TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

Mã đề thi: MADETHI

ĐỀ THI CUỐI HỌC KỲ I (2024-2025)
MÔN: LẬP TRÌNH PYTHON CHO MÁY HỌC
MÃ LỚP: MALOP
Thời gian: 90 phút

(Đề thi gồm có 12 trang. Sinh viên không được sử dụng tài liệu. Sinh viên làm bài trực tiếp trên đề. Sinh viên được sử dụng máy tính cầm tay.)

STT:	**ĐIỂM**	**CÁN BỘ COI THI**
MSSV:		
HỌ VÀ TÊN:		
PHÒNG THI:		

BẢNG TRẢ LỜI TRẮC NGHIỆM

1 (A) (B) (C) (D)	**6** (A) (B) (C) (D)	**11** (A) (B) (C) (D)	**16** (A) (B) (C) (D)	**21** (A) (B) (C) (D)
2 (A) (B) (C) (D)	**7** (A) (B) (C) (D)	**12** (A) (B) (C) (D)	**17** (A) (B) (C) (D)	**22** (A) (B) (C) (D)
3 (A) (B) (C) (D)	**8** (A) (B) (C) (D)	**13** (A) (B) (C) (D)	**18** (A) (B) (C) (D)	**23** (A) (B) (C) (D)
4 (A) (B) (C) (D)	**9** (A) (B) (C) (D)	**14** (A) (B) (C) (D)	**19** (A) (B) (C) (D)	**24** (A) (B) (C) (D)
5 (A) (B) (C) (D)	**10** (A) (B) (C) (D)	**15** (A) (B) (C) (D)	**20** (A) (B) (C) (D)	**25** (A) (B) (C) (D)

I. CÂU HỎI TRẮC NGHIỆM (5 điểm; 0.2 điểm/câu; sinh viên **chọn một hoặc nhiều đáp án đúng dựa theo yêu cầu của từng câu hỏi** và tô vào **BẢNG TRẢ LỜI TRẮC NGHIỆM**. Đối với những câu hỏi có nhiều đáp án đúng, sinh viên cần tô và chỉ tô tất cả đáp án đúng để được trọn vẹn điểm. Nếu tô thiếu hoặc sai đáp án, sinh viên sẽ không được tính điểm.)

Câu 1. *[Một đáp án đúng] (G2)* Xét mô hình hồi quy Log-Log với phương trình:

$$\ln y = \beta_0 + \beta_1 \ln x + \epsilon$$

trong đó y là tiền lương, x là số năm đi học (education), và ϵ là sai số ngẫu nhiên. Giả sử dữ liệu thực tế cho thấy hệ số $\beta_1 = 0.5$ có ý nghĩa thống kê cao ($p < 0.05$). Điều nào sau đây là đúng?

A. Tăng số năm đi học thêm 1 đơn vị sẽ làm tăng lương y thêm 0.5 đơn vị.

B. Tăng số năm đi học thêm 1% sẽ làm lương y tăng thêm 0.5%.

C. Tăng số năm đi học thêm 10% sẽ làm lương y tăng thêm 5%.

D. Tăng số năm đi học thêm 1% sẽ làm lương y tăng thêm 5%.

Câu 2. *[Một đáp án đúng] (G2)* Giả sử bạn đang phân tích dữ liệu doanh thu (biến y) của một chuỗi cửa hàng và nhận thấy y có phân phối lệch phải (positively skewed). Bạn đã dùng phép biến đổi $y_{\log} = \ln(y)$ và xây dựng mô hình hồi quy tuyến tính trên không gian log. Giả sử mục tiêu là tối ưu MAE (Mean Absolute Error), khi có $\widehat{y}_{\log}$, làm cách nào chuyển về không gian gốc y?

A. Dùng trực tiếp $y_{\log}$ làm giá trị dự đoán cho y.

B. Lấy $\exp\big(\widehat{y}_{\log}\big)$ để chuyển về không gian y.

C. Lấy $\exp\big(\widehat{y}_{\log} + \sigma^2/2\big)$, với σ^2 là phương sai lỗi trong không gian log.

D. Nhân $\widehat{y}_{\log}$ với $\ln(10)$ trước khi đổi về không gian y.

Câu 3. *[Một đáp án đúng] (G2)* Xét mô hình hồi quy tuyến tính với hai biến độc lập X_1 và X_2, cùng với biến tương tác $X_1 \times X_2$:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) + \epsilon$$

Trong đó, ϵ là sai số ngẫu nhiên. Giả sử bạn ước lượng được các hệ số hồi quy và nhận thấy rằng hệ số β_3 có ý nghĩa thống kê cao. Điều này ngụ ý điều gì về mối quan hệ giữa X_1, X_2, và Y?

A. Ảnh hưởng của X_1 lên Y không phụ thuộc vào X_2.

B. Ảnh hưởng của X_2 lên Y không phụ thuộc vào X_1.

C. Ảnh hưởng của X_1 lên Y thay đổi theo giá trị của X_2.

D. Không có mối quan hệ tương tác giữa X_1 và X_2 trong việc ảnh hưởng đến Y.

Câu 4. *[Một đáp án đúng] (G2)* Trong bài toán hồi quy, giả sử y là dự đoán của mô hình, t là giá trị thực, và $y^* = \mathbb{E}[t|x]$ là giá trị dự đoán tối ưu theo Bayes (giá trị kỳ vọng của t với điều kiện x). Lỗi kỳ vọng được phân tách thành:

$$\mathbb{E}[(y-t)^2] = (\mathbb{E}[y] - y^*)^2 + \text{Var}(y) + \text{Var}(t),$$

với $(\mathbb{E}[y] - y^*)^2$ là bias, $\text{Var}(y)$ là variance, và $\text{Var}(t)$ là Bayes error. Giả sử chúng ta áp dụng phương pháp bagging trên tập dữ liệu huấn luyện để xây dựng một mô hình dự đoán. Xét các phát biểu sau, khẳng định nào là đúng:

A. Bagging giúp giảm bias của mô hình bằng cách trung bình hóa dự đoán từ các bộ mô hình con được xây dựng trên các tập dữ liệu bootstrap.

B. Bagging không thay đổi bias và variance của mô hình chính, nhưng có thể giảm Bayes error bằng cách sử dụng các tập dữ liệu con bootstrap.

C. Bagging không ảnh hưởng đến bias nhưng có thể giảm variance bằng cách trung bình hóa dự đoán từ các mô hình con.

D. Bagging không thay đổi bias, không giảm Bayes error, nhưng giảm variance của mô hình bằng cách trung bình hóa dự đoán từ các bộ mô hình con.

Câu 5. *[Nhiều đáp án đúng] (G2)* Một mô hình bị hiện tượng quá khớp (**overfitting**) khi:

A. Mô hình quá phức tạp (số lượng tham số nhiều, bậc của tham số cao, hoặc hàm phi tuyến) và do đó rất linh hoạt khi học/huấn luyện.

B. Mô hình quá hạn chế các tham số khi huấn luyện và do đó bị giới hạn về khả năng biểu diễn của mô hình.

C. Mô hình thường dự đoán không chính xác ngay cả trên các mẫu huấn luyện.

D. Mô hình tập trung quá nhiều vào chi tiết nhiễu của tập huấn luyện, dẫn đến không có tính tổng quát.

Câu 6. *[Một đáp án đúng] (G2)* Giả sử bạn có một tập dữ liệu gồm các giá trị X = 3, 7, 8, 5, 12, 14, 21, 13, 18. Hãy tính các phân vị Q_1, Q_2, Q_3 theo phương pháp sử dụng nội suy (interpolation) như trong `np.percentile` của thư viện NumPy. Kết quả đúng là:

A. $Q_1 = 6.5$, $Q_2 = 11.0$, $Q_3 = 14.0$.
B. $Q_1 = 7.0$, $Q_2 = 12.0$, $Q_3 = 15.0$.
C. $Q_1 = 7.0$, $Q_2 = 13.0$, $Q_3 = 14.0$.
D. $Q_1 = 7.0$, $Q_2 = 12.0$, $Q_3 = 14.0$.

Câu 7. *[Một đáp án đúng] (G2)* Căn cứ đoạn mã Python sau đây:

```
import pandas as pd
df_a = pd.DataFrame({
    'Name': ['Alice', 'Bob', 'Charlie'],
    'Age': [25, 30, 35],
    'Score': [85, 90, 95]},
    index=[1, 2, 3])
df_b = pd.DataFrame({
    'Name': ['Alice', 'Charlie', 'Eve'],
    'Country': ['USA', 'UK', 'Canada'],
    'Salary': [50000, 60000, 70000]},
    index=[1, 3, 4])
```

Nên viết dòng lệnh nào để tạo ra `df_c` có dữ liệu như sau:

	Name	Age	Score	Country	Salary
1	Alice	25	85	USA	50000
2	Bob	30	90	NaN	NaN
3	Charlie	35	95	UK	60000
4	Eve	NaN	NaN	Canada	70000

A. `df_c = pd.concat([df_a, df_b], axis=1)`

B. `df_c = pd.merge(df_a, df_b, how='inner', on='Name')`

C. `df_c = pd.merge(df_a, df_b, how='outer', on='Name')`

D. `df_c = pd.concat([df_a, df_b], axis=0)`

Câu 8. *[Một đáp án đúng] (G2)* Sau khi thực thi đoạn mã Python sau đây sử dụng thư viện `numpy`, giá trị của biến `result` bằng:

```
import numpy as np
a = np.array([1, 2, 3])
b = np.array([4, 5, 6])
c = np.array([7, 8, 9])
result = np.dot(a, b) - np.dot(a, c)
```

A. -24 B. -6 C. -18 D. 0

Câu 9. *[Một đáp án đúng] (G2)* Vai trò của hệ số tương quan (correlation coefficient) là:

A. Đo lường độ nghiêng (skew) của dữ liệu. B. Đo lường mức độ phân tán của dữ liệu.

C. Đo lường mức độ liên quan tuyến tính giữa hai biến. D. Xác định độ lệch chuẩn của một biến.

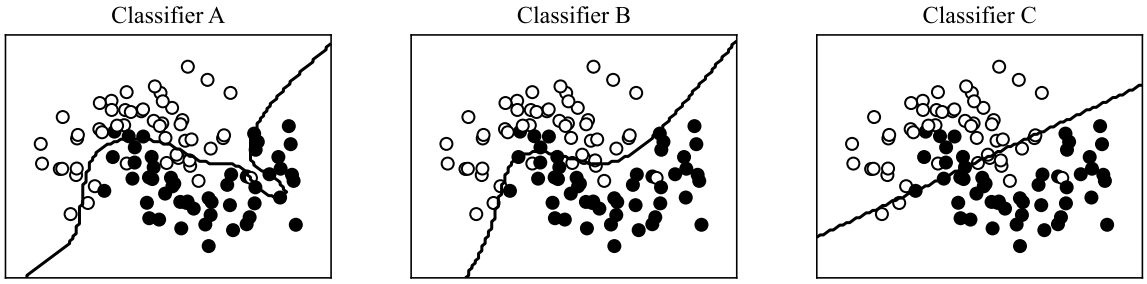
Câu 10. *[Một đáp án đúng] (G2)* Mục đích của phân tích thành phần chính (Principal Component Analysis) trong bước feature engineering là:

A. Giữ nguyên số chiều của dữ liệu nhưng thay đổi giá trị các thành phần.

B. Tăng số chiều của dữ liệu.

C. Giảm số chiều của dữ liệu.

Câu 11. *[Một đáp án đúng] (G2)* Trong bài toán phân loại nhị phân, sau khi huấn luyện bằng ba mô hình học máy khác nhau, chúng ta thu được ba bộ phân loại: Classifier A, Classifier B và Classifier C. Đường biên quyết định (decision boundary) của các bộ phân loại này trên tập kiểm tra được minh họa như sau:



Khẳng định nào dưới đây mô tả chính xác hiện tượng xảy ra với các bộ phân loại:

A. Classifier A gặp hiện tượng underfitting, Classifier B được huấn luyện phù hợp (well-fitted), và Classifier C gặp hiện tượng overfitting.

B. Classifier A gặp hiện tượng overfitting, Classifier B được huấn luyện phù hợp (well-fitted), và Classifier C gặp hiện tượng underfitting.

C. Classifier A được huấn luyện phù hợp (well-fitted), Classifier B gặp hiện tượng underfitting, và Classifier C gặp hiện tượng overfitting.

D. Classifier A gặp hiện tượng underfitting, Classifier B gặp hiện tượng overfitting, và Classifier C được huấn luyện phù hợp (well-fitted).

Câu 12. *[Nhiều đáp án đúng] (G2)* Với một tập huấn luyện cố định, bằng cách tuần tự thêm các tham số để tăng tính linh hoạt cho mô hình, chúng ta có khả năng quan sát thấy:

A. Sự chênh lệch lớn hơn giữa lỗi huấn luyện và lỗi kiểm thử.

B. Sự chênh lệch nhỏ hơn giữa lỗi huấn luyện và lỗi kiểm thử.

C. Lỗi huấn luyện tăng lên hoặc ổn định.

D. Lỗi huấn luyện giảm xuống.

Câu 13. *[Một đáp án đúng] (G2)* Giả sử bạn đang xây dựng một mô hình dự đoán khả năng trả nợ dựa trên dữ liệu từ **50 chi nhánh ngân hàng khác nhau**, với mục tiêu phân loại khách hàng thành hai nhóm: **"trả nợ đúng hạn"** và **"trễ hạn"**. Các lớp này **bị mất cân bằng**. Chiến lược kiểm định chéo (cross-validation) nào trong `sklearn` phù hợp nhất để đánh giá mô hình dự đoán trên dữ liệu từ **các chi nhánh ngân hàng không có trong tập huấn luyện**?

A. `GroupKFold`.
B. `KFold`.
C. `StratifiedKFold`.
D. `StratifiedGroupKFold`.

Câu 14. *[Nhiều đáp án đúng] (G2)* Chọn các phát biểu đúng về `RandomizedSearchCV` và `GridSearchCV` dưới đây:

A. `RandomizedSearchCV` có chi phí tính toán cố định thông qua siêu tham số `n_iter`.

B. `GridSearchCV` có chi phí tính toán lớn khi số lượng siêu tham số tăng lên.

C. Cả `GridSearchCV` và `RandomizedSearchCV` đều sử dụng phân phối xác suất để chọn giá trị siêu tham số.

D. `RandomizedSearchCV` cho phép kiểm tra tất cả tổ hợp của tập hợp siêu tham số.

Câu 15. *[Một đáp án đúng] (G2)* Với tập dữ liệu một tỷ mẫu có độ phức tạp cao, đặc trưng output có quan hệ phi tuyến tính với đặc trưng input, và bài toán phân loại 10 lớp đối tượng. Chúng ta nên sử dụng các kỹ thuật nào sau đây để đạt hiệu quả cao và tối ưu hóa tài nguyên tính toán:

A. Mô hình Logistic Regression, Holdout Validation và đánh giá bằng độ đo Root Mean Squared Error (RMSE).

B. Mô hình Random Forest, Holdout Validation và đánh giá bằng độ đo Accuracy.

C. Mô hình Random Forest, Cross-Validation và đánh giá bằng độ đo Accuracy.

D. Mô hình Logistic Regression, Cross-validation và đánh giá bằng độ đo Root Mean Squared Error (RMSE).

Câu 16. *[Nhiều đáp án đúng] (G2)* Chọn các phát biểu đúng:

A. Bagging kết hợp các mô hình cơ sở (predictors) một cách đồng thời.

B. Cả bagging và boosting đều kết hợp nhiều mô hình cơ sở (predictors).

C. Boosting kết hợp các mô hình cơ sở (predictors) một cách tuần tự.

D. Cả bagging và boosting chỉ làm việc với cây quyết định.

Câu 17. *[Một đáp án đúng] (G2)* Target encoding khác với one-hot encoding như thế nào và những rủi ro tiềm ẩn của nó là gì?

A. Target encoding thay thế các các giá trị phân loại bằng một số nguyên cố định, có thể làm phát sinh thêm bộ nhớ của máy tính.

B. Target encoding thay thế các giá trị phân loại (categories) bằng giá trị trung bình của biến mục tiêu (target variable) cho các categories đó, có thể dẫn đến rò rỉ dữ liệu (data leakage).

C. Target encoding chuyển đổi các biến phân loại thành các véctơ nhị phân, dẫn đến dữ liệu có số chiều cao.

D. Target encoding biến đổi các biến phân loại thành một chuỗi các giá trị số dựa trên tần suất của chúng, có thể làm mô hình quá khớp dữ liệu.

Câu 18. *[Nhiều đáp án đúng] (G2)* Kiểm định chéo (cross-validation):

A. Huấn luyện mô hình nhanh hơn so với phương pháp chia dữ liệu thành hai tập train/test (hay còn gọi là hold-out validation).

B. Là một phương pháp để tìm siêu tham số phù hợp.

C. Đo lường một cách tổng quát hiệu suất của mô hình.

D. Thường được sử dụng với các bộ dữ liệu nhỏ.

Câu 19. *[Nhiều đáp án đúng] (G2)* Các biện pháp để hạn chế overfitting cho phương pháp học dựa trên cây quyết định:

A. Giới hạn độ sâu của cây (max depth).

B. Sử dụng phương pháp học kết hợp (ensemble learning) với nhiều cây quyết định.

C. Áp dụng phương pháp cắt tỉa (pruning).

D. Giảm số lượng tham số của mô hình.

Câu 20. *[Nhiều đáp án đúng] (G2)* Lựa chọn đặc trưng là:

A. Tăng cường đặc trưng cho bài toán.

B. Tăng cường số mẫu cho bài toán.

C. Lựa chọn các đặc trưng tương quan tuyến tính với các đặc trưng khác để loại bỏ.

D. Tìm các đặc trưng không liên quan đến biến dự đoán (output) để loại bỏ.

Câu 21. *[Một đáp án đúng] (G2)* Xét phương pháp K-Nearest Neighbors (`KNNImputer` với các tham số mặc định trong `sklearn`), phát biểu nào sau đây đúng:

A. Giá trị bị thiếu của mỗi đặc trưng được dự đoán từ K điểm dữ liệu gần nhất.

B. Giá trị dự đoán được tính theo trọng số là khoảng cách đến điểm xa nhất.

C. Tốc độ thực thi nhanh và tối ưu chi phí tính toán nhất trong các phương pháp thay thế giá trị bị thiếu.

D. Giá trị bị thiếu của mỗi đặc trưng được gán bằng -0.1.

Câu 22. *[Một đáp án đúng] (G2)* Trong bài toán phân loại bệnh nhân, cần xác định một người có bệnh hay không từ hồ sơ bệnh án của những người từ 8 tuổi đến 100 tuổi. Đặc trưng tuổi trong bài toán nên được:

A. Mã hóa theo One-hot Encoding.
B. Mã hóa theo Label Encoding.
C. Mã hóa theo Ordinal Encoding.
D. Giữ nguyên.

Câu 23. *[Một đáp án đúng] (G2)* Khi nào việc sử dụng giá trị trung vị để thay thế dữ liệu bị thiếu sẽ **có lợi hơn** so với việc sử dụng giá trị trung bình. Tại sao?

A. Khi phần trăm dữ liệu bị thiếu rất thấp, vì giá trị trung vị có nhiều thông tin hơn.

B. Khi dữ liệu thuộc kiểu phân loại, vì giá trị trung vị có thể xử lý dữ liệu không phải dạng số tốt hơn.

C. Khi dữ liệu tuân theo phân phối chuẩn, vì giá trị trung vị thể hiện xu hướng trung tâm tốt hơn.

D. Khi dữ liệu chứa nhiều giá trị ngoại lai, vì giá trị trung vị ít nhạy cảm với các giá trị ngoại lai hơn.

Câu 24. *[Một đáp án đúng] (G2)* Mô hình nào dưới đây nhạy cảm với các giá trị ngoại lai (outlier) và tại sao?

A. Hồi quy tuyến tính, vì các giá trị ngoại lai có thể làm lệch đáng kể trọng số của các đặc trưng.

B. Các thuật toán ở các đáp án còn lại đều bị ảnh hưởng bởi giá trị ngoại lai.

C. Naive Bayes, vì nó dựa trên các giả định xác suất không ổn định trước các giá trị ngoại lai.

D. Cây quyết định, vì chúng chia dữ liệu dựa trên các giá trị ngưỡng.

Câu 25. *[Một đáp án đúng] (G2)* `Normalizer` và `StandardScaler` trong `sklearn` thường được sử dụng thay thế cho nhau, nhưng chúng phục vụ các mục đích khác nhau. Khi nào nên ưu tiên sử dụng `StandardScaler` (chuẩn hóa về giá trị trung bình bằng 0 và phương sai bằng 1) hơn là `Normalizer` (chuẩn hóa các mẫu về véctơ đơn vị)?

A. Khi thuật toán giả định dữ liệu phân phối chuẩn.

B. Khi sử dụng các thuật toán không nhạy cảm với đặc trưng nhiễu (ví dụ như Decision Tree).

C. Khi dữ liệu cần được biến đổi thành không gian phi tuyến tính.

D. Khi dữ liệu có phân phối lệch.

II. CÂU HỎI TỰ LUẬN (5 điểm)

Câu 1. (0.5 điểm) *(G2, G3)* Một ngân hàng có dữ liệu bao gồm 100 biến đầu vào và một biến mục tiêu (biến cần dự đoán). Sau khi áp dụng phương pháp chọn lọc đặc trưng, ngân hàng xác định được 20 biến đầu vào có ý nghĩa thống kê cao. Liệu ngân hàng có nên đưa 20 biến này vào mô hình thực tế hay không? Hãy nêu ngắn gọn lý do cho quyết định của bạn.

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Câu 2. (0.5 điểm) *(G2, G3)* Một công ty có dữ liệu chuỗi thời gian (time series) về doanh số hàng tháng từ tháng 1/2018 đến tháng 12/2022. Công ty muốn xây dựng một mô hình dự đoán doanh số của tháng kế tiếp dựa trên dữ liệu lịch sử. Theo bạn, cách chia dữ liệu thành tập train, dev và test nào để đánh giá đúng tính ổn định của mô hình (model stability)? Hãy mô tả cụ thể khoảng thời gian của mỗi tập và giải thích lý do.

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Câu 3. (2 điểm) *(G2, G3)* Bạn được yêu cầu xây dựng mô hình hồi quy cho dữ liệu sau, dự đoán giá trị của `y_test` từ các đặc trưng trong `X_test`. Bộ dữ liệu huấn luyện và kiểm tra được cung cấp như sau:

- `X_train` (61, 122): Mảng với 61 mẫu dữ liệu và 122 đặc trưng (features).
- `y_train` (61,): Mảng với 61 giá trị mục tiêu (output).
- `X_test` (61, 122): Mảng với 61 mẫu dữ liệu kiểm tra.
- `y_test` (61,): Dữ liệu mục tiêu kiểm tra.

Trong bộ dữ liệu này, không phải tất cả các cột dữ liệu đều cần thiết để dự đoán giá trị mục tiêu. Một số cột có thể không đóng góp nhiều vào việc dự đoán và có thể gây nhiễu (ví dụ: cột chứa thông tin không liên quan trực tiếp đến mục tiêu). Kết quả dự đoán cuối cùng được xây dựng dựa trên các đặc trưng đã được lựa chọn phù hợp nhằm tối ưu hóa chỉ số RMSE (Root Mean Squared Error). Yêu cầu và ràng buộc:

- Bạn chỉ được sử dụng một mô hình duy nhất và không được khai báo thêm thư viện.
- Huấn luyện mô hình sao cho đạt được độ chính xác tốt nhất có thể. Tính và in ra RMSE trên tập kiểm tra.
- Được phép sử dụng GridSearch nhưng chỉ được phép cho một siêu tham số duy nhất.
- In ra chỉ số của các cột chắc chắn đã bị loại bỏ bởi mô hình sau khi huấn luyện.

Hãy hoàn thành khối lệnh *### BEGIN SOLUTION ... ### END SOLUTION* trong đoạn mã sau đây:

```
import numpy as np
import pandas as pd
from sklearn.metrics import mean_squared_error
from sklearn.svm import SVR
from sklearn.linear_model import RandomForestRegressor
from sklearn.linear_model import Ridge, Lasso, LinearRegression
from sklearn.model_selection import GridSearchCV
# Đọc tập dữ liệu huấn luyện
train_data = np.load('train/train_data.npz')
X_train = train_data['X_train'] # (61, 122)
y_train = train_data['y_train'] # (61,)
# Đọc tập dữ liệu kiểm tra
X_test = np.load('test/X_test.npz')['X_test'] # (61, 122)
y_test = pd.read_json("test/y_test.json", lines=True).squeeze() # (61,)
### BEGIN SOLUTION
# Huấn luyện mô hình
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
.....
```

```
.....
# In ra các cột chắc chắn bị loại bỏ bởi mô hình
.....
.....
.....
.....
.....
# Tính và in ra RMSE trên tập kiểm tra
.....
.....
.....
.....
.....
### END SOLUTION
```

Câu 4. (2 điểm) *(G2, G3)* Bạn được yêu cầu xây dựng một mô hình dự đoán mức độ nghiện Internet (PCIAT) của trẻ em và thanh thiếu niên, dựa trên dữ liệu hoạt động thể chất. Biết rằng, mỗi trẻ em hoặc thanh thiếu niên được thu thập dữ liệu từ các thiết bị đo lường và bảng khảo sát. Thông tin input bao gồm hai tập dữ liệu sau:

1. Dữ liệu bảng (Tabular Data):

- Mã người tham gia (`int`);
- Thông tin nhân khẩu học: tuổi (`int`), giới tính (`string`);
- Điểm đánh giá toàn cầu trẻ em (`float`);
- Các chỉ số thể chất: chiều cao (`float`), cân nặng (`float`), huyết áp (`float`);
- Điểm FitnessGram: các bài kiểm tra sức bền, linh hoạt và sức mạnh (`List[float]`);
- Thang đo rối loạn giấc ngủ (`float`);
- Điểm kiểm tra mức độ nghiện Internet (PCIAT): (`int`, từ 0 đến 3: 0 - không nghiêm trọng, ..., 3 - nghiêm trọng).

2. Dữ liệu cảm biến (Actigraphy Data):

- Mã người tham gia (`int`);
- Dữ liệu gia tốc 3 trục (`X, Y, Z: List[float]`);
- Giá trị ENMO (Euclidean Norm Minus One) (`float`);
- Giá trị góc Z (`float`);
- Cờ báo không đeo thiết bị (`int`, 0: đang đeo, 1: không đeo);
- Độ sáng môi trường (`float`);
- Điện áp pin (`float`);

- Thời gian và ngày thu thập dữ liệu (`string`);
- Thứ trong tuần (`int`, 1: Thứ Hai, ..., 7: Chủ Nhật).

Mỗi người tham gia có thể có nhiều bản ghi dữ liệu cảm biến, vì dữ liệu cảm biến là dữ liệu chuỗi (time-series data). Do đó, mỗi người tham gia có thể có một hoặc nhiều dòng dữ liệu cảm biến, với mỗi dòng tương ứng với một thời điểm hoặc một khoảng thời gian cụ thể trong quá trình thu thập dữ liệu. Kết quả của mô hình dự đoán được sẽ giao động rất nhiều, do đó phương sai (variance) của độ chính xác trên tập kiểm thử khá cao. Hãy trình bày quy trình xây dựng mô hình dự đoán mức độ nghiện Internet, bao gồm:

- Phương pháp thu thập và đánh giá dữ liệu (0.5 điểm);
- Phương pháp tiền xử lý dữ liệu (0.5 điểm);
- Thiết kế cặp (đặc trưng input, giá trị output mong muốn) và mô hình máy học phù hợp để giải quyết (có giải thích) (0.5 điểm);
- Thiết kế độ đo, quy trình đánh giá hiệu suất mô hình khách quan (0.5 điểm).

Lưu ý: Cần chú trọng chứng minh tính hợp lý của mô hình. Tự do sử dụng phần cứng máy tính.

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

–HẾT–

Bảng chuẩn đầu ra môn học Lập trình Python cho Máy học:

CĐRMH	Mô tả CĐRMH
G1	Làm việc ở mức độ cá nhân và cộng tác nhóm để trình bày và giải quyết một số thuật toán học không giám sát và có giám sát.
G2	Hiểu và giải thích được các khái niệm, thuật ngữ liên quan tới các quy trình xây dựng mô hình máy học, một số phương pháp phân tích, tiền xử lý dữ liệu, một số mô hình máy học có giám sát, không giám sát, đánh giá mô hình.
G3	Ứng dụng các lý thuyết, mô hình và thuật toán học có giám sát và không giám sát vào giải quyết các bài toán trong thực tế.

Duyệt đề Khoa/Bộ Môn **Giảng viên ra đề**