Họ, tên thí sinh:..... Mã số: .....

Câu 1: Lấy mẫu là gì

A. Là quá trình phân tích tập dữ liệu con trong tập dữ liệu lớn
B. Là quá trình biến đổi từ dữ liệu lớn thành một tập dữ liệu con
C. Là quá trình xử lý cắt gọn dữ liệu
D. Là quá trình chọn ra một tập con trong tập dữ liệu lớn

Câu 2: Median của một tập hợp là gì

A. Không đáp án nào đúng
B. Giá trị độ lệch chuẩn của tập hợp
C. Giá trị có độ lớn nằm chính giữa tập hợp
D. Giá trị trung bình của tập hợp

Câu 3: Missing data có thể được chia ra thành mấy loại

A. 6 B. 3 C. 4 D. 5

Câu 4: Một cửa hàng sách ước lượng rằng: Trong tổng số khách hàng đến cửa hàng, có 30% khách cần hỏi nhân viên bán hàng, 20% khách mua sách, và 15% khác thực hiện cả 2 điều trên. Gặp ngẫu nhiên 1 khách trong nhà sách. Tính xác suất để người này không thực hiện 2 điều trên

A. 0.60 B. 0.55 C. 0.65 D. 0.70

Câu 5: Đâu là các phương pháp giảm chiều dữ liệu

A. Filter
B. Embedded
C. Tất cả các phương pháp trên
D. Wrapper

Câu 6: Overfitting và Underfitting xảy ra khi nào

A. Trong quá trình suy luận
B. Trong quá trình phân tích
C. Tất cả các quá trình trên
D. Trong quá trình huấn luyện

Câu 7: Thư viện python nào được sử dụng để crawl dữ liệu từ internet

A. Beautifulsoup B. Numpy C. Keras D. Sci kit learn

Câu 8: Một cửa hàng sách ước lượng rằng: Trong tổng số khách hàng đến cửa hàng, có 30% khách cần hỏi nhân viên bán hàng, 20% khách mua sách, và 15% khác thực hiện cả 2 điều trên. Gặp ngẫu nhiên 1 khách trong nhà sách. Tính xác suất để người này không mua sách, biết rằng người này có hỏi nhân viên bán hàng

A. 0.6 B. 0.5 C. 0.4 D. 0.3

Câu 9: Lĩnh vực nào sau đây là sự tổng hợp của các kỹ năng: Machine learning, hacking skill, maths and stats, substantive research

A. Data Analysis
B. Descriptive Analytics
C. Data Science
D. Không có đáp án nào đúng

Câu 10: CNN trong học sâu là viết tắt của từ gì

A. Convolutional neural network
B. Tất cả các phương án trên đều sai
C. Coventional neural network
D. Convertible neural network

Câu 11: Lợi ích của việc lấy mẫu

A. Giảm thời gian tính toán khi tập mẫu có độ lớn giảm đi
B. Giảm không gian lưu trữ
C. Tối ưu khả năng hiển thị của tập mẫu
D. Tăng độ chính xác của các thử nghiệm phân tích

Câu 12: Khi crawl dữ liệu sử dụng API, định dạng file nào sau đây được sử dụng

A. XML và JSON B. HTML C. XML D. JSON

Câu 13: Nhìn theo góc nhìn của hệ quản trị cơ sở dữ liệu, có thể chia dữ liệu thành mấy nhóm

A. 4 B. 1 C. 3 D. 2

Câu 14: Học tăng cường thường áp dụng vào lĩnh vực nào

A. Các nền tảng mua sắm trực tuyến B. Game
C. Giám sát giao thông D. Hệ gợi ý

Câu 15: Dữ liệu sau khi xử lý có đặc điểm gì

A. Không đáp án nào đúng
B. Dữ liệu có thể sử dụng trực tiếp bằng các thuật toán học máy
C. Dữ liệu chưa sẵn sàng cho việc phân tích
D. Dữ liệu khó có thể sử dụng cho phân tích

Câu 16: Đâu KHÔNG phải là công dụng của PCA

A. Tối ưu hóa không gian mẫu
B. Phân tích sâu hơn, và đơn giản hơn sau khi được áp dụng PCA
C. Giải thích, trực quan hóa dữ liệu
D. Tìm mối quan hệ giữa các biến trong dữ liệu

Câu 17: Đánh giá mô hình học máy hồi quy thì sử dụng độ đo gì

A. Recall B. F1 score C. Precision D. MSE

Câu 18: Mục đích của giảm chiều dữ liệu

A. Tăng độ chính xác của phương pháp phân tích
B. Giảm thời lượng tính toán
C. Giảm không gian lưu trữ
D. Tất cả các phương án trên

Câu 19: Tại sao học sâu lại trở nên phổ biến

A. Tất cả các phương án trên
B. Độ chính xác cao
C. Huấn luyện được với dữ liệu lớn
D. Khả năng tính toán của máy tính ngày càng tăng

Câu 20: F1 score được tính dựa trên các giá trị nào

A. ROC B. Precision & Recall
C. Recall D. Precision

Câu 21: Các phương pháp giảm chiều dữ liệu theo hướng trích chọn đặc trưng là

A. Kernal PCA B. PCA
C. Tất cả các phương án trên D. LDA

Câu 22: Tại sao phải biến đổi, co giãn dữ liệu về dạng phân phối chuẩn

A. Vì dữ liệu phân phối chuẩn nhẹ hơn khi lưu trữ
B. Vì dữ liệu phân phối chuẩn cho phân tích chính xác
C. Vì dữ liệu phân phối chuẩn đồng nhất
D. Tất cả các phương án trên

Câu 23: Dữ liệu chứng khoán của một công ty chứng khoán được thu thập hàng năm, để dự đoán giá trị chứng khoán trong tương lai, kỹ thuật nào phải được sử dụng

A. Clustering B. Không có đáp án nào đúng
C. Regression D. Classification

Câu 24: Biến ngẫu nhiên liên tục là gì

A. Là biến ngẫu nhiên có giá trị có thể có của nó được xếp thành dãy hữu hạn hoặc vô hạn
B. Là biến ngẫu nhiên có miền giá trị liên tục biến đổi
C. Là biến ngẫu nhiên có giá trị có thể có của nó có thể lấp đầy khoảng giá trị mà nó có
D. Tất cả các phát biểu trên đều đúng

Câu 25: SVM là thuật toán học máy dạng nào

A. Học chuyển tiếp B. Học có giám sát
C. Học không giám sát D. Học tăng cường

Câu 26: Lợi ích của data transformation

A. Khả năng tương thích trên các nền tảng khác nhau
B. Dữ liệu nhất quán
C. Dễ sử dụng dữ liệu hơn
D. Tất cả các nhận định trên

Câu 27: Phương pháp giảm chiều dữ liệu nào sau đây liên quan đến các phương pháp học máy

A. Tất cả các phương án trên
B. Filter
C. Embedded
D. Wrapper

Câu 28: Đâu không phải là một phép toán hợp lệ giữa hai vector

A. Nhân có hướng **B.** Trừ **C.** Chia **D.** Cộng

Câu 29: Dữ liệu được thu thập từ hành vi xem trang web của một ngân hàng. Kỹ thuật nào sẽ được sử dụng để tìm những pages được xem phổ biến trong cùng một lượt xem website.

A. Clustering
B. Association Rules
C. Regression
D. Classification

Câu 30: Công việc nào sau đây được thực hiện bởi Data Scientist

A. Tất cả các đáp án trên
B. Đánh giá kết quả
C. Định nghĩa câu hỏi
D. Tạo ra những đoạn mã nguồn có thể được triển khai lại bằng ngôn ngữ khác

Câu 31: Nhận định nào sau đây về dữ liệu là đúng

A. Cần có một lượng lớn dữ liệu để có được kết quả thú vị
B. Cần một lượng dữ liệu vừa đủ và một phương pháp phân tích tốt để có được kết quả thú vị
C. Không cần một lượng dữ liệu lớn để có được kết quả thú vị
D. Dữ liệu càng nhiều thì tất cả các phương pháp học máy sẽ càng tốt hơn

Câu 32: Học máy truyền thống khác học sâu ở điểm nào

A. Loại dữ liệu
B. Phương pháp học
C. Độ sâu của dữ liệu
D. Bộ trích chọn đặc trưng

Câu 33: Missing data có thể xuất hiện do:

A. Lỗi chương trình
B. Người dùng quên điền khảo sát
C. Dữ liệu mất trong quá trình chuyển thủ công
D. Tất cả các phương án trên

Câu 34: Trong các loại phân phối của dữ liệu, phân phối nào gặp nhiều nhất trong tự nhiên

A. Phân phối liên tục
B. Phân phối ngẫu nhiên
C. Phân phối chuẩn
D. Phân phối hai phía

Câu 35: Đâu là những gánh nặng của việc thu thập dữ liệu lớn

A. Dữ liệu lớn tồn tại nhiễu đo lường cao
B. Dữ liệu lớn làm việc lưu trữ dữ liệu bị quá tải
C. Tất cả các điều trên
D. Dữ liệu lớn làm chậm vòng lặp quá trình xử lý

Câu 36: Đâu KHÔNG phải là một trong các nhóm phương pháp học máy

A. Học có giám sát
B. Học không giám sát
C. Học tăng cường
D. Học tự do

Câu 37: Giảm chiều dữ liệu sẽ giảm đặc tính nào sau đây của dữ liệu

A. Stochastics **B.** Collinearity **C.** Entropy **D.** Performance

Câu 38: Để xử lý missing data, phương pháp nào sau đây là đúng

A. Xóa bỏ missing data
B. Lấy ngẫu nhiên giá trị nào đó điền vào vị trí missing data
C. Nhân bản missing data
D. Tất cả các phương án trên

Câu 39: Machine learning là một nhánh của

A. Data mining
B. Data learning
C. Artificial Intelligence
D. Deep learning

Câu 40: Cho 2 vector A(10,15), B(5,6), tính khoảng cách giữa 2 vector

A. 11.3
B. 8.1
C. 10.3
D. 9.2

Câu 41: Ngôn ngữ nào được sử dụng phổ biến trong lĩnh vực Data Science

A. C/C++
B. Java
C. Ruby
D. Python

Câu 42: Machine learning là một nhánh của

A. Deep learning
B. Data mining
C. Data learning
D. Artificial Intelligence

Câu 43: Nhận định nào dưới đây là đúng

A. Không cần khả năng lập trình xuất sắc để thành công trong lĩnh vực Data Science
B. Người làm trong lĩnh vực Data Science phải có bằng cấp liên quan về lập trình
C. Người làm trong lĩnh vực Data Science buộc phải dùng các công cụ lưu trữ dữ liệu như SQL, MySQL
D. Cần Khả năng lập trình xuất sắc để thành công trong lĩnh vực Data Science

Câu 44: Chỉ ra câu phát biểu đúng

A. Raw data là dữ liệu có được sau các bước xử lý dữ liệu
B. Raw data là nguồn dữ liệu gốc chưa qua xử lý
C. Không đáp án nào đúng
D. Preprocessed data là nguồn dữ liệu gốc

Câu 45: Một công ty có tập dữ liệu về hành vi mua hàng của khách hàng, họ muốn phân nhóm khách hàng thì thuật toán nào nên được lựa chọn

A. Regression
B. Clustering
C. Association
D. Classification

Câu 46: Công việc liên quan đến việc biểu diễn trực quan dữ liệu dạng hình ảnh

A. Visualization
B. Analyzing
C. Preprocessing
D. Fill missing value

Câu 47: Đâu là kỹ năng cần thiết của người làm trong lĩnh vực Data Science

A. Machine Learning
B. Tất cả các kỹ năng trên
C. Data Visualization
D. Statistics

Câu 48: Một công ty cần tuyển 4 nhân viên. Có 8 người gồm 5 nam, 3 nữ nộp hồ sơ ứng tuyển, mỗi người có cơ hội được tuyển như nhau. Tính xác suất để 4 người đc tuyển, Có 3 nữ, biết rằng có ít nhất 1 nữ đã được tuyển

A. 1/14
B. 1/15
C. 1/12
D. 1/13

Câu 49: Một công ty cần tuyển 4 nhân viên. Có 8 người gồm 5 nam, 3 nữ nộp hồ sơ ứng tuyển, mỗi người có cơ hội được tuyển như nhau. Tính xác suất để 4 người đc tuyển có không quá 2 nam

A. 0.7
B. 0.5
C. 0.8
D. 0.6

Câu 50: Tính giá trị trung bình của tập hợp sau {70, 70, 80, 85,85,90,95,95,100,100}

A. 110
B. 85,95, và 100
C. 87
D. 30

Câu 51: Nhận định nào sau đây về dữ liệu là đúng

A. Cần một lượng dữ liệu vừa đủ và một phương pháp phân tích tốt để có được kết quả thú vị
B. Cần có một lượng lớn dữ liệu để có được kết quả thú vị
C. Dữ liệu càng nhiều thì tất cả các phương pháp học máy sẽ càng tốt hơn
D. Không cần một lượng dữ liệu lớn để có được kết quả thú vị

Câu 52: Người làm trong lĩnh vực data science dành phần lớn thời gian để

A. Chuẩn bị dữ liệu
B. Đóng gói dữ liệu
C. Làm giàu dữ liệu
D. Phân tích dữ liệu

Câu 53: Đâu không phải là một công cụ cho phân tích dữ liệu thống kê

A. Linear & Non-Linear Regression
B. ANOVA
C. Logistic Regression
D. Histogram

---------- HẾT ----------