

3. IMPUTER

Mục tiêu IMPUTER: được sử dụng để xử lý các giá trị thiếu trong dữ liệu, tức là các ô dữ liệu mà không có giá trị (có thể là NaN, null, hoặc các giá trị không có ý nghĩa khác). Khi dữ liệu bị thiếu, việc sử dụng Imputer giúp điền các giá trị thay thế vào những vị trí thiếu này để có thể tiếp tục phân tích dữ liệu một cách chính xác.

- [Thực hiện trên hệ thống theo các bước hướng dẫn sau:](#)

Thực hiện trên hệ thống theo các bước hướng dẫn sau:

Bước 1: Vào hệ thống chọn module Explore, chọn tập dữ liệu cần tổng hợp và tạo mới một tập dữ liệu tổng hợp từ IMPUTER

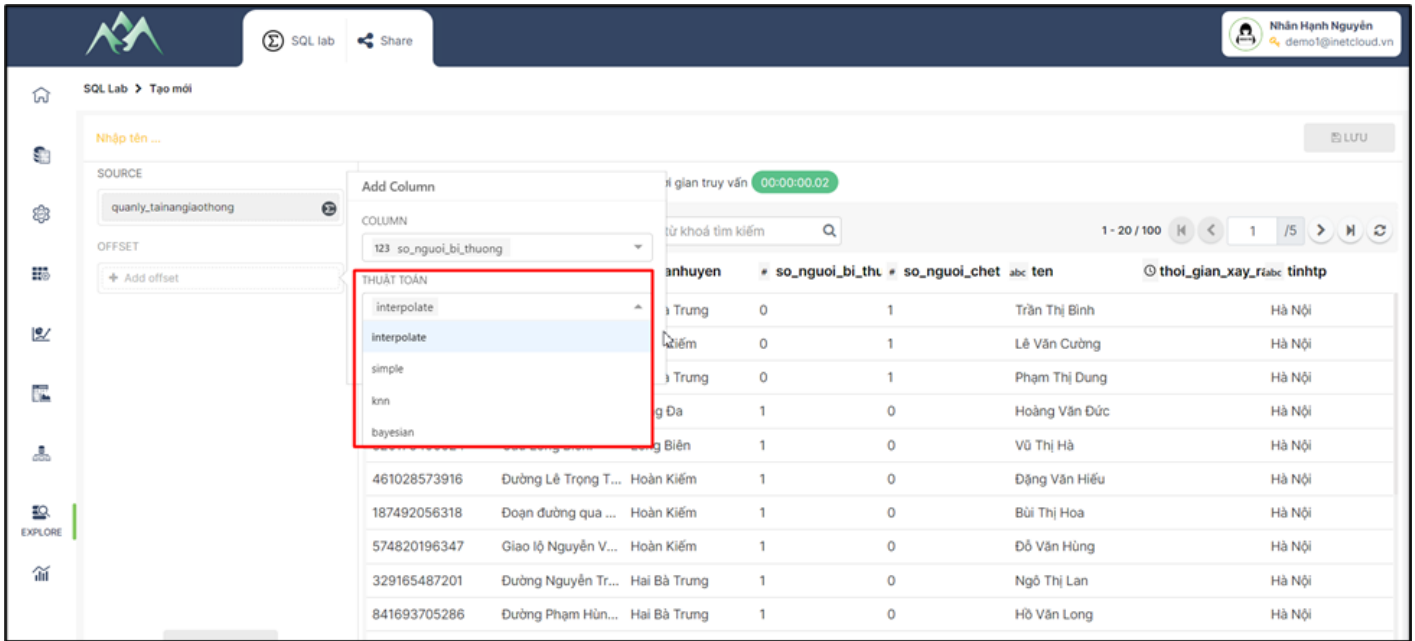
The screenshot shows the SQL Lab interface. On the left, the 'EXPLORE' icon is highlighted in a red box. The main area displays a dataset named 'storagehub.hanh_chinh.quanly_t...' with 8 columns. The 'IMPUTER' tool is selected from the 'MORE TOOLS' dropdown menu, also highlighted in a red box. The data table is as follows:

abc	cccc	abc	địa điểm tại na	abc	quanhuyen	#	so_nguoi_bi_thu	#	so_nguoi_chet	abc	ten	thoi_gian_xay
693156748023			Ngã tư Kim Liên - ...		Hai Bà Trưng	0	1				Trần Thị Bình	
918304582715			Giao lộ Lê Duẩn - ...		Hoàn Kiếm	0	1				Lê Văn Cường	Hà Nội
572639104820			Cầu Thanh Trì.		Hai Bà Trưng	0	1				Phạm Thị Dung	Hà Nội
239714865301			Đường Trần Phú - ...		Đống Đa	1	0				Hoàng Văn Đức	Hà Nội
820173495624			Cầu Long Biên.		Long Biên	1	0				Vũ Thị Hà	Hà Nội
461028573916			Đường Lê Trọng T...		Hoàn Kiếm	1	0				Đặng Văn Hiếu	Hà Nội
187492056318			Đoạn đường qua ...		Hoàn Kiếm	1	0				Bùi Thị Hoa	Hà Nội
574820196347			Giao lộ Nguyễn V...		Hoàn Kiếm	1	0				Đỗ Văn Hùng	Hà Nội
329165487201			Đường Nguyễn Tr...		Hai Bà Trưng	1	0				Ngô Thị Lan	Hà Nội
841693705286			Đường Phạm Hùn...		Hai Bà Trưng	1	0				Hồ Văn Long	Hà Nội
509728436162			Giao lộ Nguyễn Tr...		Hai Bà Trưng	1	0				Dương Thị Mai	Hà Nội
267490158374			Đường Phùng Hư...		Hai Bà Trưng	1	0				Đào Văn Nam	Hà Nội
361985247092			Đoạn đường qua ...		Hoàn Kiếm	1	0				Lý Thị Ngọc	Hà Nội

Bước 2: Chọn cột dữ liệu để thuật toán Imputer giúp điền các giá trị thay thế vào những vị trí thiếu.

The screenshot shows the SQL Lab interface with the 'Add Column' dialog box open. The 'COLUMN' field is set to '123 so_nguoi_bi_thuong', which is highlighted in a red box. The 'Add offset' button is also highlighted in a red box. The data table is the same as in the previous screenshot.

Bước 3: Chọn thuật toán xử lý cho cột dữ liệu đã chọn.



Trong đó:

- o Thuật toán nội suy (interpolation) được sử dụng để ước lượng giá trị tại các điểm trung gian dựa trên các giá trị đã biết của một tập dữ liệu.
- o Thuật toán Simple được dùng để xử lý các bài toán như tìm kiếm tuyến tính, sắp xếp nổi bọt, hoặc tính giai thừa.
- o Thuật toán KNN là một thuật toán dựa trên khoảng cách để phân loại hoặc dự đoán giá trị của một điểm dữ liệu mới dựa trên các điểm dữ liệu đã biết trong không gian đa chiều.
- o Thuật toán Payesian là một trong những thuật toán phân loại phổ biến nhất trong học máy, đặc biệt là đối với các bài toán phân loại văn bản và lọc thư rác.

Bước 4: “RUN” chạy test dữ liệu được tổng hợp.

