K-Nearest Neighbor(KNN) Algorithm for Machine Learning

* K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
* K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
* K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
* K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
* K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.

## Why do we need a K-NN Algorithm?

Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram:

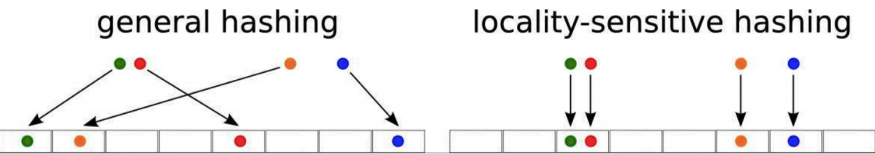


# Locality Sensitive Hashing

The task of finding nearest neighbours is very common. You can think of applications like finding duplicate or similar documents, audio/video search. Although using brute force to check for all possible combinations will give you the exact nearest neighbour but it’s not scalable at all. Approximate algorithms to accomplish this task has been an area of active research. Although these algorithms don’t guarantee to give you the exact answer, more often than not they’ll be provide a good approximation. These algorithms are faster and scalable.

Locality sensitive hashing (LSH) is one such algorithm. LSH has many applications, including:

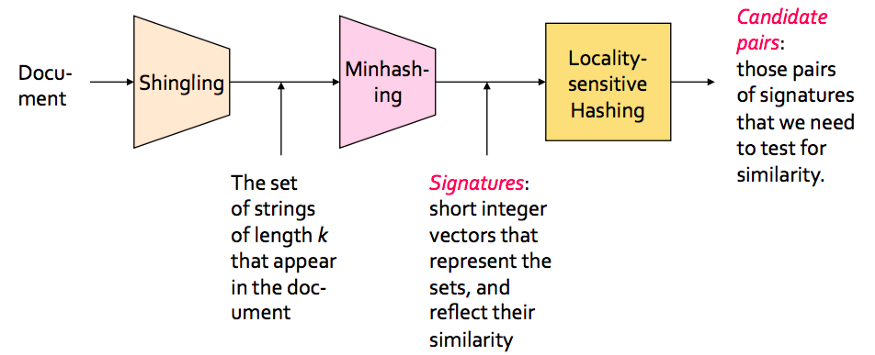
* Near-duplicate detection: LSH is commonly used to deduplicate large quantities of documents, webpages, and other files.
* Genome-wide association study: Biologists often use LSH to identify similar gene expressions in genome databases.
* Large-scale image search: Google used LSH along with PageRank to build their image search technology [VisualRank](https://research.google.com/pubs/pub34634.html" \t "_blank).
* Audio/video fingerprinting: In multimedia technologies, LSH is widely used as a fingerprinting technique A/V data.



**LSH** refers to a family of functions (known as LSH families) to hash data points into buckets so that data points near each other are located in the same buckets with high probability, while data points far from each other are likely to be in different buckets. This makes it easier to identify observations with various degrees of similarity.

In the context of this problem, we can break down the LSH algorithm into 3 broad steps:

1. **Shingling**
2. **Min hashing**
3. **Locality-sensitive hashing**

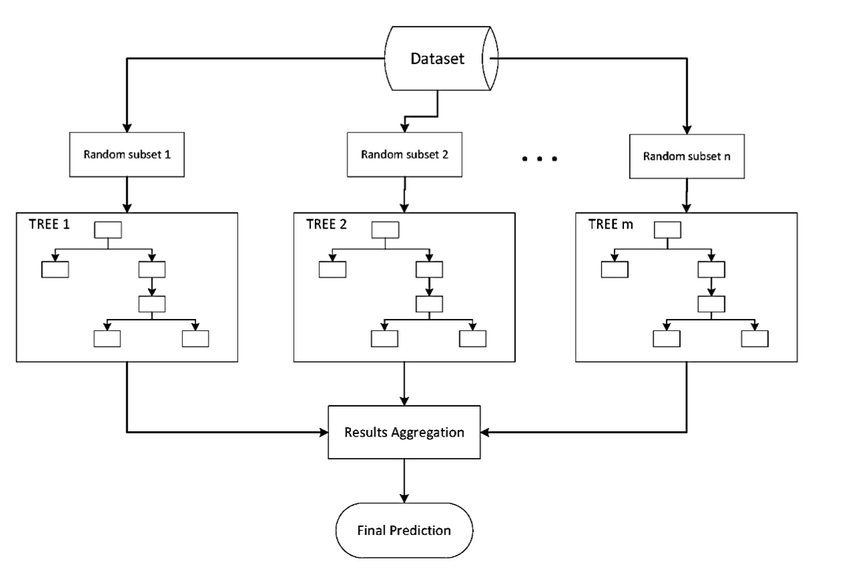


# Ensemble Methods

**Ensemble methods** is a machine learning technique that combines several base models in order to produce one optimal predictive model

**Types of Ensemble Methods**

1. ***BAGG***ing, or ***B***ootstrap ***AGG***regating. **BAGG**ing gets its name because it combines ***B***ootstrapping and ***Agg***regation to form one ensemble model. Given a sample of data, multiple bootstrapped subsamples are pulled. A Decision Tree is formed on each of the bootstrapped subsamples. After each subsample Decision Tree has been formed, an algorithm is used to aggregate over the Decision Trees to form the most efficient predictor. The image below will help explain:



2. **Random Forest**

Random forests: Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model

* As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output
* The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting

Boosting

* Boosting refers to a family of algorithms that can convert weak learners to strong learners. The main principle of boosting is to fit a sequence of weak learners

Stacking

* Stacking is an ensemble learning technique that combines multiple classification or regression models via a meta-classifier or a meta- regressor

**Meta Learning**

* Meta-learning, also known as “learning to learn”, intends to design models that can learn new skills or adapt to new environments rapidly with a few training examples
* In practice, very closely related to multi-task learning
* Common approach to meta learning:

Model-Based

* + Memory-Augmented Neural Networks
  + Meta Networks

Metric-Based

* + Convolutional Siamese Neural Network
  + Matching Networks
  + Relation Network
  + Prototypical Networks

Optimization-Based

* + LSTM Meta-Learner
  + Temporal Discreteness
  + Reptile

**Model-Based**

Model-based meta-learning models updates its parameters rapidly with a few training steps, which can be achieved by its internal architecture or controlled by another meta-learner mode

* Memory-Augmented Neural Networks: The model is known as MANN short for Memory-Augmented Neural Networks, which is expected to encode new information fast and thus to adapt to new tasks after only a few samples, it fits well for meta-learning
* Meta Networks: Meta Networks (MetaNet) learns a meta-level knowledge across tasks and shifts its inductive biases via fast parameterization for rapid generalization

**Metric-Based**

The core idea in metric-based meta-learning is similar to nearest neighbors algorithms, which weight is generated by a kernel function. It aims to learn a metric or distance function over objects. The notion of a good metric is problem-dependent. It should represent the relationship between inputs in the task space and facilitate problem solving

* Convolutional Siamese Neural Network: Siamese neural network is composed of two twin networks whose output is jointly trained. There is a function above to learn the relationship between input data sample pairs. The two networks are the same, sharing the same weight and network parameters
* Matching Networks: Matching Networks learn a network that maps a small labelled support set and an unlabelled example to its label, obviating the need for fine-tuning to adapt to new class types

**Optimization-Based:**

What optimization-based meta-learning algorithms intend for is to adjust the optimization algorithm so that the model can be good at learning with a few examples

* Temporal Discreteness: MAML, short for Model-Agnostic Meta- Learning, is a fairly general optimization algorithm, compatible with any model that learns through gradient descent
* Reptile: Reptile is a remarkably simple meta-learning optimization algorithm, given that both rely on meta-optimization through gradient descent and both are model-agnostic