**Aim:**

To implement classification using Naïve Bayes Algorithm using scikit learn

**Theory:**

The dataset is divided into two parts, namely, **feature matrix** and the **response vector**.

* Feature matrix contains all the vectors(rows) of dataset in which each vector consists of the value of **dependent features**. In above dataset, features are ‘Outlook’, ‘Temperature’, ‘Humidity’ and ‘Windy’.
* Response vector contains the value of **class variable**(prediction or output) for each row of feature matrix. In above dataset, the class variable name is ‘Play golf’.

Assumption:

The fundamental Naive Bayes assumption is that each feature makes an:

* Independent
* Equal(None of the attributes is irrelevant and assumed to be contributing equally to the outcome.)

contribution to the outcome.

**Note:** The assumptions made by Naive Bayes are not generally correct in real-world situations. In-fact, the independence assumption is never correct but often works well in practice.

**Bayes’ Theorem**

Bayes’ Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes’ theorem is stated mathematically as the following equation:



Now, with regards to our dataset, we can apply Bayes’ theorem in following way:



where, y is class variable and X is a dependent feature vector (of size n) where:



**Naive Assumption**

Now, if any two events A and B are independent, then,

P(A,B) = P(A) x P(B)

Applying:



which can be expressed as:



Now, as the denominator remains constant for a given input, we can remove that term:



Now, we need to create a classifier model. For this, we find the probability of given set of inputs for all possible values of the class variable y and pick up the output with maximum probability. This can be expressed mathematically as:



**Experiment:**

**Naïve Bayes Problem Definition:**

Classification Problem : Classify the flowers into three types as Iris Setosa, Iris Virginica and Iris Versicolor given Sepal Length, Sepal Width, Petal Length and Petal Width.

Data Analysis Report :

Confusion Matrix:

|  |  |  |
| --- | --- | --- |
| 13 | 0 | 0 |
| 0 | 15 | 0 |
| 0 | 1 | 17 |

Accuracy = 97.83 %

**Conclusion:** The data analysis was performed on IRIS dataset and the following results were identified:

* Using all the features, Naïve Bayes Algorithm gave 97.82 % Accuracy
* No. of mislabeled classification is 1