Lab4_NaiveBayesClassifier_Report

1. What's the difference between the three Naive Bayes classifiers (BernoulliNB, GaussianNB, and MultinomialNB)?
   Bernoulli Naïve Bayes
   - It is used for discrete data.
   - Assumes each feature is **binary features** yes or no (1 or 0). (e.g., word is present or not).
   - Use case: Text classification using binary indicators (whether a specific word appears or not), Spam detection (spam or not), Classify movie review (positive or negative)

   Gaussian Naïve Bayes
   - Assumes each feature is **continuous (real-valued) data** and follow a normal distribution.
   - During training, the model learns the **mean and variance** of each feature.
   - Use case: Fraud detection, medical diagnosis, sensor readings, transaction amounts.

   Multinomial Naïve Bayes
   - Designed for **discrete count features** (e.g., how many times a word appears)..
   - Assumes each feature counts or frequencies.
   - All feature values must be **non-negative**.
   - Use case: Bag-of-Words or Term Frequency in NLP

2. Find and compare the precision and recall for all three Naïve Bayes classifiers.

Model	Precision(Fraud)	Recall(Fraud)	F1-score	ROC-AUC
GaussianNB	0.38	0.86	0.53	0.97
BernoulliNB	0.93	0.76	0.84	0.96

   Multinomial Naïve Bayes classifiers due to this fraud dataset values are negative and continuous date that not supported.

3. Which model would you choose for fraud detection, and why?
   In our results, GaussianNB (GNB) has very low precision (**0.38),** which means it generates many false alarms, but it achieves high recall (**0.86)** meaning it catches most fraud cases.
   BernoulliNB (BNB) misses some fraud cases compared to GNB but it achieves much higher precision resulting in fewer false alarms, while still maintaining solid recall **(0.76)**.
   For fraud decision, the important metric is usually recall (catch as many as frauds as possible).
   However, in this dataset, I would choose **Bernoulli Naïve Bayes** because it more balanced, with **excellent precision (0.93)** avoiding excessive false alarms and still **fairly high recall (0.76)**. This balance is reflected in the overall **F1-score (0.84), the best among the models.**

   Additionally, the data is highly imbalance 98% non-fraud, 2% fraud. In such cases, precision and recall are critical to evaluate.

   Finally, MultinomialNB is designed for **count-based features** (e.g text word counts). Since fraud datasets typically consist of continuous or binary features and negative, **MultinomialNB is not a natural/optimal choice here**.