**Performance Assessment**

OFM4 — OFM4 Task 1: Clustering Techniques

Bader Ale

Department of Information Technology, Western Governors University

D212 Data Mining II

2024

Table of Contents

[Part I: Research Question 3](#_Toc179017286)

[Part II: Technique Justification 3](#_Toc179017287)

[Part III: Data Preparation 4](#_Toc179017288)

[Part IV: Analysis 4](#_Toc179017289)

[Part V: Data Summary and Implications 4](#_Toc179017290)

# Part I: Research Question

The purpose of the analysis using k-means clustering is to answer the following “*In the context of a major hospital chain, "Can we identify distinct patient clusters based on demographic and clinical factors that are more likely to experience hospital readmission within 30 days?".*  For this task, k-means clustering was used on the given medical data set. This unsupervised learning algorithm was used to group patients according to the following common features: Age, Vitamin D levels during hospital stay and Readmission status. These clusters can help medical administrators understand patterns in these clusters that show them a high risk for readmission, which would form a proper basis on which effective targeted management interventions could be based.

# Part II: Technique Justification

The K-means algorithm is a kind of partitioning method where clusters are formed by the similarities of data points, in this case, patients, based on some predefined features such as age, comorbidities, length of stay, and location of discharge. First, the algorithm randomly picks some centroids and then assigns patients to the nearest centroid according to Euclidean distance. The centroids are recalculated when all the patients have been assigned to their respective clusters. The process repeats until the centroids no longer move significantly-when, in other words, the clusters are stable (scikit-learn developers, n.d.).

For this this k-means model, three variables were chosen for the clusters: Age, Vitamin D levels, and the target variable ReAdmission. The purpose of this analysis was to group patients into well-defined clusters that possess similar characteristics, hence allowing an analysis of interactions and influences these factors have on readmission.  
The k-means algorithm represents each patient as a data point in three-dimensional space, whose dimensions are Age, Vitamin D levels, and ReAdmission status. In an iterative process, it assigns each patient to a cluster by computing a distance measure between the data point representing a patient and the centroid of each cluster. It starts with the centroids chosen arbitrarily, but in every subsequent step, the centroid is updated to actually represent an 'mean' value of patients for a cluster. This process is repeated until convergence occurs; the centroids are not changing significantly anymore.We would expect the result of this to be the clear formation of patient clusters that reflect distinct relationships between Age, Vitamin D levels, and ReAdmission. For instance, one cluster may consist of a higher group of aged patients with low Vitamin D levels who are more predisposed to readmission, while another cluster can represent the younger patients who have higher Vitamin D levels and thus will present less likelihood of readmission. The information shall allow the administrators to understand how such factors as age and Vitamin D level result in readmissions to elaborate more focused preventive measures for high-risk groups.

One of the important assumptions of the k-means clustering algorithm is that clusters are spherical and equally sized. What this means is that the algorithm takes on the assumption that data points inside each cluster are evenly distributed around a central point, called a centroid, and the distance between points and their centroid is relatively uniform across all clusters. This works well when the clusters of data are near symmetric. The algorithm can sometimes fail to capture the exact underlying structure of the data set for oddly shaped data or of varying cluster size (scikit-learn developers, n.d.).

# Part III: Data Preparation

# Part IV: Analysis

# Part V: Data Summary and Implications

Works Cited

scikit-learn developers. (n.d.). *2.3. Clustering*. Retrieved October 2024, from Unsupervised Learning: https://scikit-learn.org/stable/modules/clustering.html#k-means