


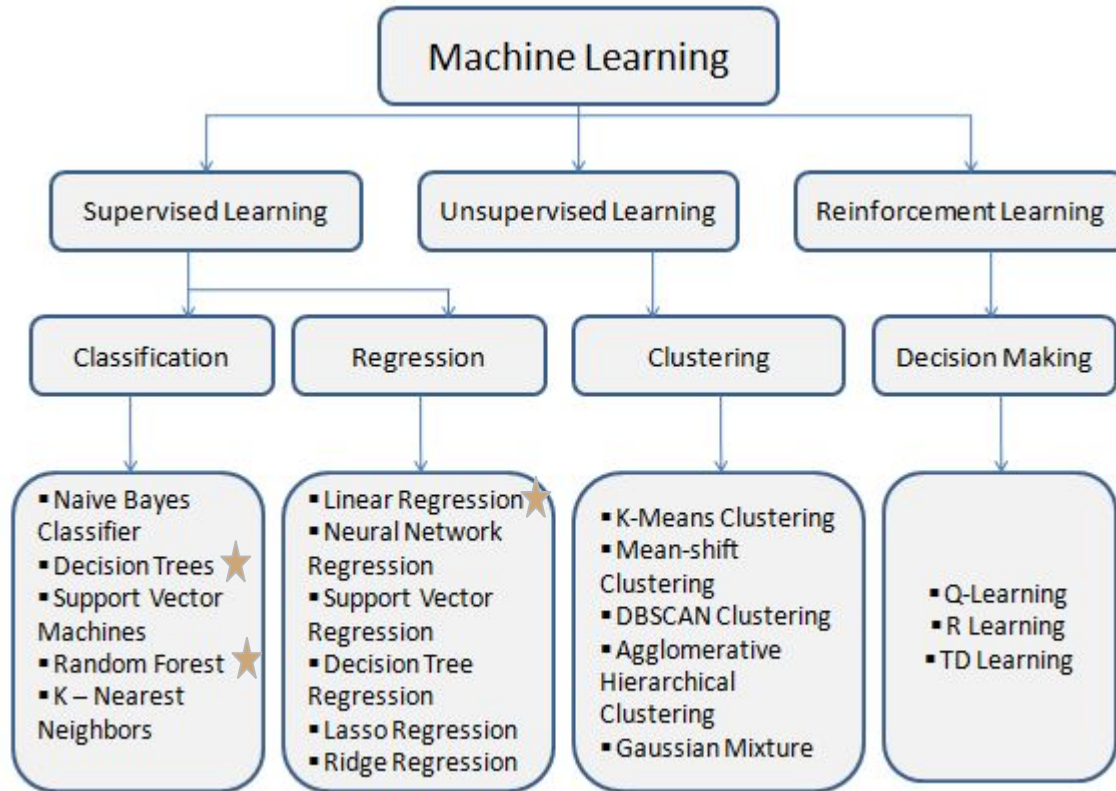
Analysis of the ML-Based Screening Models and their Features



Types of Machine Learning Models Used

- ★ Logistic Regression
- ★ Random Forest
- ★ XGBoost
- ★ Explainable Boost

What shows a model's effectiveness: Sensitivity, specificity, precision, negative predictive value, accuracy, ROC curve, etc.



Explainable ★
boosting machine

<https://www.analyticsvidhya.com/blog/2021/03/everything-you-need-to-know-about-machine-learning/>

Classification vs. Regression

Classification:

- Mapping function from input to get output
- Discrete class labels
- Data needs to have labels first
- Can have both discrete and real-valued variables

Regression:

- Estimating the mapping function
- Quantitative value
- Prediction based on “features”

<https://www.springboard.com/blog/data-science/regression-vs-classification/>

Logistic Regression

Uses the relationship between binary outcome and predictor variables for binary classification

Maximum likelihood

Multivariate logistic regression

- tests to see if a variable's effect on the prediction is significant
- model coefficients

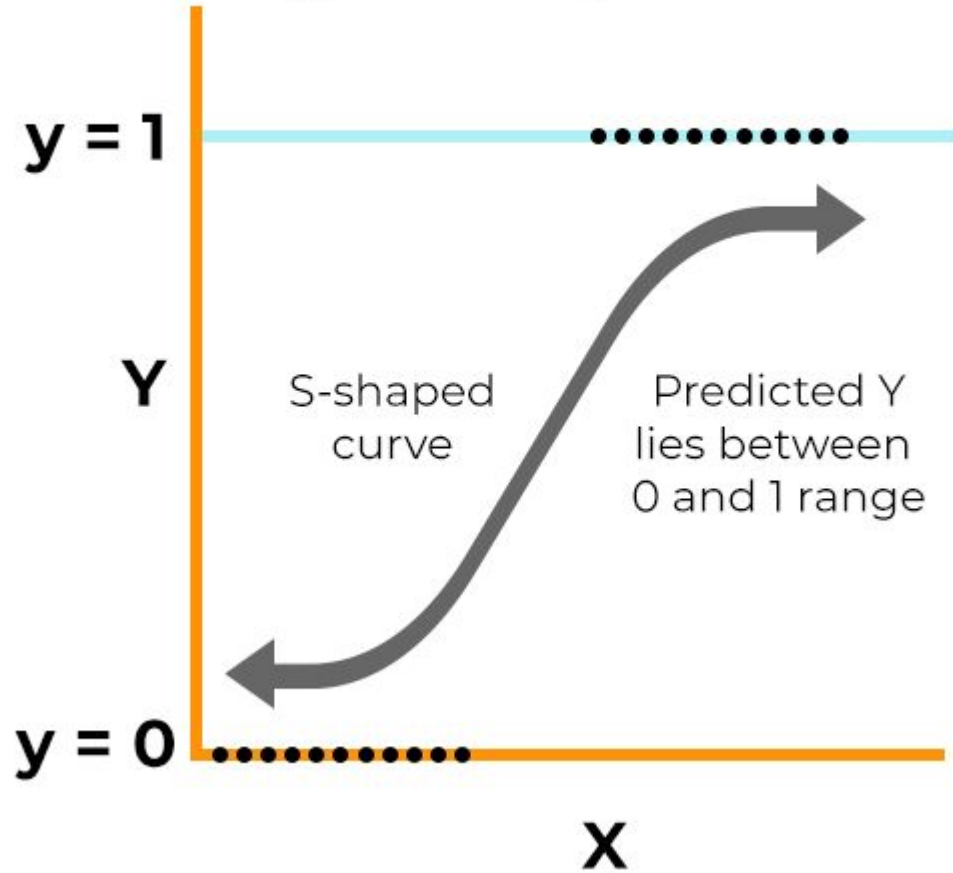
Limitation(s): better for basic relationships

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10569817/>

<https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/>

/

Logistic Regression



<https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/>

Random Forest

Model that maximizes the prediction of multiple decision trees

Factors: Node size, number of trees, number of features

More prominent features = better predictions from trees

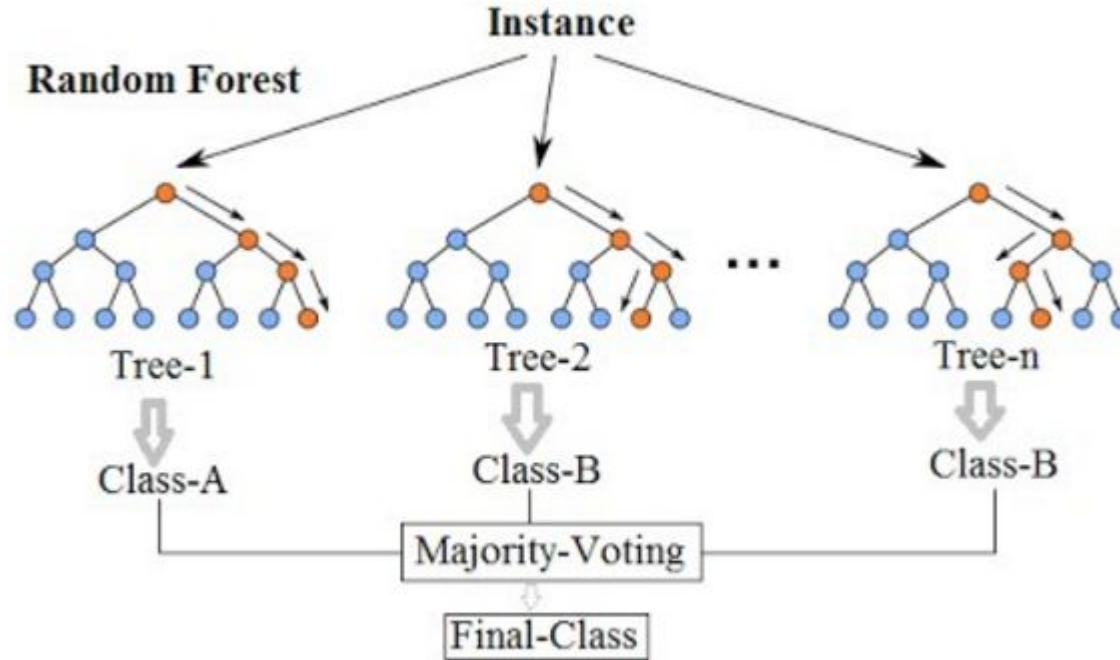
Bootstrapping and Aggregating

Limitation(s): time, amount of resources, difficult to comprehend

<https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8957986/>

Random Forest Simplified



<https://williamkoehrsen.medium.com/random-forest-simple-explanation-377895a60d2d>

XGBoost

A collection of gradient boosted decision trees
Boosting - each tree is improved from the next by using the residual errors from the previous

- Lower the loss function of each tree

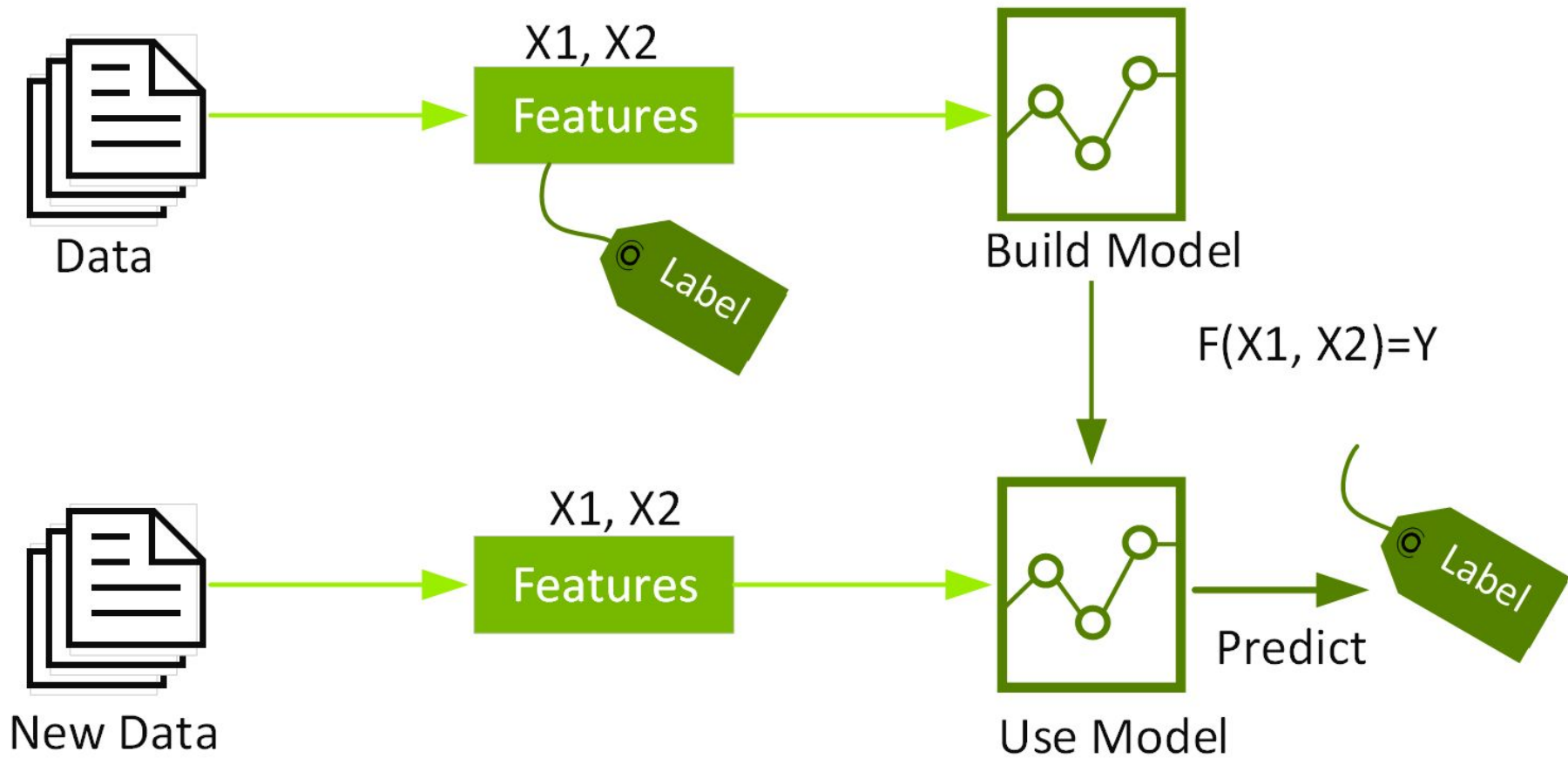
SHAP values = evaluating the features that promote the prediction

- Greater the value, the better the feature

Better for 'non-linear' data

Limitation(s): prone to overfitting, needs structured data, limited to loss data

<https://www.nvidia.com/en-us/glossary/xgboost/>



Explainable Boosting Machine

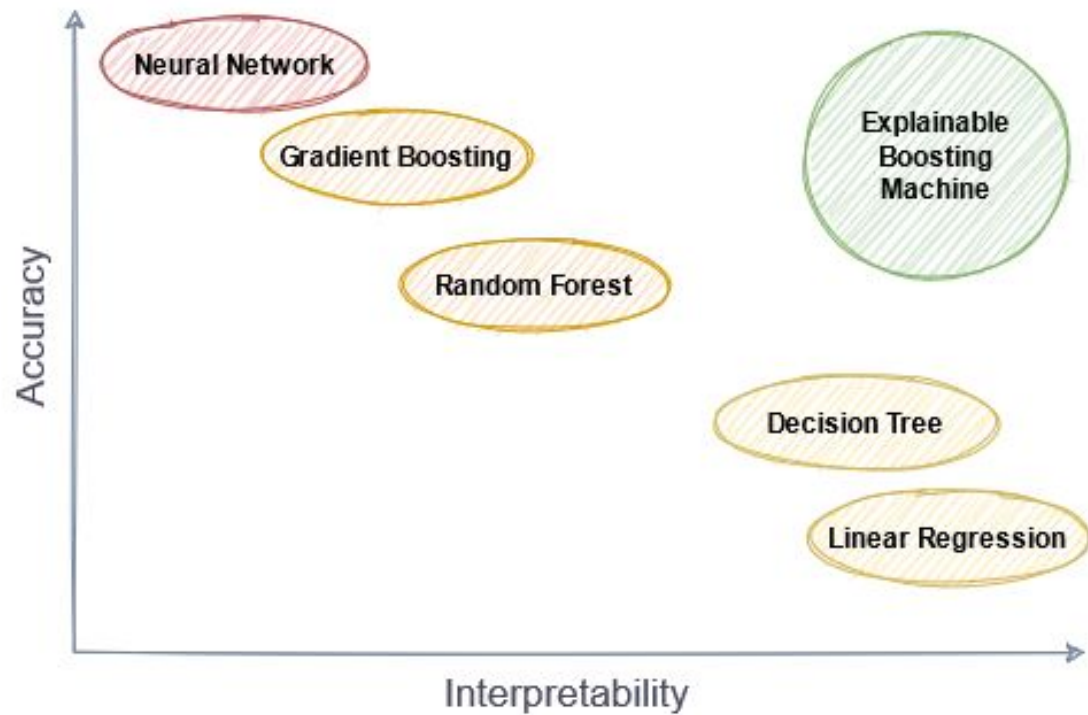
Glassbox tree-based model best for human interpretation

Uses gradient boosting and combines a number of decision trees (similar to XGBoost)

Generates “human-readable” results + provides ‘explanations’ for the predictions

Tests one feature at a time

Limitation(s): only explains the outer surface, not clear



<https://towardsdatascience.com/the-explainable-boosting-machine-f24152509ebb>

CHATGPT Opinion

Explainable Boosting Machine is the best.

- Good balance between predictive performance and interpretability

Random Forest is a close runner-up.

GROUPS 1-6

1: 3 yrs - 2.5 yrs

2: 3 yrs - 2 yrs

3: 3 yrs - 1.5 yrs

4: 3 yrs - 1 yr

5: 3 yrs - 6 months

6: 3 yrs - Date of Diagnosis

Common Features (in both models)

- ★ Platelet mean volume
- ★ Single live birth
- ★ Carbon dioxide
- ★ Glucose
- ★ Lymphocytes
- ★ Body Height
- ★ Glomerular Filtration Rate
- ★ Influenza
- ★ Cytopathology
- ★ Transvaginal echography
- ★ Leiomyoma
- ★ Measles virus

Mean Platelet Volume

Leading feature - found in groups L1 - L5, X1, X3, X4

Measurement of the average size of blood platelets

Higher MPV = bigger platelets = faster platelet circulation = inflammation

Endometriosis is a chronic, inflammatory disease.

Glucose in Serum/Urine/Blood

Serum - L1, X1

- Low = **inflammation**
- High = risk for Type 1 diabetes, connected to higher insulin levels and endo

Urine - L3, X3

- High = sign of type 1 diabetes

Blood - L5

- Low glucose levels = higher insulin levels (??)
 - Oxidative stress can worsen the condition

Carbon Dioxide/Oxygen in Blood

Carbon Dioxide

- Arterial cord: higher level indicates a higher hydrogen concentration (L1, x1)
 - Symptom: pH of arterial cord blood (L1, X1)
 - Study shows that CO₂ could potentially reduce adhesion
- Venous cord*

Oxygen

- Arterial cord: insufficient amount = hypoxia
 - Study shows a correlation with endometriosis
- Venous cord: insufficient amounts can lead to abnormal SvO₂ (L2, x2)

Glomerular Filtration Rate

Was at first positive in the first groups but then switched to negative as more time progressed (within the later groups)

Glomerular Filtration Rate $< 1.73 \text{ m}^2$ = chronic kidney disease

- Lower risk for women with endometriosis?
 - Can be used to determine if the kidneys can be preserved during surgery for ureteral endometriosis
- Higher risk for women with endometrial cancer

Questions to explore further...

- Is there a way to interconnect all of these symptoms to endometriosis?
- Why are features more prominent in one model than the other?
- If one model deems a factor as positive and the other deems it as negative, which one do we go with?
- If a feature goes from positive to negative, how do we attest it?