Experiment No. 6

Aim: To apply the Naive Bayes machine learning algorithm for classification tasks and assess accuracy, precision, and recall.

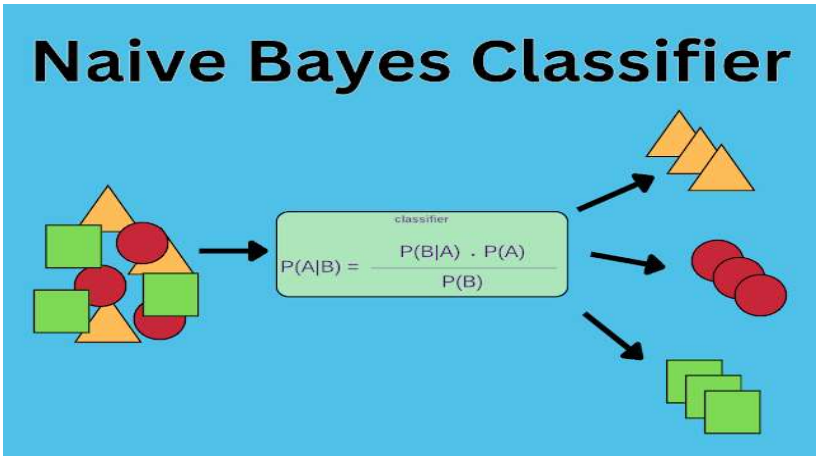
Platform used: Google Colab

Theory:

Naive Bayes Algorithm

The Naive Bayes algorithm is a probabilistic machine learning model primarily used for classification tasks. It operates based on Bayes' Theorem, a fundamental concept in probability theory, which allows for the calculation of the probability of an event based on prior knowledge of conditions that might be related to the event.

The "Naive" aspect of the algorithm stems from its core assumption: all features are conditionally independent of each other, given the class label. While this independence assumption is often unrealistic in real-world scenarios, Naive Bayes algorithms frequently yield surprisingly effective results, particularly in domains like text classification and spam detection.



How it Works:

Bayes' Theorem Application: Naive Bayes calculates the posterior probability of a class given a set of features, using the formula:

P(Class|Features) = [P(Features|Class) * P(Class)] / P(Features)

Where:

- P(Class|Features) is the posterior probability of the class given the features.
- P(Features|Class) is the likelihood of the features given the class.
- P(Class) is the prior probability of the class.
- P(Features) is the prior probability of the features.

Key Characteristics:

Simplicity and Efficiency:

Naive Bayes is computationally efficient, making it suitable for large datasets and real-time applications.

Generative Model:

It models the distribution of inputs for a given class, rather than directly learning the decision boundary between classes.

Handles Missing Data:

It can handle missing values by simply ignoring the corresponding features during probability calculations.

Robust to Irrelevant Features:

While the independence assumption is a simplification, the algorithm can still perform well even with irrelevant features.

Types of Naive Bayes Model

There are three types of Naive Bayes Model :

1. Gaussian Naive Bayes

In Gaussian Naive Bayes, continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution. A Gaussian distribution is also called Normal distribution. When plotted, it gives a bell shaped curve which is symmetric about the mean of the feature values.

2. Multinomial Naive Bayes

Multinomial Naive Bayes is used when features represent the frequency of terms (such as word counts) in a document. It is commonly applied in text classification, where term frequencies are important.

3. Bernoulli Naive Bayes

Bernoulli Naive Bayes deals with binary features, where each feature indicates whether a word appears or not in a document. It is suited for scenarios where the presence or absence of terms is more relevant than their frequency. Both models are widely used in document classification tasks.

Advantages of Naive Bayes Classifier:

- Easy to implement and computationally efficient.
- Effective in cases with a large number of features.
- Performs well even with limited training data.

- It performs well in the presence of categorical features.
- For numerical features data is assumed to come from normal distributions.

Disadvantages of Naive Bayes Classifier

- Assumes that features are independent, which may not always hold in real-world data.
- Can be influenced by irrelevant attributes.
- May assign zero probability to unseen events, leading to poor generalization.

Applications of Naive Bayes Classifier

- Spam Email Filtering: Classifies emails as spam or non-spam based on features.
- Text Classification: Used in sentiment analysis, document categorization, and topic classification.
- Medical Diagnosis: Helps in predicting the likelihood of a disease based on symptoms.
- Credit Scoring: Evaluates creditworthiness of individuals for loan approval.
- Weather Prediction: Classifies weather conditions based on various factors.

Conclusion: Thus, we successfully applied the Naive Bayes machine learning algorithm for classification tasks and assessed the accuracy, precision, and recall.