Classification in Python

As a marketing manager, you want a set of customers who are most likely to purchase your product. This is how you can save your marketing budget by finding your audience.

As a loan manager, you need to identify risky loan applications to achieve a lower loan default rate. This process of classifying customers into a group of potential and non-potential customers or safe or risky loan applications is known as a classification problem.

Classification is a two-step process, learning step and prediction step. In the learning step, the model is developed based on given training data. In the prediction step, the model is used to predict the response for given data.

**Types Of Classifiers**--------

Naive Bayes Classifier

Decision Tree Classifier

Nearest Neighbor Classifier

Support Vector Classifier

Naive Bayes Classifier

What is a classifier?

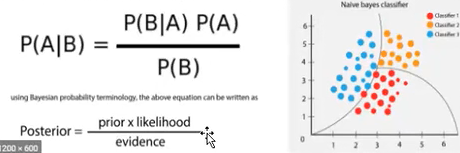
A classifier is a machine learning model that is used to discriminate different objects based on certain features.

Principle of Naive Bayes Classifier:

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem.

**Bayes Theorem**:

In machine learning, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.



The variable y is the class satiable(play golf), which represents if it is suitable to play golf or not given the conditions. Variable X represent the parameters/features.

X is given as,

**X = (X1, X2, X3,...,xn)**

Here x\_1,x\_2...x\_n.represent the features, ie they can be mapped to outlook, temperature, humidity and windy. By substituting for X and expanding using the chain rule we get,



Now, you can obtain the values for each by looking at the dataset and substitute them into the equation. For all

entries in the dataset, the denominator does not change, it remain static. Therefore, the denominator can be removed and a proportionality can be introduced.



In our case, the class variable(y) has only two outcomes, yes or no. There could be cases where the classification could be multivariate. Therefore, we need to find the class with maximum probability.



Using the above function, we can obtain the class, given the predictors.

X is given as,





**Example**

Let's build a classifier that predicts whether I should play tennis given the forecast. It takes four attributes to describe the forecast; namely, the outlook, the temperature, the humidity, and the presence or absence of wind. Furthermore the values of the four attributes are qualitative (also known as categorical). They take on the values shown below.

**Outlook E [Sunny, Overcast, Rainy]**

**Temperature Є [Hot, Mild, Cool]**

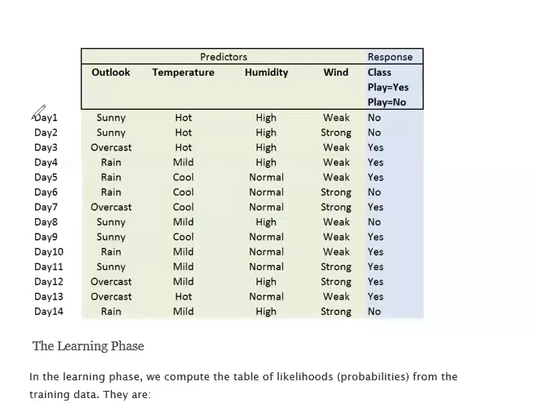
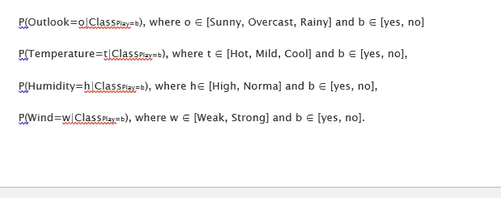
**Humidity Є [High, Normal]**

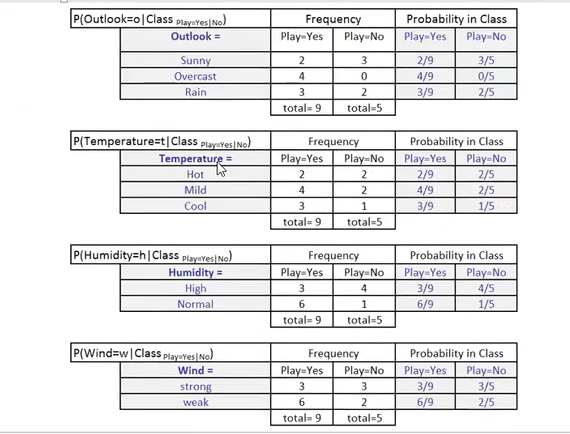
**Windy = [Weak, Strong]**

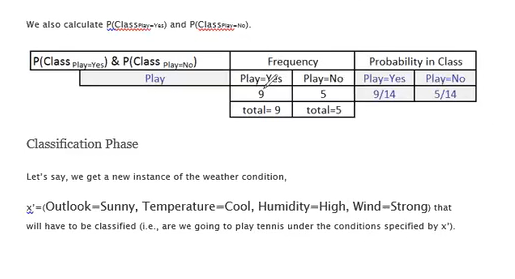
The class label is the variable, Play and takes the values yes or no.

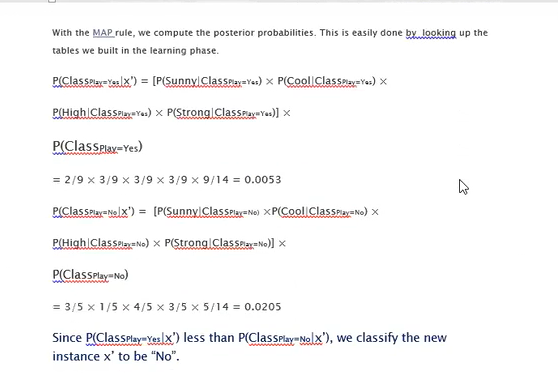
**Play ⤇ [Yes, No]**

We read-in training data below that has been collected over 14 days.









**Types of Naive Bayes Classifier:**

Multinomial Naive Bayes:

This is mostly used for document classification problem, ie whether a document belongs to the category of sports, politics, technology etc. The features/predictors used by the classifier are the frequency of the words present in the document.

Gaussian Naive Bayes:

When the predictors take up a continuous value and are not discrete, we assume that

these values are sampled from a gaussian distribution.