LỜI NÓI ĐẦU

Phân cụm là kỹ thuật rất quan trọng trong khai phá dữ liệu, nó thuộc lớp các phương pháp ***Unsupervised Learning*** (học không giám sát) trong Machine Learning. Trong thuật toán này, ta không biết được nhãn của dữ liệu mà chỉ có dữ liệu đầu vào. Thuật toán unsupervised learning sẽ dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó, ví dụ như phân cụm (clustering) hoặc giảm số chiều của dữ liệu (dimension reduction) để dễ dàng lưu trữ và tính toán dữ liệu. Phân cụm là quá trình nhóm các điểm dữ liệu trong cơ sở dữ liệu thành các cụm sao cho những điểm dữ liệu trong cùng một cụm có độ tương đồng lớn và những điểm không cùng một cụm có sự tương đồng là rất nhỏ.

K-Means là thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật phân cụm. Tư tưởng chính của thuật toán K-Means là tìm cách phân nhóm các đối tượng đã cho vào k cụm sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm là nhỏ nhất.

Báo cáo này gồm 2 chương, trong đó chương 1 nghiên cứu tổng quan về bài toán phân cụm, một số độ đo thường được sử dụng trong phân cụm đó là khoảng cách Euclide và khoảng cách Manhattan. Cũng ở chương này chúng em trình bày một số phương pháp phổ biến trong phân cụm đó là phương pháp phân hoạch, phương pháp phân cấp. Cụ thể hơn ở chương 2 chúng em đi nghiên cứu về thuật toán K-Means, đây là một trong những thuật toán thuộc phương pháp phân hoạch.

Trong chương 2, chúng em nghiên cứu về thuật toán K-Means và ứng dụng của nó. Trong thuật toán K-Means việc xác định số cụm K là rất khó, do đó chúng em đã tìm hiểu và nghiên cứu phương pháp Elbow, đây là phương pháp giúp xác định số cụm K, sao cho phân cụm đạt kết quả tốt. Từ đó chúng em áp dụng thuật toán K-Means kết hợp với sử dụng thuật toán Elbow vào bài toán phân vùng ảnh. Kết quả thu được thông qua ngôn ngữ lập trình Python.