Đề thi:

R PROGRAMMING LANGUAGE FOR DATA SCIENCE

Thời gian làm bài : từ khi nhận đề đến 20h00, Chủ Nhật ngày 01/08/2021

Đọc kỹ các thông tin dưới đây trước khi làm bài :

- Đổi tên folder **De_thi_k267_HV** được cung cấp thành **LDS7_HoVaTen_Cuoi_Ky** với HoVaTen thay bằng họ và tên của HV rồi lần lượt làm các câu vào trong folder này. Upload folder này vào folder **LDS7_ONLINE_HoVaTen** đã share trên Google Drive, lưu tất cả bài làm vào để GV chấm điểm.
- Đến deadline, HV gửi mail cho giáo viên kèm link của folder **LDS7_HoVaTen_Cuoi_Ky**, HV không gửi bài thi sẽ không có điểm thi.
- HV được sử dụng tài liệu.
- HV sẽ bị trừ điểm nếu bài làm giống nhau.

Chú ý, với mỗi câu:

- Lần lượt thực hiện các bước làm bài như đã được hướng dẫn trong lớp.
- Mỗi câu là 1 file, các yêu cầu nhận xét kết quả trong từng câu được viết trong cell dưới định dạng Markdown.

1. Sales of shampoo over a three year – Time series Analysis (1.0 điểm)

- *Tạo tập tin: **question_1.ipynb** (toàn bộ code của câu 1 sẽ được viết trong file này)*
- Cho dữ liệu **sales-of-shampoo-over-a-three-year.csv** chứa thông tin bán shampoo trong 3 năm từ 2016 đến 2018
- Yêu cầu:
    1. Đọc dữ liệu
    2. Xem thông tin chung từ dữ liệu: head(), số dòng, số cột, str()…
    3. Chuyển dữ liệu này thành Time Series object => in Time Series object.
    4. Vẽ Time Series object vừa tạo.
    5. Thực hiện việc decompostition, nhận xét.
    6. Thực hiện việc dự báo và vẽ biểu đồ so sánh với thực tiễn.
    7. Dự đoán lượng sales cho 6 tháng tiếp theo.

2. Normal – Binormial Distribution (0.5 điểm)

- *Tạo tập tin: **question_2.ipynb** (toàn bộ code của câu 2 sẽ được viết trong file này)*
- Thực hiện các yêu cầu sau :
    1. Giả sử chỉ số IQ thường được phân phối với giá trị trung bình là 100 và độ lệch chuẩn là 15.
        a. Vậy tỷ lệ bao nhiêu phần trăm người có IQ nhỏ hơn 125 ?
        b. Vậy tỷ lệ bao nhiêu phần trăm người có IQ lớn hơn 110 ?
        c. Vậy tỷ lệ bao nhiêu phần trăm người có IQ trong khoảng từ 110 và 125 ?
    2. Xúc xắc có 6 mặt :
        a. Tìm xác suất để có được 2 lần mặt 4 nút trong 5 lần đổ xúc xắc.
        b. Có bao nhiêu mặt 4 nút khi có xác suất 25% xuất hiện khi một xúc xắc được đổ 50 lần?

3. Marketing (1.5 điểm)

- *Tạo tập tin: **question_3.ipynb** (toàn bộ code của câu 3 sẽ được viết trong file này)*
- Cho dữ liệu trong tập tin **marketing.csv**
- Thực hiện các yêu cầu sau:

1. Yêu cầu 1: Sử dụng Linear Regression để thực hiện việc **dự đoán sales** dựa trên thuộc tính **youtube**.

Gợi ý các bước thực hiện:

a. Đọc dữ liệu
b. Xem thông tin chung từ dữ liệu: head(), số dòng, số cột, str(), summary()
c. Vẽ biểu đồ quan sát mối liên hệ giữa sales và youtube
d. Tiền xử lý dữ liệu
e. Kiểm tra và xử lý outliers
f. Tạo train:test từ dữ liệu data với tỉ lệ 70:30 hoặc 80:20
g. Thực hiện Linenear Regression với train data.
h. In summary của model
i. Dự đoán y_test_predict từ test data => so sánh y_test_pred với y_test
j. Tính Mean Square Error (mse), r^2 cho train, r^2 cho test. Nhận xét.
k. Tìm Coefficients, Intercept
l. Cho youtube lần lượt: x <- c(100, 200, 300) => dự đoán sales.
m. Trực quan hóa kết quả.

2. Yêu cầu 2: Sử dụng Linear Regression để **dự đoán sales** dựa trên các thuộc tính (youtube, facebook, newspaper) do học viên tự lựa chọn (chọn 2 hoặc 3 thuộc tính).

Gợi ý các bước thực hiện: tương tự như yêu cầu 1, không có phần yêu cầu dự đoán mới.

4. Mushroom (1.0 điểm)

- *Tạo tập tin: **question_4.ipynb** (toàn bộ code của câu 4 sẽ được viết trong file này)*
- Cho dữ liệu mushroom trong tập tin **mushrooms.csv** chứa thông tin của các mẫu nấm, nấm ăn được và không ăn được.
  - Dữ liệu có thể tham khảo và download tại: https://www.kaggle.com/jnduli/decision-tree-classifier-for-mushroom-dataset/data

Data Infromation : Bộ dữ liệu chứa 23 thuộc tính. Thuộc tính **"class"** là class attribute (output).

- **class: edible=e, poisonous=p**
- cap-shape: bell=b, conical=c, convex=x, flat=f, knobbed=k, sunken=s
- cap-surface: fibrous=f, grooves=g, scaly=y, smooth=s
- cap-color: brown=n, buff=b, cinnamon=c, gray=g, green=r, pink=p, purple=u, red=e, white=w, yellow=y
- bruises: bruises=t, no=f
- odor: almond=a, anise=l, creosote=c, fishy=y, foul=f, musty=m, none=n, pungent=p, spicy=s
- gill-attachment: attached=a, descending=d, free=f, notched=n
- gill-spacing: close=c, crowded=w, distant=d

- gill-size: broad=b,narrow=n
- gill-color: black=k, brown=n, buff=b, chocolate=h, gray=g, green=r, orange=o, pink=p, purple=u, red=e, white=w, yellow=y
- stalk-shape: enlarging=e, tapering=t
- stalk-root: bulbous=b, club=c, cup=u, equal=e, rhizomorphs=z, rooted=r, missing=?
- stalk-surface-above-ring: fibrous=f, scaly=y, silky=k, smooth=s
- stalk-surface-below-ring: fibrous=f, scaly=y, silky=k, smooth=s
- stalk-color-above-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
- stalk-color-below-ring: brown=n, buff=b, cinnamon=c, gray=g, orange=o, pink=p, red=e, white=w, yellow=y
- veil-type: partial=p,universal=u
- veil-color: brown=n, orange=o, white=w, yellow=y
- ring-number: none=n, one=o,two=t
- ring-type: cobwebby=c, evanescent=e, flaring=f, large=l, none=n, pendant=p, sheathing=s, zone=z
- spore-print-color: black=k, brown=n, buff=b, chocolate=h, green=r, orange=o, purple=u, white=w, yellow=y
- population: abundant=a, clustered=c, numerous=n, scattered=s, several=v, solitary=y
- habitat: grasses=g, leaves=l, meadows=m, paths=p, urban=u, waste=w, woods=d

- Yêu cầu: Sử dụng **cả Logistic Regression và Decision Tree** để thực hiện việc xác định một mẫu nấm là **nấm ăn được** hay **nấm độc** dựa vào các thông tin còn lại. Trong hai thuật toán trên thì thuật toán nào phù hợp hơn cho bộ dữ liệu này? Vì sao?
- Gợi ý các bước thực hiện cho từng thuật toán :
  1. Đọc dữ liệu.
  2. Xem thông tin dữ liệu: head(), số dòng, số cột, summary()…
  3. Tiền xử lý dữ liệu (nếu cần).
  4. Tạo train và test từ dữ liệu.
  5. Xây dựng model với train.
  6. In summary của model.
  7. Dự đoán y_pred từ test => so sánh với y_test.
  8. Đánh giá model.
  9. Trực quan hóa model.

5. Ageinc (1.0 điểm)

- *Tạo tập tin: **question_5.ipynb** (toàn bộ code của câu 5 sẽ được viết trong file này)*
- Cho dữ liệu **ageinc_g.csv** chứa thông tin 1000 khách hàng gồm : income, age, gender
- Yêu cầu: Đọc dữ liệu, chuẩn hóa dữ liệu (nếu cần) và sử dụng KMeans để thực hiện việc **phân cụm** dữ liệu dựa trên hai cột là **income** và **age**.
- Gợi ý các bước thực hiện:
  1. Đọc dữ liệu.
  2. Xem thông tin data: head(), số dòng, số cột, summary().

3. Tiền xử lý dữ liệu (nếu cần).
4. Vẽ hình để xem xét mối liên hệ giữa các thuộc tính. Cho nhận xét dựa trên biểu đồ.
5. Xây dựng model từ dữ liệu income và age.
6. Tìm kết quả => có bao nhiêu cụm => mẫu nào thuộc cụm nào?
7. Vẽ hình (với mỗi cụm là một màu) => xem kết quả.
8. Đưa ra một số nhận xét dựa trên kết quả.

--- ☺ Chúc các bạn làm bài tốt ☺ ---