Machine Learning (CS60050) - Assignment 2 Report

Amit Kumar - 20CS30003

Nikhil Saraswat - 20CS10039

Tasks

In this report, we analyze a Naive Bayes Classifier trained on theDataset_A.
The tasks performed with regard to this are:

1) Randomly divide Dataset A into 80% for training and 20% for testing. Encode categorical variables using appropriate encoding method (in-built function allowed).

2) A feature value is considered as an outlier if its value is greater than mean + 3 x standard deviation ($\mu + 3 \times \sigma$). A sample having maximum such outlier features must be dropped. Print the final set of features formed. Normalize the features as required.

3) Train the Naïve Bayes Classifier using 10-fold cross validation (no packages to be used for Naïve Bayes Classifier). Print the final accuracy.

4) Train the Naïve Bayes Classifier using Laplace correction on the same train and test split. Print the final accuracy.

5) Prepare a report including all your results

Dataset

The dataset has the following data fields (features) :

- id - unique ID
- Gender - Gender of customer
- Ever_Married - Marital status of customer
- Age - Age of customer
- Graduated - Is the customer a graduate
- Profession - profession of the customer
- Work_Experience - Work_Experience in years
- Spending_Score - Spending Score of customer
- Family_size - Number of family members of the customer customer (including the customer)
- Var_1 - Anonymised Category for the customer
- Segmentation - Customer Segment of the customer

The dataset has 8068 training examples and is available in the '.csv' format.

Results

1. Naive Bayes Classifier (without Laplace Correction)

Dataset	Accuracy
Train	46.975806451612904
Test	48.759305210918114

2. Naive Bayes Classifier with Laplace Correction

Dataset	Accuracy
Train	52.17121588089329
Test	51.736972704714645

3. 10-Fold Cross Validation

Accuracy (over 10-Folds)
50.37220843672456
50.86848635235732
53.349875930521094
50.744416873449126
49.007444168734494
51.488833746898266
50.49627791563276
51.86104218362283
50.86848635235732
50.0

Mean Accuracy	50.90570719602978

Max Accuracy	53.349875930521094

Instructions to Execute the Code

- Ensure you are using the latest version of Python3, and install all dependencies. pip install -r requirements.txt
- Execute the file main.py
  python main.py
- The relevant output will be displayed on the terminal or console.