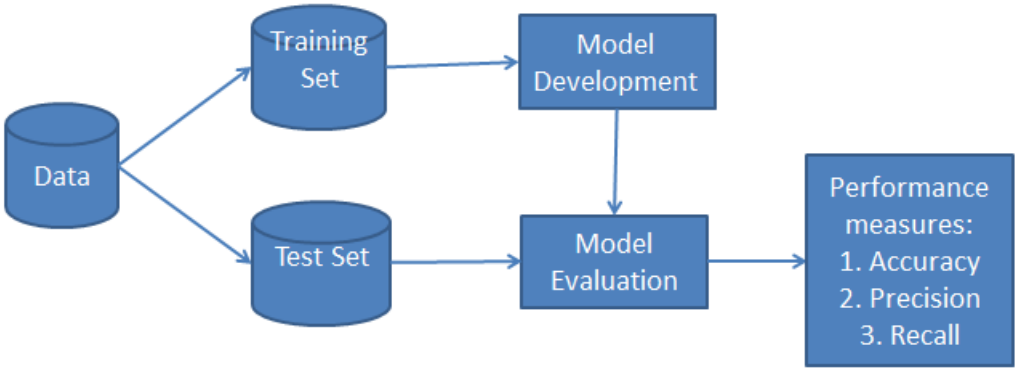
Bài thực hành

Phân loại dữ liệu với cây quyết định (Decision Tree)

1. Ví dụ minh họa sử dụng giải thuật cây quyết định (Decision Tree)

Trong ví dụ này, học viên làm quen với:

- Đọc dữ liệu từ các tập tin json
- Nắm được các bước phân loại dữ liệu



- Sử dụng cây quyết định (Decision Tree) để phân loại dữ liệu
- Biểu diễn cây quyết định

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

# Đọc tập tin json chứa tập dữ liệu iris
iris =
pd.read_json('https://raw.githubusercontent.com/ltdaovn/dataset/master/iris.j
son')
print('Dataset info:\n', iris.info)
X = iris.drop(columns=['species'])
y = iris.species

# 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 với giải thuật Cây quyết định
model = tree.DecisionTreeClassifier(criterion="gini")
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 xac cua mo hinh voi nghi thuc kiem tra hold-out: %.3f" %
accuracy_score(y_test, y_pred))

# Hiển thị cây
tree.plot_tree(model.fit(X, y))
plt.show()
```

2. Ví dụ minh họa xử lý dữ liệu được lưu trong CSDL MySQL.

Trong ví dụ này, học viên làm quen với:

- Kết nối dữ liệu đến máy chủ MySQL.
- Viết các câu truy vấn để lấy dữ liệu.

```
# Tập dữ liệu iris được lưu tại máy chủ MySQL có các thông số như sau:

#Username: uiiYzHajDl
#Database name: uiiYzHajDl
#Password: y6kbl8Na7i
#Server: remotemysql.com
#Port: 3306

# Nạp các gói thư viện cần thiết
from mysql import connector     # Có thể cần cài thêm thư viện mysql-
connector-python-rf
import pandas as pd
from sklearn import tree
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
import matplotlib.pyplot as plt
import seaborn as sns


def select_data(cursor):
    # Returns the full dataset of features and outputs
    # from the iris table.
    cursor.execute('''
        SELECT
            id,
            sepal_length,
            sepal_width,
            petal_length,
```

```
            petal_width,
            target_observed
        FROM
            iris
    ''')
    ids, X, Y = [], [], []
    for row in cursor.fetchall():
        ids.append(row[0])
        X.append([float(z) for z in row[1:-1]])
        Y.append(float(row[-1]))
    return ids, X, Y


# Kết nối đến máy chủ
connection = connector.connect(
        host='remotemysql.com',
        database='uiiYzHajDl',
        user='uiiYzHajDl',
        password='y6kbl8Na7i'
)
cursor = connection.cursor(buffered=True)

# Get iris data
ids, X, y = select_data(cursor)
columns=["Petal Length","Petal Width","Sepal Length","Sepal Width"];
X = pd.DataFrame(X, columns=columns)
print(X.head())

# Huấn luyện mô hình, đánh giá mô hình tương tự ví dụ 1
```

3. Hãy viết chương trình phân loại các tập dữ liệu sau với giải thuật cây quyết định (Decision Tree): Breast Cancer Wisconsin, Wine, Optical recognition of handwritten digits dataset. Khi chạy cần tuân theo nghi thức kiểm tra chéo k-fold, nhớ thay đổi các tham số như hàm phân hoạch và kích thước nhỏ nhất của nút mà cây quyết định không phân hoạch tiếp. Ghi nhận kết quả. Chú ý các tập dữ liệu có thể tìm thấy trong gói thư viện scikit.