Report on Naive Bayes Classifier

1. Introduction

Naive Bayes Algorithm:

The Naive Bayes classifier is a probabilistic model based on Bayes' Theorem. It assumes that the features are conditionally independent given the class label. The algorithm computes the probability of each class based on the observed features and selects the class with the highest probability.

Objectives:

The objective was to implement a Naive Bayes classifier to predict class labels based on the provided feature values. The model was trained on a dataset, tested for accuracy, and evaluated using performance metrics.

2. Methodology

Data Preparation:

The dataset was loaded and preprocessed using pandas. The target variable (Yes or No) was identified. Class priors were calculated as the frequency of each class in the dataset. Likelihoods for each feature given the class label were computed, applying Laplace smoothing to handle zero probabilities.

Implementation:

- **Training**: The `train` method calculates class priors and feature likelihoods for each class.
- **Prediction**: The `classify_instance` method calculates posterior probabilities for each class and selects the class with the highest posterior.
- **Testing**: The `classify_test_set` method evaluates the classifier on a test set and calculates accuracy.

3. Results

The classifier achieved an overall accuracy of 93% on the test set.
Test Results:

Instance	Actual Class	Predicted Class	Correct
1	No	No	True
2	No	No	True
3	Yes	Yes	True
4	Yes	Yes	True
5	Yes	Yes	True
6	No	No	False
7	Yes	Yes	True
8	No	No	True
9	Yes	Yes	True
10	Yes	Yes	True
11	Yes	Yes	True
12	Yes	Yes	True
13	Yes	Yes	True
14	No	No	True

Accuracy: 93%

4. Discussion

Analysis:

The Naive Bayes classifier achieved a high accuracy of 93%. A misclassification occurred in Instance 6, where the actual class was 'No', but the predicted class was 'Yes'. This suggests that feature values for this instance may have been more aligned with the 'Yes' class in terms of probability. Possible Reasons for Misclassifications:

- **1. Independence Assumption:** The Naive Bayes algorithm assumes feature independence, which may not hold for all datasets. This assumption can lead to misclassifications if the features are highly correlated.
- **2. Rare Feature Values:** Instances with rare combinations of feature values may not be well represented in the training data, leading to potential misclassification.
- **3. Imbalanced Data:** The classifier may have a bias toward the majority class ('Yes') if the data is imbalanced.

Limitations:

- The independence assumption may limit the model's ability to capture interactions between features.
- The classifier may be sensitive to imbalanced data distributions.

5. Conclusion

The Naive Bayes classifier performed well with an overall accuracy of 93%. It correctly classified the majority of the instances but misclassified a few, likely due to the assumption of feature independence and potential data imbalances.

Lessons Learned:

This project demonstrated how to implement Naive Bayes from scratch, handle zero probabilities with Laplace smoothing, and evaluate the model's performance effectively.