Đề thi: Phân Tích Dữ Liệu

Giảng viên: Trần Sơn Hải

Thời lượng: 60 phút

Tổng số câu hỏi: 40 Câu hỏi trắc nghiệm

Hướng dẫn:

- Đọc kỹ từng câu hỏi và chọn câu trả lời đúng nhất.
- Khoanh tròn hoặc đánh dấu vào lựa chọn của bạn.

Tên sinh viên:	
Mã sinh viên:	
Lớp:	
Phòng thi:	
Thời gian thi:	

Trả lời (Khoanh tròn và đáp án đúng nhất):

1	A	B	C	D	**11**	A	B	C	D	**21**	A	B	C	D	**31**	A	B	C	D
2	A	B	C	D	**12**	A	B	C	D	**22**	A	B	C	D	**32**	A	B	C	D
3	A	B	C	D	**13**	A	B	C	D	**23**	A	B	C	D	**33**	A	B	C	D
4	A	B	C	D	**14**	A	B	C	D	**24**	A	B	C	D	**34**	A	B	C	D
5	A	B	C	D	**15**	A	B	C	D	**25**	A	B	C	D	**35**	A	B	C	D
6	A	B	C	D	**16**	A	B	C	D	**26**	A	B	C	D	**36**	A	B	C	D
7	A	B	C	D	**17**	A	B	C	D	**27**	A	B	C	D	**37**	A	B	C	D
8	A	B	C	D	**18**	A	B	C	D	**28**	A	B	C	D	**38**	A	B	C	D
9	A	B	C	D	**19**	A	B	C	D	**29**	A	B	C	D	**39**	A	B	C	D
10	A	B	C	D	**20**	A	B	C	D	**30**	A	B	C	D	**40**	A	B	C	D

Phần I: Giới thiệu về Phân tích dữ liệu và Python cơ bản (10 câu)

1. Mục tiêu chính của phân tích dữ liệu là gì? A. Chỉ thu thập dữ liệu lịch sử. B. Chuyển đổi dữ liệu thành thông tin có ý nghĩa để đưa ra quyết định sáng suốt và dự đoán. C. Lưu trữ dữ liệu vĩnh viễn trên ổ cứng. D. Tạo các bảng tính phức tạp.
2. "Data analytics" và "data analysis" khác nhau như thế nào? A. Chúng là các thuật ngữ hoàn toàn giống nhau. B. "Data analysis" là một phần con của "data analytics", tập trung vào việc trích xuất ý nghĩa từ dữ liệu, trong khi "data analytics" bao gồm các quy trình rộng hơn như khoa học dữ liệu và kỹ thuật dữ liệu. C. "Data analytics" chỉ liên quan đến dữ liệu có cấu trúc, còn "data analysis" thì không. D. "Data analysis" là một thuật ngữ cũ hơn không còn được sử dụng.
3. Theo định nghĩa, "Predictive data analytics" là gì? A. Thực hành tóm tắt dữ liệu lịch sử. B. Sử dụng dữ liệu quá khứ để xác định các mẫu và xu hướng, và sử dụng thông tin đó để dự đoán kết quả trong tương lai. C. Đề xuất các hành động dựa trên dữ liệu để đạt được mục tiêu cụ thể. D. Chỉ thu thập dữ liệu từ các nguồn truyền thống.
4. Thư viện Python nào sau đây được sử dụng phổ biến nhất để thao tác và phân tích dữ liệu dạng bảng (DataFrames)? A. Matplotlib B. Seaborn C. Pandas D. Scikit-learn
5. Trong ngữ cảnh phân tích dữ liệu, đâu là một lợi ích chính của phân tích dữ liệu (data analytics)? A. Làm cho dữ liệu phức tạp hơn. B. Giảm hiệu quả hoạt động. C. Cải thiện việc ra quyết định và tối ưu hóa hiệu suất. D. Tăng chi phí thu thập dữ liệu.
6. "Data Quality Issues" là một thách thức phổ biến trong mô hình hóa dữ liệu. Điều này đề cập đến vấn đề gì? A. Thiếu nguồn dữ liệu. B. Dữ liệu không nhất quán, không đầy đủ hoặc không chính xác. C. Khó khăn trong việc chọn công cụ trực quan hóa. D. Chi phí cao của phần mềm phân tích dữ liệu.
7. Trong Python, kiểu dữ liệu tích hợp nào thường được sử dụng để triển khai một mảng động (resizable array)? (Gợi ý: Tương tự câu 6 đề DSA) A. `tuple` B. `set` C. `list` D. `dict`
8. "Time complexity" và "Space complexity" là các đặc điểm quan trọng khi đánh giá thuật toán. Khi đánh giá thuật toán trong phân tích dữ liệu, yếu tố nào **KHÔNG** thường được coi là một đặc điểm chính? (Gợi ý: Tương tự câu 2 đề DSA) A. Time Complexity B. Space Complexity C. Readability D. Hardware Compatibility
9. Đâu là bước đầu tiên trong quá trình triển khai phân tích dự đoán (Predictive Analytics)? A. Đánh giá và xác thực mô hình. B. Lựa chọn và huấn luyện mô hình. C. Thu thập và tiền xử lý dữ liệu. D. Ứng dụng kết quả vào kinh doanh.
10. "Data Analytics" được định nghĩa là gì? A. Tập hợp các thủ tục sử dụng các phương pháp thống kê khác nhau để chuyển đổi dữ liệu thành thông tin có ý nghĩa. B. Chỉ là quá trình tóm tắt dữ liệu lịch sử. C. Một ngôn ngữ lập trình cụ thể. D. Một công cụ phần cứng cho việc lưu trữ dữ liệu.

Phần II: Xử lý và tiền xử lý dữ liệu với Pandas và NumPy (10 câu)

11. Cho DataFrame `df` của Pandas, phương thức nào dùng để hiển thị 5 dòng đầu tiên của DataFrame? A. `df.tail()` B. `df.info()` C. `df.head()` D. `df.describe()`
12. Để kiểm tra tổng số giá trị thiếu (NaN) trong mỗi cột của một DataFrame `df` trong Pandas, bạn sẽ sử dụng đoạn mã nào? A. `df.info()` B. `df.sum()` C. `df.isnull().sum()` D. `df.dropna()`

13. Mục đích của `df.describe()` trong Pandas là gì? A. Hiển thị kiểu dữ liệu của mỗi cột. B. Cung cấp thống kê mô tả cho các cột số (ví dụ: trung bình, độ lệch chuẩn, min, max, tứ phân vị). C. Liệt kê tên tất cả các cột. D. Xóa các hàng có giá trị thiếu.
14. Giả sử bạn có DataFrame `df` và muốn tạo một cột mới `TotalSales` bằng cách nhân cột `Quantity` với cột `Price`. Đoạn code Python đúng là gì? A. `df['TotalSales'] = df['Quantity'] + df['Price']` B. `df['TotalSales'] = df['Quantity'] * df['Price']` C. `df.add_column('TotalSales', df['Quantity'] * df['Price'])` D. `df.new_column('TotalSales', 'Quantity' * 'Price')`
15. Khi xử lý các bản ghi trùng lặp trong Pandas, phương thức `drop_duplicates()` sẽ làm gì theo mặc định? A. Giữ tất cả các bản sao trùng lặp. B. Chỉ giữ bản sao đầu tiên của các bản ghi trùng lặp và xóa các bản còn lại. C. Chỉ giữ bản sao cuối cùng của các bản ghi trùng lặp. D. Báo lỗi nếu có bản ghi trùng lặp.
16. Trong phân tích dữ liệu, việc chuyển đổi biến "gender" (nam/nữ) thành các giá trị số (0/1) trước khi đưa vào mô hình học máy là một ví dụ về kỹ thuật nào? A. Xử lý giá trị thiếu. B. Kỹ thuật đặc trưng (Feature Engineering) hoặc mã hóa (Encoding). C. Tổng hợp dữ liệu. D. Trực quan hóa dữ liệu.
17. Giả sử bạn muốn tính tổng doanh thu theo từng `Category` trong DataFrame `df`. Đoạn code nào sau đây là đúng? A.`df.sum('Sales').groupby('Category')` B.`df.groupby('Category').sum()['Sales']` C.`df['Sales'].groupby('Category').sum()` D.`df.groupby('Sales')['Category'].sum()`
18. Nếu bạn cần tính hệ số tương quan giữa các biến số numeric trong một DataFrame `df`, bạn sẽ sử dụng phương thức nào? A. `df.describe()` B. `df.corr()` C. `df.info()` D. `df.pivot_table()`
19. Khi bạn sử dụng `dropna()` trong Pandas, điều gì sẽ xảy ra? A. Các giá trị thiếu được điền bằng 0. B. Các giá trị thiếu được điền bằng giá trị trung bình. C. Các hàng hoặc cột chứa giá trị thiếu sẽ bị xóa. D. Các giá trị thiếu được chuyển đổi thành một chuỗi văn bản.
20. Để lưu một Pandas DataFrame `df` thành một tệp CSV mà không bao gồm chỉ mục hàng, bạn sử dụng đoạn mã nào? A. `df.to_csv('output.csv', include_index=False)` B. `df.to_csv('output.csv', index=False)` C. `df.save_csv('output.csv')` D. `df.export_csv('output.csv', index=False)`

Phần III: Trực quan hóa dữ liệu với Matplotlib và Seaborn (10 câu)

21. Loại biểu đồ nào thích hợp nhất để hiển thị phân phối của một biến liên tục duy nhất (ví dụ: tuổi của hành khách)? A. Biểu đồ cột (Bar Chart) B. Biểu đồ tròn (Pie Chart) C. Biểu đồ phân tán (Scatter Plot) D. Biểu đồ tần suất (Histogram)
22. Mục đích chính của "Data Visualization" là gì? A. Làm phức tạp hóa dữ liệu. B. Trình bày dữ liệu dưới dạng hình ảnh để tạo điều kiện hiểu và thu thập thông tin chi tiết. C. Chỉ để lưu trữ dữ liệu. D. Thực hiện các phép tính phức tạp.
23. Trong một biểu đồ đường (line chart), trục x (x-axis) thường đại diện cho điều gì? A. Các danh mục (Categories) B. Tần số (Frequency) C. Thời gian (Time) D. Độ lớn (Magnitude)
24. Biểu đồ nào là lý tưởng để hiển thị xu hướng hoặc thay đổi theo thời gian? A. Biểu đồ cột (Bar Chart) B. Biểu đồ đường (Line Chart) C. Biểu đồ tròn (Pie Chart) D. Biểu đồ phân tán (Scatter Plot)

25. Thư viện Python nào sau đây được biết đến với khả năng tạo các biểu đồ thống kê hấp dẫn và thường được xây dựng trên Matplotlib? A. Plotly B. Bokeh C. Seaborn D. D3.js
26. Khi nào thì biểu đồ phân tán (Scatter Plot) là lựa chọn tốt nhất? A. Để so sánh các giá trị giữa các danh mục khác nhau. B. Để hiển thị mối quan hệ giữa hai biến số số học (numerical variables). C. Để hiển thị tỷ lệ của một tổng thể. D. Để minh họa các phần tử của một danh sách.
27. Mục đích của một biểu đồ hộp (Box Plot) trong trực quan hóa dữ liệu là gì? A. Chỉ hiển thị giá trị trung bình và trung vị. B. Hiển thị phạm vi và phân phối của dữ liệu, bao gồm các tứ phân vị và giá trị ngoại lai. C. Làm nổi bật các điểm dữ liệu riêng lẻ. D. Biểu diễn các mối quan hệ phân cấp.
28. Biểu đồ nào phù hợp nhất để so sánh các phần với một tổng thể (parts to a whole)? A. Biểu đồ phân tán (Scatter Plot) B. Biểu đồ cột (Bar Chart) C. Biểu đồ tròn (Pie Chart) D. Biểu đồ bong bóng (Bubble Chart)
29. "Visual Clutter" trong trực quan hóa dữ liệu đề cập đến điều gì? A. Dữ liệu không đầy đủ hoặc không chính xác. B. Sử dụng quá nhiều màu sắc. C. Tải quá nhiều yếu tố hoặc thông tin không cần thiết vào biểu đồ, khiến nó khó hiểu. D. Chọn sai loại biểu đồ.
30. Theo các phương pháp hay nhất (Best Practices) cho trực quan hóa dữ liệu hiệu quả, điều gì nên được thực hiện? A. Sử dụng càng nhiều màu càng tốt để đa dạng. B. Giữ tỷ lệ "data-to-ink" thấp (hạn chế mực không chứa thông tin). C. Tránh gắn nhãn các trục để đơn giản. D. Sử dụng hiệu ứng 3D để làm cho biểu đồ hấp dẫn hơn.

Phần IV: Phân tích thống kê và Giới thiệu Mô hình học máy (10 câu)

31. Trong thống kê mô tả, giá trị nào được tính bằng cách lấy tổng của tất cả các giá trị trong một tập dữ liệu và chia cho số lượng giá trị? A. Trung vị (Median) B. Yếu vị (Mode) C. Trung bình (Mean) D. Độ lệch chuẩn (Standard Deviation)
32. Loại phân tích dữ liệu nào tập trung vào việc tóm tắt và khám phá dữ liệu để hiểu các đặc điểm của nó ("What happened?")? A. Phân tích Dự đoán (Predictive Analytics) B. Phân tích Chẩn đoán (Diagnostic Analytics) C. Phân tích Mô tả (Descriptive Analytics) D. Phân tích Đề xuất (Prescriptive Analytics)
33. Trong hồi quy tuyến tính (Linear Regression), mục tiêu chính là gì? A. Phân loại dữ liệu thành các nhóm. B. Dự đoán một kết quả liên tục dựa trên các biến đầu vào. C. Xác định các giá trị ngoại lai trong tập dữ liệu. D. Tìm các mẫu ẩn trong dữ liệu không được gắn nhãn.
34. "Regression", "Classification" và "Clustering" khác nhau về "Input" như thế nào? A. Regression và Classification đều yêu cầu dữ liệu có nhãn, trong khi Clustering không yêu cầu nhãn. B. Regression yêu cầu dữ liệu có nhãn, còn Classification và Clustering thì không. C. Chỉ có Clustering yêu cầu dữ liệu có nhãn. D. Cả ba kỹ thuật đều không yêu cầu dữ liệu có nhãn.
35. Mục tiêu chính của thuật toán phân loại (Classification) là gì? A. Dự đoán các giá trị số. B. Nhóm các điểm dữ liệu tương tự nhau. C. Xác định các giá trị ngoại lai trong tập dữ liệu. D. Phân loại dữ liệu thành các danh mục rời rạc.
36. "Unsupervised learning" là gì trong ngữ cảnh phân tích dữ liệu? A. Kỹ thuật học máy sử dụng dữ liệu có nhãn để huấn luyện mô hình. B. Kỹ thuật học máy nơi các biến đầu ra không được cung cấp và thuật toán cố gắng tìm các cấu trúc hoặc mẫu ẩn trong dữ liệu. C.

Một phương pháp để dự đoán các giá trị liên tục. D. Một cách để đánh giá hiệu suất mô hình bằng các giá trị thực tế.

37. Trong hồi quy (Regression), thuật ngữ "residuals" (phần dư) đề cập đến điều gì? A. Các giá trị được dự đoán trong mô hình. B. Sự khác biệt giữa các giá trị quan sát và giá trị được dự đoán. C. Các biến độc lập trong mô hình. D. Các giá trị ngoại lai trong tập dữ liệu.
38. Thuật toán nào sau đây là một kỹ thuật phân cụm (Clustering) dựa trên mật độ? A. K-Means B. Hồi quy tuyến tính (Linear Regression) C. DBSCAN D. Cây quyết định (Decision Tree)
39. Ma trận nhầm lẫn (Confusion Matrix) được sử dụng cho mục đích gì trong phân loại (Classification)? A. Để dự đoán các giá trị số. B. Để đánh giá hiệu suất của một mô hình phân loại. C. Để nhóm các điểm dữ liệu tương tự nhau. D. Để xác định mối tương quan giữa các biến.
40. Theo các slide bài giảng, đâu là một trong những ứng dụng thực tế của phân tích dự đoán (Predictive Analytics) trong lĩnh vực Sales & Marketing? A. Chỉ để lưu trữ hồ sơ bán hàng. B. Tối ưu hóa các chiến dịch tiếp thị và xác định các cơ hội bán hàng tiềm năng. C. Giảm thiểu tương tác với khách hàng. D. Chỉ phân tích dữ liệu lịch sử mà không đưa ra dự đoán.

----------hết----------