Dữ liệu:

* Em ưu tiên dùng dữ liệu 1 lần / ngày, do dữ liệu đồng bộ hơn, đầy đủ cho mỗi điểm dữ liệu, và dễ xử lý dữ liệu hơn.

Về mô hình:

* Em phát triển trước mô hình dự đoán Độ kiềm dựa trên các input đầu vào. (Chưa phải là mô hình dự báo ngày tiếp theo)

Xử lý dữ liệu, từ dữ liệu thô ban đầu, sẽ đi qua các bước xử lý sau:

* Làm sach dữ liệu, xử lý các dữ liệu khuyết, dữ liệu lỗi, xử lý các ký tự tiếng Việt(đổi tên)
* Lược bỏ các dữ liệu sẽ không dùng trong mô hình
* Gom nhóm dữ liệu theo từ ao vụ.
* Mã hóa dữ liệu dạng phân loại (non-numeric), dùng mã hóa dạng OnehotEncoder, ví dụ từ Season chứa các giá trị là ‘Dry’ và ‘Rainy’ sẽ được mã hóa thành 2 cột ‘Season\_Dry’ và cột ‘Season\_rainy’ chứa giá trị 0 và 1
* Sau khi gom nhóm theo từng vụ ao, đối với mỗi ao riêng biệt, kết hợp các thông số vật lý của ngày hiện tại và độ kiềm của ngày hôm trước làm input để dự đoán độ kiềm của ngày hiện tại.
* Tách dữ liệu ngẫu nhiên thành 2 phần, phần train và phần test.
* Scale dữ liệu về scale chuẩn (chuẩn hóa dữ liệu)

Mô hình:

Output: Độ kiềm.

Input:

* Thông số cố định: season, loại ao, công nghệ, giống, ngày thả, diện tích ao, tuổi tôm
* Thông số dễ đo: Nhiệt độ, pH, Độ mặn, TDS, Độ đục, DO, mực nước, độ trong
* Thông số phòng thí nghiệp: độ màu, độ kiềm ngày hôm trước

Mô hình RandomForest (Classic machine learning model):

* Với các thông số được tuning dựa trên bộ:
  + n\_estimators': [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000]
  + 'max\_features': ['auto', 'sqrt']
  + 'max\_depth': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None]
  + 'min\_samples\_split': [2, 5, 10]
  + 'min\_samples\_leaf': [1, 2, 4]

Tìm được bộ thông số cho kết quả tốt nhất: 'n\_estimators': 200, 'min\_samples\_split': 10, 'min\_samples\_leaf': 5, 'max\_features': 10, 'max\_depth': 400,

A graph with blue dots

Description automatically generated

Mô hình ANN, với các kiểm nghiệm như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | RMSE | MAE | MAPE | R2 |
| 1 Hidden layer optimizer = adam | 21.519 | 13.613 | 8.69% | 0.466 |
| 2 Hidden layer optimizer = adam | 21.852 | 13.913 | 8.99% | 0.449 |
| 3 Hidden layer optimizer = adam | 21.239 | 13.475 | 8.71% | 0.48 |
| 1 Hidden layer optimizer = SGD | 21.51 | 14.099 | 9.18% | 0.466 |
| 2 Hidden layer optimizer = SGD | 21.457 | 13.69 | 8.72% | 0.469 |
| 3 Hidden layer optimizer = SGD | 22.215 | 14.301 | 9.24% | 0.435 |
| 1 Hidden layer optimizer = Nadam | 21.593 | 13.639 | 8.73% | 0.462 |
| 2 Hidden layer optimizer = Nadam | 21.794 | 13.865 | 8.77% | 0.452 |
| 3 Hidden layer optimizer = Nadam | 22.045 | 14.008 | 9.05% | 0.439 |