**KHOA HỌC DỮ LIỆU PYTHON**

**BÀI GIẢNG 1: GIỚI THIỆU VỀ KHOA HỌC DỮ LIỆU**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Nội dung câu hỏi | Lựa chọn/Số lượng câu | Đáp án đúng | Không xáo |
| 1 | Khoa học dữ liệu là gì? | A. Một lĩnh vực chỉ tập trung vào lập trình  B. Lĩnh vực kết hợp toán học, khoa học máy tính và kiến thức chuyên ngành  C. Một ngành chỉ nghiên cứu về thống kê  D. Chỉ là quá trình trực quan hóa dữ liệu | B |  |
| 2 | Thành phần nào KHÔNG thuộc quy trình khoa học dữ liệu? | A. Thu thập dữ liệu  B. Phân tích dữ liệu  C. Phát triển phần cứng máy tính  D. Trực quan hóa dữ liệu | C |  |
| 3 | Khoa học dữ liệu trong kinh tế và kinh doanh khác Khoa học dữ liệu thông thường như thế nào? | A. Chỉ sử dụng dữ liệu tài chính  B. Tập trung vào tối ưu hóa mô hình học máy hơn  C. Ưu tiên giải thích, ra quyết định và hiểu bản chất kinh tế  D. Không cần trực quan hóa dữ liệu | C |  |
| 4 | Lĩnh vực Kỹ thuật dữ liệu tập trung vào giải quyết bước nào trong chu trình khoa học dữ liệu? | A. Tập trung vào xác định, tổ chức, và thu thập dữ liệu  B. Trực quan hóa dữ liệu  C. Tất cả các bước: từ xác định vấn đề đến truyền đạt kết quả  D. Đánh giá mô hình học máy | A |  |
| 5 | Kể chuyện bằng dữ liệu là kỹ năng quan trọng nhất của ai? | A. Lập trình viên  B. Chuyên gia phân tích dữ liệu  C. Kỹ sư phần mềm  D. Kỹ sư dữ liệu | B |  |
| 6 | Học máy khác với lập trình truyền thống như thế nào? | A. Viết các câu lệnh phức tạp hơn hơn  B. Không cần dữ liệu đầu vào  C. Học từ dữ liệu thay vì được lập trình rõ ràng  D. Không cần kiểm thử | C |  |
| 7 | Các nguồn dữ liệu phổ biến trong kinh tế và kinh doanh bao gồm? | A. Hệ thống ERP, CRM, dữ liệu giao dịch, dữ liệu khách hàng  B. Dữ liệu vật lý thiên văn  C. Dữ liệu gen và sinh học phân tử  D. Dữ liệu mạng máy tính | A |  |
| 8 | Tại sao mô hình học máy trong kinh tế và kinh doanh lại ưu tiên khả năng giải thích? | A. Vì độ chính xác luôn là thấp  B. Vì cần tuân thủ quy định và ra quyết định có trách nhiệm  C. Vì không sử dụng được trong thực tế  D. Vì không cần dự báo | B |  |
| 9 | Vai trò chính của nhà khoa học dữ liệu là gì? | A. Chỉ lập trình ứng dụng  B. Phân tích, trích xuất và tạo giá trị từ dữ liệu  C. Quản lý cơ sở hạ tầng mạng  D. Chỉ thiết kế giao diện người dùng | B |  |
| 10 | Dữ liệu phi cấu trúc là gì? | A. Dữ liệu được tổ chức trong các bảng có hàng và cột  B. Dữ liệu không có cấu trúc cố định, ví dụ: văn bản, hình ảnh, video  C. Dữ liệu được lưu trữ trong hệ quản trị CSDL quan hệ  D. Dữ liệu luôn có dạng số và dễ xử lý | B |  |
| 11 | Mục tiêu chính của học không giám sát là gì? | A. Dự đoán đầu ra từ dữ liệu đã gán nhãn  B. Tối ưu hóa phần thưởng thông qua hành động  C. Phân nhóm, phát hiện cấu trúc ẩn trong dữ liệu  D. Tăng độ chính xác của mô hình học sâu | C |  |
| 12 | Các mô hình Học máy trong Khoa học dữ liệu được dùng để làm gì? | A. Tự động hóa quá trình ra quyết định dựa trên dữ liệu  B. Chỉ để thiết kế website  C. Chỉ dùng để vẽ biểu đồ dữ liệu  D. Tăng tốc độ truyền dữ liệu trong mạng | A |  |
| 13 | Python được ưa chuộng trong khoa học dữ liệu vì lý do nào? | A. Chỉ vì cú pháp ngắn gọn  B. Có nhiều thư viện hỗ trợ phân tích và học máy  C. Không cần học thuật toán cũng lập trình được  D. Chỉ chạy nhanh hơn các ngôn ngữ khác | B |  |
| 14 | Trực quan hóa dữ liệu giúp ích gì trong khoa học dữ liệu? | A. Giảm dung lượng lưu trữ dữ liệu  B. Giúp hiểu và truyền đạt thông tin từ dữ liệu một cách dễ dàng  C. Chỉ phục vụ mục đích làm đẹp báo cáo  D. Loại bỏ hoàn toàn dữ liệu nhiễu | B |  |
| 15 | Nguồn dữ liệu nào sau đây là dữ liệu phi cấu trúc? | A. Bảng chấm công nhân viên  B. Ảnh và video từ mạng xã hội  C. Bảng điểm sinh viên trong Excel  D. Danh sách sản phẩm trong cơ sở dữ liệu quan hệ | B |  |

**BÀI GIẢNG 2: PYTHON CƠ BẢN**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Nội dung câu hỏi | Lựa chọn/Số lượng câu | Đáp án đúng | Không xáo |
| 1 | Python là ngôn ngữ lập trình loại nào? | A. Biên dịch, kiểu tĩnh  B. Thông dịch, kiểu động  C. Chỉ dùng cho phát triển web  D. Chỉ dùng cho trí tuệ nhân tạo | B |  |
| 2 | Ai là người phát triển ngôn ngữ Python? | A. Dennis Ritchie  B. Guido van Rossum  C. James Gosling  D. Bjarne Stroustrup | B |  |
| 3 | Đặc điểm nổi bật của Python là gì? | A. Cú pháp phức tạp, khó đọc  B. Đơn giản, dễ học, mã nguồn mở  C. Chỉ chạy trên hệ điều hành Linux  D. Yêu cầu biên dịch trước khi chạy | B |  |
| 4 | Kiểu dữ liệu nào KHÔNG có trong Python? | A. List  B. Tuple  C. Dictionary  D. Array tĩnh kiểu cố định như C | D |  |
| 5 | Làm thế nào để in ra màn hình trong Python? | A. console.log('Hello')  B. printf('Hello')  C. System.out.println('Hello')  D. print('Hello') | D |  |
| 6 | Dấu thụt lề (indentation) trong Python có vai trò gì? | A. Chỉ để làm đẹp mã  B. Không quan trọng, có thể bỏ qua  C. Xác định cấu trúc khối lệnh  D. Chỉ áp dụng trong hàm | C |  |
| 7 | Câu lệnh nào dùng để khai báo một hàm trong Python? | A. function myFunc():  B. def myFunc():  C. func myFunc():  D. declare myFunc(): | B |  |
| 8 | Biến trong Python được khai báo như thế nào? | A. Phải chỉ định kiểu dữ liệu trước  B. Tự động gán kiểu khi gán giá trị  C. Dùng từ khóa var  D. Không cần khai báo, chỉ dùng trực tiếp | B |  |
| 9 | Thư viện nào phổ biến trong Python cho tính toán khoa học? | A. NumPy  B. React  C. Laravel  D. Spring Boot | A |  |
| 10 | Khi chạy file Python có đuôi mở rộng là gì? | A. .py  B. .java  C. .cpp  D. .js | A |  |

**BÀI GIẢNG 3: CÁC THU VIỆN PYTHON ĐƯỢC DÙNG PHỔ BIẾN TRONG KHOA HỌC DỮ LIỆU**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Nội dung câu hỏi | Lựa chọn/Số lượng câu | Đáp án đúng | Không xáo |
| 1 | Thư viện NumPy trong Python chủ yếu được sử dụng cho mục đích gì? | A. Xử lý dữ liệu dạng bảng  B. Tính toán khoa học với mảng nhiều chiều  C. Vẽ biểu đồ dữ liệu  D. Xây dựng giao diện web | B |  |
| 2 | Pandas được sử dụng để làm gì trong khoa học dữ liệu? | A. Tạo ứng dụng di động  B. Xử lý và phân tích dữ liệu dạng bảng  C. Huấn luyện mô hình học sâu  D. Lập trình nhúng | B |  |
| 3 | Matplotlib là thư viện phục vụ cho mục đích nào? | A. Trực quan hóa dữ liệu  B. Tăng tốc tính toán song song  C. Xử lý ngôn ngữ tự nhiên  D. Quản lý dữ liệu lớn | A |  |
| 4 | Seaborn được xây dựng dựa trên thư viện nào? | A. TensorFlow  B. PyTorch  C. Matplotlib  D. OpenCV | C |  |
| 5 | Scikit-learn chủ yếu hỗ trợ gì? | A. Phân tích thống kê mô tả  B. Các thuật toán học máy (Machine Learning)  C. Xử lý hình ảnh  D. Tạo dashboard trực quan | B |  |
| 6 | TensorFlow và PyTorch là các thư viện nổi tiếng trong lĩnh vực nào? | A. Xử lý văn bản  B. Phát triển ứng dụng web  C. Học sâu (Deep Learning)  D. Phân tích dữ liệu kinh tế | C |  |
| 7 | Thư viện OpenCV thường được sử dụng trong lĩnh vực nào? | A. Xử lý ảnh và thị giác máy tính  B. Phân tích dữ liệu kinh tế  C. Trực quan hóa dữ liệu  D. Tối ưu hóa mô hình toán học | A |  |
| 8 | Thư viện Statsmodels được sử dụng chủ yếu để làm gì? | A. Phân tích dữ liệu thời gian thực  B. Thống kê và mô hình hóa dữ liệu  C. Huấn luyện mạng nơ-ron  D. Quản lý cơ sở dữ liệu | B |  |
| 9 | Plotly khác với Matplotlib ở điểm nào? | A. Plotly hỗ trợ tương tác với biểu đồ  B. Plotly chỉ dùng để huấn luyện mô hình  C. Plotly không hỗ trợ vẽ biểu đồ 3D  D. Plotly chỉ chạy trên Linux | A |  |
| 10 | Khi cần xử lý dữ liệu lớn với Spark trong Python, ta thường dùng thư viện nào? | A. Dask  B. PySpark  C. Joblib  D. Numba | B |  |

**BÀI GIẢNG 4: LẬP VÀ LƯU TRỮ GIỮ LIỆU VỚI PYTHON (KO THẤY CÓ CÂU HỎI)**

**BÀI GIẢNG 5: LÀM SẠCH VÀ CHUẨN BỊ DỮ LIỆU TRONG KHOA HỌC DỮ LIỆU**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| STT | Nội dung câu hỏi | Lựa chọn/Số lượng câu | Đáp án đúng | Không xáo |
| 1 | Phương thức nào dùng để phát hiện dữ liệu thiếu trong pandas? | A. isna() hoặc isnull()  B. missing()  C. empty()  D. nan\_check() | A |  |
| 2 | Phương thức fillna() được sử dụng để làm gì? | A. Loại bỏ dữ liệu thiếu  B. Thay thế dữ liệu thiếu  C. Phát hiện dữ liệu thiếu  D. Đếm dữ liệu thiếu | B |  |
| 3 | MinMaxScaler đưa dữ liệu về khoảng giá trị nào? | A. [-1, 1]  B. [0, 1]  C. [0, 100]  D. [-100, 100] | B |  |
| 4 | Phương thức nào dùng để loại bỏ hàng trùng lặp? | A. remove\_duplicates()  B. drop\_duplicates()  C. delete\_duplicates()  D. unique() | B |  |
| 5 | Trong pandas, để chuyển chuỗi về chữ thường ta sử dụng? | A. .str.lowercase()  B. .str.lower()  C. .str.downcase()  D. .str.small() | B |  |
| 6 | Label Encoding phù hợp nhất với loại dữ liệu nào? | A. Dữ liệu số liên tục  B. Dữ liệu nominal  C. Dữ liệu ordinal  D. Dữ liệu thời gian | C |  |
| 7 | StandardScaler chuẩn hóa dữ liệu có Mean và Standard Deviation là bao nhiêu? | A. Mean=1, Std=0  B. Mean=0, Std=1  C. Mean=0.5, Std=0.5  D. Mean=100, Std=10 | B |  |
| 8 | Khi nào nên sử dụng RobustScaler thay vì MinMaxScaler? | A. Khi dữ liệu có nhiều outliers  B. Khi dữ liệu đã chuẩn hóa  C. Khi dữ liệu là categorical  D. Khi dữ liệu có kích thước nhỏ | A |  |

**BÀI GIẢNG 6: SẮP XẾP VÀ BIẾN ĐỔI DỮ LIỆU (KO THẤY CÓ CÂU HỎI)**

**BÀI GIẢNG 7: TRỰC QUAN HOÁ DỮ LIỆU (KO THẤY CÓ CÂU HỎI)**

**BÀI GIẢNG 8: CÁC MÔ HÌNH HỌC MÁY CÓ GIÁM SÁT (KO THẤY CÓ CÂU HỎI)**