

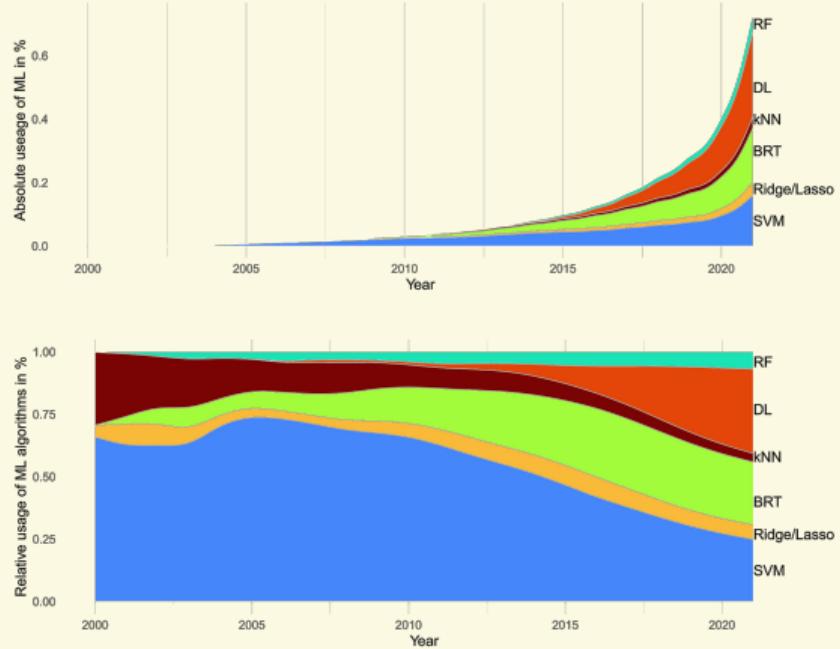
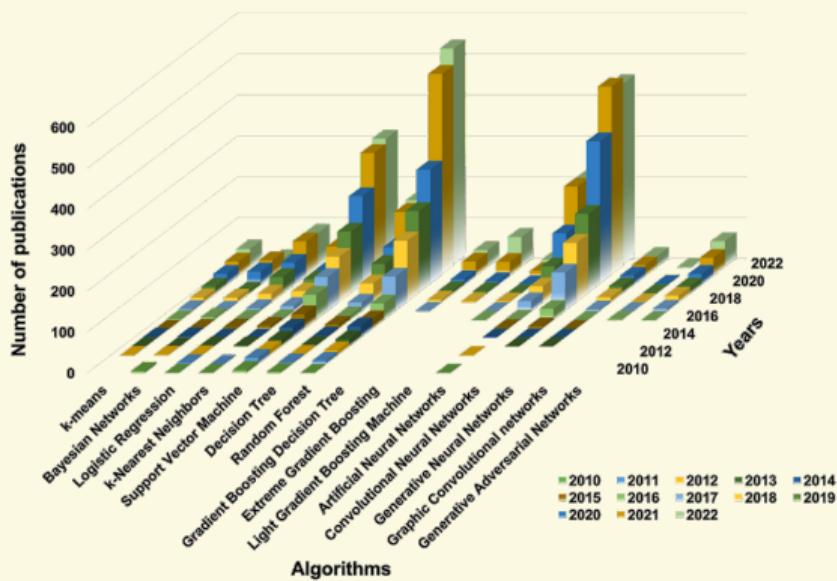
ECOML WORKSHOP:

MACHINE LEARNING FOR ECOLOGICAL RESEARCH

JEREMY FORSYTHE
& JAZLYNN HALL



MACHINE LEARNING PUBLICATIONS IN ECOLOGY

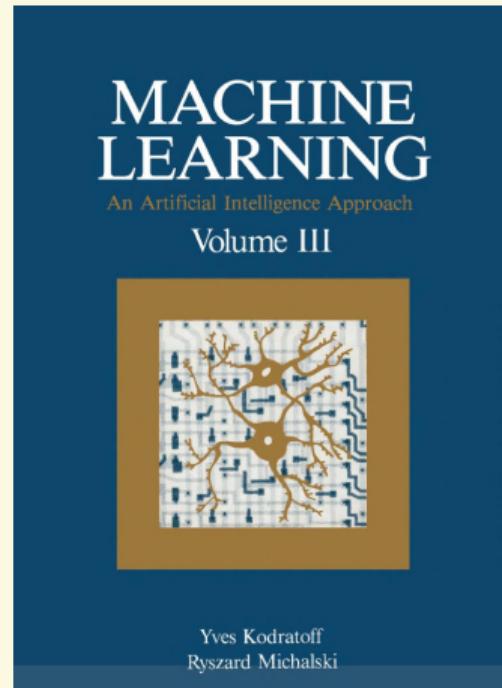


Cui et. al. (2023). Advances and applications of machine learning and deep learning in environmental ecology. *Environmental Pollution* Vol 335. Pichler, M., & Hartig, F. (2023). Machine learning and deep learning - A review for ecologists. *Methods in Ecology and Evolution*, 14, 994–1016.

MACHINE LEARNING DEFINITION

Widely Accepted Definition:

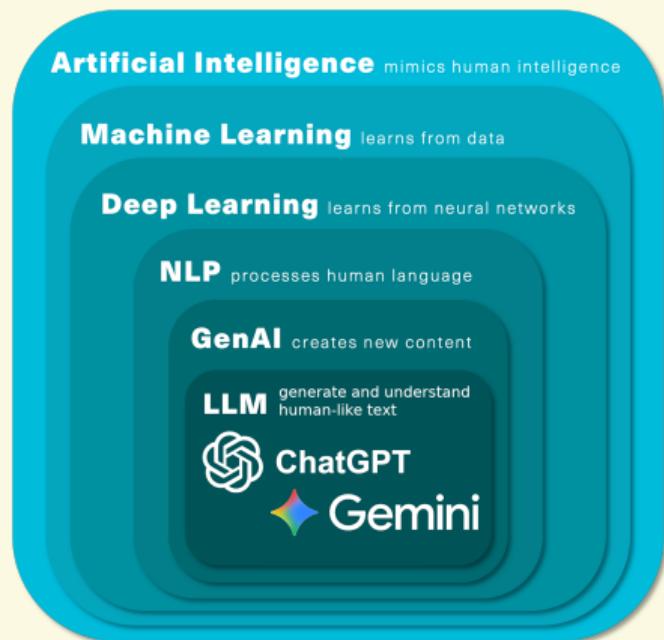
A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E .



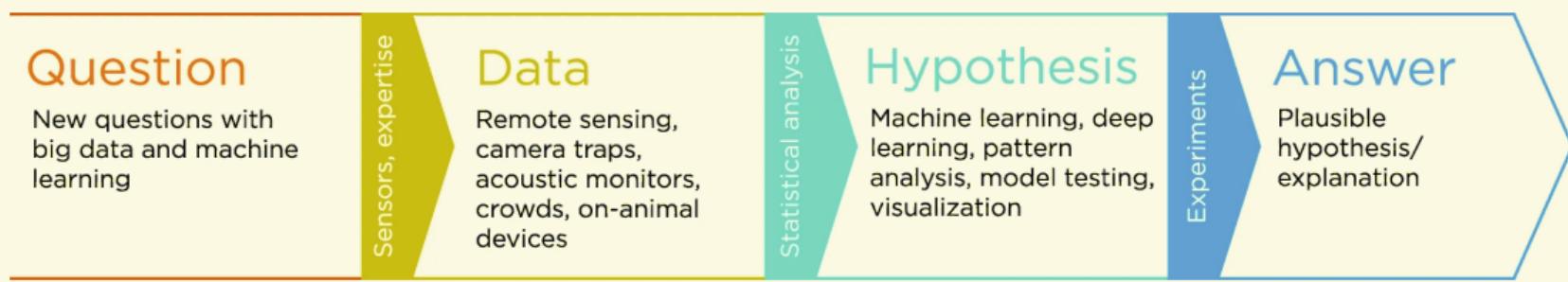
MACHINE LEARNING DEFINITION

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 .



SCIENTIFIC METHOD



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

CLASSICAL STATISTICS VS ML : DIFFERENT SIDES. SAME COIN.

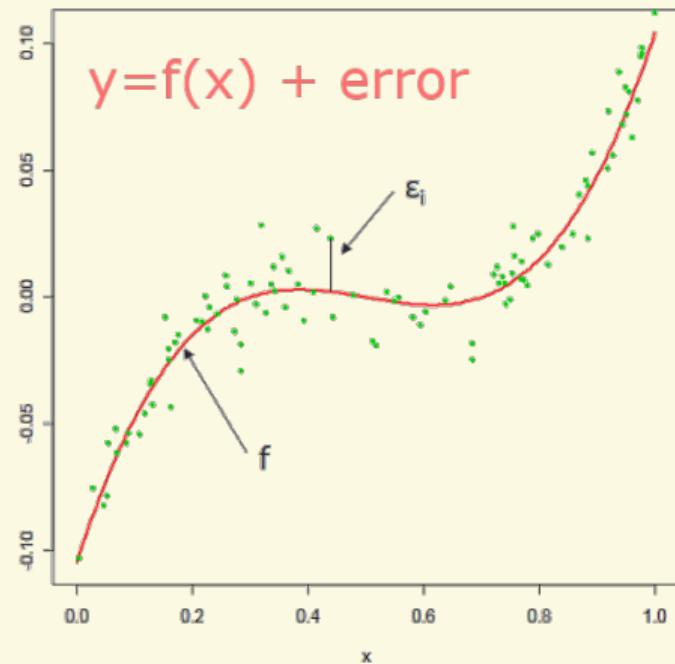
Example: One of the most common tasks we have as ecologists. Regression.

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

Given $x_i = (x_{i1}, x_{i2}, \dots, x_{in})$

and $y = y_1, y_2, \dots, y_n$:

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 VS ML : APPROACH TO MODELING

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

(R^2)

Machine Learning

- **Predictive Power:** How well does the model perform on new, unseen data?
- **Metrics:**

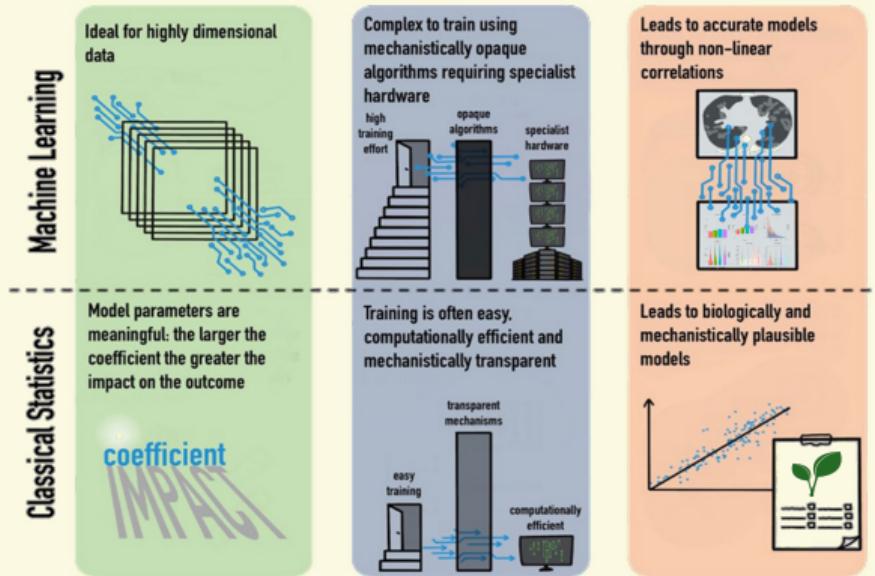
Cross-Validation Accuracy

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



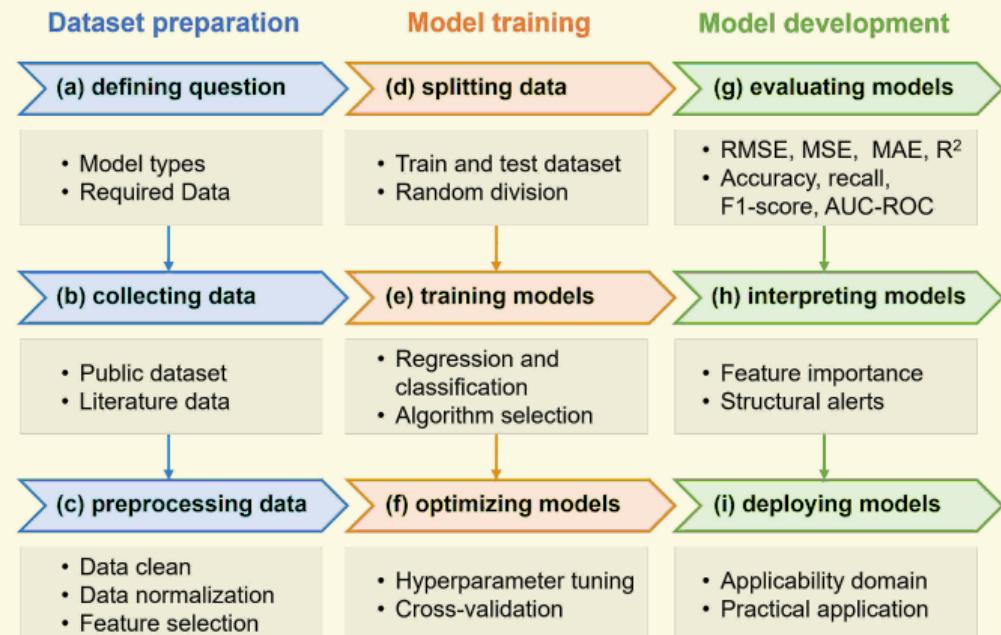
“All models are wrong, but some are useful.” - George Box

MACHINE LEARNING BASICS

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

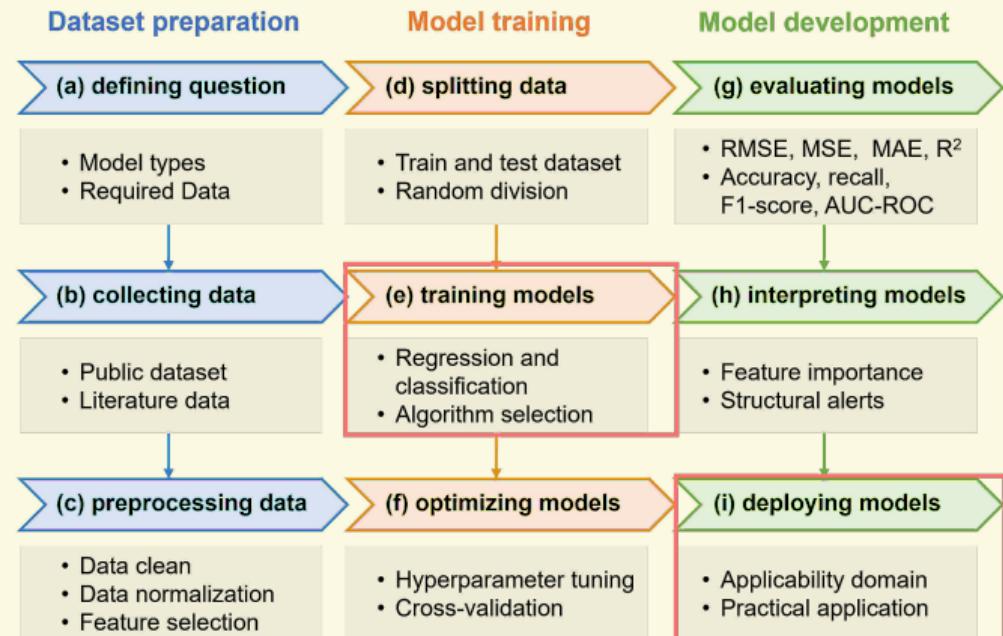
Task (T):

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.



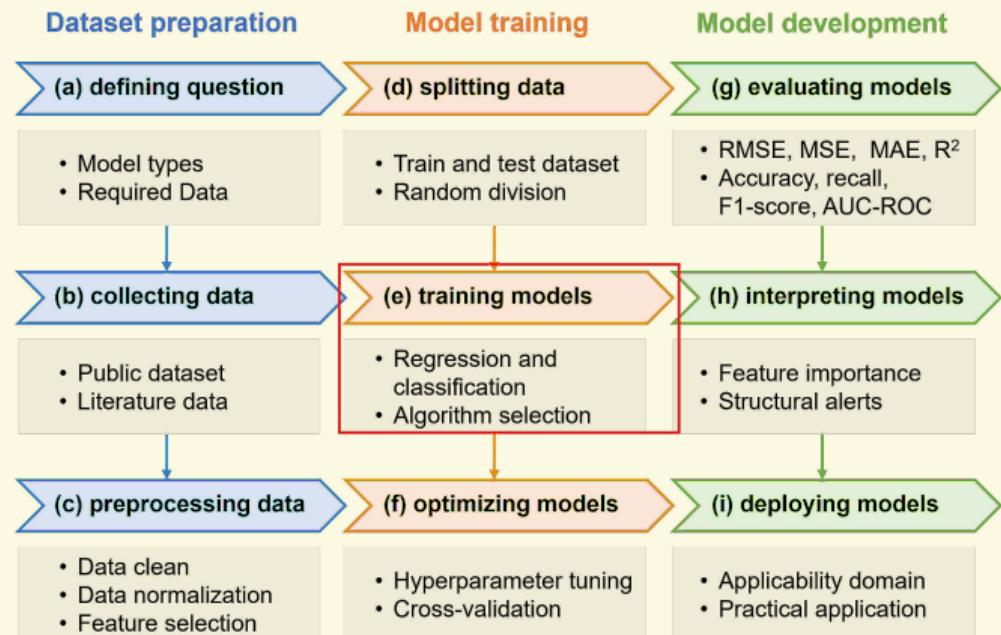
MACHINE LEARNING BASICS : EXPERIENCE

Experience (E):

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

Supervised or Unsupervised Learning

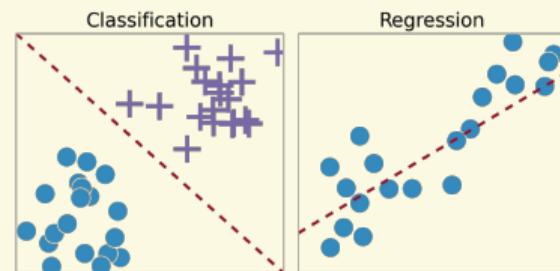
Cui et. al. (2023). Advances and applications of machine learning and deep learning in environmental ecology. Environmental Pollution Vol 335.



SUPERVISED VS UNSUPERVISED LEARNING

Overview

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

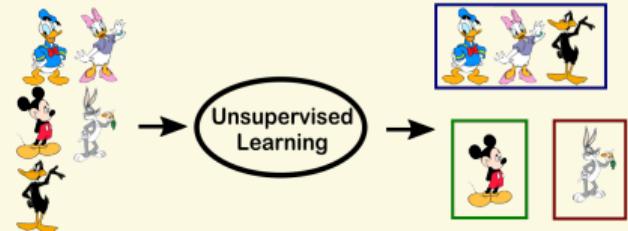


Supervised Learning

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.



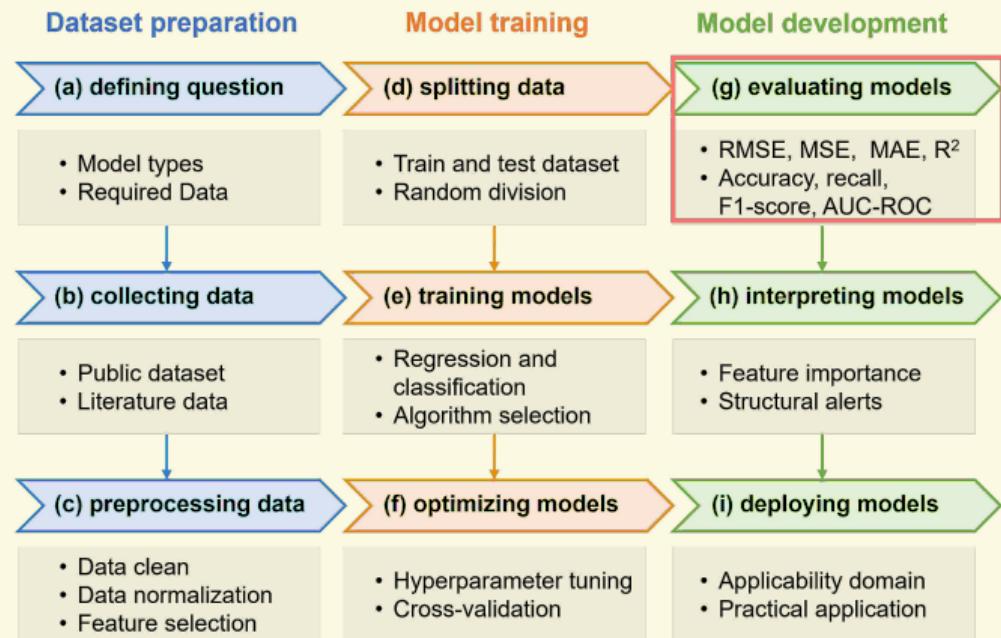
MACHINE LEARNING BASICS : PERFORMANCE MEASUREMENTS

Performance Measure (P):

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

For example, in regression you could use: **RMSE, R², 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

Top 15 Machine Learning Algorithms

Un-Supervised Learning

K-Means
Clustering

01

Principal
Component
Analysis

02

Apriori
Algorithm

03

Reinforcement Learning

AdaBoost
(Adaptive
Boosting)

01

LightGBM

02

Long
Short-Term
Memory Net-
works (LSTM)

03

Semi-Supervised Learning

Artificial
Neural
Networks
(ANN)

01

Supervised Learning Algorithms

Naïve Bayes
Classifier
Algorithm

01

Support Vector
Machine (SVM)
Algorithm

02

Linear
Regression
Algorithm

03

Logistic
Regression
Algorithm

04

Decision
Trees
Algorithm

05

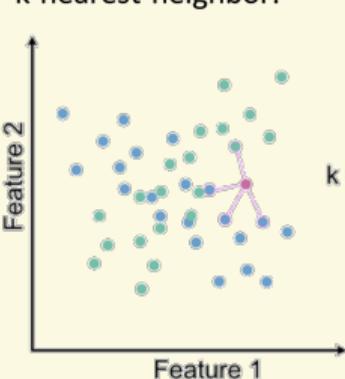
Random
Forests
Algorithm

06

K-Nearest
Neighbours
Algorithm

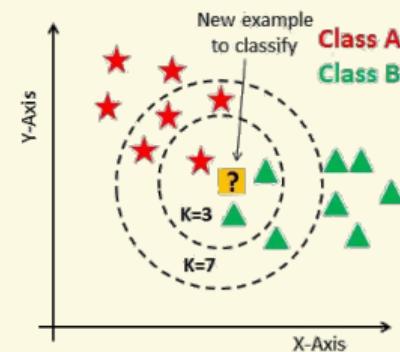
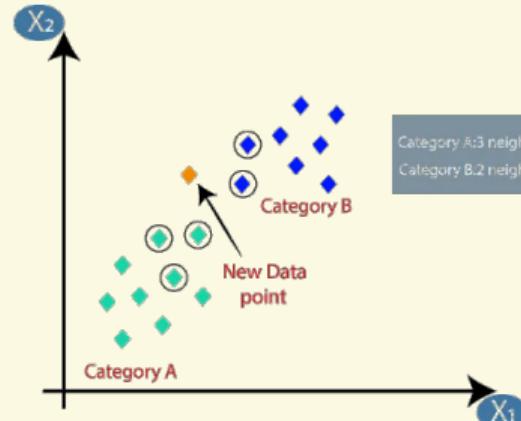
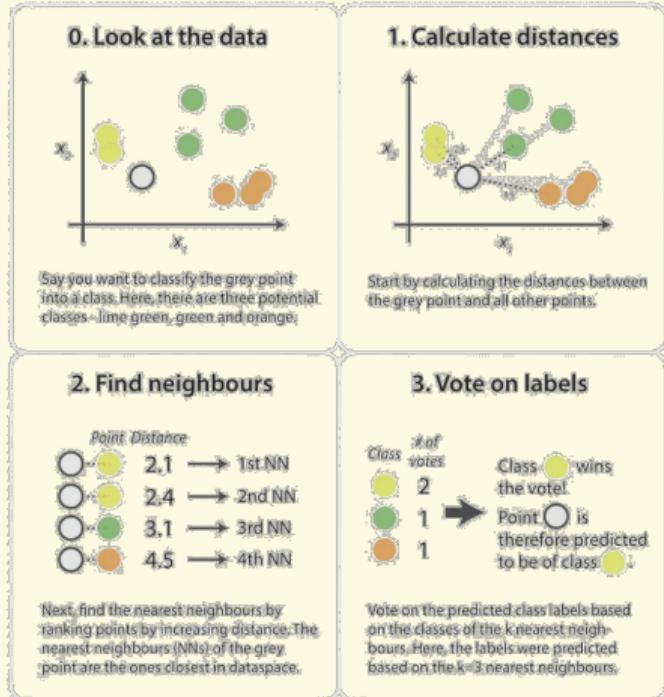
07

MACHINE LEARNING MODEL OVERVIEWS

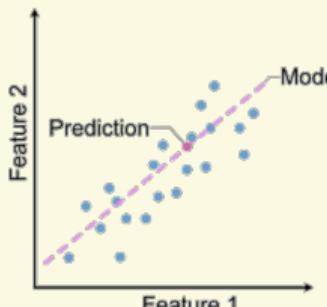
| Machine learning model | Description | Data Type | Application areas |
|---|---|--|--|
| <p>k-nearest-neighbor:</p>  | <p>K nearest neighbors in feature space decide response (e.g., by majority voting)</p> <ul style="list-style-type: none">+ simple+ no training- scales poorly- high dimensionality | <p>Tabular Data:</p> <ul style="list-style-type: none">- Classification- Regression | <p>species identification decision making mortality invasive ecosystem biodiversity remote sensing <small>species distribution extinction</small></p> |

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 OVERVIEWS

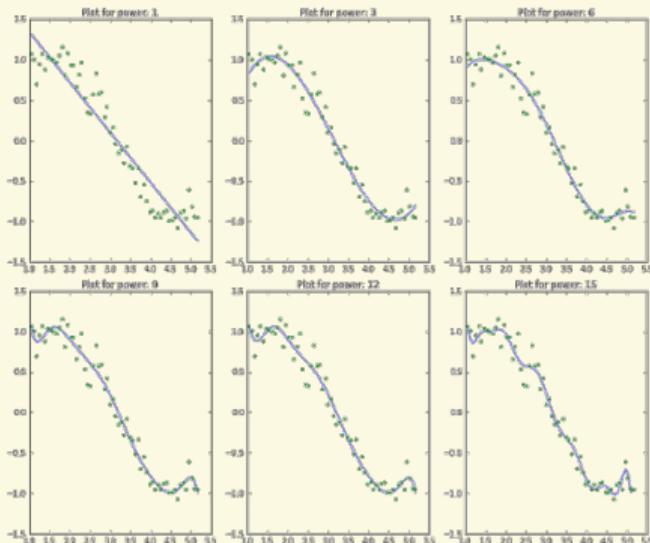
| Machine learning model | Description | Data Type | Application areas |
|---|--|--|--|
| Lasso, Ridge Regression:  | <p>Regression models with regularized coefficients</p> <ul style="list-style-type: none">+ highly interpretable+ few observations- Not very flexible | <p>Tabular data:</p> <ul style="list-style-type: none">- Classification- Regression | <p>remote sensing ecological network biodiversity</p> <p>invasive mortality ecosystem</p> <p>decision making extinction species distribution functional trait species interaction</p> |

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

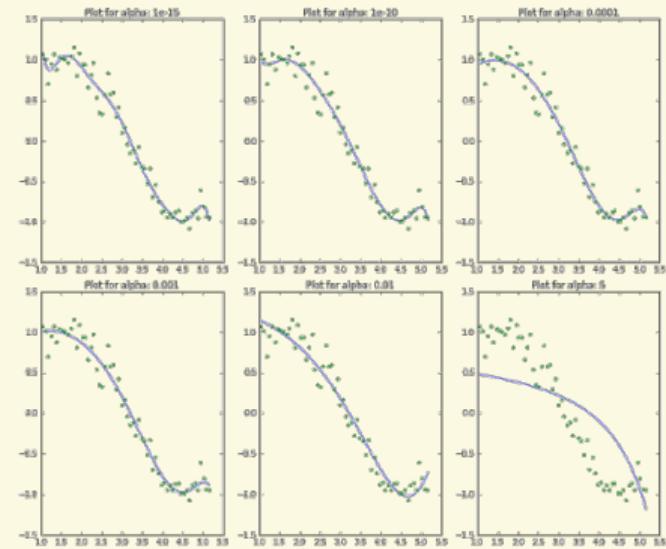
LASSO & RIDGE REGRESSION

Polynomial Regression (No Penalty)

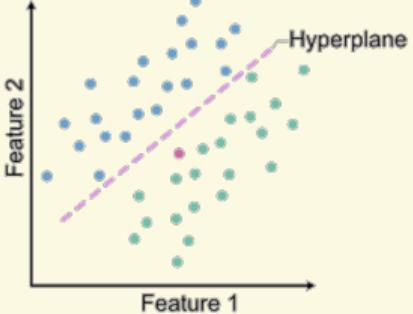
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots + \beta_n x^n + \epsilon$$



Ridge Regression (Overfitting Penalty)

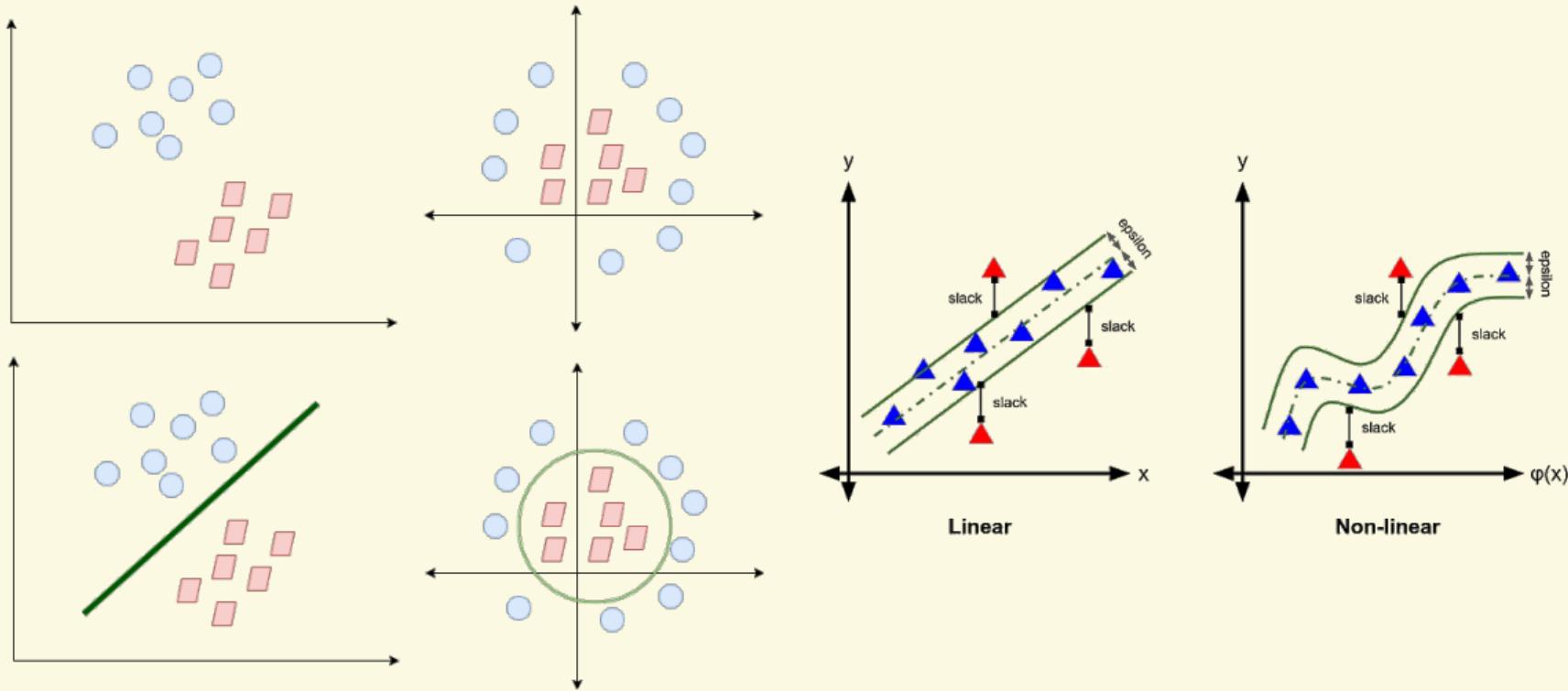


MACHINE LEARNING MODEL OVERVIEWS

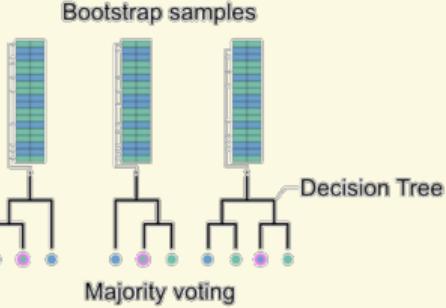
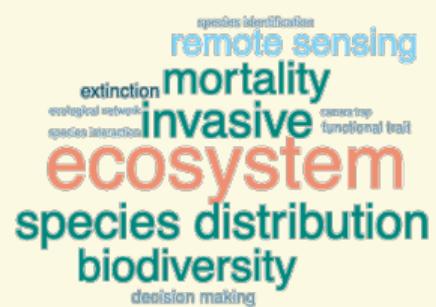
| Machine learning model | Description | Data Type | Application areas |
|---|---|--|---|
| Support vector machines:  | <p>Hyperplane is optimized to separate response classes</p> <ul style="list-style-type: none">+ fast and memory efficient+ high dimensional data- kernel dependent- no probabilities | <p>Tabular data:</p> <ul style="list-style-type: none">- Classification- Regression | <p>ecological network species distribution decision making remote sensing invasive mortality ecosystem biodiversity</p> <p>camera trap species identification definition functional trait species interaction</p> |

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



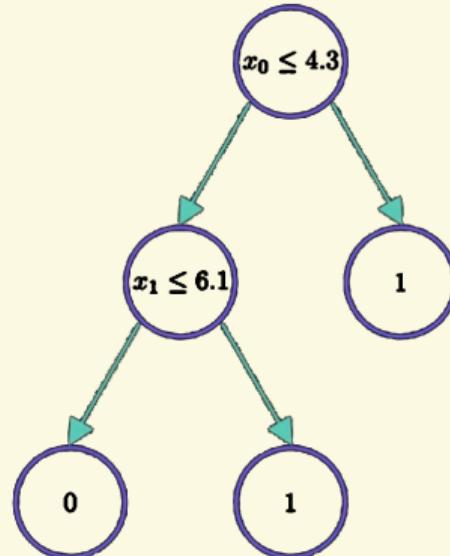
MACHINE LEARNING MODEL OVERVIEWS

| Machine learning model | Description | Data Type | Application areas |
|--|---|--|---|
| <p>Random Forest:</p>  <p>Bootstrap samples</p> <p>Decision Tree</p> <p>Majority voting</p> | <p>N decision (regression) trees are fitted on bootstrap samples. Split variable is selected from random subset of variables</p> <ul style="list-style-type: none">+ flexible+ robust (e.g. outliers)+ few hyper-parameters(+) variable importance- scales poorly | <p>Tabular Data:</p> <ul style="list-style-type: none">- Classification- Regression |  |

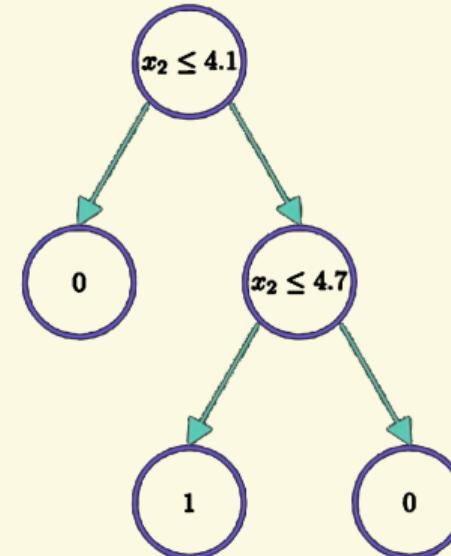
Pichler, M., & Hartig, F. (2023). Machine learning and deep learning - A review for ecologists. *Methods in Ecology and Evolution*, 14, 994–1016.

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 |



RANDOM FOREST

| <i>id</i> | 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 |

| <i>id</i> |
|-----------|
| 2 |
| 0 |
| 2 |
| 4 |
| 5 |
| 5 |

| <i>id</i> |
|-----------|
| 2 |
| 1 |
| 3 |
| 1 |
| 4 |
| 4 |

| <i>id</i> |
|-----------|
| 4 |
| 1 |
| 3 |
| 0 |
| 0 |
| 2 |

| <i>id</i> |
|-----------|
| 3 |
| 3 |
| 2 |
| 5 |
| 1 |
| 2 |

New Data:

| | | | | |
|-----|-----|-----|-----|-----|
| 2.8 | 6.2 | 4.3 | 5.3 | 5.5 |
|-----|-----|-----|-----|-----|

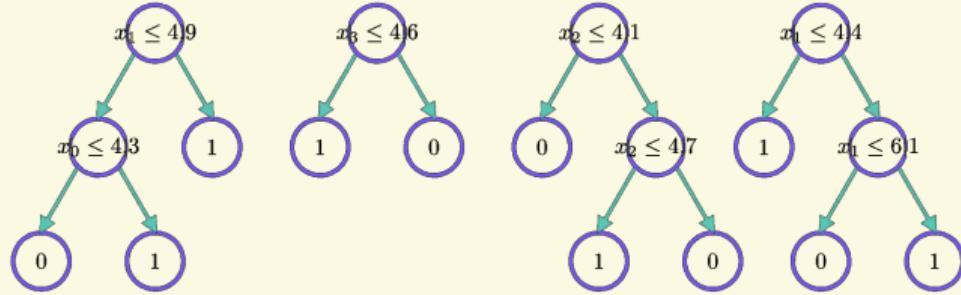
Bootstrap + Aggregating
(Bagging)

x_0, x_1

x_2, x_3

x_2, x_4

x_1, x_3



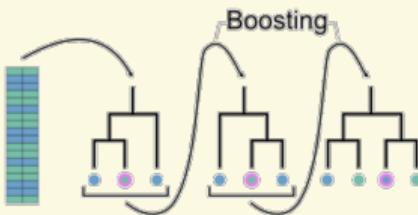
1

0

1

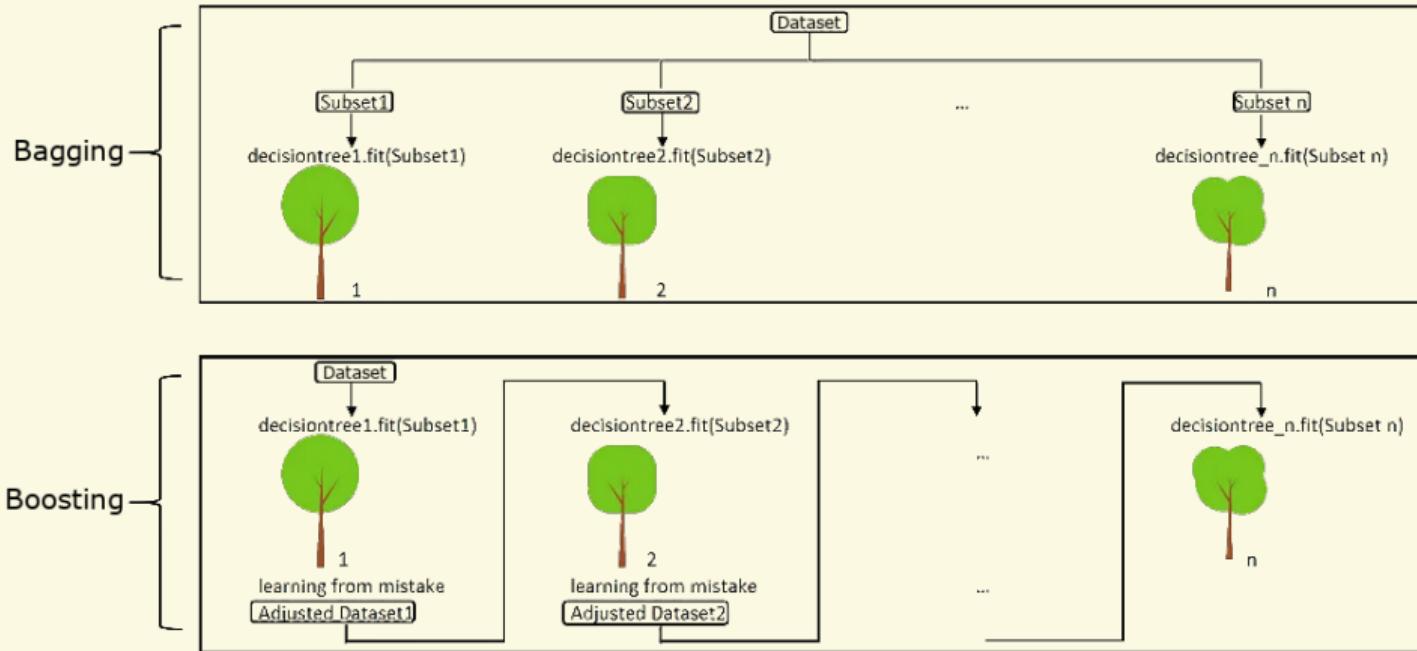
1

MACHINE LEARNING MODEL OVERVIEWS

| Machine learning model | Description | Data Type | Application areas |
|---|--|--|---|
| Boosted Regression Trees:  | <p>N trees are fitted sequentially to minimize an overall loss function</p> <ul style="list-style-type: none">+ flexible(+) variable importance- many hyper-parameters- high complexity | <p>Tabular Data:</p> <ul style="list-style-type: none">- Classification- Regression | <p>functional traits decision making ecological contexts invasive ecosystem species distribution mortality biodiversity remote sensing extinction species distribution species interaction</p> |

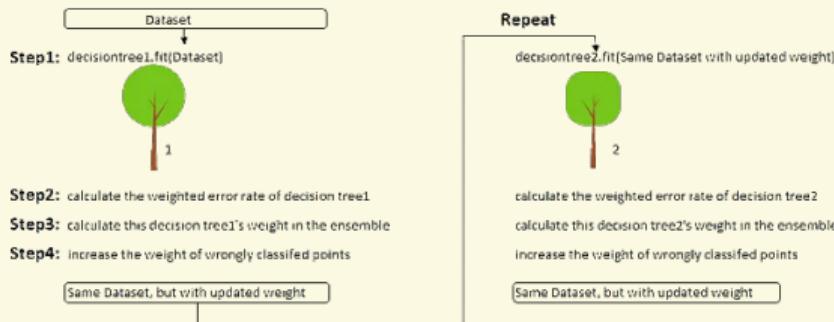
Pichler, M., & Hartig, F. (2023). Machine learning and deep learning - A review for ecologists. *Methods in Ecology and Evolution*, 14, 994–1016.

RANDOM FOREST : BOOSTING VS BAGGING

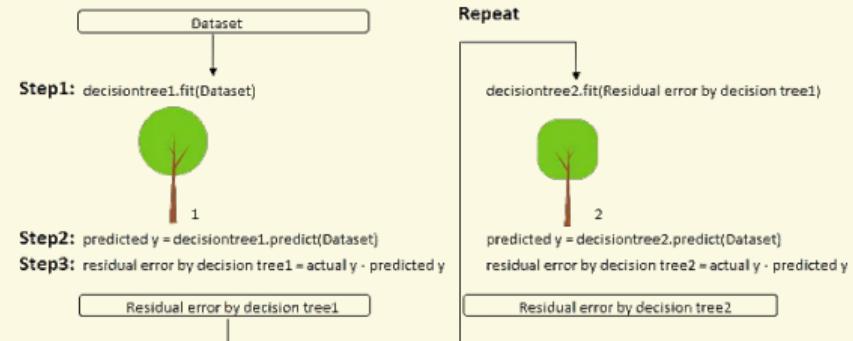


ADAPTIVE VS. GRADIENT BOOSTING

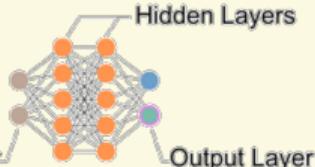
Adaptive Boosting:



Gradient Boosting:



MACHINE LEARNING MODEL OVERVIEWS

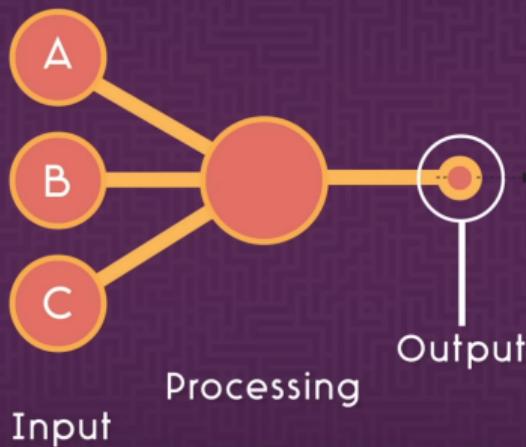
| Machine learning model | Description | Data Type | Application areas |
|--|---|--|---|
| Deep Neural Networks:  | <p>Input (features) are passed through many hidden layers. Last layer maps into response space</p> <ul style="list-style-type: none">+ flexible+ adaptive to different tasks- many hyper-parameters- computationally expensive | <p>Tabular Data:</p> <ul style="list-style-type: none">- Classification- Regression | <p>species identification</p> <p>biological invasions</p> <p>remote sensing</p> <p>biodiversity</p> <p>mortality</p> <p>invasive ecosystem</p> <p>decision making</p> <p>extinction</p> <p>species distribution</p> |

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

NEURAL NETWORKS

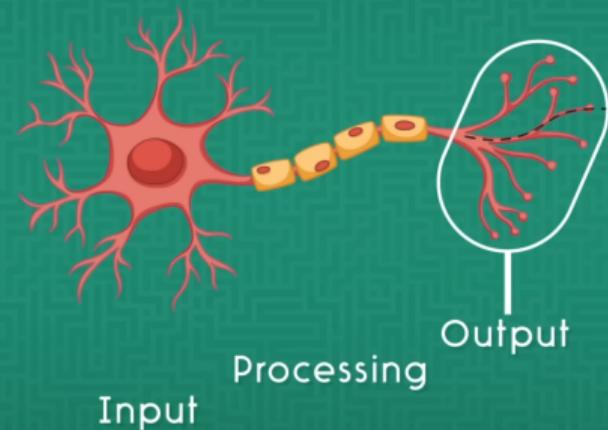
Artificial Neuron

(software)

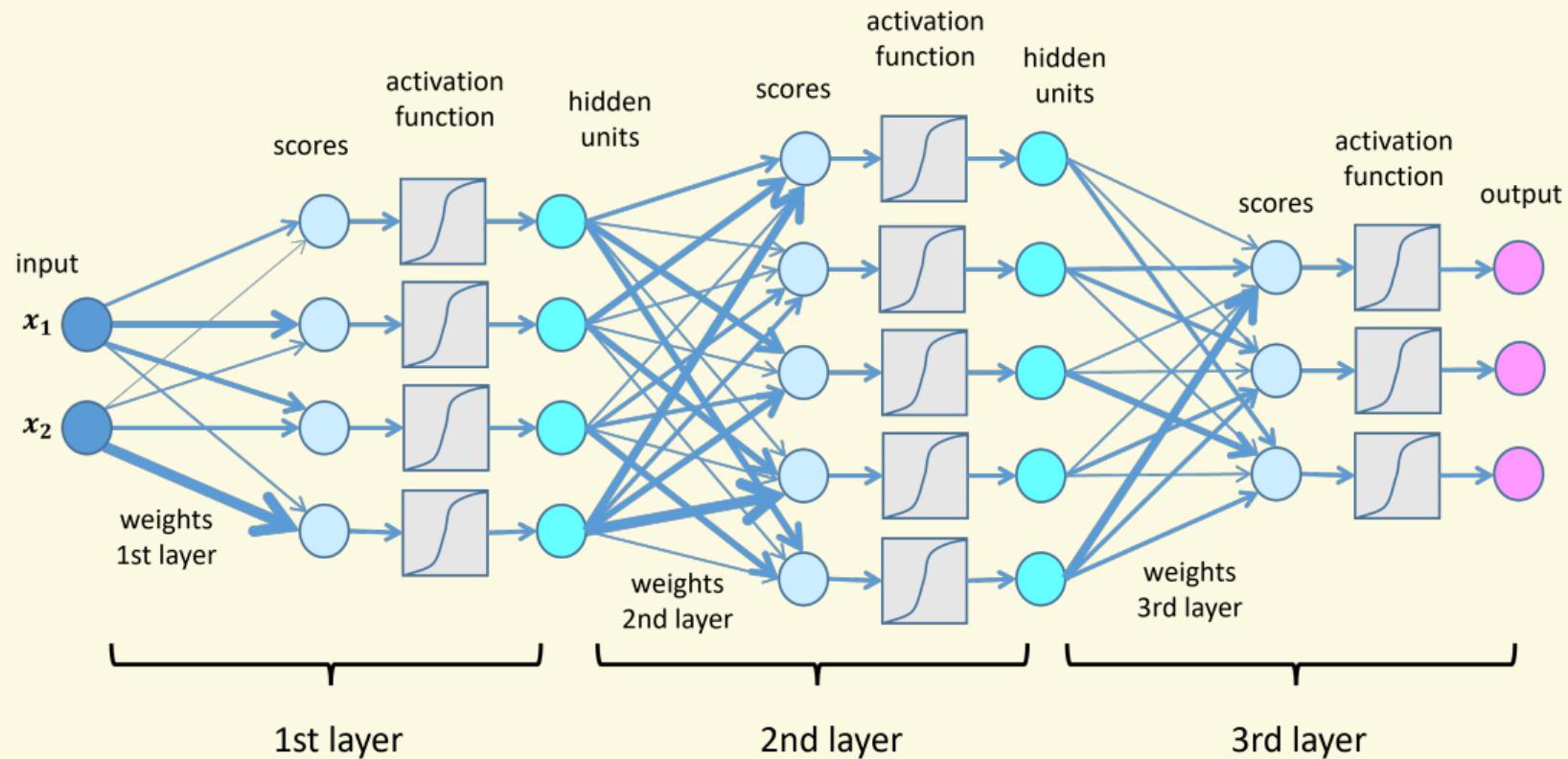


Human Neuron

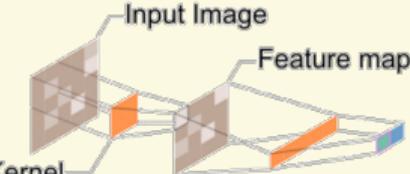
(hardware)



NEURAL NETWORKS

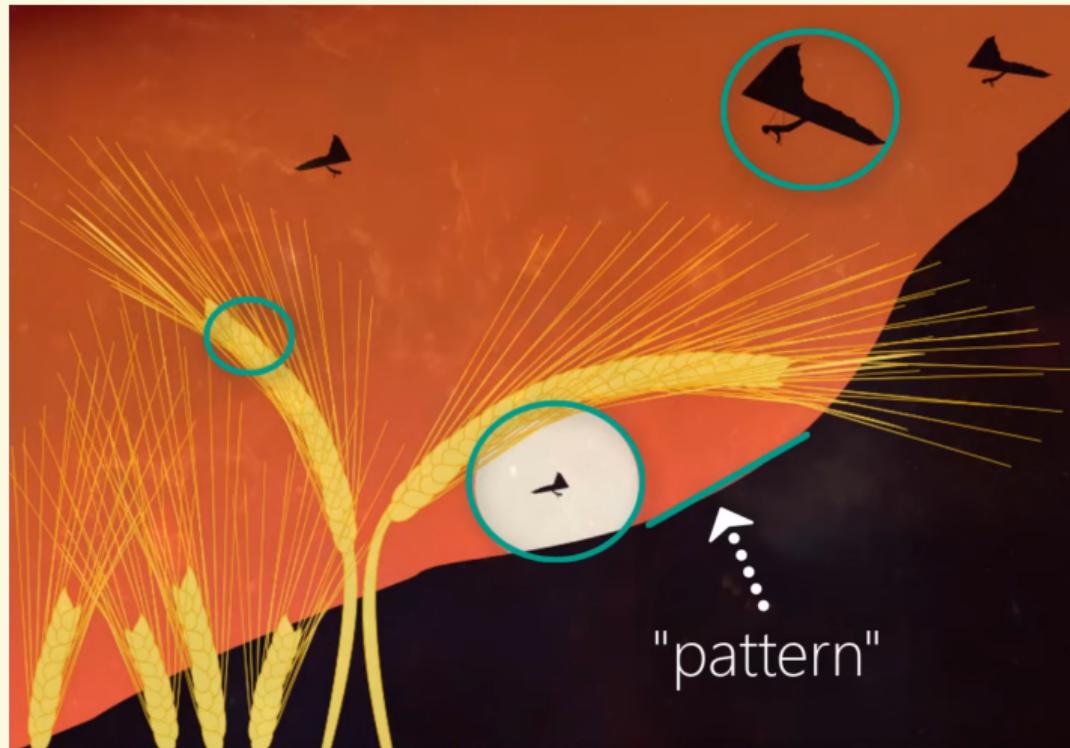


MACHINE LEARNING MODEL OVERVIEWS

| Machine learning model | Description | Data Type | Application areas |
|---|---|--|---|
| <p>Convolutional Neural Networks:</p>  <p>Input Image</p> <p>Kernel</p> <p>Feature map</p> | <p>Small kernels (filters) processes images before passing it to fully connected layers</p> <ul style="list-style-type: none">+ flexible+ detecting shapes and edges- many hyper-parameters- computationally expensive | <p>Images:</p> <ul style="list-style-type: none">- Classification- Object detection | <p>functional trait extinction decision making remote sensing invasive mortality species identification species distribution ecosystem biodiversity camera trap species detection</p> |

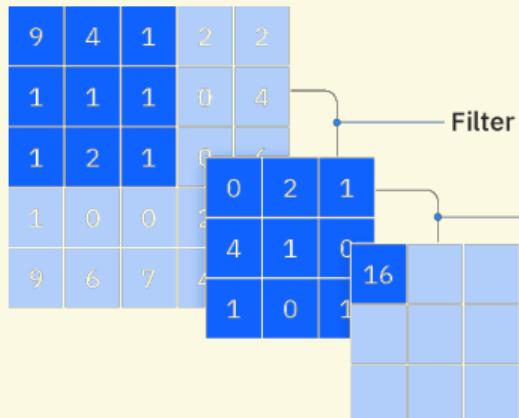
Pichler, M., & Hartig, F. (2023). Machine learning and deep learning - A review for ecologists. *Methods in Ecology and Evolution*, 14, 994–1016.

CONVOLUTIONAL NEURAL NETWORKS



CONVOLUTIONAL NEURAL NETWORKS

Input image



CONVOLUTIONAL NEURAL NETWORKS

The screenshot shows a Microsoft Excel spreadsheet titled "conv-example". The data is organized into several columns:

- Input:** A column of binary values (0.0 or 1.0) representing input features.
- Conv1:** A column of numerical values representing the output of the first convolutional layer.
- filter 1/1:** A column of numerical values representing the output of the first filter in the first convolutional layer.
- filter 2/1:** A column of numerical values representing the output of the second filter in the first convolutional layer.
- filter 3/1:** A column of numerical values representing the output of the third filter in the first convolutional layer.
- filter 4/1:** A column of numerical values representing the output of the fourth filter in the first convolutional layer.
- filter 5/1:** A column of numerical values representing the output of the fifth filter in the first convolutional layer.
- filter 6/1:** A column of numerical values representing the output of the sixth filter in the first convolutional layer.
- filter 7/1:** A column of numerical values representing the output of the seventh filter in the first convolutional layer.
- filter 8/1:** A column of numerical values representing the output of the eighth filter in the first convolutional layer.
- filter 9/1:** A column of numerical values representing the output of the ninth filter in the first convolutional layer.
- filter 10/1:** A column of numerical values representing the output of the tenth filter in the first convolutional layer.
- filter 11/1:** A column of numerical values representing the output of the eleventh filter in the first convolutional layer.
- filter 12/1:** A column of numerical values representing the output of the twelfth filter in the first convolutional layer.
- filter 13/1:** A column of numerical values representing the output of the thirteenth filter in the first convolutional layer.
- filter 14/1:** A column of numerical values representing the output of the fourteenth filter in the first convolutional layer.
- filter 15/1:** A column of numerical values representing the output of the fifteenth filter in the first convolutional layer.
- filter 16/1:** A column of numerical values representing the output of the sixteenth filter in the first convolutional layer.
- filter 17/1:** A column of numerical values representing the output of the seventeenth filter in the first convolutional layer.
- filter 18/1:** A column of numerical values representing the output of the eighteenth filter in the first convolutional layer.
- filter 19/1:** A column of numerical values representing the output of the nineteenth filter in the first convolutional layer.
- filter 20/1:** A column of numerical values representing the output of the twentieth filter in the first convolutional layer.
- filter 21/1:** A column of numerical values representing the output of the twenty-first filter in the first convolutional layer.
- filter 22/1:** A column of numerical values representing the output of the twenty-second filter in the first convolutional layer.
- filter 23/1:** A column of numerical values representing the output of the twenty-third filter in the first convolutional layer.
- filter 24/1:** A column of numerical values representing the output of the twenty-fourth filter in the first convolutional layer.
- filter 25/1:** A column of numerical values representing the output of the twenty-fifth filter in the first convolutional layer.
- filter 26/1:** A column of numerical values representing the output of the twenty-sixth filter in the first convolutional layer.
- filter 27/1:** A column of numerical values representing the output of the twenty-seventh filter in the first convolutional layer.
- filter 28/1:** A column of numerical values representing the output of the twenty-eighth filter in the first convolutional layer.
- filter 29/1:** A column of numerical values representing the output of the twenty-ninth filter in the first convolutional layer.

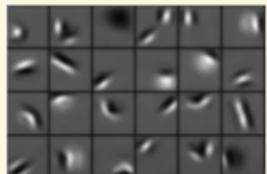
The Excel interface includes standard ribbon tabs (File, Home, Insert, Page Layout, Formulas, Data, Review, View) and various toolbars for cell selection, text alignment, and data manipulation.

CONVOLUTIONAL NEURAL NETWORKS



CONVOLUTIONAL NEURAL NETWORKS

Low-level features



Edges, dark spots

Mid-level features



Eyes, ears, nose

High-level features



Facial structure

Faces



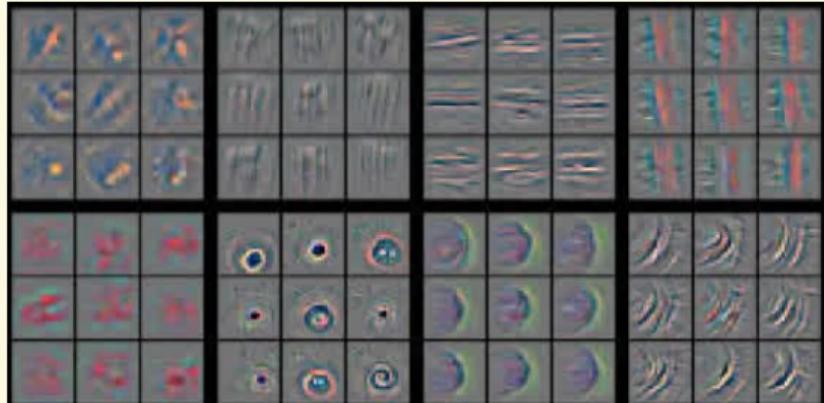
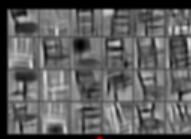
Cars



Elephants



Chairs



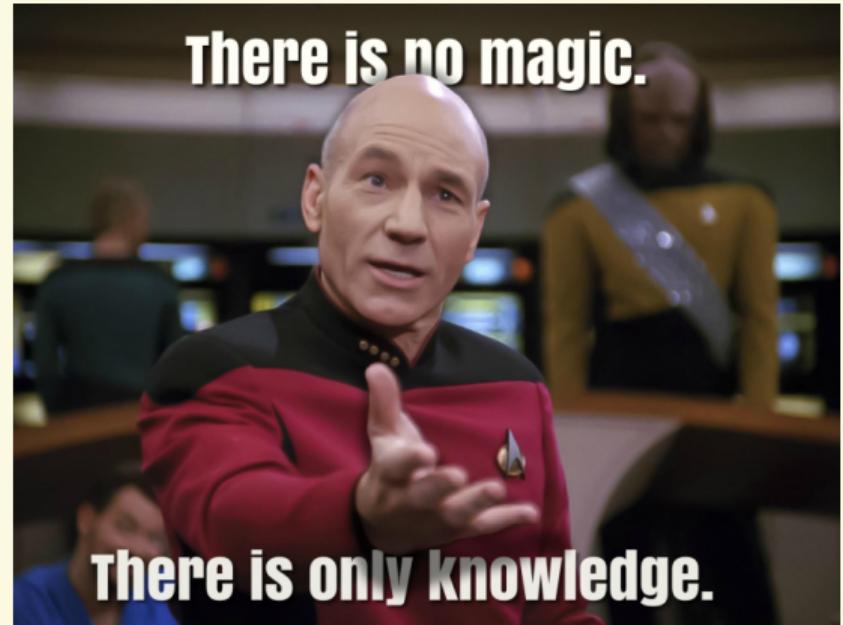
CONVOLUTIONAL NEURAL NETWORKS



THERE IS NO MAGIC...

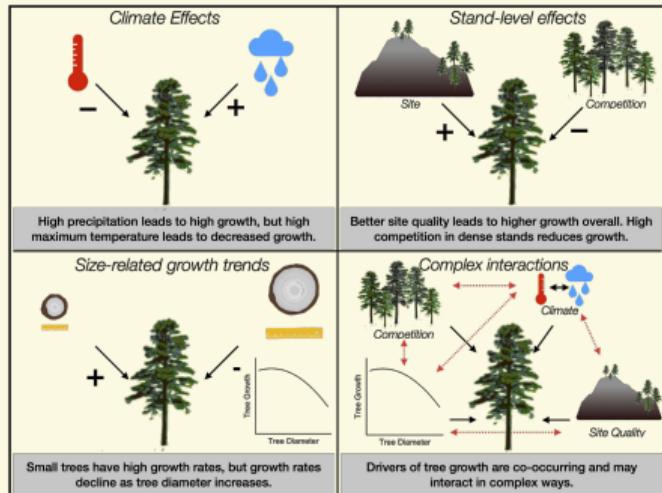
There is no magic here, no one model is always better than another.

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

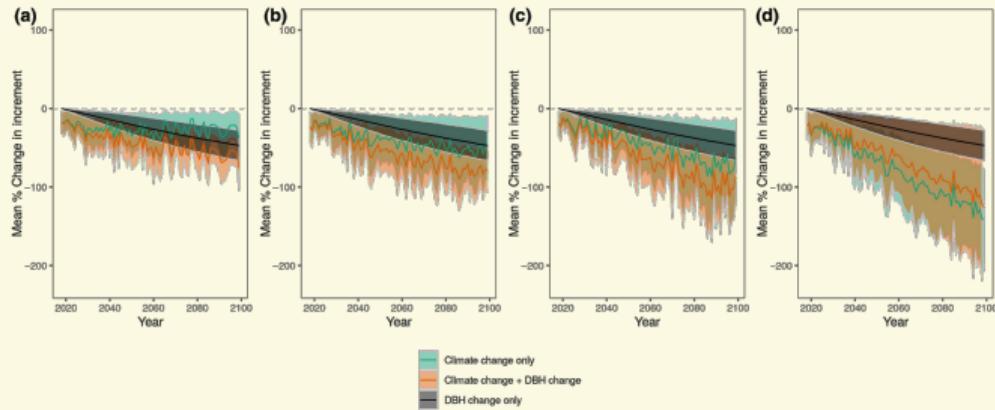


PICKING A MODEL: BIG PICTURE GOALS?

Inference: Explanatory Power

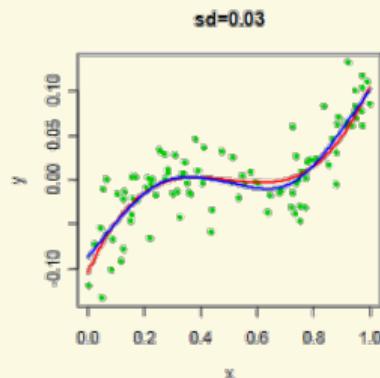
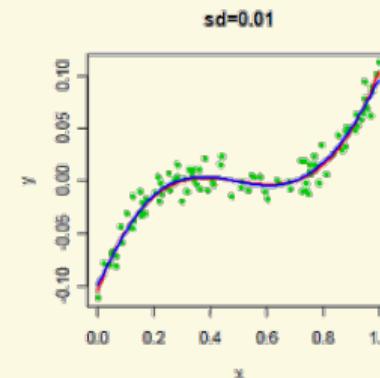
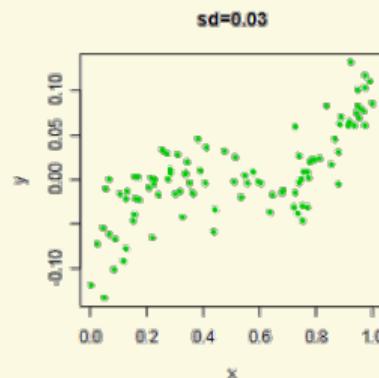
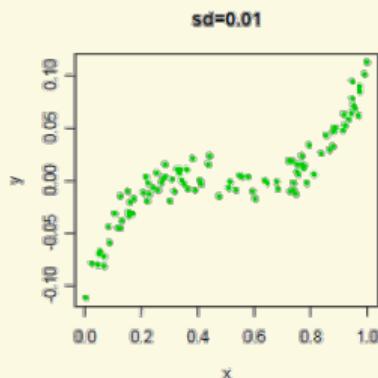
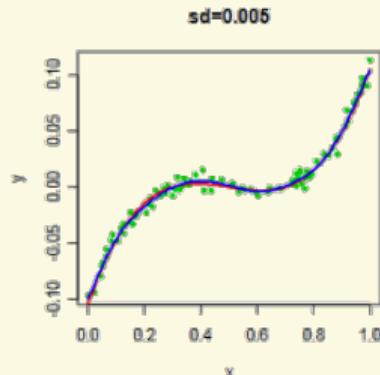
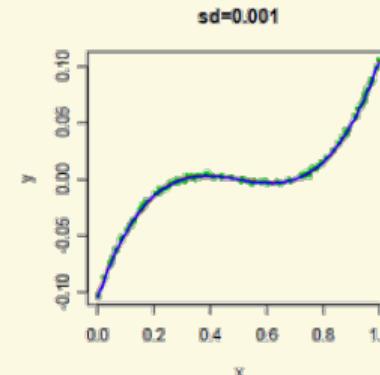
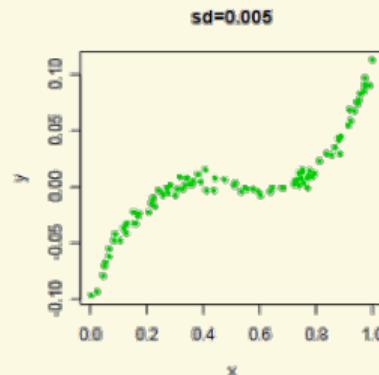
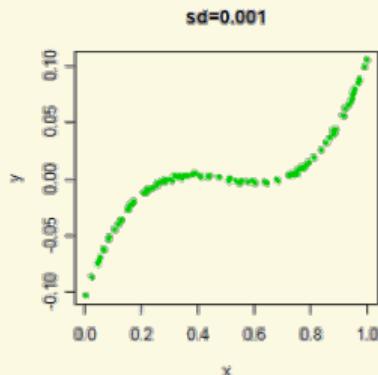


Prediction: Predictive Power

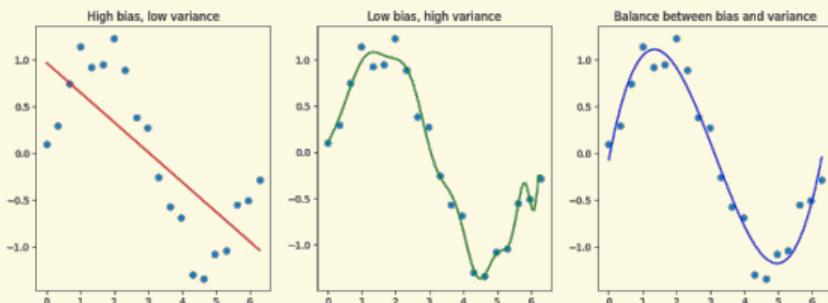


Heilman et. al. (2022). Ecological forecasting of tree growth: Regional fusion of tree-ring and forest inventory data to quantify drivers and characterize uncertainty. *Global Change Biology*, 28, 24422460.

PICKING A MODEL: BIAS & OVERFITTING



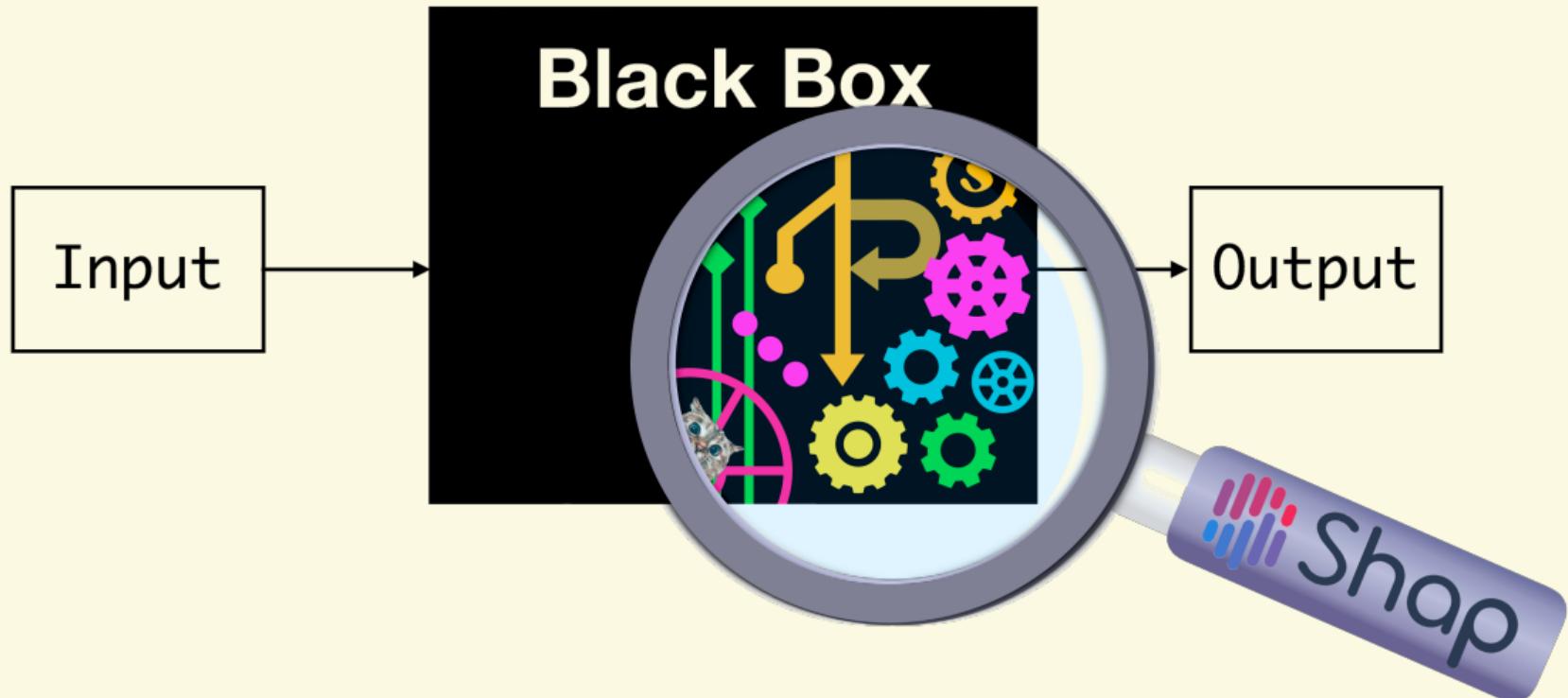
PICKING A MODEL: BIAS & OVERFITTING



Bias-Variance Tradeoff

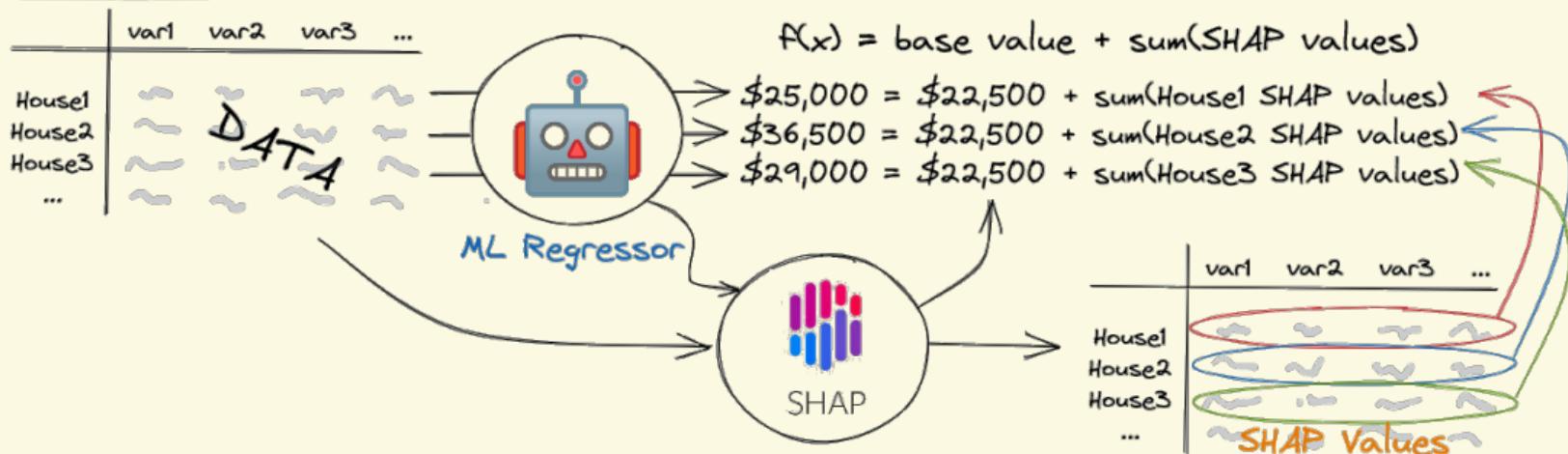
- Model is too simple, it will have very few parameters then it may have high bias and low variance.
- On the other hand if the model has large number of parameters then its going to have high variance and low bias.
- Our job is to strike the correct balance.

INTERPRETABLE MACHINE LEARNING



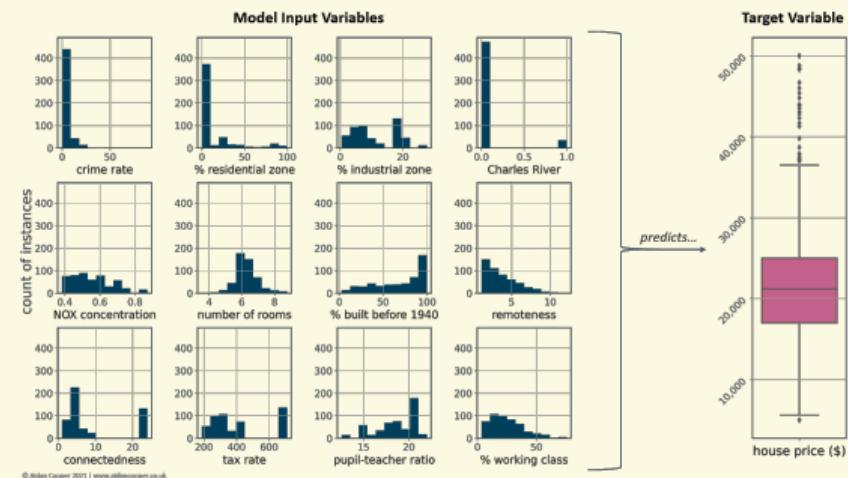
SHAP (SHAPLEY ADDITIVE exPLANATIONS)

Regression:



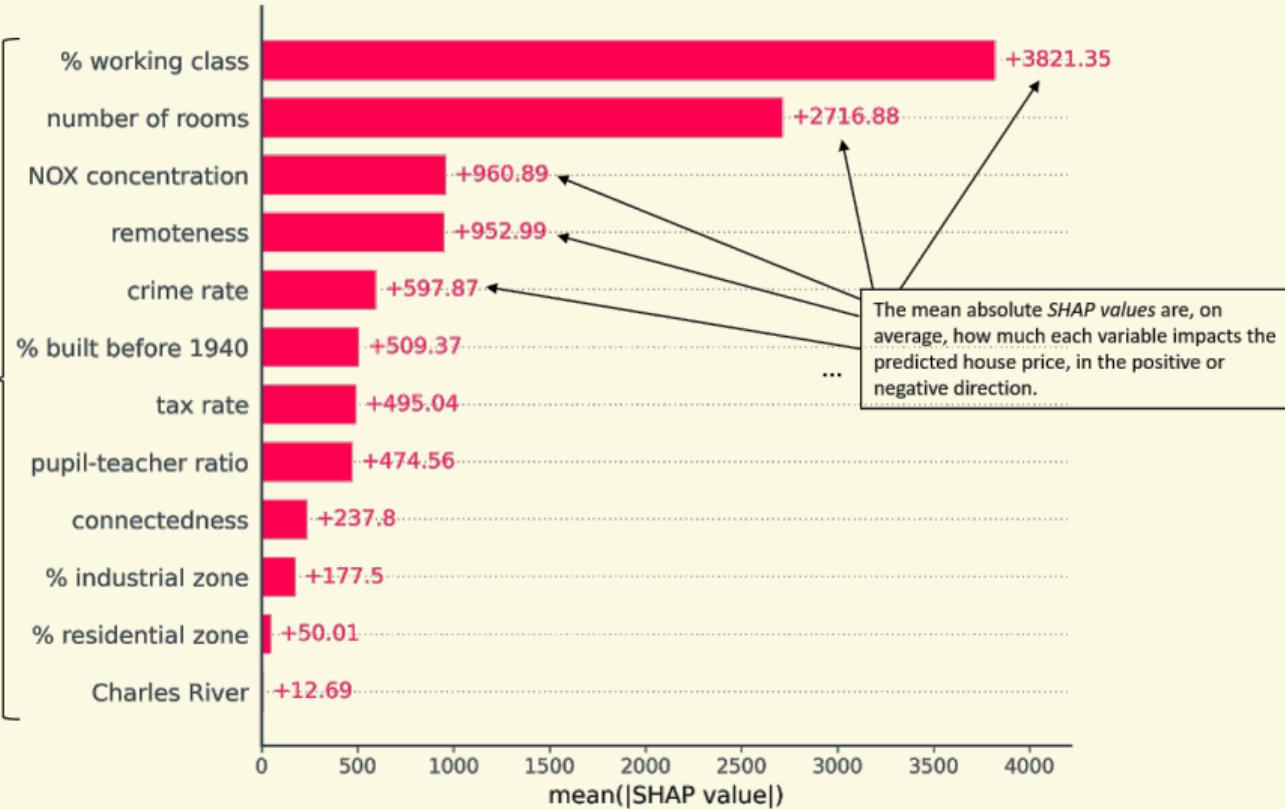
SHAP (SHAPLEY ADDITIVE exPLANATIONS)

| 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)

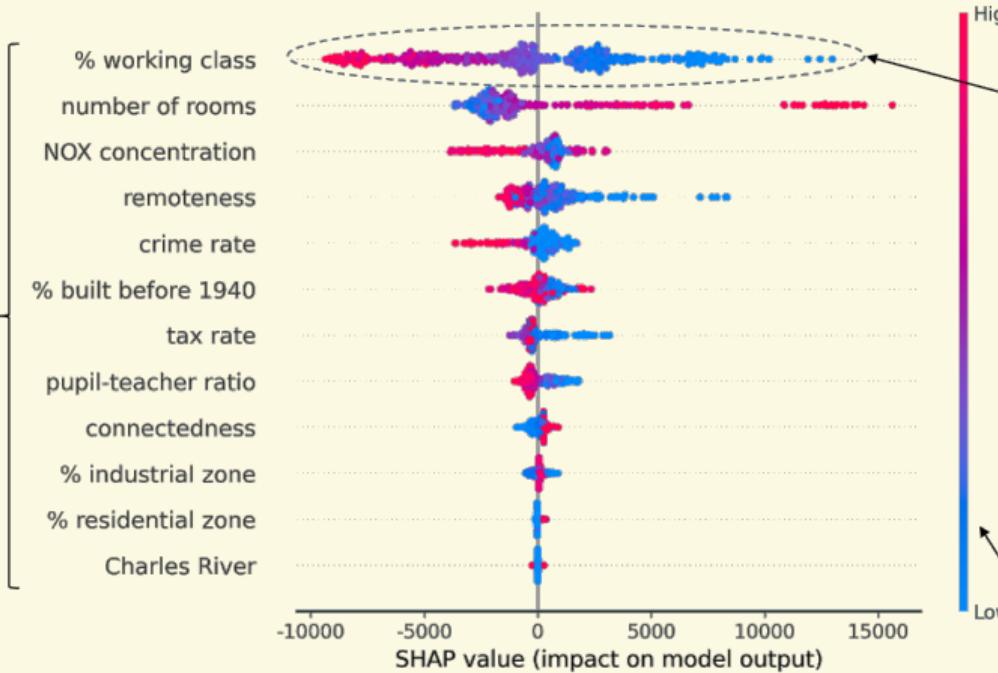
These are the input variables, ranked from top to bottom by their mean absolute SHAP values for the entire dataset – i.e. the average magnitude of each variable's impact on the predicted house price across all instances.



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 its 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.

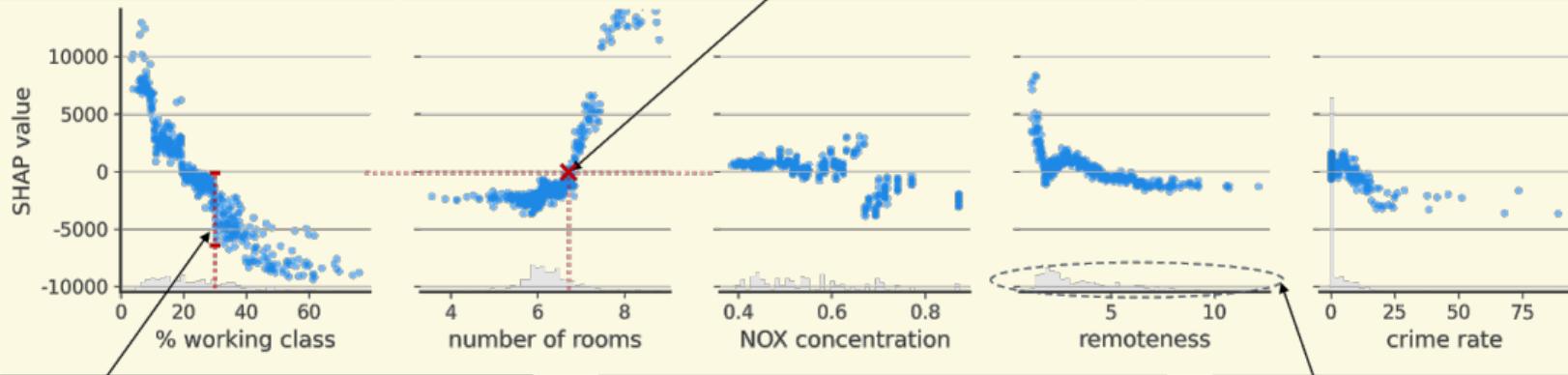
Examining the colour distribution horizontally along the x-axis for each variable provides insights into the general relationship between a variable's raw values and its SHAP values.

SHAP (SHAPLEY ADDITIVE exPLANATIONS)

In a **dependence plot**, every instance (i.e. row) of the dataset appears as its 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 **X**.

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%).

MACHINE LEARNING BASICS : RECIPE

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

