

# ML Models 2

Bias & Variance  
Cross Validation  
Ensemble Models

کارگروه هوش مصنوعی شرکت پتروشیمی مهاباد

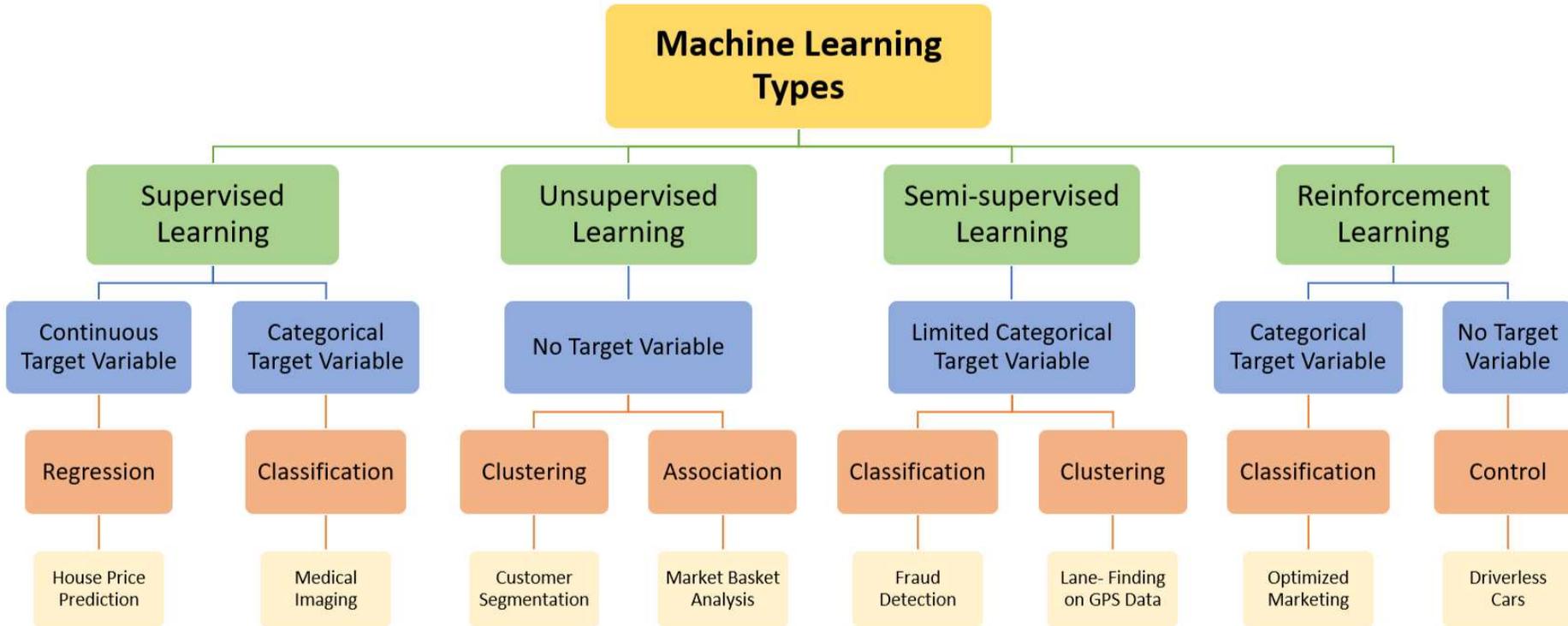
محمد ستوده

زمستان ۱۴۰۴

# ML

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

<b>Artificial Intelligence (AI)</b>
<b>1. Cognitive &amp; Symbolic AI</b>
1.1. Knowledge Representation (KR)
1.2. Automated Reasoning
1.3. Planning & Decision Making
1.4. Cognitive Modeling
<b>2. Machine Learning (ML)</b>
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
<b>3. Data Mining &amp; Predictive Analytics</b>
3.1. Association Rule Mining
3.2. Sequential Pattern Mining
3.3. Temporal / Time-series Analysis
3.4. Predictive Modeling
3.5. Anomaly Detection
3.6. Graph Mining
3.7. Recommender Systems
<b>4. Natural Language Processing (NLP)</b>
4.1. Linguistic Processing
4.2. Semantic Processing
4.3. Pragmatic / Discourse Processing
4.4. Text Generation
4.5. Speech & Audio
4.6. Multimodal NLP
<b>5. Computer Vision (CV)</b>
5.1. Image Processing
5.2. Object Detection & Recognition
5.3. 3D Vision
5.4. Video Understanding
5.5. Multimodal Vision
<b>6. Robotics &amp; Embodied AI</b>
6.1. Perception
6.2. Control & Motion Planning
6.3. Manipulation & Locomotion
6.4. Human-Robot Interaction (HRI)
6.5. Reinforcement Learning in Robotics
<b>7. Multi-Agent Systems (MAS)</b>
7.1. Agent Architectures
7.2. Distributed Coordination
7.3. Communication & Negotiation
7.4. Swarm Intelligence
7.5. Game-Theoretic Modelling
<b>8. Generative AI</b>
8.1. Generative Models
8.2. Text Generation (LLMs)
8.3. Image / Video / Audio Synthesis
8.4. Code Generation
8.5. Multimodal Generation
<b>9. Affective &amp; Social Computing</b>
9.1. Emotion Recognition
9.2. Sentiment Analysis
9.3. Personality Modelling
9.4. Empathic Agents
<b>10. Explainable &amp; Ethical AI</b>
10.1. Explainable AI (XAI)
10.2. Fairness & Bias Mitigation
10.3. AI Governance & Trust
10.4. Privacy-preserving AI
10.5. Causal Inference
<b>11. Applied &amp; Domain-specific AI</b>
11.1. Medical AI
11.2. Industrial AI
11.3. Financial AI
11.4. Social & Recommender Systems
11.5. Education AI
11.6. Environmental / Climate AI



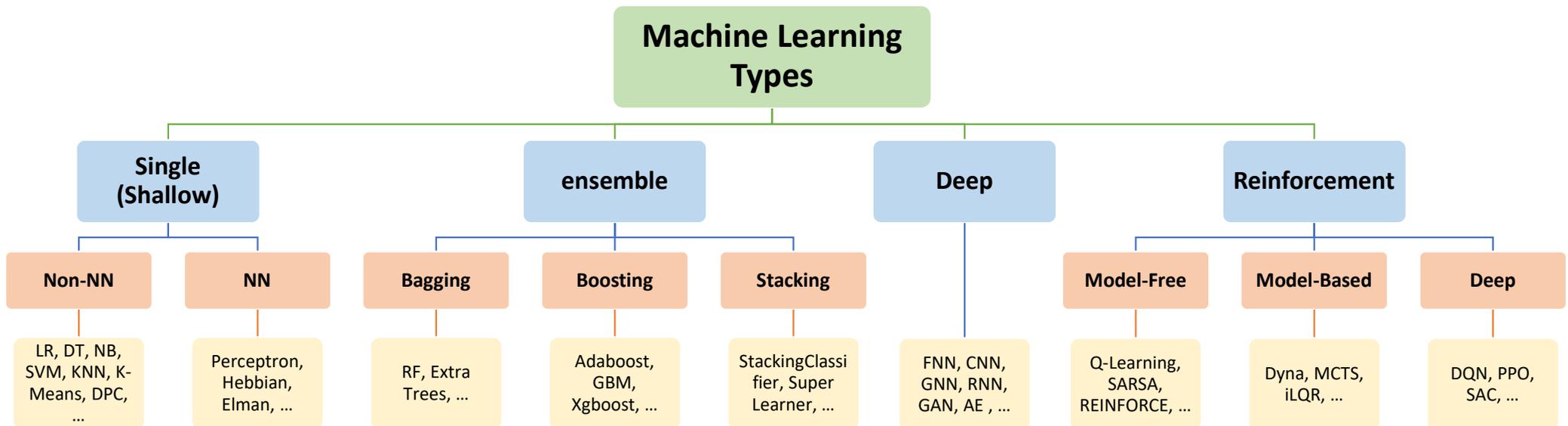
# ML Models Taxonomy

«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)

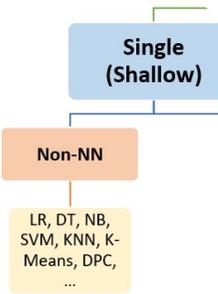
# ML Models Taxonomy

« 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
- NB
- KNN
- SVM

- Liner Regression

- Unsupervised

- K-Means
- DBSCAN

درخت تصمیم:

نایف بیز:

K نزدیکترین همسایگی:

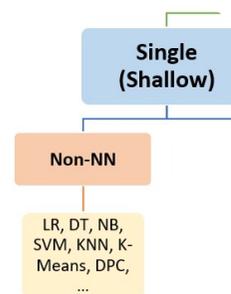
ماشین بردار پشتیبان:

رگرسیون خطی:

**:K-Means**

**:DBSCAN**

# Single Shallow Non-NN



## • Supervised

- DT
- NB
- KNN
- SVM

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

**نایف بیز:** ضرب احتمال های ساده ی 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 نزدیک ترین نمونه های هر کلاس

## • Liner Regression

**رگرسیون خطی:** کدام خط / صفحه کمترین فاصله ی مجموع از همه ی نقاط را دارد؟

## • Unsupervised

- K-Means
- DBSCAN

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

**DBSCAN:** تشخیص خوشه ها را بر اساس تراکم نقاط

# Single Shallow Non-NN

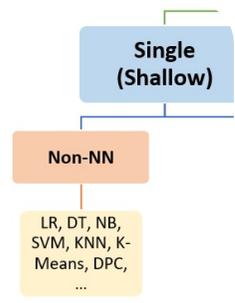
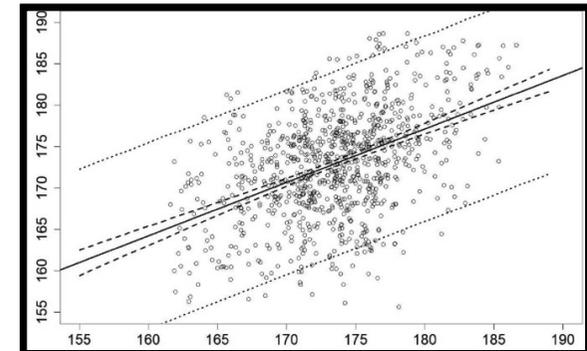
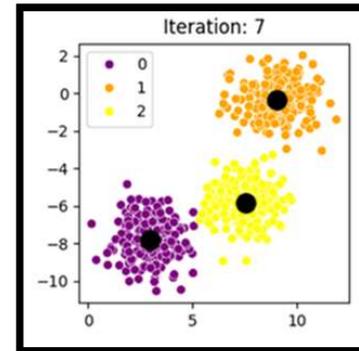
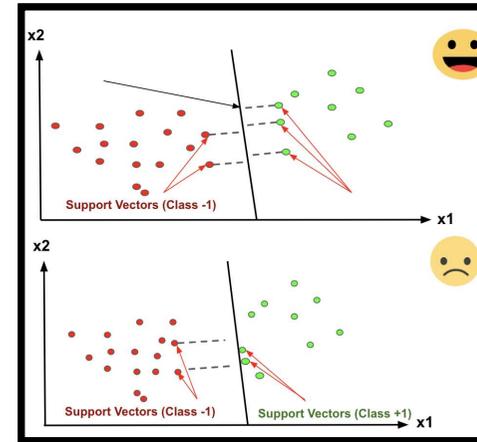
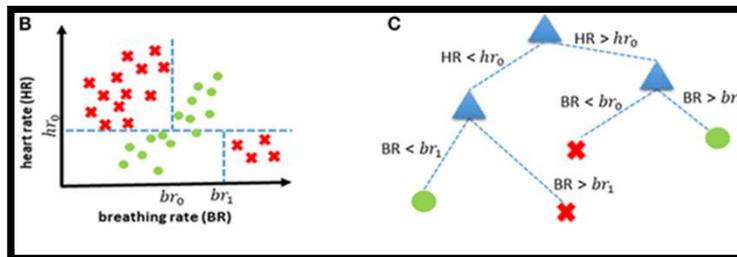
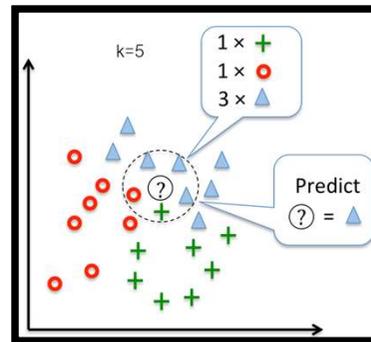
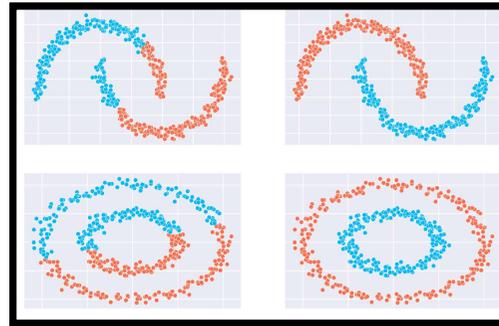
- Supervised

- DT
- NB
- KNN
- SVM

- Liner Regression

- Unsupervised

- K-Means
- DBSCAN



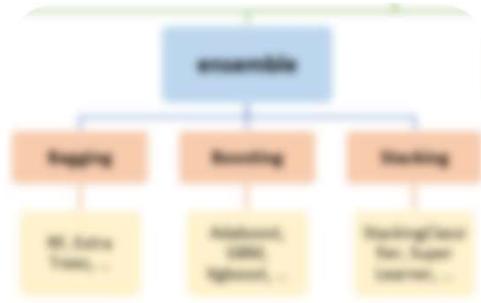
Outlook	Temp	Humidity	Windy	Play?
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	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?
query:	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	Result
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%



**بایاس در داده**

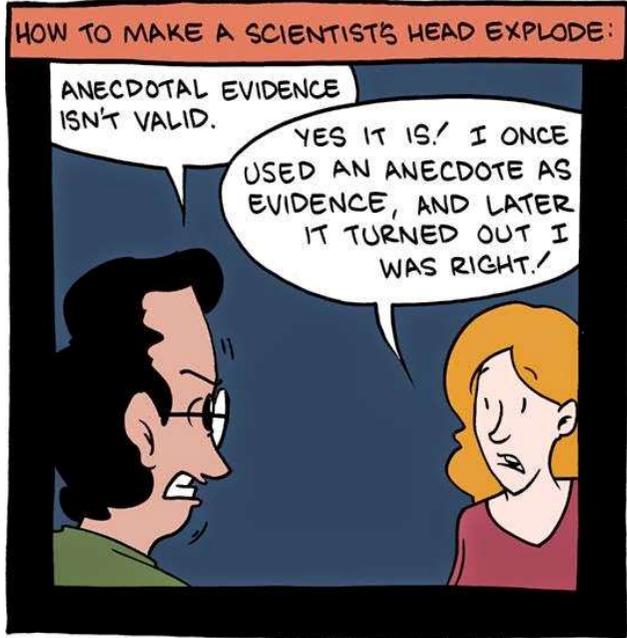
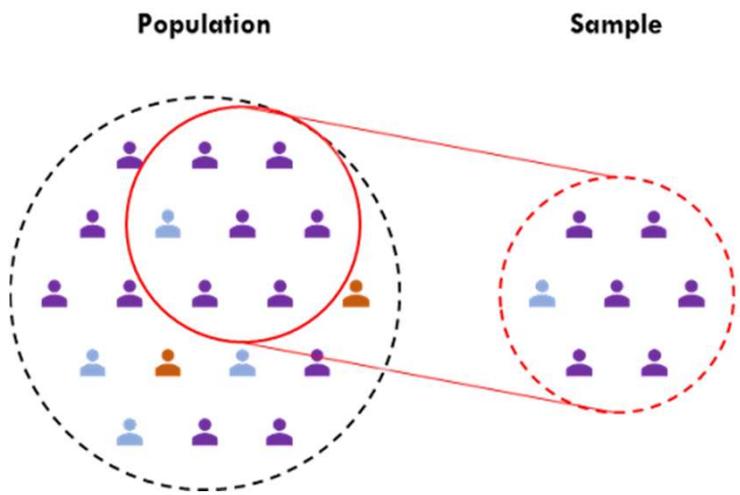
تعمیم جزء به کل!

Hasty Generalization Fallecy

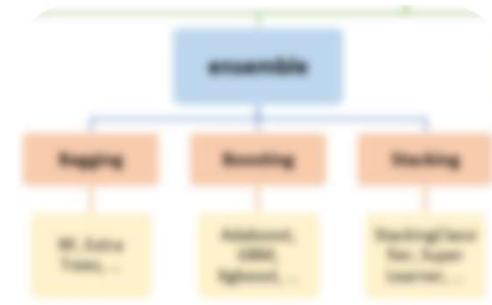
# Bias & Variance



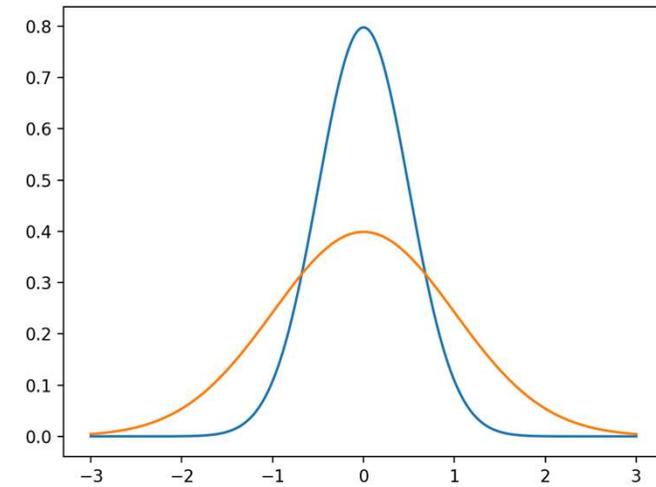
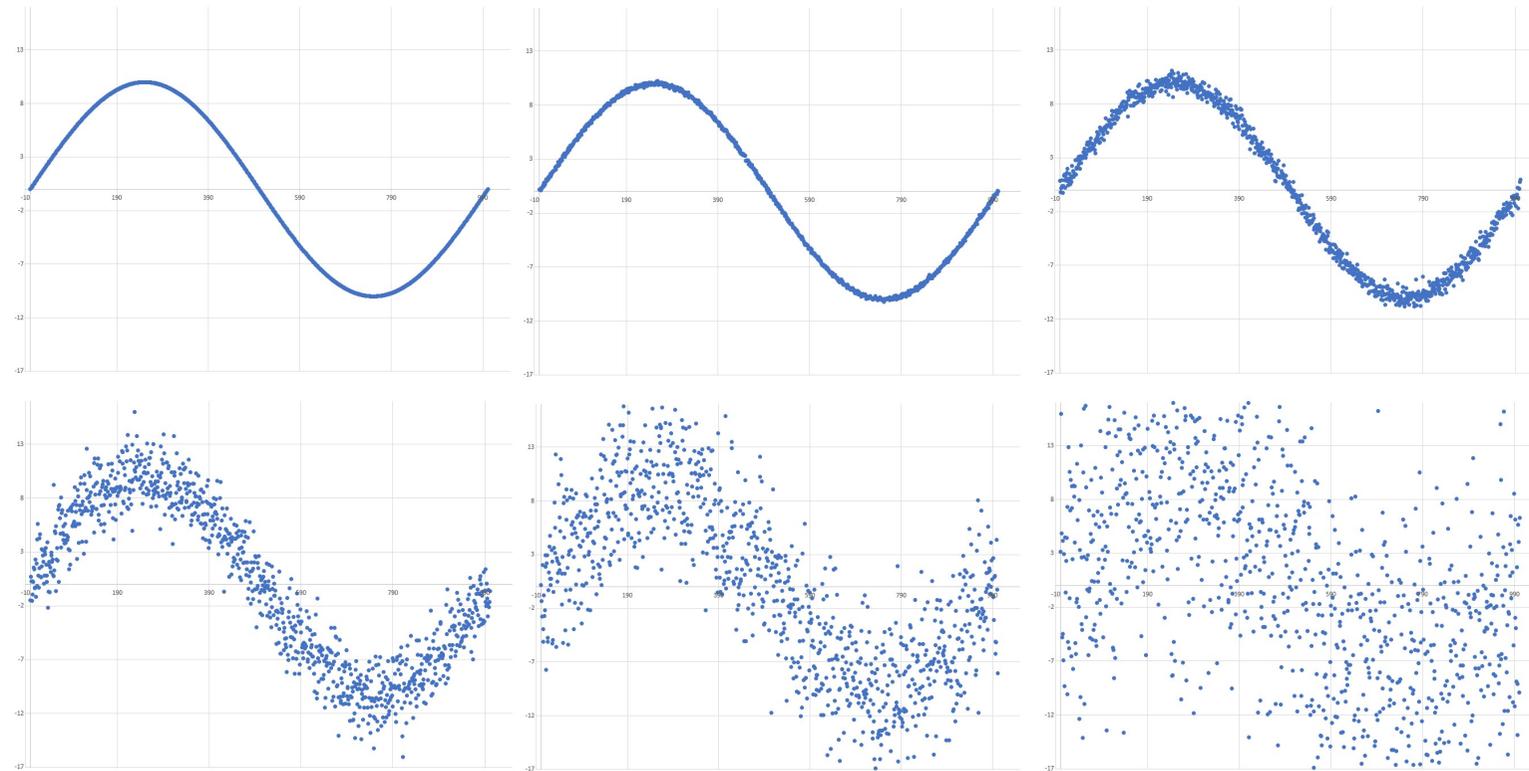
وقتی مشت نمونه خروار نیست!



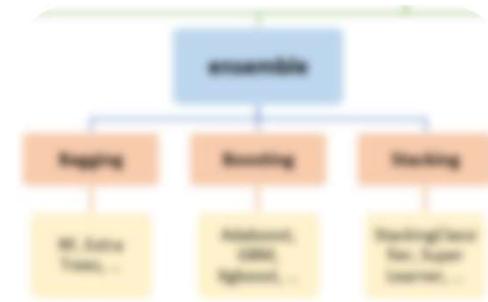
# Bias & Variance



واریانس در داده

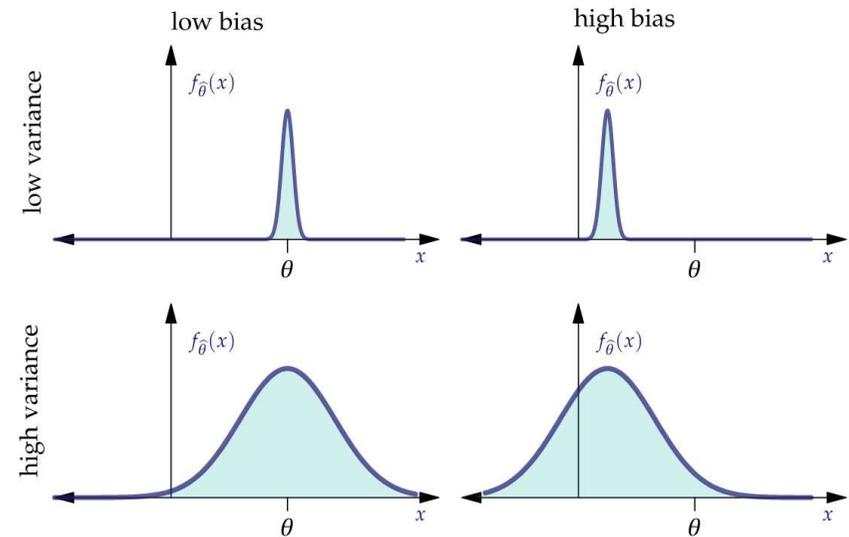
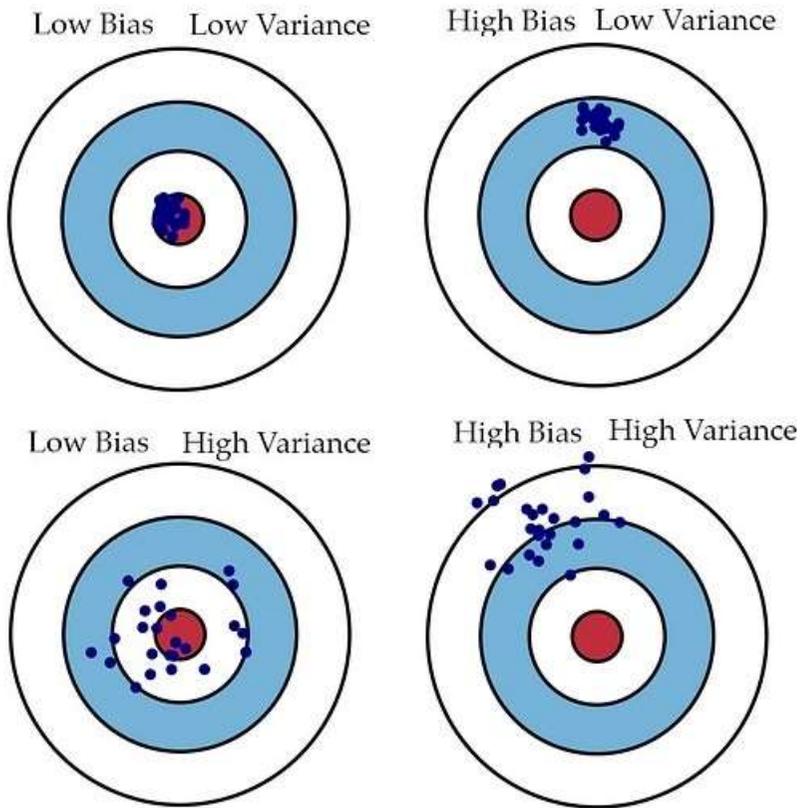


# Bias & Variance

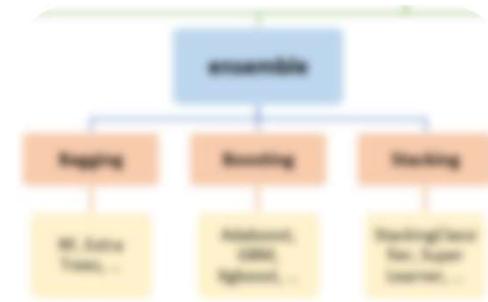


## بایاس و واریانس در مدل

پراکندگی و دقت خروجی  
مدل نسبت به واقعیت



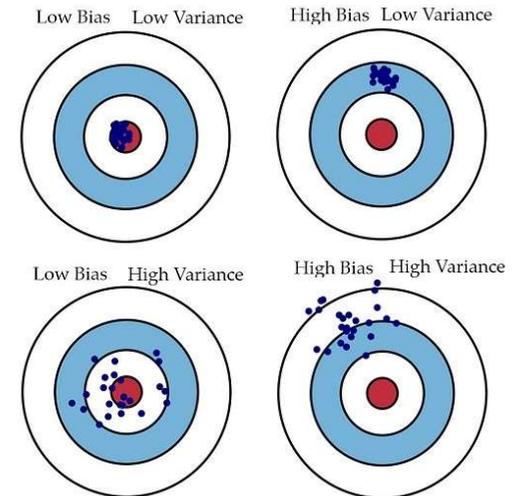
# Bias & Variance



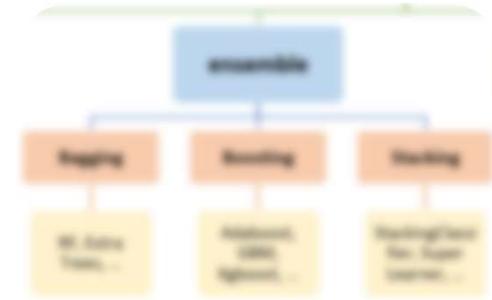
## بایاس و واریانس در مدل

پراکندگی و دقت خروجی  
مدل نسبت به واقعیت

	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none"> <li>• High training error</li> <li>• Training error close to test error</li> <li>• High bias</li> </ul>	<ul style="list-style-type: none"> <li>• Training error slightly lower than test error</li> </ul>	<ul style="list-style-type: none"> <li>• Very low training error</li> <li>• Training error much lower than test error</li> <li>• High variance</li> </ul>
Regression illustration			
Classification illustration			



# Bias & Variance



درک سطحی

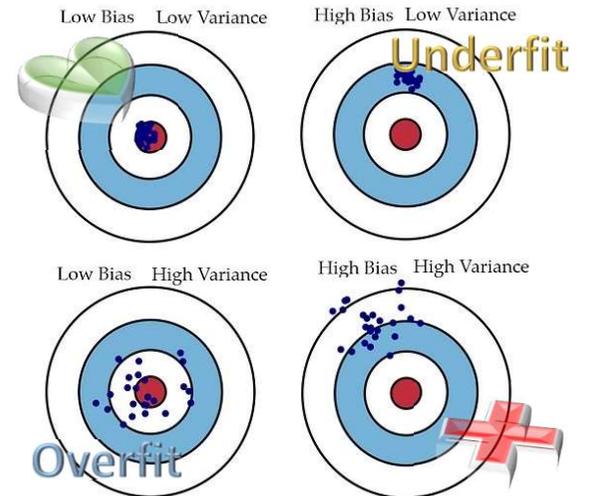
فهمیده

حفظ کرده

بایاس و واریانس در مدل

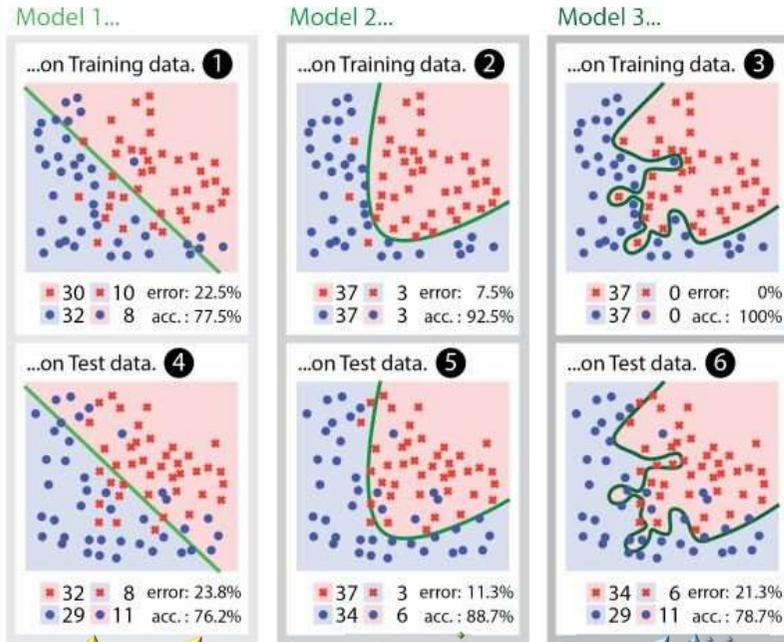
پراکندگی و دقت خروجی  
مدل نسبت به واقعیت

	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none"> <li>High training error</li> <li>Training error close to test error</li> <li>High bias</li> </ul>	<ul style="list-style-type: none"> <li>Training error slightly lower than test error</li> </ul>	<ul style="list-style-type: none"> <li>Very low training error</li> <li>Training error much lower than test error</li> <li>High variance</li> </ul>
Regression illustration			
Classification illustration			

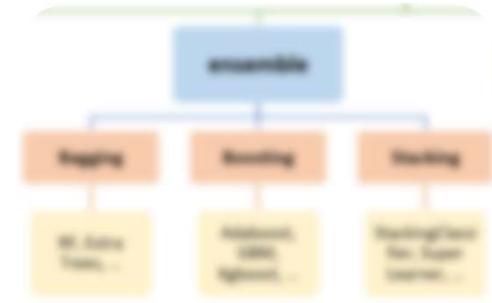
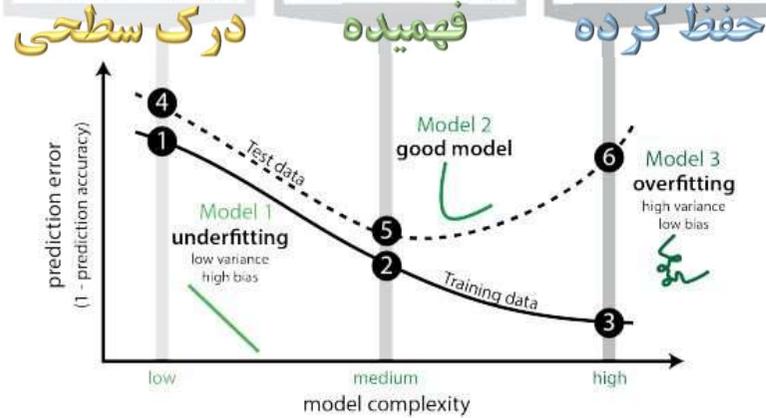


## پیچیدگی مدل

آموزش

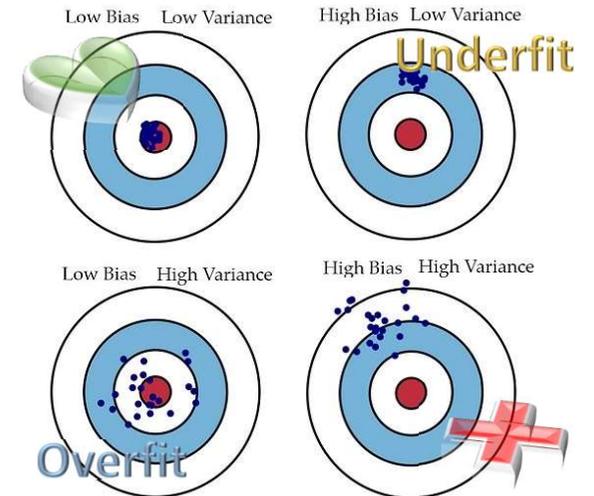


تست



## بایاس و واریانس در مدل

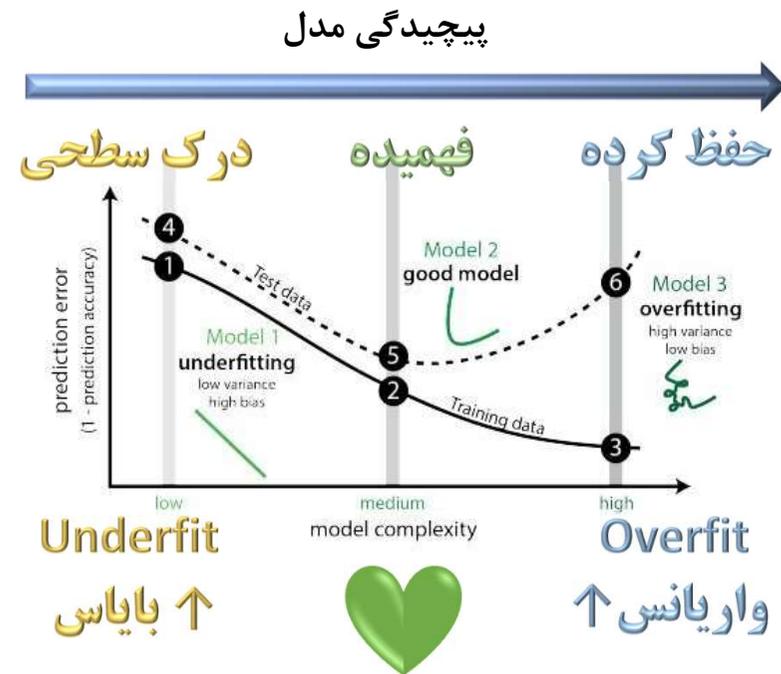
پراکندگی و دقت خروجی  
مدل نسبت به واقعیت



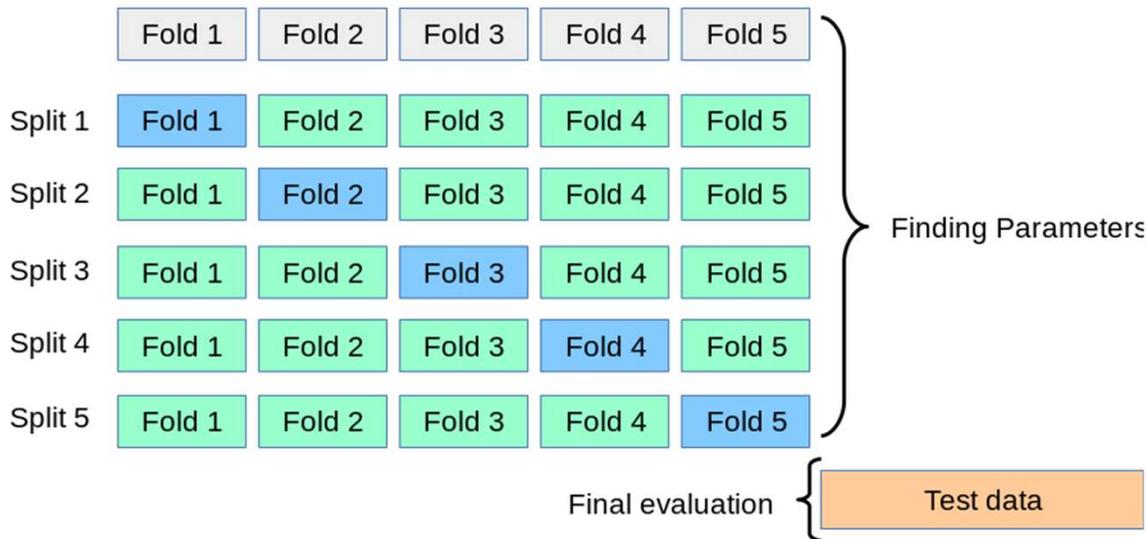
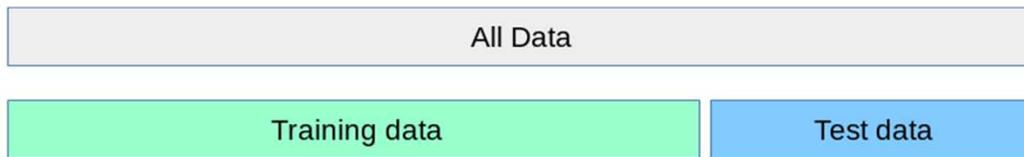
# Bias & Variance

وارianس	بایاس	روش
کمک به مدیریت	کمک به مدیریت	Cross-Validation
↓↓↓	→	Bagging
↓ (اغلب)	↓↓	Boosting
↓	↓	Stacking
↑	↓	افزایش پیچیدگی مدل
↓	↑	کاهش پیچیدگی مدل
↓	↑ (کمی)	Regularization (L1/L2)
↓	↑ (ممکن است)	Feature Selection
↓	↓	Feature Engineering خوب
↓	→	افزایش داده
↓	→	حذف نویز
↓	↑ / →	Early Stopping
↓	↑ / →	Dropout

روش‌های مدیریت  
بایاس و واریانس



# Cross Validation



CV مدل را قوی تر نمی کند؛ بلکه تخمینِ بهتر و انتخابِ **هایپر پارامتر** مناسب را ممکن می سازد که منجر به مدلی با بایاس / واریانس مطلوب تر می شود.

انواع:

## k-fold CV

Stratified k-fold CV

Leave-One-Out CV (LOOCV)

Repeated k-fold CV

Hold-out (train/validation/test split)

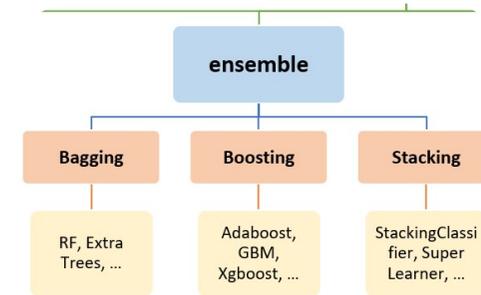
Time-series (forward chaining/ rolling-origin) CV

Nested CV

Monte Carlo (Random) CV / ShuffleSplit

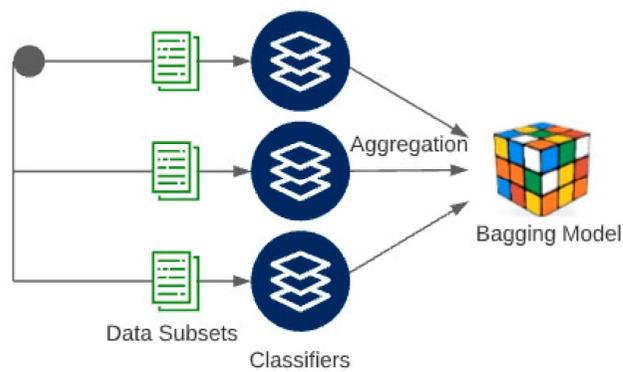
Group k-fold CV

# Ensemble Models

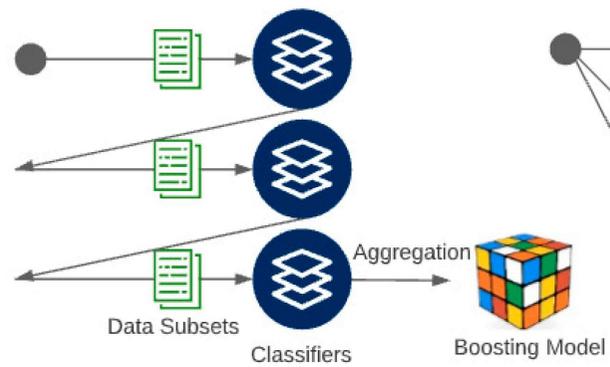


- Ensemble Strategies

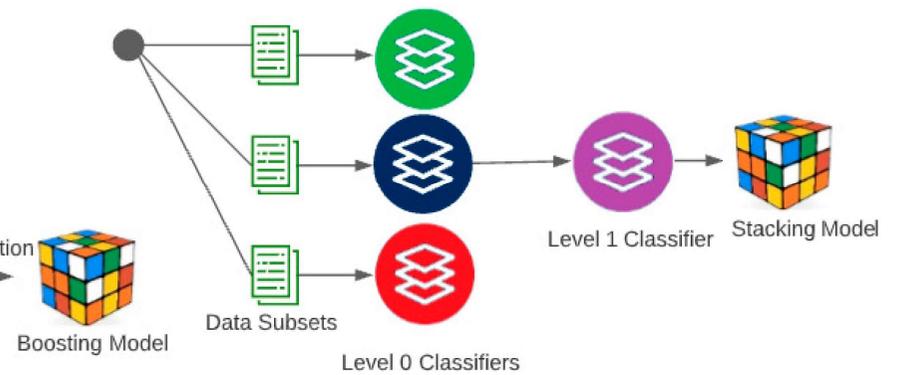
- Bagging
- Boosting
- Stacking



(A) Bagging

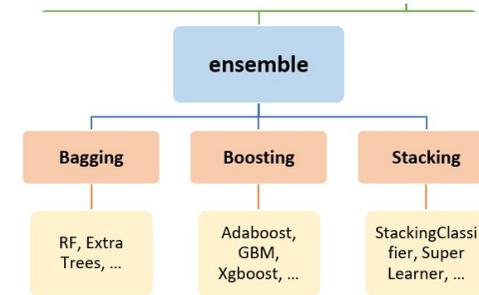


(B) Boosting



(C) Stacking

# Ensemble Models



## Ensemble Strategies

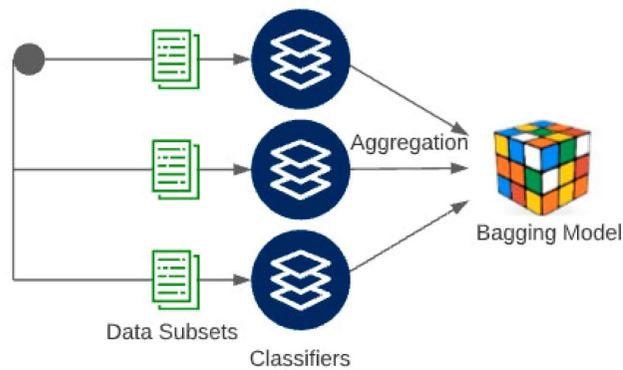
- Bagging (bootstrap aggregating)
- Boosting
- Stacking

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

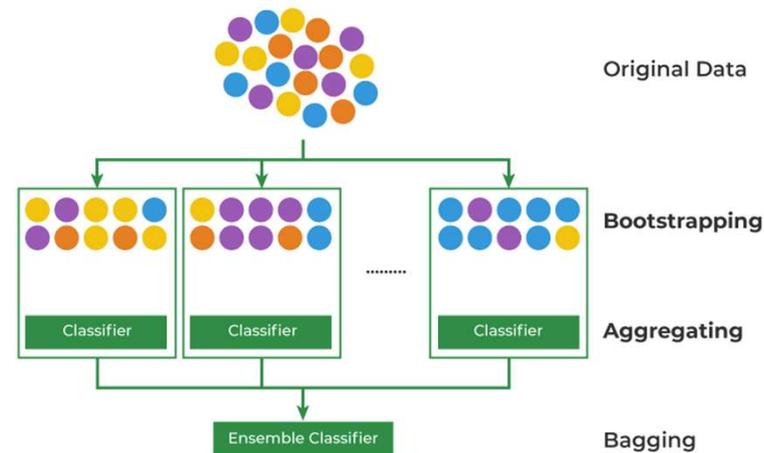
آموزش موازی چند مدل مشابه

داده‌ها با نمونه برداری تصادفی و جایگذاری مجدد

خروجی با رای گیری



(A) Bagging

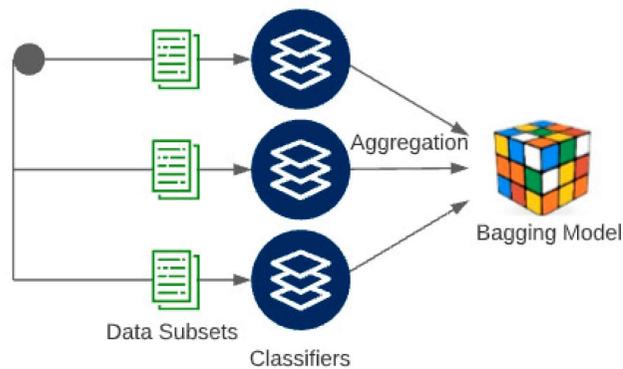


هدف: کاهش واریانس و افزایش پایداری مدل بدون افزایش بایاس قابل توجه.

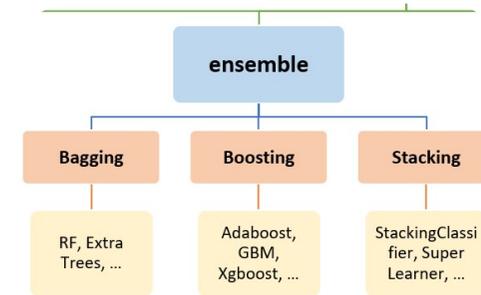
# Ensemble Models

- Ensemble Strategies

- Bagging (bootstrap aggregating)
- Boosting
- Stacking



(A) Bagging

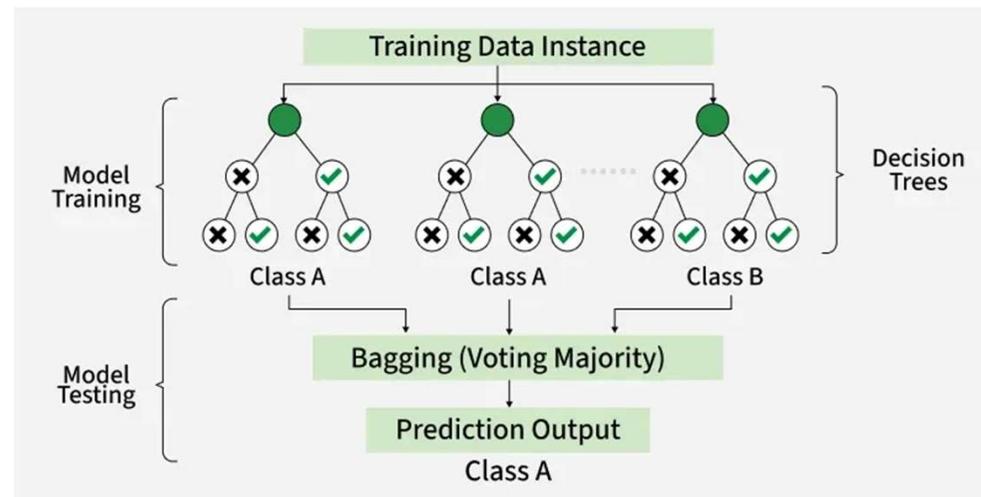


مدل‌های معروف / معمول بگینگ

Random Forest

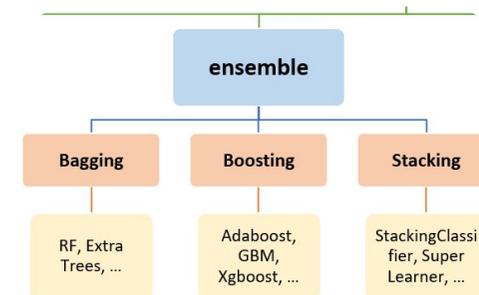
Bagging with k-NN

Bagging with SVM



هدف: کاهش واریانس و افزایش پایداری مدل بدون افزایش بایاس قابل توجه.

# Ensemble Models



## بوستینگ

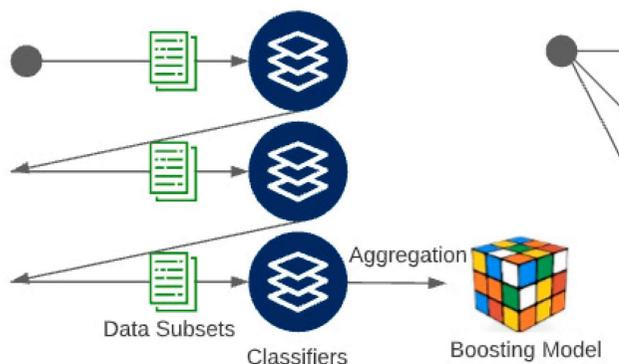
آموزش سری مدل‌ها. تمرکز بر نمونه‌های سخت.

- 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.

## • Ensemble Strategies

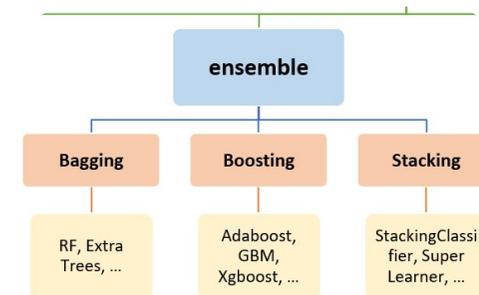
- Bagging
- Boosting
- Stacking



## (B) Boosting

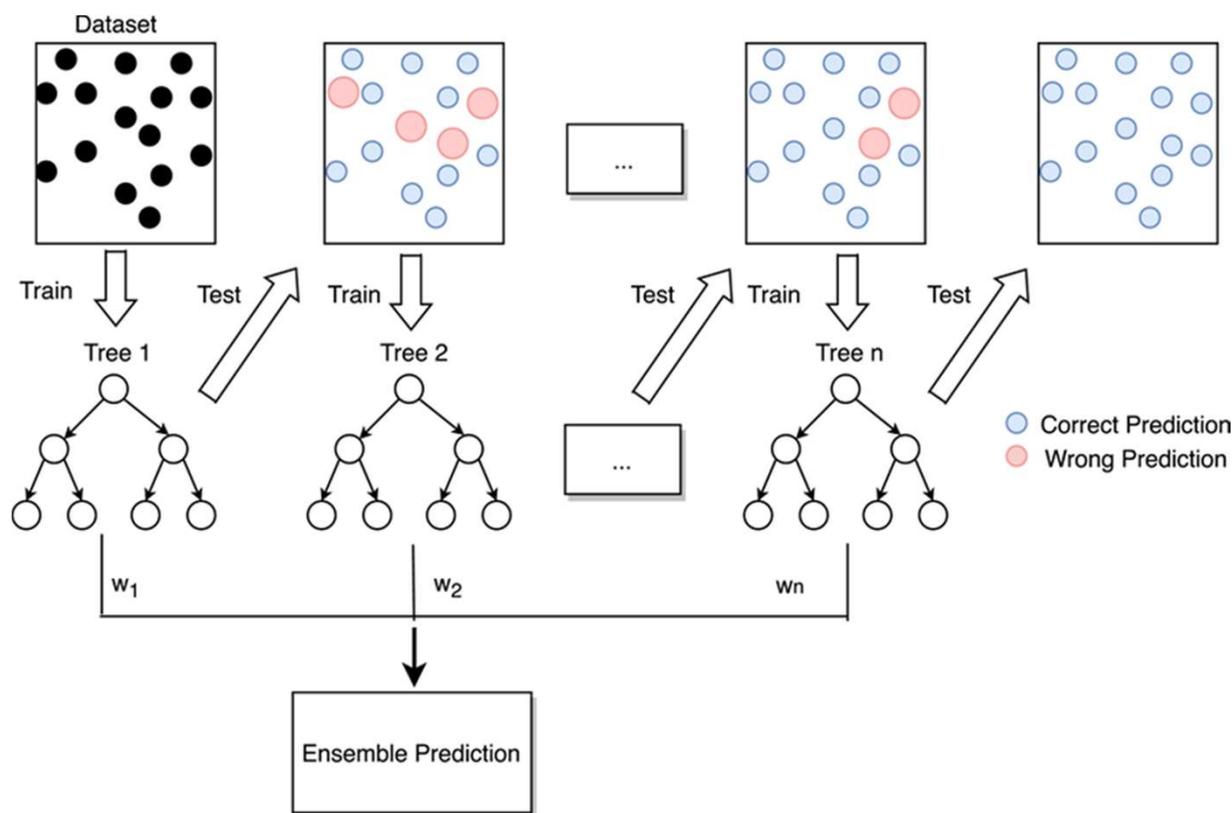
هدف: کاهش بایاس و خطای کلی با ترکیب سری مدل‌های ضعیف که هر کدام روی نمونه‌های مشکل‌تر تمرکز می‌کنند.

# Ensemble Models



## بوستینگ

آموزش سری مدل‌ها. تمرکز بر نمونه‌های سخت.

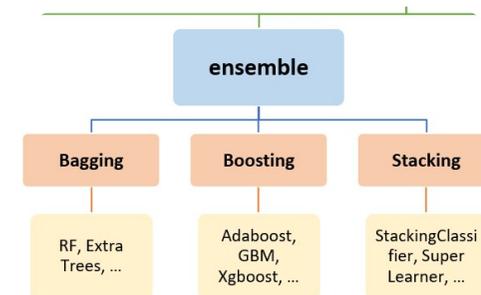


- آغاز با مدل پایه‌ی ساده (Weak learner)، معمولاً با وزن‌های برابر برای نمونه‌ها.
- در هر تکرار:
  - آموزش مدل پایه روی داده‌های آموزش با توجه به وزن‌های نمونه‌ها یا خطاهای قبلی.
  - افزایش وزن نمونه‌های دشوار.
  - ترکیب خروجی مدل جدید با مدل‌های قبلی.
  - وزن‌دهی مدل‌ها بر اساس دقت.
- ادامه تا رسیدن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

### (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

## مدل های معروف / معمول بوستینگ

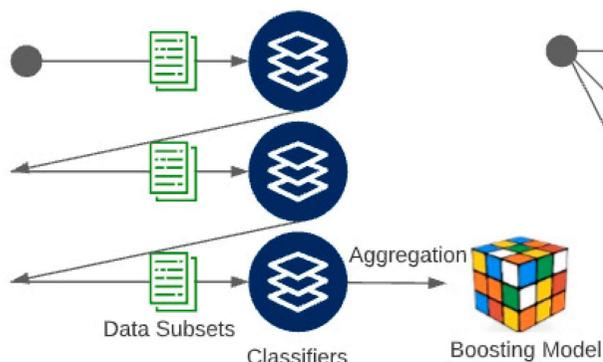
AdaBoost

Gradient Boosting Machines (GBM)

XGBoost

LightGBM

CatBoost

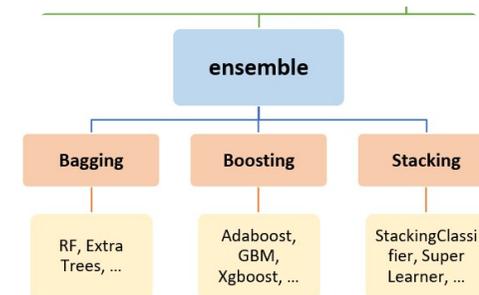


CatBoost	LightGBM	XGBoost	GBM	AdaBoost	مدل
سريع	بسیار سریع	سريع	کند	متوسط	ویژگی
عالی	عالی	عالی	عالی	خوب	سرعت
بزرگ	بسیار بزرگ	بزرگ	متوسط	متوسط	دقت
بسیار مقاوم	تنظیم پذیر	تنظیم پذیر	نیاز به تنظیم	نسبتاً مقاوم	حجم داده
					Overfitting

### (B) Boosting

هدف: کاهش بایاس و خطای کلی با ترکیب سری مدل های ضعیف که هر کدام روی نمونه های مشکل تر تمرکز می کنند.

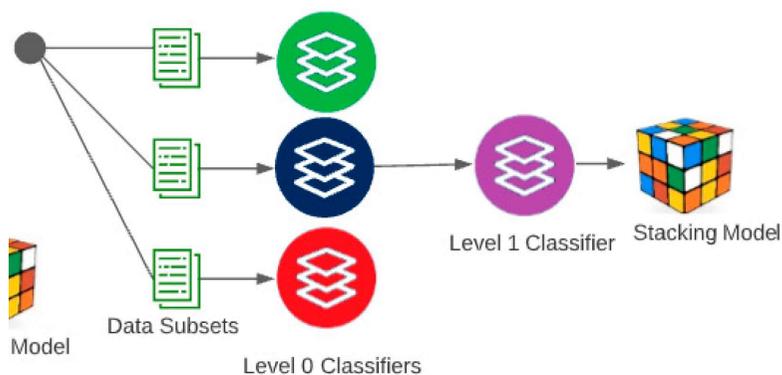
# Ensemble Models



## ستکینگ (دو لایه مدل)

### Ensemble Strategies

- Bagging
- Boosting
- **Stacking**



#### • لایه یک:

- چند مدل پایه متنوع، یادگیری مستقل از هم و مستقیماً 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

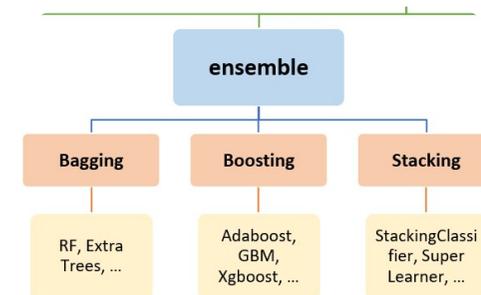
#### • لایه دو:

- یک مدل (متامدل / فرامدل) یادگیری از خروجی مدل‌های لایه قبل (معمولاً 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.

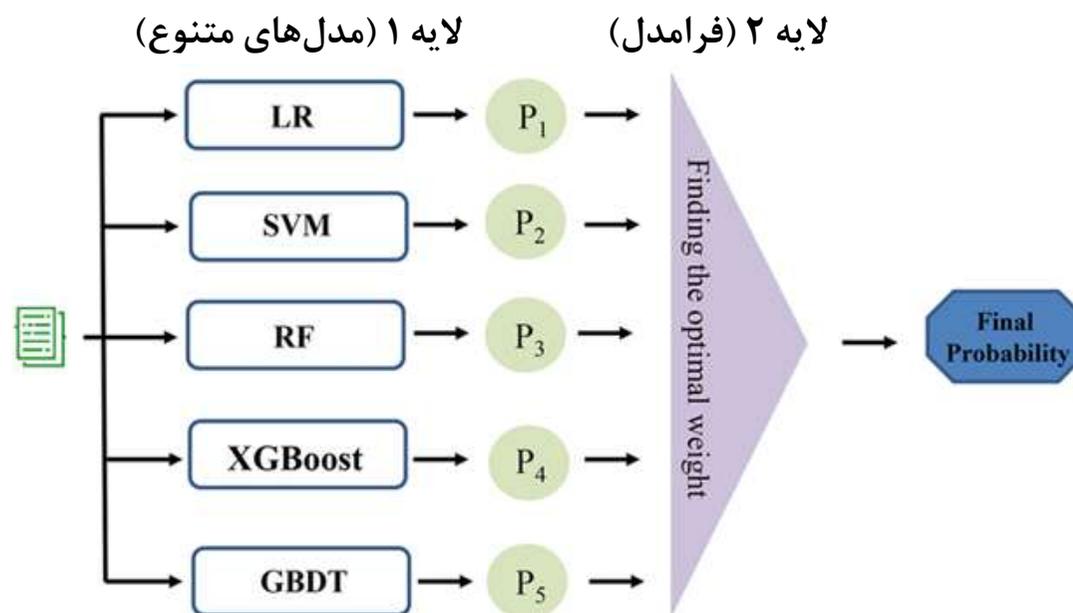
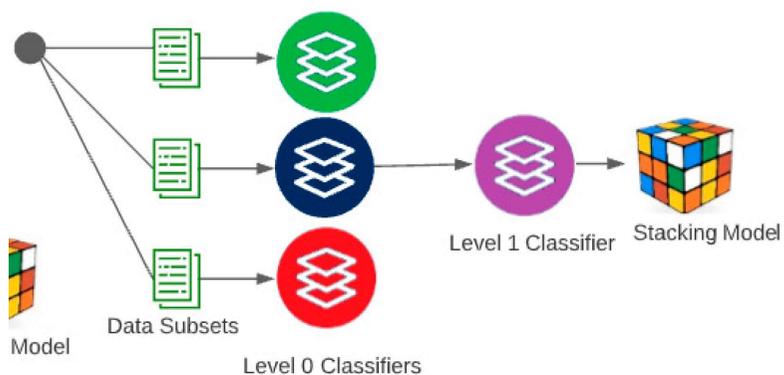
# Ensemble Models



ستکینگ (دو لایه مدل)

## • Ensemble Strategies

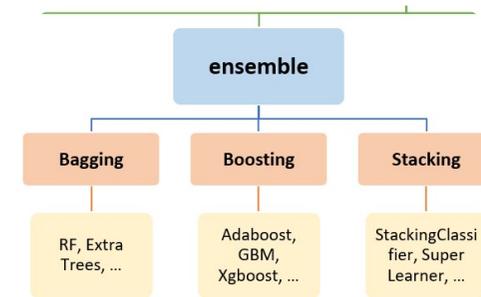
- Bagging
- Boosting
- Stacking



### (C) Stacking

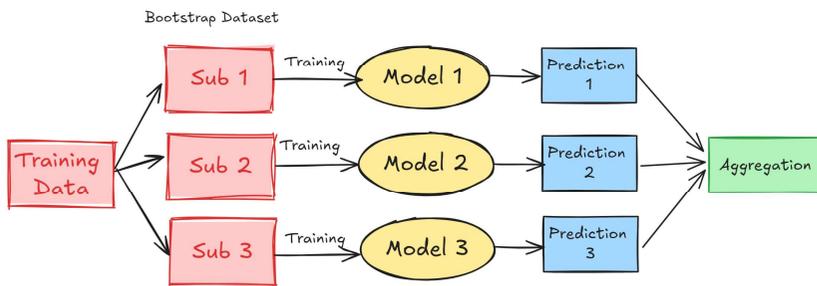
هدف: کاهش همزمان بایاس (با تنوع مدل های پایه) و واریانس (با ترکیب فرامدل) برای بهبود قابلیت تعمیم مدل نهایی.

# Ensemble Models



- Ensemble Strategies Bagging, Boosting, Stacking

## Bagging (Bootstrap Aggregation)

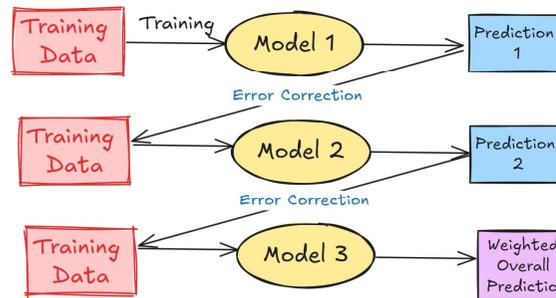


**Process** Aggregate multiple models trained on different subsets bootstrap data.

**Predictions** Averaging (regression) or Voting (classification).

**Example** Random Forest, Bagged Decision Trees.

## Boosting

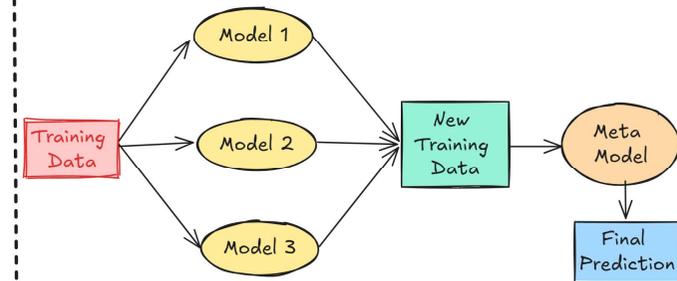


Builds models sequentially and perform error correction based on previous model.

Weighted sum of predictions based on every model.

AdaBoost, Gradient Boosting Machines (GBM), XGBoost.

## Stacking

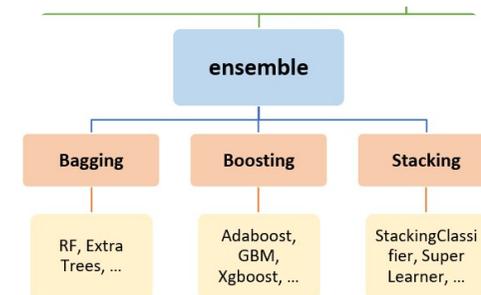


Combines predictions of multiple models using a meta model.

Meta model give final prediction based on the multiple models prediction

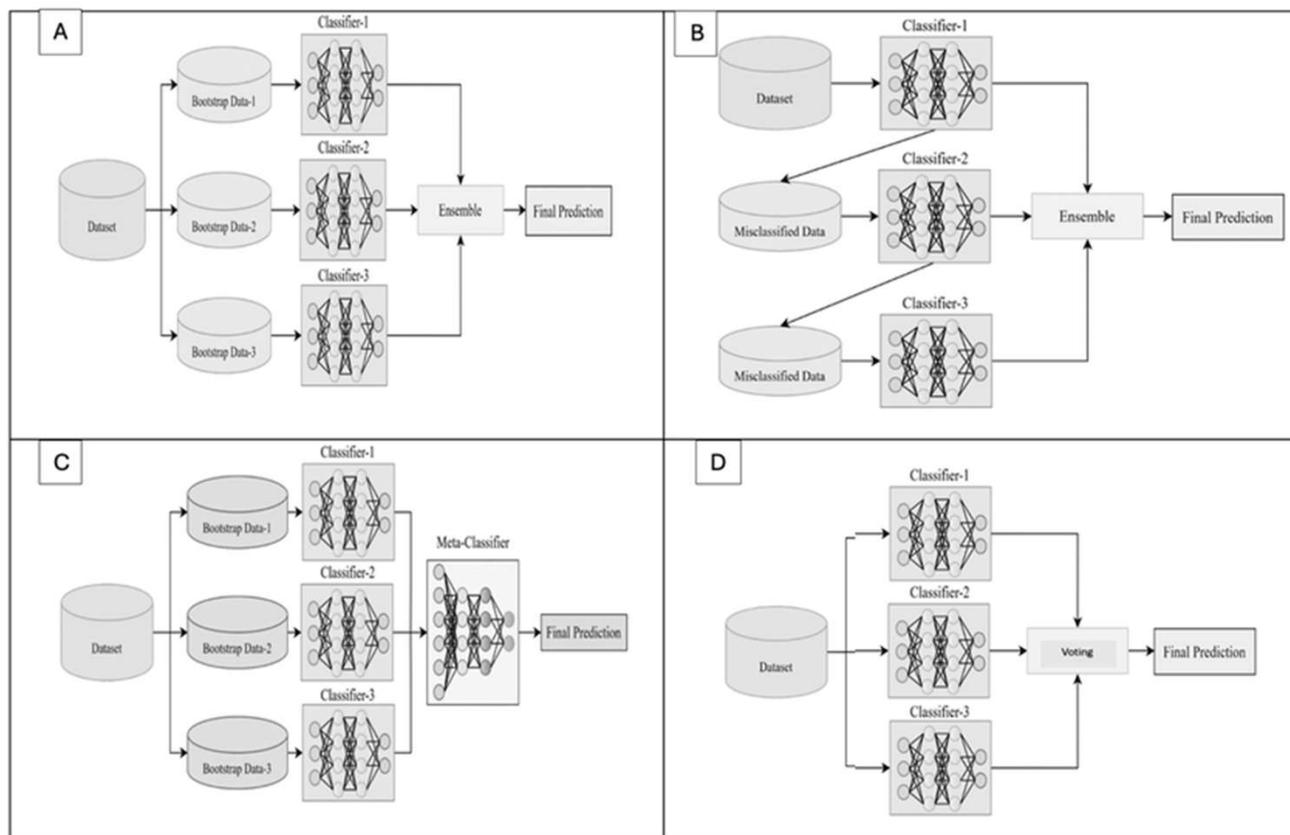
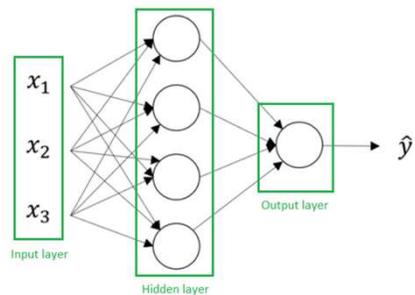
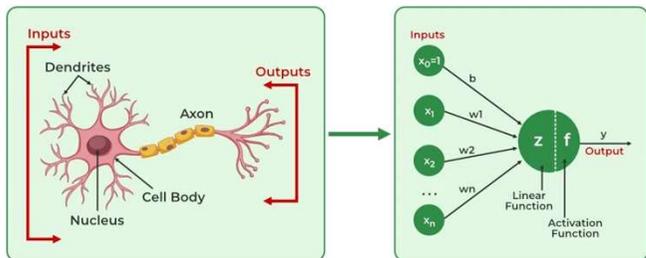
Logistic Regression on top of multiple classifiers

# Ensemble Models (NN)



## Ensemble Strategies

- Bagging
- Boosting
- Stacking



Thank You!

