2025

4Geeks Academy: data science cohort 12

DAY 20: DECISION TREES

TODO

DECISION TREES	Model details, applications and hyperparameters
DECISION TREE PROJECT	Work on Decision Tree Project Tutorial (Decision Tree Algo. module), plan to finish before class Monday
LINEAR REGRESSION PROJECTS	Try to submit the linear regression projects (banking marketing and health demographics) tonight, if you haven’t done so already

TOPICS

01 DECISION TREE MODELS

02 APPLICATIONS

03 HYPERPARAMETERS

DECISION TREES

WHAT

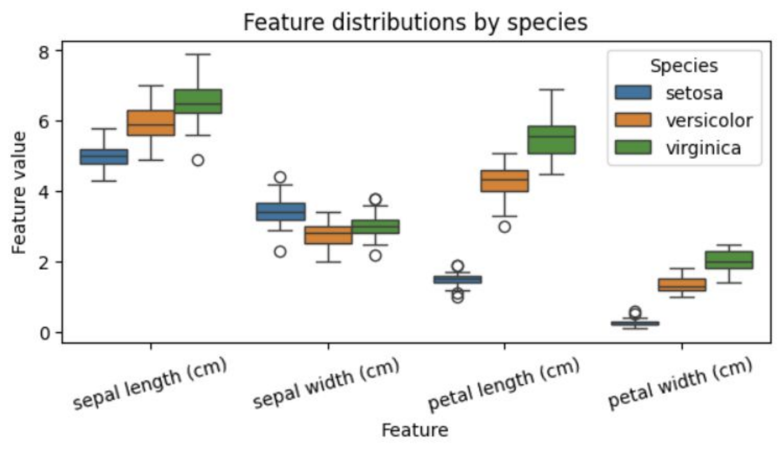
Models data using a sequence of ‘decisions’ based on input feature values

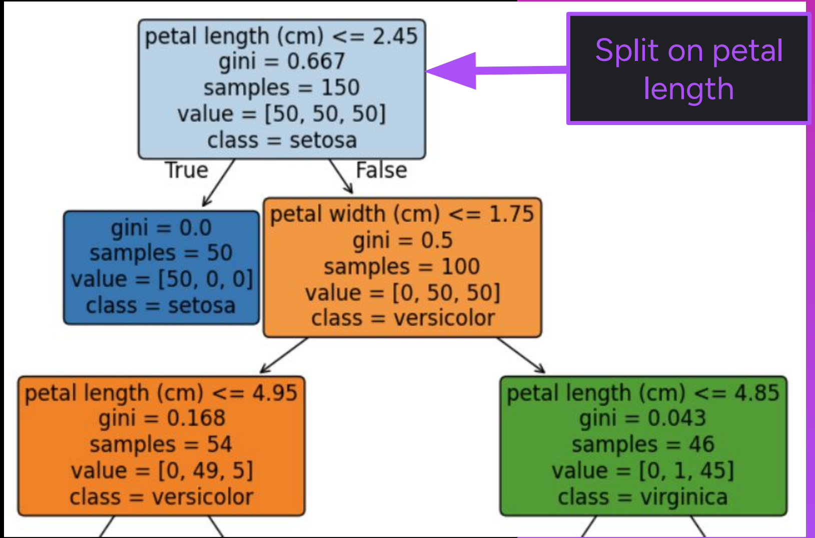
WHY

- More powerful than linear or logistic regression
- Can be used for classification and regression problems
- Intuitive - reflects how human make decisions with data
- Interpretable - shows exactly how/why a specific prediction was made

HOW

EX: Classification with iris dataset





Split on petal length

DECISION TREES

WHAT

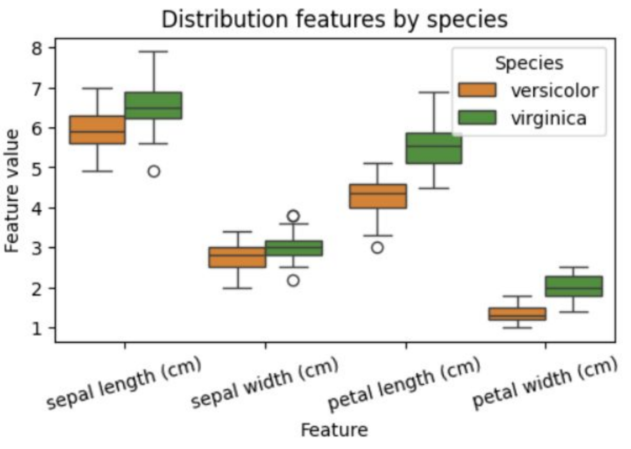
Models data using a sequence of ‘decisions’ based on input feature values

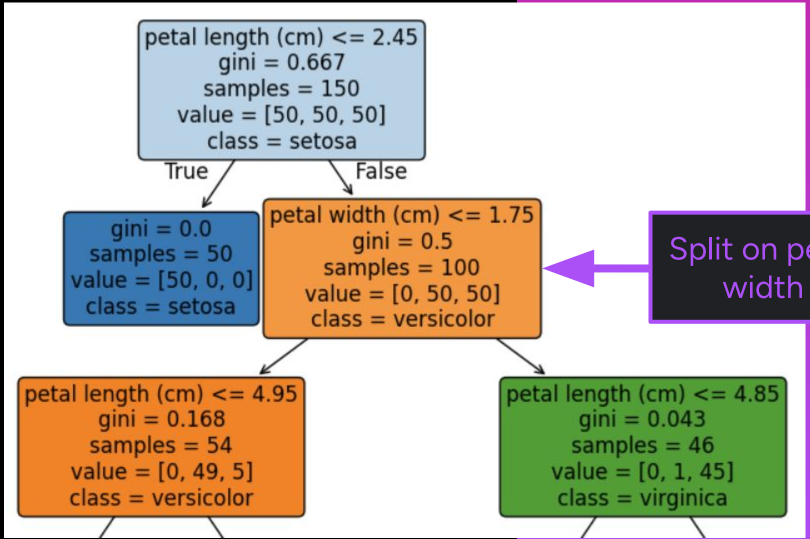
WHY

- More powerful than linear or logistic regression
- Can be used for classification and regression problems
- Intuitive - reflects how human make decisions with data
- Interpretable - shows exactly how/why a specific prediction was made

HOW

EX: Classification with iris dataset





petal length (cm) <= 2.45
gini = 0.667
samples = 150
value = [50, 50, 50]
class = setosa

gini = 0.0
samples = 50
value = [50, 0, 0]
class = setosa

petal width (cm) <= 1.75
gini = 0.5
samples = 100
value = [0, 50, 50]
class = versicolor

petal length (cm) <= 4.95
gini = 0.168
samples = 54
value = [0, 49, 5]
class = versicolor

petal length (cm) <= 4.85
gini = 0.043
samples = 46
value = [0, 1, 45]
class = virginica

Split on petal width

APPLICATIONS

TYPES

- **Scikit-learn** `DecisionTreeClassifier()`: for classification problems
- **Scikit-learn** `DecisionTreeRegressor()`: for regression problems

PROS

- Works well for datasets with nonlinear relationships
- Catches interactions between features
- Can handle categorical and continuous data
- Can handle missing data
- Robust to outliers

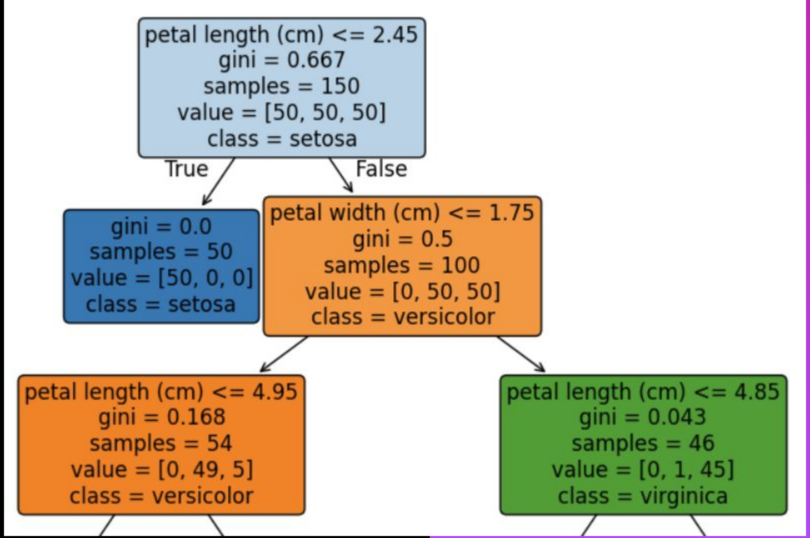
CONS

- Prone to overfitting
- Regression trees don’t extrapolate outside of training label range
- Sensitive to imbalanced classes
- Can be computationally expensive for large datasets

HYPERPARAMETERS

- `max_depth`: how many splits deep will the tree go?
- `min_samples_split`: minimum sample remaining in a leaf to keep splitting
- `max_features`: maximum features to consider for splitting at each node
- `max_leaf_nodes`: maximum number of leaf nodes to create
- `min_impurity_decrease`: minimum gain in score to split

Optimizing these parameters can help with overfitting!


petal length (cm) <= 2.45
gini = 0.667
samples = 150
value = [50, 50, 50]
class = setosa
True
False
gini = 0.0
samples = 50
value = [50, 0, 0]
class = setosa
petal width (cm) <= 1.75
gini = 0.5
samples = 100
value = [0, 50, 50]
class = versicolor
petal length (cm) <= 4.95
gini = 0.168
samples = 54
value = [0, 49, 5]
class = versicolor
petal length (cm) <= 4.85
gini = 0.043
samples = 46
value = [0, 1, 45]
class = virginica