Cheat Sheet: Where To Use Which Algorithm

Analytics Vidhya

Supervised Machine Learning Algorithms

- Supervised ML Algorithms are used where a clear target variable is present
- If the target variable is continuous, we use regression algorithms
- If the target variable is categorical, we use classification algorithms
- Some algorithms work for both classification and regression

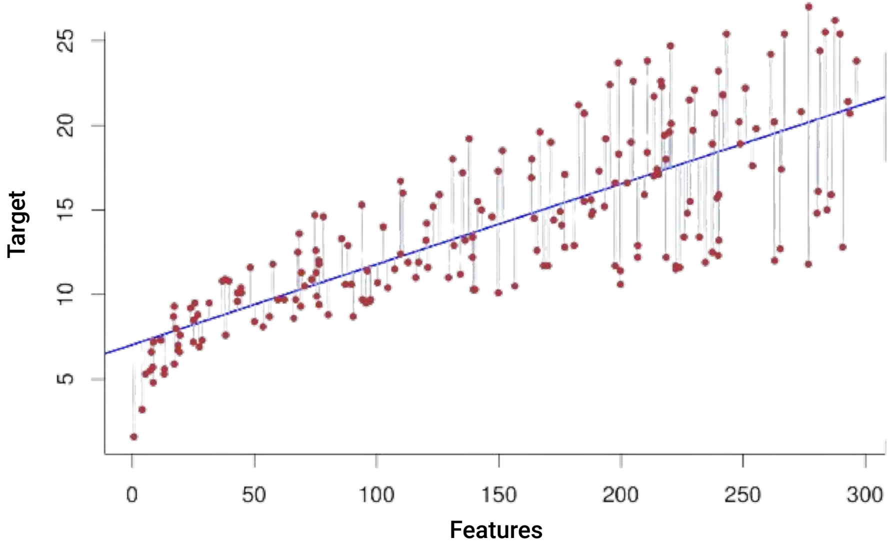
Linear Regression

What is it?

This model assumes that there is a linear relationship between the Target variable and the features.

When to use?

Used for predicting continuous values based on a set of independent variables. Suitable for situations where there is a linear relationship between the variables and an explainable model is required.



Can be used for Regression Only

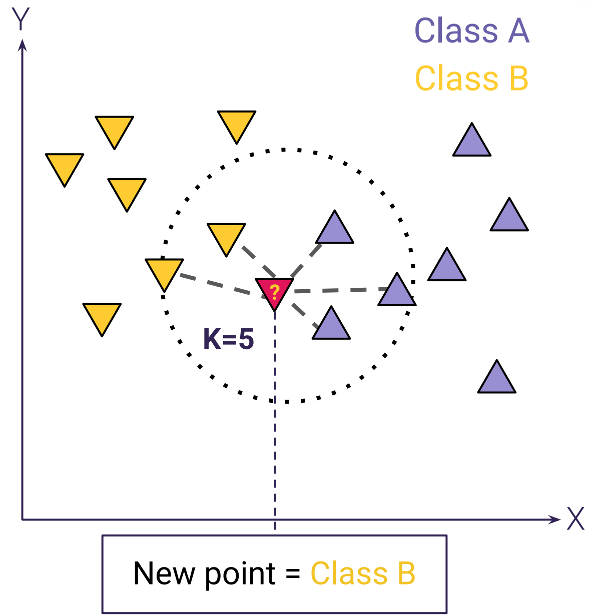
K-Nearest Neighbors (KNN)

What is it?

The KNN algorithm classifies new data based on the class of its neighbors. The algorithm looks at K closest neighbors based on the distance and attributes the class based on the majority of neighbors.

When to use?

It is suitable for small to medium-sized datasets datasets with non-linear relationships.



New point = Class B

Can be used for Classification and Regression

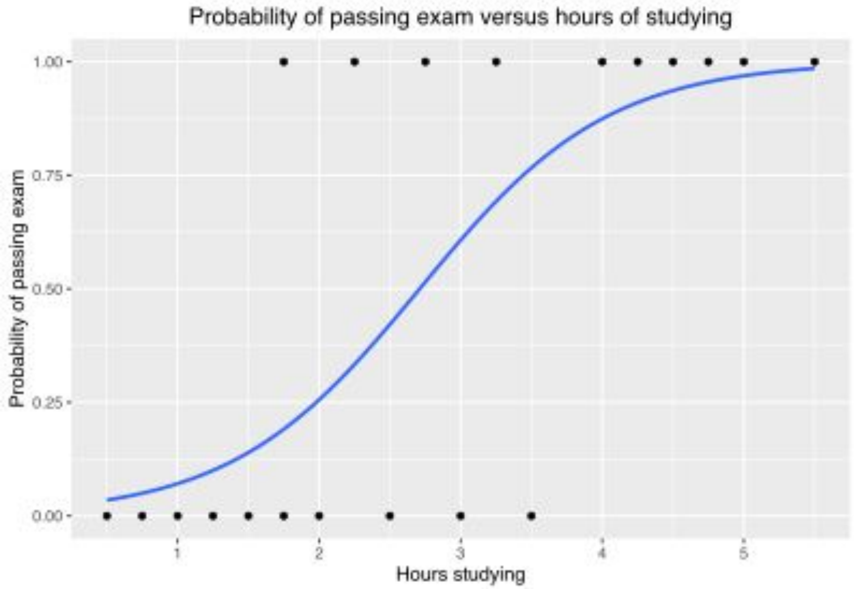
Logistic Regression (Classification)

What is it?

It is a modified version of linear regression algorithm. The end output of Linear Regression model is converted to binary i.e. it is either 1 or 0, using a sigmoid function.

When to use?

Ideal for binary classification problems, predicting outcomes that fall into one of two classes. It can be extended for multi-class classification as well.



Can be used for Classification only

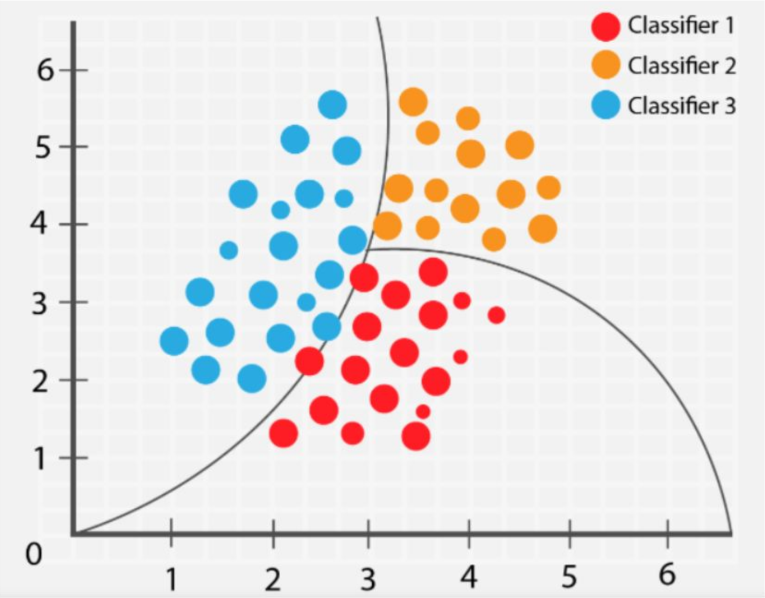
Naive Bayes

What is it?

It is one of the fastest algorithms used to predict a class of datasets.This algorithm works on the Bayes' theorem on conditional probability.

When to use?

It is popularly used for text classification problems because text classification problems involve huge datasets.



Can be used for Classification only

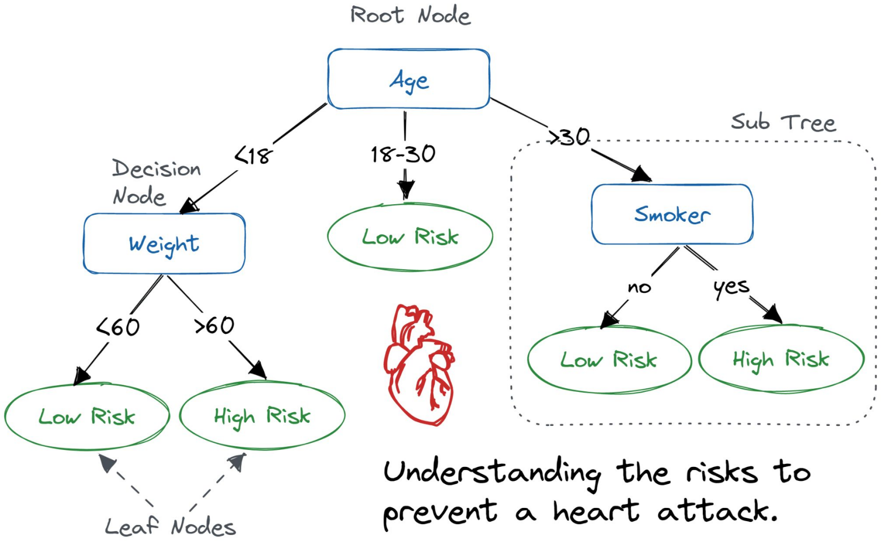
Decision Tree

What is it?

The model is designed similar to the structure of a tree starting with root nodes, splitting into branches and then ending with leaf nodes as decisions are made.

When to use?

They are suitable for large datasets with complex interactions and non-linear relationships.



Understanding the risks to prevent a heart attack.

Can be used for Classification and Regression

Analytics Vidhya

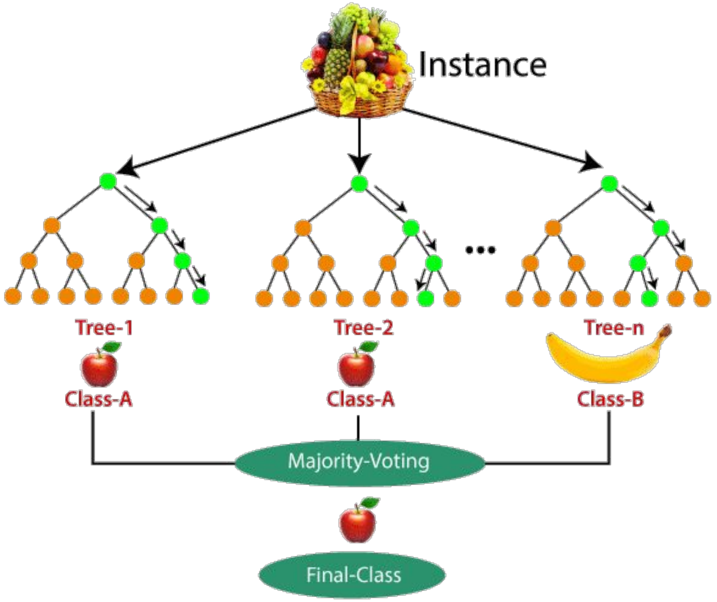
Random Forest

What is it?

Multiple decision trees are made and the output of each decision tree is considered. The final decision is taken based on the majority for classification problems and average for regression problems.

When to use?

Random Forest can be tried in situations where a decision tree is suffering from overfitting problems.



Can be used for Classification and Regression

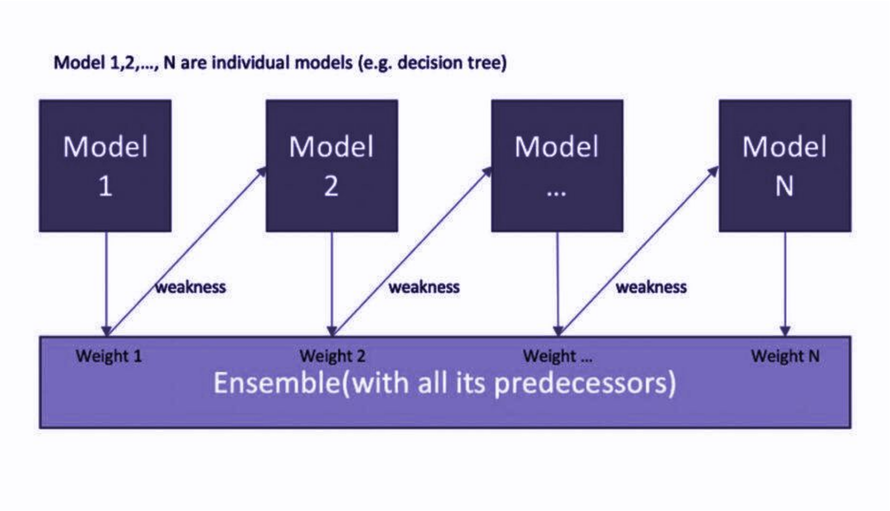
Adaptive Boosting (Ada Boost)

What is it?

It builds a model in sequential steps. It starts by giving equal weightage to all the data points. Then with each step it assigns higher weights to points that are wrongly classified so that the model is forced to perform better on them.

When to use?

This technique is preferred when high performance on evaluation metrics is required.



Model 1,2,..., N are individual models (e.g. decision tree)
Model 1
Model 2
Model ...
Model N
weakness
weakness
weakness
Weight 1
Weight 2
Weight ...
Weight N
Ensemble(with all its predecessors)

Can be used for Classification and Regression

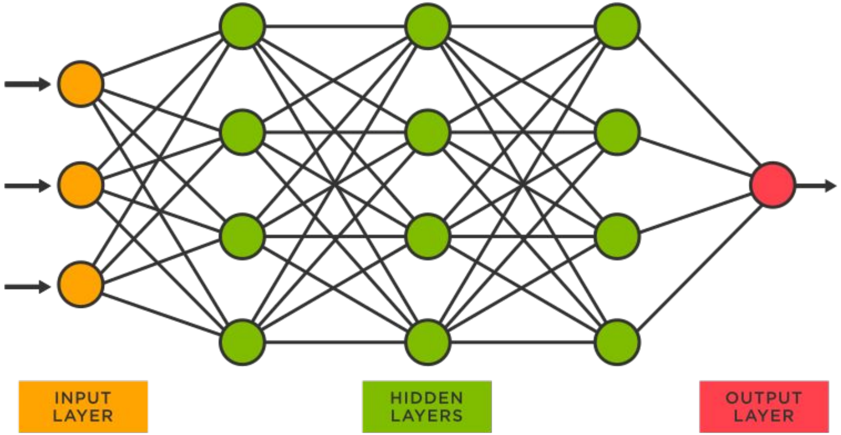
Neural Network

What is it?

It is a complex model that was initially inspired and designed by the way a human brain operates.

When to use?

Neural Networks are preferred in scenarios where complex patterns and relationships exist within the data, and when large amounts of labeled data are available for training.



Can be used for Classification and Regression

Unsupervised Machine Learning Algorithms

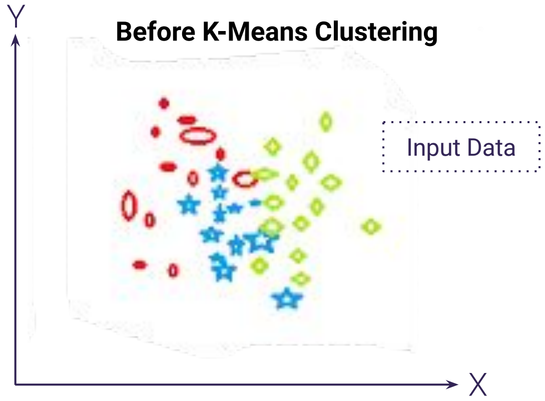
K-Means

What is it?

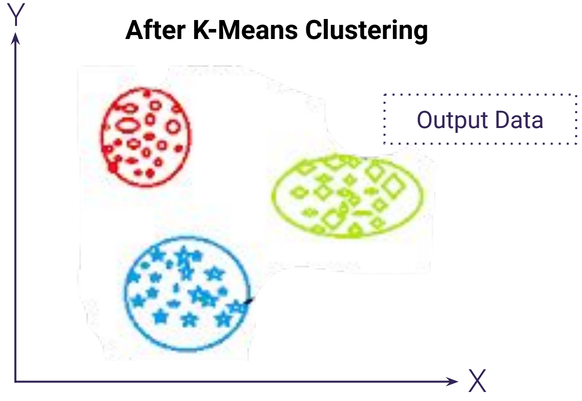
This algorithm takes an unlabeled dataset as input, then proceeds to group the dataset into k-number of clusters.

When to use?

This algorithm is more suitable for unsupervised problems that do not have outliers.



Before K-Means Clustering



After K-Means Clustering

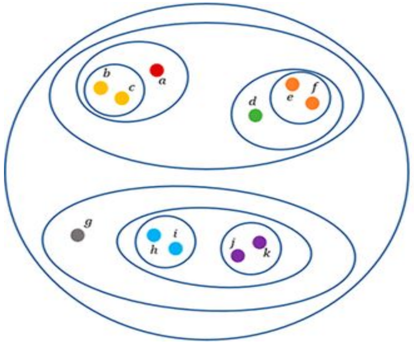
Hierarchical Cluster Analysis (HCA)

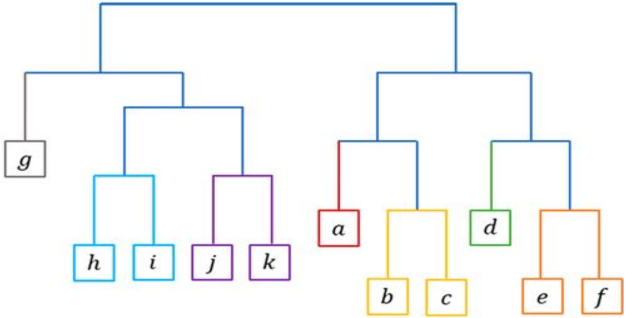
What is it?

It is used to group the unlabeled dataset into clusters. The hierarchy of clusters is in the form of a tree shaped structure called a dendrogram

When to use?

It is most suitable when the data is small in size.





Dendrogram

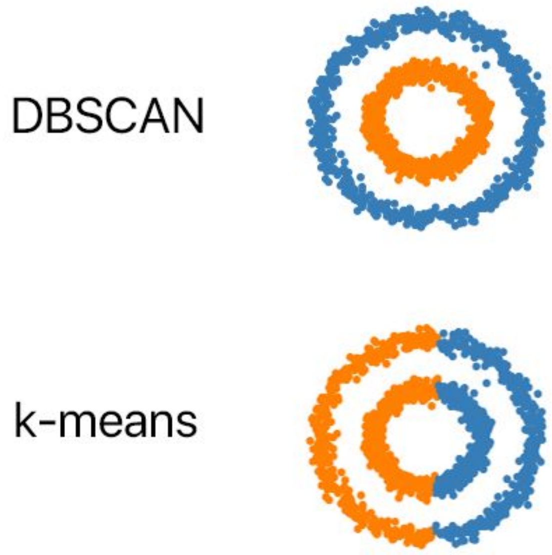
Density-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise **(DBSCAN)**

What is it?

DBSCAN clusters or groups data points based on their density. DBSCAN can automatically detect number of clusters in the data and identify outliers.

When to use?

It is most suitable for datasets that have outliers in them.



Hyperparameter Tuning in Machine Learning

By training a model with existing data, we are able to fit the model parameters. However, there is another kind of parameter, known as ***Hyperparameters***, that cannot be directly learned from the regular training process. We have to change these hyperparameters, based on their function and the impact they have on the model. This process of finding the ideal hyperparameters is called Hyperparameter Tuning. Hyperparameter Tuning can increase the performance of the models.

Hyperparameter Tuning in Machine Learning

Some important model hyperparameters include:

1. The penalty in Linear Regression Classifier i.e. L1 or L2 regularization
2. The learning rate for training a neural network.
3. The C parameter in Logistic Regression.
4. The K number of neighbors in K-nearest neighbors.
5. The tree depth in a decision tree.

Conclusion

Remember that this cheat sheet provides a general guide, and the selection of the appropriate algorithm also depends on factors such as the **nature of the problem, available data, computational resources, and domain expertise**.

It's always recommended to experiment with multiple algorithms and fine-tune them for optimal performance.