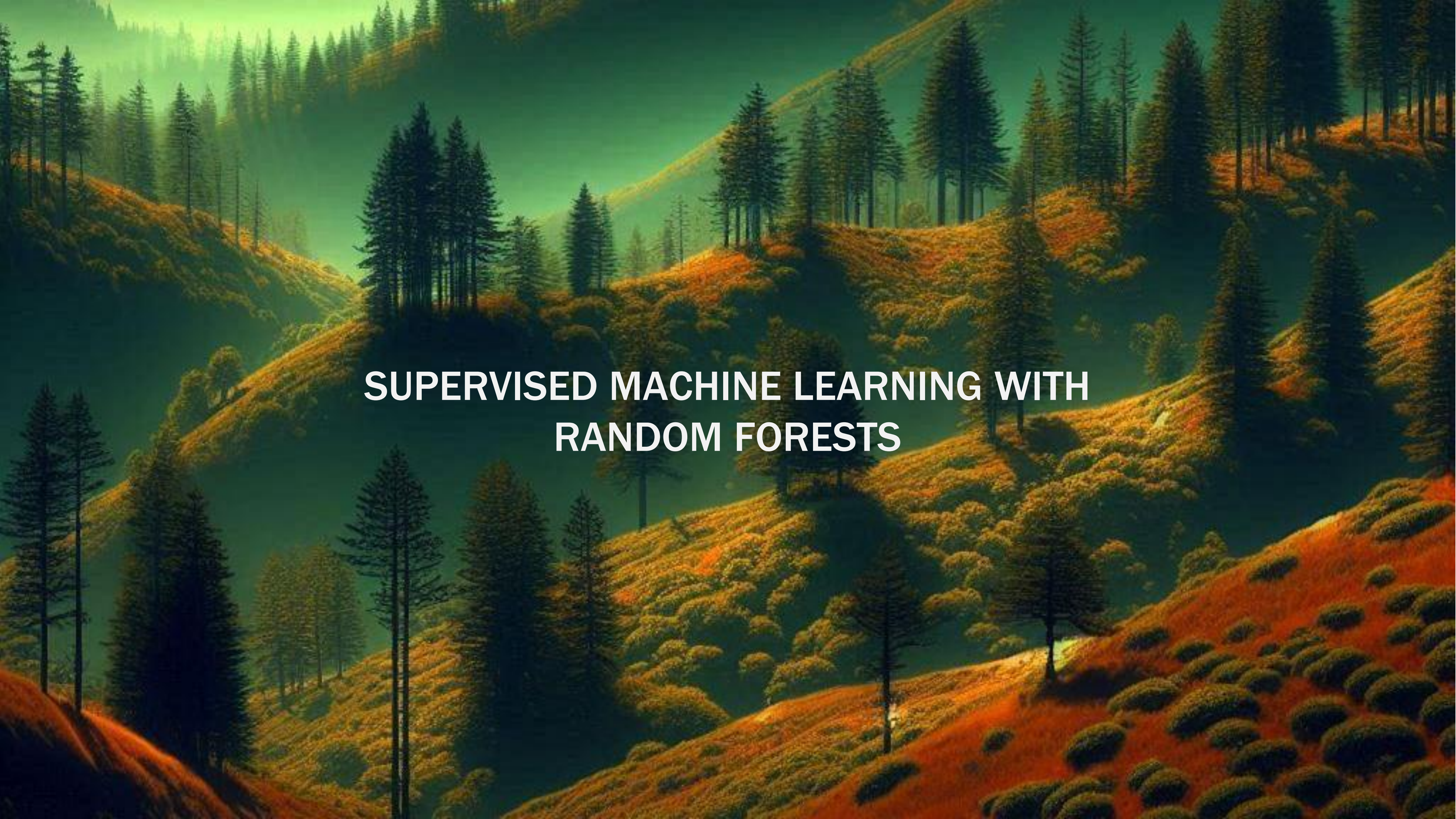


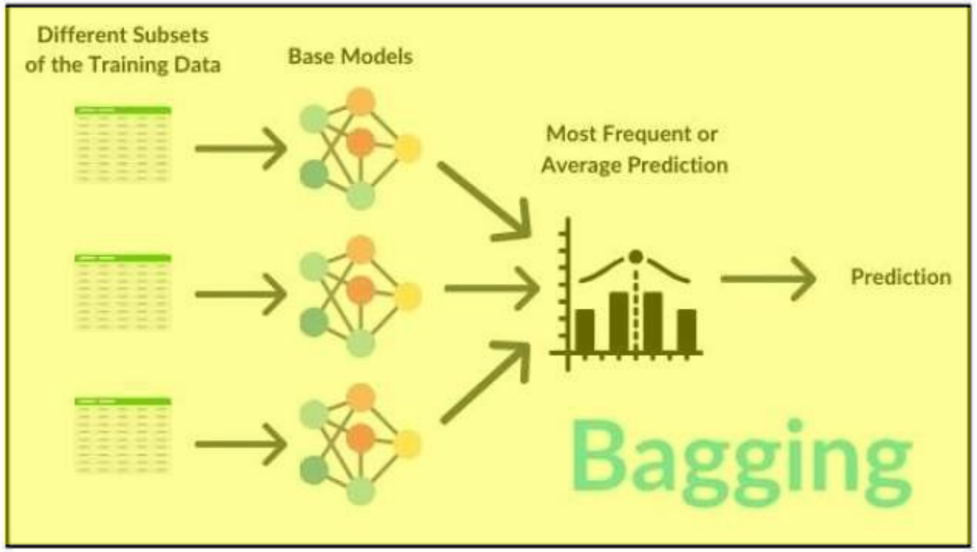
SUPERVISED MACHINE LEARNING WITH RANDOM FORESTS

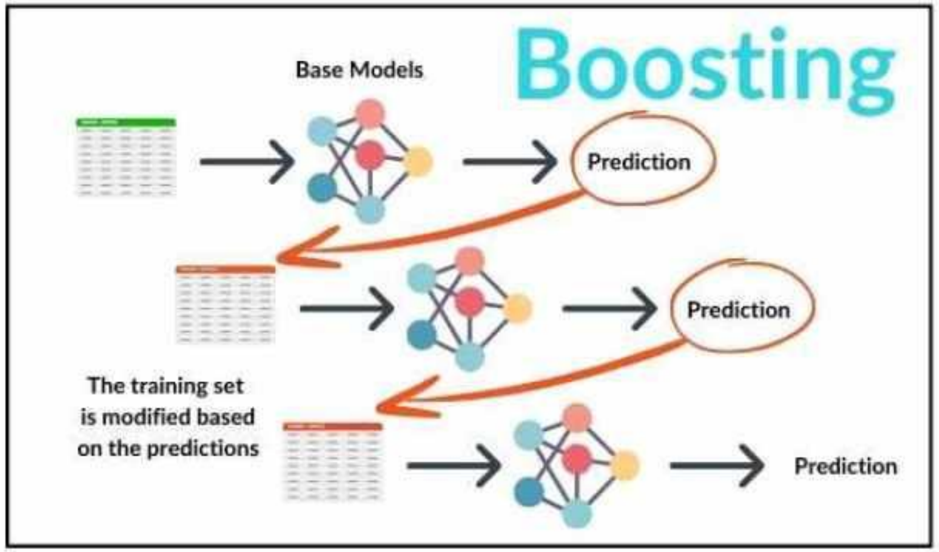
ENSEMBLE LEARNING

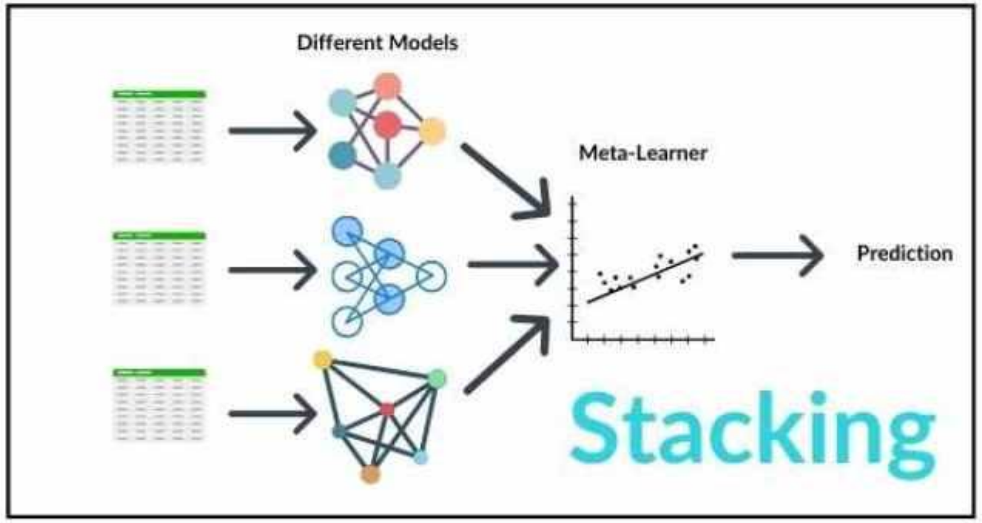
Combines multiple **“weak learners”** to create a stronger model that reduces error and improves generalization.

- **Bagging:** Trains models on bootstrapped samples; combines predictions by averaging or voting → reduces variance, prevents overfitting.
- **Boosting:** Sequentially corrects errors of prior models → reduces bias and variance for high accuracy.
- **Stacking:** Combines diverse models using a **meta-model** that learns optimal blending → leverages strengths of all models.

FLAVORS OF ENSEMBLE LEARNING: BAGGING, BOOSTING, STACKING







Random Forest

- An ensemble of randomized CART decision trees.
- Combining trees beneficial if uncorrelated.

Recall: CART Limitations

CART

RANDOM FORESTS

Random Forest Randomizations

BOOTSTRAPPING

- For each tree, data is sampled with replacement
- Sampled data: In-bag
- Otherwise: Out-of-bag
- P(In-Bag) -->

RANDOM FEATURE SELECTION

- At each split, *m* variables considered
- For classification: *m = sqrt(d)*
- For regression: *m = (1/3)d*

$$1 - (1 - \frac{1}{n})^n \rightarrow 1 - \frac{1}{e} \approx 0.632$$

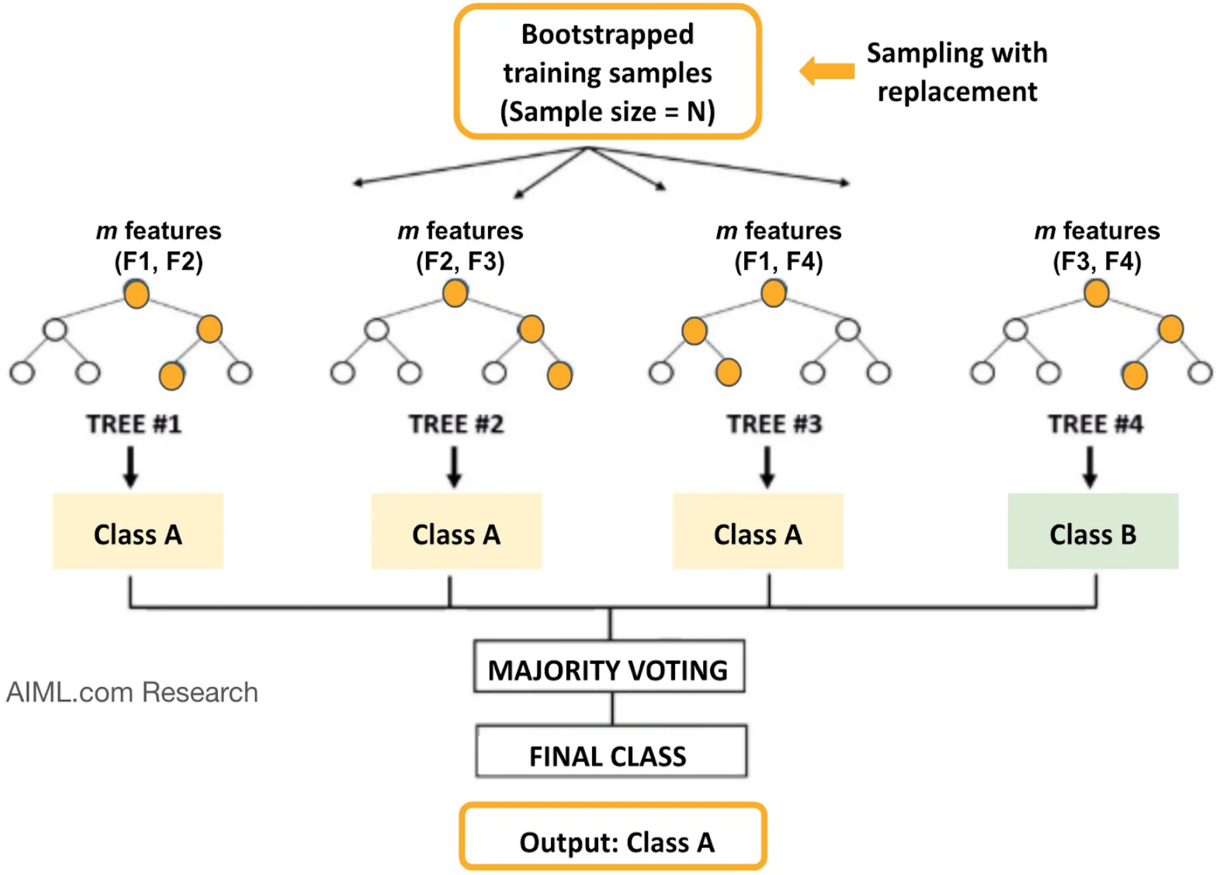
Out-of-Bag Points

- OOB pts. unseen by about 37% of trees
- Serval as an internal validation set.
- Used to estimate generalization error.
- Can serve as cross-validation for hyperparameter tuning

Random Forest Predictions

- Each tree provides one vote
- Unweighted majority voting
- Averaging for regression

Random Forest Classifier



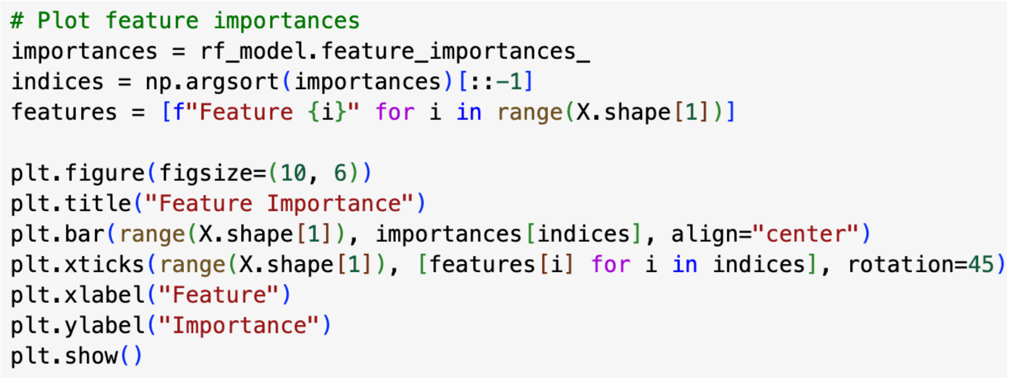
© AIML.com Research

FEATURE IMPORTANCE

Two main methods to assess feature impact on predictions:

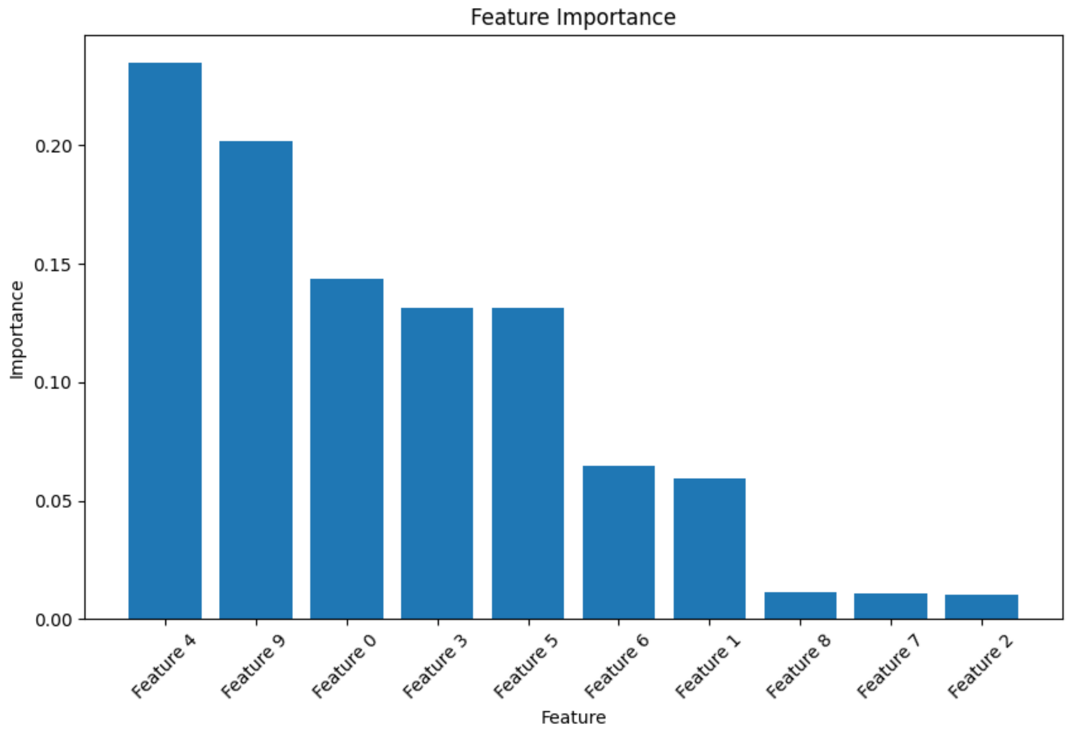
- **1. Mean Decrease in Impurity (MDI)**
  - Default method; measures how much each feature reduces impurity (Gini/entropy).
  - **Steps:** compute impurity reduction → sum across trees → normalize.
  - **Bias:** favors features with many categories or continuous values.

- **2. Permutation Importance**
  - Shuffles feature values to see how performance drops.
  - **Steps:** measure baseline → shuffle feature → re-evaluate → compute drop.
  - **Interpretation:** larger performance drop = more important feature.
  - Model-agnostic and less biased than MDI.



```
# Plot feature importances
importances = rf_model.feature_importances_
indices = np.argsort(importances)[::-1]
features = [f"Feature {i}" for i in range(X.shape[1])]

plt.figure(figsize=(10, 6))
plt.title("Feature Importance")
plt.bar(range(X.shape[1]), importances[indices], align="center")
plt.xticks(range(X.shape[1]), [features[i] for i in indices], rotation=45)
plt.xlabel("Feature")
plt.ylabel("Importance")
plt.show()
```



Random Forest Proximities

“[Proximities] are one of the most useful tools in random forests.”
--Leo Breiman

Original description:

“The proximities originally formed a NxN matrix. After a tree is grown, put all of the data, both training and oob, down the tree. If cases k and n are in the same terminal node increase their proximity by one. At the end, normalize the proximities by dividing by the number of trees.”

https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#:~:text=few%20data%20sets.-,Proximities,by%20the%20number%20of%20trees.

Proximities in a Nutshell

“Proximities don’t just measure the similarity of the variables---they also take into account the importance of the variables.”

“Two cases that have quite different predictor variables might have large proximity if they differ only on variables that are not important.”

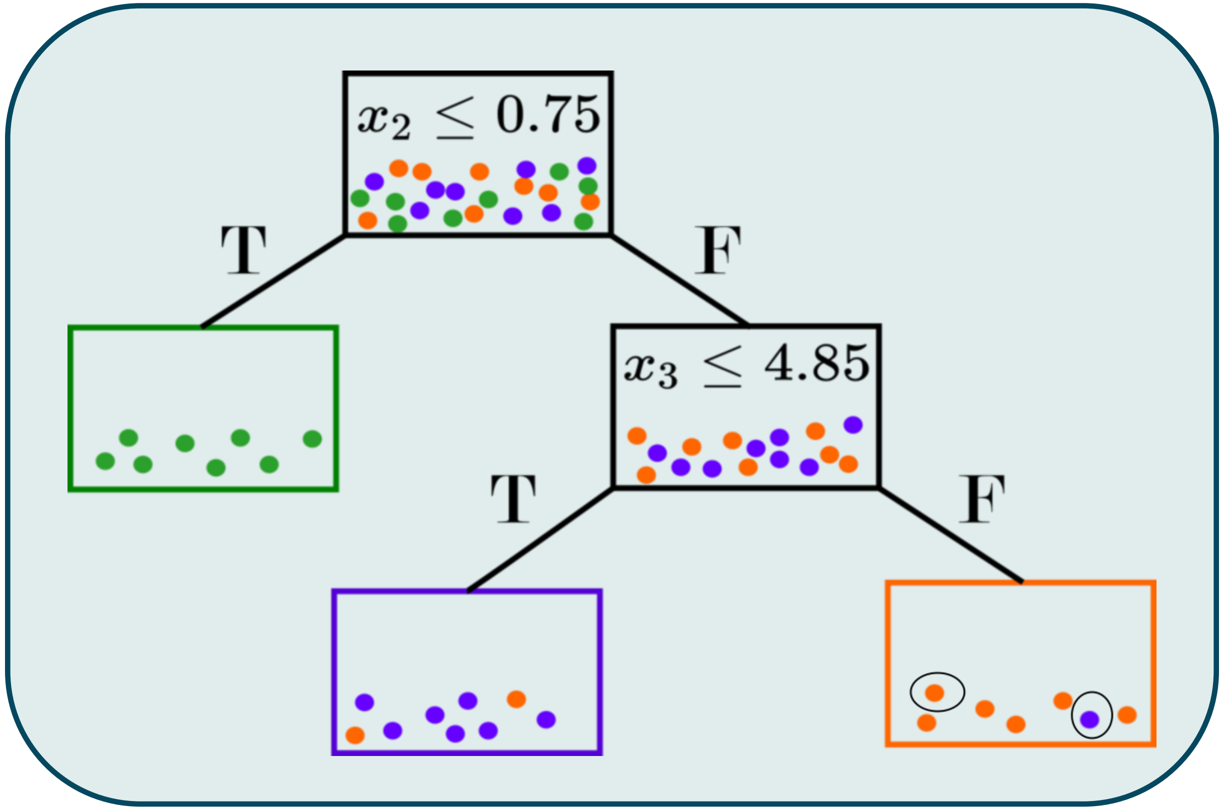
“Two cases that have quite similar values of the predictor variables might have small proximity if they differ on inputs that are important.”

--Adele Cutler

What are Random Forest Proximities

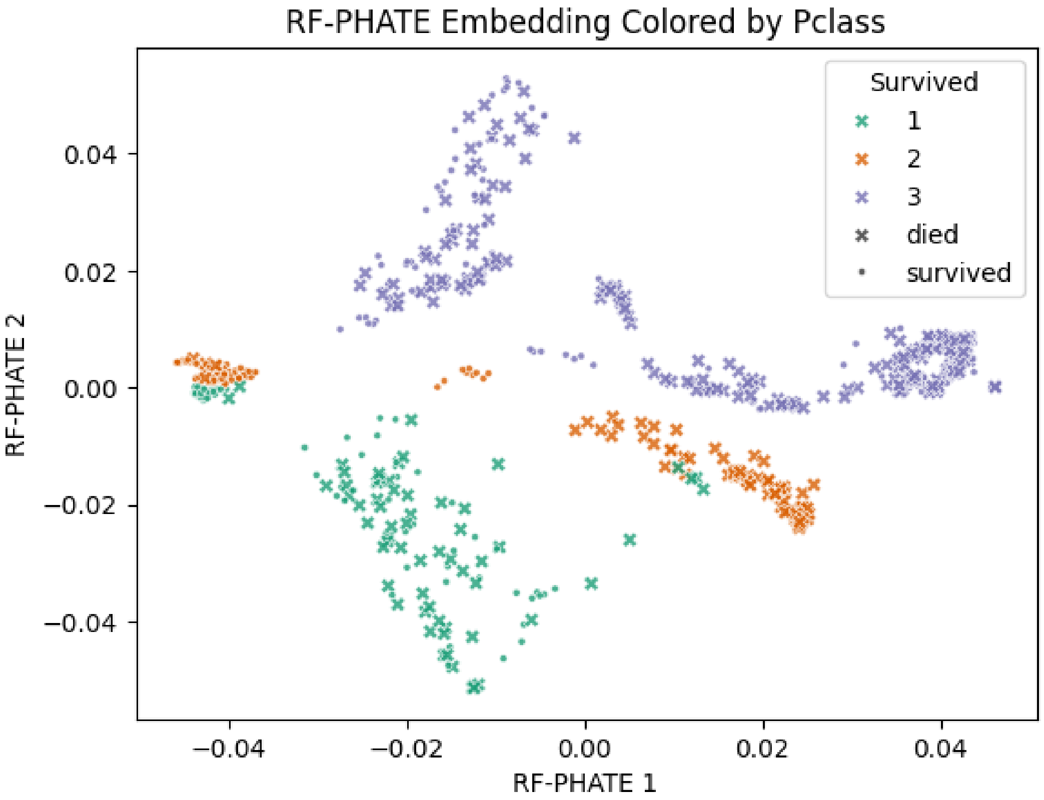
PROXIMITIES AS A METRIC

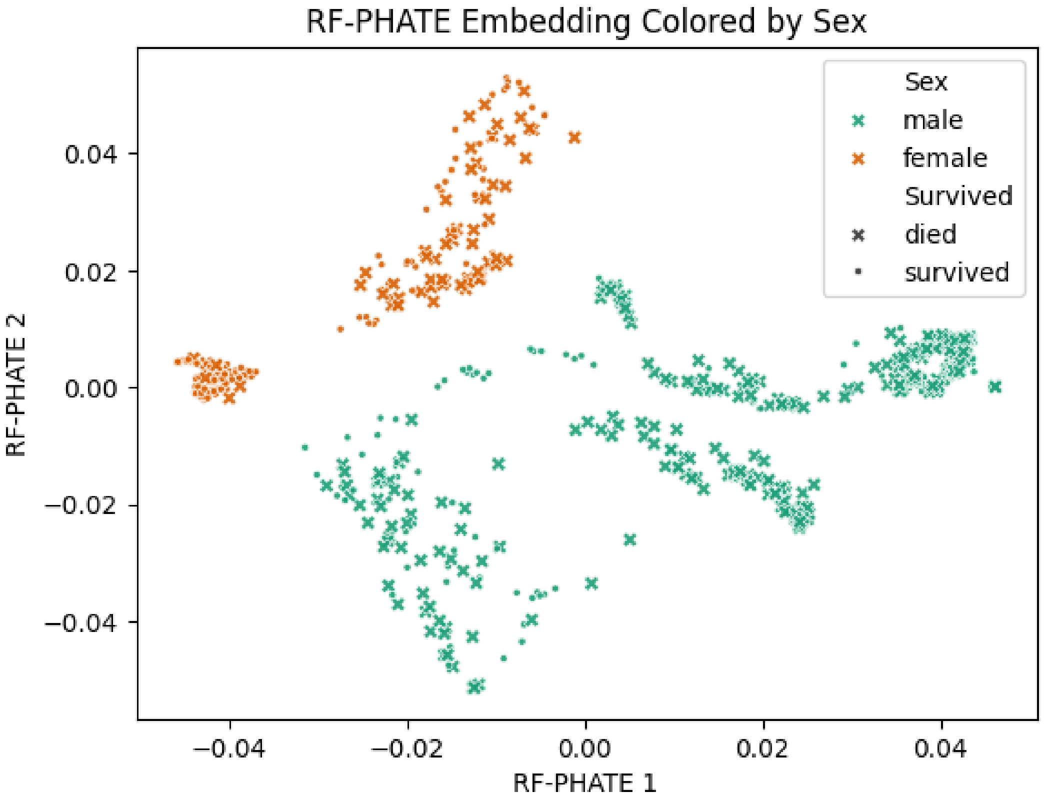
- A measure of “closeness”
- Supervised, but not class-conditional
- Based on splits (optimized over responses)
- Natural incorporation of variable importance



RF Proximities for Visualization

- RF proximities visualizing the Titanic dataset


RF-PHATE Embedding Colored by Pclass


RF-PHATE Embedding Colored by Sex