**CHECKPOINT 2**

|  |  |
| --- | --- |
| **Bộ môn:** Coding | **Năm môn:** Computer Scientist – Advanced |
| **Độ tuổi học viên:** 14 - 18 Tuổi | **Thời lượng:** 60 phút |

# **TRẮC NGHIỆM**

|  |  |
| --- | --- |
| **CÂU HỎI 1:** | Khi xây dựng câu chuyện dữ liệu, việc đầu tiên cần làm là gì? |
| A. | Chọn biểu đồ phù hợp để trình bày |
| B. | Thu thập dữ liệu từ các nguồn không kiểm chứng |
| C. | Tạo báo cáo trực tiếp mà không cần phân tích |
| D. | Xác định vấn đề hoặc câu hỏi chính cần trả lời bằng dữ liệu |
| **CÂU HỎI 2:** | Để trình bày dữ liệu về tỷ lệ phần trăm các nhóm khách hàng trong một câu chuyện dữ liệu, biểu đồ nào được sử dụng phổ biến nhất? |
| A. | Biểu đồ đường (Line Chart) |
| B. | Biểu đồ tròn (Pie Chart) |
| C. | Biểu đồ phân tán (Scatter Plot) |
| D. | Biểu đồ hộp (Box Plot) |
| **CÂU HỎI 3:** | Để hiển thị biểu đồ sau khi vẽ với thư viện matplotlib, bạn cần sử dụng lệnh nào? |
| A. | plt.plot() |
| B. | plt.show() |
| C. | plt.draw() |
| D. | plt.save() |
| **CÂU HỎI 4:** | Một nhóm nghiên cứu muốn dự đoán điểm số của học sinh dựa trên số giờ học.  Họ quyết định chia dữ liệu thành tập huấn luyện và tập kiểm tra với tỷ lệ 80:20. Điều này có ý nghĩa gì? |
| A. | Tăng kích thước của dữ liệu |
| B. | Giảm thời gian huấn luyện mô hình với tập train |
| C. | Tăng độ phức tạp của mô hình |
| D. | Đảm bảo mô hình không bị huấn luyện quá nhiều với dữ liệu train (overfit) và có thể kiểm tra độ chính xác trên dữ liệu test |
| **CÂU HỎI 5:** | Trong quá trình xây dựng câu chuyện dữ liệu, bạn cần một tập dữ liệu về dân số thế giới. Bạn sẽ sử dụng nguồn nào sau đây để tìm kiếm dữ liệu đáng tin cậy? |
| A. | Kaggle |
| B. | Wikipedia |
| C. | Các blog cá nhân |
| D. | Một tệp dữ liệu bất kỳ trên mạng mà không kiểm tra nguồn gốc |
| **CÂU HỎI 6:** | Khi huấn luyện mô hình hồi quy tuyến tính đa biến với sklearn, bạn cần chuẩn bị tập dữ liệu đầu vào X như thế nào? |
| A. | Một danh sách hai chiều (2D array) với kích thước (n, d) với n tượng trưng cho số giá trị, d tượng trưng cho số biến (số đặc trưng dữ liệu) |
| B. | Một danh sách một chiều chứa tất cả các giá trị |
| C. | Một chuỗi ký tự chứa tất cả các biến |
| D. | Một dictionary chứa các cặp key-value |
| **CÂU HỎI 7:** | Bạn muốn tạo một biểu đồ tròn (pie chart) để hiển thị tỷ lệ phần trăm đóng góp của từng mục và hiển thị các giá trị phần trăm trên biểu đồ. Bạn sẽ sử dụng tham số nào trong hàm plt.pie()? |
| A. | autopct |
| B. | explode |
| C. | shadow |
| D. | startangle |
| **CÂU HỎI 8:** | Bạn thử nghiệm hai mô hình hồi quy tuyến tính khác nhau và nhận được các chỉ số kết quả sau.   * Mô hình A: MSE=100 * Mô hình B: MSE=64   Bạn sẽ lựa chọn mô hình nào, tại sao? |
| A. | Mô hình A vì có MSE lớn hơn. |
| B. | Mô hình B vì có RMSE nhỏ hơn. |
| C. | Mô hình A vì nó ổn định hơn. |
| D. | Tất cả đều sai |
| **CÂU HỎI 9:** | Để bắt đầu xây dựng mô hình hồi quy tuyến tính đa biến với thư viện sklearn, bạn cần thực hiện bước nào đầu tiên? |
| A. | Huấn luyện mô hình bằng model.fit() |
| B. | Chuẩn bị dữ liệu đầu vào (X) và đầu ra (y) ở định dạng phù hợp |
| C. | Dự đoán giá trị đầu ra bằng model.predict() |
| D. | Sử dụng thư viện matplotlib để trực quan hóa dữ liệu |
| **CÂU HỎI 10:** | Bạn đang vẽ biểu đồ đường (line plot) để so sánh doanh thu của hai cửa hàng trong suốt một năm. Cả hai đường đều sử dụng màu mặc định và khó phân biệt. Làm thế nào bạn có thể chỉnh sửa biểu đồ để dễ dàng phân biệt các đường? |
| A. | Thêm chú thích (legend) và sử dụng các màu sắc khác nhau cho từng đường. |
| B. | Tăng độ dày của cả hai đường. |
| C. | Sử dụng biểu đồ cột thay vì biểu đồ đường. |
| D. | Đặt cả hai đường trên hai biểu đồ riêng biệt. |

# **THỰC HÀNH**

|  |  |
| --- | --- |
| **ĐỀ BÀI:** | Sử dụng dữ liệu [“housing\_data.csv”](https://drive.google.com/file/d/1JwwhslG-2zOB535bhWZ4JCk6RgBQ8UQX/view?usp=sharing) của một công ty định giá nhà.  *Biết dữ liệu có các cột sau:*   * *Size: Diện tích căn nhà (m2)* * *Age: Tuổi cùa căn nhà (năm)* * *Price: Giá bán của căn nhà (ngàn $)*   **Hãy:**   1. Xây dựng mô hình Linear Regression để dự đoán giá bán của căn nhà dựa trên hai yếu tố là Diện tích và Tuổi. 2. Đánh giá độ hiệu quả của mô hình bằng chỉ số R2. 3. Vẽ 02 biểu đồ Scatter Plot biểu diễn mối quan hệ giữa Diện tích và Tuổi với Giá nhà. Đánh giá yếu tố nào biểu thị mối tương quan rõ rệt hơn. |