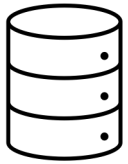
In-modeling Explainability

Explainable and Trustworthy AI

Eliana Pastor

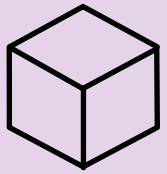
Stages of Explainability

Explainability involves the entire AI development pipeline



Pre-modelling explainability

Before building the model
- Data exploration
- Data selection
- Feature engineering



Explainable modeling

Build inherently interpretable models
- Manage the accuracy and interpretability trade-off

Post-modelling explainability

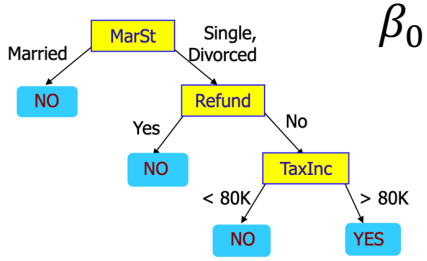
After model development
- Explaining predictions and behavior of trained models

Stages of Explainability – Explainable modelling

Design, train and adopt more interpretable/explainable models

- Adopting **an inherently explainable models**
  - does not automatically guarantee explainability (e.g., deep trees, linear models on high dimensional data)
  - Problem of explainability vs performance trade-off: interpretable models are typically less performing

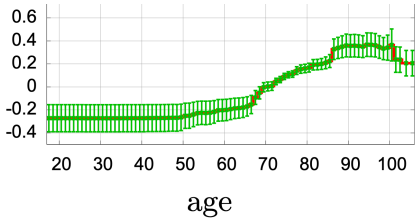
Trees



Linear models

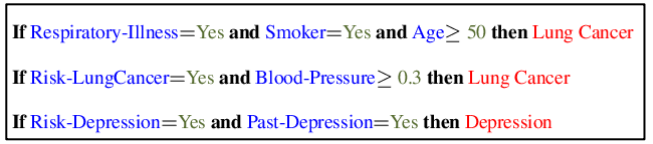
$$\beta_0 + \sum_i \beta_i x_i$$

GAMs, GA²Ms, GLMs



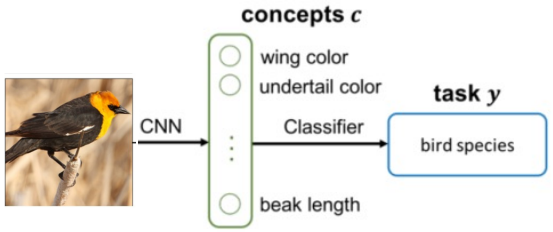
Caruana et al. "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission." KDD 2015

Decision sets - Rules



Lakkaraju et al. "Interpretable decision sets: A joint framework for description and prediction." KDD 2016

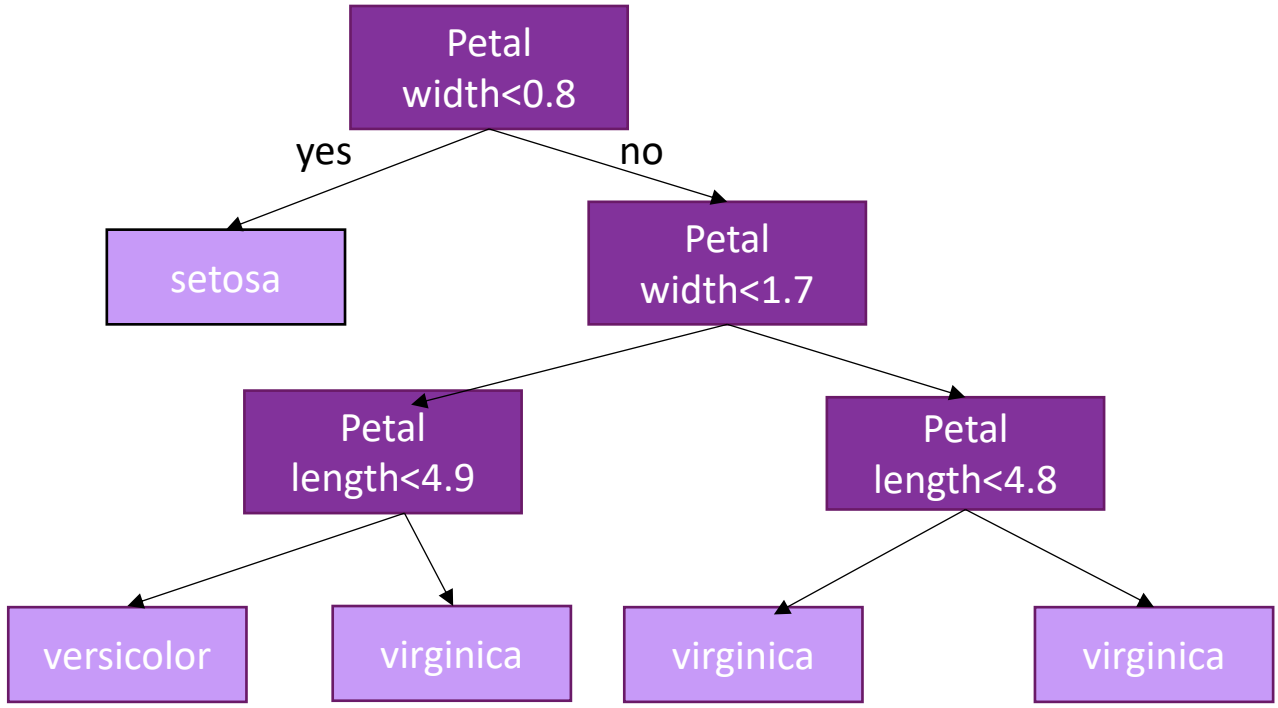
Concept-based models



Koh, Pang Wei, et al. "Concept bottleneck models." ICML 2020.

Decision trees

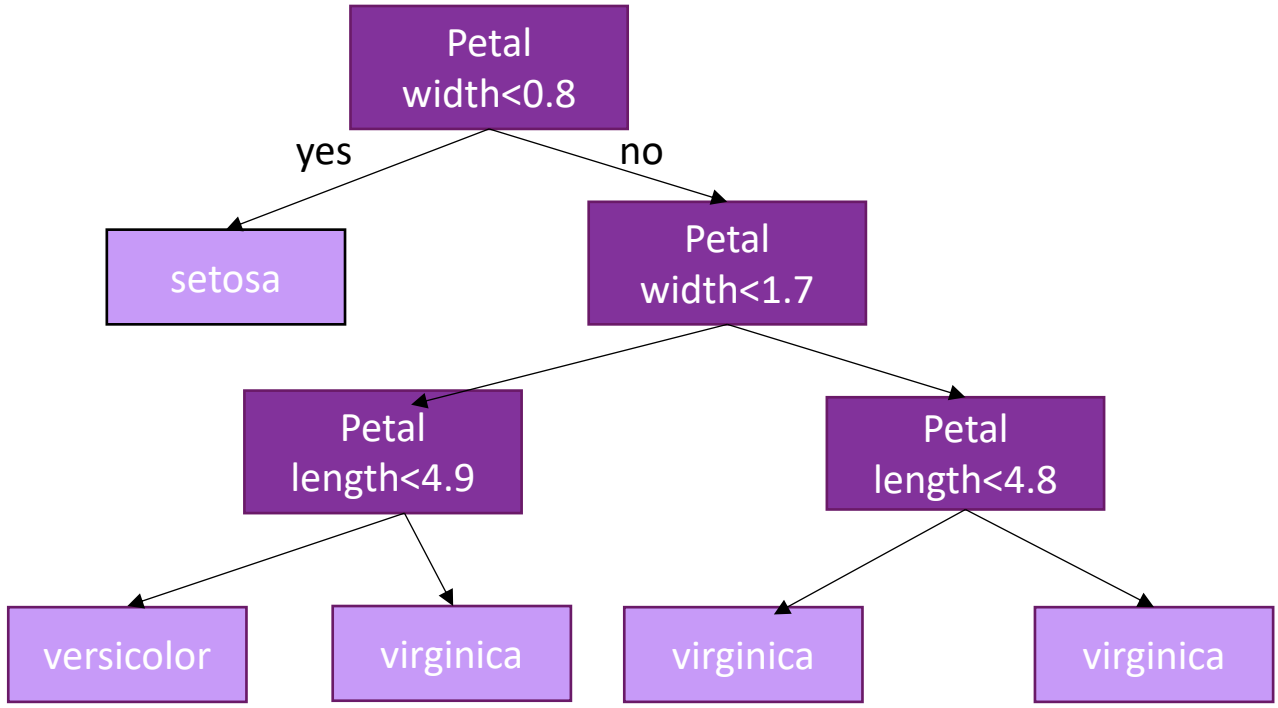
- Simple supervised models used for both classification and regression tasks.
- **Tree-like structure**
- Each **internal** node represents a decision based on a feature
- Each **leaf** node represents the outcome or the decision



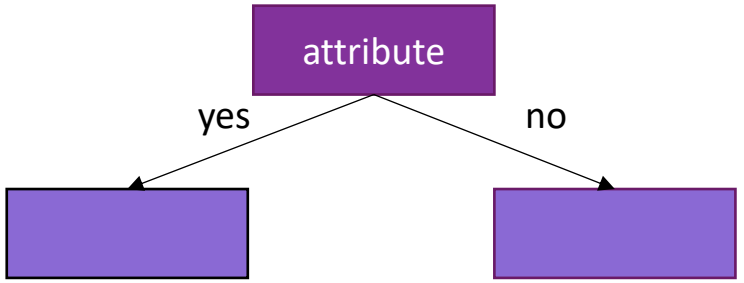
Structure - Decision trees

Structure

- Root Node: topmost node where the first decision is made
- Decision or Inner Nodes: Nodes that represent decisions or tests on attribute
- Edges: possible outcomes of a decision
- Leaf Nodes: terminal nodes that provide the final decision



Building a Decision Tree



1. Begin with the entire dataset at the root node

2. Select best splitting attribute and value based on a splitting criterion (e.g., Gini Impurity).

3. Partition the dataset into subsets based on the values of the selected attribute.

4. Recursively apply steps 2 and 3 to each subset until one of the following conditions is met:
   - All instances in the subset belong to the same class.
   - No more attributes to split on.
   - Stopping criteria (e.g., maximum depth, minimum leaf samples per leaf) are met.

5. Assign a class label to each leaf node based on the majority

Decision trees interpretability

- Decision trees offer both
  - Global interpretability
  - Local interpretability

Global interpretability for Decision trees

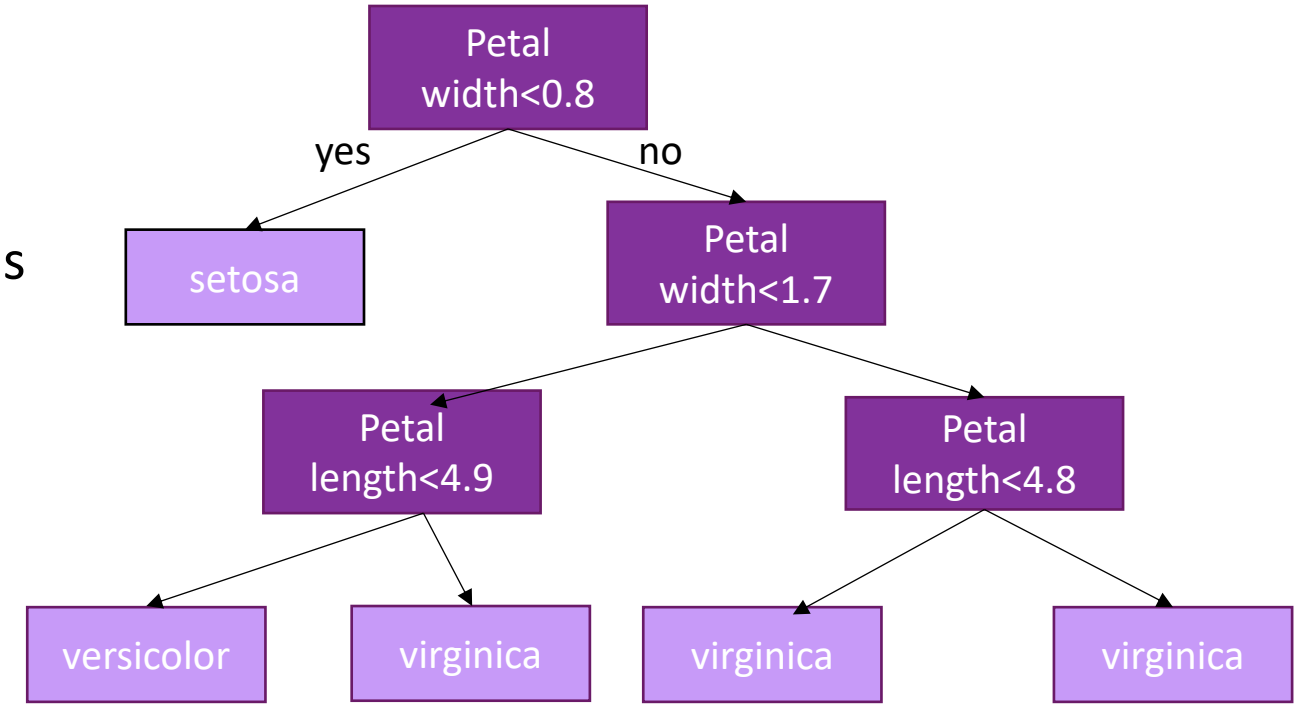
Global interpretability focuses on understanding the overall behavior and workings of the model across the entire dataset.

- Tree Structure
- Decision rules from the tree
- Feature Importance

Global interpretability for Decision trees

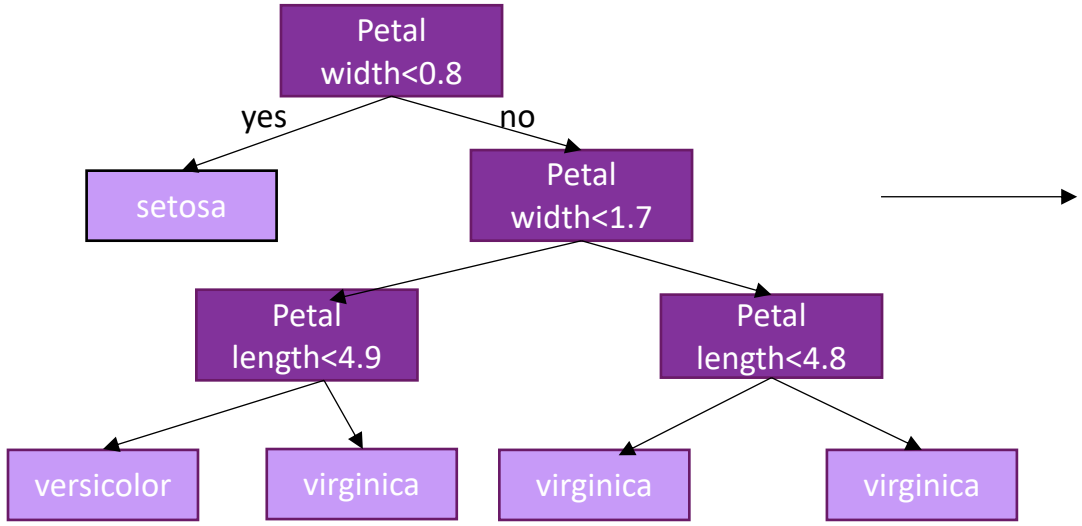
Tree structure

Analyze the decision paths of the tree models



Global interpretability for Decision trees

Decision rules from the tree

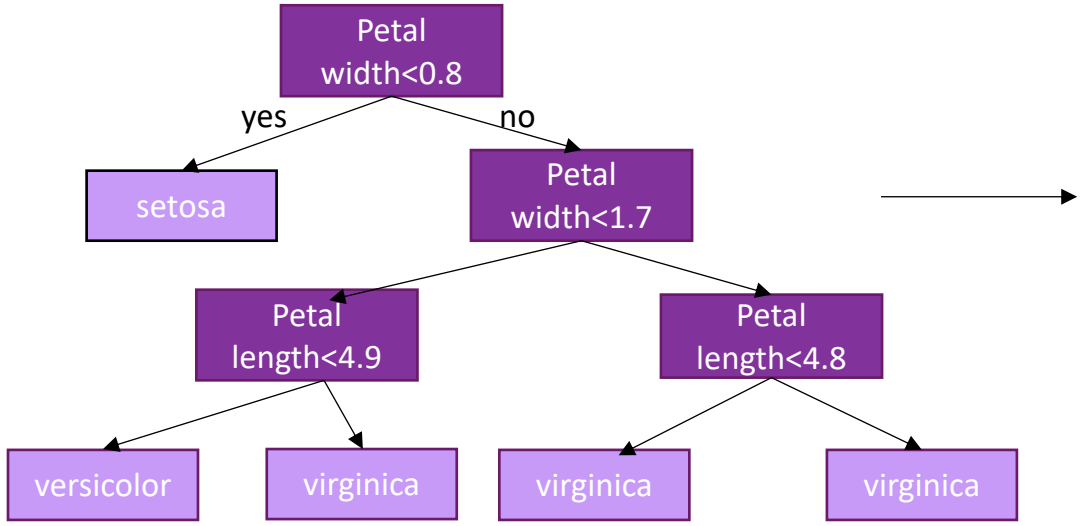


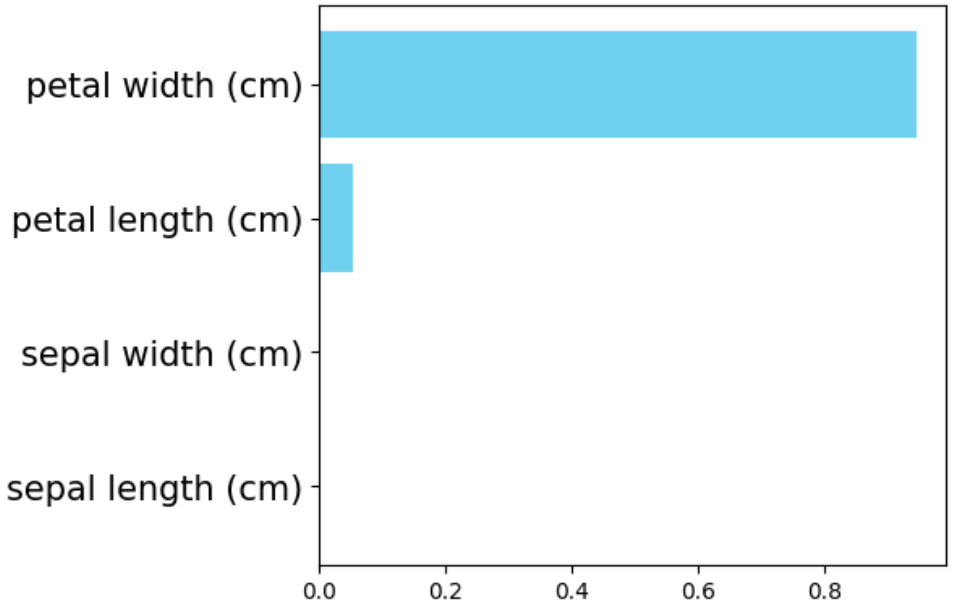
width<0.9 → setosa
width in [0.8, 1.7], length<4.9 → versicolor
width in [0.8, 1.7], length>4.9 → virginica
width > 1.7], length>4.9 → virginica

For some users, rules are more easy to understand

Global interpretability for Decision trees

Feature Importance





Feature Importance for Decision Trees

Multiple ways to compute the feature importance

- **Impurity-based feature importance**
  - The importance of a feature is the (normalized) total reduction of the impurity criterion obtained by using that feature for splitting
    - Also known as GINI importance
- **Depth-based Importance**
  - Higher importance to features that appear closer to the root node
- **Path-based Importance**
  - Features that appear more frequently in the tree have higher important

Impurity-based feature importance

Based on the notion of impurity, e.g., Gini Index

t | # Class 1: 10
Class 2: 10 | GINI = 0.5

t | # Class 1: 20
Class 2: 0 | GINI = 0

Gini Index for a given node t:

$$GINI(t) = 1 - \sum_{j} [p(j|t)]^2$$

where $p(j|t)$ is the relative frequency of class j at node t

- Maximum (1 - 1/nc) when each class occurs with equal probability, implying higher impurity degree
- Minimum (0.0) when all instances belong to one class, implying lower impurity degree

Impurity-based feature importance

The importance of a feature is computed as the (normalized) total reduction of the impurity criterion obtained by using that feature for splitting.

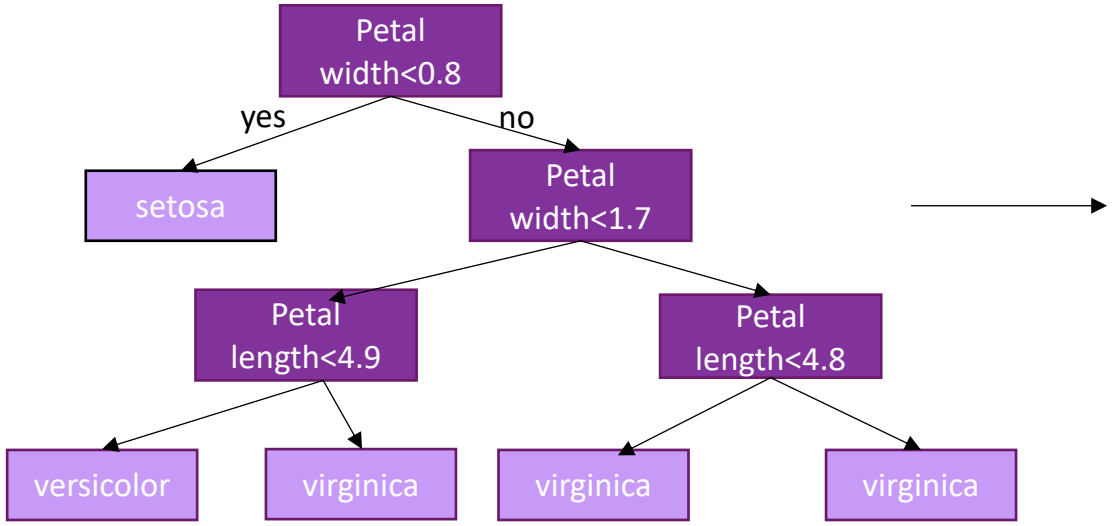
Computation

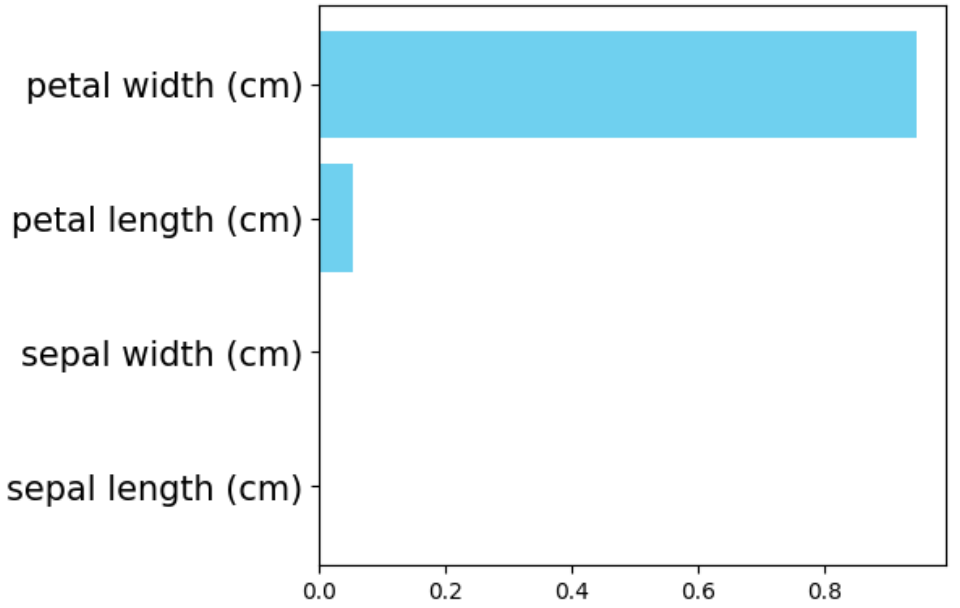
- For each split, measure how much it has reduced the impurity (e.g., Gini index) compared to the parent node
  - Difference in impurity between the parent node and its child nodes
  - Weight the difference by the number of samples in each node
  - Increment the total importance of the attribute used for the splitting by this importance
- Scale the sum of all importance in the scale 0-1
- Each feature importance indicates its relevance to overall model importance.

Impurity-based feature importance

Global interpretability

Global importance of each attribute



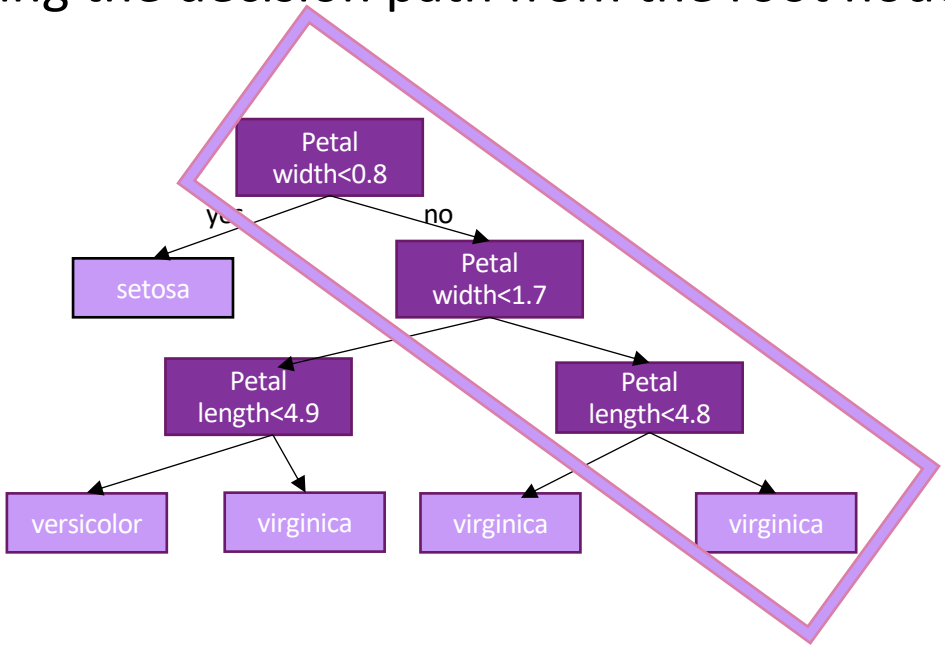


Local interpretability for Decision trees

Local interpretability refers to understanding the behavior and predictions of the model for individual instances.

It explain why a particular prediction is made for a specific input.

- **Path Explanation**: tracing the decision path from the root node to the leaf node for an instance

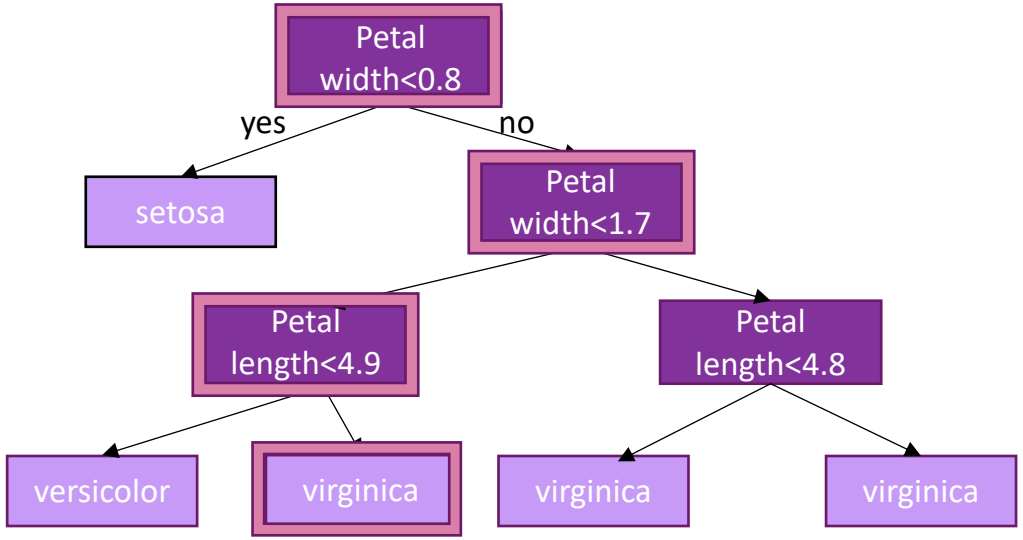


Local interpretability for Decision trees

Example

Instance

Petal width = 1.1, petal length=5, sepal width = 1, sepal lenght=1



Decision Path

Petal width<0.8 = False, petal width <1.7 = True, Petal legth < 4.9 = False

Advantages of Trees

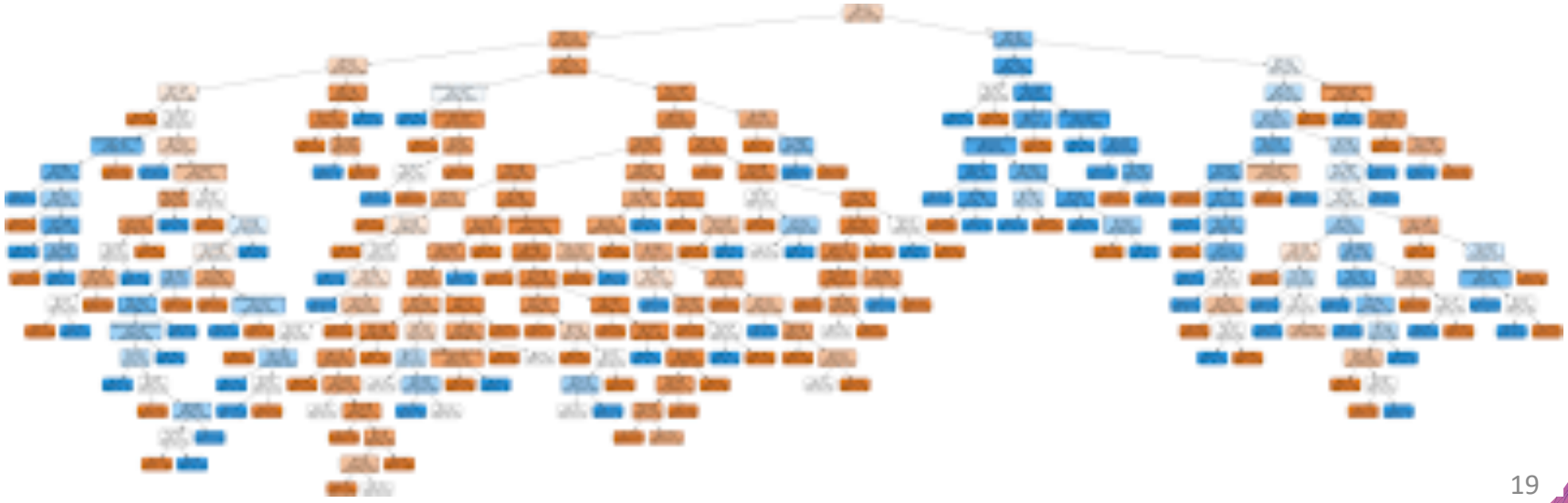
- Offer multiple insights
  - Global interpretability
  - Local interpretability
  - + Subgroup. Each path actually covers a subset of the data. Easier to understand than individual points

- (Generally) Easy to interpret, also globally
  - Human-friendly explanations, the interpretation is simple
- Tree structures offers a build-in visualization, enhance understanding
- Facilitates communication with non-technical stakeholders
- Being interpretable, users can assess if they can trust the model

Limitations of Trees

- Low accuracy compared to more complex model
  - Interpretability accuracy trade-off

- Decision trees are very interpretable – if they are small!
  - Few splitting nodes
  - Low depth



Decision rules

Classify instances by using “if…then…” rules

- **Rule: (Condition) → y** where
  - Condition is a conjunction of simple predicates
  - y is the class label

- Rule extraction
  - Rule Induction Algorithms
    - e.g., CN2 or RIPPER, explicitly generate rules from the training data based
  - Associative classifiers
  - From decision trees

Decision rules

- **Decision list**
  - Ordered decision rules
  - Prediction based on the first rule satisfying the instance

- **Decision set**
  - Independent rules
  - Rules are mutually exclusive, or there is a strategy for resolving conflicts, such as majority voting

If Respiratory-Illness=Yes **and** Smoker=Yes **and** Age≥ 50 **then** Lung Cancer

Else if Risk-Depression=Yes **then** Depression

Else if BMI ≥ 0.2 **and** Age≥ 60 **then** Diabetes

Else if Headaches=Yes **and** Dizziness=Yes, **then** Depression

Else if Doctor-Visits≥ 0.3 **then** Diabetes

Else if Disposition-Tiredness=Yes **then** Depression

Else Diabetes

If Respiratory-Illness=Yes **and** Smoker=Yes **and** Age≥ 50 **then** Lung Cancer

If Risk-LungCancer=Yes **and** Blood-Pressure≥ 0.3 **then** Lung Cancer

If Risk-Depression=Yes **and** Past-Depression=Yes **then** Depression

If BMI≥ 0.3 **and** Insurance=None **and** Blood-Pressure≥ 0.2 **then** Depression

If Smoker=Yes **and** BMI≥ 0.2 **and** Age≥ 60 **then** Diabetes

If Risk-Diabetes=Yes **and** BMI≥ 0.4 **and** Prob-Infections≥ 0.2 **then** Diabetes

If Doctor-Visits ≥ 0.4 **and** Childhood-Obesity=Yes **then** Diabetes

Lakkaraju et al. "Interpretable decision sets: A joint framework for description and prediction." KDD 2016

Global Interpretability for Decision rules

- Analyze rules themselves

If Respiratory-Illness=Yes **and** Smoker=Yes **and** Age≥ 50 **then** Lung Cancer

If Risk-LungCancer=Yes **and** Blood-Pressure≥ 0.3 **then** Lung Cancer

If Risk-Depression=Yes **and** Past-Depression=Yes **then** Depression

If BMI≥ 0.3 **and** Insurance=None **and** Blood-Pressure≥ 0.2 **then** Depression

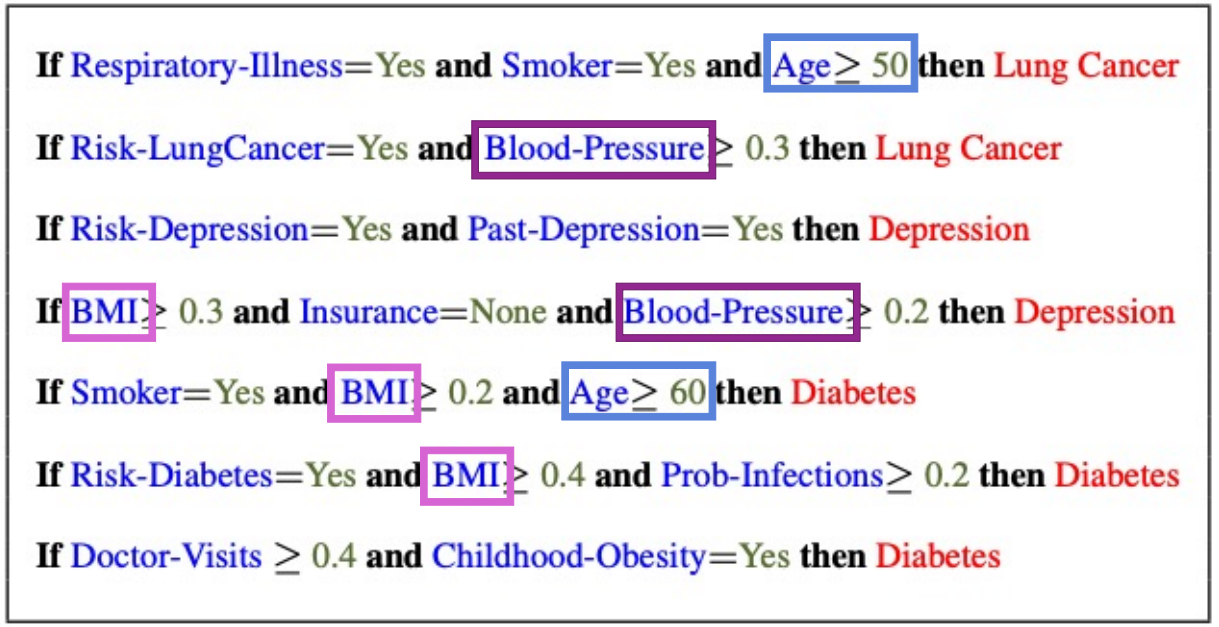
If Smoker=Yes **and** BMI≥ 0.2 **and** Age≥ 60 **then** Diabetes

If Risk-Diabetes=Yes **and** BMI≥ 0.4 **and** Prob-Infections≥ 0.2 **then** Diabetes

If Doctor-Visits ≥ 0.4 **and** Childhood-Obesity=Yes **then** Diabetes

Global Interpretability for Decision rules

- Feature importance
  - Features that appear in multiple decision rules are likely to be more important



Local Interpretability for Decision rules

- Analyze individual rule satisfying the instance

If Respiratory-Illness=Yes **and** Smoker=Yes **and** Age$\geq$ 50 **then** Lung Cancer

Advantages of Rules

- Offer multiple insights
  - Global interpretability
  - Local interpretability
  - + Subgroup. Each rule actually covers a subset of the data

- (Generally) easy to interpret, also globally
  - Human-friendly explanations, the interpretation is simple
- Expressive as tree, but more compact
  - Some users find them **more interpretable than trees**
- Facilitates communication with non-technical stakeholders
- Being interpretable, users can assess if they can trust the model

Limitations of Rules

- Often require categorical data
  - Numerical feature should be discretized
- Low accuracy compared to more complex model
  - Interpretability accuracy trade-off
- Rules are very interpretable – if they are compact!
  - Few rules
  - Short rules

Linear regression

A linear regression model predicts the target as a weighted sum of the feature inputs.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_p x_p$$

Interpret the Coefficients

- The coefficients β_i represents the change in the dependent variable for a one-unit change in the corresponding independent variable, holding all other variables constant.
  - i.e., Increasing x_i by one unit changes the estimated outcome by its β_i .
    - If β_i is positive, it indicates that as x_i increases, y also increases.
    - If β_i is negative, it indicates that as x_i increases, y decreases.
- The intercept β_i represents the value of the dependent variable when all independent variables are set to zero.
  - In some cases, the interpretation might not be meaningful, especially if zero doesn't have a practical meaning for the variables.

Example - Interpreting Linear Regression

Goal**.** Predict the salaries of individuals based on their years of experience and level of education. We want to build a linear regression model to predict salaries based on these two variables.

Model.

$$Salary = \beta_0 + \beta_1 \times Years\ of\ Experience + \beta_1 \times Level\ of\ Education$$

β_0 = 40000, β_1 = 3000, β_2 = 2000

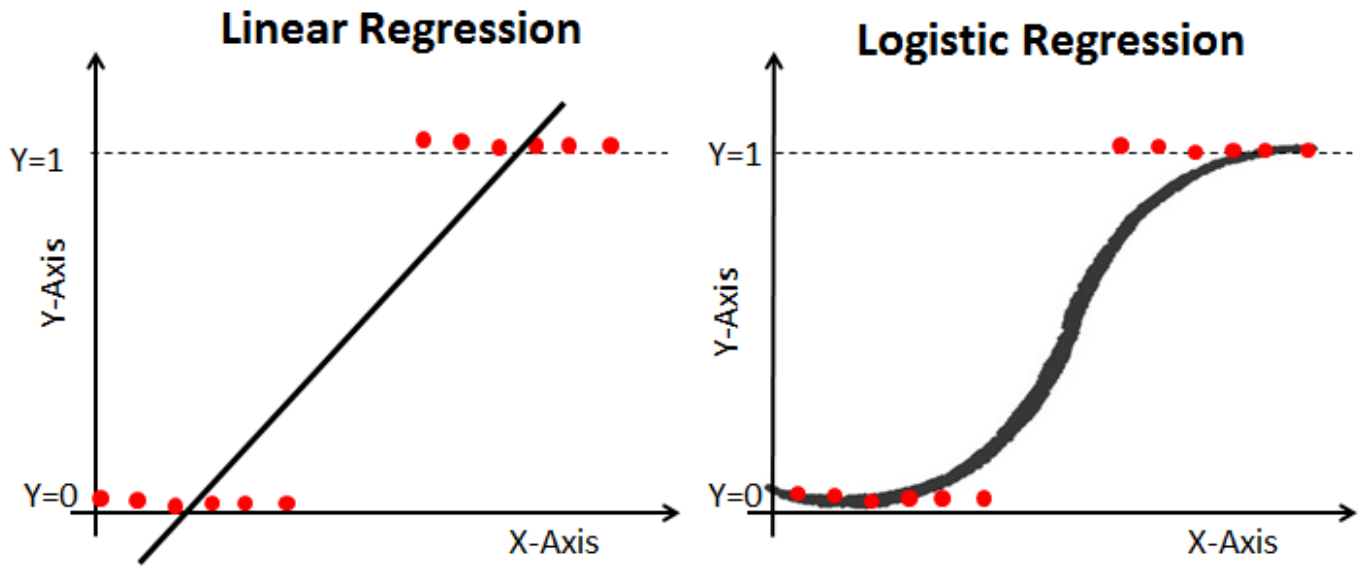
- Intercept : A person with zero years of experience and zero years of education would have a predicted salary of $40,000.
- For each additional year of experience, the predicted salary is expected to increase by $3000, holding the level of education constant.
- For each additional year of education, the predicted salary is expected to increase by $2000, holding years of experience constant.

Inherently explainable models
Logistic regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_p x_p$$

$$\text{logistic}(x) = \frac{1}{1+\exp(-x)}$$

$$P(y = 1) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_p x_p))}$$



Logistic regression

$$P(y=1) = \frac{1}{1+\exp(-(\beta_0 + \beta_1 x_1 \ + \beta_2 x_2 + .. + \beta_p x_p))}$$

Log odds

$$\ln\left(\frac{P(y=1)}{1-P(y=1)}\right) = \ln(Odds) = \beta_0 + \beta_1 x_1 \ + \beta_2 x_2 + .. + \beta_p x_p$$

- The coefficients represent the change in the log odds of the event occurring for a one-unit change in the corresponding predictor variable, holding all other variables constant.

Logistic regression

$$\frac{P(y=1)}{1-P(y=1)} = Odds = \exp(\beta_0 + \beta_1 x_1 \ + \beta_2 x_2 + .. + \beta_p x_p)$$

We compare what happens when we increase one of the feature values by 1.

We look at the ratio of the two predictions:

$$\frac{Odds_{x_j+1}}{Odds_{x_j}} = \frac{\exp(\beta_0 + \beta_1 x_1 + .. + \beta_j (x_j+1) + .. + \beta_p x_p)}{\exp(\beta_0 + \beta_1 x_1 + .. + \beta_j x_j + .. + \beta_p x_p)}$$
$$= \exp\big(\beta_j (x_j+1) - \beta_j x_j\big) = \exp(\beta_j)$$

- If we increase the value of feature x_j by one unit, the estimated odds change by a factor of $\exp(\beta_j)$

Example - Logistic regression

Let's consider a hypothetical example where we want to predict the likelihood of a student passing an exam based on the number of hours they studied and whether they attended a preparatory course.

$$P(y = 1) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 Hours\ Studied\ + \beta_2 Preparatory\ Course))}$$

β_0 = -2, β_1 = 0.8, β_2 = 1.2

- For every additional hour studied, the odds of passing the exam increase by approximately 2.22 (≈ e^0.8) times.
- Students who attended a preparatory course have odds 3.32 (≈ e^1.2) times higher of passing the exam compared to those who did not attend the preparatory course, holding the hours studied constant.

Advantages of Linear/Logistic Regression

- Simple Interpretation
  - Via coefficients

- Provide variable importance
  - Via the magnitude and sign of the coefficients

Limitations of Linear/Logistic Regression

- Low accuracy compared to more complex model
  - Interpretability accuracy trade-off
- Limited to Linear Decision Boundaries
- Impact of the number of coefficients on explainability
  - The higher, lower interpretable

Interpreting Naïve Bayes

$$P(C_k \mid x) = \frac{1}{Z} \mathrm{P}(\mathrm{C_k}) \prod_{i=1}^{n} P(x_i \mid C_k)$$

- **Feature Importance**
  - Given by the conditional probabilities of features given the class labels.
  - For each class, Naive Bayes calculates the probability of each feature occurring given that class. Higher probabilities indicate that the feature is more indicative of that class.

Advantages and Disadvantages of Naïve Bayes

Advantages

- Simple and easy to implement
- Provide feature importance

Disadvantage

- Assumption of Feature Independence
- Limited expressiveness, low performance

Instance-based classifiers - KNN

- Prediction based on the K nearest neighbors of the instance
- Explanation by example
  - Set instances, the K nearest neighbours
- Do not offer global interpretability
  - It is inherently local!

Advantages and Limitations of KNN

Advantages

- Easy to derive the explanation
- Explanation by example is close as how often human reason
- Intuitive for some data types
  - e.g., similar images

Disadvantages

- Difficult to interpret as we increase the number of features
- Other form of explanations, e.g., feature importance, could be preferred

Stages of Explainability – Explainable modelling

- **Targeting interpretability by design**
  - Design high-performing models imposing interpretability constraints to enable their interpretability
  - e.g., Explainability via regularization
    - Apply regularization to improve model explainability

$$\min_{f \in F} \sum_i Loss(f, x_i, y_i) + \text{InterpretabilityPenalty}(f),$$

$$\text{subject to Interpretability constraint}(f)$$

Cynthia Rudin et al. “Interpretable machine learning: Fundamental principles and 10 grand challenges”, 2021

Targeting interpretability by design

- Trees

$$\min_{f \in \text{ set of trees}} \frac{1}{n} \sum_i \text{Loss}(f, z_i) + C \cdot \text{Number of leaves } (f),$$

- Linear models

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_i \text{Loss}(f, z_i) + C \cdot \text{Number of nonzero terms } (f), \quad \text{subject to}$$

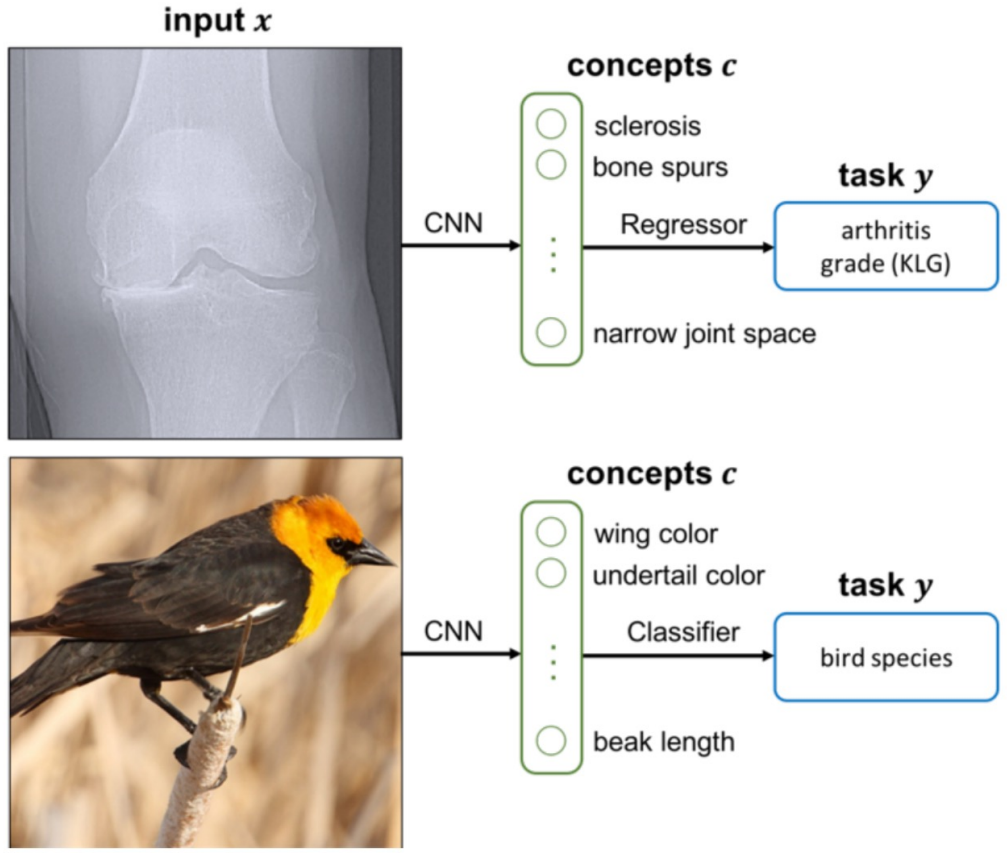
$$f \text{ is a linear model, } f(\mathbf{x}) = \sum_{j=1}^{p} \lambda_j x_j,$$

Problem: these models could still underperform compared to more complex models

Cynthia Rudin et al. “Interpretable machine learning: Fundamental principles and 10 grand challenges”, 2021
Wu, Mike, et al. "Beyond sparsity: Tree regularization of deep models for interpretability." AAAI 2018.

Targeting interpretability by design - Concept-based models

- E.g., Concept Bottleneck Models

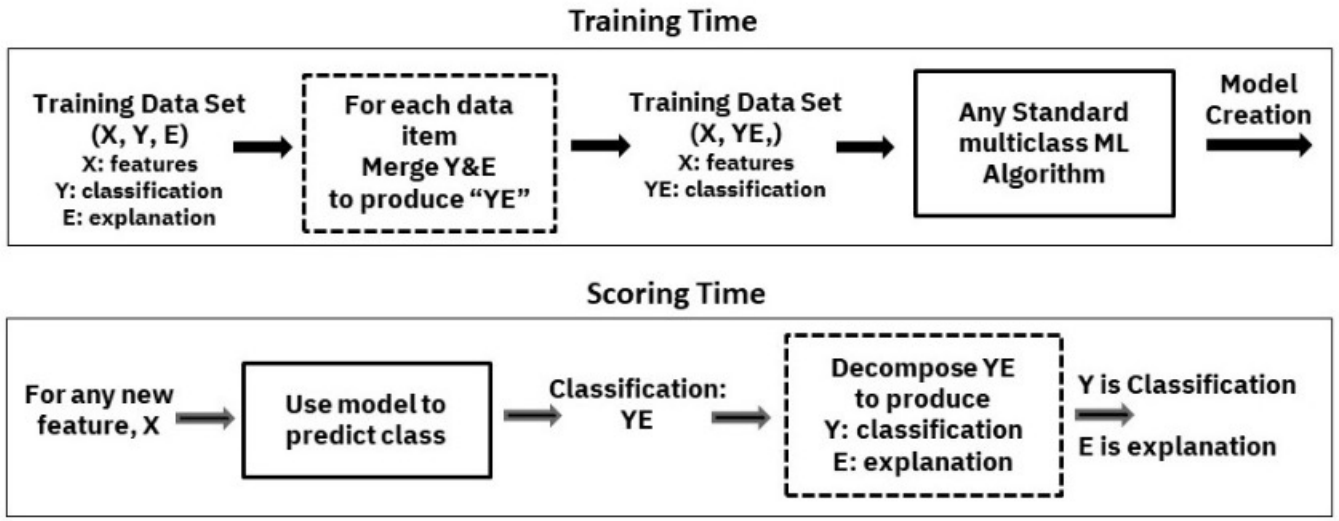


Koh, Pang Wei, et al. "Concept bottleneck models." International conference on machine learning. PMLR, 2020.

Stages of Explainability – Explainable modelling

- **Explanations-in-the-loop**

  Train AI systems **to** jointly **provide** a prediction and its **explanation**



Hind et al. TED: Teaching AI to Explain its Decisions. AIES 2019

TED - Teaching Explanations for Decisions

Train a model to jointly produce both a decision as well as an explanation

Analogy – Teaching & Learning process

- Training: As a supervisor show the new employee several example situations and teach them the correct action: approve or reject a loan application, and explain the reason for the action, such as “insufficient salary”.
- Deployment: The new employee will be able to make independent decisions on new loan applications and will give explanation based on the explanations they learned from their supervisor

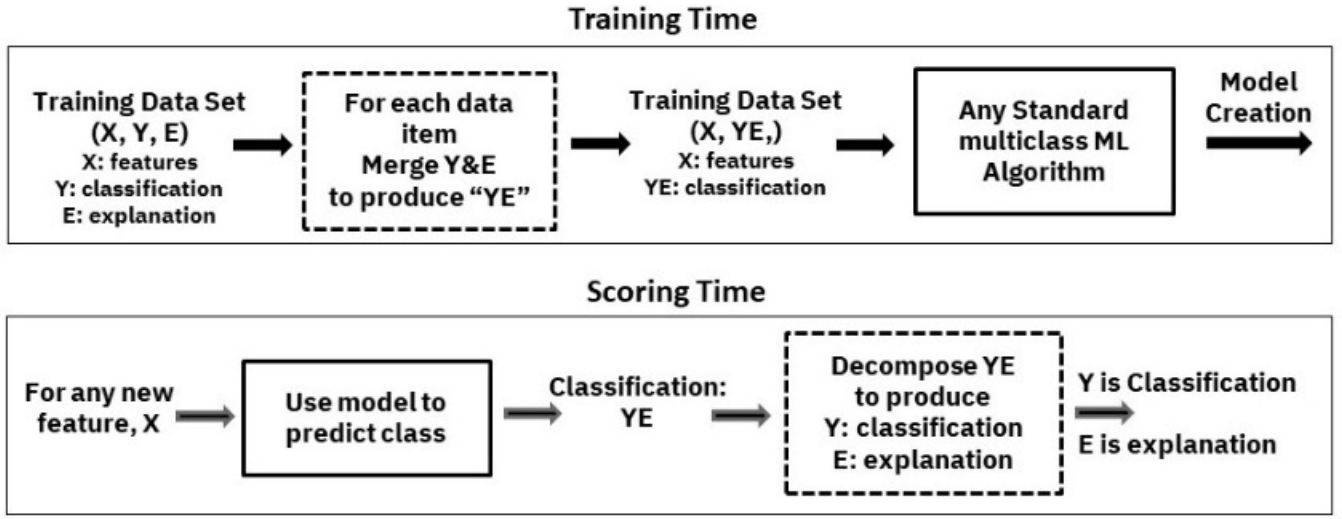
TED

- Training: Teach the model to make correct predictions but also to learn their explanations, by providing them
- Test/Deployment: For a new sample, the model generates the prediction and its explanation

Hind et al. TED: Teaching AI to Explain its Decisions. AIES 2019

TED - Teaching Explanations for Decisions

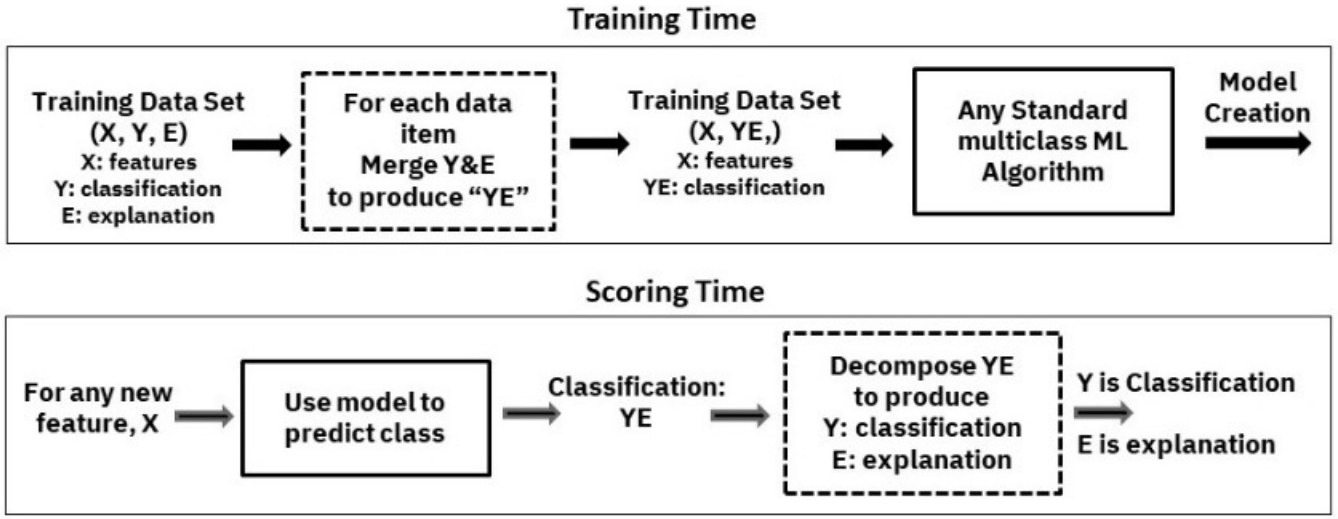
- Training data
  - X, Y, E (Explanation)
  - E is a **rationales:** human annotations that can explain labels, ground truth explanation
- Training
  - Learn Y+E from X
  - Generic classification model f



Hind et al. TED: Teaching AI to Explain its Decisions. AIES 2019

TED - Teaching Explanations for Decisions

- Test/Deployment
  - Predict Y+E from new instances
  - Decompose Y, E



Hind et al. TED: Teaching AI to Explain its Decisions. AIES 2019

Advantages of Explanation-in-the-loop

- Explainability directly in the training process
- Teach the model what important for us as human
  - Alignment to human reasoning and values
- Explanation can be tailored for the target audience

Hind et al. TED: Teaching AI to Explain its Decisions. AIES 2019

Limitations of Explanation-in-the-loop

- Require a dataset annotated with explanations
  - The paper tests the approach with synthetic rationales..
    - Rationales as rules and predict which one matches the input, encoded as an integer

- Explanations may not necessarily reflect of how model predictions were made but what humans expects

- Faithfulness to the model vs Plausibility
  - Faithfulness: whether the explanation matches the model inner working
  - Plausibility: whether the explanation matches what humans expect

Hind et al. TED: Teaching AI to Explain its Decisions. AIES 2019

Teach Me to Explain – Datasets annotated with explanations

- Goal
  - to train better models via additional training supervision
  - to train interpretable models that explain their own predictions
  - to evaluate *plausibility* of model-generated explanations by measuring their agreement with human explanations

- Multiple examples, especially for text data
  - Highlight – part of the input --> What a **wonderful** day! Sentiment: positive
  - Free text --> 'The answer is correct because the person said it with a joyful voice'
  - Structured --> e.g., constrained text/form 'Is it joyful? Yes/no' Yes, 'Is it loud?' Yes

Wiegreffe, Sarah, and Ana Marasovic. "Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing." Neurips Benchmark and Datasets 2021

References

- Lakkaraju et al. "Interpretable decision sets: A joint framework for description and prediction." KDD 2016
- Molnar, Christoph. *Interpretable machine learning* https://christophm.github.io/interpretable-ml-book/
- Wiegreffe, Sarah, and Ana Marasovic. "Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing." Neurips Benchmark and Datasets 2021
- Hind et al. TED: Teaching AI to Explain its Decisions. AIES 2019
- Koh, Pang Wei, et al. "Concept bottleneck models." International conference on machine learning. PMLR, 2020.