Câu 1: Phân phối Bernoulli và Multinomial

Cho tập dữ liệu Education.csv
[https://drive.google.com/file/d/1Gn6YWHXRuPbTUXY5HFxM5C_tJHuZxCka/view?usp=sharing]

- Trong đó:
  - Text: Chứa đoạn văn bản liên quan đến chủ đề giáo dục.
  - Label: Chứa nhãn cảm xúc của văn bản [Tích cực (Positive)/Tiêu cực (Negative)].
- Yêu cầu: Áp dụng thuật toán Naive Bayes (phân phối bernoulli và phân phối Multinomial) để dự đoán cảm xúc của văn bản là tích cực hay tiêu cực và so sánh kết quả của hai phân phối đó.

In [1]:

```
import pandas as pd
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_splitt
from sklearn.naive_bayes import BernoulliNB, MultinomialNB
from sklearn.preprocessing import LabelBinarizer
from sklearn.metrics import accuracy_score, classification_report

# Tải tập dữ liệu (thay thế bằng đường dẫn tệp cục bộ)
dataset = "D:\\học máy và ứng dụng\\Nguyenphamthanhhuan-
197ct09716_lab2\\Education.csv"
data = pd.read_csv(dataset)

print("Dữ liệu đầu tiên để kiểm tra:")
print(data.head())

# LabelBinarizer chuyển đổi 'Dương'/'Âm' thành giá trị nhị phân (1 cho
'Dương', 0 cho 'Âm')
label_binarizer = LabelBinarizer()
y = label_binarizer.fit_transform(data['Label']).ravel()  # .ravel() chuyển
đổi vector cột thành mảng 1 chiều

# Sử dụng TfidfVectorizer để trích xuất tính năng
tfidf_vectorizer = TfidfVectorizer()
X = tfidf_vectorizer.fit_transform(data['Text'])  # Chuyển đổi văn bản thành
các tính năng số

# Chia dữ liệu thành các tập training và test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Train và dự đoán bằng Bernoulli Naive Bayes
bernoulli_nb = BernoulliNB()
bernoulli_nb.fit(X_train, y_train)
y_pred_bernoulli = bernoulli_nb.predict(X_test)

# Đánh giá mô hình
print("\nBernoulli Naive Bayes Results:")
print("Accuracy:", accuracy_score(y_test, y_pred_bernoulli))
print("Classification Report:")
print(classification_report(y_test, y_pred_bernoulli))

# Train và dự đoán bằng cách sử dụng Multinomial Naive Bayes
multinomial_nb = MultinomialNB()
multinomial_nb.fit(X_train, y_train)
```

```
y_pred_multinomial = multinomial_nb.predict(X_test)

# Đánh giá Multinomial Naive Bayes model
print("\nMultinomial Naive Bayes Results:")
print("Accuracy:", accuracy_score(y_test, y_pred_multinomial))
print("Classification Report:")
print(classification_report(y_test, y_pred_multinomial))

# kết quả
print("\nSo sánh kết quả giữa Bernoulli và Multinomial Naive Bayes:")
print(f"Bernoulli Naive Bayes Accuracy: {accuracy_score(y_test,
y_pred_bernoulli)}")
print(f"Multinomial Naive Bayes Accuracy: {accuracy_score(y_test,
y_pred_multinomial)}")
```

```
Dữ liệu đầu tiên để kiểm tra:
                                                Text     Label
0  The impact of educational reforms remains unce...  positive
1  Critics argue that recent improvements in the ...  negative
2  Innovative teaching methods have led to unexpe...  positive
3  Despite budget constraints, the school has man...  positive
4  The true effectiveness of online learning plat...  negative

Bernoulli Naive Bayes Results:
Accuracy: 0.6363636363636364
Classification Report:
              precision    recall  f1-score   support

           0       0.50      1.00      0.67         4
           1       1.00      0.43      0.60         7

    accuracy                           0.64        11
   macro avg       0.75      0.71      0.63        11
weighted avg       0.82      0.64      0.62        11


Multinomial Naive Bayes Results:
Accuracy: 0.6363636363636364
Classification Report:
              precision    recall  f1-score   support

           0       0.50      1.00      0.67         4
           1       1.00      0.43      0.60         7

    accuracy                           0.64        11
   macro avg       0.75      0.71      0.63        11
weighted avg       0.82      0.64      0.62        11


So sánh kết quả giữa Bernoulli và Multinomial Naive Bayes:
Bernoulli Naive Bayes Accuracy: 0.6363636363636364
Multinomial Naive Bayes Accuracy: 0.6363636363636364
```

Câu 2: Phân phối Gaussian

Cho tập dữ liệu Drug.csv
[https://drive.google.com/file/d/1_G8oXkLlsauQkujZzJZJwibAWu5PgBXK/view?usp=sharing]

- Trong đó:
  - Age: Tuổi của bệnh nhân
  - Sex: Giới tính của bệnh nhân
  - BP: Mức huyết áp
  - Cholesterol: Mức cholesterol trong máu
  - Na_to_K: Tỷ lệ Natri và Kali trong máu
  - Drug: Loại thuốc [A/B/C/X/Y]
- Yêu cầu: Áp dụng thuật toán Naive Bayes (phân phối Gaussian) để dự đoán kết quả loại thuốc phù hợp với bệnh nhân.

In [3]:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import accuracy_score, classification_report

# Tải tập dữ liệu (thay thế bằng đường dẫn tệp cục bộ)
dataset = "D:\\học máy và ứng dụng\\Nguyenphamthanhhuan-
197ct09716_lab2\\drug200.csv"
data = pd.read_csv(dataset)

print("Dữ liệu đầu tiên để kiểm tra:")
print(data.head())

# Tiền xử lý: Chuyển đổi các tính năng phân loại thành các giá trị số
label_encoder = LabelEncoder()

# Mã hóa 'Sex', 'BP', 'Cholesterol' và 'Drug'
data['Sex'] = label_encoder.fit_transform(data['Sex'])
data['BP'] = label_encoder.fit_transform(data['BP'])
data['Cholesterol'] = label_encoder.fit_transform(data['Cholesterol'])
data['Drug'] = label_encoder.fit_transform(data['Drug'])

print("\nDữ liệu sau khi mã hóa nhãn:")
print(data.head())

# Chia tách các tính năng (X) và mục tiêu (y)
X = data[['Age', 'Sex', 'BP', 'Cholesterol', 'Na_to_K']]  # Đặc trưng
y = data['Drug']  # Biến mục tiêu: Thuốc

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Áp dụng Gaussian Naive Bayes
gnb = GaussianNB()
gnb.fit(X_train, y_train)  # Train cái model
y_pred = gnb.predict(X_test)  # Dự đoán trên tập kiểm tra

# Đánh giá mô hình Gaussian Naive Bayes
print("\nGaussian Naive Bayes Results:")
print("Accuracy:", accuracy_score(y_test, y_pred))
print("Classification Report:")
print(classification_report(y_test, y_pred))
```

```
Dữ liệu đầu tiên để kiểm tra:
```

	Age	Sex	BP	Cholesterol	Na_to_K	Drug
0	23	F	HIGH	HIGH	25.355	DrugY
1	47	M	LOW	HIGH	13.093	drugC
2	47	M	LOW	HIGH	10.114	drugC
3	28	F	NORMAL	HIGH	7.798	drugX
4	61	F	LOW	HIGH	18.043	DrugY

Dữ liệu sau khi mã hóa nhãn:

	Age	Sex	BP	Cholesterol	Na_to_K	Drug
0	23	0	0	0	25.355	0
1	47	1	1	0	13.093	3
2	47	1	1	0	10.114	3
3	28	0	2	0	7.798	4
4	61	0	1	0	18.043	0

Gaussian Naive Bayes Results:
Accuracy: 0.925
Classification Report:

	precision	recall	f1-score	support
0	1.00	0.80	0.89	15
1	0.86	1.00	0.92	6
2	0.75	1.00	0.86	3
3	0.83	1.00	0.91	5
4	1.00	1.00	1.00	11
accuracy			0.93	40
macro avg	0.89	0.96	0.92	40
weighted avg	0.94	0.93	0.92	40