Assignment 4: Clustering with a HuggingFace Dataset

Keith Tran 300240028

The dataset chosen for this assignment is rahulvyasm/medical_insurance_data from HuggingFace. It contains the medical insurance data of almost one thousand individuals. This dataset is great for clustering because 1) it contains a mix of numerical data (age, bmi, number of children, charges) and categorical data (sex, smoker or non-smoker, and region). K-Means clustering can handle different types of data. In this case, the numerical features were scaled, and the categorical features were one-hot encoded, allowing the clustering algorithm to measure the distances between them effectively. The data also lends itself to natural groupings often used by insurance agencies to create customer risk profiles and cost patterns.

The elbow method was performed on the data to find the optimal number of clusters to group the data points into (represented as k). A graph is produced, plotting the within-cluster sum of squares (WCSS) against the number of clusters. The produced curve resembles an "arm," and the point at the "elbow" represents the optimal number of clusters. It is determined from this that the optimal k is 4. After this, a Random Forest model is used to predict the clusters. The four clusters can be interpreted as the following: 1. Young, healthy, low-cost non-smokers; 2. Middle-aged families with moderate costs; 3. Young-to-middle-aged smokers with above-average costs; and 4: Older, high-risk, high-cost smokers. An insurance company may use these clusters to determine pricing, marketing, and risk management for a certain client.

Overall, the model performed very well in its predictions, with an overall accuracy of 0.9856 and average f1-score of 0.99. Visually speaking from the plotted graph, the distinct clusters are easy to identify, with slight overlap between them.