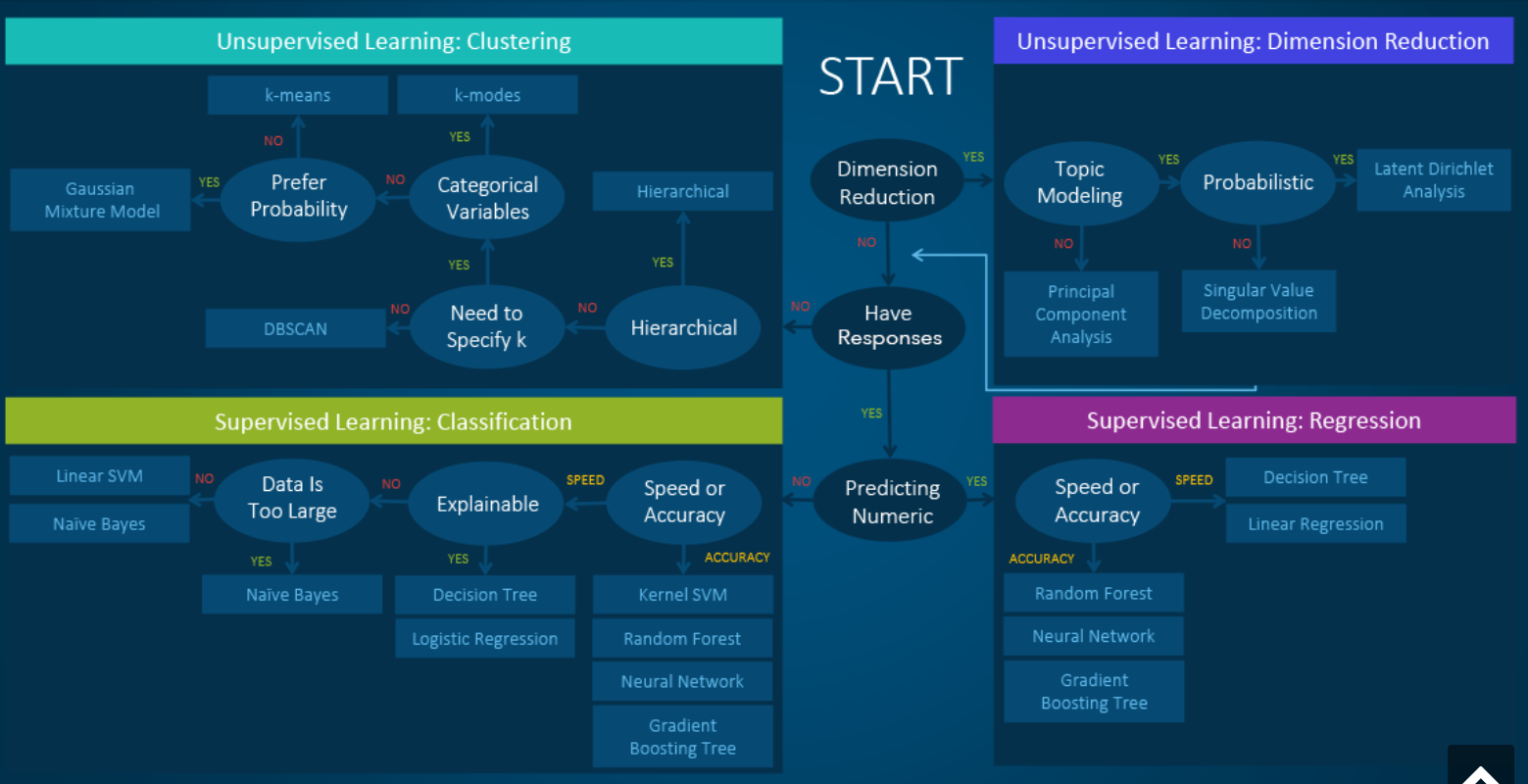
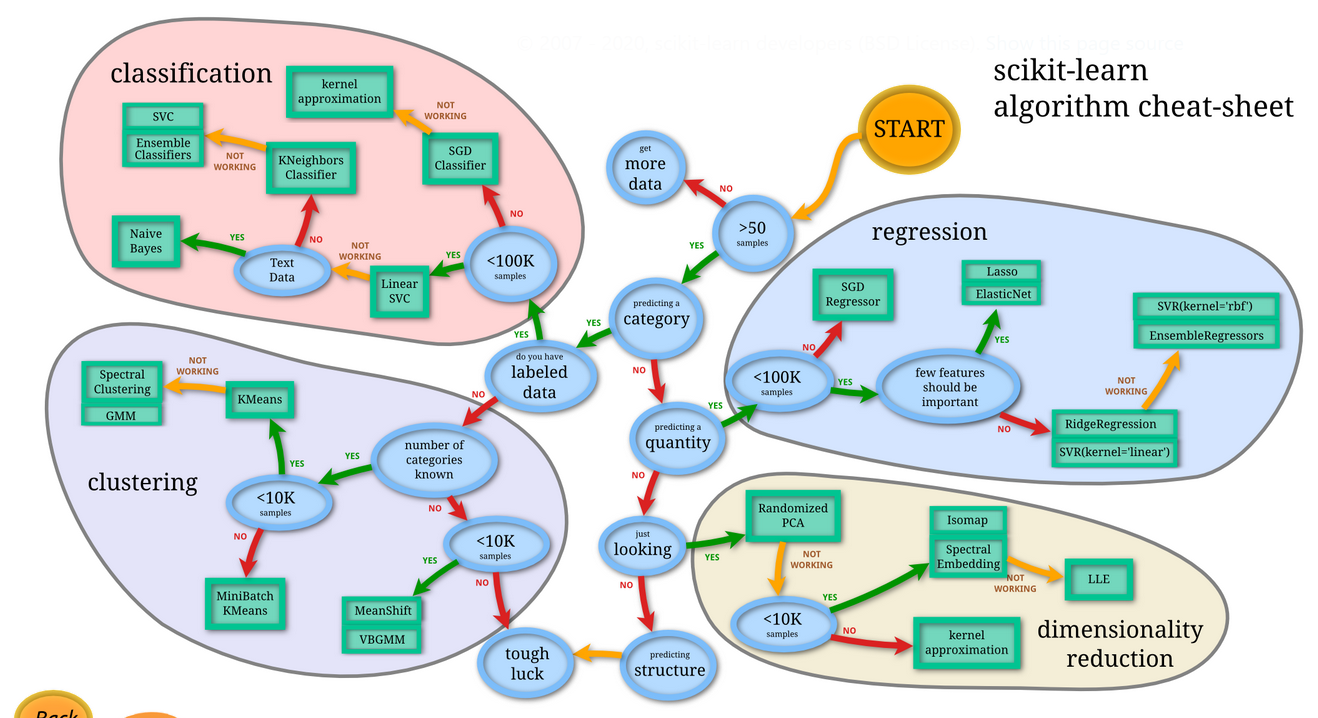
Sas cheat sheet



Python cheat sheet



What is difference between clustering and classification

Medium KNN article

https://towardsdatascience.com/how-to-find-the-optimal-value-of-k-in-knn-35d936e554eb

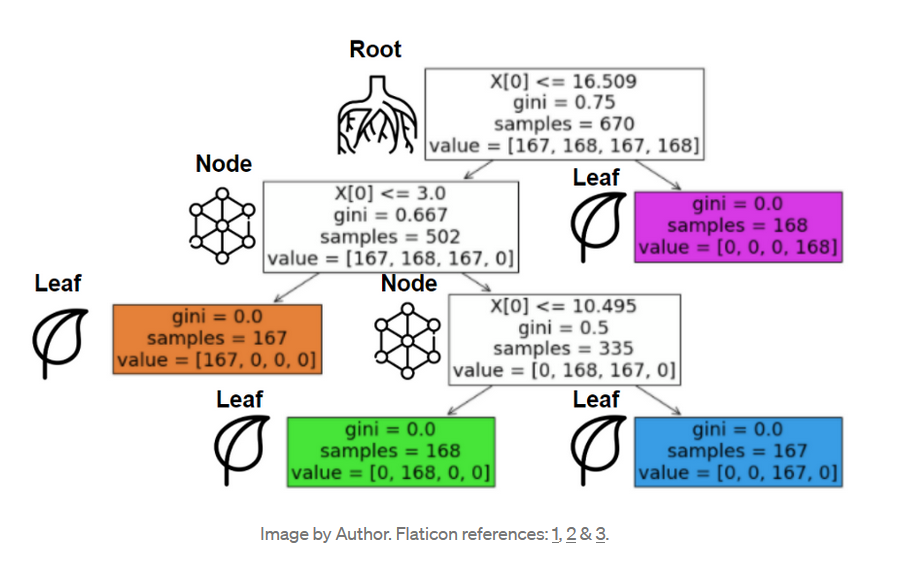
Chart, scatter chart

Description automatically generated

Normally starts from a smaller number as k, for example 2 or 3

Decision tree parameter article

<https://towardsdatascience.com/decision-trees-as-you-should-have-learned-them-99862469493e>



**Criteron parameter:**

Gini impurity: how often mistakenly labelled?

Entropy: how uncertain the outcome is?

**Splitter parameter:**

Best: best for the node

Random: poor for the node

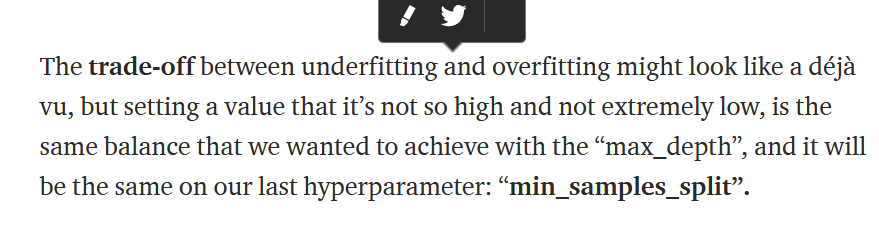
Now you know that the algorithm looks by default the best split on every iteration, but that’s not necessarily the optimal way to find the **global optimum**. Using a random split could lead to some variability, which sometimes helps us find a better tree.

**Max\_dept parameter:** **trade off overfitting vs underfitting**

According to experience, to balance between over-fitting and under-fitting

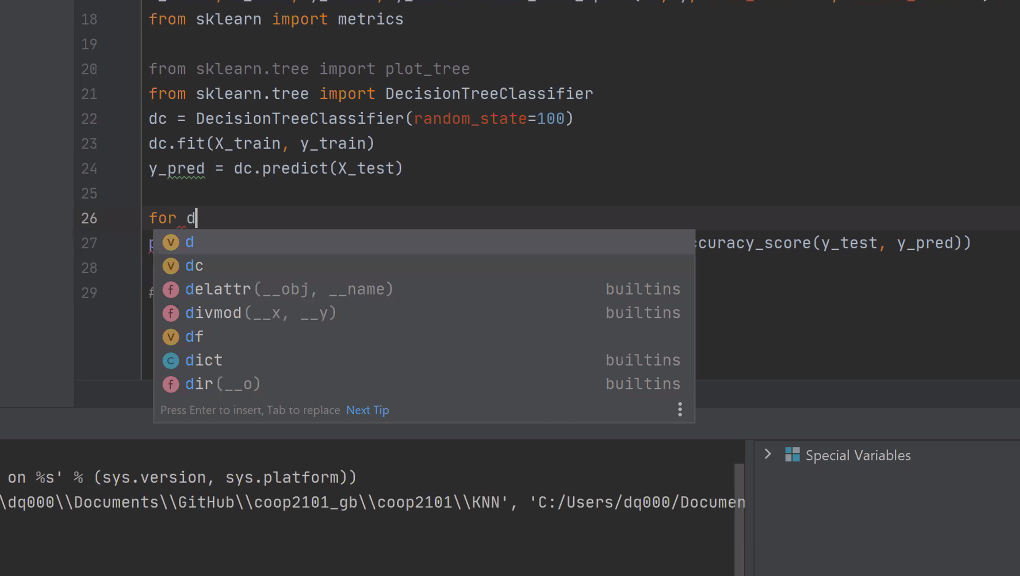
**Min\_impurity\_decrease: trade off overfitting vs underfitting**

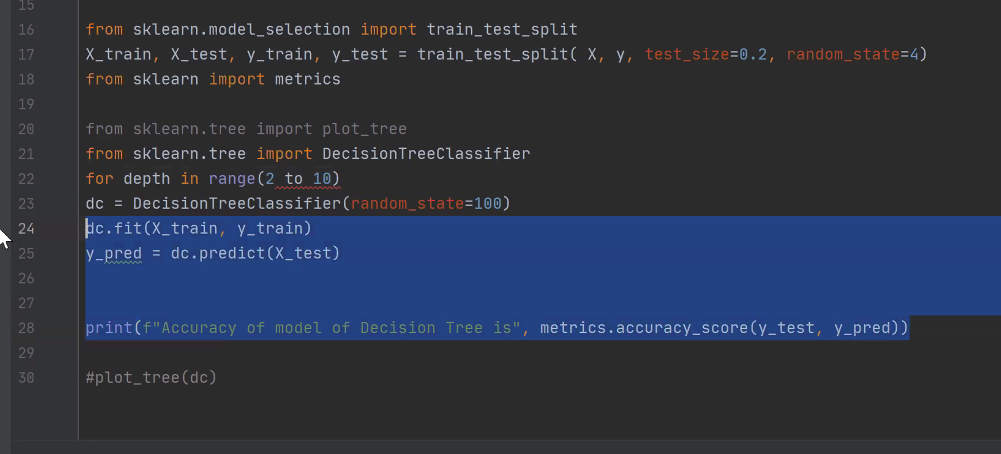
Min impurity decrease is a way to control if a split it’s really worth it. There are times that, if we keep splitting our nodes, it will do more harm than good. If we don’t use this parameter, the model could tend to overfit the data.



# Min\_samples\_split: trade off overfitting vs underfitting

# A integer or a fraction to initial sample size





Random forest