Báo Cáo Tiến Độ Đồ Án Một

---oOo---

Giáo viên hướng dẫn: cô Từ Tuyết Hồng.

Sinh viên thực hiện đồ án:

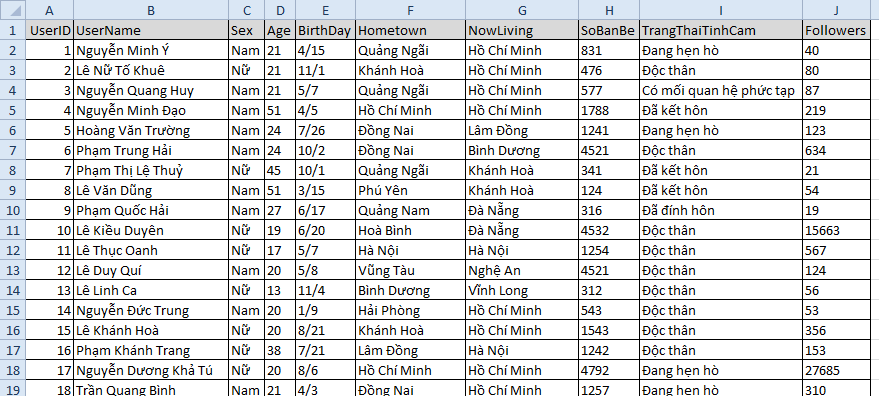
* Nguyễn Minh Ý - MSSV: 15110155 - Lớp: 15110CL3
* Lê Nữ Tố Khuê - MSSV: 15110066 - Lớp: 15110CL3

1. **Mô Tả Đồ Án:**

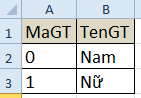
* Đề tài: Viết phần mềm phân cụm dữ liệu Facebook Users Profile.
* Thuật toán được sử dụng: Thuật toán phân cụm K-means
* Ngôn ngữ lập trình được sử dụng cho đồ án: C# (WindowForm C#)
* Hệ quản trị cơ sỡ dữ liệu được sử dụng: Microsoft SQL Server 2016
* Công nghệ sử dụng để load dữ liệu từ SQL Server: ADO.NET

1. **Quá Trình Thiết Kế Và Thực Hiện Đồ Án:**
   1. ***Khảo sát và lấy dữ liệu Facebook Users Profile thực tế***

* Khảo sát, lấy dữ liệu các Users Profile bằng cách tạo **google form** khảo sát trực tuyến lưu vào file excel (**UsersFacebook.xlsx**) như hình sau đây: (Bước đầu nhóm khảo sát được 100 users…)

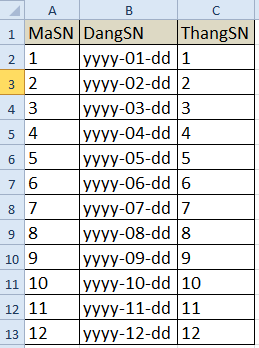


* Bước đầu nhóm thống nhất chọn **khoảng cách Euclide** để tính khoảng cách giữa hai dòng dữ liệu áp dụng vào thuật toán K-means.
* Vì dữ liệu của thuộc tính Giới Tính (**Sex**), Sinh Nhật (BirthDay), Tỉnh/Thành Phố (Hometown-Nơi sinh & Nowling-Nơi đang sinh sống và làm việc), Trạng Thái Tình Cảm (TrangThaiTinhCam) là **dữ liệu phi số** nên **cần phải được mã hóa thành số**. Nhóm thống nhất chọn cách mã hóa như các bảng tương ứng sau:
* **GioiTinh** (Mã Giới Tính, Tên Giới Tính)



* **SinhNhat** (Mã Sinh Nhật, Dạng Ngày Sinh Nhật, Tháng Sinh Nhật)

🡺Nhóm chọn cách mã hóa dữ liệu ngày sinh nhật theo số tháng

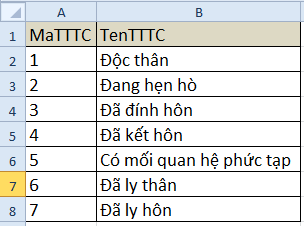


* **ThanhPho** (Mã Thành Phố, Tên Thành Phố)

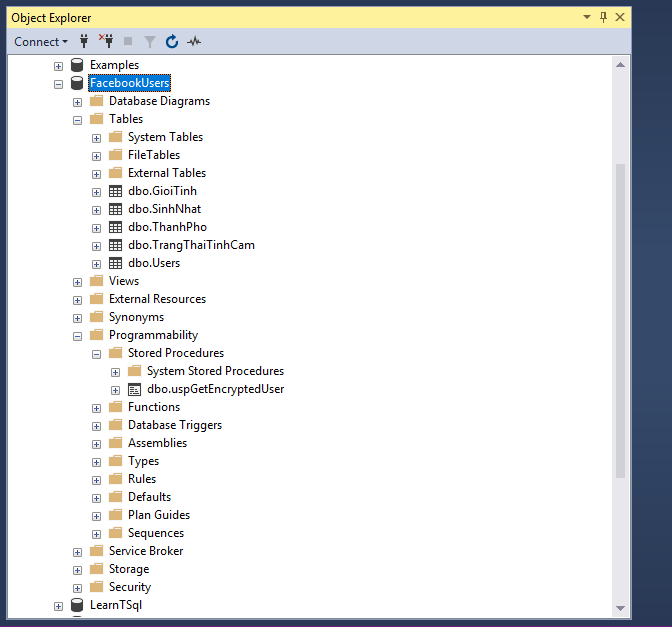
**🡺** Mã hóa cho cả Hometown và Nowliving



* **TrangThaiTinhCam** (Mã TTTC, Tên TTTC)



* 1. ***Chuẩn bị và mã hóa dữ liệu trong hệ quản trị SQL Server cho phần mềm***
* Bước tiếp theo: Import Dữ Liệu từ file Excel vào database với tên là **FacebookUsers** bằng **công cụ ImportData** phần mềm Microsoft SQL Server Management Studio, dữ liệu sau khi import như hình sau:



* Mô tả thiết kế và cơ sỡ dữ liệu trong database **FacebookUsers** gồm **5 Tables** và **1 Stored Procedure** như sau:

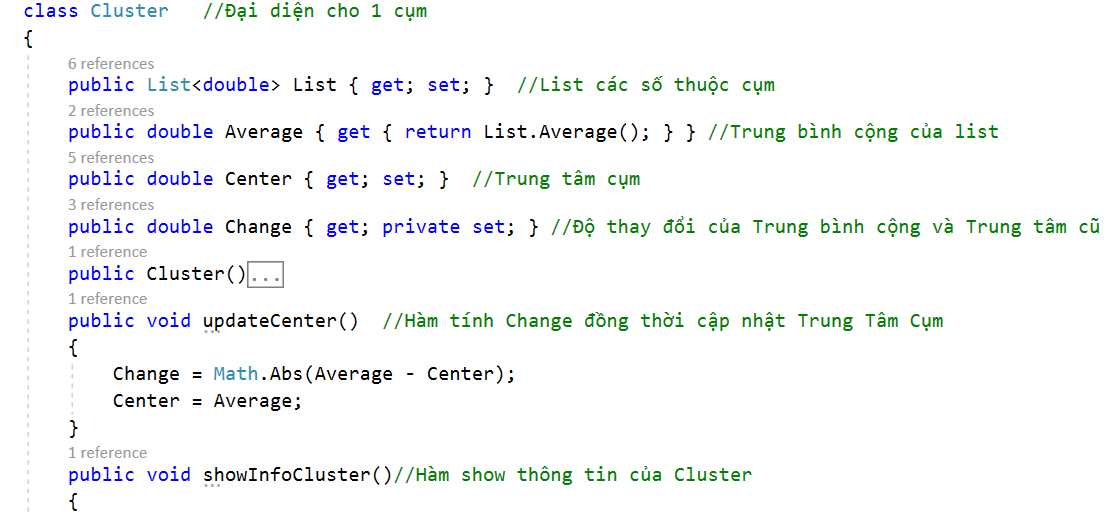
|  |  |
| --- | --- |
| Table 1:  **GioiTinh** | Design:    Records: |
| Table 2:  **SinhNhat** | Design:    Records: |
| Table 3:  **ThanhPho** | Design:    Records: |
| Table 4:  **TrangThaiTinhCam** | Design:    Records: |
| Table 5:  **Users** | Design:    Records: |
| StoredProcedure: **uspGetEncryptedUser** | Design:    **uspGetEncryptedUsers** là store procedure được nhóm tạo ra với mục đích mã hóa các dữ liệu phi số của các users thành dữ liệu số. Kết quả sau khi thực thi ta được (dữ liệu sau khi mã hóa):  UserID và UserName sẽ không tham gia vào thuật toán phân cụm nên không mã hóa! |

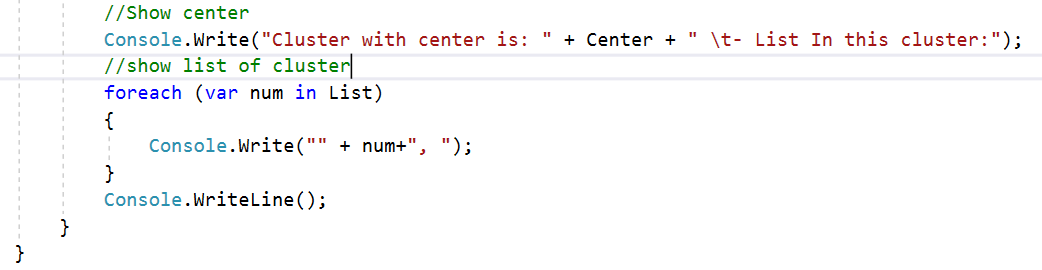
1. ***Thiết Kế Và Test Thuật Toán K-means Dùng Cho Dữ Liệu Đơn Giản***

|  |
| --- |
| **Đầu vào**: Số k cụm cần gom và cơ sở dữ liệu chứa n đối tượng.  **Đầu ra:** k cụm đã được gom.  **Thuật toán:** gồm 4 bước:   * **Bước 1:** Phân hoạch đối tượng thành k tập con ( cụm). * **Bước 2:** Tính khoảng cách giữa các đối tượng (objects) đến K tâm (dùng khoảng cách Euclide) * **Bước** **3:** Gán mỗi đối tượng cho cụm trung tâm gần nhất. * **Bước 4:** Nếu cụm không có sự thay đổi thì dừng, ngược lại quay lại bước 2 . |

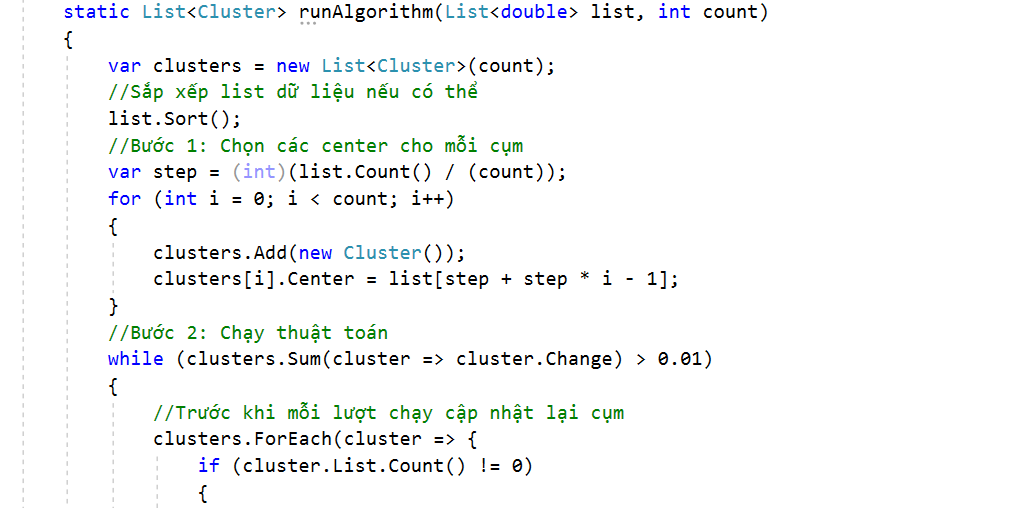
* Test Phần Code Thuật Toán Đối Với **Dữ Liệu Số Thực (Double)**

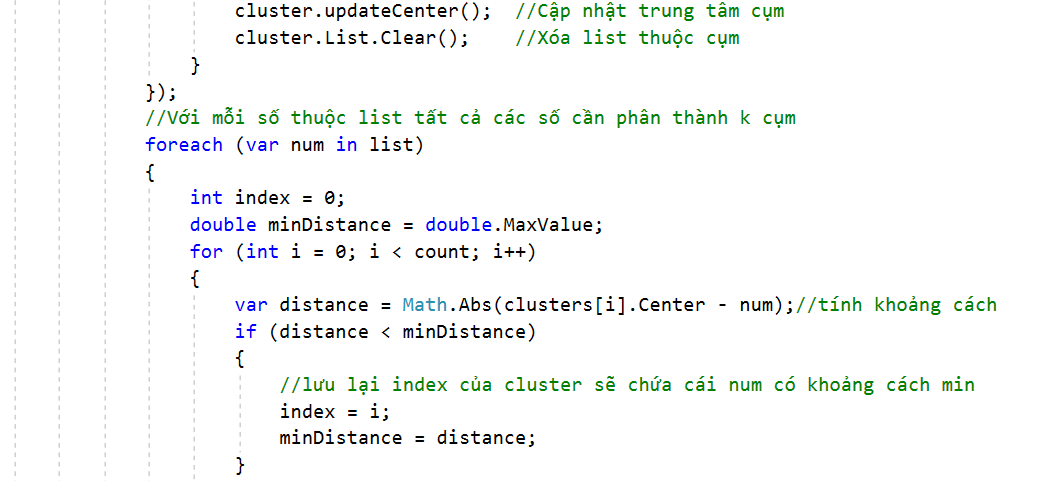
Ta có class Cluster (Đại diện cho cụm) như sau:

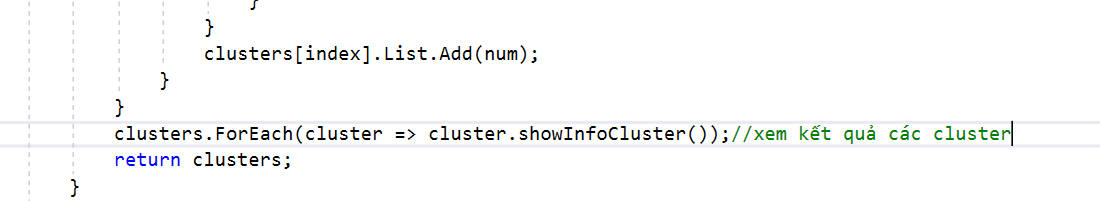




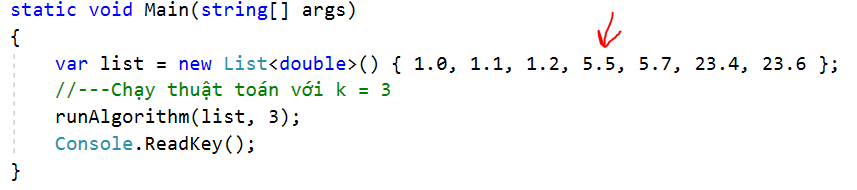
Code thuật toán được thiết kế theo giải thuật các bước trên như sau:



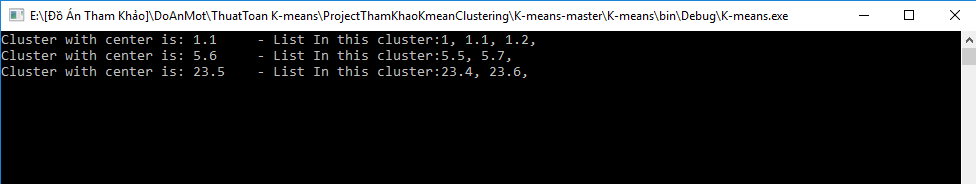




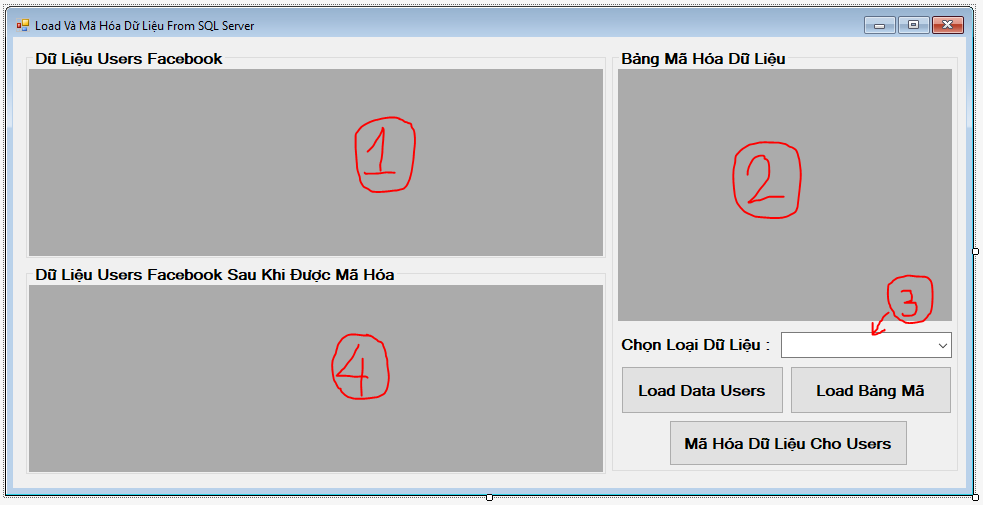
Bắt đầu test thuật toán với list các số double như sau:



Kết quả cho ra là kết quả đúng:

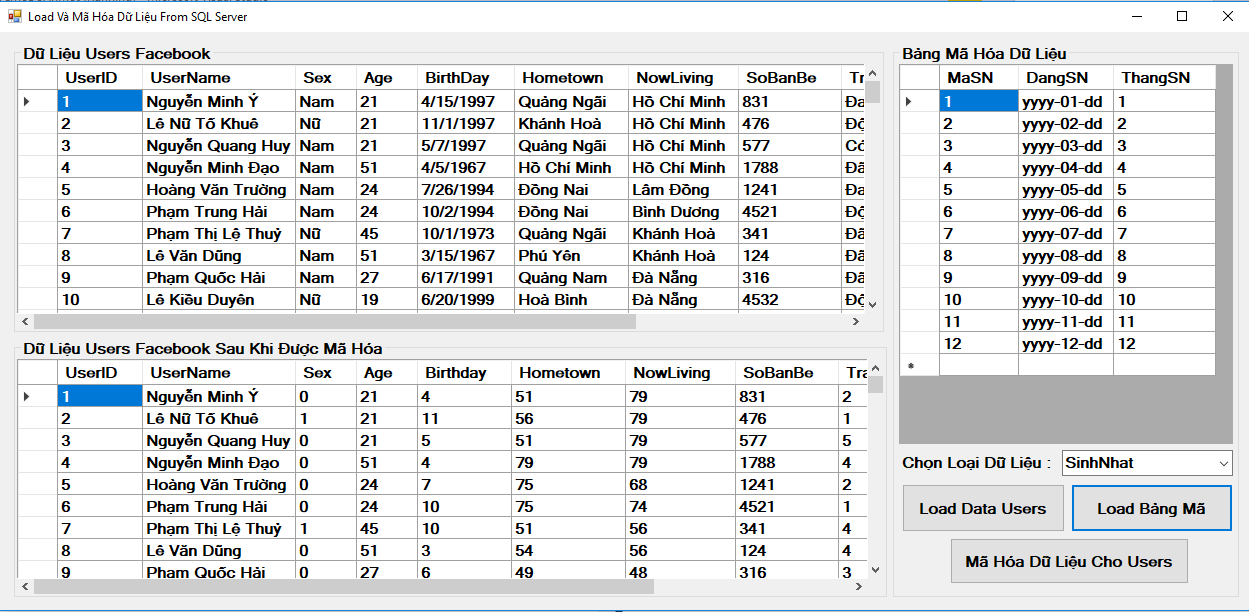


1. ***Thiết Kế Và Lập Trình Sơ Khai Cho Phần Mềm*** 
   1. **Form “Load Và Mã Hóa Dữ Liệu From SQL Server”**

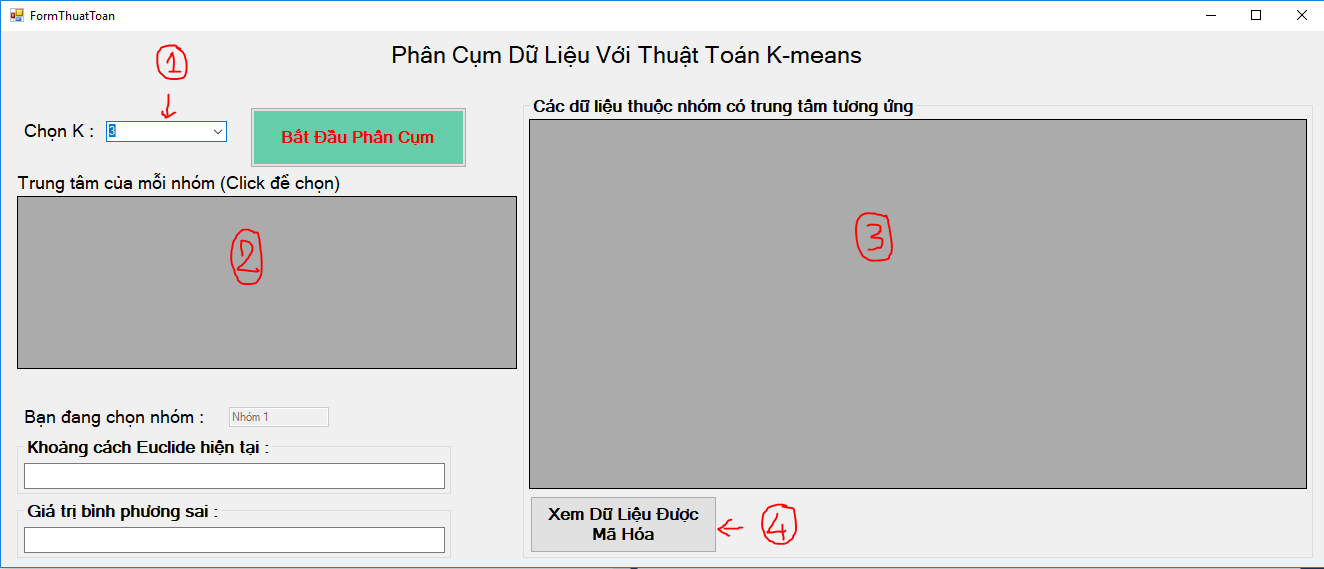


Như hình trên, Form bao gồm các thành phần chính và chức năng như sau:

* DataGirdView “Dữ Liệu Users Facebook”(Control 1): Sau khi click vào button “LoadDataUsers” thì dữ liệu Users trong database được kết nối và show dữ liệu Users lấy được ra DataGridView.
* ComboBox “Chọn Loại Dữ Liệu” (Control 3): Khi Form này được load lên thì ComboBox được add các item như: GioiTinh, ThanhPho, SinhNhat, TrangThaiTinhCam mục đích là sau khi chọn xong giá trị trong ComboBox sau đó click button “Load Bảng Mã” thì dữ liệu của bảng có tên tương ứng trong SQL Server sẽ được show lên DataGridView “Bảng Mã Hóa Dữ Liệu” (Control 2).
* DataGridView “Dữ Liệu Users Facebook Sau Khi Được Mã Hóa” (Control 4) sau khi click vào button “Mã Hóa Dữ Liệu Cho Users” thì sẽ show lên DataGridView dữ liệu users sau khi mã hóa thành số (Dùng uspGetEncryptedUsers).
* Kết quả khi Test Form:

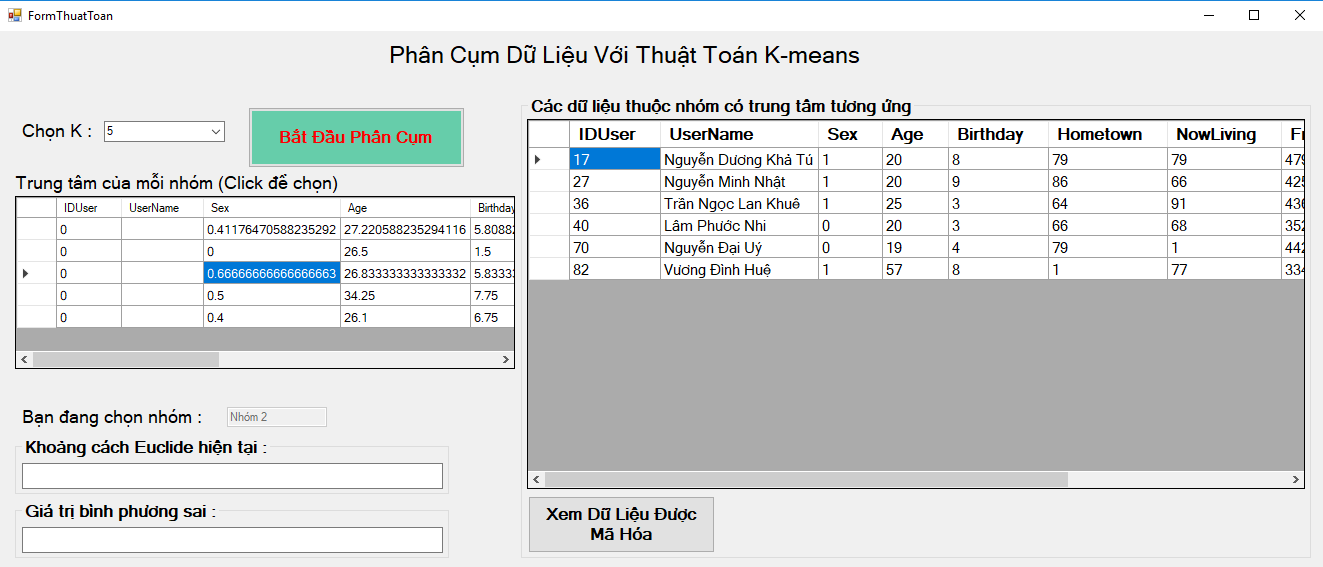


* 1. **Form Chạy Thuật Toán Phân Cụm Users Facebook**

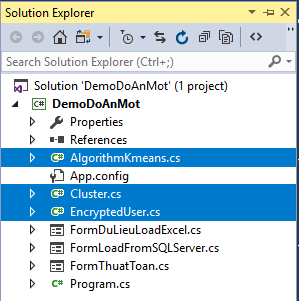


Như hình trên, Form bao gồm các thành phần chính và chức năng như sau:

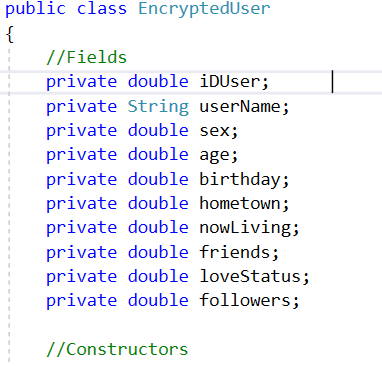
* ComboBox “Chọn K” (Control 1) sẽ có 3 giá trị K được add vào ComboBox là 3,4,5 tương ứng với số cụm ta chọn để phân ra. Giá trị mặc định là 3.
* Sau khi click vào button “Bắt Đầu Phân Cụm” thì thuật toán sẽ bắt đầu chạy với dữ liệu các users và dau khi chạy thuật toán phân cụm kết thúc k trung tâm cuối cùng của k cụm sẽ được hiển thị lên DataGridView “Trung Tâm Của Mỗi Nhóm”(Control 2)
* Khi click vào một record trung tâm trên DataGridView “Trung Tâm Của Mỗi Nhóm”(Control 2) thì các records users thuộc cụm chứa trung tâm đó sẽ được show lên trên DataGridView “Các dữ liệu thuộc nhóm có trung tâm tương ứng” (Control 3)
* Button “Xem dữ liệu được mã hóa”(Control 4) có chức năng là khi click vào sẽ hiện Form Load Dữ Liệu Từ SQL Server mà nhóm đã đề cập ở mục 5.1 trước đó.
* Kết quả khi test Form:



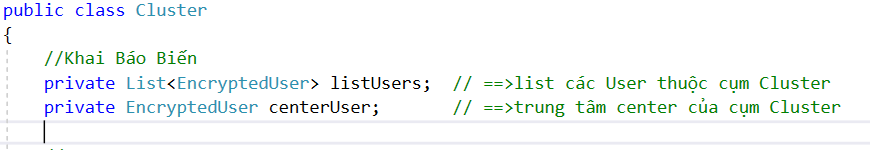
* 1. **Thiết Kế Thuật Toán Phân Cụm Cho Facebook Users**



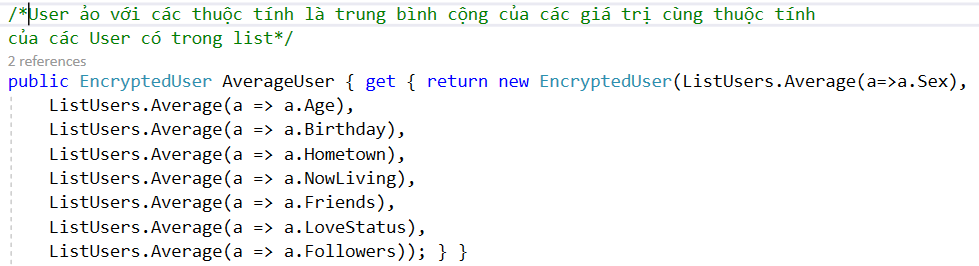
* + 1. Định nghĩa Class **EncryptedUser.cs (User Được Mã Hóa)**
* Class [EncryptedUser.cs] đại diện cho một user đã được mã hóa dữ liệu phi số thành dữ liệu số. (gồm tất cả các thuộc tính tương ứng các cột với table **Users** trong database **FacebookUsers**)
* Các fields của Class [EncryptedUser.cs] như sau:



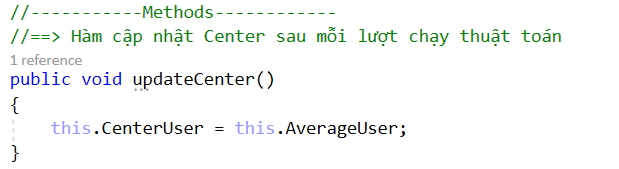
* + 1. Định nghĩa Class **Cluster.cs (Cụm)**
* Class [Cluster.cs] đại diện cho một đối tượng cụm(nhóm) sau khi phân cụm
* Class [Cluster.cs] sẽ có 2 thuộc tính là list các users thuộc cụm và trung tâm của cụm



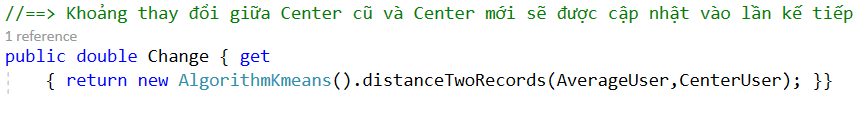
* Class này có 1 biến đặc biệt **AverageUser** khi gọi sẽ trả về 1 user ảo có giá trị thuộc tính bằng trung bình cộng của mỗi thuộc tính của các dữ liệu thuộc cụm(Đối tượng này sẽ là 1 user ảo vì nó không tồn tại trong dữ liệu nên không có UserID và UserName). Đối tượng này sẽ được cập nhật mỗi khi vòng lặp của thuật toán được thực hiện.



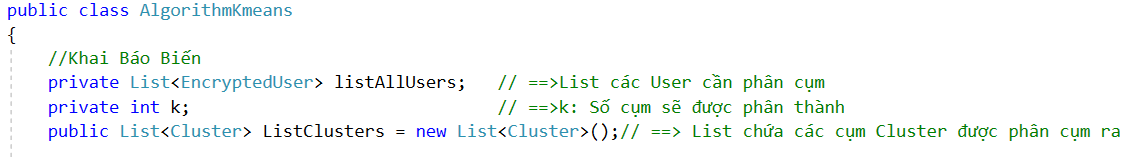
* Class có 1 method **updateCenter** sẽ cập nhật Center sau mỗi lượt chạy thuật toán



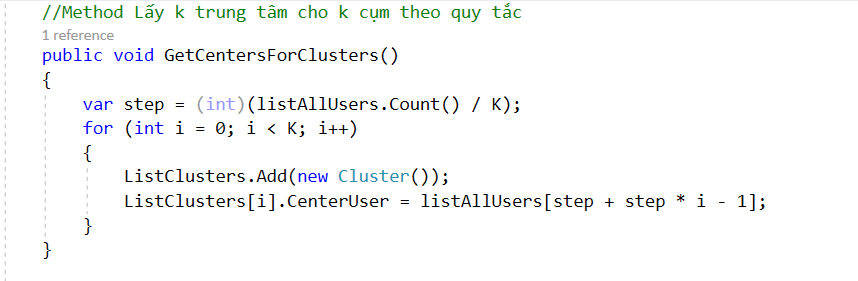
* Class này có biến đặc biệt nữa gọi là biến **Change** sẽ trả về khoảng cách của AverageUser và trung tâm cụm khi chưa được update. Ta sẽ dựa vào biến **Change** này để xét xem khi nào vòng lặp thuật toán K-mean dừng. Theo lý thuyết thì biến Change bằng 0 tức là giữa trung tâm cũ của cụm và trung bình của cụm AverageUser không có sự chệnh lệch nữa thì thuật toán sẽ dừng.



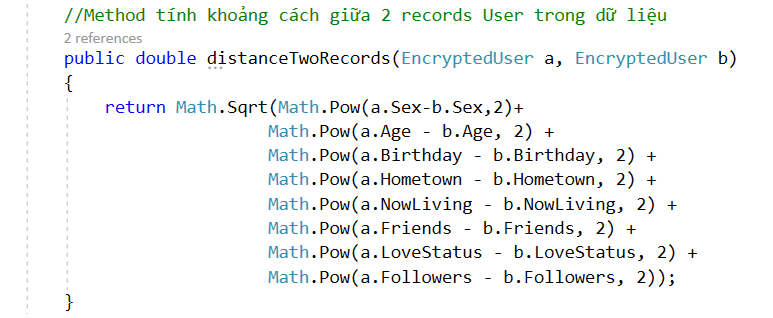
* distanceTwoRecords(…) là thuật toán được định nghĩa trong class AlgorithmKmeans được giới thiệu ngay sau đây.
  + 1. Định nghĩa Class **AlgorithmKmeans.cs (Thuật Toán K-means)**
* Class này chứa các thuộc tính như chú thích sau:



* Method lấy trung tâm cho k-cụm theo quy tắc hàng rào, cứ cách một “step” các record thì ta lấy 1 record user làm trung tâm đầu tiên của cụm



* Method tính khoảng cách Euclide của 2 EcryptedUser (User đã được mã hóa)



* Method cuối cùng cũng là method quan trọng nhất chạy thuật toán K-means

