**Thuật toán K-nearest-neighbors:**

**-** Là thuật toán phi số dùng để phân loại và hồi quy .

Một đối tượng được phân lớp dựa vào K láng giềng của nó. K là số nguyên dương được xác định trước khi thực hiện thuật toán. Người ta thường dùng khoảng cách Euclidean để tính khoảng cách giữa các đối tượng.Trong cả 2 trường hợp đầu vào bao gồm có k giá trị gần nhất thì sẽ xét tập không gian chứa k giá trị đó và đầu ra phụ thuộc vào phân loại hay hồi quy :

+ Trong phân loại k-NN, đầu ra là một thành viên lớp. Một đối tượng được phân loại theo phần lớn các nước láng giềng của nó, với đối tượng được gán cho lớp phổ biến nhất trong số những người hàng xóm gần nhất của nó (k là một số nguyên dương, thường là nhỏ). Nếu k = 1, thì đối tượng chỉ được gán cho lớp của người hàng xóm gần nhất.

+ Trong hồi quy k-NN, đầu ra là giá trị thuộc tính của đối tượng. Giá trị này là giá trị trung bình của các giá trị k của hàng xóm gần nhất.

**Thuật toán K-NN được mô tả như sau:**

1.    Xác định giá trị tham số K (số láng giềng gần nhất)

2.    Tính khoảng cách giữa đối tượng cần phân lớp (Query Point) với tất cả các đối tượng trong training data (thường sử dụng khoảng các Euclidean)

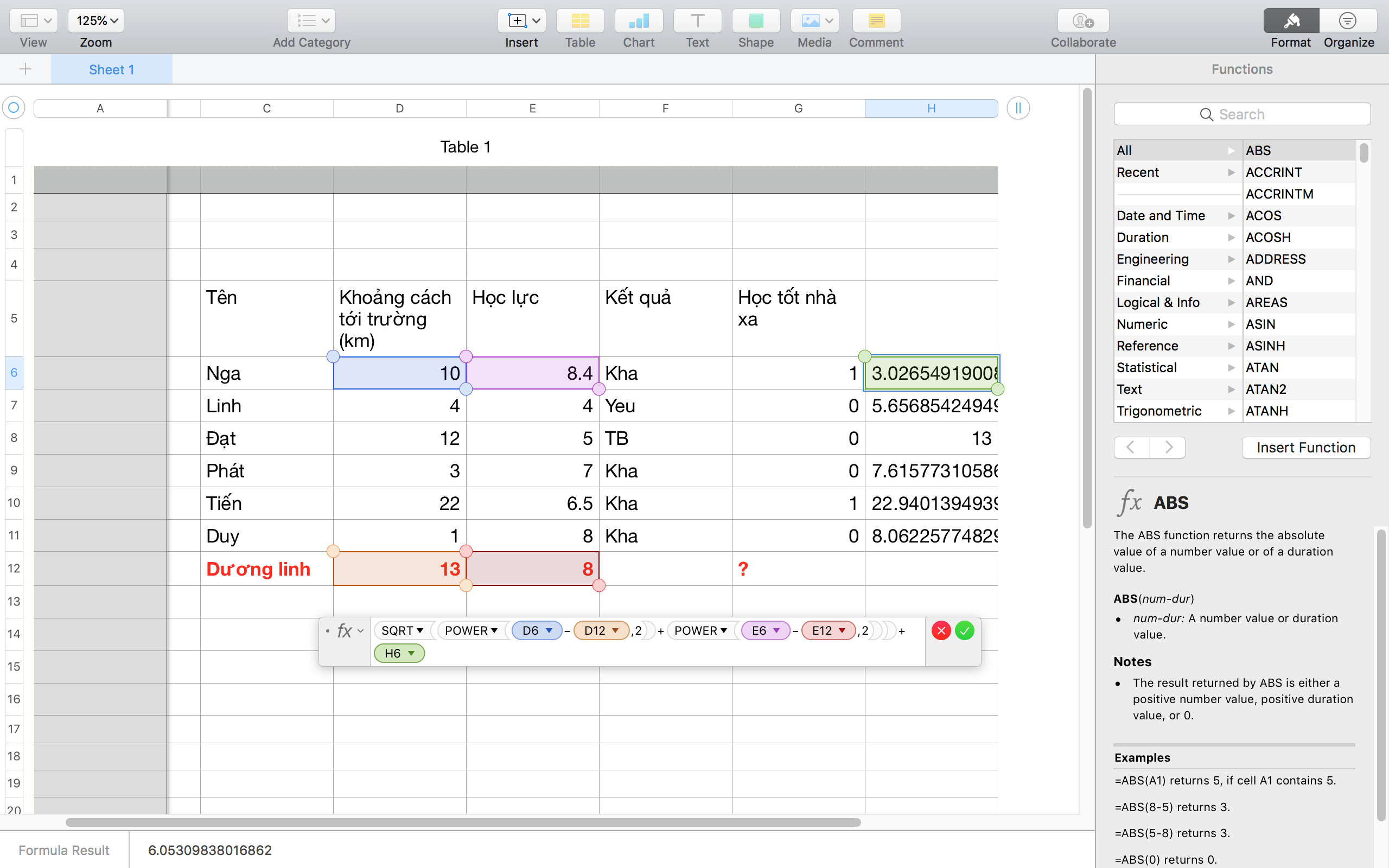
3.    Sắp xếp khoảng cách theo thứ tự tăng dần và xác định K láng giềng gần nhất với Query Point

4.    Lấy tất cả các lớp của K láng giềng gần nhất đã xác định

5.    Dựa vào phần lớn lớp của láng giềng gần nhất để xác định lớp cho Query Point.

Ví dụ: ( hình 1)

- Ta có bảng sau có số liệu về học lực và khoảng cách từ nhà đến trường nên ta dễ dàng tính được học lực và bạn ấy có khó khăn trong học tập hây không khi cho 1 bạn mới là bạn Dương Linh vào thì ta sẽ từ dữ liệu sẵn có tính ra được 1 chỉ số bên cột cuối cùng dùng công thức Euclidean và ta có thể dự đoán khi bạn Nga có chỉ số thấp nhất gần Dương Linh nhất và bạn ấy là người gặp khó khăn trong học tập vậy ta cũng có thể nhận định rằng Dương Linh cũng gặp khó khăn trong học tập.

Hình 1: