**Research Assignment 1: Unsupervised Learning and Clustering**

This paper explores the difference between supervised and unsupervised machine learning, the concept of clustering and its implementation and the importance of standardization in data science.

According to Jones et al. (2020, Introduction to Clustering), supervised learning is a machine learning technique that detects meaningful information from data already labeled and structured in some manner. The techniques used by supervised learning help predict how future data should be classified using past data that is already classified. For example, supervised learning techniques can teach a model to identify spam emails by providing the model with a large data set of emails labeled spam and non-spam. Once trained, programmers evaluate the model's ability to identify spam emails from a different data set, re-train it with more classified data, and eventually deploy it for use cases where spam email detection is needed.

By contrast, unsupervised learning is a method in machine learning that seeks out structure from unlabeled data by using various techniques to develop meaningful information despite a lack of classification (Jones et al., 2020, Introduction to Clustering). An example of unsupervised learning would be giving a model a list of invoices and asking the model to seek out potential relationships. The model might then find correlations between customer purchasing behavior and other factors such as time of year, payment methods, or store location. Unsupervised learning is distinct from supervised learning because the unsupervised model is not provided with data classifications and instead uses mathematical formulas to generate insights from the data (Jones et al., 2020, Introduction to Clustering).

One statistical approach an unsupervised learning model uses is clustering, which identifies groups of similar data within a dataset. Clustering is an important process in data science because large datasets with little to no structure can use clustering to identify where there are relationships that are not readily apparent from a tabular format (Jones et. al, 2020, Introduction to Clustering). These relationships can then be classified as subgroups and used to inform data analysis processes. There are multiple types of clustering algorithms, two of which are k-means and hierarchical clustering. K-means clustering inputs a set of data and the number of clusters, identified as k, and attempts to find the mean point of k different clusters using an iterative process of adding data to clusters and finding the mean of those clusters’ sets (Jones et al., 2020, Introduction to Clustering). An example use of the k-means clustering algorithm would be a popular restaurant looking to expand its locations across a metropolitan area. If the restaurant had a dataset containing the home addresses of everyone who ordered from their store in the past and the financing to lease three more locations, they could use the k-means clustering algorithm to find clusters of their customers across the city and use that to inform their search for the ideal property (Serrano, 2019).

Hierarchical clustering is a technique used to build tree data structures that model the relationships between data points. Agglomerative hierarchical clustering builds the tree model from the bottom up, taking the distances between data points and constructing a dendrogram. The programmer or model then labels the clusters according to the number of related data points below a certain distance threshold. Divisive hierarchical clustering takes a top-to-bottom approach, treating every data point as a single cluster and then seeking out differentiating values to split the data into smaller subgroups (Jones et al., 2020, Hierarchical Clustering).

The best use of k-means or hierarchical clustering algorithms depends on the problem and the data's complexity. One downside of the k-means clustering algorithm is its reliance on an input value for k, the number of clusters. On the other hand, hierarchical clustering requires human input when distinguishing the distance cutoff for the relationships between clusters. If the number of clusters is already known, such as in the location selection example, k-means may provide the most efficient solution. If the number of clusters is not already known, for example, if the problem requires that the model potential types of food that appear in a grocery store using only caloric content and date of expiration, hierarchical clustering algorithms may perform better at developing a more accurate number of clusters (Jones et al., 2020, Hierarchical Clustering).

Unsupervised learning models depend on clean and well-formatted data to provide valuable results. One approach that helps with this process is data standardization. This preprocessing technique reformats an input into a standardized range with a mean of zero and a standard deviation of one (Jaadi, 2020). Standardization makes the data more useful for clustering, as some data used to calculate the distance between values will have different scales. For example, suppose the input dataset contains height and weight values. Given the values used to measure height and weight, the clustering algorithm will have a much larger value for one axis than the other, which makes the relationship between the two values harder to understand. By standardizing the values for weight and height with a mean of zero and a standard deviation of one, the graph of their relationship is easier to use and understand in the context of clustering (Jaadi, 2020). Another approach to reformatting the inputs to the model is data normalization, which converts input values to a constrained scale of zero to one or negative one-to-one. Unlike standardization, the upper and lower constraints can cause outliers to skew the other values in the dataset. Standardization is best in cases where the distribution of values is known to be close to normal, where normalization is likely better to be used when a distribution is not normal or unknown (Jaadi, 2020).

**References**

Jaadi, Z. (2020, December 14) *A simple guide on when and why it is necessary to standardize data*. Built In. https://builtin.com/data-science/when-and-why-standardize-your-data

Jones, A., Kruger, C., & Johnston, B. (2020). *The Unsupervised Learning Workshop*. Packt Publishing. https://learning.oreilly.com/library/view/the-unsupervised-learning/9781800200708/B15923\_02\_Final\_SMP.xhtml#\_idParaDest-49

Serrano, L. (2019, January 28). *Clustering: K-means and Hierarchical.* YouTube. https://www.youtube.com/watch?v=QXOkPvFM6NU