**D209 Performance Assessment Task 1**

**CLASSIFICATION ANALYSIS FOR MEDICAL DATA**

Fahim A. Akbar Student ID 001434895 Masters Data Analytics (January 1, 2021) Program Mentor: Lea Yoakem (877) 435-7948x6422 [fakbar3@wgu.edu](mailto:fakbar3@wgu.edu)

**Part I: Research Question**

**A. Purpose of the Data Mining Report**

**1. Question**

Principal Component Analysis using the k-means clustering technique could be used to help answer the following: “Which patients are at a high risk of being readmitted?”

**2. Goal**

The goal of the data analysis is to accurately showcase which patients are at risk of readmission. An accuracy measure will be provided to ensure that the hospital leadership can be confident in the data. Weights for variables that the hospital should keep track of will be used to help the hospital leadership reduce readmissions.

3. **Part II: Method Justification**

**B. Reason for using *k*-means**

**1. Classification Method: *k*-means**

For this analysis, the *k*-means clustering method will be used.

The K-Nearest Neighbor (KNN) method algorithm operates by assuming that similar things exist in close proximity. KNN finds the distances between a query and all the examples in the data, selecting the specified number examples (K) closest to the query, then votes for the most frequent label, or averages the labels. The “k” value is an arbitrary numeric value that is determined by the user, and it will drive the algorithm to look at that set number of data points around the data point in question to determine its label. The algorithm carries out by setting decision boundaries in the data points, and these decision boundaries are determined by continuous variables. The data point of interest is labeled according to where it falls within those decision boundaries. For this analysis, the algorithm will be conditioned via a training set of continuous variables and match that to the outcome of either readmitted or not. Once the algorithm can determine which continuous variables classify a patient as readmitted, the model performance will then be determined on the test data set. From this we can derive a model accuracy score. The goal of this analysis is to create an algorithm that will have a model accuracy of at least 95% without over or under fitting the model. Additional expected outcomes include test data being classified according to their closest neighbors (Harrison, 2019).

**2. Assumptions of *k*-means**The main assumption of the *k*-means method is that similar things will exist in close proximity to one another. This means that an unlabeled data point will exist within close proximity to a similar labeled data point, and can be classified based on the similarity in features. (Harrison, 2019).

**3. Python packages and libraries used**

Listed below are the Python packages and libraries that will be used, and how each item supports *k*-means clustering analysis:

* Pandas – this standard import provides methods to read and visualize data. It also offers statistical tools to parse and score data.
* Numpy – this standard import provides methods to read and visualize data. It also offers statistical tools to parse and score data.
* Matplotlib – this package is used for data visualization and will provide more robust tools to visualize reports and data points
* Seaborn – this package will provide us descriptive and visually intuitive graphs, plots, and matrices
* Scikit-learn – this package will provides method and arguments for splitting, training, testing, and fitting data. This package also has arguments for predicting and classifying data as well as applying metrics for models

The code used to download these packages is provided in the “Python Package Code” document attached to this task submission.

**Part III: Data Preparation**

**C. Data preparation**

**1. Relevant data preprocessing goal**

The first step in preparing the data is to make sure that there are no missing data entries in any of the columns. Next, we will ensure that there are no duplicated columns or rows to further prevent dealing with repeated entries. For the predictive analysis, several columns in the dataset were deemed irrelevant and were subsequently dropped from the dataset (i.e latitude, longitude). The “yes/no” entries for the categorical variables will need to be converted to 1 and 0, respectively.

>Describe ONE data preprocessing goal relevant to the clustering technique from part A1.