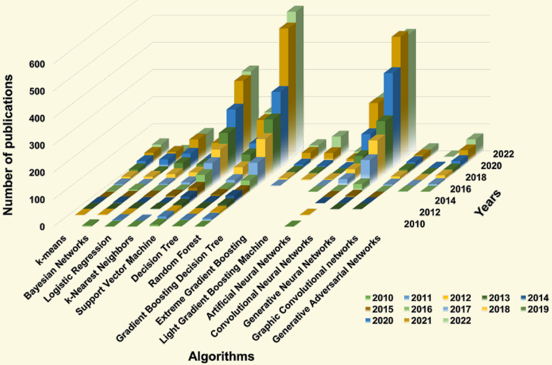
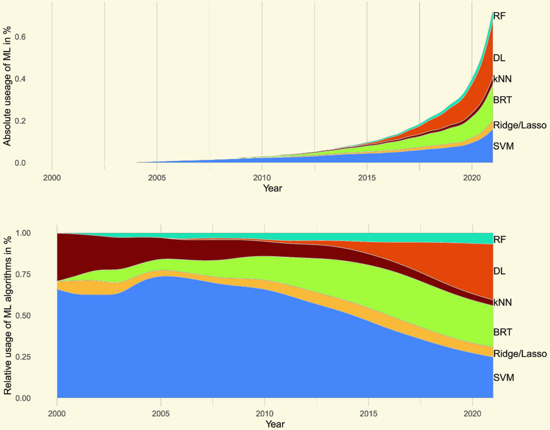
ecoML Workshop:

Machine Learning For Ecological Research

Jeremy Forsythe
& Jazlynn Hall

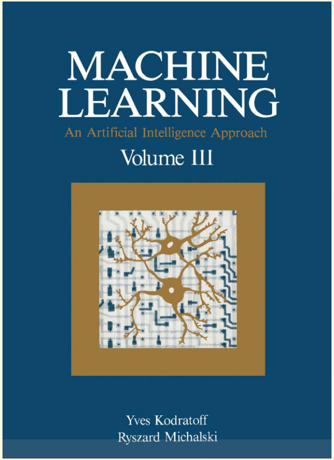


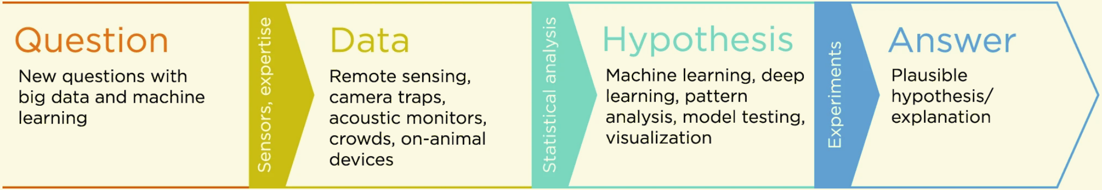


Cui et. al. (2023). Advances and applications of machine learning and deep learning in environmental ecology. Environmental Pollution Vol 335. Pichler, M., & Hartig, F. (2023). Machine learning and deep learning - A review for ecologists. Methods in Ecology and Evolution, 14, 994–1016.

Widely Accepted Definition:

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.





Tuia et. al. (2022). Perspectives in machine learning for wildlife conservation. Nat Commun 13, 792.

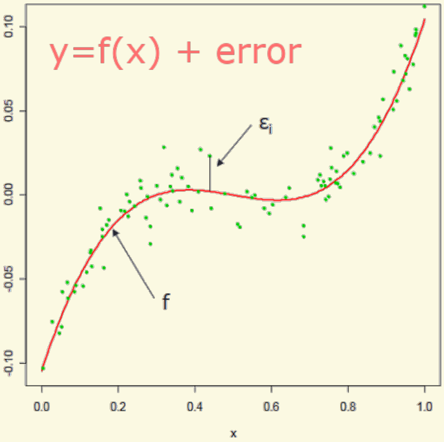
Example: One of the most common tasks we have as ecologists. Regression.

Given a set of n observations (instances or samples), **estimate** the relationship between i independent variables (predictors) and a dependent variable (target, response).

Given $x_i = (x_{i1}, x_{i2}, ..., x_in)$

and $y = y_1, y_2, ..., y_n$:

What is $y = f(x) + $?



Classical Statistics vs ML : Core Philosophy

Classical Statistics

- A subfield of mathematics.
- **Primary Goal:** Inference. Seeks the underlying relationship between output and input.
- **Benefits:** Simple, Fast, Requires Less Training Data
- **Limitations:** Constrained By Model Form, Model May Not Represent True Relationship, Based on Probabilities (p value)

Machine Learning

- A subfield of computer and data sciences.
- **Primary Goal:** Prediction. Seeks the best fit to the training data without a specified model form.
- **Benefits:** Flexible, Potentially Higher Performance
- **Limitations:** Requires More Data, Slower, Overfitting. Less Intrepretable?

Both fields use data to uncover underlying patterns.

Classical Statistics

- **Model-driven:** Start with an assumed statistical model (e.g., Linear Regression, ANOVA).
- The data is used to estimate the parameters of this pre-specified model.
- Heavy emphasis on diagnostics to check if model assumptions are met.

Machine Learning

- **Data-driven:** The algorithm learns the model structure directly from the data.
- Fewer upfront assumptions about the data distribution or relationships.
- Models are often more complex and flexible (e.g., Random Forests, Neural Networks).

Classical Statistics vs ML : Model Evaluation

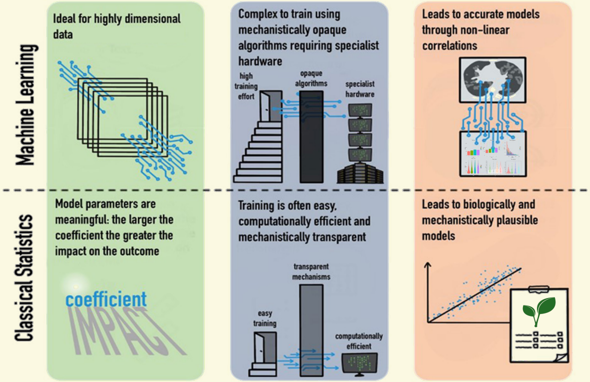
Classical Statistics

- **Goodness-of-Fit:** How well does the model explain the data it was trained on?
- **Metrics:**
  - p-values
  - Confidence Intervals
  - (Rˆ2)

Machine Learning

- **Predictive Power:** How well does the model perform on new, unseen data?
- **Metrics:**
  - Cross-Validation Accuracy
  - Confusion Matrix
  - Area Under the Curve (AUC)
  - Mean Squared Error (MSE)

Classical Statistics vs ML : A Symbiotic Relationship



Al-Hindawi et. al. 2021. A Pro-con Debate for Machine Learning vs. Traditional Statistics.

- Classical statistics and machine learning are not mutually exclusive; they are converging fields.
- Machine learning provides powerful predictive tools, especially for large and complex datasets.
- Newer advances have broken apart the "black box" and ML can now be used for inference.

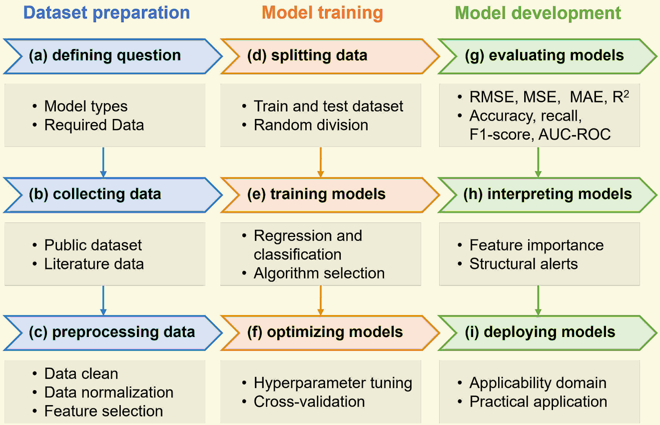
Back To Barney & Moe

"All models are wrong, but some are useful." - George Box

Widely Accepted Definition:

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Cui et. al. (2023). Advances and applications of machine learning and deep learning in environmental ecology. Environmental Pollution Vol 335.



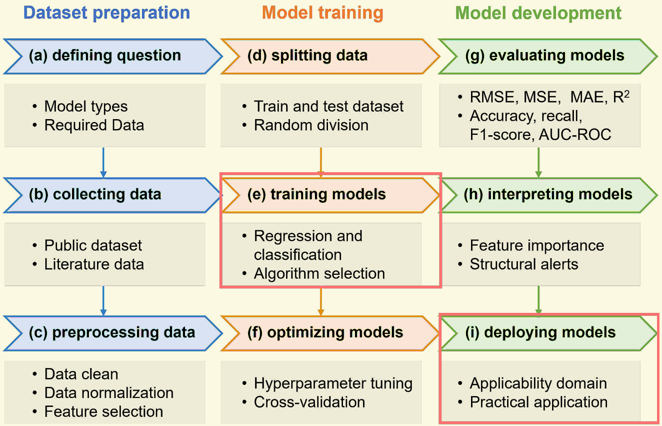
Machine Learning Basics : Task

Task (T):

Classification: Which category does x_i belong to?

Regression: Given predictor(s), estimate a corresponding numerical target variable.

Anomaly Detection: Given a set of observations, flag the unusual ones.



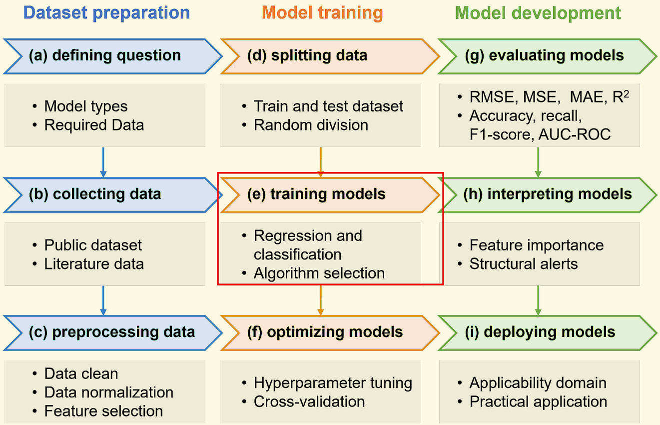
Cui et. al. (2023). Advances and applications of machine learning and deep learning in environmental ecology. Environmental Pollution Vol 335.

Experience (E):

Data: A Collection of examples, data points, observations.

Supervised or Unsupervised Learning

Cui et. al. (2023). Advances and applications of machine learning and deep learning in environmental ecology. Environmental Pollution Vol 335.

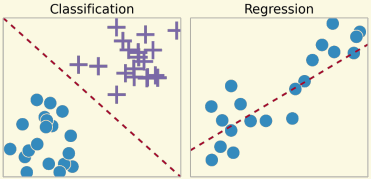


Overview

	Supervised Learning	Unsupervised Learning
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

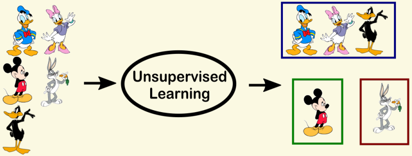
Supervised Learning

Learn a function that, given a sample of data and desired outputs, best approximates the relationship between input and output observable in the data.



Unsupervised Learning

Goal Is To Infer The Natural Structure Present Within A Set Of Data Points With Prior Expectations.

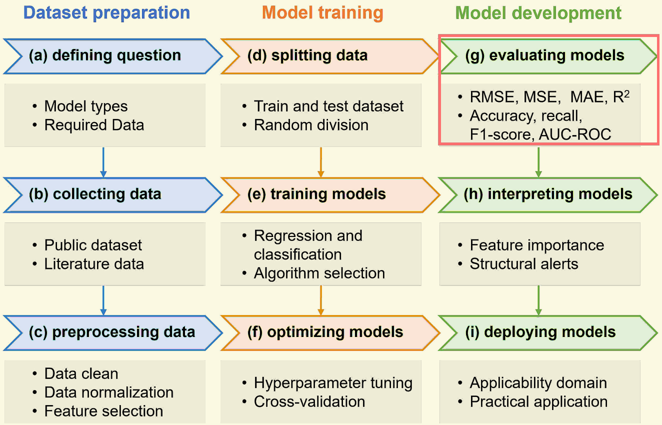


Performance Measure (P):

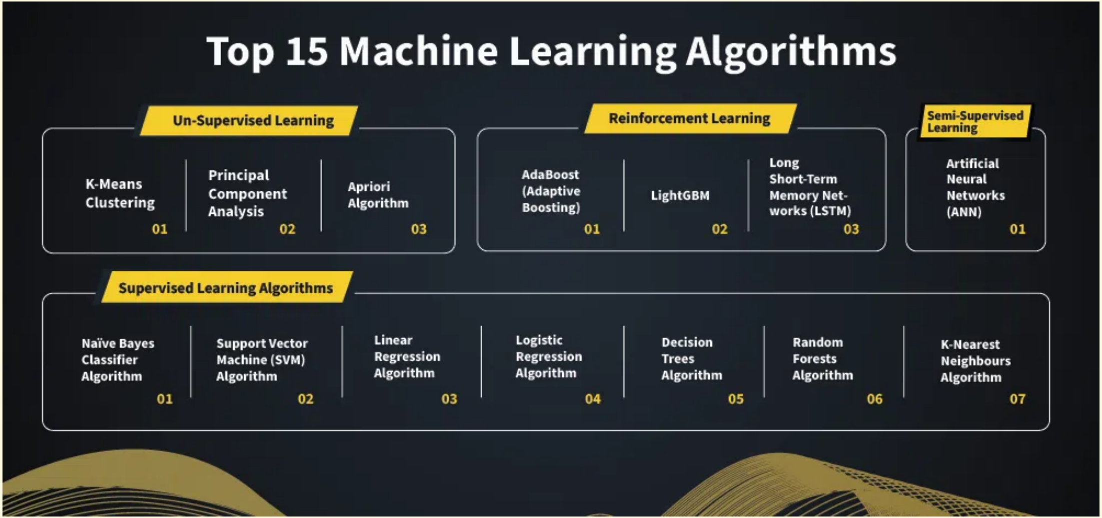
Evaluates the abilities of the machine learning system to perform the task (T).

For example, in regression you could use: **RMSE, R^2, RE**

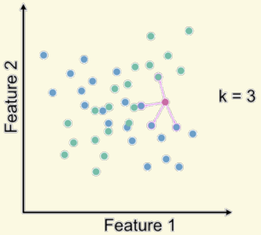
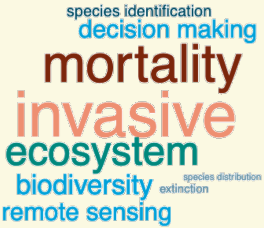
Cui et. al. (2023). Advances and applications of machine learning and deep learning in environmental ecology. Environmental Pollution Vol 335.



Machine Learning Model Overviews

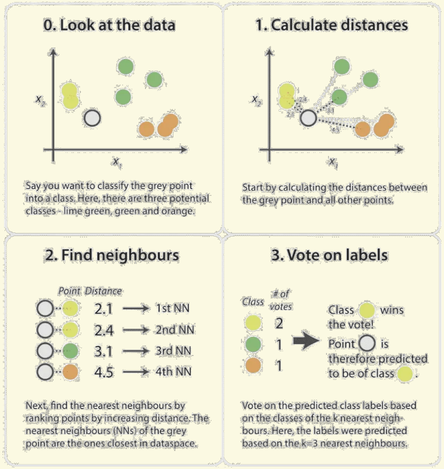


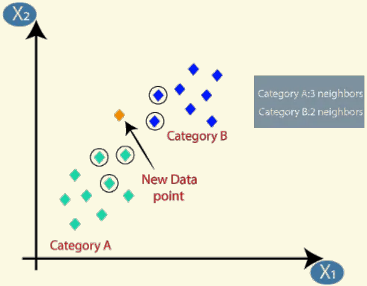
Machine Learning Model Overviews

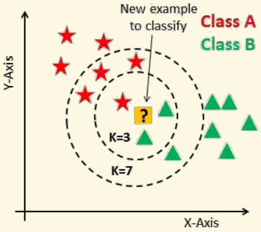
Machine learning model	Description	Data Type	Application areas
k-nearest-neighbor:  	K nearest neighbors in feature space decide response (e.g., by majority voting) + simple + no training - scales poorly - high dimensionality	Tabular Data: - Classification - Regression	

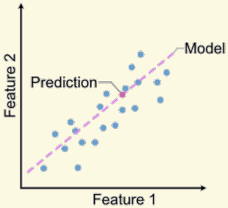
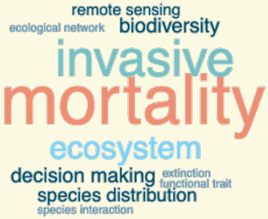
Pichler, M., & Hartig, F. (2023). Machine learning and deep learning - A review for ecologists. Methods in Ecology and Evolution, 14, 994–1016.

K Nearest Neighbor (KNN)





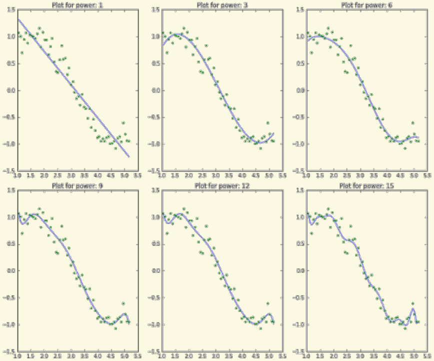


Machine learning model	Description	Data Type	Application areas
Lasso, Ridge Regression:  	Regression models with regularized coefficients + highly interpretable + few observations - Not very flexible	Tabular data: - Classification - Regression	

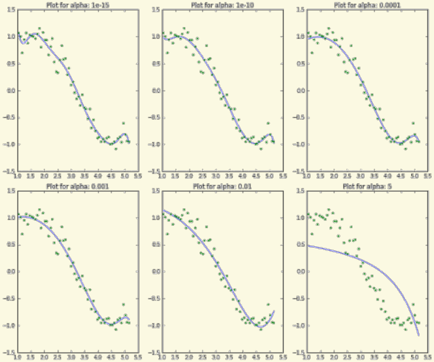
Pichler, M., & Hartig, F. (2023). Machine learning and deep learning - A review for ecologists. Methods in Ecology and Evolution, 14, 994–1016.

Polynomial Regression (No Penalty)

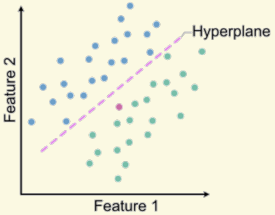
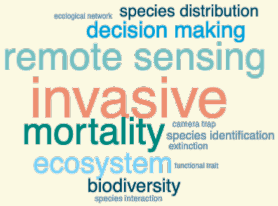
$$y = {}_0 + {}_1 x + {}_2 x^2 + {}_3 x^3 + \ldots + {}_n x^n +$$



Ridge Regression (Overfitting Penalty)

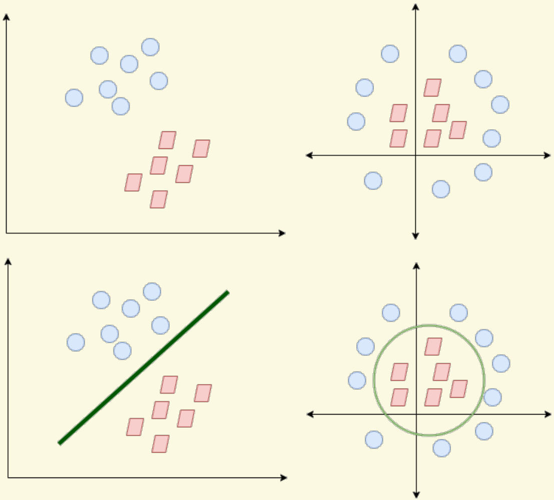


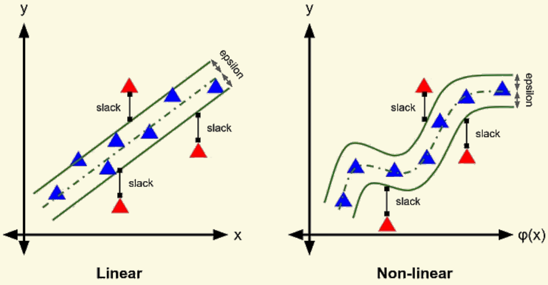
Machine Learning Model Overviews

Machine learning model	Description	Data Type	Application areas
Support vector machines: 	Hyperplane is optimized to separate response classes + fast and memory efficient + high dimensional data - kernel dependent - no probabilities	Tabular data: - Classification - Regression	

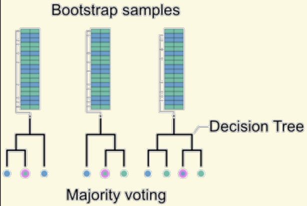
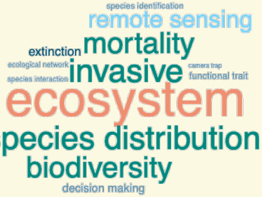
Pichler, M., & Hartig, F. (2023). Machine learning and deep learning - A review for ecologists. Methods in Ecology and Evolution, 14, 994–1016.

Support Vector Machines





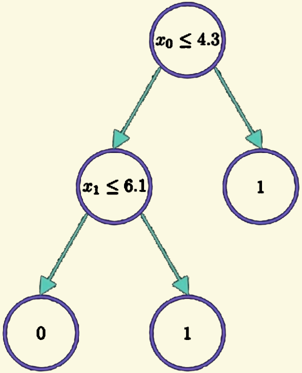
Machine Learning Model Overviews

Machine learning model	Description	Data Type	Application areas
Random Forest:  	N decision (regression) trees are fitted on bootstrap samples. Split variable is selected from random subset of variables + flexible + robust (e.g. outliers) + few hyper-parameters (+) variable importance - scales poorly	Tabular Data: - Classification - Regression	

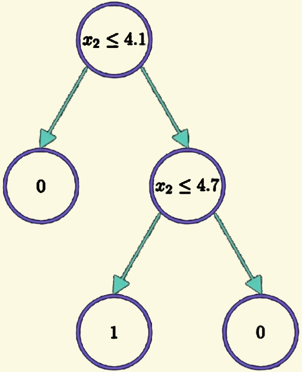
Pichler, M., & Hartig, F. (2023). Machine learning and deep learning - A review for ecologists. Methods in Ecology and Evolution, 14, 994–1016.

Decision Trees

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1



id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	6.5	4.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1



id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1

id
2
0
2
4
5
5

id
2
1
3
1
4
4

id
4
1
3
0
0
2

id
3
3
2
5
1
2

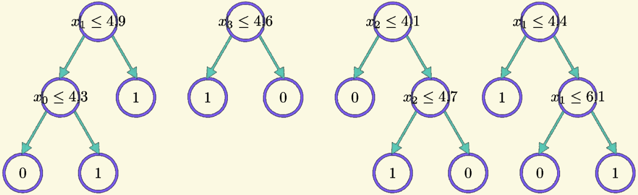
x_0, x_1 $\quad$ x_2, x_3 $\quad$ x_2, x_4 $\quad$ x_1, x_3

New Data:

2.8	6.2	4.3	5.3	5.5

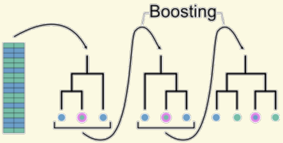
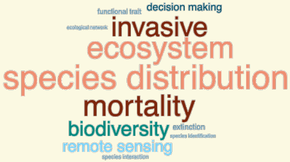
Bootstrap + Aggregating

(Bagging)



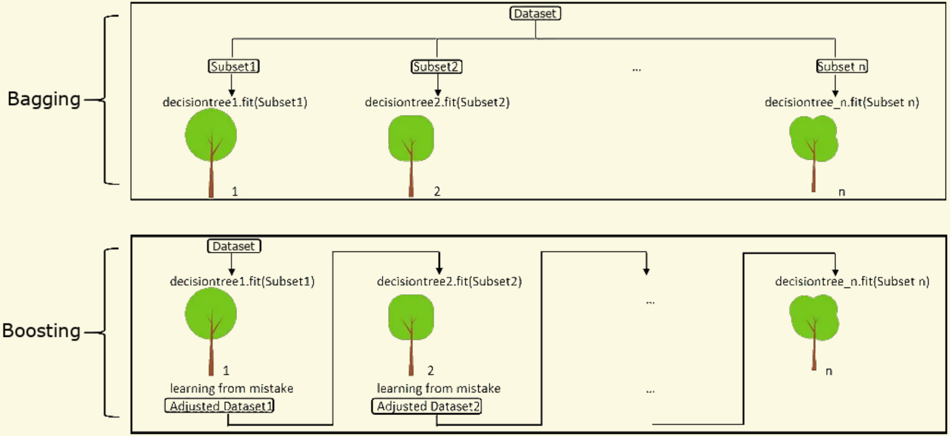
1 $\quad$ 0 $\quad$ 1 $\quad$ 1

Machine Learning Model Overviews

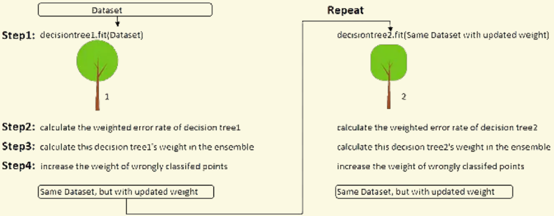
Machine learning model	Description	Data Type	Application areas
Boosted Regression Trees: 	N trees are fitted sequentially to minimize an overall loss function + flexible (+) variable importance - many hyper-parameters - high complexity	Tabular Data: - Classification - Regression	

Pichler, M., & Hartig, F. (2023). Machine learning and deep learning - A review for ecologists. Methods in Ecology and Evolution, 14, 994–1016.

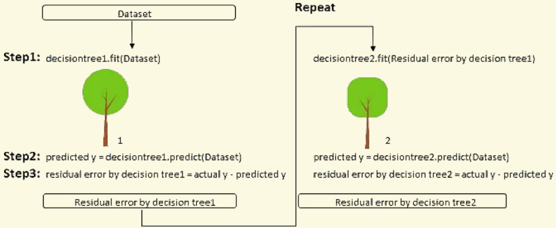
Random Forest : Boosting vs Bagging

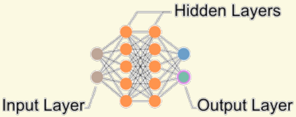
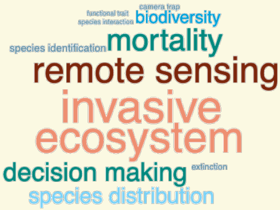


Adaptive Boosting:

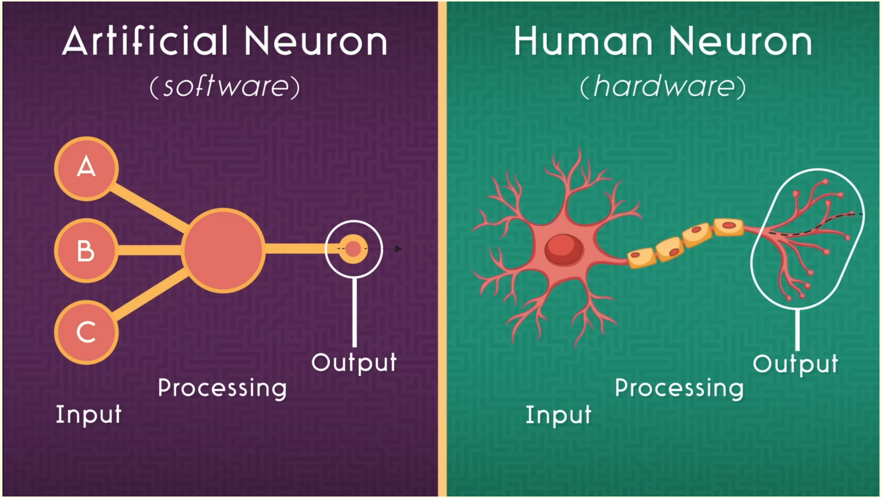


Gradient Boosting:

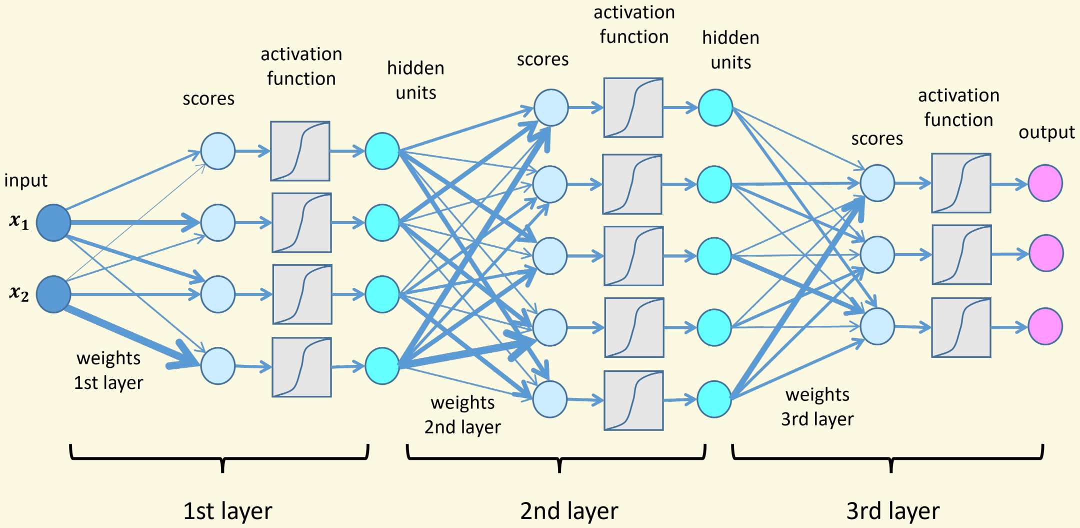


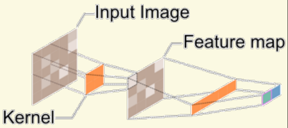
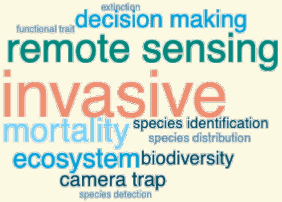
Machine learning model	Description	Data Type	Application areas
Deep Neural Networks: 	Input (features) are passed through many hidden layers. Last layer maps into response space + flexible + adaptive to different tasks - many hyper-parameters - computationally expensive	Tabular Data: - Classification - Regression	

Pichler, M., & Hartig, F. (2023). Machine learning and deep learning - A review for ecologists. Methods in Ecology and Evolution, 14, 994–1016.

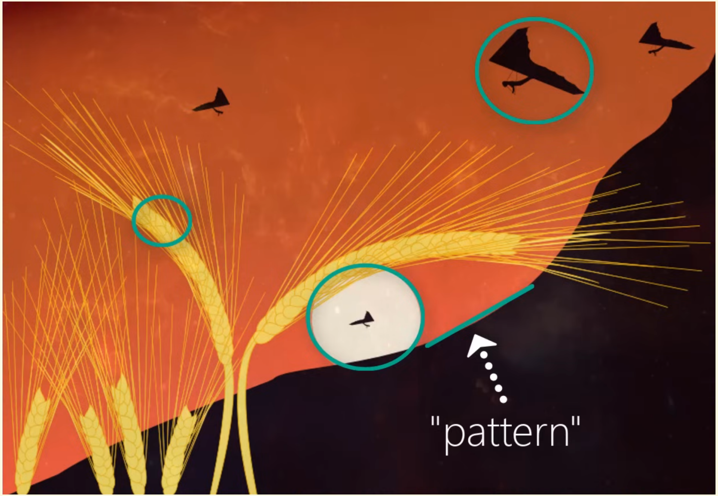

Artificial Neuron
(software)
A
B
C
Input
Processing
Output
Human Neuron
(hardware)
Input
Processing
Output

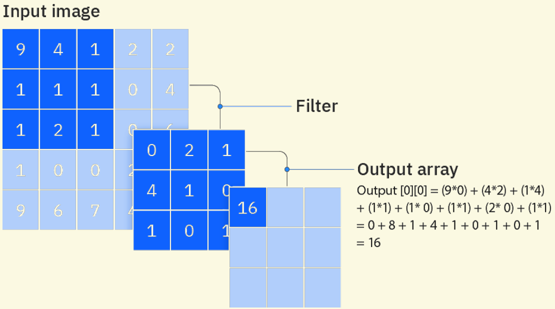
Neural Networks

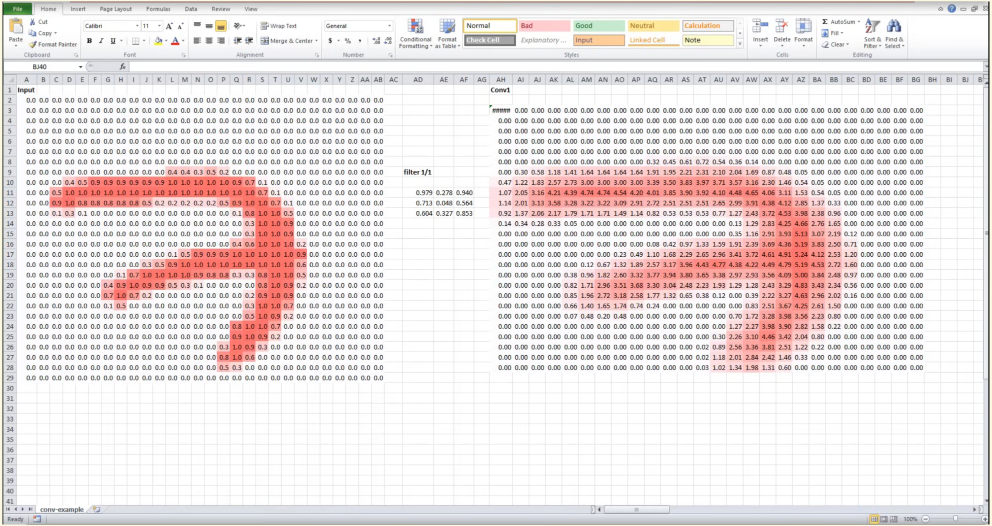


Machine learning model	Description	Data Type	Application areas
Convolutional Neural  Networks:	Small kernels (filters) processes images before passing it to fully connected layers + flexible + detecting shapes and edges - many hyper-parameters - computationally expensive	Images: - Classification - Object detection	

Pichler, M., & Hartig, F. (2023). Machine learning and deep learning - A review for ecologists. Methods in Ecology and Evolution, 14, 994–1016.


"pattern"


Input image
Filter
Output array
Output [0][0] = (9*0) + (4*2) + (1*4)
+ (1*1) + (1* 0) + (1*1) + (2* 0) + (1*1)
= 0 + 8 + 1 + 4 + 1 + 0 + 1 + 0 + 1
= 16



Input

Row	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.4	0.3	0.5	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	0.0	0.0	0.0	0.4	0.5	0.9	0.9	0.9	0.9	0.9	0.9	1.0	1.0	1.0	1.0	1.0	0.9	0.7	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	0.0	0.0	0.5	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.7	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12	0.0	0.0	0.9	1.0	0.8	0.8	0.8	0.8	0.8	0.5	0.2	0.2	0.2	0.2	0.2	0.5	0.9	1.0	1.0	0.7	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	0.0	0.0	0.1	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.8	1.0	1.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	1.0	1.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	1.0	1.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0
16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.6	1.0	1.0	1.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0
17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.5	0.9	0.9	0.9	1.0	1.0	1.0	1.0	1.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0
18	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.5	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0
19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.7	1.0	1.0	1.0	1.0	0.9	0.8	0.8	0.3	0.3	0.8	1.0	1.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0
20	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.9	1.0	0.9	0.9	0.5	0.3	0.1	0.0	0.0	0.0	0.0	0.8	1.0	0.9	0.2	0.0	0.0	0.0	0.0	0.0	0.0
21	0.0	0.0	0.0	0.0	0.0	0.0	0.7	1.0	0.7	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.9	1.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0
22	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	1.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
23	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0	0.9	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
24	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	1.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.9	1.0	0.9	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
26	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	1.0	0.9	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
27	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	1.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
28	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
29	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

filter 1/1

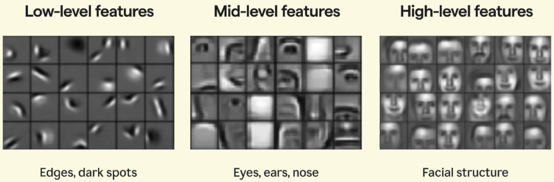
Row	AD	AE	AF
11	0.979	0.278	0.940
12	0.713	0.048	0.564
13	0.604	0.327	0.853

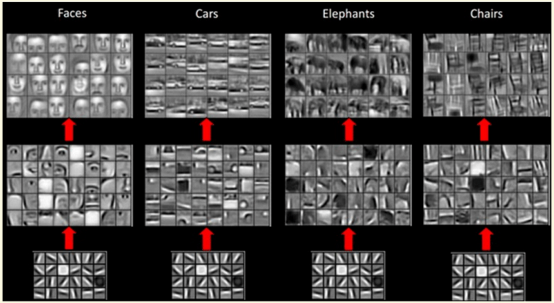
Conv1

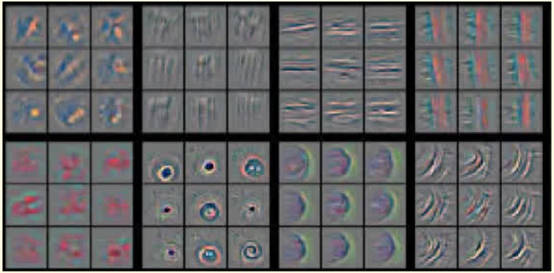
Row	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG
3	#####	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.32	0.45	0.61	0.72	0.54	0.36	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.30	0.58	1.18	1.41	1.64	1.64	1.64	1.64	1.91	1.95	2.21	2.31	2.10	2.04	1.69	0.87	0.48	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.47	1.22	1.83	2.57	2.73	3.00	3.00	3.00	3.00	3.39	3.50	3.83	3.97	3.71	3.57	3.16	2.30	1.46	0.54	0.05	0.00	0.00	0.00	0.00	0.00	0.00
11	1.07	2.05	3.16	4.21	4.39	4.74	4.74	4.54	4.20	4.01	3.85	3.90	3.92	4.10	4.48	4.65	4.06	3.11	1.53	0.54	0.05	0.00	0.00	0.00	0.00	0.00
12	1.14	2.01	3.13	3.58	3.28	3.22	3.22	3.09	2.91	2.72	2.51	2.51	2.51	2.65	2.99	3.91	4.38	4.12	2.85	1.37	0.33	0.00	0.00	0.00	0.00	0.00
13	0.92	1.37	2.06	2.17	1.79	1.71	1.71	1.49	1.14	0.82	0.53	0.53	0.53	0.77	1.27	2.43	3.72	4.53	3.98	2.38	0.96	0.00	0.00	0.00	0.00	0.00
14	0.14	0.34	0.28	0.33	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	1.29	2.83	4.25	4.66	2.76	1.65	0.00	0.00	0.00	0.00	0.00
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.35	1.16	2.91	3.93	5.13	3.07	2.19	0.12	0.00	0.00	0.00	0.00
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.42	0.97	1.33	1.59	1.91	2.39	3.69	4.36	5.19	3.83	2.50	0.71	0.00	0.00	0.00	0.00
17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.23	0.49	1.10	1.68	2.29	2.65	2.96	3.41	3.72	4.61	4.91	5.24	4.12	2.53	1.20	0.00	0.00	0.00	0.00
18	0.00	0.00	0.00	0.00	0.00	0.12	0.67	1.32	1.89	2.57	3.17	3.96	4.43	4.77	4.38	4.22	4.49	4.79	5.19	4.53	2.72	1.60	0.00	0.00	0.00	0.00
19	0.00	0.00	0.00	0.00	0.38	0.96	1.82	2.60	3.32	3.77	3.94	3.80	3.65	3.38	2.97	2.93	3.56	4.09	5.00	3.84	2.48	0.97	0.00	0.00	0.00	0.00
20	0.00	0.00	0.00	0.00	0.82	1.71	2.96	3.51	3.68	3.30	3.04	2.48	2.23	1.93	1.29	1.28	2.43	3.29	4.83	3.43	2.34	0.56	0.00	0.00	0.00	0.00
21	0.00	0.00	0.00	0.00	0.85	1.96	2.72	3.18	2.58	1.77	1.32	0.65	0.38	0.12	0.00	0.39	2.22	3.27	4.63	2.96	2.02	0.16	0.00	0.00	0.00	0.00
22	0.00	0.00	0.00	0.00	0.66	1.40	1.65	1.74	0.74	0.24	0.00	0.00	0.00	0.00	0.00	0.83	2.51	3.67	4.25	2.61	1.50	0.00	0.00	0.00	0.00	0.00
23	0.00	0.00	0.00	0.00	0.07	0.48	0.20	0.48	0.00	0.00	0.00	0.00	0.00	0.00	0.70	1.72	3.28	3.98	3.56	2.23	0.80	0.00	0.00	0.00	0.00	0.00
24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.27	2.27	3.98	3.90	2.82	1.58	0.22	0.00	0.00	0.00	0.00	0.00
25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.30	2.26	3.10	4.46	3.42	2.04	0.80	0.00	0.00	0.00	0.00	0.00	0.00
26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.89	2.56	3.36	3.81	2.51	1.22	0.22	0.00	0.00	0.00	0.00	0.00	0.00
27	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	1.18	2.01	2.84	2.42	1.46	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00
28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	1.02	1.34	1.98	1.31	0.60	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

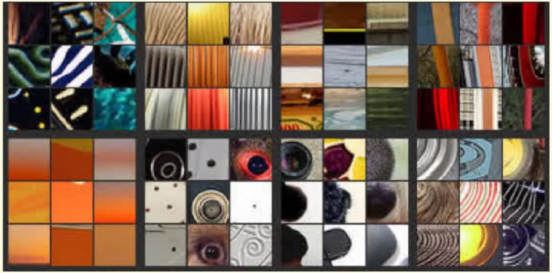
Convolutional Neural Networks

Convolutional Neural Networks





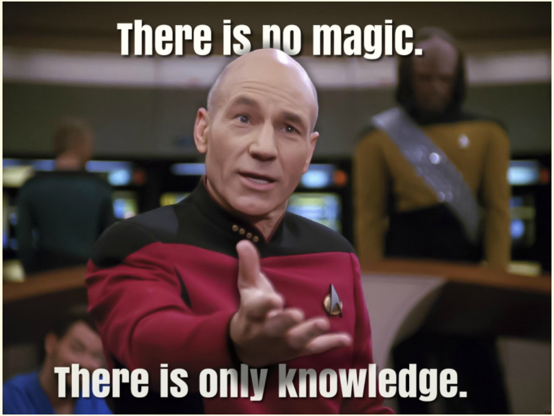




There Is No Magic...

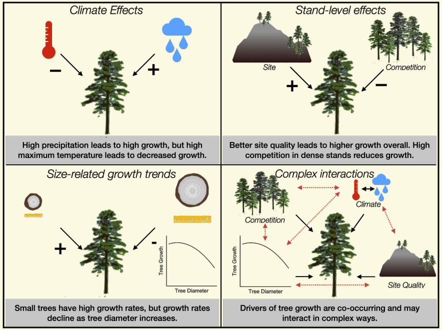
There is no magic here, no one model is always better than another.

Fit the models, make an informed choice, enjoy your life!

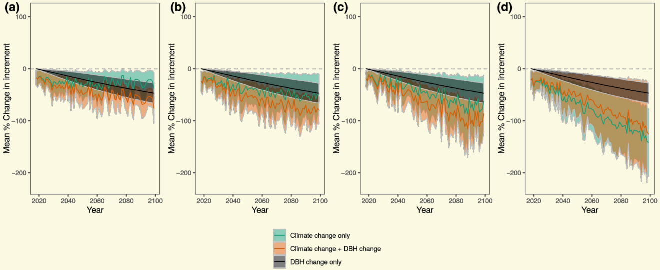


Picking a Model: Big Picture Goals?

Inference: Explanatory Power

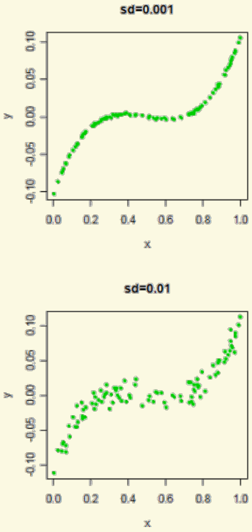


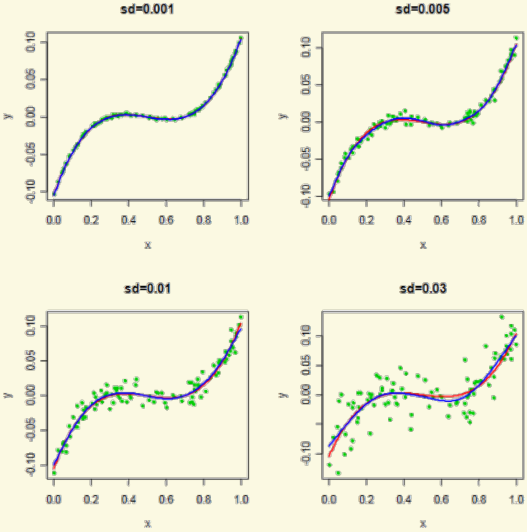
Prediction: Predictive Power



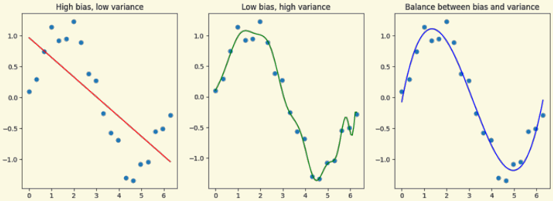
Heilman et. al. (2022). Ecological forecasting of tree growth: Regional fusion of tree-ring and forest inventory data to quantify drivers and characterize uncertainty. Global Change Biology, 28, 24422460.

Picking a Model: Bias & Overfitting



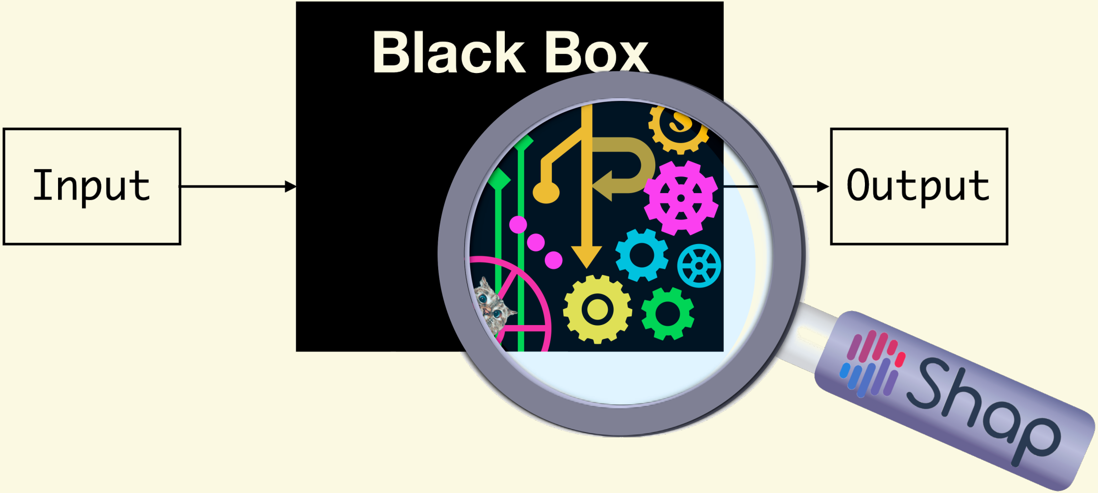


Picking a Model: Bias & Overfitting

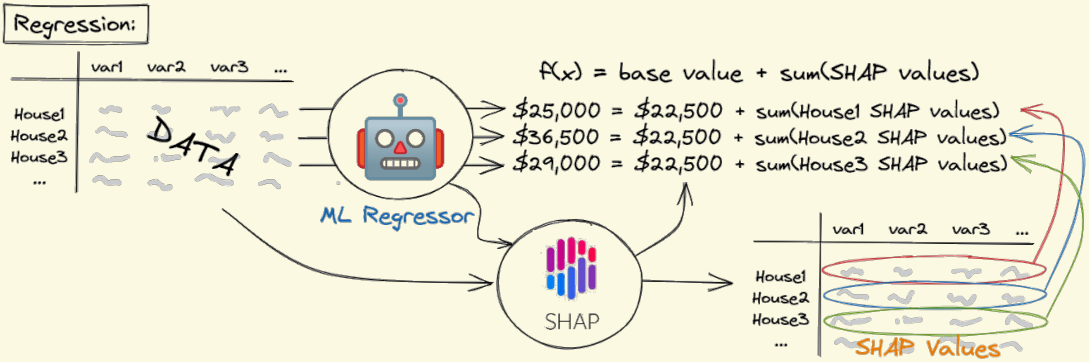


Bias-Variance Tradeoff

- Model is too simple, it will have very few parameters then it may have high bias and low variance.
- On the other hand if the model has large number of parameters then its going to have high variance and low bias.
- Our job is to strike the correct balance.

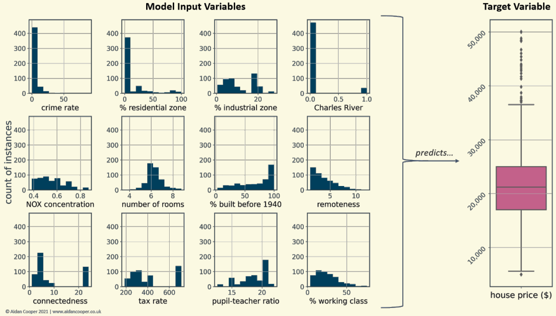

Input
Black Box
Output
Shap

SHAP (SHapley Additive exPlanations)



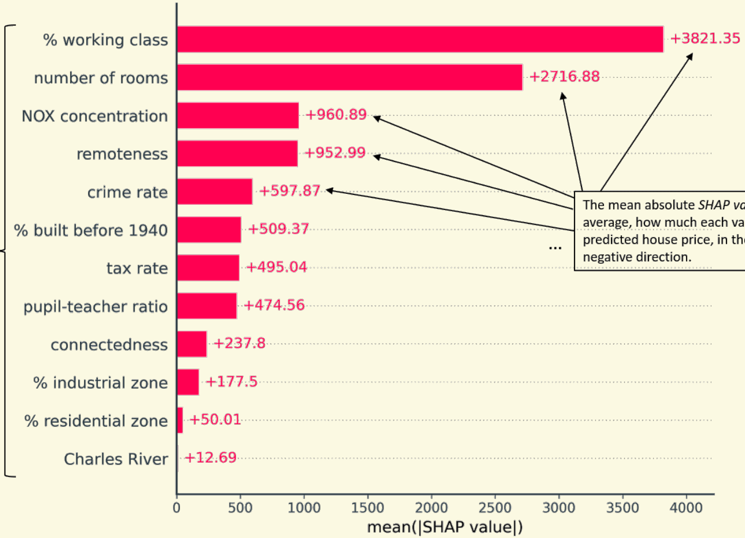
SHAP (SHapley Additive exPlanations)

Variable Name	Description
crime rate	Per capita crime rate in the town.
% residential zone	Percentage of land zoned for residential use.
% industrial zone	Percentage of land zoned for industrial use.
Charles River	1 if the house borders the Charles River; 0 otherwise.
NOX concentration	Nitric oxides concentration (parts per 10 million).
number of rooms	The average number of rooms per house in the housing unit.
% built before 1940	The proportion of houses built prior to 1940 in the unit.
remoteness	A measure of how far the housing is from employment centres (higher is more remote).
connectedness	A measure of how good the local road connections are (higher is more connected).
tax rate	Property tax rate per $10,000 of house value.
pupil-teacher ratio	The pupil:teacher ratio of the town.
% working class	Percentage of the population that is working class.



SHAP (SHapley Additive exPlanations)

These are the input variables, ranked from top to bottom by their mean absolute *SHAP values* for the entire dataset – i.e. the average magnitude of each variable's impact on the predicted house price across all instances.

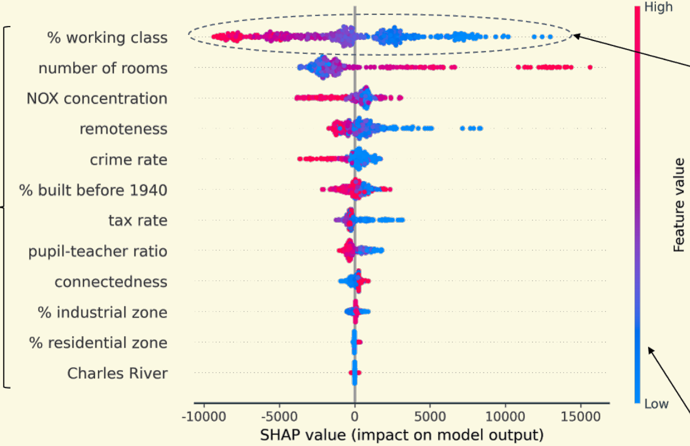


The mean absolute *SHAP values* are, on average, how much each variable impacts the predicted house price, in the positive or negative direction.

SHAP (SHapley Additive exPlanations)

These are the input variables, ranked from top to bottom by their mean absolute *SHAP values* for the entire dataset.

Note: this ranking is exactly the same as for the bar plot.



In a **beeswarm plot**, for each variable, every instance (i.e. row) of the dataset appears as it's own point. The points are distributed horizontally along the *x*-axis according to their *SHAP* value. In places where there is a high density of *SHAP values*, the points are stacked vertically.

Examining how the *SHAP* values are distributed reveals how a variable may influence the model's predictions.

The colour bar corresponds to the raw values (not to be confused with the *SHAP values*) of the variables for each instance (i.e. point) on the graph.

If the value of a variable for a particular instance is relatively high, it appears as a red dot. Relatively low variable values appear as blue dots.

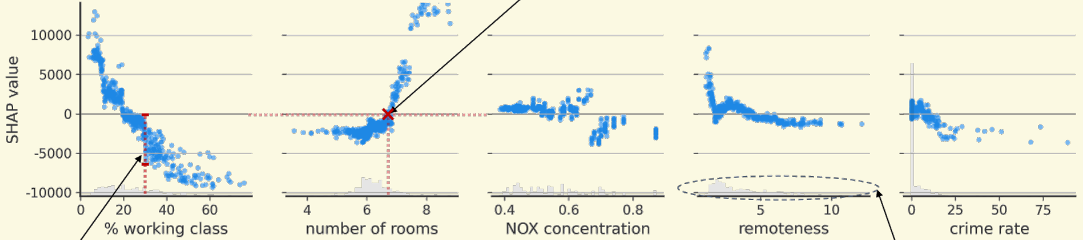
Examining the colour distribution horizontally along the *x*-axis for each variable provides insights into the general relationship between a variable's raw values and its *SHAP values*.

SHAP (SHapley Additive exPlanations)

In a **dependence plot**, every instance (i.e. row) of the dataset appears as it's own point. The points are presented as a scatterplot of a variable's *SHAP values* versus the variables underlying raw values.

SHAP values above the *y*=0 line lead to predictions of higher house prices, whereas those below it are associated with lower house price predictions. The raw variable value at which the distribution of *SHAP values* cross the *y*=0 line tells you the threshold at which the model switches from predicting lower to higher house prices. For *number of rooms*, this is at approximately 6.8 rooms, as marked by the ×.

With all five plots on the same *y*-scale, the extent of the vertical distribution of the *SHAP values* indicates how much relative influence each variable has on predictions. *% working class* has a much wider range of *SHAP values* than *crime rate*.



The vertical spread of *SHAP values* at a fixed raw variable value is due to *interaction effects* with other variables. For example, here we see that houses with a *% working class* of 30% can have *SHAP values* that range from $0 to -$6,500 depending on the other data for those particular instances.

The shapes of the distributions of points provide insights into the relationship between a variable's values and its *SHAP values.* For *% working class*, we see a negative, linear relationship across the full range of variable values. For *number of rooms*, we see that *SHAP values* are mostly flat between 4 and 6.5 rooms, but then increase sharply for higher room counts.

The inset histograms just above the *x*-axis display the distributions of raw variable values. We should be cautious not to overinterpret regions of the dependence plot where the underlying data is sparse (e.g. *crime rates* over 25%).

1. Split Data Into **Training** and **Test** Datasets. (Cross Validation)
2. **Fit** Candidate Models on Training Dataset
3. **Assess** Performance of Candidate Models on Testing Dataset Using Same Set of Metrics
4. **Choose** Final Model Form
5. **Fit** Final Model Form
6. **Interpret** Model Output / **Predict** Future Responses
7. Walk Away Feeling **Empowered**

Classification

And

Regression

Training