Dữ liệu khảo sát gồm 18 thông số đo môi trường và 1 chỉ số WQI

Bước 1: Clean dữ liệu: làm sao em clean dữ liệu em ghi vài dòng và cho kết quả giúp chị nhé.

Thuận:

Dữ liệu mẫu đã khá sạch, đúng format, nên công việc clean dữ liệu sẽ đơn giản. Một số công việc sẽ làm:

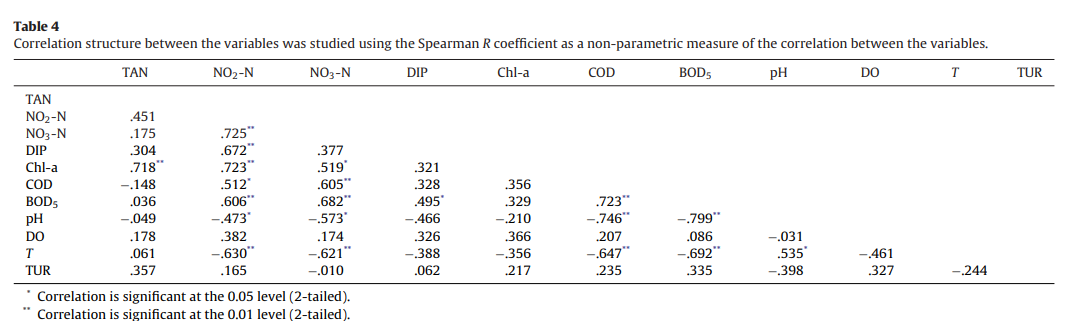
* Xác định các giá trị 0 (trong cột nitrat, nitrite, thay nước,...) là dữ liệu có giá trị = 0, hay không có dữ liệu?
* Xử lý dữ liệu cột “size”, xử lý tùy thuộc vào đặc tính cột này, có thể điền đầy (giá trị mean, median, giá trị gần nhất, nội suy,...), hoặc dữ nguyên vào thêm cột trạng thái.
* Cân nhắc loại bỏ cột “Số con” nếu không chứa nhiều thông tin.

Bước 2: Kiểm tra phân phối của các trường dữ liệu bằng Kolmogorov-Smirnov hay Shapiro-Wilk

* Cho chị kết quả distribution của từng thông và kết luận là phân phối chuẩn hay phân phối không chuẩn

Bước 3: Xác định mối tương quan của các thông số

* Cho chị kết quả xác định mối tương quan của các thông số (hình như em hay xuất) và bảng kết quả như hình dưới



Theo tài liệu lý thuyết thống kê thì các biến có phân phối chuẩn thì chạy pearson và phân phối không chuẩn chạy spearson. Thành ra, em chạy cả 2 case cho tiện hen.

Thuận:

Bước 2 và bước 3 là phần khảo sát dữ liệu:

* Sử dụng các thư viện để kiểm tra kstest và correlation của các cột.
* Có thể khai thác thêm nhiều thông tin trong quá trình khai phá dữ liệu: biểu đồ distribution của từng cột (các cột quan trọng), giá trị nhỏ nhất, giá trị lớn nhất, các giá trị bất thường, các dữ liệu lặp,...

Tiền xử lý dữ liệu:

* Xử lý các dữ liệu dạng phân loại (categorical) thành dạng số,
* Thường hóa(normalize) hoặc chuẩn hóa (standardize) dữ liệu
* Xử lý dữ liệu thời gian (ngày/tháng/năm)
* ...
* So sánh các phương pháp tiền xử lý bằng kết quả sau khi chạy mô hình để lựa chọn phương pháp tối ưu nhất

Bước 4: chạy mô hình

Case 1: input là pH, độ đục, Amoni, Nitrit và Ca ra output là WQI

Case 2: theo như cách em chọn thường ngày để đưa ra mô hình.

Case 3: mô hình dự báo sớm cho ngày kế tiếp và 1 tuần sau đó. Thông số cần dự báo là Nhiệt độ, độ mặn, DO, kiềm, input thì theo em khảo sát được

Nhưng chị cần kết quả khảo sát của hidden layer, neurons & epochs của từng case theo kiểu

|  |  |  |
| --- | --- | --- |
| Hidden layer | RMSE | R2 |
| 2 |  |  |
| 3 |  |  |
| 4 |  |  |

* Số layer được chọn

Tương tự cho neurons và epochs.

Xây dựng mô hình cho các case:

* Có thể tạo nhiều model để so sánh kết quả, bắt đầu từ các model baseline dùng các thuật toán học máy cổ điển
* Case 2: có thể chọn 1 số các cột có correlation lớn với output để làm input cho model. Hoặc có thể tạo model với tất cả các thuộc tính. Sau đó so sánh kết quả, lựa chọn
* Đối với mô hình dự đoán. Xác định xem nhiệt độ có bị ảnh hưởng trực tiếp từ nhiệt độ môi trường hay không? Nếu có, có thể sử dụng các dữ liệu dự báo môi trường để tham khảo
* Ngoài ra, em sẽ tham khảo thêm các bài báo liên quan cũng lĩnh vực để tham khảo các model của họ.