Naive Bayes Practices

I. Phân loại Naive Bayes sử dụng thư viện sklearn trên dataset nhỏ

Bài toán dự đoán rủi ro tín dụng

- Cho danh sách những người vay tiền với các đặc trưng được quan sát bao gồm: Người vay có sở hữu nhà ở hay không (Home Owner); Tình trạng hôn nhân (Marital Status); Thu nhập hàng năm (Annual Income); và Nhận định người vay có bị vỡ nợ hay không (Defaulted Borrower). Trong đó Defaulted Borrower là nhãn lớp cho biết người vay có trả được khoản tiền đã vay hay không? Dữ liệu được lưu trong file credit_data.txt.
- Yêu cầu: Sử dụng thuật toán Naïve Bayes để dự đoán xác suất một người có vỡ nợ hay không dựa vào các đặc trưng Home Owner, Marital Status và Annual Income.

1. Import các thư viện cần thiết

In [2]:
```
import pandas as pd
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.naive_bayes import GaussianNB, MultinomialNB, BernoulliNB, CategoricalNB, ComplementNB
```

2. Cấu hình dữ liệu

In [3]:
```
data = pd.DataFrame([
    ["Yes", "Single", "High", "No"],
    ["No", "Married", "High", "No"],
    ["No", "Single", "Low", "No"],
    ["Yes", "Married", "High", "No"],
    ["No", "Divorced", "Low", "Yes"],
    ["No", "Married", "Low", "No"],
    ["Yes", "Divorced", "High", "No"],
    ["No", "Single", "Low", "Yes"],
    ["No", "Married", "Low", "No"],
```

```
    ["No", "Single", "Low", "Yes"]
], columns=["Home Owner", "Marital Status", "Annual Income", "Defaulted Borrower"])
```

In [4]:
```
# Yêu cầu 1: Chuẩn bị dữ liệu cho mô hình học máy (X: lưu giá trị của các cột đặc trưng; y: lưu giá trị cột nhãn)
######
# Code


######
```

In [5]:
```
# Yêu cầu 2: Chuẩn hóa dữ liệu từ dạng phân loại sang dạng số học.
# Gợi ý: Có thể sử dụng các kỹ thuật Label Encoding, One-Hot Encoding,...
######
# Code


######
```

3. Training và Testing mô hình Naive Bayes với các phân phối xác suất khác nhau

In [6]:
```
# Yêu cầu 3: Chuẩn bị một mẫu dữ liệu mới và chuẩn hóa về giá trị số để kiểm tra kết quả phân lớp của mô hình
######
# Code


######
```

In [7]:
```
# Yêu cầu 4: Huấn luyện và kiểm tra mô hình Gaussian Naive Bayes (sử dụng phân phối chuẩn Gauss)
######
# Code


######
```

In [8]:
```
# Yêu cầu 5: Huấn luyện và kiểm tra mô hình Multinomial Naive Bayes (sử dụng phân phối đa thức)
######
# Code
```

```
######
```

In [9]:
```
# Yêu cầu 6: Huấn luyện và kiểm tra mô hình mô hình Bernoulli Naive Bayes (sử dụng phân phối Bernoulli)
######
# Code


######
```

In [10]:
```
# Yêu cầu 7: Huấn luyện và kiểm tra mô hình Categorical Naive Bayes (Categorical distribution - mở rộng của phân phối
######
# Code


######
```

In [11]:
```
# Yêu cầu 8: Huấn luyện và kiểm tra mô hình Complement Naive Bayes (điều chỉnh của phân phối đa thức)
######
# Code


######
```

II. Phân loại Naive Bayes sử dụng thư viện sklearn trên dataset lớn (Dữ liệu: credit_data.csv)

- Bài toán: Dự đoán điểm tín dụng của khách hàng khi vay vốn sử dụng Naïve Bayes.
- Mục tiêu:
  - Xây dựng được mô hình Naïve Bayes sử dụng thư viện sklearn.
  - Ứng dụng và hiểu cách áp dụng mô hình Naïve Bayes vào giải quyết bài toán thực tế (ví dụ: dự đoán điểm tín dụng).
  - Sử dụng độ đo Accuracy để đánh giá chất lượng mô hình.
  - Thay đổi các phân bố xác suất (phân phối chuẩn, phân phối đa thức, phân phối Bernoulli) để chọn ra bộ phận lớp Naive Bayes phù hợp với dữ liệu.
- Dữ liệu:

    - Tập các đặc trưng của khách hàng và điểm tín dụng tương ứng trong một khoảng thời gian nhất định.
    - Tập các nhãn (Cột "Risk"): Gồm 2 loại nhãn "Good" và "Bad". Trong đó "Good" biểu thị khách hàng có khả năng trả nợ đúng hạn, "Bad" biểu thị khách hàng có khả năng vỡ nợ.
- Loại bài toán: Phân loại
    - Input: n vector đã mã hóa của các khách hàng.
    - Output: nhãn y là một trong 2 nhãn trên.

1. Import các thư viện cần thiết

In [12]:
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, MinMaxScaler
from sklearn.naive_bayes import GaussianNB, MultinomialNB, BernoulliNB
from sklearn.metrics import accuracy_score
```

2. Load dữ liệu

In [13]:
```
# Yêu cầu 1: Đọc dữ liệu từ file csv, hiển thị 5 dòng đầu tiên trong dataset
######
# Code


######
```

3. Tiền xử lý dữ liệu

3.1. Hiểu dữ liệu

In [14]:
```
# Yêu cầu 2: Hiển thị thông tin tổng quan của dataset (số dòng, số cột, tên các cột, kiểu dữ liệu)
# Viết nhận xét về kết quả thu được
######
# Code


######
```

In [15]:
```
# Yêu cầu 3: Hiển thị thông tin các thống kê mô tả cơ bản của các đặc trưng
# Gợi ý: Sử dụng hàm describe()
######
# Code


######
```

3.2. Kiểm tra missing values

In [16]:
```
# Yêu cầu 4: Thống kê tổng số giá trị bị thiếu trong dataset, liệt kê giá trị bị thiếu trong mỗi cột.
# Viết nhận xét cho kết quả thu được
######
# Code


######
```

3.3. Xử lý missing values

In [17]:
```
# Yêu cầu 5: Lựa chọn và áp dụng phương pháp xử lý các giá trị bị thiếu
######
# Code


######
```

3.4. Mã hóa các đặc trưng rời rạc

In [18]:
```
# Yêu cầu 6: Chuyển đổi các giá trị rời rạc về giá trị số
# Gợi ý: Có thể áp dụng một trong các kỹ thuật: Label Encoding, One-Hot Encoding, Ordinal Encoding, Target Encoding
######
# Code


######
```

4. Chia dữ liệu thành 2 phần: Training và Testing

In [19]:
```
# Yêu cầu 7: Chia dữ liệu thành 80% dùng để test và 20% dùng để train.
######
# Code


######
```

5. Training and Testing Naive Bayes model

In [20]:
```
# Yêu cầu 8: Chuẩn hóa toàn bộ dữ liệu về một phạm vi nhất định (Data Scaling)
# Gợi ý: Có thể sử dụng StandardScaler, MinMaxScaler, Robust Scaling
######
# Code


######
```

In [21]:
```
# Training & Testing
# Yêu cầu 9: Sử dụng thư viện sklearn để xây dựng một mô hình Naive Bayes và kiểm tra kết quả phân lớp dựa trên độ đo
######
# Code


######
```

6. Thay đổi phân bố xác suất để chọn được bộ phân lớp phù hợp với dữ liệu

In [22]:
```
# Gợi ý: Khảo sát các phân bố xác suất như phân phối chuẩn (Gauss), phân phối đa thức, phân phối Bernoulli và chọn ra
# Mở rộng (Tùy chọn): Điều chỉnh các tham số để cho kết quả tốt hơn (VD: giá trị làm mịn alpha trong phân phối đa thứ
######
# Code


######
```