HOW TO CHOOSE THE RIGHT MACHINE LEARNING ALGORITHM?

With cheat sheets

@sanjushasuresh

There is no master algorithm for all situations. Choosing an algorithm depends on the following questions:

- How much data do you have, and is it continuous or categorical?
- Is the problem related to classification, association, clustering, or regression?
- Predefined variables (labeled), unlabeled, or mix?
- What is the goal?

Follow these guidelines:

- If accuracy is a concern, test different algorithms and cross-validate them.
- If the training dataset is small, use models that have low variance and high bias. (Model might become underfit)

Parametric or linear ML algorithms often have low variance and high bias. A few of them are:

1. Linear Regression
2. Linear Discriminant Analysis
3. Logistic Regression

- If the training dataset is large, use models that have high variance and low bias. (Model might become overfit)

Non-parametric or non-linear ML algorithms often have high variance and low bias. A few of them are:

1. Decision Trees
2. K-Nearest Neighbors
3. Support Vector Machines

The best model is one where bias and variance are both low. This is called the bias-variance tradeoff.

@sanjushasuresh

Classification is used when your target is categorical.

Examples of classification problems include:

- Predicting yes or no
- Estimating gender
- Breed of an animal
- Type of color

Regression is used when your target variable is continuous and a value needs to be predicted.

Examples of regression problems include:

- Estimating sales and price of a product
- Predicting the score of a team
- Predicting the weather
- Sales forecasting

Clustering is used when data needs to be organized to find patterns.

Examples of clustering problems include:

- Identifying fake news
- Spam filtering
- Anomaly detection
- Customer Segmentation

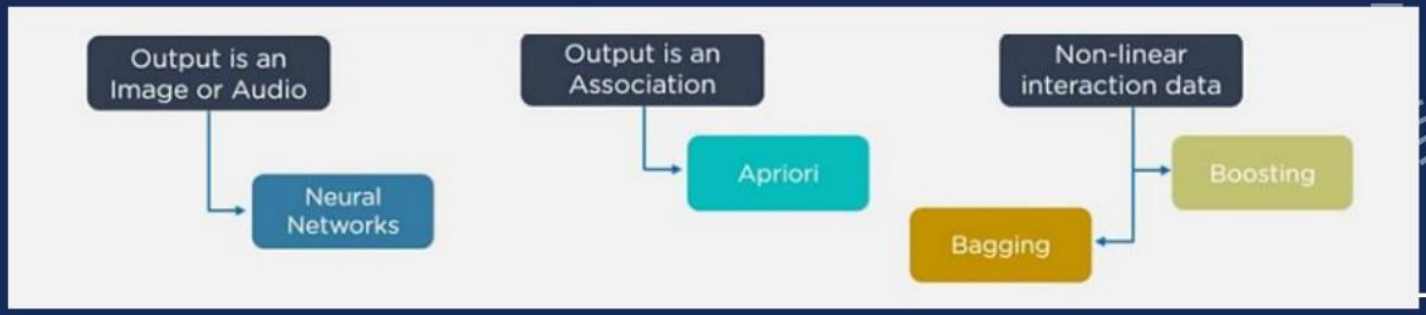
Association is used to discover interesting relations between variables in large datasets.

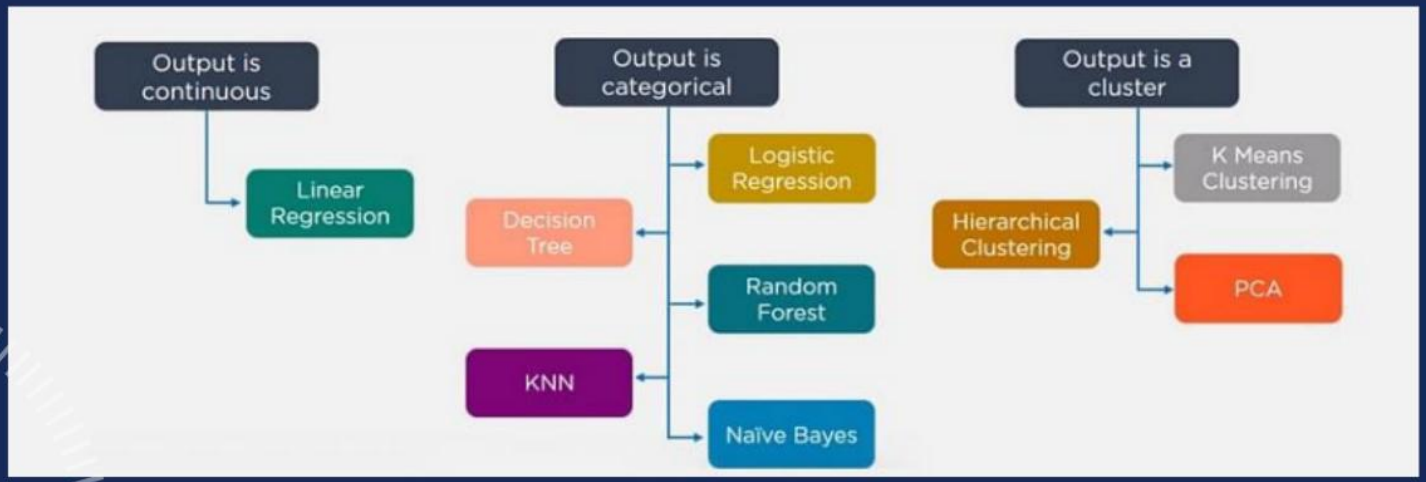
Examples of association problems include:

- Market basket analysis
- Web usage mining
- Continuous production
- Medical Diagnosis

@sanjushasuresh

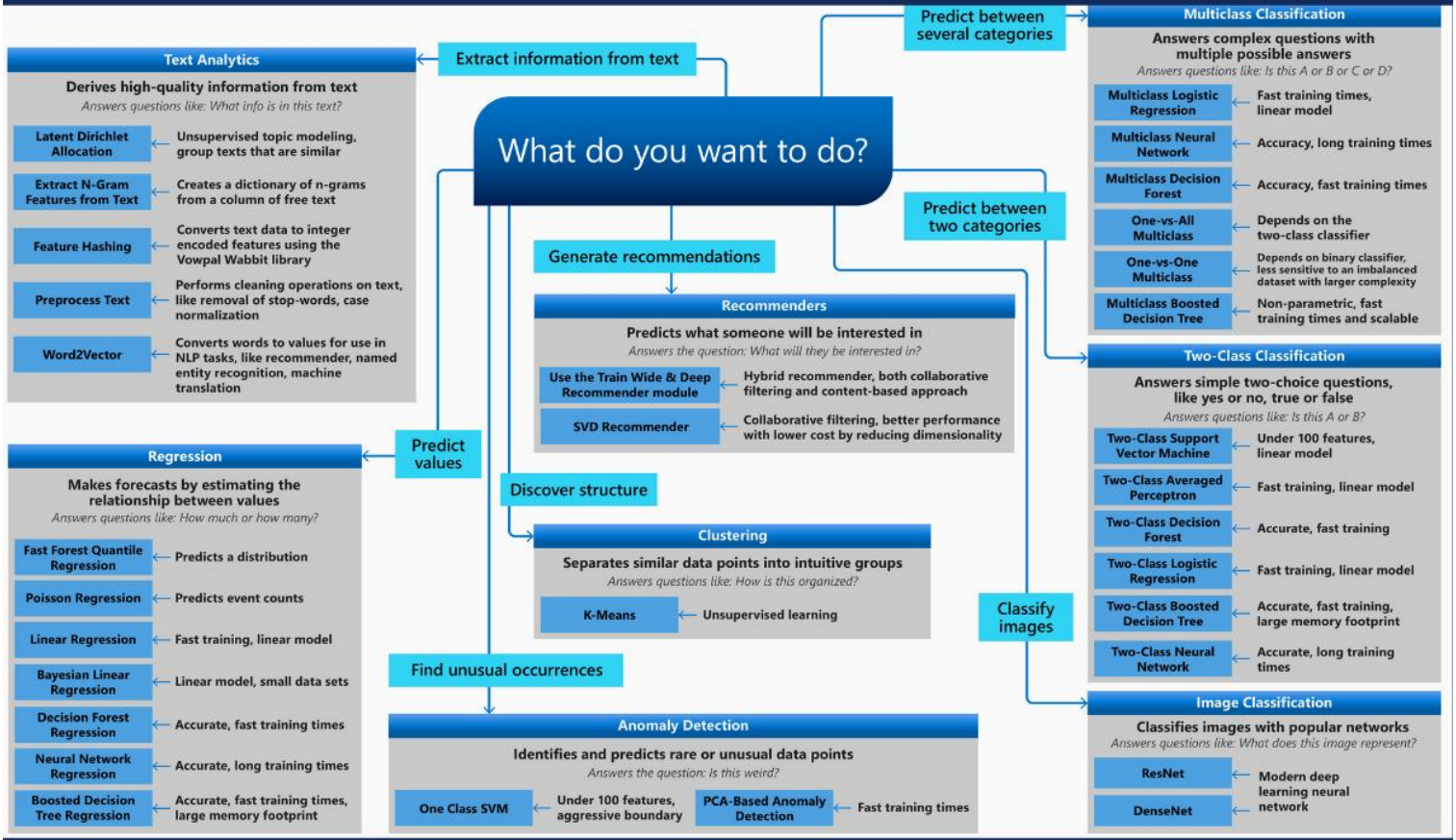
Based on the output, the following algorithms can be used:



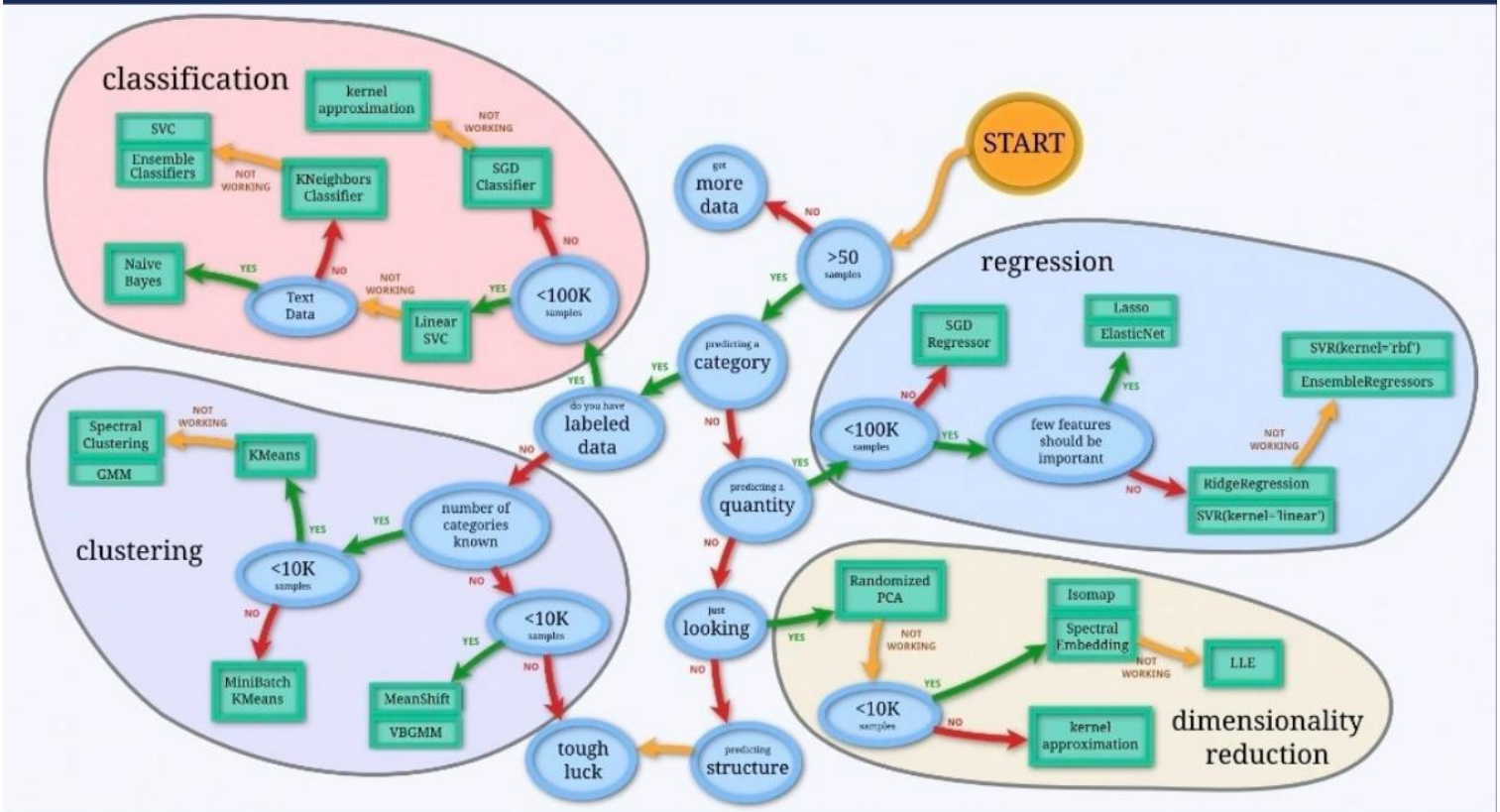


@sanjushasuresh

Below are two cheat sheets for algorithm selection:



@sanjushasuresh



@sanjushasuresh