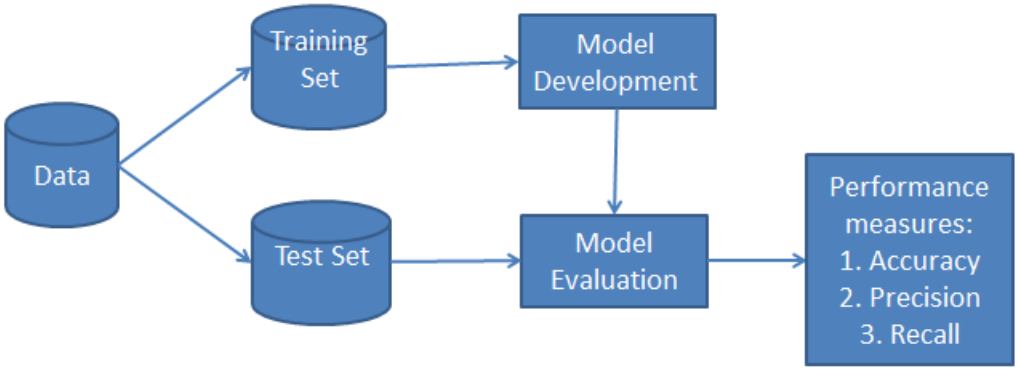
Bài thực hành

Phân loại dữ liệu với giải thuật Bayes thơ ngây (Naive Bayes)

1. Ví dụ minh họa sử dụng giải thuật Bayes thơ ngây (Naive Bayes)

Trong ví dụ này, học viên làm quen với:

- Nắm được các bước phân loại dữ liệu



- Sử dụng giải thuật Bayes thơ ngây (Naive Bayes) để phân loại dữ liệu.
- Đánh giá hiệu quả bằng ma trận confusion.

```
# Nạp các gói thư viện cần thiết
import pandas as pd
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
import matplotlib.pyplot as plt
import seaborn as sns

# Đọc dữ liệu iris từ UCI (https://archive.ics.uci.edu/ml/datasets/Iris)
# hoặc từ thư viện scikit-learn
# Tham khảo https://scikit-
learn.org/stable/auto_examples/datasets/plot_iris_dataset.html

from sklearn import datasets
from sklearn.model_selection import train_test_split
iris = datasets.load_iris()
#print(iris)
columns=["Petal Length","Petal Width","Sepal Length","Sepal Width"];
X = pd.DataFrame(iris.data, columns=columns)
y = iris.target
print(X.describe())

# Sử dụng nghi thức kiểm tra hold-out
# Chia dữ liệu ngẫu nhiên thành 2 tập dữ liệu con:
# training set và test set theo tỷ lệ 70/30
from sklearn.model_selection import train_test_split
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
#print(X_train.shape, y_train.shape)
#print(X_test.shape, y_test.shape)

# Xây dựng mô hình phân loại dữ liệu
model = GaussianNB()
model.fit(X_train, y_train)

# Dự đoán nhãn tập kiểm tra
y_pred = model.predict(X_test)
#print(y_pred)

# Tính độ chính xác
print("Do chinh xác cua mo hinh voi nghi thuc kiem tra hold-out: %.3f" %
accuracy_score(y_test, y_pred))

# Xây dựng confusion-matrix. Tham khảo:
# https://www.python-course.eu/confusion_matrix.php
cm = confusion_matrix(y_test, y_pred)

# Chuyển confusion-matrix về data frame phục vụ cho việc vẽ đồ thị
cm_df = pd.DataFrame(cm,
                     index=['setosa', 'versicolor', 'virginica'],
                     columns=['setosa', 'versicolor', 'virginica'])

plt.figure(figsize=(5.5, 4))
sns.heatmap(cm_df, annot=True)
plt.title('Decision Tree \nAccuracy:{0:.3f}'.format(accuracy_score(y_test,
y_pred)))
plt.ylabel('True label')
plt.xlabel('Predicted label')
plt.show()
```

2. Học viên làm theo các ví dụ trong trang web sau.

https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn

3. Hãy viết chương trình phân loại các tập dữ liệu sau với giải thuật Bayes thơ ngây (Naive Bayes): Breast Cancer Wisconsin, Wine, Optical recognition of handwritten digits dataset. Đánh giá hiệu quả bằng ma trận confusion. Ghi nhận kết quả. Chú ý các tập dữ liệu có thể tìm thấy trong gói thư viện scikit.