Đề thi:

R PROGRAMMING LANGUAGE FOR DATA SCIENCE

Thời gian làm bài: từ khi nhận đề đến 23h30, Chủ Nhật ngày 17/09/2023

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

- HV tạo một folder là **LDS7_K287_HoVaTen_Cuoi_Ky** (nằm trong folder **LDS7_K287_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_K287_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 làm demo/ bài tập trong lớp.
- Tiền xử lý dữ liệu (nếu cần)
- Mỗi câu là 1 file, các yêu cầu nhận xét kết quả trong từng câu được viết trong cell dưới định dạng Markdown.

1. Cubic_zirconia (1.0 điểm)

- *Tạo tập tin: question_1.ipynb (toàn bộ code của câu 1 sẽ được viết trong file này)*
- Cho dữ liệu cubic_zirconia.csv chứa giá và các thuộc tính khác của gần 27.000 mẫu đá zirconia (là một loại kim cương giá rẻ với nhiều đặc điểm giống như kim cương). Bao gồm:

Variable Name	Description
carat	Carat weight of the cubic zirconia.
cut	Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
color	Colour of the cubic zirconia.With D being the best and J the worst.
clarity	Cubic zirconia Clarity refers to the absence of the Inclusions and Blemishes. (In order from Best to Worst, FL = flawless, I3= level 3 inclusions) FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1, I2, I3
depth	The Height of a cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.
table	The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
price	The Price of the cubic zirconia.
x	Length of the cubic zirconia in mm.
y	Width of the cubic zirconia in mm.
z	Height of the cubic zirconia in mm.

- Yêu cầu:
  1. Đọc dữ liệu cubic_zirconia.csv và đưa vào dataframe zirconia. Xem thông tin dữ liệu với head(), tail(), str(), summary().
  2. Tạo zirconia_sub từ zirconia chỉ chứa các cột: 'carat', 'cut', 'color', 'clarity', 'depth', 'table', 'price'. Cho biết dữ liệu có bao nhiêu dòng, cột?

3. Trong zirconia_sub dữ liệu thiếu/ null không? Có dữ liệu trùng không ? Nếu có, hãy xóa bỏ tất cả các dòng có chứa dữ liệu thiếu, dữ liệu trùng. Cho biết dữ liệu lúc này còn bao nhiêu dòng? In thống kê chung.

Từ câu 4. trở đi, sử dụng dữ liệu zirconia_sub:

4. Vẽ biểu đồ phân phối tần suất của 'price'. Nhận xét.
5. Thực hiện các thống kê cơ bản cho 'depth', 'table', 'price' (mean, median, mode, max, min, range).
6. Cho biết các giá trị ở phân vị thứ 5%, 30%, 60% và 95% của 'price'. Biểu diễn phân vị và giá trị tương ứng trên biểu đồ.
7. Vẽ boxplot cho 'price', 'carat'. 'price', 'carat' có outlier hay không? Nếu có thì mỗi thuộc tính có bao nhiêu outlier? (Biết outlier dưới < Q1 – 1.5 * IQR và outlier trên > Q3 + 1.5 * IQR)
8. Vẽ pie chart thống kê mẫu đá theo từng cột 'cut', 'color'. Nhận xét.
9. Tính phương sai, độ lệch chuẩn, skewness và kurtosis của tất cả các thuộc tính số trong zirconia_sub. Nhận xét cho từng thuộc tính.
10. Tính giá trị covariance, correlation giữa 'price' và 'carat'. Nhận xét.
11. Vẽ biểu đồ thể hiện mối quan hệ giữa 'carat', 'price' theo 'color'. Nhận xét biểu đồ.
12. Cho biết số lượng mẫu có giá trị 'price'>=10000 USD, xác suất để một viên đá có 'price'>=10000 USD là bao nhiêu?
13. Xác suất để một viên đá có color = 'D' và cut = 'Premium' là bao nhiêu?

2. Giá gạo Việt Nam xuất khẩu (1.0 điểm)

- *Tạo tập tin: question_2.ipynb (toàn bộ code của câu 2 sẽ được viết trong file này)*
- Cho dữ liệu **Export_rice_prices_5percent_broken_vn.csv**, cung cấp giá gạo (5% tấm) xuất khẩu của Việt Nam từ tháng 01/2004 đến tháng 09/2022
- Yêu cầu:
  1. Đọc dữ liệu từ tập tin.
  2. In một số thông tin chung từ dữ liệu: head(), 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, trực quan và nhận xét.
  6. Thực hiện việc dự báo và vẽ biểu đồ so sánh với thực tiễn. Nhận xét.
  7. Dự đoán giá gạo cho 12 tháng tiếp theo.

3. Marketing (1.0 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. In 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.
  - Hoặc bạn 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).

Attribute Information:

- **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 và đưa vào dataframe data.
2. In 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 data.
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. In thông tin dữ liệu: 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 ☺ ---