Saemi Ramirez

011926418

D212 PA 2 Dimensionality Reduction Methods

9/28/2024

WGU

1. Research Question
   1. Research question and clustering technique  
      My research question is “Can the number of dimensions of the WGU medical dataset be reduced by using PCA (Principal Component Analysis)?”
   2. Define a goal  
      The goal of the dimension reduction on the dataset using PCA is to remove unnecessary columns and make the dataset more efficient for the future analysis.
2. Reasons of chosen PCA
   1. How PCA analyzes the selected dataset & Expected outcome
      1. Standardize the continuous variables in the dataset
      2. Identify the correlations between the variables using covariance matrix
      3. Identify the principal components by computing the eigenvectors/eigenvalues of the covariance matrix
      4. Decide how many principal components to maintain by creating feature vector
      5. Recast the data along the principal component axes (Jaadi)
      6. Expected outcome is to reduce the features of the dataset from the Kaiser criterion method to get the best number of important principal components.
   2. One assumption of PCA  
      PCA operates on the assumption that the principal components with higher variance are the most significant, while those with lower variance are treated as noise and often disregarded. The development of PCA was influenced by the Pearson correlation coefficient framework, where it was initially assumed that only the axes with high variance would be transformed into principal components. (Vadapalli)
3. Perform data preparation
   1. Identify the continuous data set variables

The following variables are all numeric: Population, Children, Age, Income, VitD\_levels, Doc\_visits, Full\_meals\_eaten, vitD\_supp, Initial\_days, TotalCharge, and Additional\_charges

* + 1. First, the object columns were removed and printed the remaining columns  
       A screenshot of a computer code

       Description automatically generated
    2. Removed additional columns that are not relevant for the future analysis such as Caseorder, Zip, Lat, Lng, Item1, Item2, Item3, Item4, Item5, Item6, Item7, and Item8  
       A white background with text

       Description automatically generated
    3. The dataset remained with 11 columns including Population, Children, Age, Income, VitD\_levels, Doc\_visits, Full\_meals\_eaten, vitD\_supp, Initial\_days, TotalCharge, and Additional\_charges.  
       A screenshot of a computer code

       Description automatically generated
  1. Standardize the continuous dataset variables
     1. Called StandardScaler and used fit\_transform function to normalize the dataset, then added the column names  
        A screenshot of a computer

        Description automatically generated
  2. Provide a copy of the cleaned data set

‘scaled\_df.csv’ is submitted

1. Perform PCA
   1. Matrix of all the principal components
      1. Called PCA and used fit\_transform function with scaled dataset, then generated the matrix using the DataFrame with pca.components\_, column names, and index from PC1 to PC11  
         A screenshot of a computer

         Description automatically generated
   2. Total number of principal components
      1. Kaiser criterion method was used to reduce the dimensions of the scaled medical dataset
      2. calculated the variance of 11 columns using pca.explained\_variance\_  
         A screenshot of a computer

         Description automatically generated
      3. Graphed Kaiser criterion scree plot  
         A graph with a line

         Description automatically generated
      4. I picked 5 as a number of components from the scree plot
   3. Variance of each of the principal components

Following are the variances of the 5 principal components  
A screenshot of a computer code

Description automatically generated  
The explained variance ratio is shows as following  
A close-up of a number

Description automatically generated

* 1. Total variance

Total variance is approximately 62%  
A screenshot of a computer

Description automatically generated

* 1. Summarize the results  
     I started with 11 continuous variables and 11 PCs for PCA. After using the Kaiser criterion method, I decided to have 5 principal components which represent approximately 62% of the original dataset. 5 Principal components were selected from the Kaiser criterion method because it was important it if it was greater than 1 which is the standard deviation after normalizing the data

1. Panopto Link: <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=091a5e77-5480-4520-b5d1-b1f900f5b62c>
2. Third-party code to support the analysis

N/A

1. Third-party in-text citations and references
   1. Jaadi, Zakaria. *Principal Components Analysis (PCA): A Step-by-Step Explanation*. Builtin. (February 23, 2024). <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>.
   2. Vadapalli, Pavan. *PCA in Machine Learning: Assumptions, Steps to Apply & Applications.* upGrad. (February 15, 2024). <https://www.upgrad.com/blog/pca-in-machine-learning>.