HOMEWORK 3

TEAM: ARATRIK CHANDRA 23M0786
SAYANTAN BISWAS 23M0806

TITLE: NAIVE BAYES CLASSIFICATION REPORT

PROBLEM STATEMENT

The objective of this project is to perform Naive Bayes classification on a dataset with ten features. These features are generated from various probability distributions, including Gaussian, Bernoulli, Laplace, Exponential, and Multinomial Distributions. The goal is to train a Naive Bayes classifier and evaluate its performance on both training and validation data.

APPROACH

1. DATASET

The dataset consists of ten features, where each pair of features (X1, X2), (X3, X4), (X5, X6), (X7, X8), and (X9, X10) are generated independently from different probability distributions. Understanding the distribution of these features is crucial for building the Naive Bayes classifier.

A. (X1, X2) are drawn independently from two different univariate Gaussian distributions.
B. (X3, X4) are random variables drawn independently from two different Bernoulli Distributions
C. (X5, X6) are random variables drawn independently from two different Laplace Distributions
D. (X7, X8) are random variables drawn independently from two different Exponential Distribution
E. (X9, X10) are random variables drawn independently from two different Multinomial Distributions.

2. MODEL TRAINING

We implemented a Naive Bayes classifier in Python, leveraging the Naive Bayes algorithm based on Bayes' theorem. The classifier makes strong independence assumptions between features, which is why it is called "naive."
The `fit` method was used to train the model. It takes the feature matrix `X` and the target variable `y` as input.

The `fit` method calculates prior probabilities for each class and fits different types of distributions (Gaussian, Bernoulli, Laplace, Exponential, Multinomial) to different feature ranges.

3. MODEL PREDICTION

The `predict` method is used to make predictions on new data after the model has been trained. For each class, the method calculates the log-likelihood of the data given the parameters of the distributions fitted during training.
The class with the maximum log-posterior probability is predicted for each data point.

4. MODEL EVALUATION

We evaluated the model's performance using various metrics, including precision, recall, F1 score, and accuracy.

A. Precision measures the ability of the model to make accurate positive predictions.
B. Recall measures the model's ability to correctly identify all relevant instances.
C. F1 score balances precision and recall, providing a single metric for model evaluation.
D. Accuracy measures the overall proportion of correct predictions.

ESTIMATED PARAMETERS

RESULTS

1. Training Accuracy: 90.14%
2. Validation Accuracy: 90.23%

The model performs consistently well on both the training and validation datasets, indicating that it has learned to generalize effectively.

F1 SCORES BY CLASS

1. F1 Score for Class 1: 88.10%
2. F1 Score for Class 2: 87.85%
3. F1 Score for Class 3: 94.68%

The F1 scores for each class reflect the model's ability to classify instances accurately for each class, and they demonstrate the model's effectiveness in handling different types of data distributions.

CONCLUSION

In this project, we successfully applied a Naive Bayes classifier to a dataset with ten features generated from various probability distributions. The classifier demonstrated strong performance in terms of accuracy and F1 scores across different classes. This approach is particularly useful for handling data with various distribution types, making it versatile for a wide range of classification tasks.