Regression vs Classification in Machine Learning Explained!

What is Regression?

Regression algorithms predict continuous value from the provided input.
A supervised learning algorithm uses real values to predict quantitative data like income, height, weight, scores or probability.

What is Classification?

A procedure in which a model or a function separates the data into discrete values, i.e., multiple classes of datasets using independent features, is called classification.

Classification algorithms are used to predict/Classify the discrete values such as Male or Female, True or False, Spam or Not Spam, etc.

Types of Regression

1. Linear Regression

Most preferable and simple to use, it applies linear equations to the datasets. Using a straight line, the relationship between two quantitative variables i.e., one independent and another dependent, is modeled in simple linear regression.

2. Polynomial Regression

To find or model the non-linear relationship between an independent and a dependent variable is called polynomial regression. It is specifically used for curvy trend datasets.

3. Logistic Regression

Commonly known as the logit model, Logistic Regression understands the probable chances of the occurrence of an event.

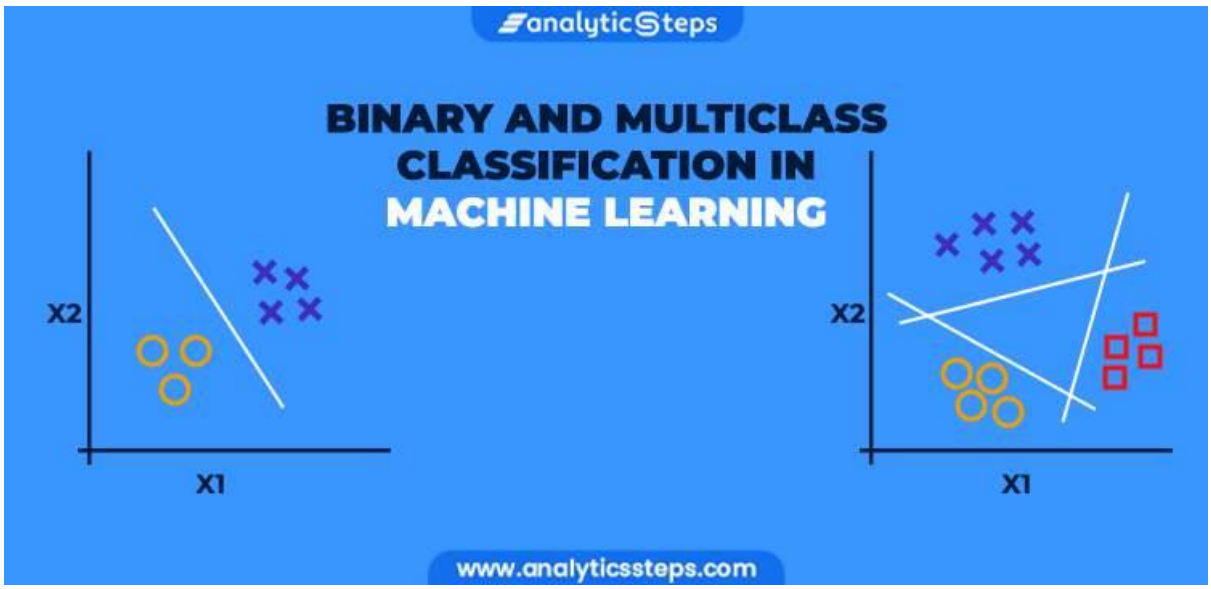
Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class.

it's referred to as regression because it takes the output of the linear regression function as input and uses a sigmoid function to estimate the probability for the given class.

Types of Classification

1. Binary Classification

Binary classification is the task of classifying the elements of a set into two groups (each called class) on the basis of a classification rule.



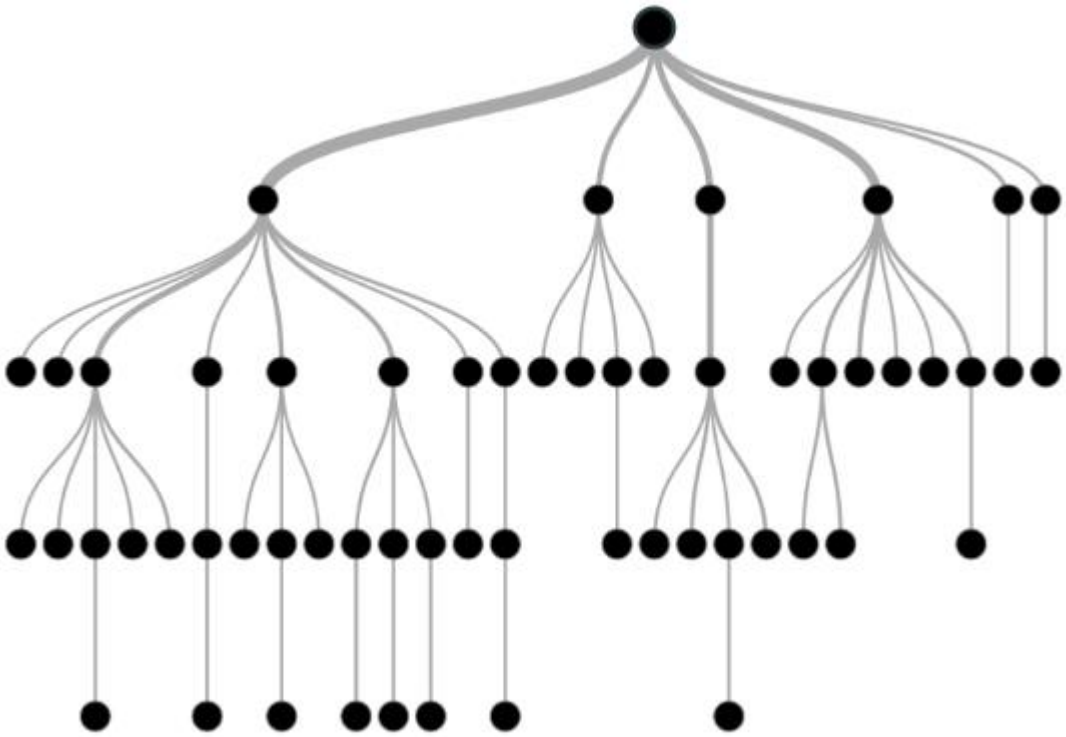
2. Multi-class Classification

In machine learning, multi-class classification provides more than two outcomes of the model.

What is a Decision Tree?

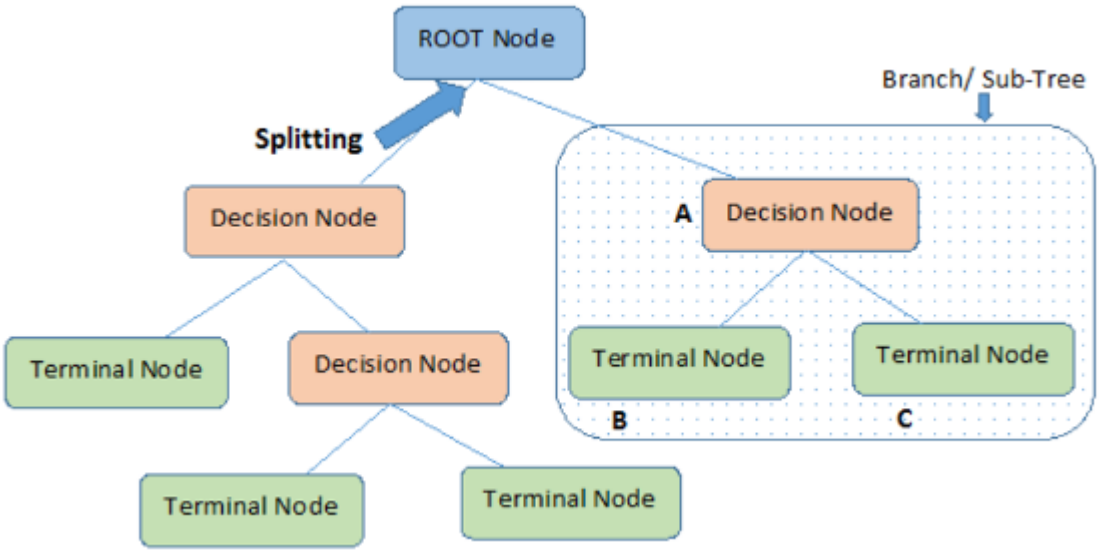
A decision tree is a predictive model that uses a flowchart-like structure to make decisions based on input data.

It divides data into branches and assigns outcomes to leaf nodes. Decision trees are used for classification and regression tasks, providing easy-to-understand models.



Decision Tree Terminologies

- ***Root Nodes*** – It is the node present at the beginning of a decision tree from this node the population starts dividing according to various features.
- ***Decision Nodes*** – the nodes we get after splitting the root nodes are called Decision Node
- ***Leaf Nodes*** – the nodes where further splitting is not possible are called leaf nodes or terminal nodes
- ***Sub-tree*** – just like a small portion of a graph is called sub-graph similarly a sub-section of this decision tree is called sub-tree.
- ***Pruning*** **– is nothing but cutting down some nodes to stop overfitting.**



Guide on Support Vector Machine (SVM) Algorithm

What is a Support Vector Machine?

It is a supervised machine learning problem where we try to find a hyperplane that best separates the two classes. **Note:** Don't get confused between SVM and logistic regression. Both the algorithms try to find the best hyperplane, but the main difference is logistic regression is a probabilistic approach whereas support vector machine is based on statistical approaches.

Logistic Regression vs Support Vector Machine

SVM works best when the dataset is small and complex. It is usually advisable to first use logistic regression and see how does it performs, if it fails to give a good accuracy you can go for SVM without any kernel.

Types of Support Vector Machine Algorithms

1. Linear SVM

When the data is perfectly linearly separable only then we can use Linear SVM. Perfectly linearly separable means that the data points can be classified into 2 classes by using a single straight line (if 2D).

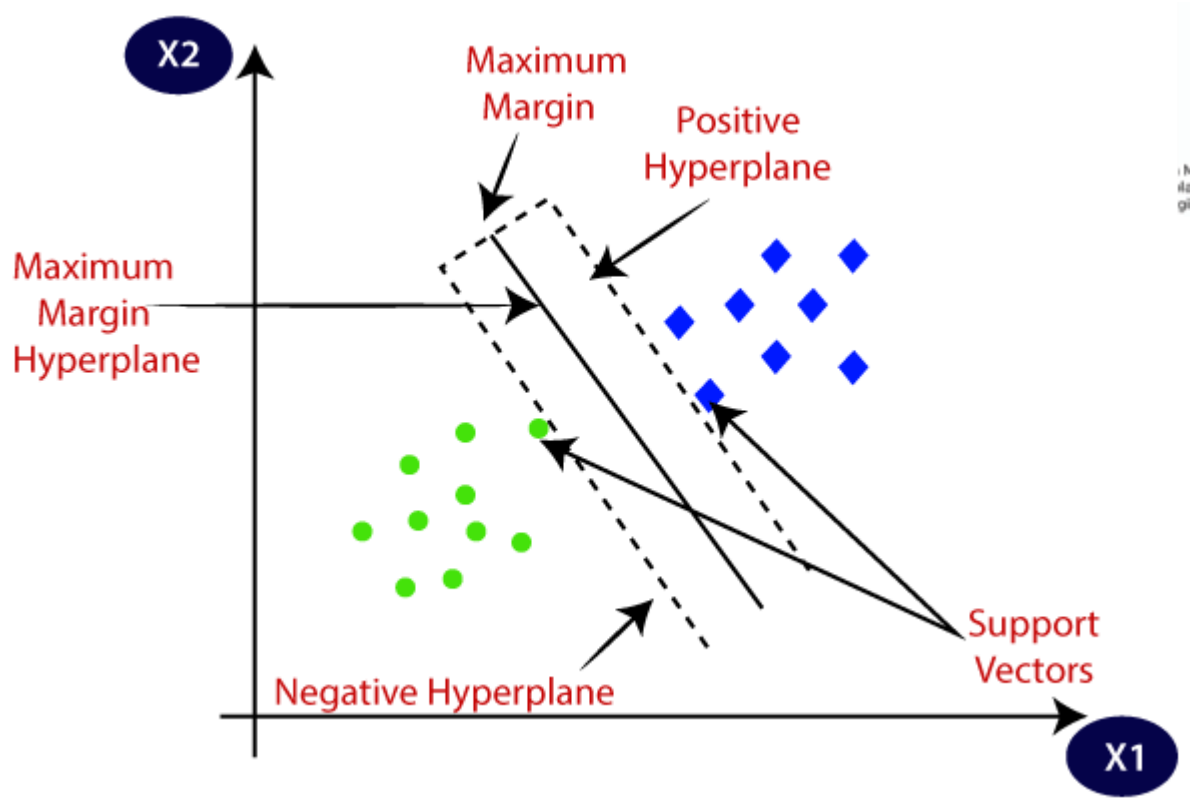
2. Non-Linear SVM

When the data is not linearly separable then we can use Non-Linear SVM, which means when the data points cannot be separated into 2 classes by using a straight line (if 2D) then we use some advanced techniques like kernel tricks to classify them. In most real-world applications we do not find linearly separable datapoints hence we use kernel trick to solve them.

Important Terms

Now let's define two main terms which will be repeated again and again in this article:

- **Support Vectors:** These are the points that are closest to the hyperplane. A separating line will be defined with the help of these data points.
- **Margin:** it is the distance between the hyperplane and the observations closest to the hyperplane (support vectors). In SVM large margin is considered a good margin. There are two types of margins **hard margin** and **soft margin.** I will talk more about these two in the later section.



What is Random Forest Algorithm?

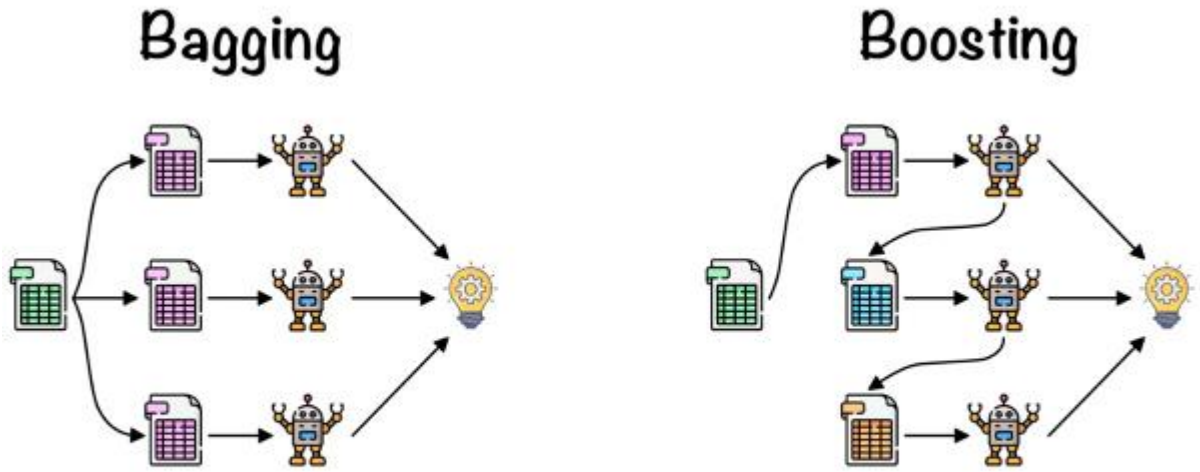
Random Forest is one of the most popular and commonly used algorithms by Data Scientists. Random forest is a ***Supervised*** ***Machine Learning Algorithm*** that is ***used widely in Classification and Regression problems***. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

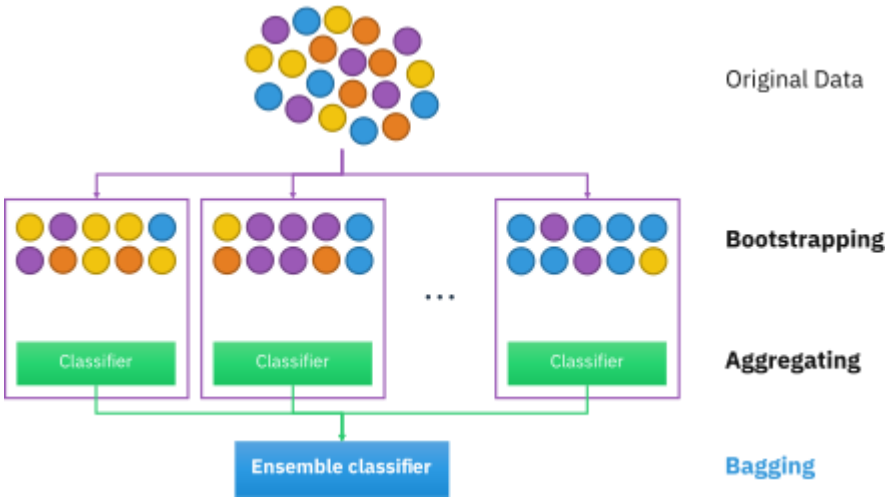
Ensemble simply means combining multiple models. Thus, a collection of models is used to make predictions rather than an individual model.

Ensemble uses two types of methods:

1. **Bagging**– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest.

2. **Boosting**– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.



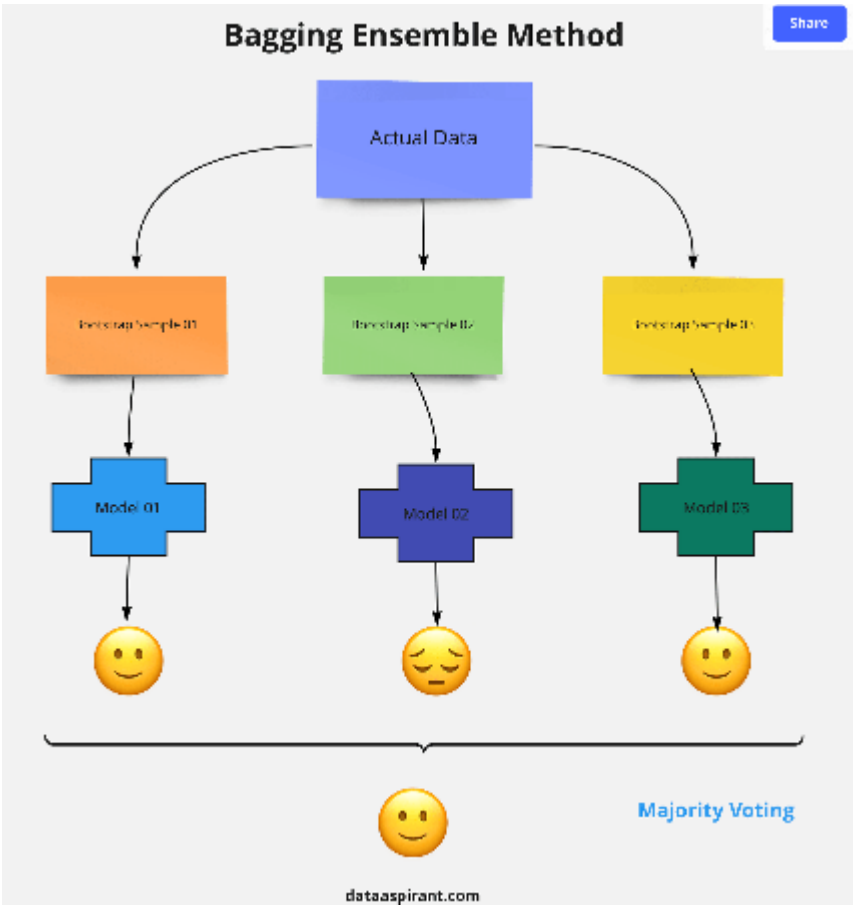


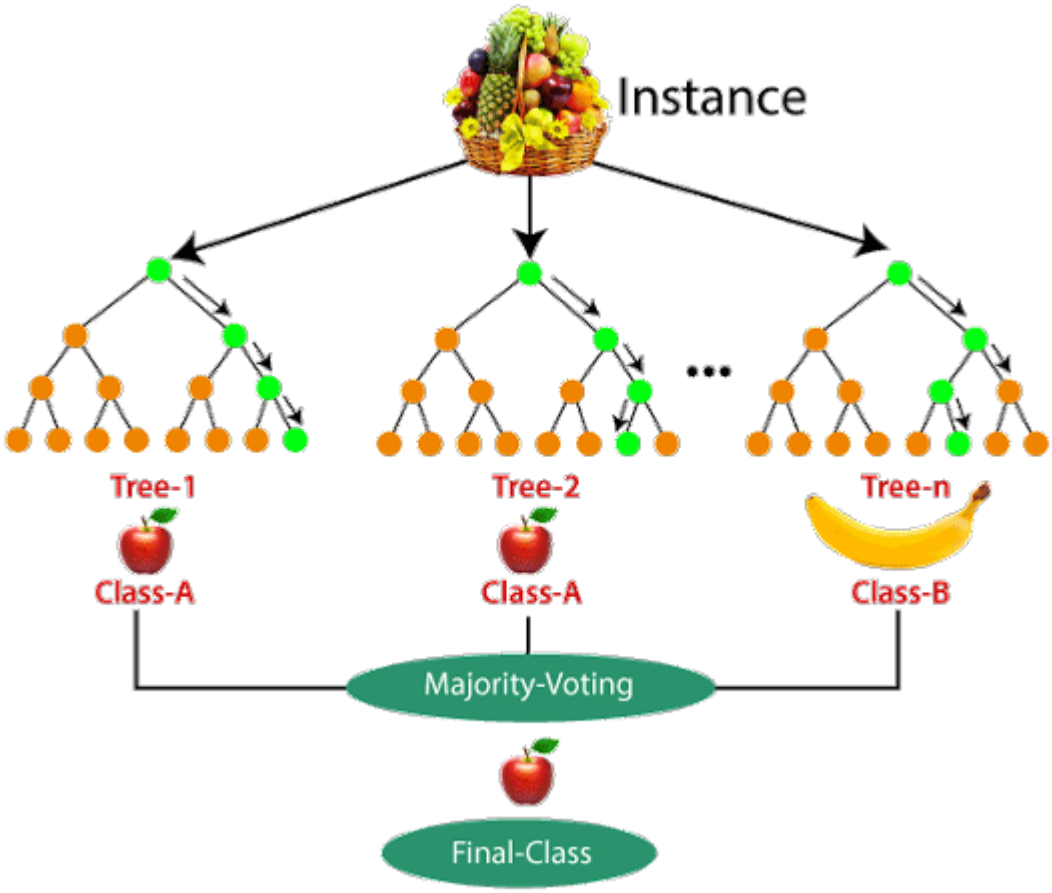
Original Data

Bootstrapping

Aggregating

Bagging


Bagging Ensemble Method
Share
Actual Data
Model 01
Model 02
Majority Voting
dataaspirant.com


Instance
Tree-1
Tree-2
Tree-n
Class-A
Class-A
Class-B
Majority-Voting
Final-Class

Difference Between Decision Tree and Random Forest

Random forest is a collection of decision trees; still, there are a lot of differences in their behavior.

Decision trees	Random Forest
1. Decision trees normally suffer from the problem of overfitting if it's allowed to grow without any control.	1. Random forests are created from subsets of data, and the final output is based on average or majority ranking; hence the problem of overfitting is taken care of.
2. A single decision tree is faster in computation.	2. It is comparatively slower.

Important Hyperparameters in Random Forest

Hyperparameters are used in random forests to either enhance the performance and predictive power of models or to make the model faster.

Hyperparameters to Increase the Predictive Power

n_estimators: Number of trees the algorithm builds before averaging the predictions.

max_features: Maximum number of features random forest considers splitting a node.

mini_sample_leaf: Determines the minimum number of leaves required to split an internal node.

criterion: How to split the node in each tree? (Entropy/Gini impurity/Log Loss)

max_leaf_nodes: Maximum leaf nodes in each tree

Hyperparameters to Increase the Speed
n_jobs: it tells the engine how many processors it is allowed to use. If the value is 1, it can use only one processor, but if the value is -1, there is no limit.

random_state: controls randomness of the sample. The model will always produce the same results if it has a definite value of random state and has been given the same hyperparameters and training data.

Naive Bayes Classifier Explained:

Where is Naive Bayes Used?

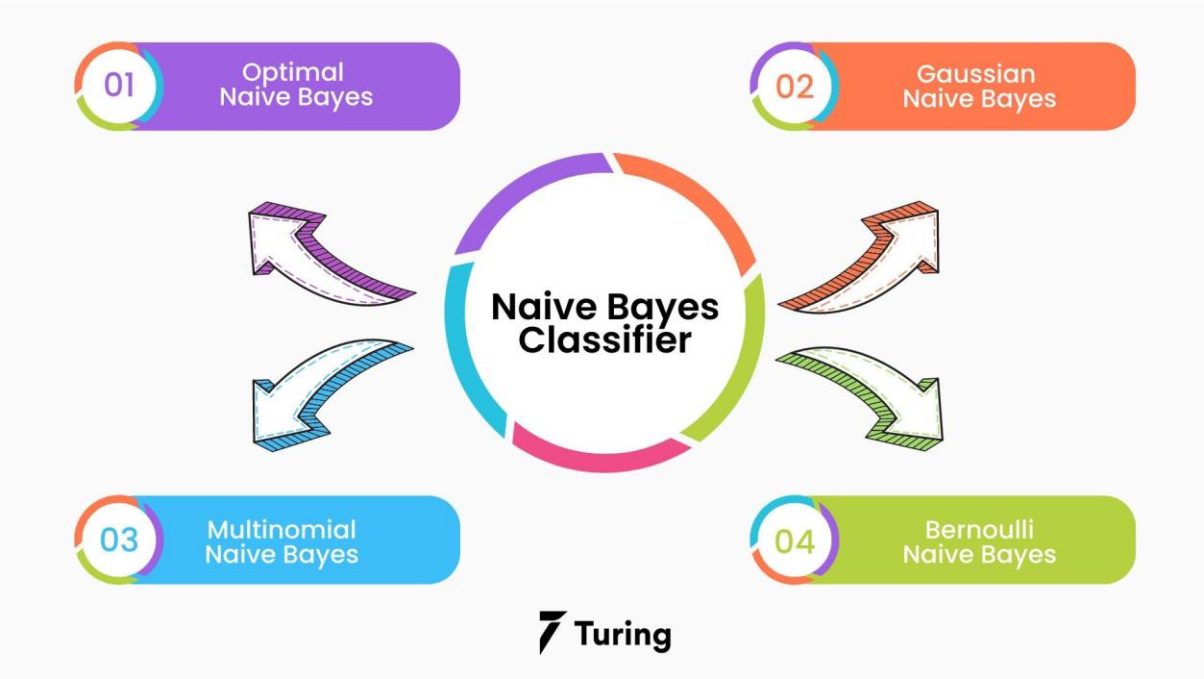
You can use Naive Bayes for the following things:

Face Recognition

As a classifier, it is used to identify the faces or its other features, like nose, mouth, eyes, etc.

Weather Prediction

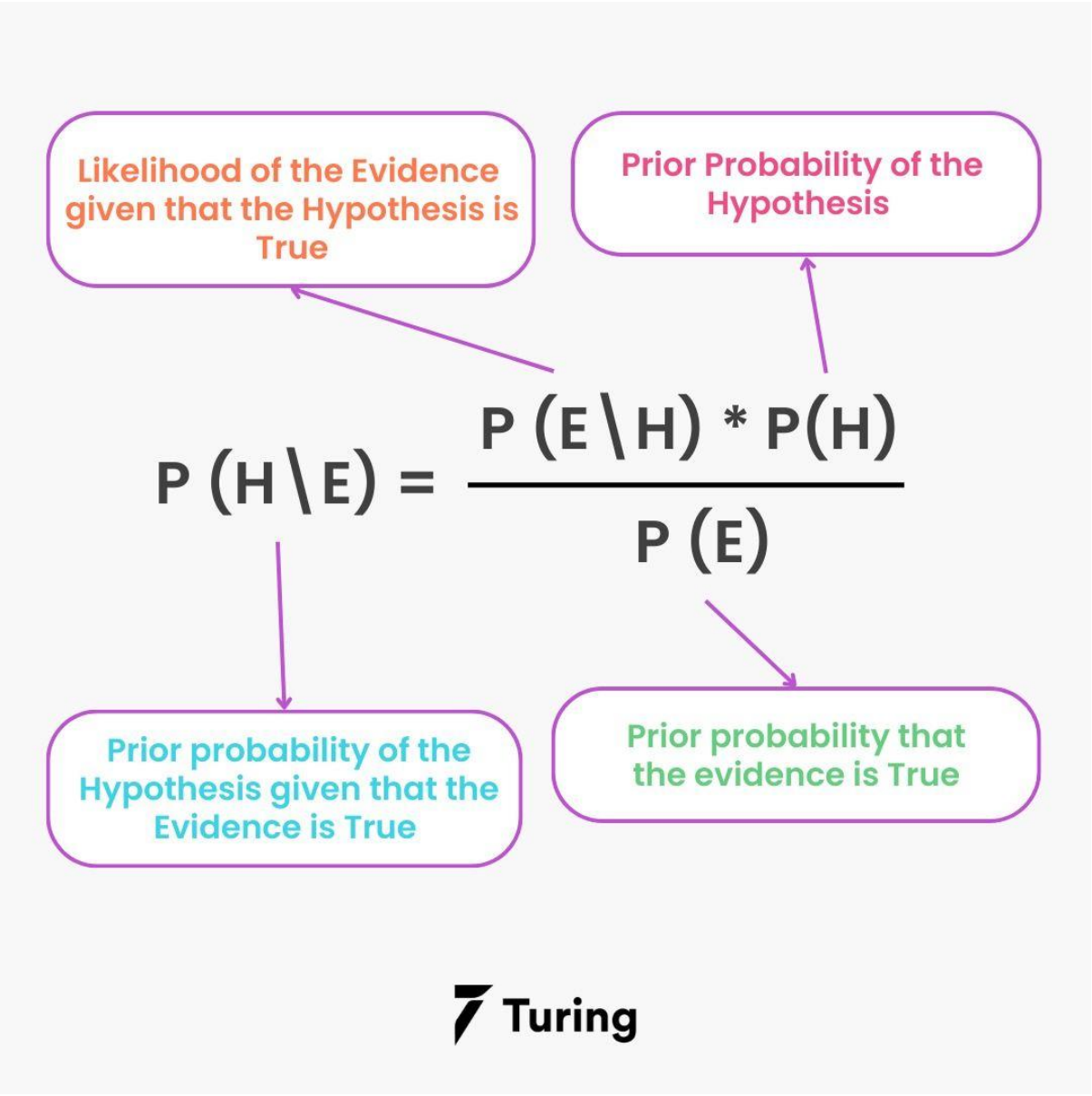
It can be used to predict if the weather will be good or bad.



Understanding Naive Bayes Classifier

Based on the Bayes theorem, the Naive Bayes Classifier gives the conditional probability of an event A given event B.

- **Gaussian Naive Bayes:** gaussiannb is used in classification tasks and it assumes that feature values follow a gaussian distribution.
- **Multinomial Naive Bayes:** It is used for discrete counts.



Bayes Theory

Bayes Theory works on coming to a hypothesis (H) from a given set of evidence (E). It relates to two things: the probability of the hypothesis before the evidence P(H) and the probability after the evidence P(H|E). The Bayes Theory is explained by the following equation:

```
P(H|E) = (P(E|H} * P(H))/P(E)
```

In the above equation,

- P(H|E) denotes how event H happens when event E takes place.
- P(E|H) represents how often event E happens when event H takes place first.
- P(H) represents the probability of event X happening on its own.
- P(E) represents the probability of event Y happening on its own.

Conditional Probability

Conditional probability is a subset of probability. It reduces the probability of becoming dependent on a single event. You can compute the conditional probability for two or more occurrences.

When you take events X and Y, the conditional probability of event Y is defined as the probability that the event occurs when event X is already over. It is written as P(Y|X). The mathematical formula for this is as follows:

```
P(Y|A) = P(X and Y) /P(X)
```