Machine Learning

- Gradient Descent Algorithm
- Linear Regression
- Non-Linear Regression
- Logistic Regression
- Decision Trees
  - Regression Trees
  - Classification Trees
- Clustering Algorithms
  - K-Means
  - Hierarchical clustering
  - DB-Scan
  - Mean Shift
  - GMM
- Support Vector Machine

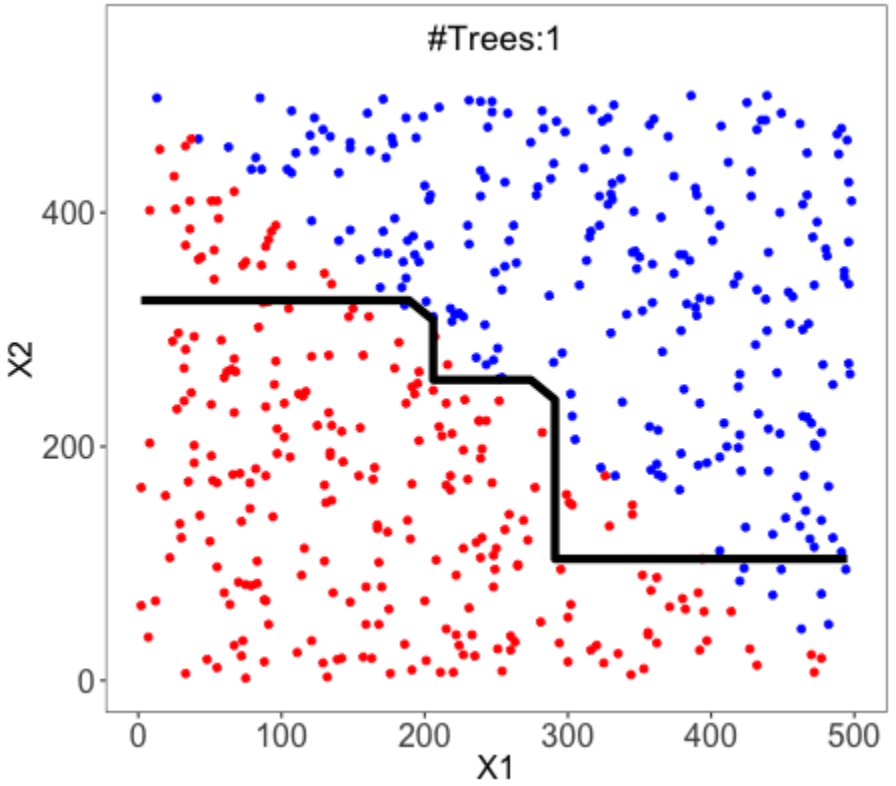
Deep Learning

- MLP
- CNN

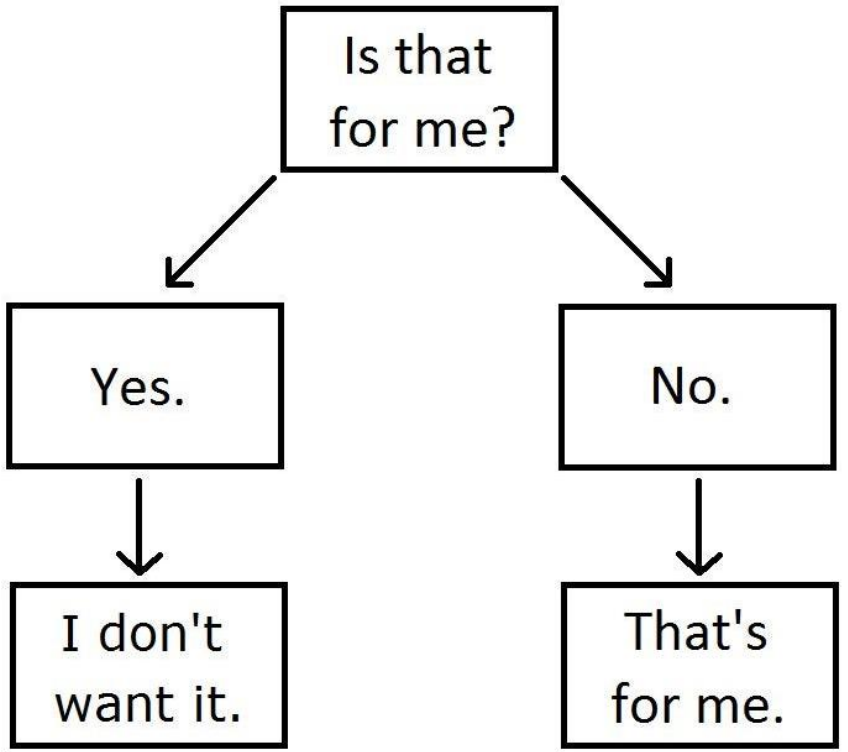
Datasets

- Breast Cancer Wisconsin
- MIMIC-III
- Framingham Heart Study
- Alzheimer’s Disease Neuroimaging Initiative
- Drug discovery
- Microbiome

Tree Based Models



C4 - Decision Trees

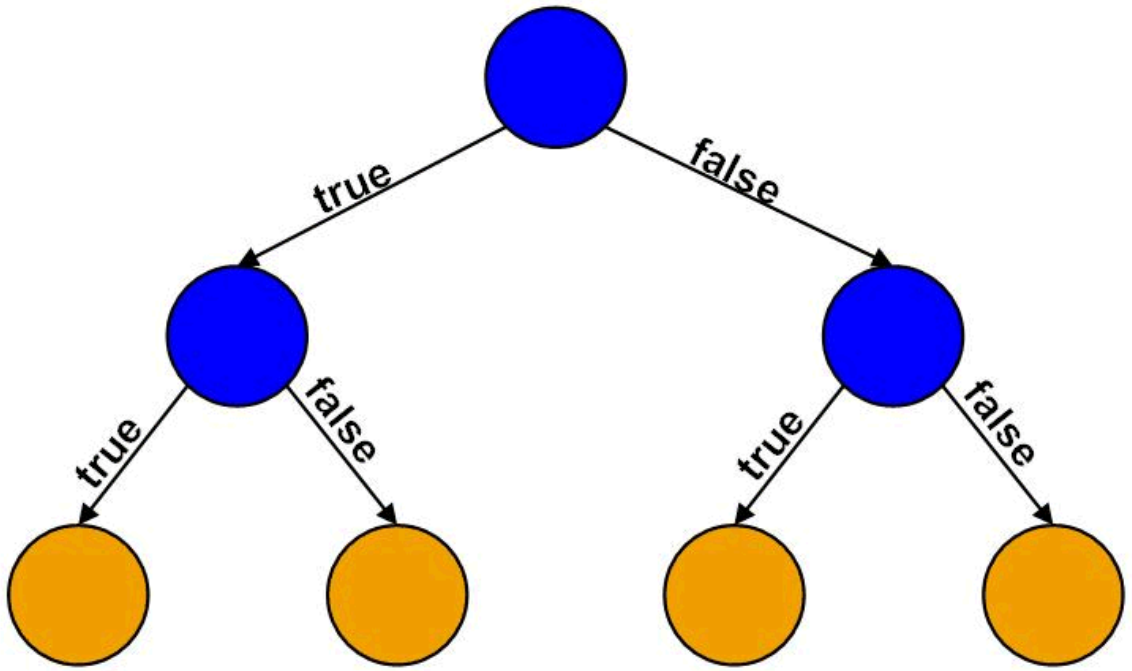


Decision Tree

GOAL: Learn if else-questions with each question involving one feature and one split-point

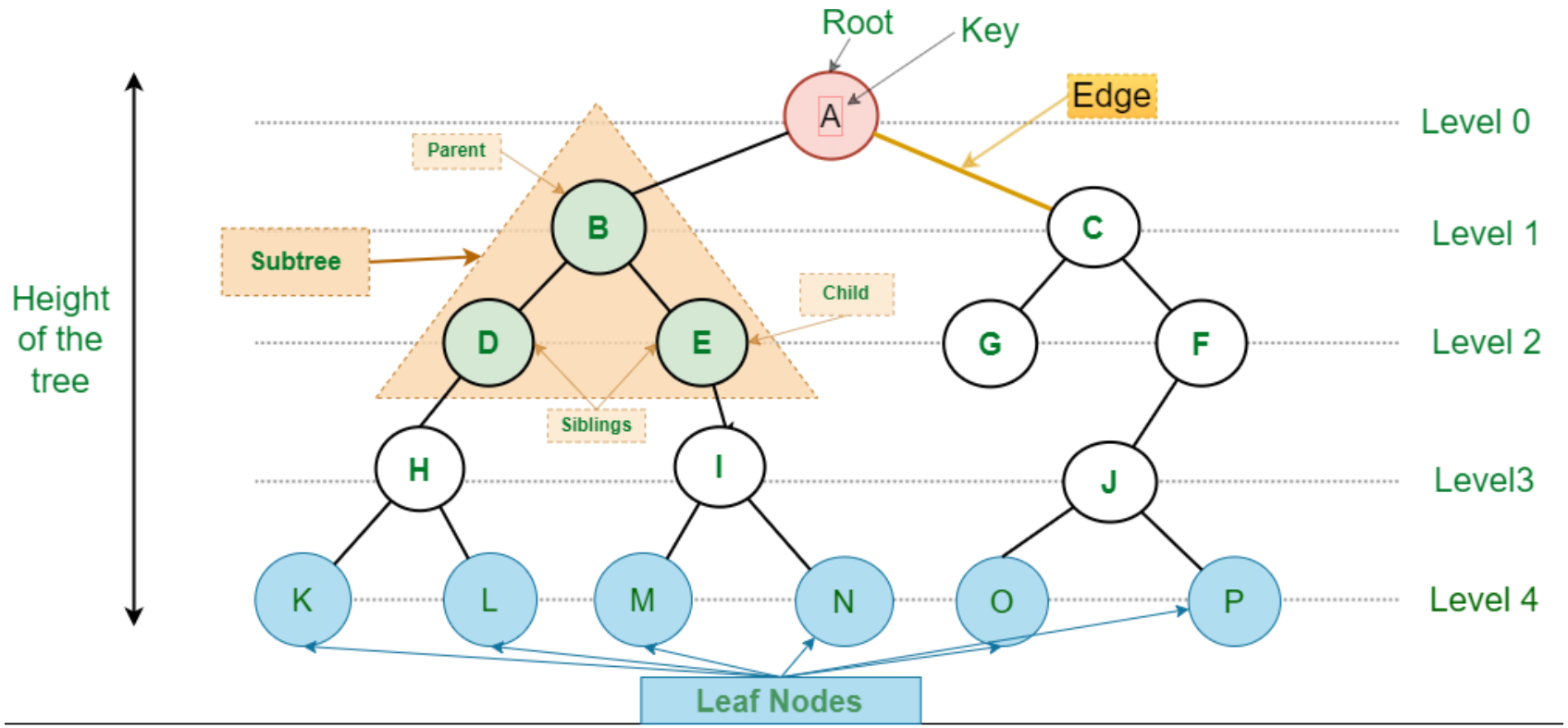
HOW:

- Divides the feature-space (N-dimensional) into regions where all instances in one region are assigned to only one class-label (discrete or continue).



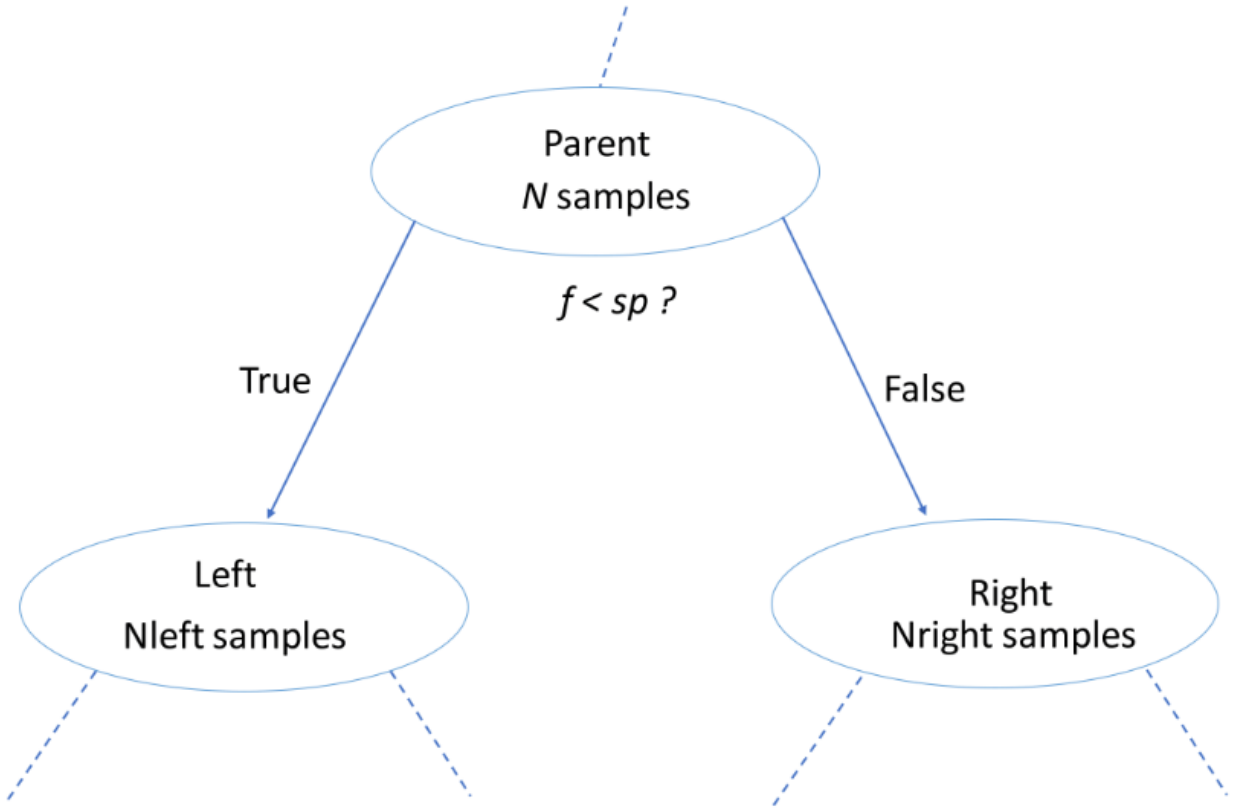
Decision Tree building blocks

- **DT**: data structure consisting of a hierarchy of nodes (individual units).
- **NODE**:
  - **ROOT**: no-parent node
  - **INTERNAL NODE**: question giving rise to two children nodes
  - **LEAF**: prediction(discrete or continue), no children nodes.



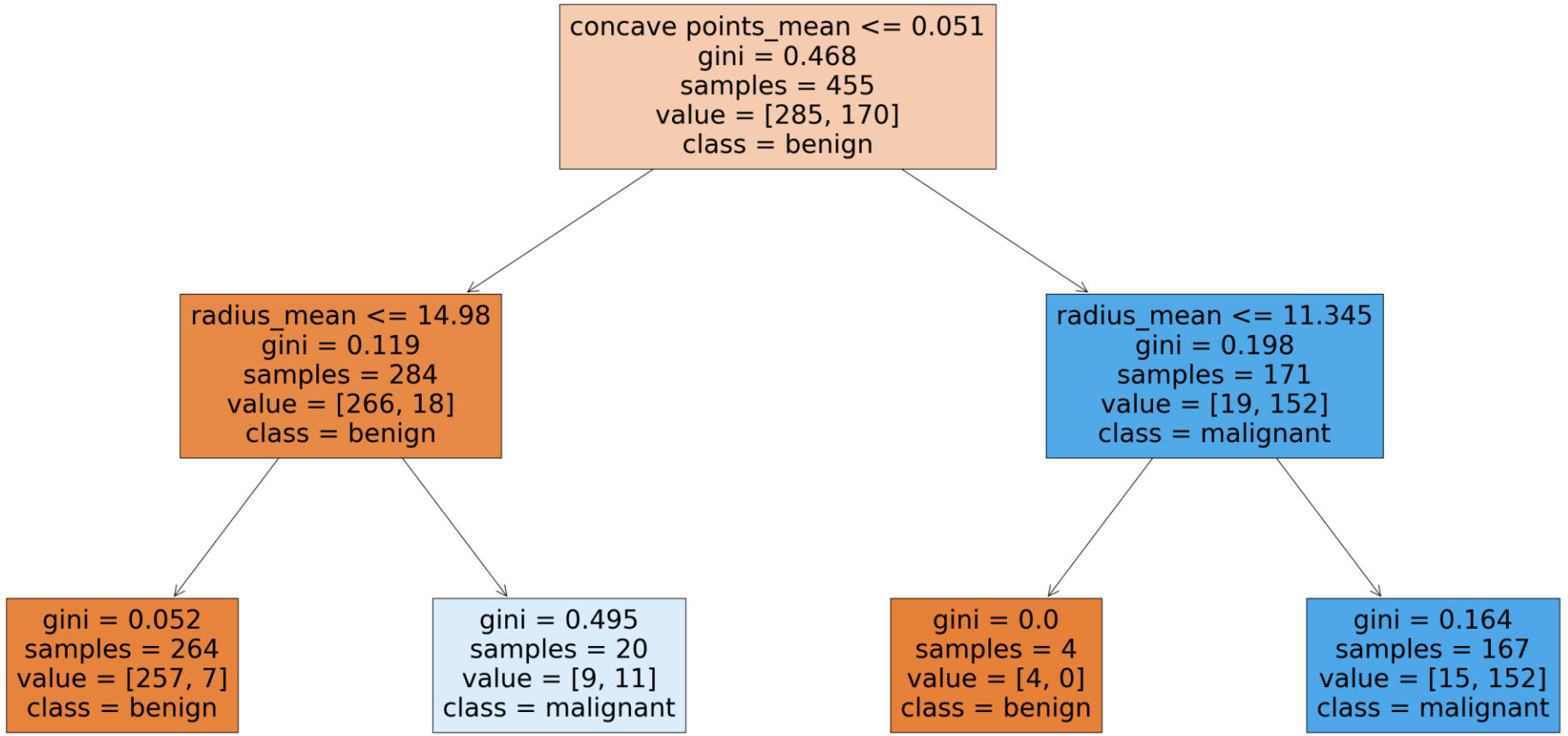
Decision Tree construction/learning

- Nodes are grown recursively.
- The obtention of an internal node or a leaf depends on the state of its predecessors.
- At each node, split the data based on:
  - Feature f and split-point sp to maximize criteria (gini-index, info-gain).

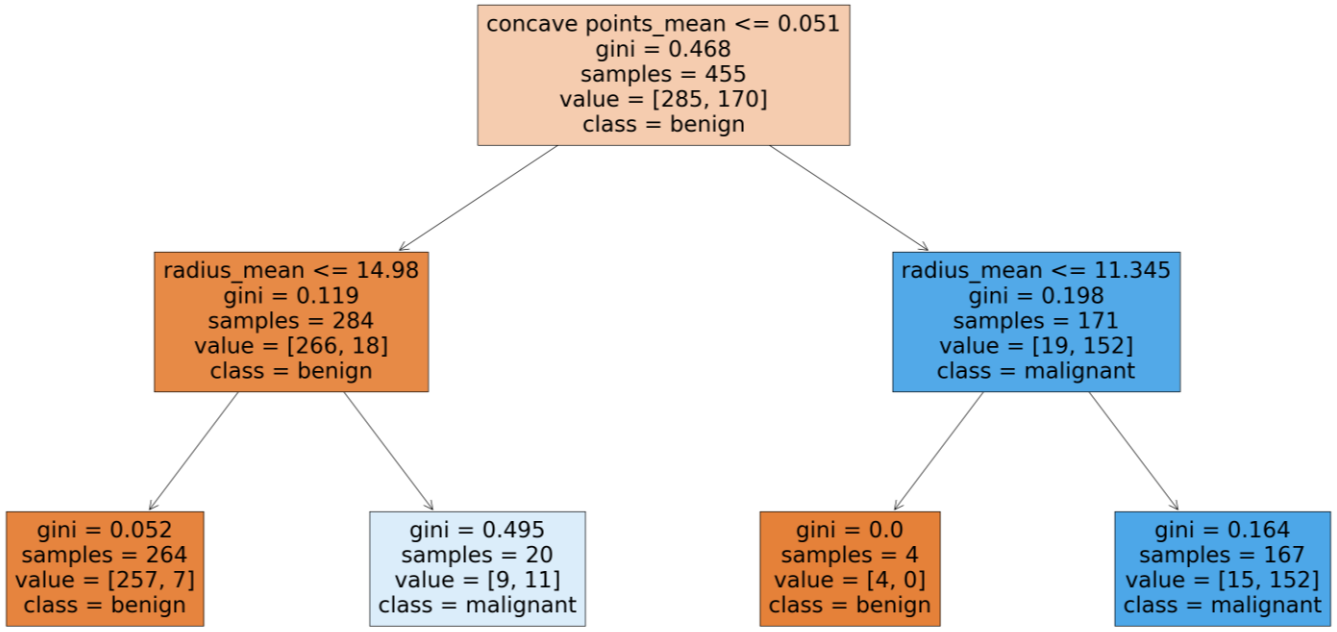


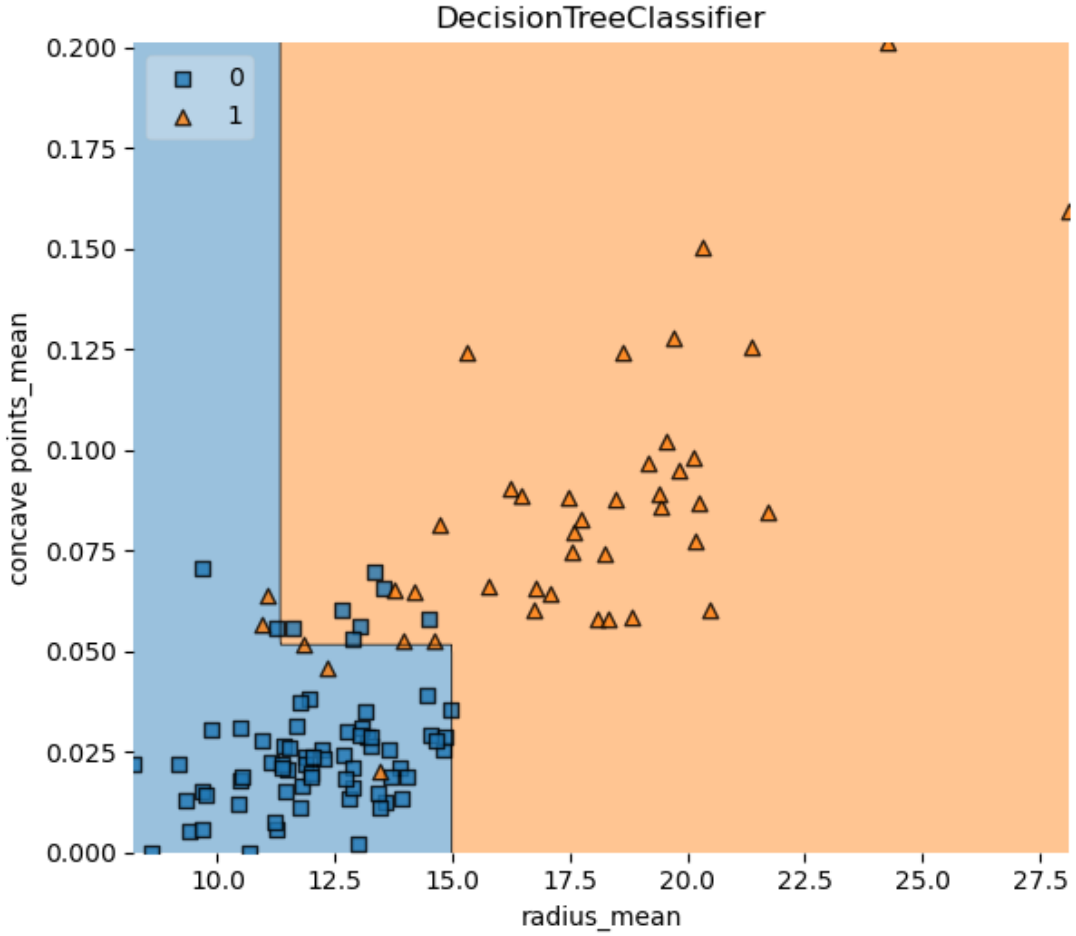
Decision Regions

Decision regions are separated by surfaces called decision-boundaries (SVM)

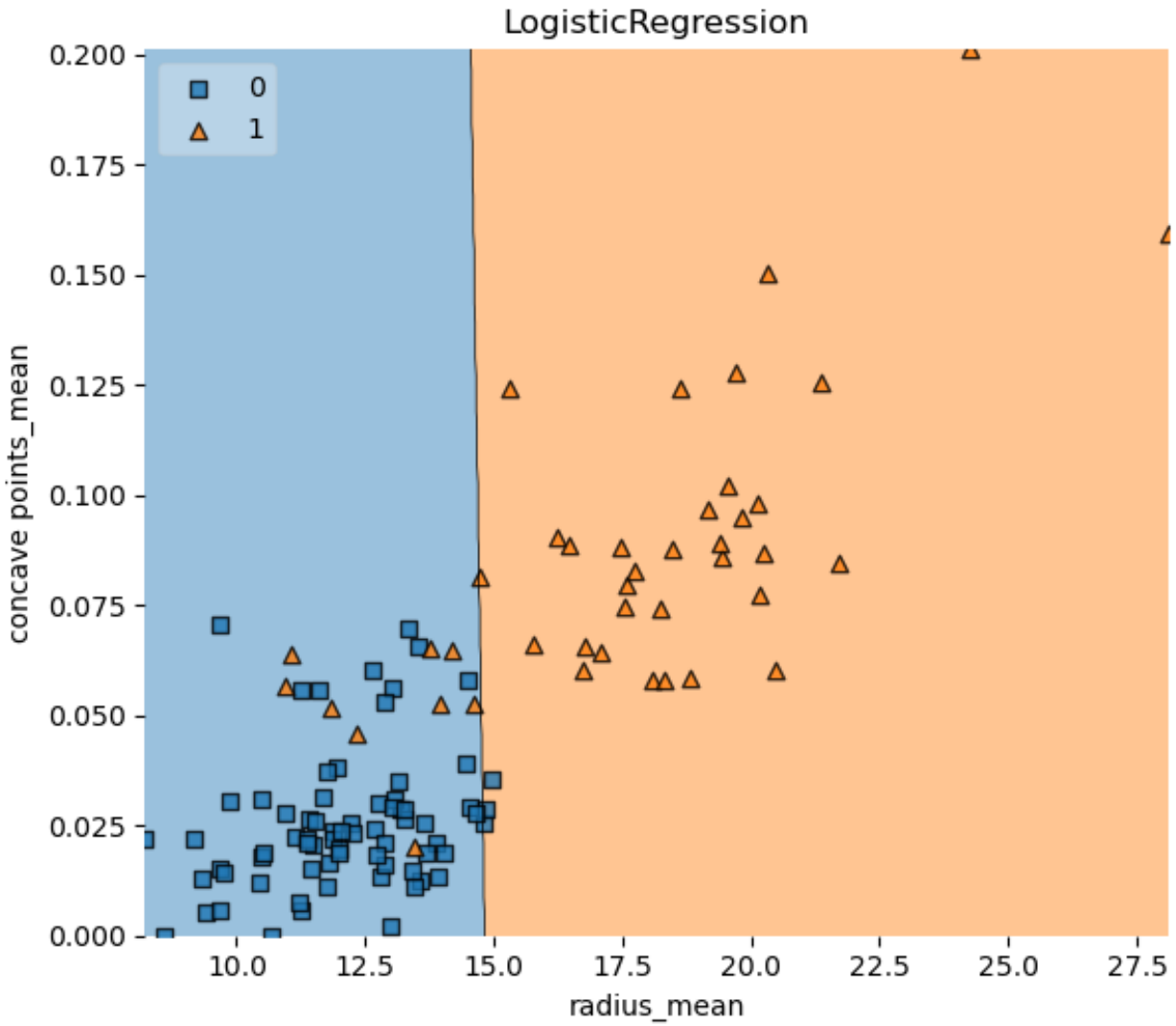


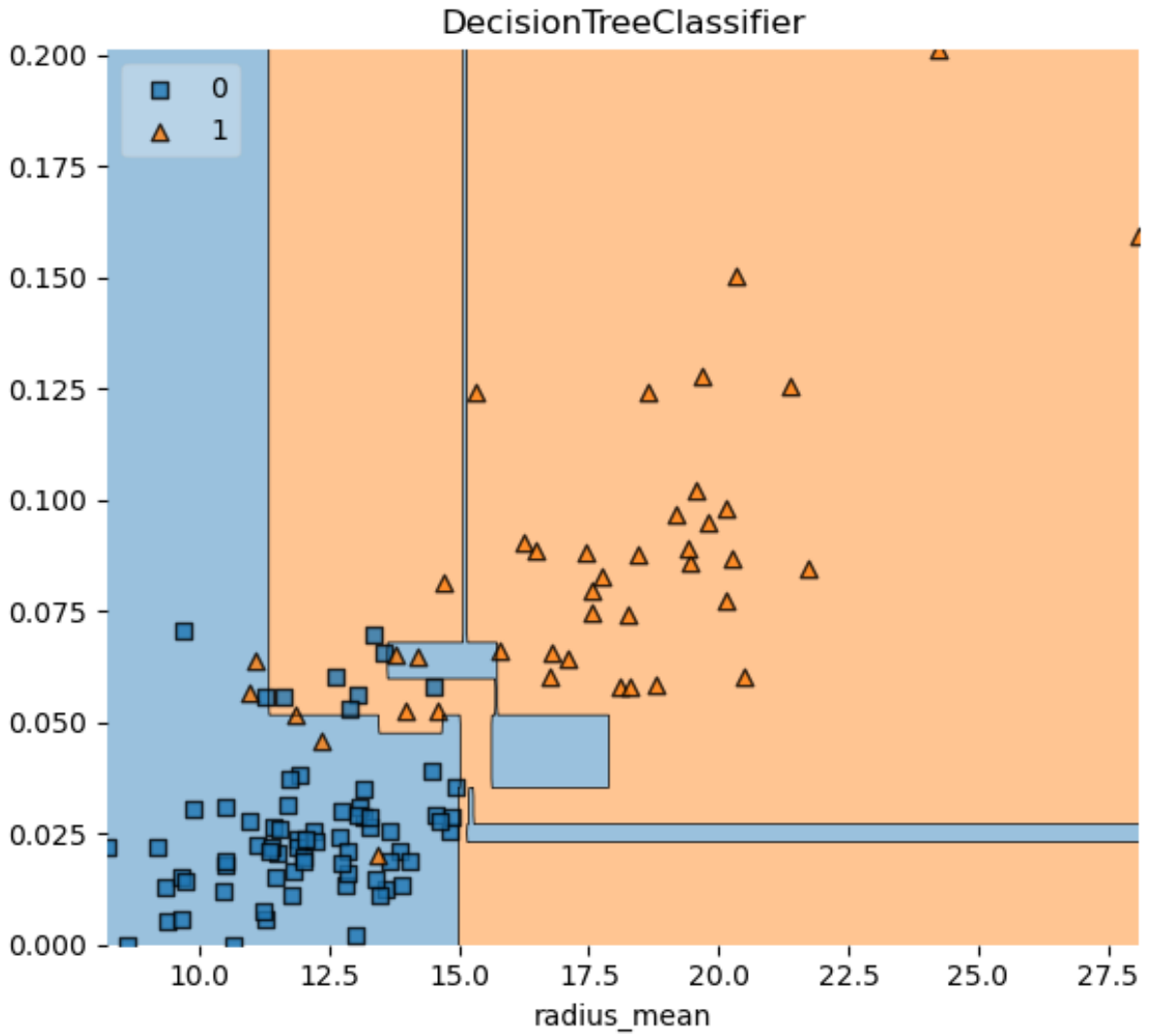
Decision boundaries



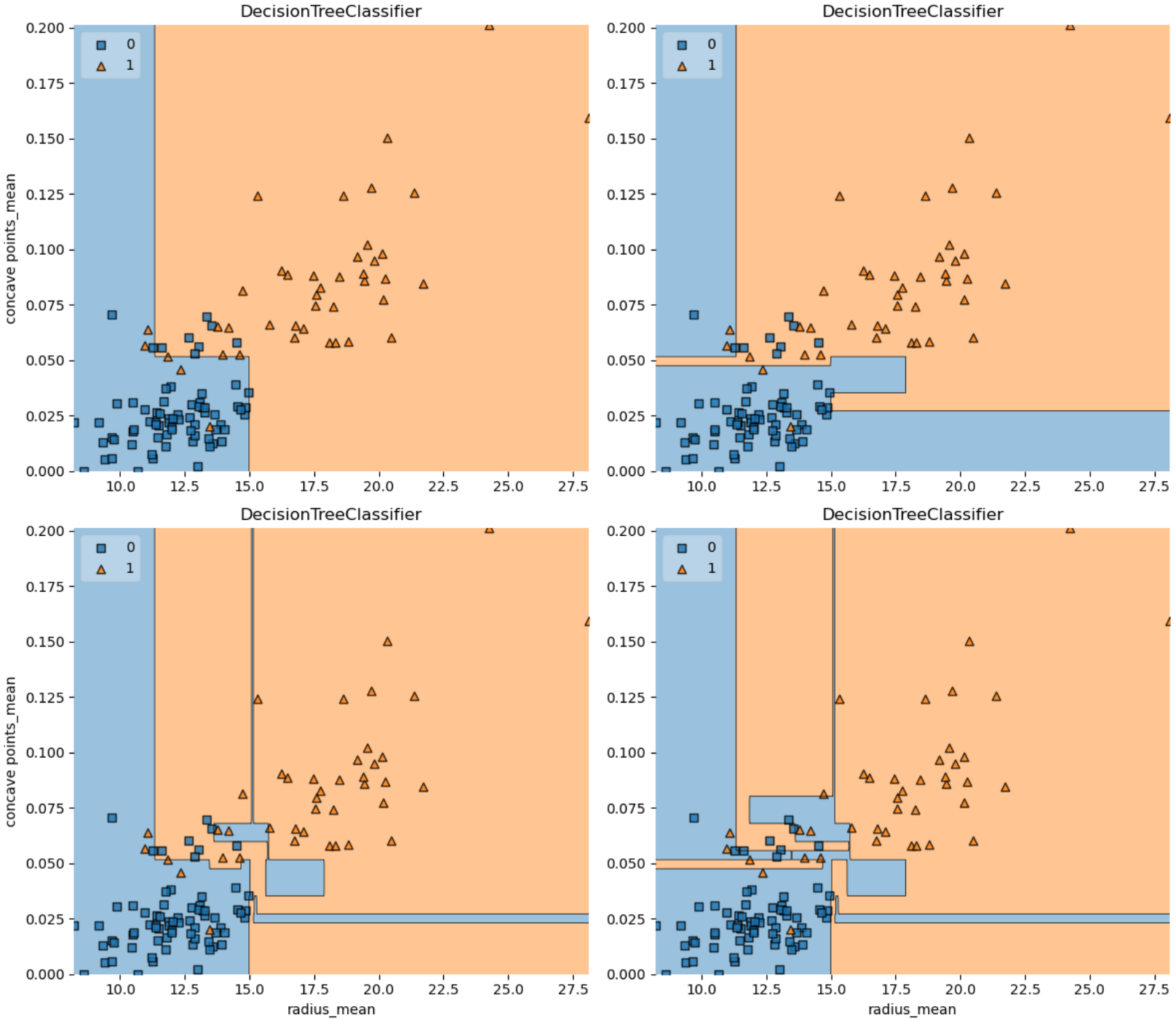


Classifier (binary)

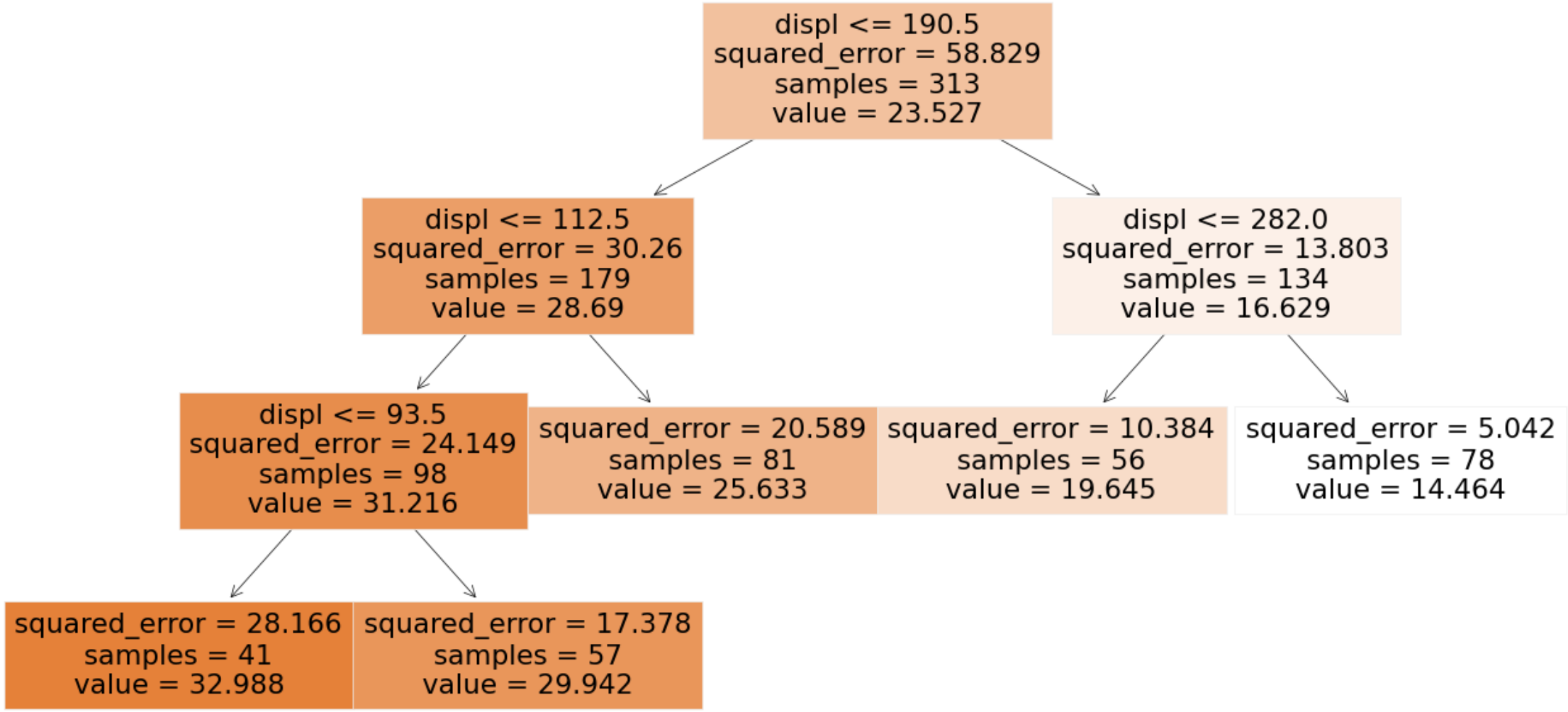




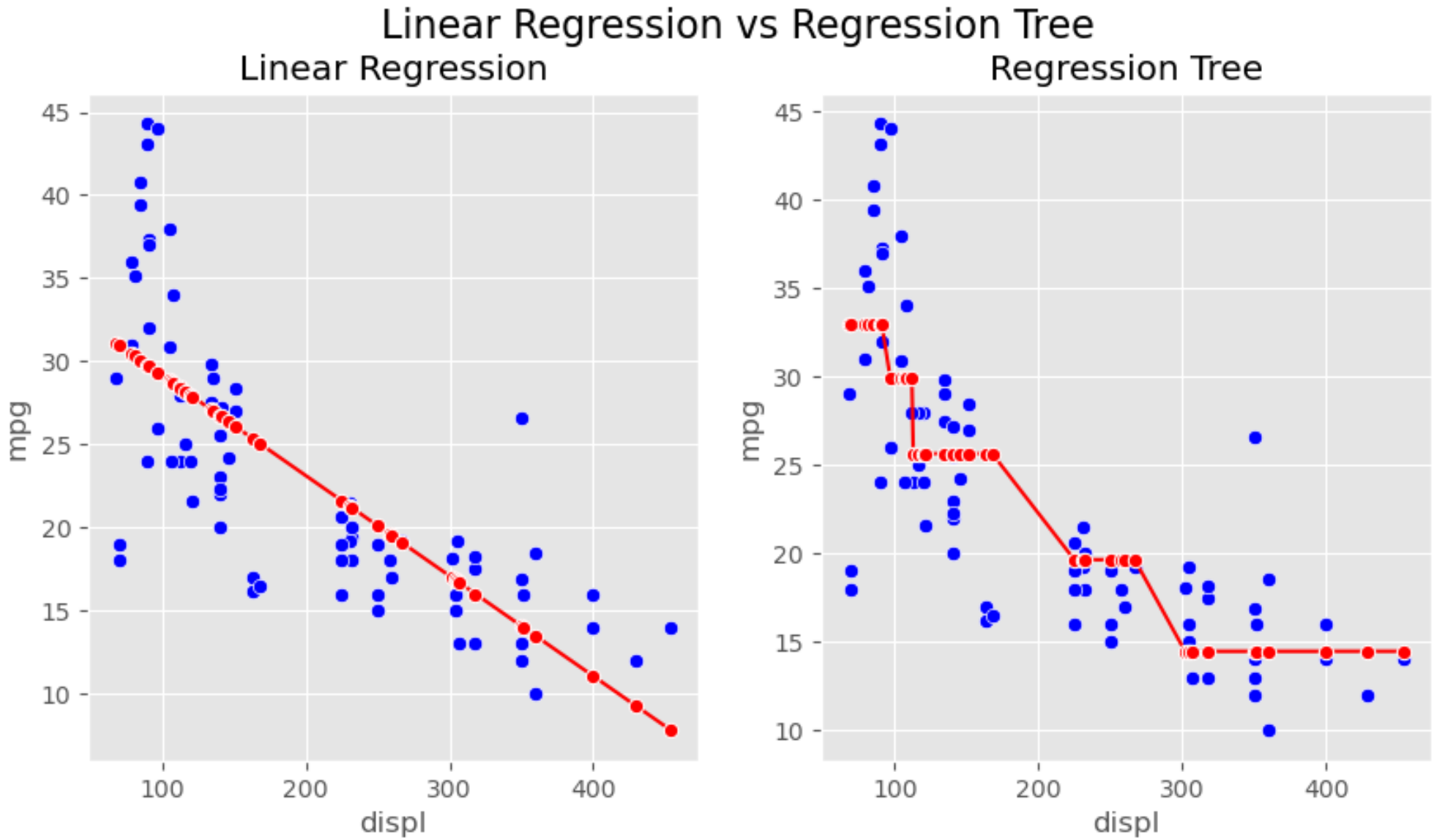
Model complexity



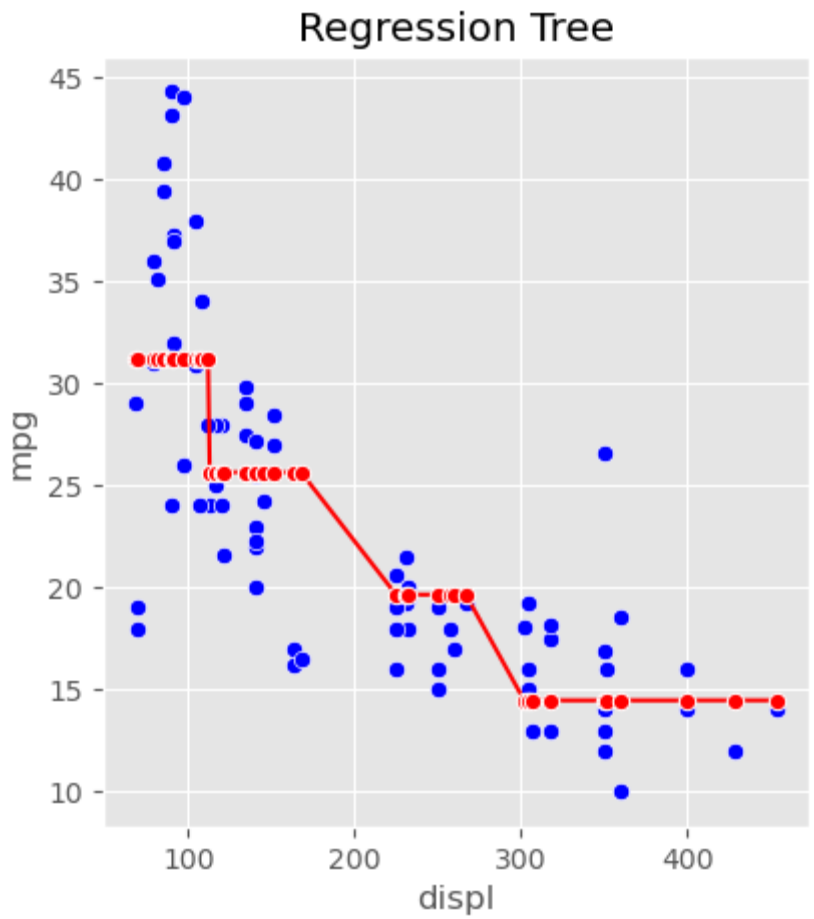
Regressor

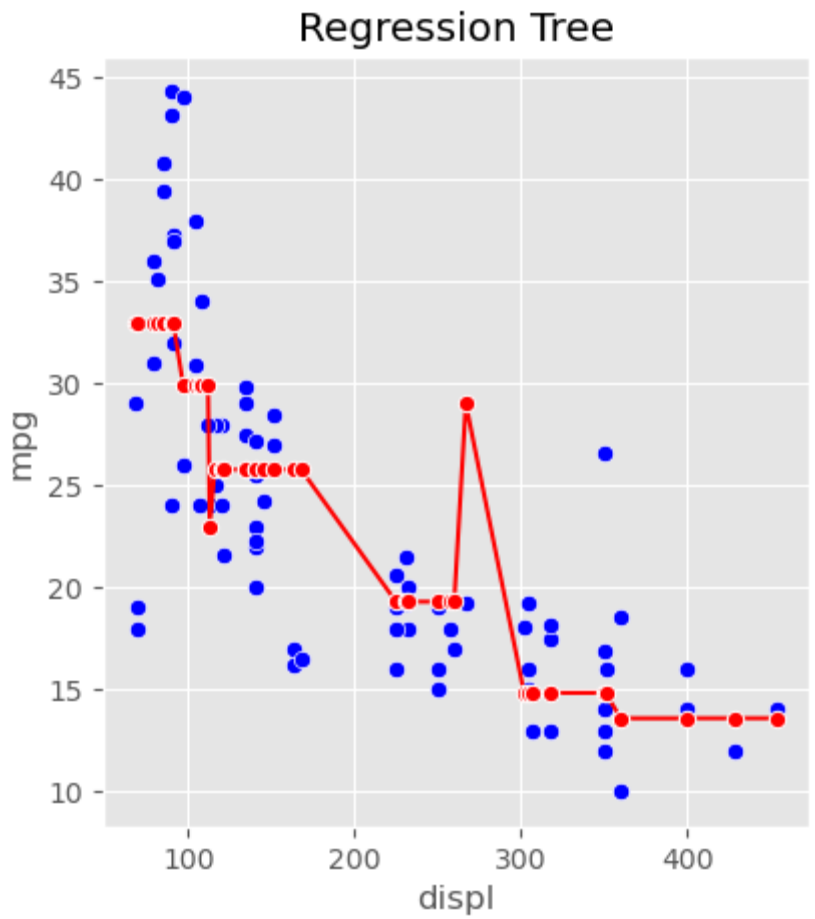

displ <= 190.5
squared_error = 58.829
samples = 313
value = 23.527
displ <= 112.5
squared_error = 30.26
samples = 179
value = 28.69
displ <= 282.0
squared_error = 13.803
samples = 134
value = 16.629
displ <= 93.5
squared_error = 24.149
samples = 98
value = 31.216
squared_error = 20.589
samples = 81
value = 25.633
squared_error = 10.384
samples = 56
value = 19.645
squared_error = 5.042
samples = 78
value = 14.464
squared_error = 28.166
samples = 41
value = 32.988
squared_error = 17.378
samples = 57
value = 29.942

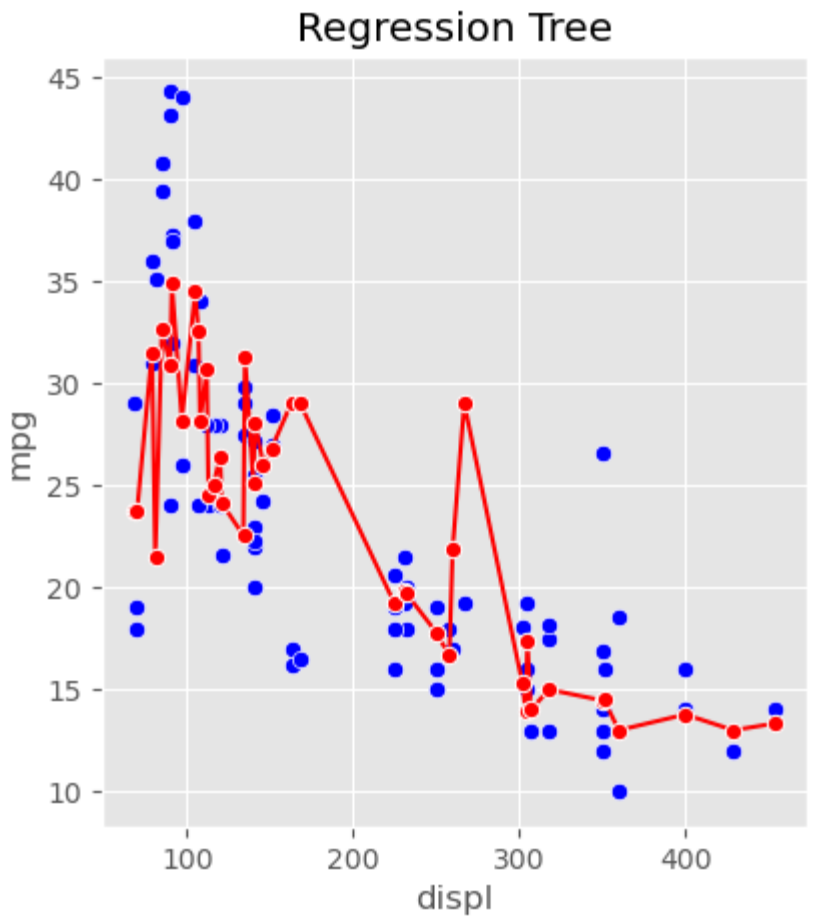
Regressor


Linear Regression vs Regression Tree
Linear Regression
Regression Tree
mpg
displ
45
40
35
30
25
20
15
10
100
200
300
400

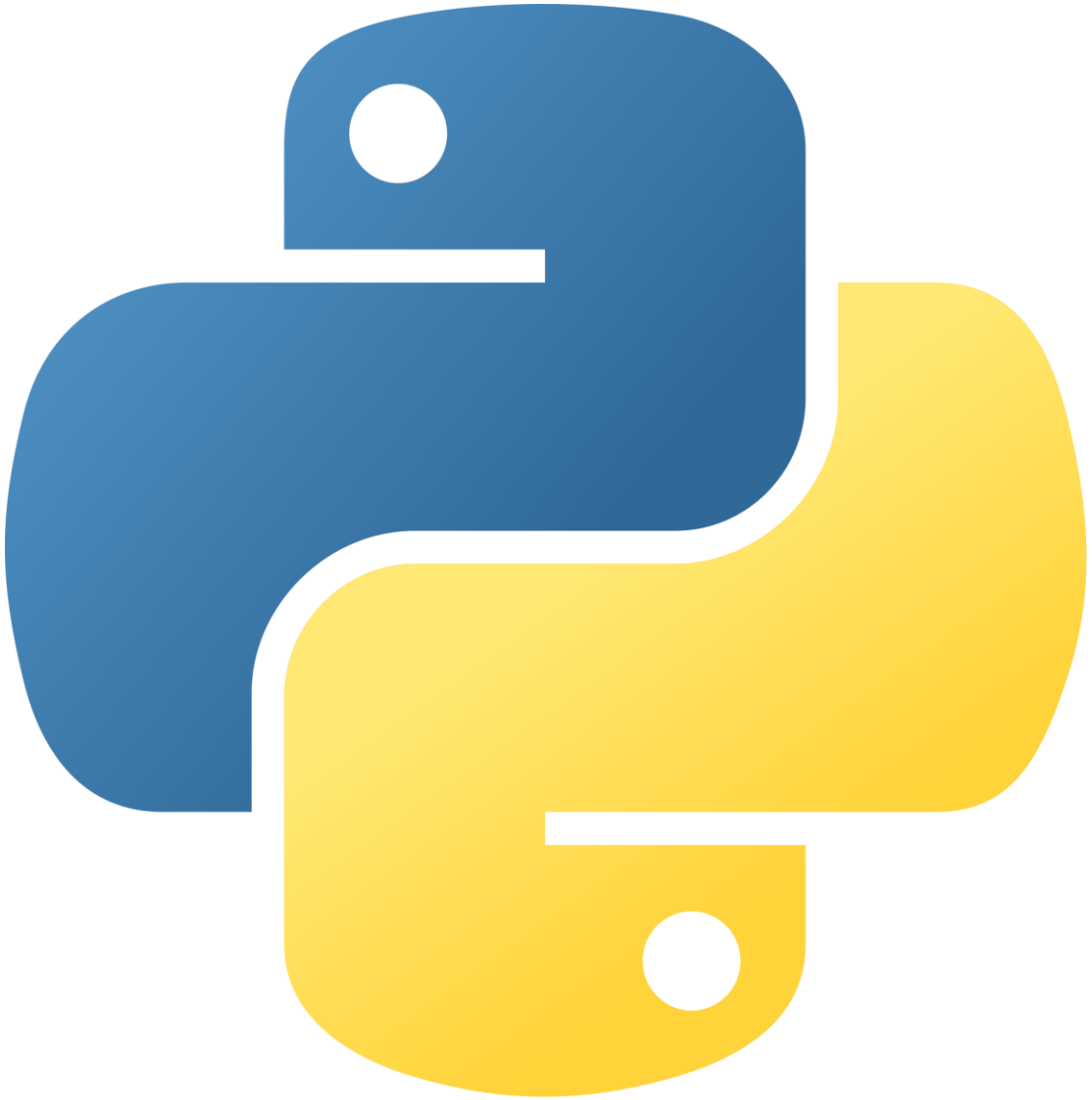
Model complexity







Code



Bibliography

https://github.com/kamperh/data414
https://polakowo.io/datadocs/docs/machine-learning/tree-based-models
https://towardsdatascience.com/why-random-forests-outperform-decision-trees-1b0f175a0b5
https://www.statistics.cool/post/why-do-random-forests-work/
https://luisvalesilva.com/datasimple/random_forests.html

Information

.

- **Low Probability Event**: High Information (*surprising*).
- **High Probability Event**: Low Information (*unsurprising*).

> *The basic intuition behind information theory is that learning that an unlikely event has occurred is more informative than learning that a likely event has occurred.*

— Page 73, Deep Learning, 2016.

Rare events are more uncertain or more surprising and require more information to represent them than common events.

Information

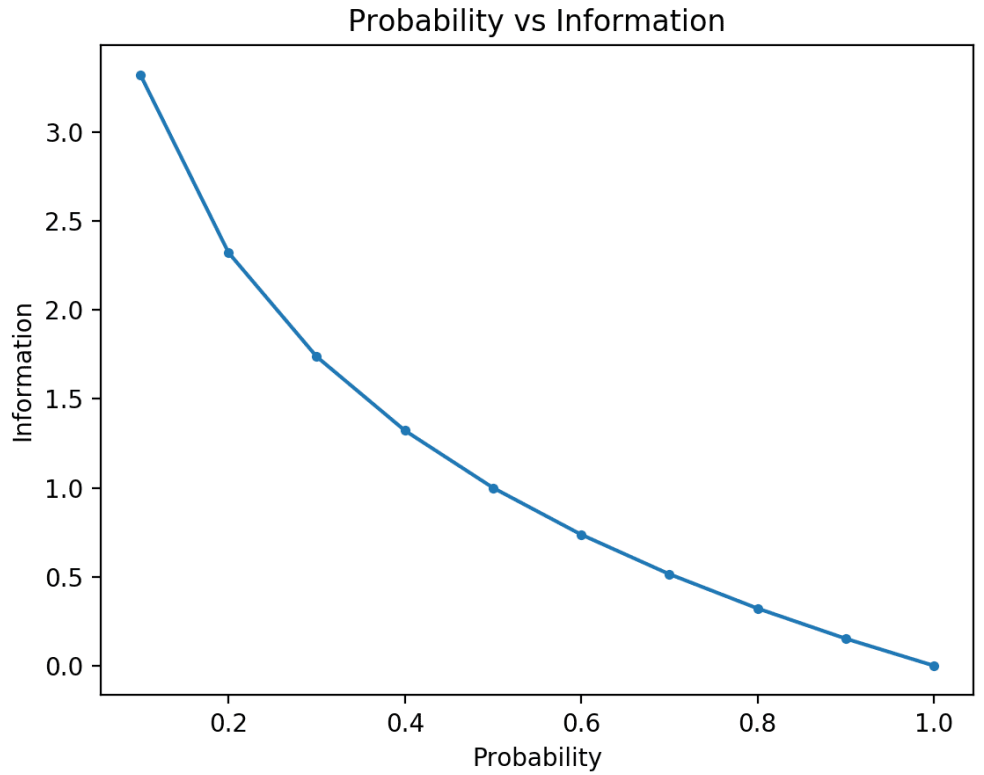
.

The calculation of information is often written as *h()*; for example:

- h(x) = -log(p(x))

The negative sign ensures that the result is always positive or zero.

Information will be zero when the probability of an event is 1.0 or a certainty, e.g. there is no surprise.



Information Gain

TODO

Entropy

TODO

more tree based models

TODO