**CÁC PHƯƠNG PHÁP, KỸ THUẬT, THUẬT TOÁN SỬ DỤNG**

1. **Kỹ thuật tiền xử lý dữ liệu:**

Tiền xử lý dữ liệu là tập hợp các kỹ thuật được sử dụng trước khi áp dụng vào các mô hình phân cụm. Đây là một trong những vấn đề quan trọng nhất trong quá trình khám phá tri thức từ dữ liệu (Knowledge Discovery from Data - KDD)(cite) . Vì dữ liệu khách hàng thường không hoàn hảo, chứa các bất thường và sự dư thừa, nên không thể áp dụng trực tiếp cho quá trình chạy mô hình.

**1.1. StringIndexer: (cite)**

StringIndexer là một lớp quan trọng trong thư viện pyspark.ml.feature của PySpark, được sử dụng để biến đổi dữ liệu danh mục (categorical data) thành dữ liệu số (numerical data). Việc biến đổi này giúp các thuật toán học máy và mô hình thống kê có thể xử lý và phân tích dữ liệu hiệu quả hơn.

StringIndexer hoạt động bằng cách gán một chỉ số số duy nhất cho mỗi giá trị duy nhất trong biến dữ liệu danh mục. Ví dụ: giả sử ta có một biến gender với giá trị "Male" và "Female". StringIndexer sẽ gán chỉ số 0 cho "Male" và chỉ số 1 cho "Female".

**Lợi ích của StringIndexer:**

**Cải thiện hiệu suất:** Các thuật toán học máy và mô hình thống kê thường hoạt động hiệu quả hơn với dữ liệu số. StringIndexer giúp biến đổi dữ liệu danh mục thành dữ liệu số, từ đó tăng hiệu suất của các thuật toán.

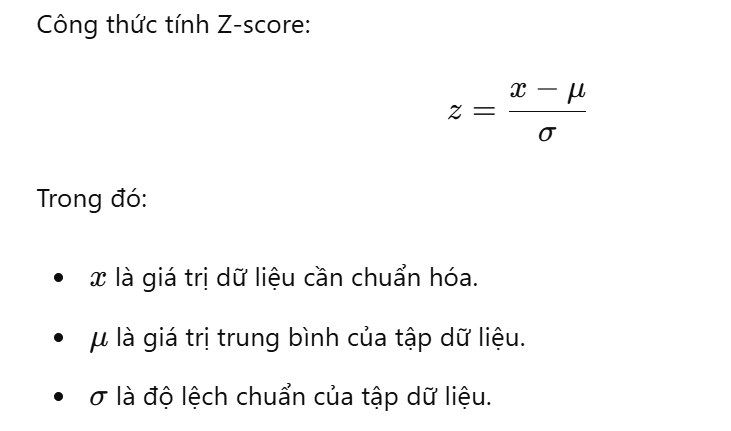
**Đơn giản hóa:** StringIndexer giúp đơn giản hóa quá trình tiền xử lý dữ liệu danh mục, loại bỏ nhu cầu thực hiện các bước xử lý phức tạp khác.

**Tăng tính linh hoạt:** StringIndexer có thể được sử dụng với nhiều loại dữ liệu danh mục khác nhau, bao gồm các biến có nhiều giá trị (multi-class categorical data).

**1.2. Chuẩn hóa bằng Z-score: (cite)**

Chuẩn hóa dữ liệu là một bước quan trọng trong quá trình tiền xử lý dữ liệu. Nó giúp đưa các đặc trưng về cùng một khoảng giá trị để dễ dàng so sánh và phân tích. Có 2 phương pháp chuẩn hóa dữ liệu phổ biến là chuẩn hóa Min-Max và chuẩn hóa Z-core. Trong đồ án này, nhóm thực hiện chuẩn hóa Z-score. Phương pháp này đưa các giá trị dữ liệu về một phân phối chuẩn có trung bình bằng 0 và độ lệch chuẩn bằng 1 bằng cách trừ đi giá trị trung bình của dữ liệu và chia cho độ lệch chuẩn.

Công thức chuẩn hóa Z-core:



Ý nghĩa:

* Giá trị Z-score > 0 cho biết giá trị lớn hơn giá trị trung bình.
* Giá trị Z-score < 0 cho biết giá trị nhỏ hơn giá trị trung bình.
* Giá trị Z-score gần 0 cho biết giá trị gần với giá trị trung bình.

Việc chuẩn hóa dữ liệu giúp loại bỏ ảnh hưởng của tỷ lệ và biến đổi các giá trị dữ liệu về dạng thống nhất, từ đó làm cho việc phân tích và xử lý dữ liệu trở nên dễ dàng hơn, cải thiện hiệu suất của các thuật toán học máy và mô hình thống kê. Z-score chuẩn hóa dữ liệu để đảm bảo rằng mỗi đặc trưng đóng góp tương tự vào mô hình, đặc biệt hữu ích trong các mô hình yêu cầu dữ liệu chuẩn hóa như K-means,