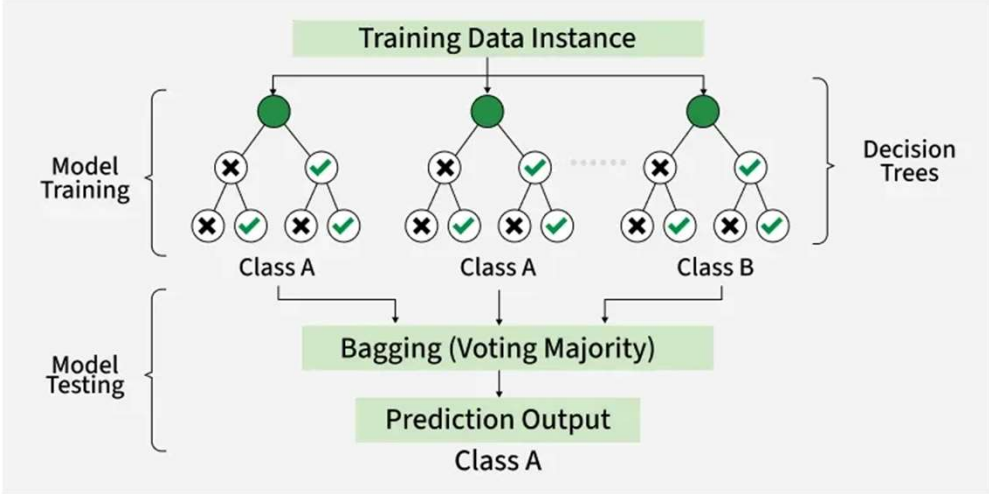
Random Forest

Bagging with k-NN

Bagging with SVM

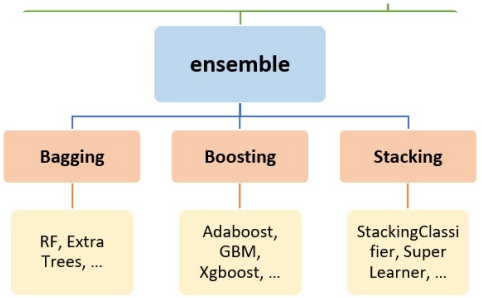


(A) Bagging

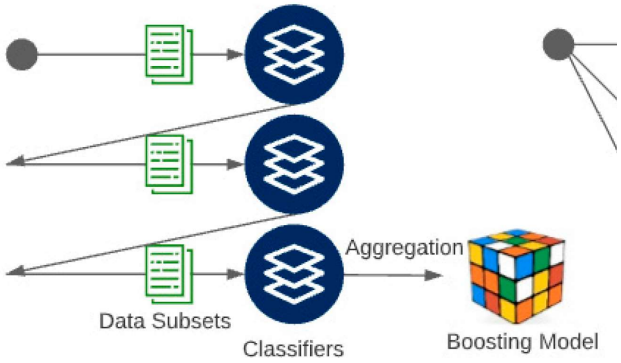


هدف: کاهش واریانس و افزایش پایداری مدل بدون افزایش بایاس قابل‌توجه.

Ensemble Models



- Ensemble Strategies
  - Bagging
  - Boosting
  - Stacking



بوستینگ

آموزش سری مدل‌ها. تمرکز بر نمونه‌های سخت.

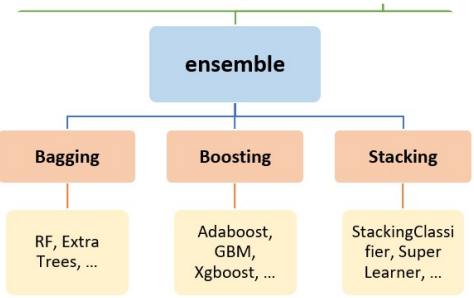
- آغاز با مدل پایه‌ی ساده (Weak learner)، معمولاً با وزن‌های برابر برای نمونه‌ها.
- در هر تکرار:
  - آموزش مدل پایه روی داده‌های آموزش با توجه به وزن‌های نمونه‌ها یا خطاهای قبلی.
  - افزایش وزن نمونه‌های دشوار.
  - ترکیب خروجی مدل جدید با مدل‌های قبلی.
  - وزن‌دهی مدل‌ها بر اساس دقت.
- ادامه تا رسیدن به تعداد مشخص یا توقف براساس خطا.

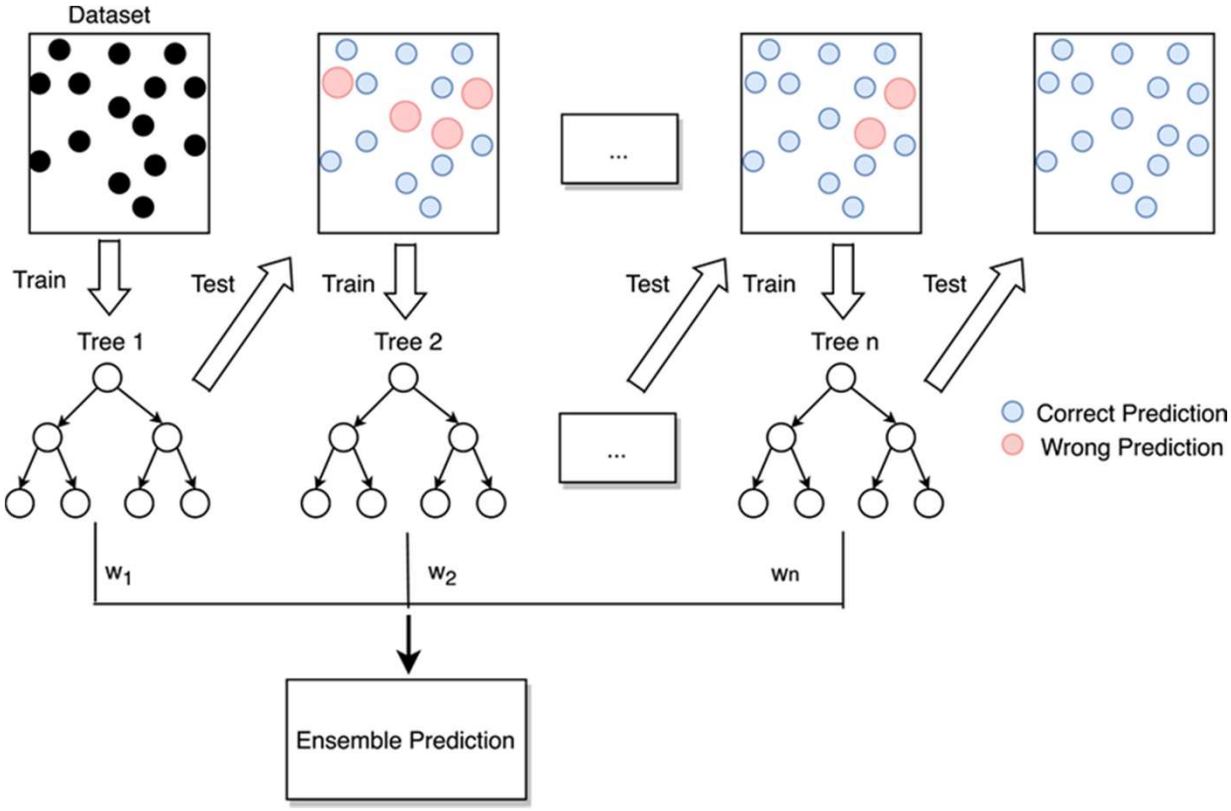
ترکیب نهایی مدل‌ها یک مدل قوی می‌سازد که خطاهای مدل‌های ضعیف را تصحیح کرده است.

(**B**) Boosting

هدف: کاهش بایاس و خطای کلی با ترکیب سری مدل‌های ضعیف که هر کدام روی نمونه‌های مشکل‌تر تمرکز می‌کنند.

Ensemble Models





(**B**) Boosting

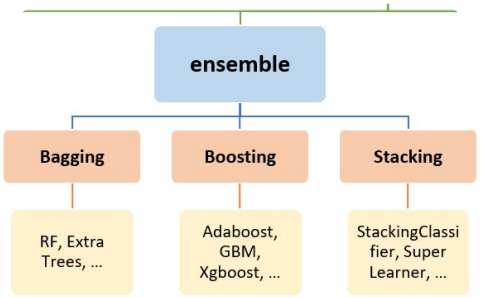
بوستینگ

آموزش سری مدل‌ها. تمرکز بر نمونه‌های سخت.

- آغاز با مدل پایه‌ی ساده (Weak learner)، معمولاً با وزن‌های برابر برای نمونه‌ها.
- در هر تکرار:
  - آموزش مدل پایه روی داده‌های آموزش با توجه به وزن‌های نمونه‌ها یا خطاهای قبلی.
  - افزایش وزن نمونه‌های دشوار.
  - ترکیب خروجی مدل جدید با مدل‌های قبلی.
  - وزن‌دهی مدل‌ها بر اساس دقت.
- ادامه تا رسیدن به تعداد مشخص یا توقف براساس خطا.

هدف: کاهش بایاس و خطای کلی با ترکیب سری مدل‌های ضعیف که هر کدام روی نمونه‌های مشکل‌تر تمرکز می‌کنند.

Ensemble Models



- Ensemble Strategies
  - Bagging
  - Boosting
  - Stacking

مدل‌های معروف/معمول بوستینگ

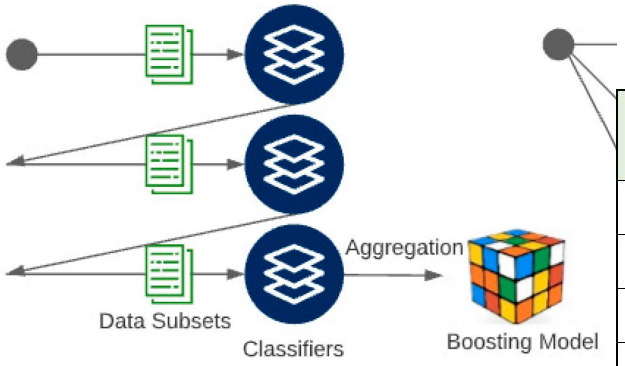
AdaBoost

Gradient Boosting Machines (GBM)

XGBoost

LightGBM

CatBoost

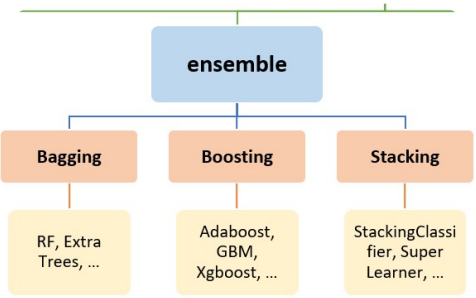


مدل / ویژگی	AdaBoost	GBM	XGBoost	LightGBM	CatBoost
سرعت	متوسط	کند	سریع	بسیار سریع	سریع
دقت	خوب	عالی	عالی	عالی	عالی
حجم داده	متوسط	متوسط	بزرگ	بسیار بزرگ	بزرگ
Overfitting	نسبتاً مقاوم	نیاز به تنظیم	تنظیم‌پذیر	تنظیم‌پذیر	بسیار مقاوم

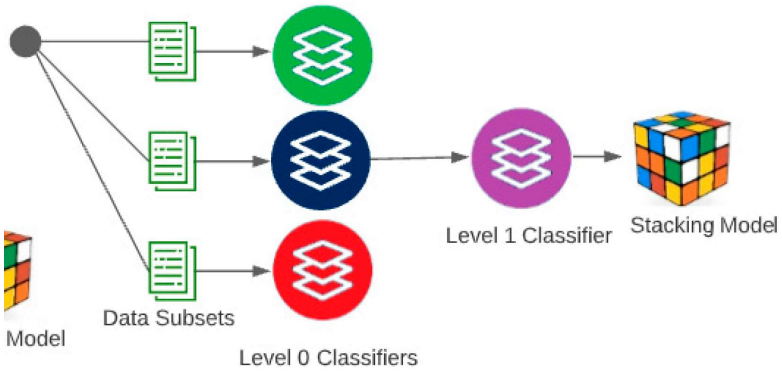
(**B**) Boosting

هدف: کاهش بایاس و خطای کلی با ترکیب سری مدل‌های ضعیف که هر کدام روی نمونه‌های مشکل‌تر تمرکز می‌کنند.

Ensemble Models



- Ensemble Strategies
  - Bagging
  - Boosting
  - Stacking



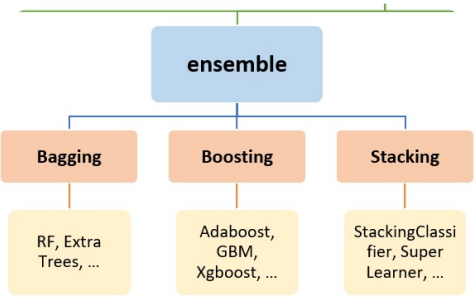
ستکینگ (دو لایه مدل)

- لایه یک:
  - چند مدل پایه **متنوع**، یادگیری مستقل از هم و مستقیما از خود داده‌های دیتاست
    - درخت تصمیم: Random Forest, Extra Trees
    - مدل‌های خطی: Logistic Regression, Ridge/Lasso
    - روش‌های فاصله‌ای: K-NN, SVM
    - شبکه‌های عصبی: MLP ساده
    - مدل‌های بوستینگ: XGBoost, LightGBM, CatBoost
    - مدل‌های بیزی: Naive Bayes
- لایه دو:
  - یک مدل (متامدل/ فرامدل) یادگیری از خروجی مدل‌های لایه قبل (معمولا از لیست زیر)
    - رگرسیون لجستیک: ساده و مؤثر برای طبقه‌بندی
    - XGBoost/LightGBM: برای الگوهای پیچیده‌تر
    - شبکه عصبی ساده: اگر داده کافی باشد
    - رگرسیون خطی: برای مسائل رگرسیون
    - مدل‌های خطی: Logistic Regression, Ridge/Lasso

(C) Stacking

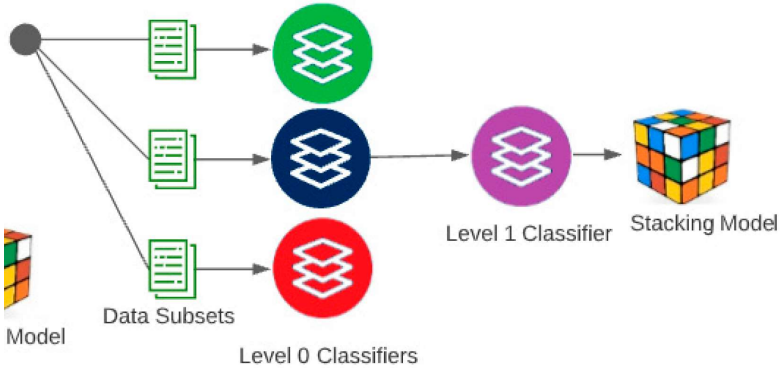
هدف: کاهش همزمان بایاس (با تنوع مدل‌های پایه) و واریانس (با ترکیب فرامدل) برای بهبود قابلیت تعمیم مدل نهایی.

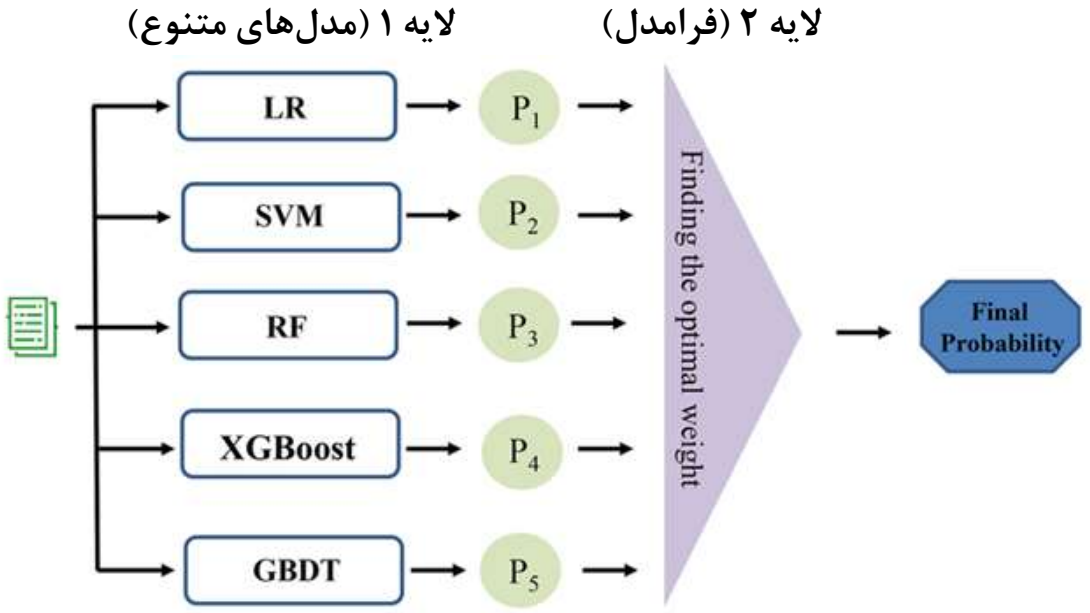
Ensemble Models



- Ensemble Strategies
  - Bagging
  - Boosting
  - Stacking

ستکینگ (دو لایه مدل)

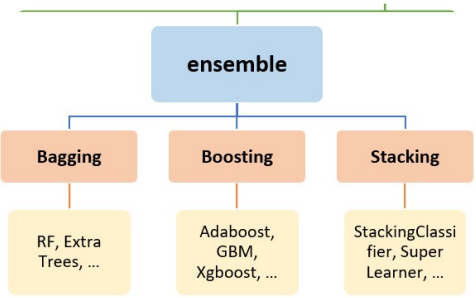




(C) Stacking

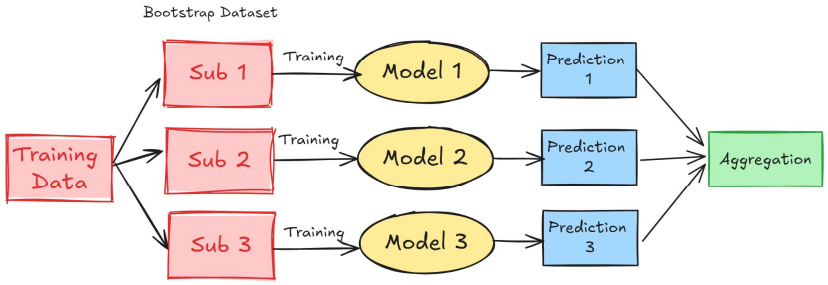
هدف: کاهش همزمان بایاس (با تنوع مدل‌های پایه) و واریانس (با ترکیب فرامدل) برای بهبود قابلیت تعمیم مدل نهایی.

Ensemble Models



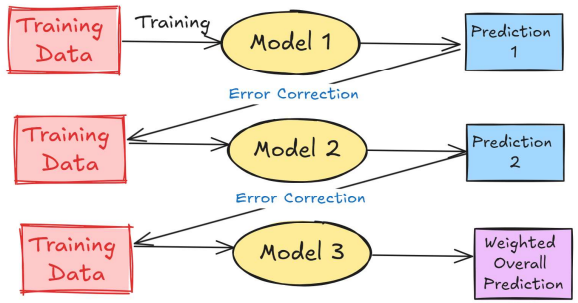
- Ensemble Strategies Bagging, Boosting, Stacking

Bagging (Bootstrap Aggregation)



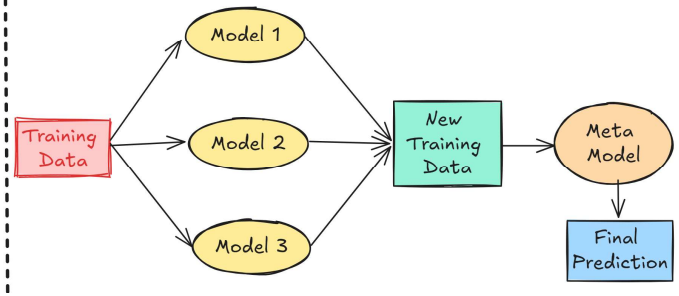
Process	Aggregate multiple models trained on different subsets bootstrap data.
Predictions	Averaging (regression) or Voting (classification).
Example	Random Forest, Bagged Decision Trees.

Boosting



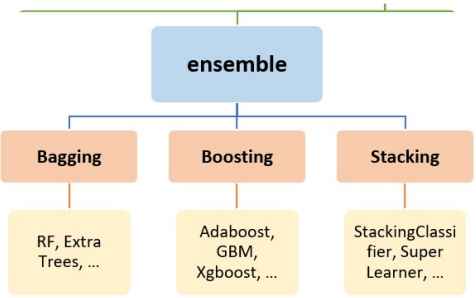
Builds models sequentially and perform error correction based on previous model.
Weighted sum of predictions based on every model.
AdaBoost, Gradient Boosting Machines (GBM), XGBoost.

Stacking

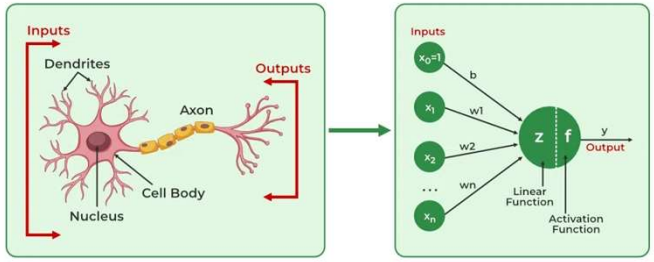


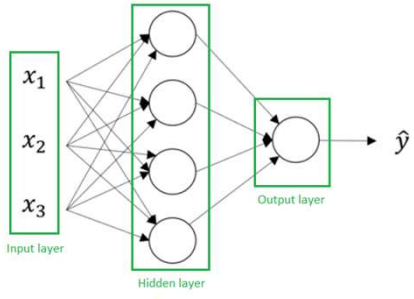
Combines predictions of multiple models using a meta model.
Meta model give final prediction based on the multiple models prediction
Logistic Regression on top of multiple classifiers

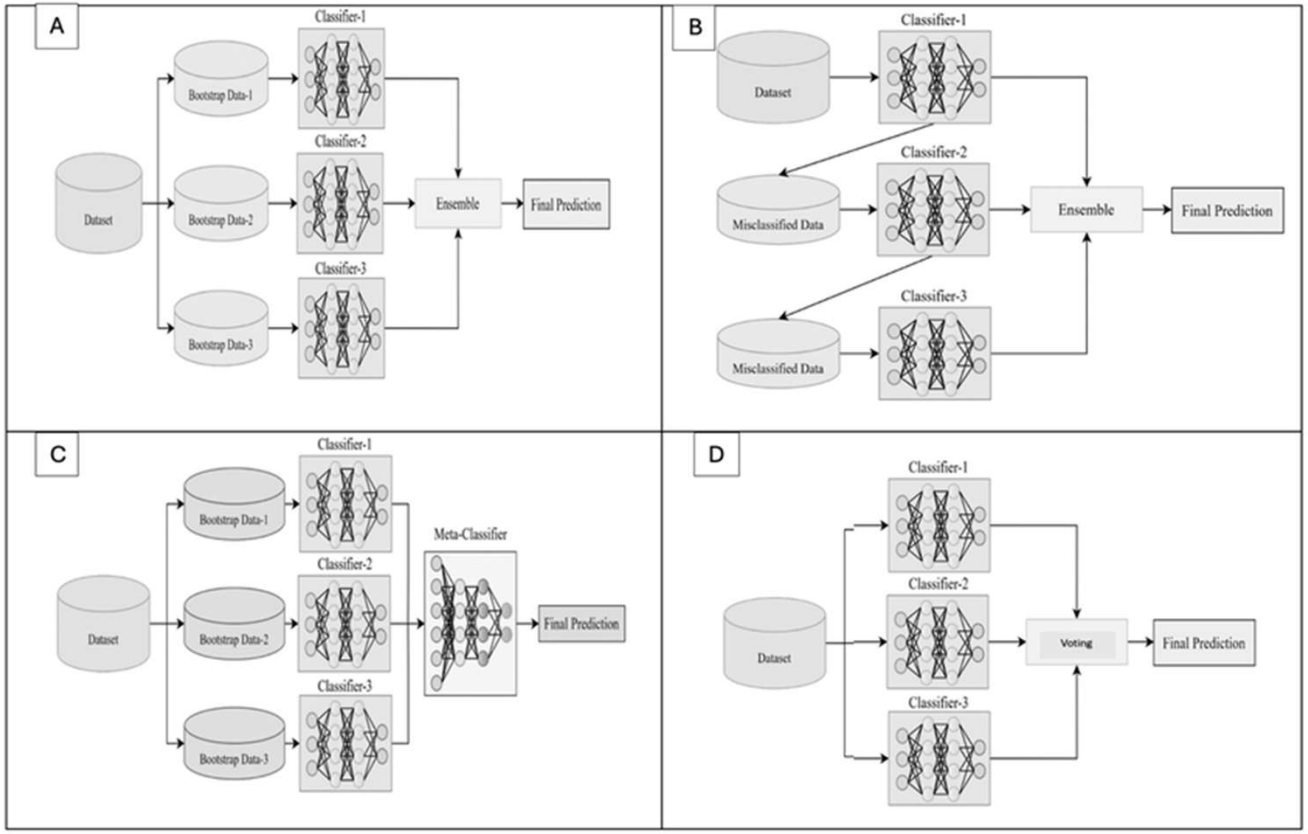
Ensemble Models (NN)



- Ensemble Strategies
  - Bagging
  - Boosting
  - Stacking







Thank You!