Tình hình lọc dữ liệu

# Sơ đồ lọc dữ liệu

Diagram

Description automatically generated

# 2. Sau khi lọc xong

* Trên tập dữ liệu hợp lệ, 0 là giá trị nhỏ nhất và 10 là giá trị lớn nhất.
* Trên tập dữ liệu trung bình, 0 là giá trị nhỏ nhất và 10 là giá trị lớn nhất.
* Số lượng chất theo ACTIVITY được gắn sẵn.

|  |  |
| --- | --- |
| **ACTIVITY** | **Count of ACTIVITY** |
| Active | 1090 |
| Inactive | 10 |
| Inconclusive | 50 |
| Unspecified | 268 |

* Ngoài ra, đây là bảng đếm số lần trùng lặp của các chất hợp lệ, ví dụ như các chất chỉ xuất hiện 1 lần trong số các chất hợp lệ là 631, 2 lần là 105, …

Chart

Description automatically generated

* Nhận xét
  + Có 4 chất Active và 1 chất Inactive nhưng chỉ số IC50 có dạng như sau:

Table

Description automatically generated

Đề xuất của em là loại bỏ dấu ‘< >’ ở trước giá trị và sử dụng dữ liệu.

* + Phần dữ liệu Unspecified có 246 cột có mô tả IC50 > 10. Em đề xuất sử dụng những chất này với nhãn là ‘Inactive’.

Table

Description automatically generated

* + 50 dữ liệu gắn nhãn ‘Inconclusive’, em muốn xin ý kiến của cô

Table, Excel

Description automatically generated

* + Các chất Inactive nhưng có IC50 bằng 0, theo em hiểu gốc nó không có số liệu, nên em đề nghị lấy IC50 bằng 10 với tất cả các chất này.

Table

Description automatically generated with medium confidence

* + Do tập dữ liệu có gắn nhãn Active so với Inactive là rất lớn nên em đề xuất xây dựng mô hình hồi quy nhằm dự đoán IC50 thì sẽ có giá trị hơn so với sử dụng mô hình phân loại Active hoặc Inactive trong bài báo gốc. Em xin ý kiến cô.